SIAM J. OPTIM.
Vol. 13, No. 1, pp. 1–23

SEMIDEFINITE PROGRAMS: NEW SEARCH DIRECTIONS, SMOOTHING-TYPE METHODS, AND NUMERICAL RESULTS*

CHRISTIAN KANZOW† AND CHRISTIAN NAGEL†

Abstract. Motivated by some results for linear programs and complementarity problems, this paper gives some new characterizations of the central path conditions for semidefinite programs. Exploiting these characterizations, some smoothing-type methods for the solution of semidefinite programs are derived. The search directions generated by these methods are automatically symmetric, and the overall methods are shown to be globally and locally superlinearly convergent under suitable assumptions. Some numerical results are also included which indicate that the proposed methods are very promising and comparable to several interior-point methods. Moreover, the current method seems to be superior to the smoothing method recently proposed by Chen and Tseng [*Non-interior continuation methods for solving semidefinite complementarity problems*, Technical report, Department of Mathematics, University of Washington, Seattle, 1999].

Key words. semidefinite programs, smoothing-type methods, Newton's method, global convergence, superlinear convergence

AMS subject classifications. 90C22, 90C46

PII. S1052623401390525

1. Introduction.

In this paper we describe an algorithm for the solution of semidefinite programs (SDPs). Using some standard notation that will be defined formally at the end of this section, a semidefinite program is a constrained optimization problem that is typically given in primal form by

$$\min C \bullet X \quad \text{subject to (s.t.)} \quad A_i \bullet X = b_i, \quad i = 1, \ldots, m, \; X \succeq 0, \tag{1.1}$$

or in its dual form by

$$\max b^T \lambda \quad \text{s.t.} \quad \sum_{i=1}^{m} \lambda_i A_i + S = C, \quad S \succeq 0; \tag{1.2}$$

here, the vector $b \in \mathbb{R}^m$ as well as the symmetric matrices $C \in \mathbb{R}^{n \times n}$ and $A_i \in \mathbb{R}^{n \times n}$ $(i = 1, \ldots, m)$ are the given data, whereas the symmetric matrix $X \in \mathbb{R}^{n \times n}$ denotes the variable for the primal SDP (1.1), and the vector $\lambda \in \mathbb{R}^m$ together with the symmetric matrix $S \in \mathbb{R}^{n \times n}$ denote the variables of the dual SDP (1.2).

It is easy to see that the (primal) SDP is a convex minimization problem. Under a suitable constraint qualification, this SDP is therefore equivalent to its optimality conditions. These optimality conditions can be written as follows:

$$\begin{array}{c} \sum_{i=1}^{m} \lambda_i A_i + S = C, \\ A_i \bullet X = b_i \qquad \forall i = 1, \ldots, m, \\ X \succeq 0, \quad S \succeq 0, \quad XS = 0. \end{array} \tag{1.3}$$

Motivated by the groundbreaking work of Nesterov and Nemirovskii [24], several authors suggest solving the optimality conditions (1.3) by (primal-dual) interior-point

*Received by the editors June 5, 2001; accepted for publication (in revised form) December 21, 2001; published electronically May 15, 2002.
http://www.siam.org/journals/siopt/13-1/39052.html

†Institute of Applied Mathematics and Statistics, University of Würzburg, Am Hubland, 97074 Würzburg, Germany (kanzow@mathematik.uni-wuerzburg.de, nagel@mathematik.uni-wuerzburg.de).

methods. These interior-point methods typically consider the following perturbation of the optimality conditions (1.3), usually called the central path conditions:

$$\begin{array}{l} \sum_{i=1}^m \lambda_i A_i + S = C, \\ A_i \bullet X = b_i \qquad \forall i = 1, \ldots, m, \\ X \succ 0, \quad S \succ 0, \quad XS = \tau^2 I, \end{array} \tag{1.4}$$

where τ denotes a positive parameter. (Note that we parameterize the central path conditions by τ^2 instead of τ.) Typical interior-point methods now apply a Newton-type method to (a symmetrized version of) the equations within the central path conditions, dealing with the $X \succ 0$ and $S \succ 0$ constraints explicitly by a suitable line search. The interested reader is referred to [15, 2, 28, 33], for example.

The method to be discussed here is also a Newton-type method. However, before applying Newton's method, we first reformulate the optimality conditions or the central path conditions as a nonlinear system of equations. This reformulated system does not contain any explicit inequality constraints like $X \succeq 0, S \succeq 0$ or $X \succ 0, S \succ 0$, and Newton's method applied to this system automatically generates symmetric search directions without any further transformations (unlike interior-point methods).

We believe that our method is of particular interest for the solution of some difficult combinatorial optimization problems. In fact, SDPs are known to provide very good lower bounds for some combinatorial problems. However, solving such a semidefinite relaxation by an interior-point method within a branch-and-bound strategy may not result in the most efficient way to solve the underlying combinatorial problem, since the solution of one semidefinite relaxation may not be used as a starting point for a neighboring problem because interior-point methods require strictly feasible starting points. On the other hand, the method to be presented here does not have such a restriction regarding its starting point.

Our method may be viewed as a generalization of some smoothing-type methods for linear programs and complementarity problems to the framework of SDPs. While such a generalization has already been suggested in a recent paper by Chen and Tseng [8], we stress that there are differences between that paper and ours. For example, we present a new characterization of the central path conditions which may be viewed as the basis for our method. Furthermore, our method is based on an essentially smooth reformulation of the optimality conditions (1.3) themselves (and this is what we really want to solve), while Chen and Tseng [8] consider a reformulation of the central path conditions. This may also explain why our approach seems to give better numerical results than the one from [8].

The organization of this paper is as follows. Section 2 contains some new characterizations of the central path conditions (1.4). These characterizations are based on a certain function ϕ, whose further properties are discussed in section 3. Our algorithm is described in section 4, and its global and local convergence properties are analyzed in section 5. We then present some very promising numerical results in section 6 and close this manuscript with some final remarks in section 7.

Throughout this paper, we use the following notation: For two matrices $A, B \in \mathbb{R}^{n\times n}$, we define the scalar product $A \bullet B := \langle A, B\rangle := \mathrm{tr}(AB^T)$, where $\mathrm{tr}(C) := \sum_{i=1}^n c_{ii}$ denotes the trace of a matrix $C \in \mathbb{R}^{n\times n}$. (Warning: The related symbol $\circ$ is used for the composition of two mappings; it does not denote the Hadamard product of two matrices!) We denote by $\mathcal{S}^{n\times n}, \mathcal{S}_+^{n\times n}$, and $\mathcal{S}_{++}^{n\times n}$ the sets of symmetric, symmetric positive semidefinite, and symmetric positive definite matrices, respectively, of dimension $n \times n$. We also write $A \succeq 0$ and $A \succ 0$ to indicate that A belongs to $\mathcal{S}_+^{n\times n}$

and $\mathcal{S}^{n\times n}_{++}$, respectively. Furthermore, $A \succeq B$ or $A \succ B$ means that $A - B \succeq 0$ or $A - B \succ 0$. If $A \succeq 0$, we denote by $A^{1/2}$ the unique positive semidefinite square root of A. In our analysis, we will use both the spectral norm $\|A\|_2$ and the Frobenius norm $\|A\|_F$ for a matrix $A \in \mathbb{R}^{n\times n}$. We endow the vector space $\mathbb{R}^{n\times n} \times \mathbb{R}^m \times \mathbb{R}^{n\times n}$ with the norm

$$\|\!|(X, \lambda, S)|\!\| := \sqrt{\|X\|_F^2 + \|\lambda\|_2^2 + \|S\|_F^2}.$$

We use the same symbol for the norm

$$\|\!|(X, \lambda, S, \tau)|\!\| := \sqrt{\|X\|_F^2 + \|\lambda\|_2^2 + \|S\|_F^2 + \tau^2}$$

in the vector space $\mathbb{R}^{n\times n} \times \mathbb{R}^m \times \mathbb{R}^{n\times n} \times \mathbb{R}$.

2. Reformulations of the central path. The aim of this section is to give two new reformulations of the central path conditions (1.4) for SDPs. These reformulations can be obtained by generalizing existing reformulations for linear programs and complementarity problems in a suitable way.

Before we deal with the central path conditions (1.4), however, we first consider the optimality conditions (1.3). In order to motivate our approach, let us define a mapping $\varphi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ by

$$\varphi(a, b) := a + b - \sqrt{a^2 + b^2}.$$

This mapping was introduced by Fischer [13] and is usually called the Fischer–Burmeister function. It is well known (and easy to verify) that it has the following property:

$$\varphi(a, b) = 0 \iff a \geq 0, b \geq 0, ab = 0. \tag{2.1}$$

Now let us define a mapping $\phi : \mathcal{S}^{n\times n} \times \mathcal{S}^{n\times n} \to \mathcal{S}^{n\times n}$ by

$$\phi(X, S) := X + S - (X^2 + S^2)^{1/2}, \tag{2.2}$$

which is an obvious extension of the definition of φ, with the arguments being symmetric matrices rather than two real numbers. It has been shown by Tseng [30, Lemma 6.1] that the mapping ϕ has a property similar to (2.1), namely,

$$\phi(X, S) = 0 \iff X \succeq 0, S \succeq 0, XS = 0. \tag{2.3}$$

In the following, we will include a proof for this equivalence. We stress that our proof is somewhat different from the one given by Tseng [30] and that a similar technique will later be used to prove our new characterizations of the central path conditions. To verify the equivalence (2.3), we will exploit the following simple result from Alizadeh [1, Lemma 2.9].

LEMMA 2.1. *Let $X, S \in \mathcal{S}^{n\times n}_+$ be two symmetric positive semidefinite matrices. Then $XS = 0$ if and only if $X \bullet S = 0$.*

PROPOSITION 2.2. *Let ϕ be the Fischer–Burmeister function defined in* (2.2). *Then*

$$\phi(X, S) = 0 \iff X \succeq 0, S \succeq 0, XS = 0.$$

Proof. First assume that $X \succeq 0, S \succeq 0, XS = 0$ holds. This implies $XS + SX = 0$ and therefore

$$(X + S)^2 = X^2 + S^2.$$

Using the fact that X and S are symmetric positive semidefinite, it follows that

$$X + S = (X^2 + S^2)^{1/2},$$

since the square root of a symmetric and positive semidefinite matrix is uniquely defined within the space of symmetric and positive semidefinite matrices. Obviously, this implies $\phi(X, S) = 0$.

Conversely, assume that $\phi(X, S) = 0$ holds for two symmetric matrices $X, S \in \mathcal{S}^{n\times n}$. This means that $X + S = (X^2 + S^2)^{1/2}$. Squaring both sides of this equation gives

$$X^2 + S^2 = (X + S)^2 \quad \text{and} \quad X + S \in \mathcal{S}^{n\times n}_+.$$

This is equivalent to

$$XS + SX = 0 \quad \text{and} \quad X + S \in \mathcal{S}^{n\times n}_+. \tag{2.4}$$

Let $X = Q^T DQ$, with $Q \in \mathbb{R}^{n\times n}$ orthogonal and $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, be the spectral decomposition of the symmetric matrix X. Then (2.4) can be rewritten as

$$Q^T DQS + SQ^T DQ = 0 \quad \text{and} \quad Q^T DQ + S \in \mathcal{S}^{n\times n}_+.$$

If we premultiply this equation by Q and postmultiply it by Q^T, we obtain

$$DQSQ^T + QSQ^T D = 0 \quad \text{and} \quad D + QSQ^T \in \mathcal{S}^{n\times n}_+.$$

Using the abbreviation $A := QSQ^T$, we get

$$DA + AD = 0 \quad \text{and} \quad D + A \in \mathcal{S}^{n\times n}_+. \tag{2.5}$$

Componentwise, this can be rewritten as

$$(\lambda_i + \lambda_j)a_{ij} = 0 \quad \text{and} \quad D + A \in \mathcal{S}^{n\times n}_+ \tag{2.6}$$

for all $i, j = 1, \ldots, n$. In particular, taking $i = j$, we obtain $2\lambda_i a_{ii} = 0$ and $\lambda_i + a_{ii} \geq 0$ for all $i = 1, \ldots, n$. Obviously, this implies $\lambda_i \geq 0$ for all $i = 1, \ldots, n$, which in turn means that X is positive semidefinite. Using a symmetric argument (based on a spectral decomposition of S), we see that S is also positive semidefinite.

To see that $XS = 0$, we observe that (2.4) implies $X \bullet S = \operatorname{tr}[XS] = \frac{1}{2}\operatorname{tr}[XS + SX] = 0$. In view of Lemma 2.1, we therefore have $XS = 0$. □

We now want to modify the definition of ϕ so that it can be used to characterize the central path conditions (1.4). To this end, let $\tau \geq 0$ be any nonnegative number that will be viewed as a parameter within this section. Then define $\varphi_\tau : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ by

$$\varphi_\tau(a, b) := a + b - \sqrt{a^2 + b^2 + 2\tau^2}.$$

This is the so-called smoothed Fischer–Burmeister function since it is obviously continuously differentiable for every $\tau > 0$ and since it coincides with the Fischer–Burmeister

function φ for $\tau = 0$. The mapping φ_τ was introduced in [19] and has the following interesting property:

$$\varphi_\tau(a, b) = 0 \iff a \geq 0, b \geq 0, ab = \tau^2.$$

This simple observation was made in [19], and it shows that several smoothing-type methods for linear programs and related problems are closely related to interior-point methods.

We now generalize the smoothed Fischer–Burmeister function φ_τ in an obvious way: Define $\phi_\tau : \mathcal{S}^{n\times n} \times \mathcal{S}^{n\times n} \to \mathcal{S}^{n\times n}$ by

$$\phi_\tau(X, S) := X + S - (X^2 + S^2 + 2\tau^2 I)^{1/2}. \tag{2.7}$$

Then we can state the following result.

PROPOSITION 2.3. *Let $\tau > 0$ be any positive number, and let ϕ be defined by* (2.7). *Then*

$$\phi_\tau(X, S) = 0 \iff X \succ 0, S \succ 0, XS = \tau^2 I.$$

Proof. First assume that $X \succ 0, S \succ 0, XS = \tau^2 I$ holds. This implies $XS + SX = 2\tau^2 I$ and therefore $(X + S)^2 = X^2 + S^2 + 2\tau^2 I$. Using the fact that X and S are symmetric positive definite, it follows that $X + S = (X^2 + S^2 + 2\tau^2 I)^{1/2}$. This, in turn, implies $\phi_\tau(X, S) = 0$.

Conversely, let $\phi_\tau(X, S) = 0$ for two symmetric matrices $X, S \in \mathcal{S}^{n\times n}$. This means that $X + S = (X^2 + S^2 + 2\tau^2 I)^{1/2}$. Squaring both sides of this equation gives

$$X^2 + S^2 + 2\tau^2 I = (X + S)^2 \quad \text{and} \quad X + S \in \mathcal{S}^{n\times n}_{++}.$$

This is equivalent to

$$XS + SX = 2\tau^2 I \quad \text{and} \quad X + S \in \mathcal{S}^{n\times n}_{++}. \tag{2.8}$$

Let $X = Q^T DQ$, with $Q \in \mathbb{R}^{n\times n}$ orthogonal and $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, be the spectral decomposition of the symmetric matrix X. Following the proof of Proposition 2.2 and using the abbreviation $A := QSQ^T$, we see that (2.8) can be rewritten as

$$DA + AD = 2\tau^2 I \quad \text{and} \quad D + A \in \mathcal{S}^{n\times n}_{++}. \tag{2.9}$$

Componentwise, this becomes

$$(\lambda_i + \lambda_j)a_{ij} = 2\tau^2 \delta_{ij} \quad \text{and} \quad D + A \in \mathcal{S}^{n\times n}_{++} \tag{2.10}$$

for all $i, j = 1, \ldots, n$, where δ_{ij} is the standard Kronecker symbol. In particular, taking $i = j$, we obtain $2\lambda_i a_{ii} = 2\tau^2$ and $\lambda_i + a_{ii} > 0$ for all $i = 1, \ldots, n$. Obviously, this implies $\lambda_i > 0$ for all $i = 1, \ldots, n$. Hence the symmetric matrix X is positive definite. In a similar way (using a spectral decomposition of S), we can show that S is also positive definite.

In order to verify that $XS = \tau^2 I$, we observe that (2.10) implies $a_{ij} = 0$ for all $i \neq j$, since $\lambda_i + \lambda_j > 0$ according to our previous argument. Hence A is a diagonal matrix. In particular, we therefore have $DA = AD$. Consequently, we obtain from (2.9) that $DA = \tau^2 I$. Premultiplying this equation by Q^T and postmultiplying it by Q gives $XS = Q^T DQS = Q^T DAQ = \tau^2 I$. ∎

We next want to introduce a second function with properties similar to those of the (smoothed) Fischer–Burmeister function. To this end, let

$$\varphi(a,b) := 2\min\{a,b\}$$

for $a, b \in \mathbb{R}$. For obvious reasons, this mapping is called the minimum function. It is easy to see that it satisfies the equivalence

$$\varphi(a,b) = 0 \iff a \geq 0, b \geq 0, ab = 0.$$

In order to extend its definition to the class of symmetric matrices, it is helpful to reformulate the minimum function in the following way:

$$\varphi(a,b) = 2\min\{a,b\} = a + b - |a-b| = a + b - \sqrt{(a-b)^2}.$$

Motivated by the expression on the right-hand side, we now define the function $\phi : \mathcal{S}^{n\times n} \times \mathcal{S}^{n\times n} \to \mathcal{S}^{n\times n}$ by

$$\phi(X,S) := X + S - ((X-S)^2)^{1/2}. \tag{2.11}$$

It turns out that this function shares the property (2.3) with the Fischer–Burmeister function from (2.2). This observation is similar to the one made by Tseng [30, Lemma 2.1] and can alternatively be verified by following the proof of Proposition 2.2. We skip the details here and just state the result.

PROPOSITION 2.4. *Let ϕ be the minimum function defined in* (2.11). *Then*

$$\phi(X,S) = 0 \iff X \succeq 0, S \succeq 0, XS = 0.$$

We now want to modify the definition of the minimum function in such a way that we get a characterization of the central path conditions (1.4). To this end, we first recall that there is a suitable modification of the minimum function for scalar variables, namely,

$$\varphi_\tau(a,b) := a + b - \sqrt{(a-b)^2 + 4\tau^2},$$

where τ denotes a nonnegative number. This smoothed minimum function is usually called the Chen–Harker–Kanzow–Smale smoothing function in the literature [6, 19, 25], and it was noted in [19] that it has the following property for each $\tau > 0$:

$$\varphi_\tau(a,b) = 0 \iff a > 0, b > 0, ab = \tau^2.$$

This observation motivates us to define a mapping $\phi_\tau : \mathcal{S}^{n\times n} \times \mathcal{S}^{n\times n} \to \mathcal{S}^{n\times n}$ by

$$\phi_\tau(X,S) := X + S - ((X-S)^2 + 4\tau^2 I)^{1/2}. \tag{2.12}$$

It turns out that this function has the desired property.

PROPOSITION 2.5. *Let $\tau > 0$ be any positive number, and let ϕ be defined by* (2.12). *Then*

$$\phi_\tau(X,S) = 0 \iff X \succ 0, S \succ 0, XS = \tau^2 I.$$

Proof. It is easy to see that $X \succ 0, S \succ 0, XS = \tau^2 I$ implies $\phi_\tau(X,S) = 0$. Conversely, if $\phi_\tau(X,S) = 0$ for two matrices $X, S \in \mathcal{S}^{n\times n}$, we get $X + S = ((X - S)^2 + 4\tau^2 I)^{1/2}$ and therefore $(X-S)^2 + 4\tau^2 I = (X+S)^2$ and $X + S \in \mathcal{S}^{n\times n}_{++}$. This is

equivalent to $XS+SX = 2\tau^2 I$ and $X+S \in \mathcal{S}^{n\times n}_{++}$. Hence we can follow the argument from the proof of Proposition 2.3 in order to show that $X \succ 0, S \succ 0$, and $XS = \tau^2 I$ holds. □

Let ϕ_τ denote either the smoothed Fischer–Burmeister function from (2.7) or the smoothed minimum function from (2.12). Then define a mapping $\Phi_\tau : \mathcal{S}^{n\times n} \times \mathbb{R}^m \times \mathcal{S}^{n\times n} \to \mathcal{S}^{n\times n} \times \mathbb{R}^m \times \mathcal{S}^{n\times n}$ by

$$\Phi_\tau(X,\lambda,S) := \begin{pmatrix} \sum_{i=1}^m \lambda_i A_i + S - C \\ A_i \bullet X - b_i \ \ (i=1,\ldots,m) \\ \phi_\tau(X,S) \end{pmatrix}. \tag{2.13}$$

Then Propositions 2.3 and 2.5 immediately give the following new characterization of the central path conditions (1.4) for SDPs.

THEOREM 2.6. *Let Φ_τ be defined by* (2.13)*, with ϕ given by* (2.7) *or* (2.12)*, and let $\tau > 0$. Then the following statements are equivalent:*

(a) (X,λ,S) *satisfies the central path conditions* (1.4).

(b) (X,λ,S) *is a solution of the nonlinear system of equations* $\Phi_\tau(X,\lambda,S) = 0$.

We close this section by noting that some further properties of the functions ϕ discussed in this section may be found in [7, 8, 14, 26, 27, 32].

3. Properties of ϕ. In this section we will state some properties of the functions ϕ_τ introduced in the previous section. In particular, we will show that these functions are differentiable (in the sense of Fréchet).

However, in contrast to the approach of the previous section, we will view the nonnegative number τ as an independent variable from now on. In order to make this clear in our notation, we set $\phi(X,S,\tau) := \phi_\tau(X,S)$; i.e., we now write

$$\phi(X,S,\tau) := X + S - (X^2 + S^2 + 2\tau^2 I)^{1/2} \tag{3.1}$$

for the smoothed Fischer–Burmeister function from (2.7), and

$$\phi(X,S,\tau) := X + S - ((X-S)^2 + 4\tau^2 I)^{1/2} \tag{3.2}$$

for the smoothed minimum function from (2.12). Taking τ as a variable rather than a parameter is motivated by some computational considerations and will be explained in more detail in our next section when we present our smoothing-type method for the solution of the optimality conditions (1.3).

We begin our analysis of the functions ϕ with the following two results, whose proofs can be found in [8, Lemma 1 and Corollary 1].

LEMMA 3.1. *Let ϕ denote one of the functions defined in* (3.1) *or* (3.2)*. Then, for any $X, S \in \mathcal{S}^{n\times n}$ and any $\tau > \nu > 0$, we have*

$$\begin{aligned} \kappa\,(\tau-\nu) I &\succeq \phi(X,S,\nu) - \phi(X,S,\tau) \succ 0, \\ \kappa\,\tau I &\succeq \phi(X,S,0) - \phi(X,S,\tau) \succ 0, \end{aligned}$$

where κ denotes a (known) positive constant independent of X, S, τ, and ν.

COROLLARY 3.2. *Let ϕ be given by* (3.1) *or* (3.2)*, and let κ be the constant from Lemma* 3.1*. Then the following inequalities hold:*

(a) $\|\phi(X,S,\nu) - \phi(X,S,\tau)\|_F \le \kappa\sqrt{n}(\tau-\nu) \quad \forall\, X,S \in \mathcal{S}^{n\times n}\ \forall\, \tau > \nu > 0$.

(b) $\|\phi(X,S,0) - \phi(X,S,\tau)\|_F \le \kappa\sqrt{n}\,\tau \quad \forall\, X,S \in \mathcal{S}^{n\times n}\ \forall\, \tau > 0$.

We next want to show that the two functions ϕ from (3.1) and (3.2) are continuously differentiable in their arguments X, S, and τ, at least under suitable assumptions. This result was essentially given by Chen and Tseng [8, Lemma 2] (who,

however, view τ as a parameter) and can alternatively be derived from the recent paper [26] by Sun and Sun.

Here we give a somewhat different proof for the differentiability of the functions ϕ. The reason is that, at least in our opinion, the proof given in, e.g., [8] is not very constructive, in the sense that it is not clear how to obtain the somewhat complicated formulas for the derivatives of the functions ϕ. We hope that the reader will find our approach more constructive. It is based on the following lemma from [16, section 7.2].

LEMMA 3.3. *Let $A \in \mathcal{S}^{n\times n}_{++}$, $B \in \mathcal{S}^{n\times n}_{+}$ be two given matrices. Then*

$$\|A^{1/2} - B^{1/2}\|_2 \le \|A^{-1/2}\|_2 \cdot \|A - B\|_2 .$$

We are now in the position to derive a formula for the derivatives of the mappings ϕ. To be specific, assume that ϕ denotes the smoothed Fischer–Burmeister function from (3.1). We have to show that

$$\|\phi(X+U, S+V, \tau+\mu) - \phi(X,S,\tau) - \nabla\phi(X,S,\tau)(U,V,\mu)\|_2 = o(\|\!|(U,V,\mu)|\!\|)$$

holds for all $(U,V,\mu) \in \mathcal{S}^{n\times n}\times\mathcal{S}^{n\times n}\times\mathbb{R}$ tending to $(0,0,0)$, where $\nabla\phi(X,S,\tau)$ denotes a suitable linear operator standing for the derivative of ϕ at the point (X,S,τ). To this end, we decompose the mapping ϕ into $\phi(X,S,\tau) = \phi_1(X,S,\tau) - \phi_2(X,S,\tau)$ with

$$\phi_1(X,S,\tau) := X + S, \qquad \phi_2(X,S,\tau) := (X^2 + S^2 + 2\tau^2 I)^{1/2}. \tag{3.3}$$

Then it is easy to see that ϕ_1 is differentiable with $\nabla\phi_1(X,S,\tau)(U,V,\mu) = U + V$. The situation for ϕ_2 is more complicated. Let us define

$$E := \left(X^2 + S^2 + 2\tau^2 I\right)^{1/2}$$

and assume that E is positive definite. Let

$$L_E[X] := EX + XE \tag{3.4}$$

denote the corresponding Lyapunov operator. Then the positive definiteness of E guarantees that the Lyapunov equation $L_E[X] = H$ has a unique solution within the set of symmetric matrices for every $H \in \mathcal{S}^{n\times n}$; cf. [17, Theorem 2.2.3]. Hence we can define the inverse L_E^{-1} of L_E; i.e., $L_E^{-1}[H]$ denotes the unique element X satisfying $EX + XE = H$. Let us further define the matrix

$$D := \left((X+U)^2 + (S+V)^2 + 2(\tau+\mu)^2 I\right)^{1/2}.$$

An easy calculation shows that $D^2 - E^2 = L_E[D-E] + (D-E)^2$. Applying L_E^{-1} to this equation and rearranging terms yields

$$\begin{aligned} E - D &= L_E^{-1}[(D-E)^2 - (D^2 - E^2)] \\ &= L_E^{-1}[(E-D)^2 - (XU + UX + SV + VS + 4\tau\mu I + U^2 + V^2 + 2\mu^2 I)]. \end{aligned}$$

Using the linearity of L_E^{-1} then gives

$$\begin{aligned} &\phi_2(X+U, S+V, \tau+\mu) - \phi_2(X,S,\tau) - \nabla\phi_2(X,S,\tau)(U,V,\mu) \\ &= -\nabla\phi_2(X,S,\tau)(U,V,\mu) - (E-D) \\ &= -\nabla\phi_2(X,S,\tau)(U,V,\mu) + L_E^{-1}[XU + UX + SV + VS + 4\tau\mu I] \\ &\quad + L_E^{-1}[U^2 + V^2 + 2\mu^2 I] - L_E^{-1}[(E-D)^2]. \end{aligned} \tag{3.5}$$

Obviously, we have $\|L_E^{-1}[U^2+V^2+2\mu^2 I]\|_F = O(\|\!|(U,V,\mu)|\!\|^2)$. In view of Lemma 3.3, we also have

$$\begin{aligned}\|(E-D)^2\|_F &\le \|E-D\|_F^2 \le \gamma_1\|E^2-D^2\|_F^2\\ &= \gamma_1\|XU+UX+SV+VS+4\tau\mu I+U^2+V^2+2\mu^2 I\|_F^2\\ &= O(\|\!|(U,V,\mu)|\!\|^2)\end{aligned}$$

for some constant $\gamma_1 > 0$ independent of U, V, and μ. This implies

$$\|L_E^{-1}[(E-D)^2]\|_F = O(\|\!|(U,V,\mu)|\!\|^2).$$

Therefore, setting

$$\nabla\phi_2(X,S,\tau)(U,V,\mu) := L_E^{-1}[XU+UX+SV+VS+4\tau\mu I],$$

it follows immediately from (3.5) that ϕ_2 is differentiable at (X,S,τ). This, in turn, implies that ϕ itself is differentiable at this point. This proves the main part of the first statement in the following result.

THEOREM 3.4. *Let $X, S \in \mathcal{S}^{n\times n}$ be two given matrices and $\tau \in \mathbb{R}_+$.*

(a) *If ϕ is given by (3.1) and $X^2+S^2+2\tau^2 I \succ 0$, then ϕ is continuously differentiable in (X,S,τ) with*

$$\nabla\phi(X,S,\tau)(U,V,\mu) = U+V-L_E^{-1}\left[XU+UX+SV+VS+4\tau\mu I\right], \tag{3.6}$$

where $E := (X^2+S^2+2\tau^2 I)^{1/2}$.

(b) *If ϕ is given by (3.2) and $(X-S)^2+4\tau^2 I \succ 0$, then ϕ is continuously differentiable in (X,S,τ) with*

$$\nabla\phi(X,S,\tau)(U,V,\mu) = U+V-L_E^{-1}\left[(X-S)(U-V)+(U-V)(X-S)+8\tau\mu I\right], \tag{3.7}$$

where $E := ((X-S)^2+4\tau^2 I)^{1/2}$.

Proof. (a) The differentiability of the smoothed Fischer–Burmeister function follows from our preceding discussion. Since $E = (X^2+S^2+2\tau^2 I)^{1/2} \succ 0$ is continuous in (X,S,τ) by Lemma 3.3, it is readily seen that $\nabla\phi(X,S,\tau)$ is continuous in (X,S,τ); see also [8]. Hence ϕ is continuously differentiable in (X,S,τ). Part (b) can be verified in a similar way. □

Note that Theorem 3.4 implies that if $\tau > 0$, then both functions ϕ are continuously differentiable everywhere.

4. Description of the algorithm. We now want to exploit our previous results to obtain a suitable algorithm for the solution of the optimality conditions (1.3) and, therefore, for the solution of the underlying primal and dual SDPs. The most obvious way would be to utilize the mapping

$$\Phi(X,\lambda,S) := \begin{pmatrix} \sum_{i=1}^m \lambda_i A_i + S - C \\ A_i \bullet X - b_i \ \ (i=1,\ldots,m) \\ \phi(X,S) \end{pmatrix},$$

with ϕ being the Fischer–Burmeister function (2.2) or the minimum function (2.11), since then Propositions 2.2 and 2.4 immediately imply that

$$(X^*,\lambda^*,S^*) \text{ solves (1.3)} \iff (X^*,\lambda^*,S^*) \text{ solves } \Phi(X,\lambda,S) = 0.$$

However, solving the nonlinear system of equations $\Phi(X, \lambda, S) = 0$ is a nontrivial task because ϕ, and therefore Φ, is nonsmooth in general. Hence we do not follow this idea here, although some recent theoretical results [26, 7, 14, 27] indicate that such an approach might be possible.

The next idea is to replace the nondifferentiable mapping Φ by the smooth function

$$\Phi_\tau(X, \lambda, S) := \begin{pmatrix} \sum_{i=1}^m \lambda_i A_i + S - C \\ A_i \bullet X - b_i \ \ (i = 1, \ldots, m) \\ \phi_\tau(X, S) \end{pmatrix},$$

where ϕ_τ denotes either the smoothed Fischer–Burmeister function from (2.7) or the smoothed minimum function from (2.12). This (specialized to the framework of SDPs) is precisely the approach followed by Chen and Tseng [8], although they have not observed the equivalence between the nonlinear system of equations $\Phi_\tau(X, \lambda, S) = 0$, on the one hand, and the central path conditions (1.4), on the other hand; cf. Theorem 2.6.

In this paper we follow an idea by Jiang [18] (in the context of nonlinear complementarity problems) and view τ as an independent variable. To this end, we define the mapping $\Theta : \mathcal{S}^{n\times n} \times \mathbb{R}^m \times \mathcal{S}^{n\times n} \times \mathbb{R} \to \mathcal{S}^{n\times n} \times \mathbb{R}^m \times \mathcal{S}^{n\times n} \times \mathbb{R}$ by

$$\Theta(X, \lambda, S, \tau) := \begin{pmatrix} \sum_{i=1}^m \lambda_i A_i + S - C \\ A_i \bullet X - b_i \ \ (i = 1, \ldots, m) \\ \phi(X, S, \tau) \\ \tau \end{pmatrix}, \tag{4.1}$$

where ϕ denotes one of the functions given by (3.1) or (3.2). Apart from the fact that τ is an independent variable rather than a parameter, the function Θ also differs from the function Φ_τ because we have added one more line so that

$$\Theta(X, \lambda, S, \tau) = 0 \tag{4.2}$$

becomes a square system of equations. This additional line immediately implies $\tau = 0$, so that the system (4.2) is equivalent to the optimality conditions (1.3) themselves (and not to the central path conditions (1.4)). This might be an advantage compared with the reformulation $\Phi_\tau(X, \lambda, S) = 0$, since we really want to solve the optimality conditions (1.3) and not the central path conditions (1.4). Furthermore, it follows from Theorem 3.4 that Θ is a continuously differentiable function at any point (X, λ, S, τ) with $\tau > 0$, and the positivity of τ will be guaranteed automatically by our method. This is an advantage compared with the nonsmooth reformulation $\Phi(X, \lambda, S) = 0$. Moreover, according to our numerical experience with some related methods for the solution of linear programs (cf. [10, 11, 9]), the reformulation (4.2) has the best numerical behavior. It also has some better theoretical properties in the context of linear complementarity problems (see Burke and Xu [5, 4]), although it is currently not clear whether this can be extended to SDPs.

The main idea of our algorithm is to solve the system of equations (4.2) by Newton's method. Global convergence of this method is achieved by following a suitable neighborhood of the central path. The neighborhood used here is given by

$$\mathcal{N}(\beta) = \left\{ (X, \lambda, S, \tau) \;\middle|\; A_i \bullet X = b_i \quad \forall i = 1, \ldots, m, \right. \\ \left. \sum_{i=1}^m \lambda_i A_i + S = C, \ \|\phi(X, S, \tau)\|_F \le \beta\tau \right\},$$

where β denotes a positive number. Local fast convergence will be guaranteed by using a suitable predictor step. To simplify the formulation of our algorithm as well as the notation used in the subsequent analysis, let us introduce the abbreviation $W^k := (X^k, \lambda^k, S^k)$, where k denotes the iteration index. We are now in a position to give a formal statement of our smoothing-type method for the solution of SDPs.

ALGORITHM 4.1.

(S.0) *(Initialization)*
Choose $W^0 = (X^0, \lambda^0, S^0) \in \mathcal{S}^{n\times n} \times \mathbb{R}^m \times \mathcal{S}^{n\times n}$ *with* $\sum_{i=1}^m \lambda_i^0 A_i + S^0 = C$ *and* $A_i \bullet X^0 = b_i$ $(i = 1, \ldots, m)$. *Choose* $\tau_0 > 0$, $\beta > 0$ *with* $\|\phi(X^0, S^0, \tau_0)\|_F \le \beta\tau_0$, *and set* $k := 0$. *Choose* $\hat\sigma, \alpha_1, \alpha_2 \in (0, 1)$.

(S.1) *(Predictor step)*
Let $(\Delta W^k, \Delta\tau_k) = (\Delta X^k, \Delta\lambda^k, \Delta S^k, \Delta\tau_k) \in \mathcal{S}^{n\times n} \times \mathbb{R}^m \times \mathcal{S}^{n\times n} \times \mathbb{R}$ *be a solution of the system*

$$\nabla\Theta(W^k, \tau_k) \begin{pmatrix} \Delta W \\ \Delta\tau \end{pmatrix} = -\Theta(W^k, \tau_k). \tag{4.3}$$

If $\|\phi(X^k + \Delta X^k, S^k + \Delta S^k, 0)\|_F = 0$, *STOP.*
Otherwise, if $\|\phi(X^k + \Delta X^k, S^k + \Delta S^k, \tau_k)\|_F > \beta\tau_k$, *then let* $\hat W^k := W^k$, $\hat\tau_k := \tau_k$, *and* $\eta_k := 1$; *else let* $\eta_k = \alpha_1^s$, *where* s *is the nonnegative number with*

$$\begin{aligned} &\|\phi(X^k + \Delta X^k, S^k + \Delta S^k, \alpha_1^r\tau_k)\|_F \le \beta\tau_k\alpha_1^r, \quad r = 0, 1, 2, \ldots, s, \\ &\|\phi(X^k + \Delta X^k, S^k + \Delta S^k, \alpha_1^{s+1}\tau_k)\|_F > \beta\tau_k\alpha_1^{s+1}, \end{aligned}$$

and set

$$\hat\tau_k := \eta_k\tau_k \quad \textit{and} \quad \hat W^k := \begin{cases} W^k & \textit{if } s = 0, \\ W^k + \Delta W^k & \textit{otherwise.} \end{cases}$$

(S.2) *(Corrector step)*
Let $(\Delta\hat W^k, \Delta\hat\tau_k) = (\Delta\hat X^k, \Delta\hat\lambda^k, \Delta\hat S^k, \Delta\hat\tau_k)$ *be a solution of*

$$\nabla\Theta(\hat W^k, \hat\tau_k) \begin{pmatrix} \Delta\hat W \\ \Delta\hat\tau \end{pmatrix} = -\Theta(\hat W^k, \hat\tau_k) + \begin{pmatrix} 0 \\ (1 - \hat\sigma)\hat\tau_k \end{pmatrix}. \tag{4.4}$$

Let $\hat\eta_k$ *be the maximum of the numbers* $1, \alpha_2, \alpha_2^2, \ldots,$ *with*

$$\|\phi(\hat X^k + \hat\eta_k\Delta\hat X^k, \hat S^k + \hat\eta_k\Delta\hat S^k, \hat\tau_k + \hat\eta_k\Delta\hat\tau_k)\|_F \le (1 - \hat\sigma\hat\eta_k)\beta\hat\tau_k. \tag{4.5}$$

Set $W^{k+1} := \hat W^k + \hat\eta_k\Delta\hat W^k$, $\tau_{k+1} := (1 - \hat\sigma\hat\eta_k)\hat\tau_k$, $k \leftarrow k + 1$, *and go to* (S.1).

It can easily be seen that all iterates (X^k, λ^k, S^k) and $(\hat X^k, \hat\lambda^k, \hat S^k)$ generated by Algorithm 4.1 are feasible for the optimality conditions (1.3) in the sense that

$$\sum_{i=1}^m \lambda_i^k A_i + S^k = C, \qquad A_i \bullet X^k = b_i \quad (i = 1, \ldots, m) \tag{4.6}$$

and $\sum_{i=1}^m \hat\lambda_i^k A_i + \hat S^k = C$ and $A_i \bullet \hat X^k = b_i\,(i = 1, \ldots, m)$ hold for all $k \in \mathbb{N}$. Moreover, we will see below that all matrices X^k, S^k and $\hat X^k, \hat S^k$ are automatically symmetric; cf. section 6. This is in contrast to interior-point methods, which first have to symmetrize the central path conditions in order to guarantee that they get

symmetric search directions. On the other hand, our matrices are not necessarily positive definite or positive semidefinite.

Note that the predictor step (S.1) will be responsible for the local fast convergence of Algorithm 4.1, whereas the corrector step (S.2) will be used in order to prove global convergence.

The termination criterion used in (S.1) is justified by Propositions 2.2 and 2.4. Together with our previous note on the feasibility of the iterates, these results imply that

$$\left\|\phi(X^k + \Delta X^k, S^k + \Delta S^k, 0)\right\|_F = 0 \iff W^k + \Delta W^k \text{ is a solution of (1.3).}$$

For our theoretical analysis of Algorithm 4.1, we will always assume that this criterion never holds so that Algorithm 4.1 generates an infinite sequence. Furthermore, the updating rule for τ_{k+1} in (S.2) is equivalent to the more standard formula $\tau_{k+1} = \hat{\tau}_k + \hat{\eta}_k \Delta\hat{\tau}_k$; this observation follows immediately from the last row of the linear system (4.4) in the corrector step, which gives $\Delta\hat{\tau}_k = -\hat{\sigma}\hat{\tau}_k$.

Finally, we stress that we have to solve a linear system of equations in both the predictor and the corrector steps, with possibly different matrices $\nabla\Theta(W, \tau)$, and this is more costly than what is usually done by interior-point methods. However, an easy inspection of our subsequent analysis shows that all convergence results remain true for the following modification of Algorithm 4.1: If the predictor step has been accepted with $\eta_k < 1$, then skip the corrector step; i.e., set $W^{k+1} := W^k + \Delta W^k, \tau_{k+1} := \eta_k \tau_k, k \leftarrow k+1$, and return to (S.1). This modified algorithm either has to solve only one linear system of equations in the predictor step or it has to solve two systems, but then these two systems have the same coefficient matrix. This modification has been implemented in order to obtain the numerical results in section 6.

We now start to analyze the properties of Algorithm 4.1 more formally. The aim of the remaining part of this section will be to show that Algorithm 4.1 is well defined. To this end, we first want to show that the linear systems (4.3) and (4.4) have a unique solution. In order to verify this statement, we need some further properties of the Lyapunov operator from (3.4). These properties are therefore summarized in our next result.

LEMMA 4.2. *Let $A, B \in \mathcal{S}^{n\times n}_{++}$ be two symmetric positive definite matrices, and let L_A, L_B be the corresponding Lyapunov operators defined by* (3.4)*, with L_A^{-1}, L_B^{-1} denoting their inverses. Then the following statements hold:*

(a) *L_A and L_B are self-adjoint.*
(b) *L_A^{-1} and L_B^{-1} are self-adjoint.*
(c) *$L_A \circ L_B$ and $L_B \circ L_A$ are strongly monotone.*
(d) *$L_A^{-1} \circ L_B$ and $L_B^{-1} \circ L_A$ are strongly monotone.*

Proof. (a) We have to verify only that L_A is self-adjoint. This follows directly from the fact that

$$\begin{aligned} L_A[X] \bullet Y &= \operatorname{tr}(L_A[X]Y) = \operatorname{tr}\left((AX + XA)Y\right) = \operatorname{tr}(AXY) + \operatorname{tr}(XAY) \\ &= \operatorname{tr}(XYA) + \operatorname{tr}(XAY) = \operatorname{tr}\left(X(AY + YA)\right) = \operatorname{tr}(XL_A[Y]) = X \bullet L_A[Y] \end{aligned}$$

for all $X, Y \in \mathcal{S}^{n\times n}$.

(b) We show that L_A^{-1} is self-adjoint. Noting that L_A^{-1} is the inverse of L_A and exploiting part (a), we obtain

$$L_A^{-1}[X] \bullet Y = L_A^{-1}[X] \bullet L_A\big[L_A^{-1}[Y]\big] = L_A\big[L_A^{-1}[X]\big] \bullet L_A^{-1}[Y] = X \bullet L_A^{-1}[Y]$$

for all $X, Y \in \mathcal{S}^{n\times n}$.

(c) Using the first statement, we obtain

$$\begin{aligned}(L_A \circ L_B[X]) \bullet X &= L_B[X] \bullet L_A[X] = \operatorname{tr}(L_B[X]L_A[X]) \\ &= \operatorname{tr}((BX+XB)(AX+XA)) \\ &= \operatorname{tr}(BXAX + XBAX + BXXA + XBXA) \\ &= \operatorname{tr}(2BXAX + X^2(BA+AB)) \\ &= 2\operatorname{tr}(BXAX) + \operatorname{tr}(X^2(BA+AB)) \\ &= 2\|B^{1/2}XA^{1/2}\|_F^2 + \operatorname{tr}(X(BA+AB)X)\end{aligned} \tag{4.7}$$

for all $X \in \mathcal{S}^{n\times n}$. Since $BA\,(AB)$ is similar to $B^{1/2}AB^{1/2}\,(A^{1/2}BA^{1/2})$ and A, B are symmetric positive definite, it follows that BA and AB have real and positive eigenvalues. Hence the symmetric matrix $BA+AB$ is positive definite. Consequently, $X(BA+AB)X$ is positive semidefinite, so that

$$\operatorname{tr}\big(X(BA+AB)X\big) \geq 0 \tag{4.8}$$

for all $X \in \mathcal{S}^{n\times n}$. Furthermore, since the mapping $X \mapsto \|B^{1/2}XA^{1/2}\|_F$ defines a norm and all norms are equivalent in finite-dimensional spaces, there exists a constant $\mu > 0$ such that

$$\|B^{1/2}XA^{1/2}\|_F \geq \mu \|X\|_F \tag{4.9}$$

for all $X \in \mathcal{S}^{n\times n}$. Putting together the inequalities (4.7)–(4.9), we obtain

$$(L_A \circ L_B[X]) \bullet X \geq 2\|B^{1/2}XA^{1/2}\|_F^2 \geq 2\mu^2 \|X\|_F^2,$$

i.e., $L_A \circ L_B$ is strongly monotone on $\mathcal{S}^{n\times n}$. In order to see that $L_B \circ L_A$ is also strongly monotone, we just have to change the roles of A and B.

(d) Since L_A is self-adjoint by part (a), we obtain for every $X \in \mathcal{S}^{n\times n}$ (by setting $Y := L_A^{-1}[X]$)

$$\big(L_A^{-1} \circ L_B[X]\big) \bullet X = \big(L_A^{-1} \circ L_B \circ L_A[Y]\big) \bullet L_A[Y] = \big(L_B \circ L_A[Y]\big) \bullet Y.$$

However, $L_B \circ L_A$ is strongly monotone by part (c). Hence (d) follows from (c). ∎

In order to see that the linear systems (4.3) and (4.4) have a unique solution, we will show that the linear mapping $\nabla\Theta(X, \lambda, S, \tau)$ is invertible. To this end, we state the following standard assumption.

ASSUMPTION 4.3. *The matrices A_i $(i = 1, \ldots, m)$ are linearly independent.*

Exploiting Lemma 4.2 and Assumption 4.3, we can now show that $\nabla\Theta(X, \lambda, S, \tau)$ is a bijection, i.e., it is both one-to-one and onto. Note that this implies that the predictor direction $(\Delta X^k, \Delta\lambda^k, \Delta S^k, \Delta\tau_k)$ and the corrector direction $(\Delta\hat{X}^k, \Delta\hat{\lambda}^k, \Delta\hat{S}^k, \Delta\hat{\tau}_k)$ are well-defined.

PROPOSITION 4.4. *Suppose that Assumption* 4.3 *holds. Then the linear mapping $\nabla\Theta(X, \lambda, S, \tau)$, with ϕ given by* (3.1) *or* (3.2)*, is bijective for all $(X, \lambda, S, \tau) \in \mathcal{S}^{n\times n} \times \mathbb{R}^m \times \mathcal{S}^{n\times n} \times \mathbb{R}_{++}$.*

Proof. We only consider the case in which the function ϕ is given by (3.1). The proof for the smoothed minimum function is similar and therefore omitted here.

Let $(X, \lambda, S, \tau) \in \mathcal{S}^{n\times n} \times \mathbb{R}^m \times \mathcal{S}^{n\times n} \times \mathbb{R}_{++}$ be fixed. Since $\nabla\Theta(X, \lambda, S, \tau)$ is a linear mapping from the finite-dimensional vector space $\mathcal{S}^{n\times n} \times \mathbb{R}^m \times \mathcal{S}^{n\times n} \times \mathbb{R}$ into itself, we only have to verify that this mapping is one-to-one. To this end, it is

sufficient to show that the system $\nabla\Theta(X,\lambda,S,\tau)(\Delta X,\Delta\lambda,\Delta S,\Delta\tau) = (0,0,0,0)$ or, equivalently, the system

$$\sum_{i=1}^m \Delta\lambda_i A_i + \Delta S = 0, \tag{4.10}$$

$$A_i \bullet \Delta X = 0 \quad (i = 1,\dots,m), \tag{4.11}$$

$$\nabla\phi(X,S,\tau)(\Delta X,\Delta S,\Delta\tau) = 0, \tag{4.12}$$

$$\Delta\tau = 0 \tag{4.13}$$

has $(\Delta X,\Delta\lambda,\Delta S,\Delta\tau) = (0,0,0,0)$ as its only solution. From (4.13) we immediately obtain $\Delta\tau = 0$. Setting $E := (X^2 + S^2 + 2\tau^2 I)^{1/2}$, we therefore get from (4.12) and Theorem 3.4 that

$$\Delta X + \Delta S - L_E^{-1}\left[X\Delta X + \Delta X\, X + S\Delta S + \Delta S\, S\right] = 0.$$

Applying L_E to both sides of the equation and rearranging terms yields

$$L_{E-X}[\Delta X] + L_{E-S}[\Delta S] = 0.$$

Since $E - S \succ 0$ (see [30, Lemma 6.1(c)] for a formal proof), the inverse L_{E-S}^{-1} exists, and we get

$$L_{E-S}^{-1} \circ L_{E-X}[\Delta X] + \Delta S = 0. \tag{4.14}$$

Using (4.10) and (4.11) and taking the scalar product with ΔX yields

$$0 = L_{E-S}^{-1} \circ L_{E-X}[\Delta X] \bullet \Delta X - \sum_{i=1}^m \Delta\lambda_i \underbrace{A_i \bullet \Delta X}_{=0} = L_{E-S}^{-1} \circ L_{E-X}[\Delta X] \bullet \Delta X. \tag{4.15}$$

Using the fact that $E - X \succ 0$ and $E - S \succ 0$, it follows from Lemma 4.2(d) that the operator $L_{E-S}^{-1} \circ L_{E-X}$ is strongly monotone. Therefore, (4.15) immediately gives $\Delta X = 0$. This implies $\Delta S = 0$ by (4.14). The assumed linear independence of the matrices A_i and (4.10) shows that $\Delta\lambda = 0$, and this completes the proof. □

Based on the previous results, it is possible to show that Algorithm 4.1 is well defined under Assumption 4.3. Since the proof is rather standard, we skip it here and refer the reader to the preprint version of this paper [20] for further details.

THEOREM 4.5. *Algorithm* 4.1 *is well defined under Assumption* 4.3. *Furthermore, the iterates* $W^k = (X^k,\lambda^k,S^k)$ *and* τ_k *and* $\hat{W}^k = (\hat{X}^k,\hat{\lambda}^k,\hat{S}^k)$ *and* $\hat{\tau}_k$ *belong to the neighborhood* $\mathcal{N}(\beta)$.

5. Global and local superlinear convergence. We first state the main global convergence result for Algorithm 4.1. Again, its proof is more or less standard, so we skip it here and refer the interested reader once more to the preprint version [20] for more details. The only thing we note here is that the updating rules for the smoothing parameter τ in Algorithm 4.1 guarantee that this parameter is monotonically decreasing and positive at all iterations.

THEOREM 5.1. *If the sequence* $\{W^k\} = \{(X^k,\lambda^k,S^k)\}$ *generated by Algorithm* 4.1 *has an accumulation point, then the sequence* $\{\tau_k\}$ *converges to zero. In particular, every accumulation point of the sequence* $\{W^k\}$ *is a solution of the optimality conditions* (1.3).

We next investigate the local properties of Algorithm 4.1. Our aim is to show that the sequence $\{\tau_k\}$ converges superlinearly to zero. Since this result depends on certain properties of the predictor step in Algorithm 4.1, we first state the following assumption.

ASSUMPTION 5.2. *The sequence* $\{\tau_k\}$ *generated by Algorithm* 4.1 *converges to zero, and we have*

$$\left\| \begin{pmatrix} \Delta W^k \\ \Delta \tau_k \end{pmatrix} \right\| = O(\tau_k), \tag{5.1}$$

where $(\Delta W^k, \Delta \tau_k)$ *denotes the search direction computed in* (4.3).

In order to justify Assumption 5.2, we first note that Theorem 5.1 provides a sufficient condition for the sequence $\{\tau_k\}$ to converge to zero. To understand the second condition, assume that the sequence of inverse operators $\nabla\Theta(W^k, \tau_k)^{-1}$ remains bounded for $k \to \infty$. Then we obtain from the linear system (4.3) that (5.1) holds, provided that the right-hand side in (4.3) is of the order $O(\tau_k)$. This, however, is rather obvious since the feasibility of the iterates (cf. (4.6)) together with the fact that all iterates belong to the neighborhood $\mathcal{N}(\beta)$ (cf. Theorem 4.5) show that

$$\begin{aligned} \left\| \Theta(W^k, \tau_k) \right\| &= \sqrt{\|\phi(X^k, S^k, \tau_k)\|_F^2 + \tau_k^2} \leq \left\| \phi(X^k, S^k, \tau_k) \right\|_F + \tau_k \\ &\leq \beta\tau_k + \tau_k = O(\tau_k). \end{aligned}$$

In addition, such a relation also holds if we replace the right-hand side in (4.3) by $-\Theta(W^k, 0)$, since then Corollary 3.2 and Theorem 4.5 imply

$$\begin{aligned} \left\| \Theta(W^k, 0) \right\| &= \left\| \phi(X^k, S^k, 0) \right\|_F \\ &\leq \left\| \phi(X^k, S^k, \tau_k) - \phi(X^k, S^k, 0) \right\|_F + \left\| \phi(X^k, S^k, \tau_k) \right\|_F \\ &\leq \kappa\sqrt{n}\tau_k + \beta\tau_k \\ &= O(\tau_k), \end{aligned}$$

where $\kappa > 0$ denotes the constant from Lemma 3.1. In particular, all global and local convergence properties of Algorithm 4.1 remain true if we use this modification of the right-hand side in (4.3).

In order to state a sufficient condition for Assumption 5.2 to be satisfied, we introduce the following assumption.

ASSUMPTION 5.3. *Let* (X^*, λ^*, S^*) *be a solution of the optimality conditions* (1.3) *such that*

(a) *(Strict complementarity)* $X^* + S^* \succ 0$*;*

(b) *(Nondegeneracy) for any* $(\Delta X, \Delta\lambda, \Delta S)$ *satisfying* $\sum_{i=1}^m \Delta\lambda_i A_i + \Delta S = 0$ *and* $A_i \bullet \Delta X = 0$ $(i = 1, \ldots, m)$*, the following implication holds:*

$$X^*\Delta S + \Delta X S^* = 0 \Longrightarrow (\Delta X, \Delta S) = (0, 0).$$

Assumption 5.3(a) is rather standard, and Assumption 5.3(b) was introduced by Kojima, Shida, and Shindoh [22]. As noted in [22], Haeberly showed that this assumption is equivalent to the primal and dual nondegeneracy conditions considered by Alizadeh, Haeberly, and Overton [2].

The next result implies that Assumption 5.2 holds under Assumptions 4.3 and 5.3, provided that the iterates (X^k, λ^k, S^k) generated by Algorithm 4.1 converge to a solution (X^*, λ^*, S^*) satisfying these two conditions. The convergence of the iterates to

this single point is not at all restrictive since it is known that the two Assumptions 4.3 and 5.3 together imply that (X^*, λ^*, S^*) is the unique solution of the optimality conditions (1.3).

THEOREM 5.4. *Suppose that Assumptions* 4.3 *and* 5.3 *hold at a solution* (X^*, λ^*, S^*) *of* (1.3). *Then the linear mapping* $\nabla\Theta(X^*, \lambda^*, S^*, 0)$ *is bijective.*

Proof. We consider only the case in which ϕ is defined via the smoothed Fischer–Burmeister function from (3.1). The proof for the smoothed minimum function from (3.2) is similar.

Let us define $E := ((X^*)^2 + (S^*)^2)^{1/2}$. In view of the assumed strict complementarity, it is easy to see that E is a positive definite matrix. Hence Theorem 3.4 implies that Θ is continuously differentiable at $(X^*, \lambda^*, S^*, 0)$. In order to see that $\nabla\Theta(X^*, \lambda^*, S^*, 0)$ is bijective, we have only to verify that it is one-to-one. To this end, we consider the equation

$$\nabla\Theta(X^*, \lambda^*, S^*, 0)\,(\Delta X, \Delta\lambda, \Delta S, \Delta\tau) = (0, 0, 0, 0)$$

and show that $(\Delta X, \Delta\lambda, \Delta S, \Delta\tau) = (0, 0, 0, 0)$ is its only solution. The last row gives

$$\Delta\tau = 0. \tag{5.2}$$

Taking this into account and using Theorem 3.4, the first three block rows can be rewritten as follows:

$$\textstyle\sum_{i=1}^m \Delta\lambda_i A_i + \Delta S = 0, \tag{5.3}$$

$$A_i \bullet \Delta X = 0 \quad (i = 1, \ldots, m), \tag{5.4}$$

$$\Delta X + \Delta S - L_E^{-1}[X^*\Delta X + \Delta X X^* + S^*\Delta S + \Delta S S^*] = 0. \tag{5.5}$$

Equation (5.5) implies

$$L_{E-X^*}[\Delta X] + L_{E-S^*}[\Delta S] = 0; \tag{5.6}$$

cf. the proof of Proposition 4.4. Now, using the fact that (X^*, λ^*, S^*) is a strictly complementary solution of (1.3) so that, in particular, we have $X^*S^* = 0$, i.e., X^* and S^* commute, it follows that these two matrices can be diagonalized simultaneously by an orthogonal transformation. This means that we can find a single orthogonal matrix $Q \in \mathbb{R}^{n\times n}$ and diagonal matrices $D_X \in \mathbb{R}^{n\times n}$ and $D_S \in \mathbb{R}^{n\times n}$ such that $X^* = Q^T D_X Q$ and $S^* = Q^T D_S Q$. Taking this into account, an easy calculation shows that $E - X^* = S^*$ and $E - S^* = X^*$. Hence (5.6) can be rewritten as $S^*\Delta X + \Delta X S^* + X^*\Delta S + \Delta S X^* = 0$. Using (5.3), (5.4), and Assumption 5.3, we therefore obtain from [23, Lemma 6.2] that $(\Delta X, \Delta S) = (0, 0)$. Since the matrices A_i are linearly independent by Assumption 4.3, it follows from (5.3) that $\Delta\lambda = 0$. In view of (5.2), this completes the proof. □

We stress that Theorem 5.4 provides only a sufficient condition for Assumption 5.2 to be satisfied. Since the assumptions used in Theorem 5.4 do imply that the solution set of the optimality conditions (1.3) is just a singleton, Theorem 5.4 is somewhat restrictive. However, some recent results obtained for linear programs and complementarity problems indicate that Assumption 5.2 may also hold under weaker conditions that do not necessarily imply the unique solvability of (1.3); cf. [31] and [11].

We now start to analyze the local behavior of Algorithm 4.1, starting with the following technical result.

LEMMA 5.5. *Suppose Assumption* 5.2 *holds. Then we have*

$$\left\|\phi(X^k + \Delta X^k, S^k + \Delta S^k, \tau_k + \Delta\tau_k)\right\|_F = o(\tau_k).$$

Proof. Since $\nabla\phi(X^k, S^k, \tau_k)\left(\Delta X^k, \Delta S^k, \Delta\tau_k\right) = -\phi(X^k, S^k, \tau_k)$ by (4.3), we obtain from the integral mean value theorem that

$$\begin{aligned}
&\left\|\phi(X^k + \Delta X^k, S^k + \Delta S^k, \tau_k + \Delta\tau_k)\right\|_F \\
&= \left\| \int_0^1 \nabla\phi(X^k + \eta\Delta X^k, S^k + \eta\Delta S^k, \tau_k + \eta\Delta\tau_k) \begin{pmatrix} \Delta X^k \\ \Delta S^k \\ \Delta\tau_k \end{pmatrix} \mathrm{d}\eta + \phi(X^k, S^k, \tau_k) \right\|_F \\
&= \left\| \int_0^1 \left[\nabla\phi(X^k + \eta\Delta X^k, S^k + \eta\Delta S^k, \tau_k + \eta\Delta\tau_k) - \nabla\phi(X^k, S^k, \tau_k)\right] \begin{pmatrix} \Delta X^k \\ \Delta S^k \\ \Delta\tau_k \end{pmatrix} \mathrm{d}\eta \right\|_F \\
&\le \int_0^1 \left\| \left[\nabla\phi(X^k + \eta\Delta X^k, S^k + \eta\Delta S^k, \tau_k + \eta\Delta\tau_k) - \nabla\phi(X^k, S^k, \tau_k)\right] \begin{pmatrix} \Delta X^k \\ \Delta S^k \\ \Delta\tau_k \end{pmatrix} \right\|_F \mathrm{d}\eta \\
&= o\left(\left\| \begin{pmatrix} \Delta X^k \\ \Delta S^k \\ \Delta\tau_k \end{pmatrix} \right\| \right),
\end{aligned}$$

where the last equality follows from the continuous differentiability of the mapping ϕ. Using Assumption 5.2, we therefore get $\|\phi(X^k + \Delta X^k, S^k + \Delta S^k, \tau_k + \Delta\tau_k)\|_F = o(\tau_k)$. ∎

The main step required to prove local superlinear convergence of the sequence $\{\tau_k\}$ is contained in the following result.

LEMMA 5.6. *Suppose Assumption* 5.2 *holds, and let the constant β satisfy the inequality $\beta > \kappa\sqrt{n}$, where κ denotes the constant from Lemma* 3.1. *Then the sequence $\{\eta_k\}$ converges to zero.*

Proof. Let $\varepsilon > 0$ be arbitrarily given. Using the fact that $\Delta\tau_k = -\tau_k$ because of (4.3), we obtain from Lemma 5.5 that

$$\left\|\phi(X^k + \Delta X^k, S^k + \Delta S^k, 0)\right\| = \left\|\phi(X^k + \Delta X^k, S^k + \Delta S^k, \tau_k + \Delta\tau_k)\right\|_F = o(\tau_k).$$

Hence there is an index $K_\varepsilon \in \mathbb{N}$ such that $\|\phi(X^k + \Delta X^k, S^k + \Delta S^k, 0)\|_F \le \varepsilon\tau_k$ for all $k \ge K_\varepsilon$. Then we get for all $\eta > 0$ and all $k \ge K_\varepsilon$

$$\begin{aligned}
&\left\|\phi(X^k + \Delta X^k, S^k + \Delta S^k, \eta\tau_k)\right\|_F \\
&\le \left\|\phi(X^k + \Delta X^k, S^k + \Delta S^k, 0)\right\|_F \\
&\quad + \left\|\phi(X^k + \Delta X^k, S^k + \Delta S^k, \eta\tau_k) - \phi(X^k + \Delta X^k, S^k + \Delta S^k, 0)\right\|_F \\
&\le \varepsilon\tau_k + \kappa\sqrt{n}\eta\tau_k,
\end{aligned}$$

where the last inequality follows from Corollary 3.2. Since $\varepsilon\tau_k + \kappa\sqrt{n}\eta\tau_k \le \beta\eta\tau_k$ holds for all $\eta \ge \frac{\varepsilon}{\beta - \kappa\sqrt{n}}$, the definition of η_k shows that $\eta_k\alpha_1$ does not satisfy this inequality, i.e., $\eta_k < \varepsilon/((\beta - \kappa\sqrt{n})\alpha_1)$. Since $\beta - \kappa\sqrt{n} > 0$ by assumption and $\varepsilon > 0$ was chosen arbitrarily, this implies $\eta_k \to 0$. ∎

We are now in a position to state the main local convergence result for Algorithm 4.1.

THEOREM 5.7. *Under Assumption* 5.2 *we have* $\tau_{k+1} = o(\tau_k)$*; i.e., the smoothing parameter converges locally superlinearly to zero.*

Proof. Using Lemma 5.6 and the definition of τ_{k+1} and $\hat{\tau}_k$ in Algorithm 4.1, we obtain $\tau_{k+1} \le \hat{\tau}_k = \eta_k \tau_k = o(\tau_k)$, i.e., $\tau_k \to 0$ superlinearly. □

We close this section with a few remarks concerning Theorem 5.7: First, Theorem 5.7 still holds if we replace the right-hand side in (4.3) by $-\Theta(W^k, 0)$. This follows from the analysis carried out in [8] (so we skip the details here) and does not follow immediately from our previous discussion since Lemmas 5.5 and 5.6 depend on the fact that the right-hand side of (4.3) is given by $-\Theta(W^k, \tau_k)$. Furthermore, a more involved analysis (see [8] once again) may be used to show that $\nabla\phi$ is locally Lipschitzian. This observation may then be applied to show that we actually have a quadratic rate of convergence in Theorem 5.7. Finally (and this observation is credited to one of the referees) one can borrow a result from [27] to show that the mapping Θ is strongly semismooth, at least if Θ is defined via the minimum function. Using this fact together with a strong regularity assumption at a solution of the optimality conditions implies that our method is locally quadratically convergent without assuming the strict complementarity condition from Assumption 5.3(a). So far, however, we do not have a handy criterion for the strong regularity assumption to be satisfied.

6. Numerical results. In order to test the numerical performance of Algorithm 4.1, we implemented the method in Matlab. To simplify the programming work, we borrowed the data structure, problem input, and some linear algebra routines from the SDPT3 (version 2.1) Matlab code; see [29].

In our Matlab implementation of Algorithm 4.1, we choose ϕ to be the smoothed minimum function from (3.2). (The results for the smoothed Fischer–Burmeister function seem to be similar.) Furthermore, we take $\alpha_1 = \alpha_2 = 0.5$. The centering parameter $\hat{\sigma}$ gets updated dynamically, using a procedure suggested in [12] for the solution of linear programs.

In order to see how the Newton directions can be computed, let us first consider one iteration of the predictor step. Dropping the superscript k and using the abbreviation $R_d = C - \sum_{j=1}^m \lambda_j A_j - S$, the predictor step (4.3) (with the modification mentioned in section 5 that the right-hand side $-\Theta(W, \tau)$ gets replaced by $-\Theta(W, 0)$) becomes

$$\sum_{j=1}^m \Delta\lambda_j A_j + \Delta S = R_d, \tag{6.1}$$

$$A_i \bullet \Delta X = b_i - A_i \bullet X \quad (i = 1, \ldots, m), \tag{6.2}$$

$$\nabla\phi(X, S, \tau)(\Delta X, \Delta S, \Delta\tau) = -\phi(X, S, 0), \tag{6.3}$$

$$\Delta\tau = 0\,. \tag{6.4}$$

Writing $E := ((X - S)^2 + 4\tau^2 I)^{1/2}$ (cf. Theorem 3.4), applying the corresponding Lyapunov operator L_E on both sides of (6.3), and using (6.4), we obtain

$$L_{E-(X-S)}[\Delta X] + L_{E+(X-S)}[\Delta S] = -L_E[\phi(X, S, 0)]$$

or, equivalently,

$$\Delta X = -L^{-1}_{E-(X-S)} \left[L_{E+(X-S)}[\Delta S] + L_E[\phi(X, S, 0)] \right]. \tag{6.5}$$

Substituting ΔS from (6.1) and rearranging terms yields

$$\Delta X = \sum_{j=1}^{m} \Delta\lambda_j \, L^{-1}_{E-(X-S)} \left[L_{E+(X-S)}[A_j] \right] \\ - L^{-1}_{E-(X-S)} \left[L_{E+(X-S)} \left[R_d \right] + L_E[\phi(X,S,0)] \right].$$

Taking inner products with $A_i \, (i = 1, \ldots, m)$ and using the fact that $L^{-1}_{E-(X-S)}$ is self-adjoint by Lemma 4.2(b), we obtain from (6.2)

(6.6)
$$\sum_{j=1}^{m} \Delta\lambda_j L_{E+(X-S)}[A_j] \bullet L^{-1}_{E-(X-S)}[A_i] = b_i - A_i \bullet X \\ + \left(L_{E+(X-S)}[R_d] + L_E[\phi(X,S,0)] \right) \bullet L^{-1}_{E-(X-S)}[A_i], \qquad i = 1, \ldots, m.$$

This is a linear equation in the variables $\Delta\lambda \in \mathbb{R}^m$. After solving this system, we immediately get ΔS from (6.1). Note that ΔS is obviously symmetric, since R_d and all A_i are symmetric. In view of (6.5), ΔX can then be obtained as a solution of a Lyapunov equation with a symmetric right-hand side and is therefore also symmetric; cf. [17, Theorem 2.2.3]. The solution of this Lyapunov equation may be computed by using a spectral decomposition of $X - S$, which in turn yields a spectral decomposition of $E - (X - S)$ and which may also be used to compute the matrix from the linear system (6.6); see [17, p. 100].

The computation of the search direction in the corrector step (4.4) is similar to the one of the predictor step. The main difference is that we compute the vector $\Delta\hat{\lambda}$ by solving the linear system

(6.7)
$$\sum_{j=1}^{m} \Delta\hat{\lambda}_j L_{\hat{E}+(\hat{X}-\hat{S})}[A_j] \bullet L^{-1}_{\hat{E}-(\hat{X}-\hat{S})}[A_i] = b_i - A_i \bullet \hat{X} \\ + \left(L_{\hat{E}+(\hat{X}-\hat{S})}[\hat{R}_d] + L_{\hat{E}}[\phi(\hat{X},\hat{S},\hat{\tau})] + 8\sigma\hat{\tau}^2 I \right) \bullet L^{-1}_{\hat{E}-(\hat{X}-\hat{S})}[A_i], \quad i = 1, \ldots, m,$$

rather than (6.6), where, of course, we have used the notation $(\hat{X}, \hat{\lambda}, \hat{S}) := (\hat{X}^k, \hat{\lambda}^k, \hat{S}^k)$, $\hat{R}_d := C - \sum_{j=1}^m \hat{\lambda}_j A_j - \hat{S}$, and $\hat{E} := ((\hat{X} - \hat{S})^2 + 4\hat{\tau}^2 I)^{1/2}$. Note, however, that the corrector step is not carried out when the predictor step is accepted with $\eta_k < 1$. Hence, either the algorithm uses only a predictor step in one iteration, or the two matrices in (6.6) and (6.7) coincide.

In order to describe the way we compute our starting point (X^0, λ^0, S^0), let us call a triple (X, λ, S) *feasible* for the optimality conditions (1.3) if it satisfies the linear equations $\sum_{i=1}^m \lambda_i A_i + S = C$ (this will be called *dual feasibility*) and $A_i \bullet X = b_i$, $i = 1, \ldots, m$ (this will be called *primal feasibility*). Note that we do not require $X \succeq 0$ or $S \succeq 0$ for such a feasible triple. Of course, our starting point (X^0, λ^0, S^0) should be feasible in this sense.

To this end, we define a symmetric matrix $\mathcal{A} \in \mathbb{R}^{m \times m}$ by $\mathcal{A}_{ij} = A_i \bullet A_j$ for $i, j = 1, \ldots, m$ and solve the linear system $\mathcal{A}y = b$ to obtain $y^0 \in \mathbb{R}^m$. Then we define $X^0 = \sum_{i=1}^m y_i^0 A_i$ and compute λ^0 as a solution of the system $\mathcal{A}\lambda = (A_1 \bullet C, \ldots, A_m \bullet C)^T$. Finally, setting $S^0 = C - \sum_{i=1}^m \lambda_i^0 A_i$, we obtain a starting point (X^0, λ^0, S^0) that is obviously feasible.

Having computed this starting point, the remaining parameters of Algorithm 4.1 are initialized by

$$\tau_0 = \frac{\|\phi(X^0, S^0, 0)\|}{5} \quad \text{and} \quad \beta = \max\left\{2.1 \cdot \sqrt{n}, 1.5 \cdot \frac{\|\phi(X, S, \tau_0)\|}{\tau_0}\right\}.$$

We terminate the iteration if $\tau_k/n < 10^{-6}$ (recall that we parameterize the central path conditions by τ^2) and if the feasibility measure

$$\max\left\{\frac{\left\|\left[b_i - A_i \bullet X^k\right]_{i=1}^m\right\|_2}{\max\{1, \|b\|_2\}}, \frac{\left\|C - S^k - \sum_{i=1}^m \lambda_i^k A_i\right\|_F}{\max\{1, \|C\|_2\}}\right\}$$

is smaller than 10^{-10}. The reason for dividing τ_k by n is based on the fact that $\|\phi(X^k, S^k, 0)\|_F = O(\tau_k)$. Since we want to have $\|\phi(X^k, S^k, 0)\|_F$ small, it seems reasonable to terminate if τ_k gets small. However, getting $\|\phi(X^k, S^k, 0)\|_F$ small becomes increasingly difficult the larger the dimension of the matrices X^k and S^k are, since we take the Frobenius norm. In order to make our termination criterion more or less independent of the dimension of X^k and S^k, we therefore decided to use the above-mentioned stopping rule.

Note that, theoretically, this feasibility measure is always zero for our method. Numerically, however, the situation is different. While the dual feasibility does not really cause any troubles (mainly because S^k gets defined in such a way that the dual feasibility is zero), we sometimes observed difficulties with respect to the primal feasibility. In order to decrease the primal infeasibility, we therefore exploit a projection technique also used in SDPT3: After computing a Newton direction $(\Delta X, \Delta\lambda, \Delta S, \Delta\tau)$, we check whether the inequality $\|[A_i \bullet (X + \Delta X)]_{i=1}^m - b\| > \|[A_i \bullet X]_{i=1}^m - b\|$ holds. If this inequality is satisfied, we replace ΔX by its orthogonal projection onto the nullspace $\{U \in \mathcal{S}^{n\times n} \mid A_i \bullet U = 0,\ i = 1, \ldots, m\}$. As a consequence of this procedure, the feasibility stays close to the machine precision for all test problems.

In the SDPT3 code, there are eight test problems. The results for different sizes are shown in Tables 1–2. To compare our results with those from interior-point methods, the number of iterations for the infeasible path following algorithm from the SDPT3 package are also printed; more precisely, we present the results for the three most popular interior-point methods, namely, those based on the AHO-, HKM-, and NT-directions; see, e.g., [29] for some further details.

In Tables 1–2 we report the average iteration counts for the first ten instances of each problem using different problem dimensions. (Note that all test problems depend on some random numbers, so we decided to give the average results over ten runs for each problem.) For most smaller problems, it seems that Algorithm 4.1 needs fewer iterations than all interior-point methods. On the other hand, the termination criterion is different and not directly comparable. Furthermore, we should note that one iteration of Algorithm 4.1 is (usually) more expensive than one iteration of an interior-point method due to the fact that we have to calculate a matrix square root. In any case, we stress that the results we obtain for Algorithm 4.1 seem to be considerably better than those reported for a related method by Chen and Tseng [8].

Finally, Table 3 gives some results for the application of Algorithm 4.1 to some test problems from the SDPLIB; cf. Borchers [3]. In this table, we present for each test problem the number of iterations, the final value of the smoothing parameter τ, the relative duality gap, as well as the feasibility measure at the final iterate. Note that the duality gap is negative for many test problems because the matrices generated by

TABLE 1
Average number of iterations for small SDPs.

			AHO	HKM	NT	Alg. 4.1
Problem	n	m	Iter	Iter	Iter	Iter
random	10	10	8.2	13.5	12.6	6.5
Norm min	20	6	8.0	9.3	10.1	6.7
Cheby	20	11	7.9	9.8	9.9	5.9
Maxcut	10	10	7.4	8.2	8.4	5.5
ETP	20	10	11.8	14.7	12.2	10.8
Lovasz	10	≈ 25	7.6	8.7	8.7	9.1
LogCheby	60	6	10.4	10.9	11.1	10.8
ChebyC	40	11	7.6	8.5	9.0	5.2

TABLE 2
Average number of iterations for medium-sized SDPs.

			AHO	HKM	NT	Alg. 4.1
Problem	n	m	Iter	Iter	Iter	Iter
random	20	20	10.2	14.4	13.1	8.9
Norm min	40	11	8.5	10.1	10.7	7.7
Cheby	40	21	7.7	9.7	10.0	6.1
Maxcut	21	21	8.2	9.6	9.6	6.3
ETP	40	20	12.6	16.7	13.3	14.0
Lovasz	21	≈ 105	9.7	10.1	10.4	12.7
LogCheby	120	11	12.3	13.2	13.1	13.5
ChebyC	80	21	8.3	9.2	9.4	6.1

our method are not necessarily positive semidefinite. (This, in fact, was the reason why we had to take a different termination criterion than interior-point methods.)

7. Final remarks. We have presented two new characterizations of the central path conditions for SDPs. These characterizations were used to derive a smoothing-type method for the solution of SDPs. The search directions generated by these methods are automatically symmetric, and the method was shown to be globally and locally superlinearly convergent under suitable assumptions. The numerical results are very promising, and it will certainly be worthwhile to improve these methods. For example, it is interesting to investigate the question of how the matrix square roots could be computed in a more efficient way.

Furthermore, we note that, for the purposes of this paper, both the smoothed minimum function and the smoothed Fischer–Burmeister function can be handled in the same way, since the method discussed here has exactly the same theoretical (and similar numerical) properties for both functions. However, it was pointed out by one of the referees that, in general, these two functions might have very different properties when applied to SDPs. In fact, it has been shown in [27] that the minimum function is (strongly) semismooth, whereas it is currently not known whether this is true for the Fischer–Burmeister function. Moreover, the very recent paper [21] shows that the matrix of the linear system (6.6) is symmetric positive definite for the smoothed minimum function, whereas it is only positive definite (usually not symmetric) for the smoothed Fischer–Burmeister function.

Acknowledgement. The authors would like to thank the referees for useful comments and for pointing out some recent references on related Newton-type methods for SDPs.

TABLE 3
Selected problems from SDPLIB.

Problem	n	m	Iter	τ	Rel. gap.	Feas. measure
arch0	335	174	44	5.9e-05	-1.037695e-05	3.499271e-13
arch2	335	174	43	9.4e-05	5.195117e-05	6.384054e-13
arch4	335	174	47	1.3e-04	7.104246e-05	4.250081e-13
arch8	335	174	78	1.1e-04	-3.542500e-06	9.052898e-13
gpp100	100	101	18	9.9e-05	-3.042351e-06	2.123789e-15
gpp124-1	124	125	19	1.0e-04	-1.912270e-05	1.195466e-14
gpp124-2	124	125	19	7.0e-05	-2.615799e-06	2.176315e-15
gpp124-3	124	125	16	1.1e-04	-2.025964e-06	1.887616e-15
gpp124-4	124	125	20	6.5e-05	-4.630696e-06	1.331054e-14
gpp250-1	250	250	19	2.1e-04	-1.294246e-05	2.476407e-14
gpp250-2	250	250	17	1.9e-04	-5.718274e-06	2.130110e-14
gpp250-3	250	250	16	1.7e-04	-3.424270e-06	1.258957e-14
gpp250-4	250	250	17	2.2e-04	-2.061362e-06	3.875924e-14
mcp100	100	100	10	1.8e-06	-2.068888e-09	6.683366e-16
mcp124-1	124	124	15	2.6e-05	-5.421892e-09	5.389812e-16
mcp124-2	124	124	10	8.6e-05	-2.899952e-07	7.948236e-16
mcp124-3	124	124	9	2.9e-05	-8.401942e-08	6.089117e-16
mcp124-4	124	124	9	5.8e-07	-1.672955e-09	7.768182e-16
mcp250-1	250	250	14	5.7e-05	-9.849429e-08	9.333300e-16
mcp250-2	250	250	11	1.1e-04	-4.597849e-07	1.003759e-15
mcp250-3	250	250	11	8.7e-05	-1.589781e-07	1.081043e-15
mcp250-4	250	250	11	4.9e-05	-1.219282e-07	1.007461e-15
mcp500-1	500	500	26	3.7e-04	-5.734057e-07	1.087701e-15
mcp500-2	500	500	14	1.4e-04	-3.508263e-07	1.429353e-15
mcp500-3	500	500	11	3.7e-04	-1.295995e-06	1.526598e-15
mcp500-4	500	500	10	5.6e-05	-7.792705e-08	1.574678e-15
theta1	50	104	13	3.9e-05	-1.307451e-07	6.261965e-17
theta2	100	498	15	1.7e-05	-1.766186e-07	1.049632e-14
theta3	150	1106	15	7.8e-05	-1.009075e-06	1.998401e-15
theta4	200	1949	15	9.5e-06	-1.006602e-07	3.996803e-15
truss1	13	6	8	3.3e-09	-3.003989e-09	3.621438e-15
truss2	133	58	13	1.3e-05	-7.869353e-06	2.209316e-14
truss3	31	27	14	5.0e-06	-5.614971e-10	2.660288e-15
truss4	19	12	7	1.3e-05	-3.397473e-05	1.324462e-15
truss5	331	208	16	2.1e-04	-6.840872e-07	1.803378e-14
truss6	451	172	21	2.5e-04	-2.430952e-04	4.601362e-13
truss7	301	86	25	5.9e-05	-2.316906e-07	3.795188e-13
truss8	628	496	20	1.7e-04	-4.805725e-06	3.038575e-14

REFERENCES

[1] F. ALIZADEH, *Interior point methods in semidefinite programming with applications to combinatorial optimization*, SIAM J. Optim., 5 (1995), pp. 13–51.

[2] F. ALIZADEH, J.-P. A. HAEBERLY, AND M. L. OVERTON, *Primal-dual interior-point methods for semidefinite programming: Convergence rates, stability and numerical results*, SIAM J. Optim., 8 (1998), pp. 746–768.

[3] B. BORCHERS, *SDPLIB* 1.2, *A library of semidefinite programming test problems*, Optim. Methods Softw., 11 (1999), pp. 597–611.

[4] J. V. BURKE AND S. XU, *A non-interior predictor-corrector path-following method for LCP*, in Reformulation: Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods, M. Fukushima and L. Qi, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999, pp. 45–63.

[5] J. V. BURKE AND S. XU, *A non-interior predictor-corrector path following algorithm for the monotone linear complementarity problem*, Math. Program., 87 (2000), pp. 113–130.

[6] 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.

[7] X. CHEN, H. QI, AND P. TSENG, *Analysis of Nonsmooth Symmetric-Matrix Functions with Applications to Semidefinite Complementarity Problems*, Technical report, Department of Mathematics, University of Washington, Seattle, WA, 2000.

[8] X. Chen and P. Tseng, *Non-Interior Continuation Methods for Solving Semidefinite Complementarity Problems*, Technical report, Department of Mathematics, University of Washington, Seattle, WA, 1999.
[9] S. Engelke and C. Kanzow, *Improved smoothing-type methods for the solution of linear programs*, Numer. Math., 90 (2002), pp. 487–507.
[10] S. Engelke and C. Kanzow, *On the solution of linear programs by Jacobian smoothing methods*, Ann. Oper. Res., 103 (2001), pp. 49–70.
[11] S. Engelke and C. Kanzow, *Predictor-Corrector Smoothing Methods for the Solution of Linear Programs*, Preprint 153, Department of Mathematics, University of Hamburg, Hamburg, Germany, 2000.
[12] S. Engelke and C. Kanzow, *Predictor-Corrector Smoothing Methods for Linear Programs with a More Flexible Update of the Smoothing Parameter*, Preprint 162, Department of Mathematics, University of Hamburg, Hamburg, Germany, 2001.
[13] A. Fischer, *A special Newton-type optimization method*, Optimization, 24 (1992), pp. 269–284.
[14] M. Fukushima, Z.-Q. Luo, and P. Tseng, *Smoothing functions for second-order-cone complementarity problems*, SIAM J. Optim., 12 (2001), pp. 436–460.
[15] C. Helmberg, F. Rendl, R. J. Vanderbei, and H. Wolkowicz, *An interior-point method for semidefinite programming*, SIAM J. Optim., 6 (1996), pp. 342–361.
[16] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, U.K., 1985.
[17] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, U.K., 1991.
[18] H. Jiang, *Smoothed Fischer-Burmeister Equation Methods for the Complementarity Problem*, Technical report, Department of Mathematics, University of Melbourne, Melbourne, Australia, 1997.
[19] C. Kanzow, *Some noninterior continuation methods for linear complementarity problems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 851–868.
[20] C. Kanzow and C. Nagel, *Semidefinite Programs: New Search Directions, Smoothing-Type Methods, and Numerical Results*, Preprint 163, Department of Mathematics, University of Hamburg, Hamburg, Germany, 2001.
[21] C. Kanzow and C. Nagel, *Some Practical Aspects of a Newton-Type Method for Semidefinite Programs*, Preprint 243, Institute of Applied Mathematics and Statistics, University of Würzburg, Würzburg, Germany, 2001.
[22] M. Kojima, M. Shida, and S. Shindoh, *Local convergence of predictor-corrector infeasible-interior-point algorithms for SDPs and SDLCPs*, Math. Programming, 80 (1998), pp. 129–160.
[23] M. Kojima, M. Shida, and S. Shindoh, *A predictor-corrector interior-point algorithm for the semidefinite linear complementarity problem using the Alizadeh–Haeberly–Overton search direction*, SIAM J. Optim., 9 (1999), pp. 444–465.
[24] Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*, Studies in Appl. Math. 13, SIAM, Philadelphia, 1994.
[25] S. Smale, *Algorithms for solving equations*, in Proceedings of the International Congress of Mathematicians, Providence, 1987, AMS, Providence, RI, pp. 172–195.
[26] D. Sun and J. Sun, *Semismooth matrix valued functions*, Math. Oper. Res., 27 (2002), pp. 150–169.
[27] J. Sun, D. Sun, and L. Qi, *From Strong Semismoothness of the Squared Smoothing Matrix Function to Semidefinite Complementarity Problems*, Applied Mathematics Report AMR 00/20, School of Mathematics, University of New South Wales, Sydney, Australia, 2000.
[28] M. J. Todd, K. C. Toh, and R. H. Tütüncü, *On the Nesterov–Todd direction in semidefinite programming*, SIAM J. Optim., 8 (1998), pp. 769–796.
[29] K. C. Toh, M. J. Todd, and R. H. Tütüncü, *SDPT3—A Matlab software package for semidefinite programming, Version* 2.1, Optim. Methods Softw., 11 (1999), pp. 545–581.
[30] P. Tseng, *Merit functions for semi-definite complementarity problems*, Math. Programming, 83 (1998), pp. 159–185.
[31] P. Tseng, *Error bounds and superlinear convergence analysis of some Newton-type methods in optimization*, in Nonlinear Optimization and Related Topics, G. D. Pillo and F. Giannessi, eds., Kluwer Academic Publishers, Norwell, MA, 2000, pp. 445–462.
[32] N. Yamashita and M. Fukushima, *A new merit function and a descent method for semidefinite complementarity problems*, in Reformulation: Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods, M. Fukushima and L. Qi, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999, pp. 405–420.
[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.

SIAM J. OPTIM.
Vol. 13, No. 1, pp. 24–43

ERROR BOUNDS FOR ABSTRACT LINEAR INEQUALITY SYSTEMS*

KUNG FU NG† AND WEI HONG YANG†

Abstract. In this paper we study error bounds of the abstract linear inequality system (A, C, b): $Ax \leqslant b$, where A is a bounded linear operator from a Banach space X to a Banach space Y partially ordered by a closed convex cone C. We also give some general results on the existence of error bounds for a convex function F; we show in particular that F has an error bound if and only if the directional derivative of the distance function (to the solution set S) at each boundary point of S along any nontangential direction is bounded by the derivative of F. As an application, we prove that if C is a polyhedral cone, then the system (A, C, b) has an error bound. When Y is a Hilbert space, our results can be expressed in terms of the angles between $Ax - b - P_{-C}(Ax - b)$ and $\mathrm{Im}(A)$, or in terms of the angles between $\mathrm{Im}(A)$ and the nonvertex supporting hyperplanes of C. In the case in which $X = \mathbb{R}^n$ and C is an "ice-cream" cone, we identify exactly when (A, C, b) has an error bound.

Key words. error bound, abstract linear inequality system, polyhedral cone, angles between subspaces, "ice-cream" cone

AMS subject classifications. 90C31, 90C25, 49J52

PII. S1052623401388914

1. Introduction and preliminaries.

For convenience, we first set notations that will be used throughout the paper. X, Y denote Banach spaces, and C is a closed convex cone in Y introducing a partial order in the usual way: for $y_1, y_2 \in Y$

$$y_1 \leqslant y_2 \quad \text{if and only if} \quad y_2 - y_1 \in C.$$

F denotes a continuous convex function on X, and $S := \{x \in X : F(x) \leqslant 0\}$. A denotes a bounded linear operator on X into Y, and b denotes a vector in Y. For A, b as above, we have a continuous convex function $\varphi_{A,b}$ (henceforth denoted by φ_b if there is no need to emphasize A):

$$\varphi_b(x) = \mathrm{dist}(Ax - b, -C) \quad \forall x \in X; \tag{1.1}$$

let S_C^b denote the set of all x satisfying $\varphi_b(x) \leqslant 0$, that is,

$$S_C^b = \{x : Ax - b \leqslant 0\}. \tag{1.2}$$

We will use (A, C, b) to denote the abstract inequality system

$$Ax - b \leqslant 0. \tag{1.3}$$

It is said to have an error bound (and τ is called an error bound for the system) if there exists a $\tau > 0$ such that

$$\mathrm{dist}(x, S_C^b) \leqslant \tau \varphi_b(x) \quad \forall x \in X;$$

*Received by the editors May 4, 2001; accepted for publication (in revised form) December 19, 2001; published electronically May 15, 2002. This research was supported by a direct grant (CUHK) and an earmarked grant from the Research Grant Council of Hong Kong.
http://www.siam.org/journals/siopt/13-1/38891.html
†Department of Mathematics, The Chinese University of Hong Kong, Shatin, New Territory, Hong Kong (kfng@math.cuhk.edu.hk, whyang@math.cuhk.edu.hk).

this is clearly the case if and only if φ_b has an error bound [7, 2, 9, 12, 10, 11].

V will denote a nonempty closed convex set in X. Let $\bar{x} \in V$. If there exists a continuous linear functional f on X with norm 1 such that

$$f(\bar{x}) = \max\{f(x) : x \in V\},$$

then $\bar{x}$ is called a support point of V, and f is called a (normalized) supporting functional of V at $\bar{x}$. The set of all support points of V will be denoted by supp(v).

Given a point x in the topological boundary bd(V) of V, $h \in X\backslash\{0\}$ is called a normal (direction) of V at x if $\operatorname{dist}(x + h, V) = \|h\|$; in view of the variational inequality, this is the case if and only if

$$\operatorname{dist}(x + th, V) = t\|h\| \quad \forall t \in [0, 1]. \tag{1.4}$$

More generally, for $\gamma \in (0, 1]$, $h \in X$ is called a γ-normal of V at x if

$$\operatorname{dist}(x + th, V) \geqslant t\gamma\|h\| \quad \forall t \in [0, 1]. \tag{1.5}$$

The set of all γ-normals of V at x will be denoted by $N_V(x, \gamma)$. Note that sometimes $N_V(x)$ has been used to denote $N_V(x, 1)$, and that $N_V^1(x)$ denotes the set of all unit vectors in $N_V(x)$. Thus $N_V(x, 1)$ equals the cone generated by $N_V^1(x)$, considered in [9]. Another closely related concept concerns the so-called Bishop–Phelps cone $P(f, \gamma)$, which is defined by

$$P(f, \gamma) = \{h \in X : \gamma\|h\| \leqslant f(h)\},$$

where f is a continuous linear functional on X with norm 1. $P(f, \gamma)$ is nonempty whenever $0 < \gamma < 1$. ($P(f, 1)$ is also nonempty if X is assumed reflexive.) Note also that

$$P(f, \gamma) = \bigcap_{0<\gamma'<\gamma} P(f, \gamma'). \tag{1.6}$$

For $\gamma \in (0, 1]$ and $x \in \text{supp(V)}$, let $N_V^{'}(x, \gamma)$ denote the union

$$\bigcup_f P(f, \gamma), \tag{1.7}$$

where f runs over all normalized continuous linear functionals on X supporting V at x.

LEMMA 1.1. *Let $x \in \text{supp(V)}$ and $\gamma \in (0, 1]$. Then*

$$N_V^{'}(x, \gamma) \subseteq N_V(x, \gamma). \tag{1.8}$$

Proof. Let $0 \neq h \in N_V^{'}(x, \gamma)$; there exists an $f \in X^*$ with $\|f\| = 1$ such that f supports V at x and $\gamma\|h\| \leqslant f(h)$. Then, for any $v \in V$ and $t > 0$,

$$\|x + th - v\| \geqslant f(x + th - v) \geqslant f(th) \geqslant t\gamma\|h\|,$$

showing that $\operatorname{dist}(x + th, V) \geqslant t\gamma\|h\|$, and hence that (1.5) holds and $h \in N_V(x, \gamma)$. ∎

Recall that the tangent cone $T_V(x)$ of V at x is defined by

$$T_V(x) = \left\{h \in X : \lim_{t\to 0+} \frac{\operatorname{dist}(x + th, V)}{t} = 0\right\}. \tag{1.9}$$

LEMMA 1.2. *Let $x \in \mathrm{bd(V)}$ and $\gamma \in (0,1]$. Then*

$$N_V(x,\gamma)\backslash\{0\} \subseteq X\backslash T_V(x). \tag{1.10}$$

Proof. Let $0 \neq h \in N_V(x,\gamma)$; then (1.5) holds. It follows that

$$\frac{\mathrm{dist}(x+th,V)}{t} \geqslant \gamma\|h\| > 0 \quad \forall t \in (0,1],$$

and hence $h \notin T_V(x)$ by (1.9). □

We have observed that $N_V(x,1)$ is simply the set of all normals to V at x and has played an important role in the error bound problems (cf. [9, 12, 10]). In attempting to generalize this study to the general setting of Banach spaces (in place of finite dimensional or reflexive spaces), one must bear in mind that $N_V(x,1)$ can very well be empty for each $x \in \mathrm{supp(V)}$. The following result shows the nonemptiness of $N_V(x,\gamma)$ for many x if $\gamma < 1$. Our proof is based on a standard Banach space theory technique (cf. [13, Proposition 3.20]).

PROPOSITION 1.3. *Let $0 < \gamma < 1$, and let $y \in X\backslash V$. Then there exists $x \in \mathrm{supp(V)}$ such that*

$$y - x \in N_V^{'}(x,\gamma) \tag{1.11}$$

(in particular, $y-x \in N_V(x,\gamma)$ and $y-x \notin T_V(x)$). Moreover, if X is reflexive, then the result also holds for $\gamma = 1$.

Proof. Write d for the distance of y to V. Take $\epsilon > 0$ such that

$$\gamma \leqslant \frac{d - \sqrt{\epsilon}(d+\epsilon+\sqrt{\epsilon})}{(1+\epsilon)(d+\epsilon+\sqrt{\epsilon})}. \tag{1.12}$$

By the separation theorem, take $x_0^* \in X^*$ of norm 1 such that

$$\sup\langle x_0^*, V\rangle = \inf\langle x_0^*, y + dB\rangle = \langle x_0^*, y\rangle - d, \tag{1.13}$$

where B denotes the closed unit ball in X. Pick $x_0 \in V$ such that

$$\|x_0 - y\| \leqslant d + \epsilon. \tag{1.14}$$

Then, (1.13) and (1.14) imply that, for each $v \in V$,

$$\langle x_0^*, v\rangle \leqslant \langle x_0^*, y\rangle - d \leqslant \|x_0^*\|\|y - x_0\| + \langle x_0^*, x_0\rangle - d \leqslant \epsilon + \langle x_0^*, x_0\rangle,$$

showing that x_0^* is in the ϵ-subdifferential $\partial_\epsilon f(x_0)$ of f at x_0, where f denotes the indicator function of V (defined by $f(x) = 0$ or $+\infty$ according to $x \in V$ or $x \in X\backslash V$). By the Bronstead–Rockafellar theorem (cf. [13]), there exist $x_\epsilon \in dom\ f = V$ and

$$x_\epsilon^* \in \partial f(x_\epsilon) \qquad (\text{and thus} \quad \sup\langle x_\epsilon^*, V\rangle = \langle x_\epsilon^*, x_\epsilon\rangle \quad \text{and} \quad x_\epsilon \in \mathrm{supp(V)}) \tag{1.15}$$

such that $\|x_\epsilon - x_0\| \leqslant \sqrt{\epsilon}$, $\|x_\epsilon^* - x_0^*\| \leqslant \sqrt{\epsilon}$. Then one has from (1.14) that

$$\|x_\epsilon - y\| \leqslant d + \epsilon + \sqrt{\epsilon}, \tag{1.16}$$

and it follows from (1.13) that

$$\langle x_\epsilon^*, x_\epsilon - y\rangle = \langle x_\epsilon^* - x_0^*, x_\epsilon - y\rangle + \langle x_0^*, x_\epsilon - y\rangle \leqslant \sqrt{\epsilon}(d+\epsilon+\sqrt{\epsilon}) - d;$$

therefore we have, by (1.16), that

$$\begin{aligned}\left\langle \frac{x_\epsilon^*}{\|x_\epsilon^*\|}, y - x_\epsilon \right\rangle &\geqslant \frac{d - \sqrt{\epsilon}(d + \epsilon + \sqrt{\epsilon})}{1 + \epsilon} \\ &\geqslant \frac{d - \sqrt{\epsilon}(d + \epsilon + \sqrt{\epsilon})}{1 + \epsilon} \frac{\|x_\epsilon - y\|}{d + \epsilon + \sqrt{\epsilon}} \\ &\geqslant \gamma \|x_\epsilon - y\|,\end{aligned}$$

showing that $y - x_\epsilon \in N_V'(x_\epsilon, \gamma)$. Together with Lemmas 1.1 and 1.2, the first assertion of Proposition 1.3 is proved. Moreover, suppose that X is reflexive. Then there exists $x_0 \in V$ satisfying (1.14) with $\epsilon = 0$; hence the supremum in (1.13) is attained at x_0, and the whole argument is then seen to be valid with $\gamma = 1, \epsilon = 0$. □

2. Error bounds for convex functions. We continue to use the notations introduced in section 1; in particular, F is a continuous convex function on X, and $S = \{x : F(x) \leqslant 0\}$. To avoid triviality, we assume throughout that $\emptyset \neq S \neq X$. Suppose that G is another continuous convex function such that $G(x) \leqslant 0$ if and only if $F(x) \leqslant 0$:

$$S = \{x : F(x) \leqslant 0\} = \{x : G(x) \leqslant 0\} \tag{2.1}$$

(e.g., the function $x \to \operatorname{dist}(x, S)$ has this property of G). Recall that the directional derivative $F'(x; h) = \lim_{t \to 0+} \frac{F(x+th) - F(x)}{t}$ always exists for $x, h \in X$, and similarly for G. We begin with a general result for the existence of error bounds for F.

THEOREM 2.1. *Assume* (2.1); *let* $\gamma \in (0, 1)$ *and* $M > 0$ *be such that, for each* $x \in \operatorname{supp}(S)$ *and* $h \in N_S'(x, \gamma)$,

$$G'(x; h) \leqslant M F'(x; h). \tag{2.2}$$

Suppose that G *has an error bound: there exists* $\tau > 0$ *such that*

$$\operatorname{dist}(x, S) \leqslant \tau [G(x)]_+ \quad \forall x \in X. \tag{2.3}$$

Then F *also has an error bound:*

$$\operatorname{dist}(x, S) \leqslant \gamma^{-1} \tau M [F(x)]_+ \quad \forall x \in X. \tag{2.4}$$

Moreover, if X *is reflexive, then the result remains true for* $\gamma = 1$.

Proof. Let $y \in X \backslash S$. By Proposition 1.3 (which is valid for $\gamma \in (0, 1)$ and also for $\gamma = 1$ if X is reflexive), take $x \in \operatorname{supp}(S)$ such that $h := y - x \in N_S'(x, \gamma)$; thus by (1.5) and (1.8) one has for each $t \in (0, 1]$ that

$$\operatorname{dist}(x + th, S) \geqslant t\gamma \|h\| \geqslant t\gamma \operatorname{dist}(x + h, S) > 0, \tag{2.5}$$

showing, in particular, that $x + th \notin S$. It follows from (2.3) that

$$\tau G(x + th) \geqslant \operatorname{dist}(x + th, S) \geqslant t\gamma \operatorname{dist}(x + h, S),$$

which implies that $\tau G'(x; h) \geqslant \gamma \operatorname{dist}(x + th, S) > 0$ because $G(x) = 0$. Since F is convex and $F(x) = 0$, it follows from (2.2) that

$$M\tau F(x + h) \geqslant M\tau F'(x; h) \geqslant \tau G'(x; h) \geqslant \gamma \operatorname{dist}(x + h, S),$$

that is, $M\tau F(y) \geqslant \gamma \mathrm{dist}(y, S)$ is valid for each $y \in X\backslash S$. Therefore (2.4) is seen to hold. □

Let D be the convex function on X defined by

$$D(x) = \mathrm{dist}(x, S) \quad \forall x \in X. \tag{2.6}$$

It is trivial to see that D has an error bound $\tau = 1$. By applying the preceding theorem to $G = D$, we have the following generalization of a theorem of Lewis and Pang [9].

THEOREM 2.2. *Let $\tau > 0$ and $\gamma \in (0, 1)$. Then one has* (i) ⇒ (ii) ⇒ (iii) ⇒ (iv) ⇒ (i*) *among the following:*

(i) *$\gamma\tau$ is an error bound for F.*
(ii) *$D'(x; h) \leqslant \gamma\tau F'(x; h)$ for each $x \in \mathrm{supp}(S)$ and $h \in X\backslash T_S(x)$.*
(iii) *$D'(x; h) \leqslant \gamma\tau F'(x; h)$ for each $x \in \mathrm{supp}(S)$ and $h \in N_S(x, \gamma)$.*
(iv) *$D'(x; h) \leqslant \gamma\tau F'(x; h)$ for each $x \in \mathrm{supp}(S)$ and $h \in N'_S(x, \gamma)$.*
(i*) *τ is an error bound for F.*

Moreover, the result remains true for $\gamma = 1$ if X is reflexive.

Proof. By Theorem 2.1 (applied to $\gamma\tau, 1$ in place of M, τ), (iv) ⇒ (i*). In view of Lemmas 1.1 and 1.2, it is trivial that (ii) ⇒ (iii) ⇒ (iv). To prove (i) ⇒ (ii), let $x \in \mathrm{bd}(S)$ and $h \in X\backslash T_S(x)$. Then $D(x + th) > 0$ and so $x + th \notin S$ for all $t > 0$. By (i), one has

$$D(x + th) \leqslant \gamma\tau[F(x + th)]_+ = \gamma\tau F(x + th);$$

since $D(x) = 0 = F(x)$, it follows from taking limits (after dividing by t) that

$$D'(x; h) \leqslant \gamma\tau F'(x; h). \qquad \square$$

Remark 2.2.1. For $\gamma = 1$, it is clear from (1.4) that $D'(x; h) = \|h\|$. Thus the implications (i) ⇒ (iii) ⇒ (i*) provide an extension of [9, Theorem 1], in which the case of $X = \mathbb{R}^n$ was considered.

COROLLARY 2.3. *τ is an error bound for F if and only if*

$$\mathrm{(ii^*)} \quad \mathrm{D'(x; h) \leqslant \tau F'(x; h)} \textit{ for each } \mathrm{x \in supp(S)} \textit{ and } \mathrm{h \in X\backslash T_S(x)}.$$

Proof. Apply Theorem 2.2 and let $\gamma \to 1$. □

3. Error bounds for abstract linear inequality systems. We continue to use the notation set from section 1. In particular, A is a bounded linear operator from X into Y, which is partially ordered by a closed convex cone C. Let $b \in Y$; let φ_b, S^b_C, and (A, C, b) be defined by (1.1), (1.2), and (1.3). Note that

$$\varphi_b(x) = g(Ax - b) \quad \forall x \in X, \tag{3.1}$$

where $g : Y \to \mathbb{R}$ is defined by

$$g(y) = \mathrm{dist}(y, -C). \tag{3.2}$$

By the chain rules [3, Theorem 2.3.10],

$$\varphi'_b(x; h) = g'(Ax - b; Ah), \qquad \partial\varphi_b(x) = \{A^*y^* : y^* \in \partial g(Ax - b)\}. \tag{3.3}$$

Let $M := \mathrm{Im}(A)$ and, for any $y^* \in Y^*$, let $\|y^*\|_M$ denote the norm of the restriction $y^*|_M$ of y^* to M:

$$\|y^*\|_M := \sup\{|\langle y^*, y\rangle| : y \in M, \|y\| = 1\}.$$

Here and throughout, we assume that $A \neq 0$, so that $M \neq \{0\}$.

LEMMA 3.1. *Let* $M := \mathrm{Im}(A)$, $y^* \in Y^*$, *and let* $A^*y^* \in X^*$ *be defined by*

$$(A^*y^*)(x) = y^*(Ax) \quad \forall x \in X.$$

Then

$$\|A^*y^*\| \leqslant \|A\|\|y^*\|_M. \tag{3.4}$$

If $\mathrm{Im}(A)$ *is assumed closed, then there exists* $\lambda > 0$ *such that*

$$\|y^*\|_M \leqslant \lambda\|A^*y^*\|. \tag{3.5}$$

Proof. The proof for (3.4) is straightforward. To prove (3.5), we let

$$\lambda := \sup\{\mathrm{dist}(0, A^{-1}y) : y \in \mathrm{Im}(A), \|y\| = 1\}. \tag{3.6}$$

Since $\mathrm{Im}(A)$ is closed, it follows from the open mapping theorem that $\lambda < +\infty$. By (3.6), for each $y \in \mathrm{Im}(A)$ with $\|y\| < 1$ there exists a z with $\|z\| \leqslant \lambda$ such that $Az = y$. Then

$$|y^*(y)| = |A^*y^*(z)| \leqslant \lambda\|A^*y^*\|,$$

which implies (3.5). □

THEOREM 3.2. *Suppose that* S_C^b *is nonempty and that* $M = \mathrm{Im}(A)$ *is closed. Then* (A, C, b) *has an error bound if and only if*

$$\inf_{\substack{x \in X \setminus S_C^b \\ \eta \in \partial g(Ax-b)}} \|\eta\|_M > 0. \tag{3.7}$$

Proof. Clearly φ_b is continuous and convex, and $S_C^b = \{x \in X : \varphi_b(x) \leqslant 0\}$. Therefore (A, C, b) has an error bound if and only if φ_b also does. By [5, Theorem 2.2] and (3.3), this is the case if and only if

$$\inf_{\substack{x \in X \setminus S_C^b \\ y^* \in \partial g(Ax-b)}} \|A^*y^*\| > 0. \tag{3.8}$$

Take $\lambda > 0$ as in (3.6). Then, by (3.4) and (3.5), one has

$$\lambda^{-1} \inf_{\substack{x \in X \setminus S_C^b \\ y^* \in \partial g(Ax-b)}} \|y^*\|_M \leqslant \inf_{\substack{x \in X \setminus S_C^b \\ y^* \in \partial g(Ax-b)}} \|A^*y^*\| \leqslant \inf_{\substack{x \in X \setminus S_C^b \\ y^* \in \partial g(Ax-b)}} \|A\|\|y^*\|_M,$$

showing that (3.7) and (3.8) are equivalent. □

The following result shows that, when $\mathrm{Im}(A)$ is closed, for the consideration of the existence of error bounds of (A, C, b) one may replace X, A by $\mathrm{Im}(A)$, I_A, respectively, where I_A denotes the identity operator from $\mathrm{Im}(A)$ into Y.

COROLLARY 3.3. *Suppose that* $\mathrm{Im}(A)$ *is closed in* Y. *Then* (A, C, b) *has an error bound if and only if* (I_A, C, b) *has an error bound.*

Proof. Let $M := \mathrm{Im}(A)$. Then $M = \mathrm{Im}(I_A)$. Note also that the solution set for (I_A, C, b) equals $M \cap (b - C)$. Thus, applying Theorem 3.2 to (I_A, C, b), one finds that (I_A, C, b) has an error bound if and only if

$$\inf_{\substack{y \in M \setminus (b-C) \\ \eta \in \partial g(I_A y-b)}} \|\eta\|_M > 0,$$

which is seen to be exactly (3.7) by writing Ax for y in M. □

The next result deals with the special case in which C is a polyhedral cone: there exist $f_1, f_2, \ldots, f_r \in Y^*$ such that

$$-C = \{y \in Y : \ f_i(y) \leqslant 0 \quad \forall i = 1, 2, \ldots, r\}. \tag{3.9}$$

That is,

$$-C = \{y \in Y : \ \psi(y) \leqslant 0\},$$

where $\psi : Y \to R$ is the max function defined by

$$\psi(y) = \max f_i(y) \quad \forall y \in Y. \tag{3.10}$$

Further, define $\Psi : X \to R$ by

$$\Psi(x) = \psi(Ax - b) \quad \forall x \in X. \tag{3.11}$$

Thus

$$x \in S_C^b \iff \Psi(x) \leqslant 0 \iff (A^* f_i)(x) - f_i(b) \leqslant 0 \quad \forall i = 1, 2, \ldots, r. \tag{3.12}$$

The following theorem extends a result of Robinson [14, pp. 759–760] from the finite dimensional case to general Banach spaces.

THEOREM 3.4. *Suppose that C is a polyhedral cone in Y. Then the system (A, C, b) has an error bound for any $b \in Y$.*

Proof. By Ioffe's Theorem [8], there exists a $\tau > 0$ such that

$$\operatorname{dist}(x, S_C^b) \leqslant \tau[\Psi(x)]_+, \quad x \in X. \tag{3.13}$$

By Corollary 2.3, one has

$$D'(x; h) \leqslant \tau \Psi'(x; h), \quad h \in X \backslash T_{S_C^b}(x), \tag{3.14}$$

where $D : X \to R$ is defined by $D(x) = \operatorname{dist}(x, S_C^b)$. Letting $F(x) := \operatorname{dist}(Ax - b, -C)$ and $\xi := \max\{\|f_1\|, \ldots, \|f_r\|\}$, we claim that

$$\Psi'(x; h) \leqslant \xi F'(x; h) \quad \forall x \in \operatorname{bd}(S_C^b), \ h \in X \backslash T_{S_C^b}(x). \tag{3.15}$$

By the chain rules and (3.11), $\Psi'(x; h) = \psi'(Ax - b; Ah)$ and similarly $F'(x; h) = g'(Ax - b; Ah)$, where $g : Y \to R$ is defined by

$$g(y) = \operatorname{dist}(y, -C) \quad \forall y \in Y.$$

Hence (3.15) holds if and only if

$$\psi'(Ax - b; Ah) \leqslant \xi g'(Ax - b; Ah) \quad \forall x \in \operatorname{bd}(S_C^b), \ h \in X \backslash T_{S_C^b}(x). \tag{3.16}$$

Thus (3.15) follows from

$$\psi'(y; k) \leqslant \xi g'(y; k) \quad \forall y \in \operatorname{bd}(-C), \ k \in Y, \tag{3.17}$$

which in turn is seen to hold because $\psi(y) = g(y)$ for $y \in \operatorname{bd}(-C)$ and

$$f_i(y + tk) \leqslant f_i(y + tk - w) \leqslant \xi \|y + tk - w\| \quad \forall w \in (-C),$$

that $\psi(y + tk) \leqslant \xi g(y + tk)$ for any $k \in Y$. Therefore (3.15) is established. Combining (3.14) and (3.15), it follows from Corollary 2.3 that

$$\operatorname{dist}(x, S_C^b) \leqslant \tau \xi [F(x)]_+ \quad \forall x \in X. \qquad \square$$

4. Angle conditions. The aim of this section is to interpret the condition given in (3.7) by virtue of an angle condition, and hence to seek further consequence geometrically. The angle condition given in Theorem 4.3 is more intuitive and sometimes easier to apply (see Example 4.10.1 and Corollary 4.5, for example). While we retain all the notation introduced in the preceding section, we will assume throughout that $M := \mathrm{Im}(A)$ is closed and that the norm for Y is induced by an inner product $\langle \cdot, \cdot \rangle$; that is, Y is a Hilbert space. $M^{\perp}$ denotes the orthogonal complement of M. For any $y, z \in Y \backslash \{0\}$ we define $\measuredangle(y, z) \in [0, \frac{\pi}{2}]$ by

$$\measuredangle(y, z) = \arccos \left\{ \frac{|\langle y, z \rangle|}{\|y\| \|z\|} \right\}. \tag{4.1}$$

If $Z \neq \{0\}$ is a closed vector subspace of Y, we also define

$$\measuredangle(y, Z) = \min\{\measuredangle(y, z) : z \in Z, \|z\| = 1\}, \tag{4.2}$$

i.e.,

$$\cos \measuredangle(y, Z) = \max \left\{ \left\langle \frac{y}{\|y\|}, z \right\rangle : z \in Z, \|z\| = 1 \right\}. \tag{4.3}$$

We recall that, for $y \in Y$,

$$\|y\|_Z := \sup_{\substack{z \in Z \\ \|z\|=1}} \{\langle y, z \rangle\} = \max_{\substack{z \in Z \\ \|z\|=1}} \{\langle y, z \rangle\}.$$

LEMMA 4.1. *Let $y \in Y$ with $\|y\| = 1$, let $Z \neq \{0\}$ be a closed vector subspace of Y, and let P_Z denote the orthogonal projection of Y onto Z. Then*

$$\|y\|_Z = \|P_Z(y)\| = \cos \measuredangle(y, Z). \tag{4.4}$$

Proof. Write $y = p + q$ with $p \in Z$ and $q \in Z^{\perp}$, the orthogonal complement of Z. Then

$$\begin{aligned} \|y\|_Z &= \max\{\langle p + q, z \rangle : z \in Z, \|z\| = 1\} \\ &= \max\{\langle p, z \rangle : z \in Z, \|z\| = 1\} \\ &= \|p\|, \end{aligned}$$

since $p \in Z$ and because of the Cauchy–Schwarz inequality. Combining this with (4.3), (4.4) is seen to hold. ∎

Recall that $g : Y \to \mathbb{R}$ is defined by $g(y) = \mathrm{dist}(y, -C)$ for each $y \in Y$, and that $P_{-C}(y)$ denotes the projection of y to $-C$.

LEMMA 4.2. *$\partial g(y) = \{\frac{y - P_{-C}(y)}{\|y - P_{-C}(y)\|}\}$ for each $y \in Y \backslash (-C)$.*

The proof given in [3, Proposition 2.5.4] can easily be adopted for our infinite dimensional setting (cf. [1, pp. 522–526]).

THEOREM 4.3. *Let $b \in Y$. The system (A, C, b) has an error bound if and only if*

$$\sup_{x \in X \backslash S_C^b} \measuredangle(Ax - b - P_{-C}(Ax - b), \mathrm{Im}(A)) < \frac{\pi}{2}. \tag{4.5}$$

In particular, if $-b + C \subseteq \mathrm{Im}(A)$ (e.g., A is onto Y), then the system (A, C, b) has an error bound.

Proof. Let $M := \mathrm{Im}(A)$. By Lemma 4.2 and (4.3) we have

$$\begin{aligned}\inf_{\substack{x\in X\setminus S_C^b\\ \eta\in\partial g(Ax-b)}}\big\{\|\eta\|_M\big\} &= \inf_{x\in X\setminus S_C^b}\left\|\frac{Ax-b-P_{-C}(Ax-b)}{\|Ax-b-P_{-C}(Ax-b)\|}\right\|_M\\ &= \inf_{x\in X\setminus S_C^b}\cos\measuredangle(Ax-b-P_{-C}(Ax-b),M)\\ &= \cos\left(\sup_{x\in X\setminus S_C^b}\measuredangle\big(Ax-b-P_{-C}(Ax-b),M\ \big)\right),\end{aligned}$$

which is positive if and only if (4.5) holds. Thus Theorem 4.3 follows from Theorem 3.2. □

THEOREM 4.4. *Suppose that for some $b\in Y$,*

$$\sup_{x\in X\setminus S_C^b}\measuredangle\big(Ax-b-P_{-C}(Ax-b),\mathrm{Im}(A)\big)<\frac{\pi}{2}. \tag{4.6}$$

Then

$$\sup_{x\in X\setminus S_C^0}\measuredangle\big(Ax-P_{-C}(Ax),\mathrm{Im}(A)\big)<\frac{\pi}{2}.$$

Proof. Let $M := \mathrm{Im}(A)$. Assume that there exists a sequence $\{x_n\}$ with $Ax_n\notin -C$ for each n such that

$$\lim_{n\to\infty}\measuredangle\big(Ax_n-P_{-C}(Ax_n),M\big)=\frac{\pi}{2}.$$

Assuming without loss of generality that $Ax_n-P_{-C}(Ax_n)$ is of norm 1, it follows from (4.4) that

$$\begin{aligned}&\lim_{n\to\infty}\cos\measuredangle\big(Ax_n-P_{-C}(Ax_n),M\big)\\ &\quad=\lim_{n\to\infty}\|P_M(Ax_n-P_{-C}(Ax_n))\|\\ &\quad=0.\end{aligned}$$

Let $U:=\{y\in Y|\|y\|\leqslant 1/4\}$. Then for each $y\in U, (Ax_n-y)\notin -C$ because Ax_n is of distance 1 to $-C$. Let $\eta_n(y):=Ax_n-y-P_{-C}(Ax_n-y)$ and

$$\phi_n(y):=\frac{\|P_M(\eta_n(y))\|}{\|\eta_n(y)\|}\qquad\forall y\in U.$$

Obviously $|\eta_n(y)-\eta_n(y')|\leqslant 2\|y-y'\|$; letting $y'=0$, it follows from $\|\eta_n(0)\|=1$ that $\|\eta_n(y)\|\geqslant 1/2$ for each $y\in U$. Then $\phi_n(y)$ is a sequence of equicontinuous functions of y from U to R. Let $\{\epsilon_n\}$ be a sequence of real numbers decreasing to 0. For each n we can therefore find a large enough integer k_n such that $\frac{b}{k_n}\in U$ and

$$\left|\arccos\phi_m\left(\frac{b}{k_n}\right)-\arccos\phi_m(0)\right|\leqslant\epsilon_n\qquad\text{for each}\quad m\in N. \tag{4.7}$$

Then, by Lemma 4.1, one has

$$\begin{aligned}
&\lim_{n\to\infty} \measuredangle(A(k_n x_n) - b - P_{-C}(A(k_n x_n) - b), M) \\
&\quad = \lim_{n\to\infty} \measuredangle\left(Ax_n - \frac{b}{k_n} - P_{-C}\left(Ax_n - \frac{b}{k_n}\right), M\right) \\
&\quad = \lim_{n\to\infty} \arccos \phi_n\left(\frac{b}{k_n}\right) = \lim_{n\to\infty} \arccos \phi_n(0) \qquad (4.8) \\
&\quad = \lim_{n\to\infty} \measuredangle(Ax_n - P_{-C}(Ax_n), M) \\
&\quad = \frac{\pi}{2}.
\end{aligned}$$

Here (4.8) follows from (4.7). Note further that since $\frac{b}{k_n} \in U$, $Ax_n - \frac{b}{k_n} \notin -C$, and so $k_n x_n \in X \backslash S_C^b$. This contradicts (4.6). □

Combining Theorems 4.3 and 4.4, we have the following.

COROLLARY 4.5. *Suppose that the system (A, C, b) has an error bound for some $b \in Y$. Then the system $(A, C, 0)$ has an error bound.*

The converse of the above corollary is inexact, as the following example shows.

Example 4.5.1. Let $X = Y = \mathbb{R}^3$, $C = \{(x, y, z) \in \mathbb{R}^3 | z \geqslant \sqrt{x^2 + y^2}\}$. Let $A : X \to Y$ be defined by $A(x, y, z) = (x, 0, 0)$ for each $(x, y, z) \in X$, and let $b = (0, 1, 1)$. Then

$$S_C^b = \{(x, y, z) \in \mathbb{R}^3 | A(x, y, z) - b \in -C\} = \{(x, y, z) \in \mathbb{R}^3 : x = 0\}.$$

Let $e_k = (\frac{1}{k}, 0, 0)$. Then $\operatorname{dist}(e_k, S_C^b) = \frac{1}{k}$. Pick $h_k := (\frac{1}{k}, -1, -\sqrt{1 + \frac{1}{k^2}}) \in -C$; then

$$\operatorname{dist}(Ae_k - b, -C) \leqslant \|Ae_k - b - h_k\| = \sqrt{1 + \frac{1}{k^2}} - 1 < \frac{1}{k^2}.$$

Therefore, $\operatorname{dist}(e_k, S_C^b) \geqslant k \operatorname{dist}(Ae_k - b, -C)$, showing that the system (A, C, b) has no error bound. On the other hand, one can show that $(A, C, 0)$ does have an error bound. (This can be seen immediately by applying Proposition 5.4 because $\operatorname{Im}(A) \cap C = \{0\}$.)

Extending (4.2) and (4.3), we consider the angles between two vector subspaces (cf. [4]). Let M, N be closed vector subspaces of Y. If $M \subseteq N$ or $N \subseteq M$, then we stipulate that $\measuredangle(M, N) = 0$. Suppose that $M \nsubseteq N$ and $N \nsubseteq M$ (equivalently, $M \cap (M \cap N)^\perp$ and $N \cap (M \cap N)^\perp$ are not $\{0\}$). Then we define

$$\measuredangle(M, N) := \inf\{\measuredangle(x, y) : x \in M \cap (M \cap N)^\perp, y \in N \cap (M \cap N)^\perp \text{ and } x, y \neq 0\}.$$

Recall that C denotes a closed convex cone in Y; to avoid triviality, we assume that $\{0\} \neq C \neq Y$. We say that a hyperplane supports C if $H \cap C$ is nonempty and C lies on one side of H: there exists $y^* \in Y^* \backslash \{0\}$ such that

$$\sup_{h\in H} \langle y^*, h\rangle = \inf_{c\in C} \langle y^*, c\rangle; \qquad (4.9)$$

note that the above supremum and the infimum are attained at the origin $0 \in H \cap C$ (and so H is a vector subspace). If, in addition, $H \cap C = \{0\}$, then H is called a vertex supporting hyperplane of C; otherwise it is a nonvertex one. There exist many nonvertex supporting hyperplanes of C. For example, pick a nonzero boundary point c of C and $d \in Y \backslash C$ such that $\|d - c\| < \frac{1}{2}\|c\|$. Then $0 \neq P_C(d) \in (d - P_C(d))^\perp \cap (C)$;

hence $(d-P_C(d))^\perp$ is a nonvertex supporting hyperplane. Let C^+, C^{+i}, respectively, denote

$$C^+ = \{y \in Y : \langle y^*, c\rangle \geqslant 0 \quad \forall c \in C\},$$
$$C^{+i} = \{y \in Y : \langle y^*, c\rangle > 0 \quad \forall c \in C\backslash\{0\}\}.$$

The following result follows immediately from (4.9).

LEMMA 4.6.

(a) *$H \subset Y$ is a supporting hyperplane of C if and only if it can be expressed in the form $H = \xi^\perp$ for some $\xi \in C^+\backslash\{0\}$.*
(b) *$H \subset Y$ is a hyperplane supporting C at some $\bar{c} \in C\backslash\{0\}$ if and only if it can be expressed in the form $H = \xi^\perp$ for some $\xi \in C^+\backslash\{0\}$ with $\xi(\bar{c}) = 0$.*
(c) *$H \subset Y$ is a vertex supporting hyperplane of C if and only if it can be expressed in the form $H = \xi^\perp$ for some $\xi \in C^{+i}$.*

The following lemma follows easily from the variational inequalities.

LEMMA 4.7. *Let $\xi \in Y$ with $\|\xi\| = 1$. Let $\bar{c} \in \mathrm{bd}(C)$. Then the following statements are equivalent:*

(i) *$\xi \in -C^+$ and $\xi(\bar{c}) = 0$.*
(ii) *$\xi \in N_C^1(\bar{c})$.*
(iii) *$-\xi \in N_{-C}^1(-\bar{c})$.*

To prepare for the proof of our next theorem, we define a function $\phi : Y \to R$ by

$$\phi(y) = \max\{\langle x, y\rangle : x \in -C, \|x\| = 1\}, \quad y \in Y.$$

It is easy to see that ϕ is a convex function. Moreover, in view of the variational inequalities, one has the following claim.

LEMMA 4.8. *Let $y \in Y\backslash -C$. Then*

(i) *$\phi(y) \leqslant 0 \Leftrightarrow P_{-C}(y) = 0$.*
(ii) *$\phi(y) = 0 \Leftrightarrow P_{-C}(y) = 0$ and there exists a $\bar{x} \in -C, \bar{x} \neq 0$ such that $\langle y, \bar{x}\rangle = 0$.*
(iii) *$\phi(y) < 0 \Leftrightarrow P_{-C}(y) = 0$ and $\langle y, z\rangle < 0$ for any $z \in -C$, $z \neq 0$.*

THEOREM 4.9. *Let $b \in Y$, $x \in X\backslash S_C^b$, and $M := \mathrm{Im}(A)$. Then the following assertions hold:*

(i) *$\measuredangle(Ax - b - P_{-C}(Ax - b), M) < \frac{\pi}{2}$.*
(ii) *There exists a nonvertex supporting hyperplane H of C such that*

$$\measuredangle(Ax - b - P_{-C}(Ax - b), M) + \measuredangle(H, M) \leqslant \frac{\pi}{2}. \tag{4.10}$$

Proof. For simplicity, we write z, q for $Ax - b$ and $P_{-C}(Ax - b)$, respectively. Then for any $y \in -C$ one has by the variational inequality that

$$\langle z - q, y - q\rangle \leqslant 0.$$

By considering $y = 0$ and $y = 2q$ separately, we have $\langle z - q, q\rangle = 0$. Thus, for any $y \in -C$,

$$\langle z - q, y\rangle \leqslant \langle z - q, q\rangle = 0 < \langle z - q, z - q\rangle = \langle z - q, z\rangle. \tag{4.11}$$

If $\measuredangle(z - q, M) = \frac{\pi}{2}$, then $\langle z - q, Au\rangle = 0$ for any $u \in X$, and so

$$\langle z - q, Au - b\rangle = \langle z - q, z\rangle > 0; \tag{4.12}$$

it follows from (4.11) that $Au - b \notin -C$ for any $u \in X$; this contradicts the assumption that $S_C^b \neq \emptyset$, and proves (i). Moreover, if $q = 0$ and $b \in M$, then the first term of

the left-hand member of (4.10) is zero, and hence (4.10) is seen to hold with any nonvertex supporting hyperplane H of C.

The remainder of the proof is devoted to showing (ii), and we consider the following two cases.

Case 1: $\phi(z) \geqslant 0$. Take H to be $(z-q)^{\perp}$, which is clearly a supporting hyperplane of C and $q \in H \cap (-C)$. We claim that it is nonvertex. Indeed, if $\phi(z) > 0$, then $q \neq 0$ by Lemma 4.8(i), and thus $0 \neq q \in H \cap (-C)$; if $\phi(z) = 0$, then by Lemma 4.8(ii), $q = 0$ and there exists $\bar{x} \in (-C)\backslash\{0\}$ such that $\langle z, \bar{x}\rangle = 0$; hence $0 \neq \bar{x} \in H \cap (-C)$. Therefore our claim is proved.

Moreover, whether $q = 0$ or not, we have $P_M(z - q) \neq 0$ by (i), and, from the definitions,

$$\measuredangle(z - q, M) = \measuredangle(z - q, P_M(z - q)). \tag{4.13}$$

If $z - q \in M$, then (4.10) holds trivially. We can therefore assume that $z - q \notin M$. Write p for $P_M(z-q)$, and let $V = \operatorname{span}(z-q, p)$. Since $z-q \perp H$ and $(z-q-p) \perp M$, these two vectors are perpendicular to $H \cap M$, and hence

$$V \perp (H \cap M); \tag{4.14}$$

in particular,

$$p \perp (H \cap M). \tag{4.15}$$

Since $(z - q) \notin H$, $V \cap H$ is a one dimensional space, there exists $v \neq 0$ such that $V \cap H = \{\lambda v | \lambda \in R\}$. Then $\langle z - q, v\rangle = 0$ since $v \in H$. By (4.14) we have

$$v \perp (H \cap M). \tag{4.16}$$

Combining (4.15) and (4.16), it follows from the definition of the angle between two subspaces that

$$\measuredangle(H, M) \leqslant \measuredangle(v, p). \tag{4.17}$$

Moreover,

$$\measuredangle(z - q, p) + \measuredangle(p, v) = \measuredangle(z - q, v) = \frac{\pi}{2}, \tag{4.18}$$

since the three vectors involved lie in the same plane (the two dimensional subspace V). From (4.13), (4.17), and (4.18), we have

$$\measuredangle(z - q, M) + \measuredangle(H, M) \leqslant \frac{\pi}{2},$$

showing that (4.10) is satisfied by H.

Case 2: $\phi(z) < 0$. By Lemma 4.8, one has $q = P_{-C}(Ax - b) = 0$. We can assume that $b \notin M$ (otherwise, the results of the theorem have already been shown). Then $\phi(P_M b - b) > 0$. In fact,

$$\langle P_M b - b, Au - b\rangle = \|P_M b - b\|^2 > 0$$

for any $u \in X$ because $(P_M b - Au) \in M \perp (P_M b - b)$. Picking $\bar{u} \in S_C^b$ and letting $\bar{z} = A\bar{u} - b$, it follows that $\bar{z} \in -C$ and $\langle P_M b - b, \bar{z}\rangle > 0$; hence $\phi(P_M b - b) > 0$

by definition of ϕ. Since ϕ takes values of opposite signs at the end points of the line-segment $[z, P_M b - b] \subset M - b$, there exists a w in that line-segment such that $\phi(w) = 0$; write

$$w := tz + (1-t)(P_M b - b) \in M - b, \tag{4.19}$$

with some $t \in (0,1)$. Note that $z, w \notin M$ because $b \notin M$. Since $\phi(w) = 0$, one has that $w \notin -C$; otherwise, the definition of ϕ would imply that $\phi(w) \geqslant \langle w, \frac{w}{\|w\|}\rangle = \|w\| > 0$. According to (4.4),

$$\cos \measuredangle(w, M) = \frac{\|P_M(w)\|}{\|w\|}.$$

Since $(w - P_M(w)) \perp P_M(w)$, it follows that

$$\cot \measuredangle(w, M) = \frac{\|P_M(w)\|}{\|w - P_M(w)\|}. \tag{4.20}$$

Similarly,

$$\cot \measuredangle(z, M) = \frac{\|P_M(z)\|}{\|z - P_M(z)\|}. \tag{4.21}$$

Let $y \in M - b$: $y = Au - b$ for some $u \in X$. Note that

$$y - P_M y = Au - b - P_M(Au - b) = Au - b - [Au - P_M(b)] = P_M b - b.$$

Considering z and w separately for the above y, we arrive at

$$z - P_M z = w - P_M w = P_M b - b. \tag{4.22}$$

From (4.19) one has $\|P_M w\| = \|t P_M z\|$. Together with (4.20), (4.21), and (4.22), we have $\cot \measuredangle(z, M) > \cot \measuredangle(w, M)$, that is,

$$\measuredangle(z, M) < \measuredangle(w, M). \tag{4.23}$$

By Case 1 (applied to w in place of z) there exists a nonvertex supporting hyperplane H of C such that

$$\measuredangle(w, M) + \measuredangle(H, M) \leqslant \frac{\pi}{2}. \tag{4.24}$$

Combining (4.23) and (4.24), (4.10) is seen to hold. □

THEOREM 4.10. *Let $\mathcal{T}$ consist of all nonvertex supporting hyperplanes of C. If*

$$\inf_{H \in \mathcal{T}} \measuredangle(H, \mathrm{Im}(A)) > 0, \tag{4.25}$$

then system (A, C, b) has an error bound, provided that $S_C^b \neq \emptyset$.

Proof. Let α denote the infimum in (4.25). Then (4.10) implies that

$$\measuredangle(Ax - b - P_{-C}(Ax - b), M) \leqslant \frac{\pi}{2} - \alpha < \frac{\pi}{2} \qquad \forall x \in X \backslash S_C^b.$$

Thus the result follows immediately from Theorem 4.3. □

The following example provides an application of Theorem 4.10.

Example 4.10.1. Let $X = Y = \mathbb{R}^2$, $C = \{(x, y) \in \mathbb{R}^2 | x \geqslant 0 \text{ and } y \geqslant 0\}$. Let $A : X \to Y$ be defined by $A(x, y) = (-x, x)$. Then there are only two non–vertex supporting hyperplane (lines) of C: they are respectively defined by $x = 0$ and $y = 0$. Since the angles between $\mathrm{Im}(A)$ and these two lines are equal to $\frac{\pi}{4}$, Theorem 4.10 implies that the system (A, C, b) has an error bound for each $b \in Y$ satisfying $S_C^b \neq \emptyset$.

The following example shows that the converse of Theorem 4.10 is not true.

Example 4.10.2. Let $X = Y = \mathbb{R}^2$, $C = \{(x, y) \in \mathbb{R}^2 | x \geqslant 0 \text{ and } y \geqslant 0\}$. Let $A : X \to Y$ be defined by $A(x, y) = (x, 0)$. Let H be the line defined by $y = 0$. Then $\measuredangle(\mathrm{Im}(A), H) = 0$, and thus (4.25) is not satisfied. But by Theorem 3.4, the system (A, C, b) has an error bound for any $b \in \mathbb{R}^2$ with $S_C^b \neq \emptyset$.

The following lemma gives a simple method for computing $\measuredangle(H, \mathrm{Im}(A))$.

LEMMA 4.11. *Let H be a hyperplane of Y and $l \perp H$. Let M be a proper subspace of Y. Then*

$$\measuredangle(H, M) + \measuredangle(l, M) = \frac{\pi}{2}. \tag{4.26}$$

Proof. It is easy to see that (4.26) is equivalent to

$$\cos^2 \measuredangle(H, M) + \cos^2 \measuredangle(l, M) = 1. \tag{4.27}$$

We assume that $\|l\| = 1$. If $l \in M$ or $M \subset H$, (4.26) is obvious. We can therefore assume that $l \notin M$ and $M \not\subset H$. Let $N = (M \cap H)^\perp \cap M$. From $M \not\subset H$ we know that $N \neq \{0\}$. Since $N \cap H = (M \cap H)^\perp \cap (M \cap H) = \{0\}$ and H is a hyperplane, it follows that $\dim(N) = 1$: there exists a unit vector n_0 which linearly spans N. Because the unit sphere of a Hilbert space is weak compact, there exists $h_0 \in (M \cap H)^\perp \cap H$, $\|h_0\| = 1$, such that

$$\measuredangle(H, M) = \measuredangle(h_0, n_0). \tag{4.28}$$

We can express n_0 in the form $n_0 = \alpha l + \beta h$, where $h \in H$, $\|h\| = 1$, and $\alpha^2 + \beta^2 = 1$, $\alpha\beta \neq 0$. Since $l \perp (M \cap H)$, we have $h \perp (M \cap H)$. By definition and from (4.28), we have $|\langle n_0, h_0\rangle| \geqslant |\langle n_0, h\rangle|$, which implies that $|\beta\langle h, h_0\rangle| \geqslant |\beta\langle h, h\rangle|$. Replacing h_0 by $-h_0$, if necessary, it follows that $\beta h = \beta h_0$, and thus $n_0 = \alpha l + \beta h_0$. Thus, from (4.28),

$$\cos \measuredangle(H, M) = \cos \measuredangle(h_0, n_0) = |\beta|. \tag{4.29}$$

Note that $m \in M$ can be written in the form $m = sn_0 + tk$, where $k \in M \cap H$ and $\|k\| = 1$; hence $|\langle l, m\rangle| = |s\langle l, n_0\rangle| \leqslant \|m\| \cdot |\langle l, n_0\rangle|$. It follows from (4.3) that

$$\cos \measuredangle(l, M) = |\langle l, n_0\rangle| = |\alpha|. \tag{4.30}$$

Thus (4.27) and hence (4.26) hold. □

THEOREM 4.12. *Let C^+ denote the set of all x satisfying $\langle x, c\rangle \geqslant 0$ for each $c \in C$. If*

$$\sup_{\substack{l \in \mathrm{bd}(C^+) \\ l \neq 0}} \measuredangle(l, \mathrm{Im}(A)) < \frac{\pi}{2}, \tag{4.31}$$

then system (A, C, b) has an error bound, provided that $S_C^b \neq \emptyset$.

Proof. Let β denote the supremum in (4.31). Let $H \in \mathcal{T}$: H is a hyperplane in Y supporting C at some $\bar{c} \in C \backslash \{0\}$. Take $l \in H^\perp$ with $\|l\| = 1$. By Lemma 4.6(a)

and replacing l by $-l$ if necessary, we can suppose that $l \in C^+$ and $l(\bar{c}) = 0$. Then l must not be in the interior of C^+. Hence it follows from Lemma 4.11 that

$$\measuredangle(H, \mathrm{Im}(A)) = \frac{\pi}{2} - \measuredangle(l, \mathrm{Im}(A)) \geqslant \frac{\pi}{2} - \beta > 0,$$

showing that (4.25) holds. Therefore by Theorem 4.10, the system (A, C, b) has an error bound, provided that $S_C^b \neq \emptyset$. □

Next we give an example which shows the application of Theorem 4.12.

Example 4.12.1. Let $X = \mathbb{R}^3$, $C = \{(x, y, z) : \ z \geqslant \sqrt{x^2 + y^2}\}$. Let $A(x, y, z) = (x, y, 0)$. The boundary of C^+ is $\{(x, y, z) : \ z = \sqrt{x^2 + y^2}\}$. For $l \in \mathrm{bd}(C^+)$, $l \neq 0$, $\measuredangle(l, \mathrm{Im}(A)) = \measuredangle(l, P_{\mathrm{Im}(A)} l)$. We have

$$\sup_{\substack{l \in \mathrm{bd}(C^+) \\ l \neq 0}} \measuredangle(l, \mathrm{Im}(A)) = \frac{\pi}{4} < \frac{\pi}{2}.$$

By Theorem 4.12 we know that system (A, C, b) has an error bound for each b such that $S_C^b \neq \emptyset$.

5. The ice-cream cone. We continue to use the notations defined in the preceding section; in particular, Y is a Hilbert space: infinite dimensional or finite dimensional. We assume throughout that $M := \mathrm{Im}(A)$ is closed. Let V be a closed convex set in Y and $x \in \mathrm{bd(V)}$. $N_V(x)$ equals the normal cone of V at x in the following sense:

$$v \in N_V(x) \quad \text{if and only if} \quad \langle v, t\rangle \leqslant 0 \quad \text{for} \quad t \in T_V(x).$$

LEMMA 5.1. *Let V be a closed convex set in a Hilbert space Y and $x \in \mathrm{bd(V)}$. Let $d_V(y)$ be defined by $d_V(y) = \mathrm{dist}(y, V)$. Then for each $h \in Y$, the maxima below are attained and*

$$\max\{|\langle \xi, h\rangle| : \xi \in N_V^1(x)\} = \max\{|\langle \eta, h\rangle| : \eta \in \partial d_V(x)\}.$$

If $\max\{\langle \eta, h\rangle : \eta \in \partial d_V(x)\} \geqslant 0$ *for some h, then*

$$\max\{\langle \xi, h\rangle : \xi \in N_V^1(x)\} = \max\{\langle \eta, h\rangle : \eta \in \partial d_V(x)\}.$$

Proof. We assume that $h \neq 0$. According to the result in [6, p. 259],

$$\partial d_V(x) = N_V(x) \cap B(0, 1), \tag{5.1}$$

where $B(0, 1)$ denotes the closed unit ball in Y. In particular, $\partial d_V(x)$ is weak compact and contains $N_V^1(x)$; hence there exists an $\bar{\eta} \in \partial d_V(x)$ such that

$$\sup\{|\langle \xi, h\rangle| : \xi \in N_V^1(x)\} \leqslant \max\{|\langle \eta, h\rangle| : \eta \in \partial d_V(x)\} = \langle \bar{\eta}, h\rangle.$$

From (5.1), it follows that $\bar{\eta}$ is of norm 1, that is, $\bar{\eta} \in N_V^1(x)$. This proves the first assertion of the theorem. The proof for the second assertion is similar. □

Let $b \in Y$, and assume as before that $S_C^b \neq \emptyset$. For simplicity, let Ω denote $(b - C) \cap M$, where $M := \mathrm{Im}(A)$. Note that $\Omega \neq \emptyset$. Let I_A denote the identity operator from M into Y. Then $\Omega = \{y \in M : I_A y - b \leqslant 0\}$. Let $\mathrm{bd}_M(\Omega)$ denote the boundary of Ω relative to M, and $\mathcal{N}_\Omega^1(y)$ denote $N_\Omega^1(y) \cap M$. That is,

$$\mathcal{N}_\Omega^1(y) = \{h : h \in \mathrm{Im}(A) \text{ and } \mathrm{dist}(x + th, \Omega) = t\|h\| \ \forall t \in [0, 1]\}.$$

We have the following theorem.

THEOREM 5.2. *The following statements are equivalent:*

(i) *The system* (A, C, b) *has an error bound.*

(ii) $\inf_{x\in \mathrm{bd}(S^b_C) u\in N^1_{S^b_C}(x)} \max\{\langle \xi, Au\rangle : \xi \in N^1_{-C}(Ax-b)\} > 0.$

(iii) $\inf_{y\in \mathrm{bd}_M(\Omega) h\in \mathcal{N}^1_\Omega(y)} \max\{\langle \xi, h\rangle : \xi \in N^1_{-C}(y-b)\} > 0.$

Proof. Let $\varphi_b(x)$ and g be defined by (3.1) and (3.2). By the chain rules and [3, Proposition 2.1.2], we have

$$\varphi'_b(x;u) = g'(Ax-b;Au) = \max\{\langle \xi, Au\rangle : \xi \in \partial g(Ax-b)\} \quad \forall x, u \in X.$$

If $x \in \mathrm{bd}(S^b_C)$, then $Ax - b \in \mathrm{bd}(-C)$, and it follows from Lemma 5.1 that

$$\varphi'_b(x;u) = \max\{\langle \xi, Au\rangle : \xi \in N^1_{-C}(Ax-b)\},$$

which together with Theorem 2.2 and Remark 2.2.1 implies (i) $\Leftrightarrow$ (ii). Similarly, (iii) holds if and only if the system (I_A, C, b) has an error bound. Thus the result now follows from Corollary 3.3. □

COROLLARY 5.3. *Suppose that there exist* $\bar{y} \in \mathrm{bd}_M(\Omega)$ *and a hyperplane* H *satisfying the properties:*

(a) $N^1_{-C}(\bar{y}-b)$ *is a singleton.*

(b) $M \subset H$, *and* H *supports* $-C$ *at* $\bar{y}-b$.

Then the system (A, C, b) *has no error bound.*

Proof. Write $\bar{c}$ for $b - \bar{y}$. Then $\bar{c} \in \mathrm{bd}(C)$ and, by assumption, take a hyperplane $H \supset M$ such that it supports $-C$ at $-\bar{c}$. By Lemmas 4.6 and 4.7, there exists an $\bar{\xi}$ of norm 1 such that $\bar{\xi}(\bar{c}) = 0$ and $\bar{\xi} \in H^\perp \cap N^1_{-C}(-\bar{c})$. Hence, since $N^1_{-C}(-\bar{c})$ is a singleton by assumption,

$$\max\left\{\langle \xi, h\rangle : \xi \in N^1_{-C}(-\bar{c})\right\} = \langle \bar{\xi}, h\rangle \quad \forall h \in Y. \tag{5.2}$$

Note that, since $\bar{\xi} \in H^\perp \subset M^\perp$, $\langle \bar{\xi}, l\rangle = 0$ for each $l \in M$. Pick $\bar{h} \in \mathcal{N}^1_\Omega(\bar{y})$. Since $\mathcal{N}^1_\Omega(\bar{y}) \subset M$, it follows that $\langle \bar{\xi}, \bar{h}\rangle = 0$. Combining this with (5.2), one has

$$\max\left\{\langle \xi, \bar{h}\rangle : \xi \in N^1_{-C}(\bar{y}-b)\right\} = 0.$$

Since $\bar{y} \in \mathrm{bd}_M(\Omega)$ and $\bar{h} \in \mathcal{N}^1_\Omega(\bar{y})$, this implies that Theorem 5.2(iii) does not hold, and therefore we conclude from Theorem 5.2 that the system (A, C, b) has no error bound. □

PROPOSITION 5.4. *Suppose that* Y *is finite dimensional, and that* $\mathrm{Im}(A) \cap C = \{0\}$. *Then the system* $(A, C, 0)$ *has an error bound.*

Proof. Suppose that $(A, C, 0)$ has no error bound. Then according to Theorem 4.2, there exists a sequence $\{x_k\} \subset X \backslash S^0_C$ such that

$$\lim_{k\to\infty} \measuredangle(Ax_k - P_{-C}(Ax_k), \mathrm{Im}(A)) = \frac{\pi}{2}. \tag{5.3}$$

Without loss of generality we can assume that $\|Ax_k\| = 1$ for each k. Since Y is finite dimensional, $\{Ax_k\}$ has a cluster point, which will be denoted by z. It is clear that $\|z\| = 1$ and that $z = A\bar{x}$ for some $\bar{x} \in X$ because $\mathrm{Im}(A)$ is closed. From $\mathrm{Im}(A) \cap -C = \{0\}$ we know that $z \notin -C$ and hence that $\bar{x} \notin S^0_C$. But (5.3) implies that

$$\measuredangle(A\bar{x} - P_{-C}(A\bar{x}), \mathrm{Im}(A)) = \frac{\pi}{2},$$

contradicting Theorem 4.9(i). □

LEMMA 5.5. *Suppose that Y is finite dimensional. Let $b \in Y$. If $(\mathrm{Im}(A) - b) \cap (-C)$ is bounded and $(\mathrm{Im}(A) - b) \cap \mathrm{int}(-C) \neq \emptyset$, then the system (A, C, b) has an error bound.*

Proof. Write M for $\mathrm{Im}(A)$. Denote $(b - C) \cap M$ by Ω. By [16, Theorem 2.5] there exists a constant $\kappa > 0$ such that for any $y \in M$,

$$\begin{aligned} \mathrm{dist}(y, \Omega) = \mathrm{dist}(y - b, \Omega - b) &\leqslant \kappa\big[\mathrm{dist}(y - b, -C) + \mathrm{dist}(y - b, M - b)\big] \\ &= \kappa\big[\mathrm{dist}(y - b, -C)\big]. \end{aligned} \tag{5.4}$$

Note that this just means that the system (I_A, C, b) has an error bound, and therefore it follows from Corollary 3.3 that the system (A, C, b) has an error bound. □

We conclude our paper with a discussion of the special case in which C is an "ice-cream" cone. Here the so-called ice-cream cone is defined by $S_{ice} = f^{-1}(-\infty, 0]$, where f is defined by

$$f(x) = \sqrt{\sum_{i=1}^{n-1} x_i^2} - x_n, \qquad x \in \mathbb{R}^n. \tag{5.5}$$

Any set that is not a singleton will be referred to as a ray if it can be expressed in the form $\{y_0 + th : t \geqslant 0\}$. A ray that is an extreme subset of S_{ice} is called an extreme ray. The following lemma is stated for our easy reference; it is elementary and we need not give a proof here.

LEMMA 5.6. *Let S_{ice} be the ice-cream cone in $\mathbb{R}^n$, and let f be defined by* (5.5). *Then*

(i) *$S_{ice} = S_{ice}^+$, that is, ξ belongs to S_{ice} if and only if $\xi \in \mathbb{R}^n$ and*

$$\langle \xi, y \rangle \geqslant 0 \quad \forall y \in S_{ice}.$$

(ii) *The union of the extreme rays of S_{ice} equals the topological boundary $\mathrm{bd}(S_{ice})$.*

(iii) *If $s' \in \mathrm{bd}(S_{ice}) \backslash \{0\}$, $s \in S_{ice}$, and s is not a multiple of s', then the open line-segment (s, s') is contained in $\mathrm{int}(S_{ice})$, the topological interior of S_{ice}.*

(iv) *If $x \in \mathrm{bd}(S_{ice})$ and $x \neq 0$, then $\partial f(x) = \{\bigtriangledown f(x)\}$, where*

$$\bigtriangledown f(x) = \left(\frac{x_1}{\sqrt{\sum_{i=1}^{n-1} x_i^2}}, \ldots, -1 \right)$$

and $N^1_{S_{ice}}(x) = \left\{ \frac{\bigtriangledown f(x)}{\|\bigtriangledown f(x)\|} \right\}$. Moreover, $\langle \bigtriangledown f(x), x - x' \rangle$ is strictly positive whenever $x' \in \mathrm{int}(S_{ice})$.

(v) *If $x \in -S_{ice}$, then $\mathrm{dist}(x, S_{ice}) = \|x\|$.*

In what follows we will identify exactly when (A, S_{ice}, b) has an error bound (and when it has none), where A is a linear operator from X into Y, and $Y = \mathbb{R}^n$ is equipped with the partial order defined by S_{ice}. As in the preceding section, let $M := \mathrm{Im}(A)$. Also we use $\mathcal{K}$ to denote the cone S_{ice}. It will be shown that there are three possible cases.

Case I. M contains an interior point of $\mathcal{K}$;

Case II. $M \cap \mathcal{K} = \{0\}$;

Case III. $M \cap \mathcal{K}$ is an extreme ray of $\mathcal{K}$.

These cases are respectively dealt with in the next three theorems.

THEOREM 5.7. *Suppose that*

$$\mathrm{Im}(A) \cap \mathrm{int}(\mathcal{K}) \neq \emptyset. \tag{5.6}$$

Then the system $(A, \mathcal{K}, b)$ *has an error bound for each* $b \in \mathbb{R}^n$.

Proof. Let $b \in \mathbb{R}^n$ and $\Omega := M \cap (b - \mathcal{K})$. We proceed as in the proof of Lemma 5.5. By virtue of (5.6), one can apply [16, Theorem 2.6] (instead of [16, Theorem 2.5]) to obtain (5.4), completing the proof. □

THEOREM 5.8. *Suppose that*

$$M \cap \mathcal{K} = \{0\}. \tag{5.7}$$

Let $b \in \mathbb{R}^n$ *be such that*

$$(M - b) \cap (-\mathcal{K}) \neq \emptyset. \tag{5.8}$$

Then the following assertions hold:

(a) *If* $(M - b) \cap \mathrm{int}(-\mathcal{K}) \neq \emptyset$, *then* $(A, \mathcal{K}, b)$ *has an error bound.*

(b) *If* $b \in M$, *then* $(A, \mathcal{K}, b)$ *has an error bound.*

(c) *If* $(M - b) \cap \mathrm{int}(-\mathcal{K}) = \emptyset$ *and* $b \notin M$, *then* $(A, \mathcal{K}, b)$ *has no error bound.*

Proof. By (5.7) and [15, Corollary 8.4.1], $(M - b) \cap (-\mathcal{K})$ is bounded. Thus (a) holds by Lemma 5.2.

(b) By (5.7) and Proposition 5.4, $(A, \mathcal{K}, 0)$ has an error bound: there exists $\tau > 0$ such that

$$\mathrm{dist}(x, S^0_{\mathcal{K}}) \leqslant \tau \mathrm{dist}(Ax, -\mathcal{K}) \quad \forall x \in X, \tag{5.9}$$

where $S^0_{\mathcal{K}} := \{x \in X : Ax \in -\mathcal{K}\}$. Since $b \in M$, take $\bar{x} \in X$ such that $A\bar{x} = b$. Then $S^b_{\mathcal{K}} = \bar{x} + S^0_{\mathcal{K}}$. Writing z for $x + \bar{x}$, it follows from (5.9) that

$$\mathrm{dist}(z, S^b_{\mathcal{K}}) \leqslant \tau \mathrm{dist}(Az - b, -\mathcal{K}) \quad \forall z \in X,$$

proving (b).

To prove (c), we take $\bar{s} \in (M - b) \cap (-\mathcal{K})$ by (5.8). By assumption in (c), it follows that $\bar{s} \in \mathrm{bd}(-\mathcal{K})$. By Lemma 5.6(iv), $N^1_{-\mathcal{K}}(\bar{s})$ is certainly a singleton (consisting of $\frac{\nabla f(\bar{s})}{\|\nabla f(\bar{s})\|}$).

Moreover, by the separation theorem, take $\xi \in \mathbb{R}^n$ of norm 1 such that

$$\sup_{y \in M - b} \langle \xi, y \rangle = \langle \xi, \bar{s} \rangle \leqslant \inf_{y \in -\mathcal{K}} \langle \xi, y \rangle.$$

Since $-\mathcal{K}$ is a cone, it follows that $\langle \xi, \bar{s} \rangle = 0$, and hence that $\xi^\perp$ supports $-\mathcal{K}$ at $\bar{s}$. Note also that ξ is bounded on M and hence vanishes on M, that is, $\xi^\perp \supset M$.

We claim further that $(M - b) \cap (-\mathcal{K})$ is the singleton $\{\bar{s}\}$. Indeed, let $s \in (M - b) \cap (-\mathcal{K})$. Then the segment $[s, \bar{s}]$ is also contained in $(M - b) \cap (-\mathcal{K})$; in particular, $0 \notin [s, \bar{s}]$ as $b \notin M$. By assumption in (c), it follows that $[s, \bar{s}] \subset \mathrm{bd}(-\mathcal{K})$. By Lemma 5.6(iii), it follows that $\bar{s} = ts$ for some $t > 0$. If $t > 1$, then $\bar{s} - s \in M \cap (-\mathcal{K})$, contradicting (5.7). Similarly $t < 1$ is also not possible. Therefore $\bar{s} = s$, and our claim stands. Let $\Omega = M \cap (b - \mathcal{K})$; then $\Omega = \{\bar{s} + b\}$. By Corollary 5.3, $(A, \mathcal{K}, b)$ has no error bound. □

THEOREM 5.9. *Suppose that*

$$\{0\} \neq M \cap \mathcal{K} \subset \mathrm{bd}(\mathcal{K}). \tag{5.10}$$

Let $b \in \mathbb{R}^n$ be such that

$$(M-b)\cap(-\mathcal{K}) \neq \emptyset. \tag{5.11}$$

Then the following assertions hold:

(a) *If* $\dim(M) \neq 1$, *then* $(A, \mathcal{K}, b)$ *has no error bound.*

(b) *If* $\dim(M) = 1$, *then* $(A, \mathcal{K}, b)$ *has an error bound.*

Proof. Pick $\bar{s} \in M \cap \mathcal{K}$, $\bar{s} \neq 0$. By Lemma 5.6(iii) any element s of $M \cap \mathcal{K}$ must be a multiple of $\bar{s}$ (because by (5.10) we know that $(\bar{s}, s)$ does not contain any interior point of $\mathcal{K}$); therefore $M \cap \mathcal{K}$ is an extreme ray of $\mathcal{K}$. Separating M and $\mathrm{int}(\mathcal{K})$, one obtains a hyperplane H containing M and supporting $\mathcal{K}$ at $\bar{s}$.

(a) Suppose that $\dim(M) \neq 1$. Then the ray $M \cap \mathcal{K}$ cannot contain any relative interior point in M. Let $\Omega = M \cap (-\mathcal{K})$. Then $-\bar{s} \in \mathrm{bd}_M(\Omega)$, and $N^1_{-\mathcal{K}}(-\bar{s})$ is a singleton by Lemma 5.6(iv). Hence, by Corollary 5.3, $(A, \mathcal{K}, 0)$ has no error bound. In view of Corollary 4.5, this implies that $(A, \mathcal{K}, b)$ has no error bound for any $b \in \mathbb{R}^n$.

(b) Suppose that $\dim(M) = 1$. We consider the system $(I_A, \mathcal{K}, b)$, where b satisfies (5.11). In light of Corollary 3.3, it is sufficient to show that $(I_A, \mathcal{K}, b)$ has an error bound, that is, to show that there exists $\tau > 0$ such that

$$\mathrm{dist}(m, S) \leqslant \tau \mathrm{dist}(m - b, -\mathcal{K}) \quad \forall m \in M, \tag{5.12}$$

where $S := \{m \in M : m - b \in -\mathcal{K}\} = M \cap (b - \mathcal{K})$. Note from Lemma 5.6(v) that if L is a line containing an extreme ray R of $\mathcal{K}$, then, for each $l \in L \backslash R$,

$$\mathrm{dist}(l, \mathcal{K}) = \mathrm{dist}(l, L \cap \mathcal{K}) (= \|l\|).$$

Therefore, if $b \in M$,

$$\begin{aligned} \mathrm{dist}(m, M \cap (b - \mathcal{K})) &= \mathrm{dist}(m - b, (M - b) \cap (-\mathcal{K})) \\ &= \mathrm{dist}(m - b, -\mathcal{K}) \quad \forall m \in M. \end{aligned}$$

Therefore (5.12) holds with $\tau = 1$, and hence we may suppose henceforth that $b \notin M$.

We recall from the first paragraph of the proof that there exists $\bar{s} \in \mathcal{K}$, $\|\bar{s}\| = 1$, such that

$$M \cap (-\mathcal{K}) = \{-t\bar{s} : t \geqslant 0\}$$

is an extreme ray of $-\mathcal{K}$. By (5.11), take $s_0 \in (m - b) \cap (-\mathcal{K})$. Then it is easily shown that $s_0 + t(-\bar{s})$ also belongs to $(m - b) \cap (-\mathcal{K})$ for each $t \geqslant 0$. It follows that $(m-b)\cap(-\mathcal{K})$ is a ray; we suppose without loss of generality that its end point is s_0:

$$(m - b) \cap (-\mathcal{K}) = \{s_0 + t(-\bar{s}) : t \geqslant 0\}. \tag{5.13}$$

Since $s_0 \in M - b$ and $M - b$ is disjoint from M, s_0 is not a multiple of $\bar{s}$. Hence, for each $t > 0$, $s_0 + t(-\bar{s})$ is not multiple of $\bar{s}$. By Lemma 5.6(iii), $s_0 + t(-\bar{s}) \in \mathrm{int}(-\mathcal{K})$, and hence

$$\langle N_{-K}(s_0), -\bar{s}\rangle < 0. \tag{5.14}$$

This implies that (5.12) holds for some constant $\tau > 0$. Indeed, if not, then for each $k \in \mathbb{N}$ there exists an $m_k \in M$ such that

$$\mathrm{dist}(m_k - b, (m - b) \cap (-\mathcal{K})) > k \cdot \mathrm{dist}(m_k - b, -\mathcal{K}).$$

Writing $s_0 + t_k\bar{s}$ (with $t_k > 0$) for $m_k - b$, it follows that

$$t_k > k \cdot \mathrm{dist}(s_0 + t_k\bar{s}, -\mathcal{K}) \quad \forall k.$$

Since $-\mathcal{K}$ is convex, it follows that 0 must be a cluster point of $\{t_k\}$. This implies that $\bar{s} \in T_{-\mathcal{K}}(s_0)$ and $\langle N_{-K}(s_0), \bar{s}\rangle \leqslant 0$, contradicting (5.14). ∎

Acknowledgments. We are indebted to Professor J.-S. Pang (who suggested that we look at the ice-cream cone case) and to Professor X. Y. Zheng for many stimulating discussions.

REFERENCES

[1] J. M. Borwein, S. P. Fitzpatrick, and J. R. Giles, *The differentiability of real functions on normed linear space using generalized subgradients*, J. Math. Anal. Appl., 128 (1987), pp. 512–534.

[2] J. V. Burke and P. Tseng, *A unified analysis of Hoffman's bound via Fenchel duality*, SIAM J. Optim., 6 (1996), pp. 265–282.

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

[4] F. Deutsch, *The angle between subspaces of a Hilbert space*, in Approximation Theory, Wavelets and Applications, S. P. Singh, ed., Kluwer Academic Publishers, Norwell, MA, 1995, pp. 107–130.

[5] Y. R. He, *Error Bound for Nonlinear Complementarity Problems*, preprint, The Chinese University of Hong Kong, Shatin, Hong Kong, 2000.

[6] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms* I, Springer-Verlag, New York, 1993.

[7] A. J. Hoffman, *On approximate solutions of systems of linear inequalities*, J. Res. National Bureau of Standards, 49 (1952), pp. 263–265.

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

[9] A. Lewis and J.-S. Pang, *Error bounds for convex inequality systems*, in Generalized Convexity, Generalized Monotonicity. Proceedings of the Fifth Symposium on Generalized Convexity, Luminy, 1996, J.-P. Crouzeix, J.-E. Martinez-Legaz, and M. Volle, eds., Nonconvex Optim. Appl. 27, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1997, pp. 75–100.

[10] K. F. Ng and X. Y. Zheng, *Error bounds for lower semicontinuous functions in normed spaces*, SIAM J. Optim., 12 (2001), pp. 1–17.

[11] K. F. Ng and X. Y. Zheng, *Global Weak Sharp Minima on Banach Spaces*, preprint, The Chinese University of Hong Kong, Shatin, Hong Kong, 2001.

[12] J.-S. Pang, *Error bounds in mathematical programming*, Math. Programming, 79 (1997), pp. 299–332.

[13] R. R. Phelps, *Convex Functions, Monotone Operators and Differentiability*, Springer-Verlag, Berlin, Heidelberg, 1989.

[14] S. M. Robinson, *Stability theory for systems of inequalities. Part* I: *Linear systems*, SIAM J. Numer. Anal., 12 (1975), pp. 754–769.

[15] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[16] S. Zhang, *Global error bounds for convex conic problems*, SIAM J. Optim., 10 (2000), pp. 836–851.

SIAM J. OPTIM.
Vol. 13, No. 1, pp. 44–59

ON THE GLOBAL CONVERGENCE OF A FILTER–SQP ALGORITHM*

ROGER FLETCHER†, SVEN LEYFFER†, AND PHILIPPE L. TOINT‡

Abstract. A mechanism for proving global convergence in SQP–filter methods for nonlinear programming (NLP) is described. Such methods are characterized by their use of the dominance concept of multiobjective optimization, instead of a penalty parameter whose adjustment can be problematic. The main point of interest is to demonstrate how convergence for NLP can be induced without forcing sufficient descent in a penalty-type merit function.

The proof relates to a prototypical algorithm, within which is allowed a range of specific algorithm choices associated with the Hessian matrix representation, updating the trust region radius, and feasibility restoration.

Key words. nonlinear programming, global convergence, filter, multiobjective optimization, SQP

AMS subject classifications. 65K05, 49M37, 90C30, 90C26

PII. S105262340038081X

1. Introduction.

In Fletcher and Leyffer [5] a new technique for globalizing methods for nonlinear programming (NLP) is presented. The idea is referred to as an NLP filter and is motivated by the aim of avoiding the need to choose penalty parameters, such as would occur with the use of l_1 penalty functions or augmented Lagrangian functions. Numerical experience with the technique in a sequential quadratic programming (SQP) trust region algorithm is reported in [5] and is very promising. However, no global convergence proof is given in [5], although a number of heuristics are suggested to eliminate obvious situations in which the method might fail to converge.

This paper shows that the filter technique does provide a mechanism for forcing global convergence when used in an appropriate way. The proof relates to an NLP problem with both equations and inequality constraints and shows that there exists an accumulation point that satisfies first order (Kuhn–Tucker, or KT) conditions. The result requires that a Mangasarian–Fromowitz constraint qualification hold at the accumulation point. Other nontrivial assumptions that are made are that the Hessian matrices of the quadratic programming (QP) subproblems are uniformly bounded and that a global solution of the subproblem is found by the QP solver. None of these qualifications to the result are readily circumvented.

The proposed algorithm contains an inner iteration for calculating a suitable trust region radius. In some ways this resembles the use of a backtracking line search along a piecewise linear trajectory. This approach enables us to guarantee that certain conditions used in the convergence proof are met. To a large extent, however, the approach allows the use of conventional ideas of halving or doubling (say) the previous trust region radius.

*Received by the editors November 13, 2000; accepted for publication (in revised form) September 20, 2001; published electronically May 15, 2002.
http://www.siam.org/journals/siopt/13-1/38081.html
†Department of Mathematics, University of Dundee, Dundee DD1 4HN, Scotland, UK (fletcher@maths.dundee.ac.uk, sleyffer@maths.dundee.ac.uk).
‡Department of Mathematics, University of Namur, 61 rue de Bruxelles, B-5000 Namur, Belgium (philippe.toint@fundp.ac.be).

An interesting feature of the proof is that various of the heuristics used in [5] are shown to be unnecessary. These include the NW corner rule, the need to unblock the filter in some cases, and the consequent decision to reduce the strict upper bound on constraint infeasibility. In this paper we also use a way of defining the sufficient reduction condition slightly different from that used in [5]. Another new feature of some interest is that some points may be accepted by the algorithm, without a new entry in the filter being made. This contributes to the nonmonotonic properties of the algorithm. In common with [5], we do use a feasibility restoration technique but are not prescriptive as to how this is done.

Subsequent to the work described in this paper, there have been a number of more recent developments in regard to global convergence of filter-related methods for NLP. The authors have contributed to other papers that prove global convergence for different algorithmic structures such as an SLP-EQP approach or an approach in which approximate solutions of the SQP step are used, based on a decomposition into normal and tangential steps. Recent work of other authors proves the global convergence of filter-related methods in a variety of other contexts such as interior point and line search barrier methods. A brief discussion of these developments is given in section 4.

2. A filter–SQP algorithm. In this paper we consider an NLP problem of the form

$$P \left\{ \begin{array}{lll} \underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}} & f(\mathbf{x}) & \\ \text{subject to} & c_i(\mathbf{x}) = 0, & i \in \mathcal{E}, \\ & c_i(\mathbf{x}) \le 0, & i \in \mathcal{I}, \end{array} \right.$$

where the index sets $\mathcal{E}$ and $\mathcal{I}$ reference the equality and inequality constraints, respectively. We denote the cardinality of $\mathcal{E} \cup \mathcal{I}$ by m. We assume for the purposes of our convergence proof that all points that are sampled by the algorithm lie in a nonempty closed and bounded set X. Because the points generated by our algorithm satisfy the linear constraints of the problem, it is readily possible to ensure that this condition holds by including suitable simple upper and lower bounds on $\mathbf{x}$ among the constraints of P. The QP subproblem in our algorithm depends upon the value of the current iterate $\mathbf{x}$ and trust region radius ρ ($\rho > 0$) and is defined by

$$QP(\mathbf{x},\rho) \left\{ \begin{array}{lll} \underset{\mathbf{d}\in\mathbb{R}^n}{\text{minimize}} & q(\mathbf{d}) := \mathbf{g}^T\mathbf{d} + \frac{1}{2}\mathbf{d}^T B\mathbf{d} & \\ \text{subject to} & c_i + \mathbf{a}_i^T\mathbf{d} = 0, & i \in \mathcal{E}, \\ & c_i + \mathbf{a}_i^T\mathbf{d} \le 0, & i \in \mathcal{I}, \\ & \|\mathbf{d}\|_\infty \le \rho. & \end{array} \right.$$

where we define $\mathbf{g} = \operatorname{grad} f(\mathbf{x})$, $c_i = c_i(\mathbf{x})$, and $\mathbf{a}_i = \operatorname{grad} c_i(\mathbf{x})$. The l_∞ norm is used to define the trust region because it is readily implemented by adding simple bounds to the QP subproblem. The QP subproblem also requires the specification of a matrix B, although this plays a relatively minor part in the analysis of global convergence. For this reason we do not make the dependence on B explicit in the notation. We let $\mathbf{d}$ denote the global solution (if it exists) of $QP(\mathbf{x},\rho)$. Then we denote

$$\Delta q = q(\mathbf{0}) - q(\mathbf{d}) = -\mathbf{g}^T\mathbf{d} - \tfrac{1}{2}\mathbf{d}^T B\mathbf{d} \tag{2.1}$$

as the *predicted reduction* in $f(\mathbf{x})$, and

$$\Delta f = f(\mathbf{x}) - f(\mathbf{x}+\mathbf{d}) \tag{2.2}$$

as the *actual reduction* in $f(\mathbf{x})$. The measure of constraint infeasibility that we use in this paper is

$$h(\mathbf{c}) = \|\mathbf{c}_{\mathcal{I}}^{+}\|_1 + \|\mathbf{c}_{\mathcal{E}}\|_1, \tag{2.3}$$

where $c_i^+ = \max(0, c_i)$, using the notation that $\mathbf{c}_{\mathcal{E}}$ and $\mathbf{c}_{\mathcal{I}}$ are partitions of $\mathbf{c}$ corresponding to equality and inequality constraints, respectively.

The algorithm that we propose is iterative, and the index k is used throughout to refer to the iteration number. The sequence of points accepted by the algorithm is referred to by $\{\mathbf{x}^{(k)}\}$, and quantities derived from $\mathbf{x}^{(k)}$ are superscripted in a similar manner; for example, $h^{(k)}$ refers to $h(\mathbf{c}(\mathbf{x}^{(k)}))$, and $f^{(k)}$ to $f(\mathbf{x}^{(k)})$. The matrix B usually differs from iteration to iteration and is generally referred to as $B^{(k)}$. Within the inner loop of the iterative process, $B^{(k)}$ is a constant matrix.

We now turn to the definition of an NLP filter as introduced in [5]. The two aims in an NLP problem are to minimize $f(\mathbf{x})$ and to satisfy the constraints, that is, to minimize $h(\mathbf{c}(\mathbf{x}))$. In a filter we consider pairs of values (h, f) obtained by evaluating $h(\mathbf{c}(\mathbf{x}))$ and $f(\mathbf{x})$ for various values of $\mathbf{x}$. A pair $(h^{(i)}, f^{(i)})$ obtained on iteration i is said to *dominate* another pair $(h^{(j)}, f^{(j)})$ if and only if both $h^{(i)} \leq h^{(j)}$ and $f^{(i)} \leq f^{(j)}$, indicating that the point $\mathbf{x}^{(i)}$ is at least as good as $\mathbf{x}^{(j)}$ with respect to both measures. The NLP filter is defined to be a list of pairs $(h^{(i)}, f^{(i)})$ such that no pair dominates any other. This is illustrated by the solid lines in Figure 1. We use $\mathcal{F}^{(k)}$ to denote the set of iteration indices j $(j < k)$ such that $(h^{(j)}, f^{(j)})$ is an entry in the current filter. (In practice we do not need to store the index set $\mathcal{F}^{(k)}$; the notation is just for theoretical convenience.) A point $\mathbf{x}$ is said to be "acceptable for inclusion in the filter" if its (h, f) pair is not dominated by any entry in the filter. This is the condition that

$$\text{either} \qquad h < h^{(j)} \qquad \text{or} \qquad f < f^{(j)} \tag{2.4}$$

for all $j \in \mathcal{F}^{(k)}$. We may also wish to "include a point $\mathbf{x}$ in the filter," by which we mean that its (h, f) pair is added to the list of pairs in the filter, and any pairs in the filter that are dominated by the new pair are removed. We use the filter as an alternative to a penalty function as a means of deciding whether or not to accept a new point in an NLP algorithm.

In fact this definition of a filter is not adequate for proving convergence, as it allows points to accumulate in the neighborhood of a filter entry that has $h^{(i)} > 0$. This is readily corrected by defining a small envelope around the current filter, in which points are not accepted. This idea is suggested in the original paper of Fletcher and Leyffer [5]. A similar acceptability test is analyzed by Fletcher, Leyffer, and Toint [6] in proving the global convergence of an SLP–filter algorithm. This is the condition that a point is acceptable to the filter if its (h, f) pair satisfies

$$\text{either} \qquad h \leq \beta h^{(j)} \qquad \text{or} \qquad f \leq f^{(j)} - \gamma h^{(j)} \tag{2.5}$$

for all $j \in \mathcal{F}^{(k)}$, where β and γ are preset parameters such that $1 > \beta > \gamma > 0$, with β close to 1 and γ close to zero. Because $1 - \beta$ and γ are very small, there is negligible difference in practice between (2.5) and (2.4).

In fact, it has more recently become apparent that a slightly different form of the acceptability test, due to Chin and Fletcher [2], allows stronger convergence results to be proved, and it is this that we analyze here. In this test a pair (h, f) is acceptable if

$$\text{either} \qquad h \leq \beta h^{(j)} \qquad \text{or} \qquad f + \gamma h \leq f^{(j)} \tag{2.6}$$

for all $j \in \mathcal{F}^{(k)}$. This *slanting envelope test* ensures that pairs with the same f value have the same envelope in the f direction. This is illustrated in Figure 1, using the values $\gamma = 0.1$ and $\beta = 1 - \gamma$, although in practice a value of γ much closer to zero would be used. (Typical values that we have used are $\gamma = 10^{-5}$ and $\beta = 1 - \gamma$.) The test provides an important *inclusion property* that if a pair (h, f) is added to the filter, then the set of unacceptable points for the new filter always includes the set of unacceptable points for the old filter. This is not always the case for (2.5).

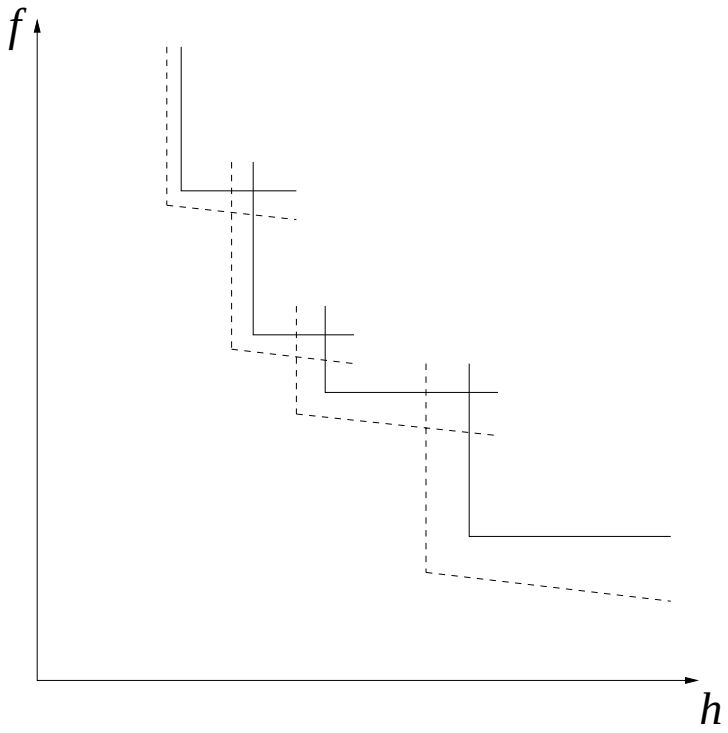


FIG. 1. *An NLP filter with slanting envelope.*

The left-hand inequality in (2.6) and also in (2.5) is an obvious way of defining a sufficient reduction in h. The right-hand inequality in (2.6) asks for a sufficient reduction in f, defined in such a way that it provides a mechanism whereby iterates are forced towards feasibility. This is shown in the following lemma and its corollary.

LEMMA 1. *Consider sequences* $\{h^{(k)}\}$ *and* $\{f^{(k)}\}$ *such that* $h^{(k)} \geq 0$ *and* $f^{(k)}$ *is monotonically decreasing and bounded below. Let constants* β *and* γ *satisfy* $0 < \gamma < \beta < 1$. *If, for all* k,

$$\textit{either} \qquad h^{(k+1)} \leq \beta h^{(k)} \qquad \textit{or} \qquad f^{(k)} - f^{(k+1)} \geq \gamma h^{(k+1)},$$

then $h^{(k)} \to 0$.

Proof. If $h^{(k+1)} \leq \beta h^{(k)}$ for all k sufficiently large, then $h^{(k)} \to 0$. Otherwise there exists an infinite subsequence $\mathcal{S}$ on which $f^{(k)} - f^{(k+1)} \geq \gamma h^{(k+1)}$. Because $f^{(k)}$ is monotonically decreasing and bounded below, it follows that $\sum_{k \in \mathcal{S}} h^{(k+1)}$ is bounded, and hence $h^{(k+1)} \to 0$ for $k \in \mathcal{S}$. But $h^{(k+1)} \leq \beta h^{(k)}$ holds for iterations $k \notin \mathcal{S}$, so it follows that $h^{(k)} \to 0$ on the main sequence. ∎

COROLLARY. *Consider an infinite sequence of iterations on which* $(h^{(k)}, f^{(k)})$ *is entered into the filter, where* $h^{(k)} > 0$ *and* $\{f^{(k)}\}$ *is bounded below. It follows that* $h^{(k)} \to 0$.

Proof. If $h^{(k+1)} \leq \beta h^{(k)}$ for all k sufficiently large, then $h^{(k)} \to 0$. Otherwise we define a subsequence $\mathcal{S}$ as follows. The initial index in $\mathcal{S}$ is the first iteration on which $h^{(k+1)} > \beta h^{(k)}$. For any $k \in \mathcal{S}$, its successor $k^+ \in \mathcal{S}$ is the least $j > k$ such that $h^{(j)} > \beta h^{(k)}$. It is a consequence of the inclusion property that $(h^{(k^+)}, f^{(k^+)})$ is acceptable to $(h^{(k)}, f^{(k)})$, even if the latter pair has been deleted from the filter on an intermediate iteration. Hence $f^{(k)} - f^{(k^+)} \geq \gamma h^{(k^+)} > 0$. Thus $f^{(k)}$ is monotonically decreasing for $k \in \mathcal{S}$ and it follows from Lemma 1 that $h^{(k)} \to 0$ for $k \in \mathcal{S}$. But intermediate iterations j such that $k < j < k^+$ have the property that $h^{(j)} \leq \beta h^{(k)}$, so it follows that $h^{(k)} \to 0$ on the main sequence. ∎

It is also convenient to allow an upper bound

$$h(\mathbf{c}(\mathbf{x})) \leq \beta u \tag{2.7}$$

$(u > 0)$ on constraint infeasibility, and this is readily implemented by initializing the filter with the entry $(u, -\infty)$. Existence of this upper bound is not necessary to the proof of convergence but is a useful practical feature that can be used to prevent iterates from becoming too infeasible. In practice we have set a large default value of $u = 10^4$, which usually has negligible impact on performance, but there are a few problems for which a much smaller value is desirable, say $u = 1$.

A common feature in a trust region algorithm for unconstrained minimization is the use of a sufficient reduction criterion

$$\Delta f \geq \sigma \Delta q, \tag{2.8}$$

where Δq is positive and $\sigma \in (0, 1)$ is a preset parameter. However, in an NLP algorithm, Δq may be negative or even zero, in which case this test is no longer appropriate. A feature of the algorithm in this paper is that it uses (2.8) only when Δq is positive. A typical value of σ that we have used is $\sigma = 0.1$.

We are now in a position to state our filter–SQP algorithm, which we do by means of the flow diagram of Figure 2. We observe that at the start of iteration k, the pair $(h^{(k)}, f^{(k)})$ is not in the current filter $\mathcal{F}^{(k)}$ but must be acceptable to it. It can be seen that there is an inner loop in which the trust region radius ρ is successively reduced until either certain tests are satisfied or the current QP subproblem becomes incompatible. (For clarity we avoid the use of the word "infeasible" in this context.) The inner loop is initialized with any value of $\rho \geq \rho^\circ$, where $\rho^\circ > 0$ is a preset parameter. The inner loop chooses a decreasing geometric sequence of values of ρ and generates corresponding values of $\mathbf{d}$, Δq, and Δf (unsubscripted). The inner loop contains a test "is $\mathbf{x}^{(k)} + \mathbf{d}$ acceptable to the filter and $(h^{(k)}, f^{(k)})$?" By this we mean that $\mathbf{x}^{(k)} + \mathbf{d}$ has to be acceptable to the filter formed of the current filter and $(h^{(k)}, f^{(k)})$, so that if $(h^{(k)}, f^{(k)})$ is subsequently entered into the filter, then $(h^{(k+1)}, f^{(k+1)})$ will still be acceptable to the new filter. When the inner iteration terminates, the current values of ρ, $\mathbf{d}$, Δq, and Δf are denoted, respectively, by $\rho^{(k)}$, $\mathbf{d}^{(k)}$, $\Delta q^{(k)}$, and $\Delta f^{(k)}$. We observe that all points that are generated by the algorithm lie in the region generated by the subset of linear constraints in the NLP problem.

Following our multiobjective thinking, we regard a step $\mathbf{d}$ that satisfies $\Delta q > 0$ as being an *f-type step* (having the primary aim of improving f, and possibly allowing an increase in h). If $\mathbf{d}$ is accepted and becomes $\mathbf{d}^{(k)}$, then an *f-type iteration* is said to have occurred. In this case we insist that the sufficient reduction condition (2.8) be

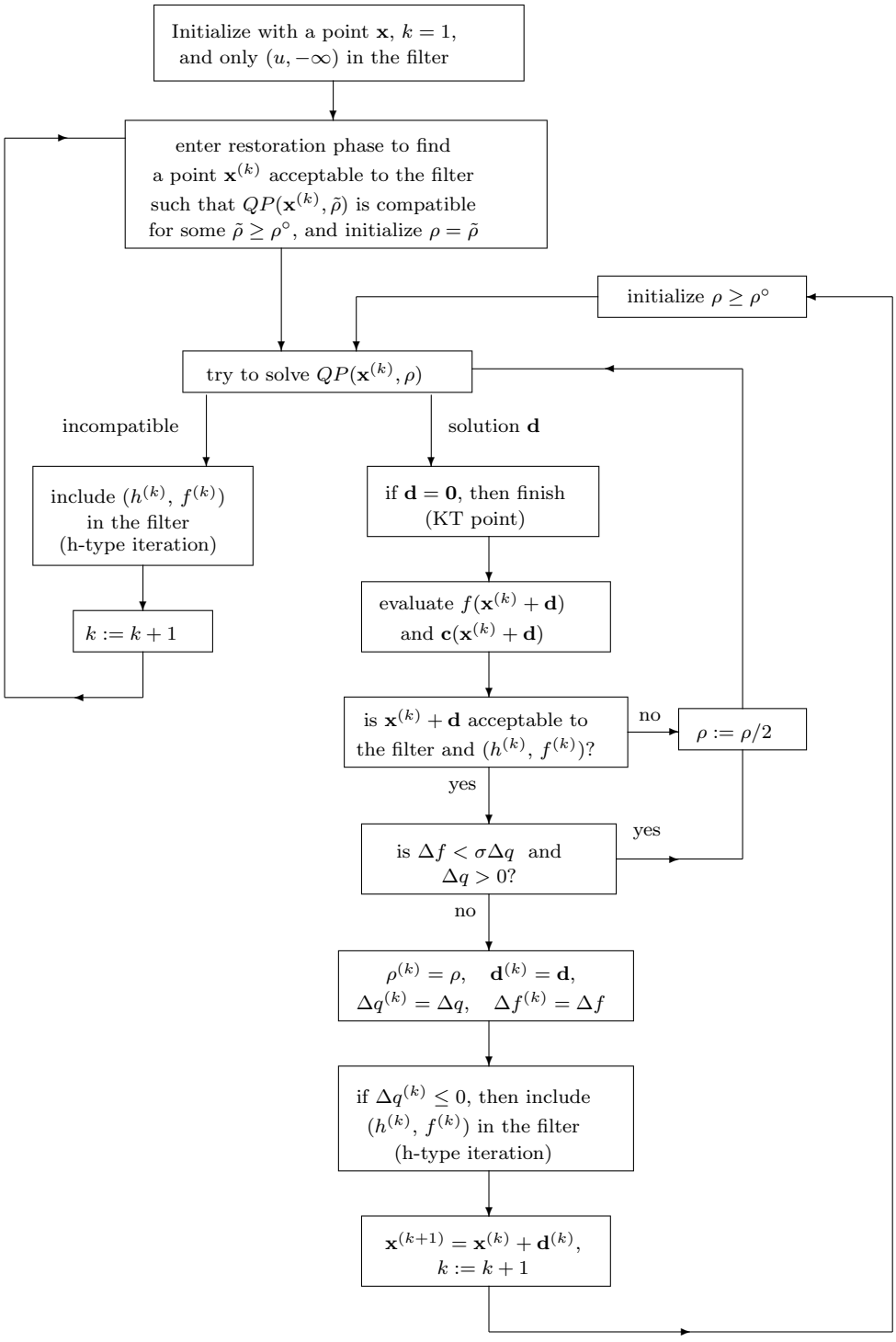


FIG. 2. *A filter–SQP algorithm.*

satisfied. Thus a necessary condition for a step $\mathbf{d}$ to give rise to an f-type iteration is that both

$$\Delta f \geq \sigma \Delta q \qquad \text{and} \qquad \Delta q > 0 \tag{2.9}$$

are satisfied. If $\Delta q^{(k)} \leq 0$, or if the current QP subproblem is incompatible, then the primary aim of the iteration is to reduce h (possibly allowing an increase in f), and we refer to the resulting iteration as an *h-type iteration*. As ρ is reduced in the inner loop, the value of Δq is reduced (a consequence of having found a global minimizer of $QP(\mathbf{x}^{(k)}, \rho)$). Thus the status of the test $\Delta q > 0$ may go from true to false, but not vice-versa. Consequently, the inner loop always samples the possibility for an f-type iteration before that of an h-type iteration. This is a key argument in the convergence proof.

This algorithm differs in one important respect from that in [5]: not all points $\mathbf{x}^{(k)}$ are included in the filter, even though they are acceptable to the filter. The point $\mathbf{x}^{(k)}$ is included in the filter at the end of the iteration if and only if that iteration is an h-type iteration. A consequence is that all the current filter entries have $h^{(j)} > 0$, $j \in \mathcal{F}^{(k)}$. This is because if $h^{(k)} = 0$, then $QP(\mathbf{x}^{(k)}, \rho)$ must be compatible, and hence, if $\mathbf{x}^{(k)}$ is not a KT point, then $\Delta q > 0$ holds. Thus if $h^{(k)} = 0$, the resulting iteration is an f-type iteration and $\mathbf{x}^{(k)}$ is not entered into the filter. It is convenient to define

$$\tau^{(k)} = \min_{j \in \mathcal{F}^{(k)}} h^{(j)} > 0. \tag{2.10}$$

It can be seen that our algorithm includes the provision for a feasibility restoration phase if the current QP subproblem becomes incompatible. Any method for solving a nonlinear algebraic system of inequalities can be used to implement this calculation, such as, for example, a Newton-like scheme for minimizing $h(\mathbf{c}(\mathbf{x}))$. The restoration phase terminates if it finds a point that both is acceptable to the filter and for which $QP(\mathbf{x},\, \rho)$ is compatible for some $\rho \geq \rho^{\circ}$. (Essentially the latter condition requires only that $QP(\mathbf{x}^{(k)}, \infty)$ be compatible, since we can always take $\rho = \infty$.) There are various existing algorithms that might be used to implement this calculation: that of Madsen [12] (with suitable changes to include inequality constraints) has a convergence proof and is close to the spirit of this paper. Alternatively, we can make use of the ideas expressed in [5], which have performed well in practice. Note that the restoration phase makes no demands on the resulting value of $f(\mathbf{x})$, which could be significantly worse than that at the previous point. If the restoration phase does terminate, then the point of termination becomes $\mathbf{x}^{(k+1)}$, and the resulting step from $\mathbf{x}^{(k)}$ to $\mathbf{x}^{(k+1)}$ is deemed to be an h-type iteration.

Of course, it may not always be possible to find a point which satisfies both the above conditions, and the restoration phase might converge to an infeasible point, for example if there exists a nonzero local minimum of $h(\mathbf{c}(\mathbf{x}))$. This is often an indication that the original problem P is incompatible. This is the situation typified by case (A) of Theorem 7 that follows in the next section. If, on the other hand, the restoration phase is converging to a feasible point, then it is usually able to terminate. This is so because $QP(\mathbf{x}, \infty)$ is usually compatible if $\mathbf{x}$ is sufficiently close to the feasible region and because $\tau^{(k)} > 0$ allows such a point to be acceptable to the filter. However, this outcome is not guaranteed, as it is possible for $QP(\mathbf{x}, \infty)$ to be incompatible for any infeasible point $\mathbf{x}$. Such an example is the pathological problem $\min(x_2 - 1)^2$ subject to $x_1^2 = 0$ and $x_1^3 = 0$, starting from $\mathbf{x} = (1,\, 0)^T$. A Newton-like iteration

for feasibility restoration is likely to converge to the feasible point $\mathbf{x} = \mathbf{0}$, which is not a solution of the NLP, without finding a point at which the QP subproblem is compatible. However, such a pathological problem (P) has the property that there exists an arbitrarily small perturbation to P for which P is incompatible. Thus in this paper we content ourselves with the possibility that the restoration phase may fail to terminate, and we regard this as an indication that the constraints of P are incompatible (in a local sense) to within round-off error.

3. A global convergence proof. In this section we present a proof of global convergence of the SQP–filter algorithm of Figure 2 when applied to problem P. We make the following assumptions.

Standard assumptions.

1. All points $\mathbf{x}$ that are sampled by the algorithm lie in a nonempty closed and bounded set X.
2. The problem functions $f(\mathbf{x})$ and $\mathbf{c}(\mathbf{x})$ are twice continuously differentiable on an open set containing X.
3. There exists an $M > 0$ such that the Hessian matrices $B^{(k)}$ satisfy $\|B^{(k)}\|_2 \leq M$ for all k.

It is a consequence of the standard assumptions that the Hessian matrices of f and the c_i are bounded on X, and without loss of generality we may assume that they also satisfy bounds $\|\nabla^2 f(\mathbf{x})\|_2 \leq M$, $\|\nabla^2 c_i(\mathbf{x})\|_2 \leq M$, $i \in \mathcal{E} \cup \mathcal{I}$, for all $\mathbf{x} \in X$.

Our global convergence theorem concerns KT necessary conditions under a Mangasarian–Fromowitz constraint qualification (MFCQ) (see, for example, Mangasarian [9]). This is essentially an extended form of the Fritz John conditions for a problem that includes equality constraints. A feasible point $\mathbf{x}^\circ$ of problem P satisfies MFCQ if and only if both (i) the vectors $\mathbf{a}_i^\circ$, $i \in \mathcal{E}$, are linearly independent and (ii) there exists a vector $\mathbf{s}$ that satisfies $\mathbf{s}^T\mathbf{a}_i^\circ = 0$, $i \in \mathcal{E}$, and $\mathbf{s}^T\mathbf{a}_i^\circ < 0$, $i \in \mathcal{A}^\circ$, where $\mathcal{A}^\circ \subset \mathcal{I}$ denotes the set of active inequality constraints at $\mathbf{x}^\circ$. Necessary conditions for $\mathbf{x}^\circ$ to solve P are that $\mathbf{x}^\circ$ is a feasible point and, if MFCQ holds, that then the set of directions

$$\{\mathbf{s} | \mathbf{s}^T\mathbf{g}^\circ < 0, \tag{3.1}$$

$$\mathbf{s}^T\mathbf{a}_i^\circ = 0, \quad i \in \mathcal{E}, \tag{3.2}$$

$$\mathbf{s}^T\mathbf{a}_i^\circ < 0, \quad i \in \mathcal{A}^\circ \ \} \tag{3.3}$$

is empty. If $\mathbf{x}^\circ$ solves P and MFCQ holds, then these conditions are equivalent to the existence of KT multipliers (although we do not use that result in this paper), and it has been shown (Gauvin [8]) that the multiplier set is bounded.

Before proving our main theorem, we need some results that describe the behavior of QP subproblems in the neighborhood of a feasible point $\mathbf{x}^\circ$ at which the vectors $\mathbf{a}_i^\circ$, $i \in \mathcal{E}$, are linearly independent. First, however, we prove two simple lemmas that enable us to handle the second order terms in the analysis.

LEMMA 2. *Consider minimizing a quadratic function $\phi(\alpha)$ $(\mathbb{R} \to \mathbb{R})$ on the interval $\alpha \in [0, 1]$ when $\phi'(0) < 0$. A necessary and sufficient condition for the minimizer to be at $\alpha = 1$ is $\phi'' + \phi'(0) \leq 0$. In this case it follows that $\phi(0) - \phi(1) \geq -\frac{1}{2}\phi'(0)$.*

Proof. Using $\phi(\alpha) = \phi(0) + \alpha\phi'(0) + \frac{1}{2}\alpha^2\phi''$, the minimizer is at $\alpha = 1$ either if $\phi'' \leq 0$ or if $\phi'' > 0$ and $-\phi'(0)/\phi'' \geq 1$, from which the result follows. ∎

LEMMA 3. *Let the standard assumptions hold, and let* $\mathbf{d}$ *be a feasible point of* $QP(\mathbf{x}^{(k)}, \rho)$. *It then follows that*

$$\Delta f \geq \Delta q - n\rho^2 M, \tag{3.4}$$

$$|c_i(\mathbf{x}^{(k)} + \mathbf{d})| \leq \tfrac{1}{2} n\rho^2 M, \qquad i \in \mathcal{E}, \tag{3.5}$$

and

$$c_i(\mathbf{x}^{(k)} + \mathbf{d}) \leq \tfrac{1}{2} n\rho^2 M, \qquad i \in \mathcal{I}. \tag{3.6}$$

Proof. These results follow from the intermediate value form of Taylor's theorem; for example,

$$f(\mathbf{x}^{(k)} + \mathbf{d}) = f^{(k)} + \mathbf{g}^{(k)T}\mathbf{d} + \tfrac{1}{2}\mathbf{d}^T \nabla^2 f(\mathbf{y})\mathbf{d},$$

where $\mathbf{y}$ denotes some point on the line segment from $\mathbf{x}^{(k)}$ to $\mathbf{x}^{(k)} + \mathbf{d}$. It follows from (2.2) and (2.1) that

$$\Delta f = \Delta q + \tfrac{1}{2}\mathbf{d}^T (B^{(k)} - \nabla^2 f(\mathbf{y}))\mathbf{d},$$

and (3.4) follows from the Hessian bounds and the inequality $\|\mathbf{d}\|_2^2 \leq n\|\mathbf{d}\|_\infty^2 \leq n\rho^2$. Also, for $i \in \mathcal{I}$, it follows that

$$c_i(\mathbf{x}^{(k)} + \mathbf{d}) = c_i^{(k)} + \mathbf{a}_i^{(k)T}\mathbf{d} + \tfrac{1}{2}\mathbf{d}^T \nabla^2 c_i(\mathbf{y}_i)\mathbf{d} \leq \tfrac{1}{2}\mathbf{d}^T \nabla^2 c_i(\mathbf{y}_i)\mathbf{d}$$

by feasibility of $\mathbf{d}$, and (3.6) then follows in a similar way. The result (3.5) follows for $i \in \mathcal{E}$ by regarding an equation as two opposed inequality constraints. ∎

LEMMA 4. *Let standard assumptions hold. If* $\mathbf{d}$ *solves* $QP(\mathbf{x}^{(k)}, \rho)$, *then* $\mathbf{x}^{(k)} + \mathbf{d}$ *is acceptable to the filter if* $\rho^2 \leq 2\beta\tau^{(k)}/(mnM)$.

Proof. It follows from (2.3), (3.5), and (3.6) that $h(\mathbf{c}(\mathbf{x}^{(k)} + \mathbf{d})) \leq \frac{1}{2} mn\rho^2 M$. If $\rho^2 \leq 2\beta\tau^{(k)}/(mnM)$, it then follows that $h(\mathbf{c}(\mathbf{x}^{(k)} + \mathbf{d})) \leq \beta\tau^{(k)}$. Hence, by the definition of $\tau^{(k)}$, the filter acceptance test (2.6) is satisfied. ∎

LEMMA 5. *Let standard assumptions hold and let* $\mathbf{x}^\circ \in X$ *be a feasible point of problem* P *at which MFCQ holds but which is not a KT point. Then there exists a neighborhood* $\mathcal{N}^\circ$ *of* $\mathbf{x}^\circ$ *and positive constants* ε, μ, *and* κ *such that for all* $\mathbf{x} \in \mathcal{N}^\circ \cap X$ *and all* ρ *for which*

$$\mu h(\mathbf{c}(\mathbf{x}))) \leq \rho \leq \kappa \tag{3.7}$$

it follows that $QP(\mathbf{x}, \rho)$ *has a feasible solution* $\mathbf{d}$ *at which the predicted reduction* (2.1) *satisfies*

$$\Delta q \geq \tfrac{1}{3}\rho\varepsilon, \tag{3.8}$$

the sufficient reduction condition (2.8) *holds, and the actual reduction* (2.2) *satisfies*

$$\Delta f \geq \gamma h(\mathbf{c}(\mathbf{x} + \mathbf{d})). \tag{3.9}$$

Proof. Since $\mathbf{x}^\circ$ is a feasible point at which MFCQ holds but it is not a KT point, it follows that the vectors $\mathbf{a}_i^\circ$, $i \in \mathcal{E}$, are linearly independent, and there exists a vector $\mathbf{s}^\circ$, for which $\|\mathbf{s}^\circ\|_2 = 1$, that satisfies (3.1), (3.2), and (3.3). We note that these

conditions imply that the cardinality $|\mathcal{E}| < n$. We use the notation $A^+ = (A^T A)^{-1} A^T$, and let $A_\mathcal{E}$ denote the matrix with columns $\mathbf{a}_i$, $i \in \mathcal{E}$, evaluated at some point $\mathbf{x}$. By linear independence and continuity there exists a neighborhood of $\mathbf{x}^\circ$ in which $A_\mathcal{E}^+$ is bounded. If $\mathcal{E}$ is not empty, we denote $\mathbf{p} = -A_\mathcal{E}^{+T} \mathbf{c}_\mathcal{E}$, which is the closest point in the linearized equality constraint manifold to $\mathbf{d} = \mathbf{0}$, and let $p = \|\mathbf{p}\|_2$. Also we denote $\mathbf{s} = (I - A_\mathcal{E} A_\mathcal{E}^+)\mathbf{s}^\circ / \|(I - A_\mathcal{E} A_\mathcal{E}^+)\mathbf{s}^\circ\|_2$, which is the closest unit vector to $\mathbf{s}^\circ$ in the null space of $A_\mathcal{E}^T$. If $\mathcal{E}$ is empty, we let $\mathbf{p} = \mathbf{0}$, $p = 0$, and $\mathbf{s} = \mathbf{s}^\circ$. It follows from (3.1) and (3.3) by continuity that there exists a (smaller) neighborhood $\mathcal{N}^\circ$ and a constant $\varepsilon > 0$ such that

$$\mathbf{s}^T \mathbf{g} \le -\varepsilon \qquad \text{and} \qquad \mathbf{s}^T \mathbf{a}_i \le -\varepsilon, \quad i \in \mathcal{A}^\circ, \tag{3.10}$$

when $\mathbf{g}$, $\mathbf{a}_i$, and $\mathbf{s}$ are evaluated for any $\mathbf{x} \in \mathcal{N}^\circ$. By definition of $\mathbf{p}$, it follows that $p = O(h(\mathbf{c}))$, and thus we can choose the constant μ in (3.7) sufficiently large so that $\rho > p$ for all $\mathbf{x} \in \mathcal{N}^\circ$.

We now consider the solution of $QP(\mathbf{x}, \rho)$ and, in particular, the line segment defined by

$$\mathbf{d}_\alpha = \mathbf{p} + \alpha(\rho - p)\mathbf{s}, \quad \alpha \in [0, 1], \tag{3.11}$$

for a fixed value of $\rho > p$. We note that $\mathbf{d}_\alpha$ satisfies the equality constraints $\mathbf{c}_\mathcal{E} + A_\mathcal{E}^T \mathbf{d} = \mathbf{0}$ of $QP(\mathbf{x}, \rho)$ for any value of α. Because the vectors $\mathbf{p}$ and $\mathbf{s}$ are orthogonal, it follows that

$$\|\mathbf{d}_1\|_2 = \sqrt{p^2 + (\rho - p)^2} = \sqrt{\rho^2 - 2\rho p + 2p^2} \le \rho$$

since $\rho > p$. Consequently $\|\mathbf{d}_1\|_\infty \le \rho$, and hence $\mathbf{d}_1$ satisfies the trust region constraint of $QP(\mathbf{x}, \rho)$.

Next we look at the inactive constraints $i \in \mathcal{I}/\mathcal{A}^\circ$. If $\mathbf{x} \in \mathcal{N}^\circ \cap X$, then there exist positive constants $\bar{c}$ and $\bar{a}$, independent of ρ, such that

$$c_i \le -\bar{c} \qquad \text{and} \qquad \mathbf{a}_i^T \mathbf{s} \le \bar{a}$$

for all vectors $\mathbf{s}$ such that $\|\mathbf{s}\|_\infty \le 1$, by the continuity of c_i and boundedness of $\mathbf{a}_i$ on X. It follows that

$$c_i + \mathbf{a}_i^T \mathbf{d} \le -\bar{c} + \rho\bar{a}, \qquad i \in \frac{\mathcal{I}}{\mathcal{A}^\circ},$$

for all vectors $\mathbf{d}$ such that $\|\mathbf{d}\|_\infty \le \rho$. Thus inactive constraints do not affect the solution to $QP(\mathbf{x}, \rho)$ if ρ satisfies $\rho \le \bar{c}/\bar{a}$.

For active inequality constraints $i \in \mathcal{A}^\circ$, we have from (3.10) and (3.11) that

$$c_i + \mathbf{a}_i^T \mathbf{d}_1 = c_i + \mathbf{a}_i^T \mathbf{p} + (\rho - p)\mathbf{a}_i^T \mathbf{s} \le c_i + \mathbf{a}_i^T \mathbf{p} - (\rho - p)\varepsilon \le 0$$

if

$$\rho \ge p + \frac{(c_i + \mathbf{a}_i^T \mathbf{p})}{\varepsilon}.$$

By the definition of $\mathbf{p}$, the right-hand side of this inequality is $O(h(\mathbf{c}))$, and thus we can choose the constant μ in (3.7) sufficiently large so that $c_i + \mathbf{a}_i^T \mathbf{d}_1 \le 0$, $i \in \mathcal{A}^\circ$. Thus $\mathbf{d}_1$ is feasible in $QP(\mathbf{x}, \rho)$ with respect to the active inequality constraints, and hence

to all the constraints, using results from above. Hence we have shown that $QP(\mathbf{x}, \rho)$ is compatible for all $\mathbf{x} \in \mathcal{N}^\circ$ and all ρ satisfying (3.7) for any value of $\kappa \leq \bar{c}/\bar{a}$.

Next we aim to obtain a bound on the predicted reduction Δq and hence show that (3.8), (2.8), and (3.9) hold. First we consider the line segment (3.11) and define $\phi(\alpha) = q(\mathbf{p} + \alpha(\rho - p)\mathbf{s})$. It follows that

$$\phi'(\alpha) = (\rho - p)\mathbf{s}^T \nabla q(\mathbf{p} + \alpha(\rho - p)\mathbf{s}) = (\rho - p)\mathbf{s}^T(\mathbf{g} + B(\mathbf{p} + \alpha(\rho - p)\mathbf{s})).$$

Hence, using (3.10), bounds on B and $\mathbf{p}$, and $\rho > p$,

$$\phi'(0) = (\rho - p)\mathbf{s}^T(\mathbf{g} + B\mathbf{p}) \leq (\rho - p)(\mathbf{s}^T B\mathbf{p} - \varepsilon) \leq (\rho - p)(Mp - \varepsilon) < (\rho - p)(M\rho - \varepsilon) \leq 0$$

if $\rho \leq \varepsilon/M$. Now $\phi'' = (\rho - p)^2 \mathbf{s}^T B\mathbf{s} \leq (\rho - p)^2 M$, and thus

$$\phi'' + \phi'(0) \leq (\rho - p)^2 M + (\rho - p)(Mp - \varepsilon) = (\rho - p)((\rho - p)M + Mp - \varepsilon) \leq 0$$

if $\rho \leq \varepsilon/M$. In this case, applying Lemma 2, the minimum value of $\phi(\alpha)$ occurs at $\alpha = 1$, and the reduction in q satisfies $\phi(0) - \phi(1) \geq -\frac{1}{2}\phi'(0)$. After adding in a contribution for the change in q along $\mathbf{p}$, we may express

$$q(\mathbf{0}) - q(\mathbf{d}_1) \geq \tfrac{1}{2}(\rho - p)(\varepsilon - \mathbf{s}^T B\mathbf{p}) + O(p) \geq \tfrac{1}{2}\rho\varepsilon + O(p).$$

Since $\mathbf{d}_1$ is feasible and $p = O(h(\mathbf{c}))$, it follows that the predicted reduction (2.1) satisfies

$$\Delta q \geq \tfrac{1}{2}\rho\varepsilon + O(h(\mathbf{c})) \geq \tfrac{1}{2}\rho\varepsilon - \xi h(\mathbf{c})$$

for some ξ sufficiently large and independent of ρ. Thus (3.8) is satisfied if $\rho \geq 6\xi h(\mathbf{c})/\varepsilon$. This condition can be achieved by making the constant μ in (3.7) sufficiently large. It follows from (3.4) and (3.8) that

$$\frac{\Delta f}{\Delta q} \geq 1 - \frac{n\rho^2 M}{\Delta q} \geq 1 - \frac{3n\rho^2 M}{\rho\varepsilon} = 1 - \frac{3n\rho M}{\varepsilon}.$$

Then, if $\rho \leq (1 - \sigma)\varepsilon/(3nM)$, it follows that (2.8) holds.

Finally, we deduce from (2.3), (3.5), (3.6), (2.8), and (3.8) that

$$f^{(k)} - f - \gamma h(\mathbf{c}(\mathbf{x} + \mathbf{d})) = \Delta f - \gamma h(\mathbf{c}(\mathbf{x} + \mathbf{d})) \geq \tfrac{1}{3}\sigma\rho\varepsilon - \tfrac{1}{2}\gamma mn\rho^2 M \geq 0$$

if $\rho \leq \frac{2}{3}\sigma\varepsilon/(\gamma mnM)$. Thus we may define the constant κ in (3.7) to be the least of $\frac{2}{3}\sigma\varepsilon/(\gamma mnM)$ and the values $(1 - \sigma)\varepsilon/(3nM)$, ε/M, and $\bar{c}/\bar{a}$, as required earlier in the proof. ∎

Now we proceed to analyze the algorithm of Figure 2. First we need a result that is similar to Lemma 2 of [6]. Here $\mathbf{x}^{(k)}$ and $B^{(k)}$ are fixed, and we consider what happens to the solution of $QP(\mathbf{x}^{(k)}, \rho)$ as ρ is reduced.

LEMMA 6. *Let the standard assumptions hold; then the inner iteration terminates finitely.*

Proof. If $\mathbf{x}^{(k)}$ is a KT point of problem P, then $\mathbf{d} = \mathbf{0}$ solves $QP(\mathbf{x}^{(k)}, \rho)$ and the algorithm terminates. Otherwise, if the inner iteration does not terminate finitely, then the rule for decreasing ρ ensures that $\rho \to 0$. Two cases need to be considered, depending on whether $h^{(k)} > 0$ or $h^{(k)} = 0$.

If $h^{(k)} > 0$ and $i \in \mathcal{E} \cup \mathcal{I}$ is an index for which $c_i^{(k)} > 0$, then for all $\mathbf{d}$ such that $\|\mathbf{d}\|_\infty \leq \rho$ it follows that

$$c_i^{(k)} + \mathbf{a}_i^{(k)T}\mathbf{d} \geq c_i^{(k)} - \rho\|\mathbf{a}_i^{(k)}\|_1 > 0$$

if either $\|\mathbf{a}_i^{(k)}\|_1 = 0$ or $\rho < c_i^{(k)}/\|\mathbf{a}_i^{(k)}\|_1$. Thus for sufficiently small ρ, constraint i cannot be satisfied and $QP(\mathbf{x}^{(k)}, \rho)$ is incompatible. A similar conclusion is obtained for $i \in \mathcal{E}$ if $c_i^{(k)} < 0$. Thus the inner iteration terminates finitely if $h^{(k)} > 0$.

If $h^{(k)} = 0$, then by a similar argument, inactive constraints at $\mathbf{x}^{(k)}$ are inactive at any point for which $\|\mathbf{d}\|_\infty \leq \rho$, for sufficiently small ρ. Thus we need consider only constraints $i \in \mathcal{E} \cup \mathcal{A}^{(k)}$. The rest of the proof is now similar to that of Lemma 5 in the case $p = 0$. Because $\mathbf{x}^{(k)}$ is not a KT point, there exists a vector $\mathbf{s}$, $\|\mathbf{s}\|_2 = 1$, and an $\eta > 0$ such that $\mathbf{s}^T\mathbf{g}^{(k)} = -\eta$, $\mathbf{s}^T\mathbf{a}_i^{(k)} = 0$, $i \in \mathcal{E}$, and $\mathbf{s}^T\mathbf{a}_i^{(k)} \leq 0$, $i \in \mathcal{A}^{(k)}$. We consider the QP-feasible line segment $\mathbf{d}_\alpha = \alpha\rho\mathbf{s}$ for $\alpha \in [0, 1]$ and construct the function $\phi(\alpha) = q(\mathbf{d}_\alpha)$. It follows that $\phi'(0) = -\rho\eta$ and $\phi'' = \rho^2\mathbf{s}^T B^{(k)}\mathbf{s} \leq \rho^2 M$. Hence if $\rho \leq \eta/M$, it follows that $\phi'' + \phi'(0) \leq 0$. It then follows from Lemma 2 that $\phi(0) - \phi(1) \geq \frac{1}{2}\rho\eta$. Therefore, by the global optimality of the solution $\mathbf{d}$ to $QP(\mathbf{x}^{(k)}, \rho)$, the actual reduction Δq also satisfies $\Delta q \geq \frac{1}{2}\rho\eta$, and if $\rho \leq (1-\sigma)\eta/(2nM)$, it follows from (3.4) that $\Delta f \geq \sigma\Delta q > 0$ and the necessary condition (2.9) for an f-type iteration is satisfied. Also, from (3.4), (3.5), and (3.6),

$$f^{(k)} - f(\mathbf{x}^{(k)} + \mathbf{d}) - \gamma h(\mathbf{c}(\mathbf{x}^{(k)} + \mathbf{d})) = \Delta f - \gamma h(\mathbf{c}(\mathbf{x}^{(k)} + \mathbf{d})) \geq \tfrac{1}{2}\sigma\rho\eta - \tfrac{1}{2}\gamma mn\rho^2 M \geq 0$$

if $\rho \leq \sigma\eta/(\gamma mnM)$. In this case it follows that $\mathbf{x}^{(k)} + \mathbf{d}$ is acceptable relative to $(h^{(k)}, f^{(k)})$. Finally, from Lemma 4, $\mathbf{x}^{(k)} + \mathbf{d}$ is acceptable to the filter if $\rho^2 \leq 2\beta\tau^{(k)}/(mnM)$. Thus, if ρ is sufficiently small, all the conditions for an f-type step are satisfied and the inner iteration terminates finitely. ∎

We are now in a position to state our main theorem.

THEOREM 7. *If standard assumptions hold, the outcome of applying the filter–SQP algorithm of Figure* 2 *is one of the following.*

(A) *The restoration phase fails to find a point* $\mathbf{x}$ *which is both acceptable to the filter and for which* $QP(\mathbf{x}, \rho)$ *is compatible for some* $\rho \geq \rho^\circ$.

(B) *A KT point of problem P is found. (*$\mathbf{d} = \mathbf{0}$ *solves* $QP(\mathbf{x}^{(k)}, \rho)$ *for some k.)*

(C) *There exists an accumulation point that is feasible and either is a KT point or fails to satisfy MFCQ.*

Proof. We need consider only the case in which neither (A) nor (B) occurs. Because the inner loop of each iteration is finite (Lemma 6), the outer iteration sequence indexed by k is infinite. All iterates $\mathbf{x}^{(k)}$ lie in X, which is bounded, so it follows that the sequence has one or more accumulation points.

First, we consider the case that the main sequence contains an infinite number of h-type iterations, and we consider this subsequence. For an h-type iteration, $(h^{(k)}, f^{(k)})$ is always entered into the filter at the completion of the iteration, so it follows from the Corollary to Lemma 1 that $h^{(k)} \to 0$ on this subsequence. It must also follow that $\tau^{(k)} \to 0$. Moreover, only h-type iterations can reset $\tau^{(k)}$, so there exists a thinner infinite subsequence on which $\tau^{(k+1)} = h^{(k)} < \tau^{(k)}$ is set. Because X is bounded, there exists an accumulation point $\mathbf{x}^\infty$ and a subsequence indexed by $k \in \mathcal{S}$ of h-type iterations for which $\mathbf{x}^{(k)} \to \mathbf{x}^\infty$, $h^{(k)} \to 0$, and $\tau^{(k+1)} = h^{(k)} < \tau^{(k)}$. One consequence is that $\mathbf{x}^\infty$ is a feasible point. If MFCQ is not satisfied at $\mathbf{x}^\infty$, then (C) is established in this case. We therefore assume that MFCQ is satisfied and consider the proposition (to be contradicted) that $\mathbf{x}^\infty$ is not a KT point. In this case, the vectors

$\mathbf{a}_i^\infty$, $i \in \mathcal{E}$, are linearly independent, and the set defined by (3.1), (3.2), and (3.3) is not empty. For sufficiently large $k \in \mathcal{S}$ it follows that $\mathbf{x}^{(k)}$ is in the neighborhood $\mathcal{N}^\infty$, as defined in Lemma 5. We show that this leads to a contradiction.

Lemma 5 provides conditions on ρ which ensure that $QP(\mathbf{x}^{(k)}, \rho)$ is compatible, and the resulting step $\mathbf{d}$ satisfies $\Delta f \geq \sigma \Delta q > 0$ and $f^{(k)} \geq f + \gamma h$, where f and h denote $f = f(\mathbf{x}^{(k)} + \mathbf{d})$ and $h = h(\mathbf{c}(\mathbf{x}^{(k)} + \mathbf{d}))$, respectively. This shows that the necessary condition (2.9) for an f-type step is satisfied, and the entry (h, f) is acceptable to (not dominated by) $(h^{(k)}, f^{(k)})$. Moreover, it follows from Lemma 4 that $\mathbf{x}^{(k)} + \mathbf{d}$ is acceptable to the filter if $\rho^2 \leq 2\beta\tau^{(k)}/(mnM)$. Thus we deduce that if ρ satisfies

$$\mu h^{(k)} \leq \rho \leq \min\left\{ \sqrt{\frac{2\beta\tau^{(k)}}{mnM}}, \kappa \right\}, \tag{3.12}$$

then (h, f) satisfies all the conditions for an f-type iteration.

Now we need to show that a value of ρ in this range will be located by the inner iteration. It follows for $k \in \mathcal{S}$ sufficiently large that $\tau^{(k)} \to 0$ and the range (3.12) becomes

$$\mu h^{(k)} \leq \rho \leq \sqrt{\frac{2\beta\tau^{(k)}}{mnM}}. \tag{3.13}$$

In the limit, because $h^{(k)} < \tau^{(k)}$, and because of the square root, the upper bound in (3.13) is more than twice the lower bound. Now consider how the inner loop of the algorithm works. Initially a value $\rho \geq \rho^\circ$ is chosen, which in the limit will be greater than the upper bound in (3.13). Then successively halving ρ in the inner loop will eventually locate a value in the interval (3.13), or to the right of this interval, which provides the conditions for an f-type step to occur. It is not possible for any value of $\rho \geq \mu h^{(k)}$ to produce an h-type step since Δq decreases monotonically as ρ decreases (this is a consequence of the global optimality of $\mathbf{d}$). Thus if $k \in \mathcal{S}$ is sufficiently large, an f-type iteration will result. This contradicts the fact that the subsequence is composed of h-type iterations. Thus $\mathbf{x}^\infty$ is a KT point and (C) is established in this case.

Next we consider the alternative case that the main sequence contains only a finite number of h-type iterations. Hence there exists an index K such that all iterations are f-type iterations for all $k \geq K$. It follows that $(h^{(k+1)}, f^{(k+1)})$ is always acceptable to $(h^{(k)}, f^{(k)})$, and also that $\Delta f^{(k)} \geq \sigma \Delta q^{(k)} > 0$, so that the sequence of function values $\{f^{(k)}\}$ is strictly monotonically decreasing for $k \geq K$. It therefore follows from Lemma 1 that $h^{(k)} \to 0$ and hence that any accumulation point $\mathbf{x}^\infty$ of the main sequence is a feasible point. Because $f(\mathbf{x})$ is bounded on X, it also follows that $\sum_{k \geq K} \Delta f^{(k)}$ is convergent. As above, we now aim to contradict the proposition that there exists an accumulation point at which MFCQ holds that is not a KT point.

Because all iterations $k \geq K$ are f-type, no filter entries are made, so $\tau^{(k)} = \tau^{(K)}$ is constant. For sufficiently large $k \geq K$ it follows that $\mathbf{x}^{(k)}$ is in the neighborhood $\mathcal{N}^\infty$ defined in Lemma 5. It follows as above that sufficient conditions for accepting an f-type step are that ρ satisfies

$$\mu h^{(k)} \leq \rho \leq \min\left\{ \sqrt{\frac{2\beta\tau^{(K)}}{mnM}}, \kappa \right\}. \tag{3.14}$$

This time the right-hand side of (3.14) is a constant, $\bar{\rho}$ say ($\bar{\rho} > 0$) independent of k, while the left-hand side converges to zero. Thus, for sufficiently large k, the upper

bound must be greater than twice the lower bound. In this case, as ρ is reduced in the inner loop, either it must eventually fall within this interval or a value to the right of the interval is accepted. Hence we can guarantee that a value $\rho^{(k)} \geq \min(\frac{1}{2}\bar{\rho}, \rho^\circ)$ will be chosen. We then deduce from (2.8) and (3.8) that $\Delta f^{(k)} \geq \frac{1}{3}\sigma\varepsilon \min(\frac{1}{2}\bar{\rho}, \rho^\circ)$, which contradicts the fact that $\sum_{k\geq K} \Delta f^{(k)}$ is convergent. Thus $\mathbf{x}^\infty$ is a KT point, and (C) is established in this case also. □

4. Discussion. Of course, the algorithm of Figure 2 is only a guide to what might be successfully implemented in practice, and it is incomplete in various ways. For example, it is necessary to make a specific choice of algorithm to implement the restoration phase. Also, the rule for adjusting ρ in the inner iteration could be more intricate, based partly on interpolation. Another possibility is to allow the pair $(h^{(k)}, f^{(k)})$ to be entered into the filter on an f-type step if $h^{(k)} \geq \tau^{(k)}$, as this does not affect the convergence proof. An overall strategic decision is that of how to specify the matrix $B^{(k)}$. One possibility is to use a Lagrangian Hessian based on exact second derivatives and estimates of Lagrange multipliers. A disadvantage of this is that the matrices $B^{(k)}$ may be indefinite, in which case finding the global minimizer of the QP subproblem is problematic. An alternative possibility is to use some quasi-Newton formula to update $B^{(k)}$, in which case it might be possible to ensure that $B^{(k)}$ is positive semidefinite, and hence any KT point of the QP subproblem is a global solution. It is also not easy to prove that $B^{(k)}$ is bounded. However, when MFCQ holds, it can be expected that Lagrange multiplier estimates are bounded and hence that $B^{(k)}$ is bounded. In practice, the algorithm has been implemented with an exact Hessian with very satisfactory performance, akin to that reported in [5]. Preliminary practical experience with a quasi-Newton form of the algorithm is also promising. There are other ways in which the potential difficulty of finding the global minimizer of the QP subproblem might be avoided, while retaining the rapid convergence normally associated with an SQP algorithm, and some of these are described later in the section.

The choice of an initial value of ρ for the inner iteration requires that the condition $\rho \geq \rho^\circ$ be satisfied but is otherwise unspecific. We envisage that in practice ρ° is close to zero (say 10^{-4}) so that the effect of this restriction is small. Thus to a large extent the algorithm of Figure 2 allows the more usual trust region procedure in which one may double or halve (say) the value of ρ from the previous iteration, setting $\rho = \rho^\circ$ only if it would otherwise be less than ρ°. The potential danger of just taking ρ from the previous iteration is that the existence of a successful f-type step may not be recognized. By starting with $\rho \geq \rho^\circ$, we ensure that ρ is greater than twice the lower bound $\mu h^{(k)}$ in the limit and hence that an f-type step will be taken if the range allows. Adjusting the trust region in this sort of way has featured in other recent work; see, for example, [10], [11] and references contained therein.

Another important aspect that we have not addressed in this paper is to consider the asymptotic behavior of the algorithm to ensure that the second order convergence property of the SQP iteration is not compromised. We have already given some thought to this, but it is not yet clear how to make progress. The algorithm in [5] allows the use of a second order correction step, although it is not clear in practice whether this is necessary or even beneficial. We shall continue to study such issues in our future work.

The referees for the paper both made the point that the link between f and $\mathbf{c}$ that is implicit in the second inequality of (2.6) is undesirable. Aesthetically we agree that it would be preferable not to have this link, although we submit that its effect

is minimal. We stress that the parameter γ is intended to be close to zero (typically 10^{-5}), so that this inequality is little different in practice from that in (2.4), in which case there is no linkage. We have successfully implemented this type of algorithm in practice, with results of a quality similar to those in [5], and changing to $f < f^{(j)}$ causes negligible difference in the outcome. It may well be desirable to take the relative scaling of f and h into account, but this is readily done.

In any event, it is by no means clear how to avoid the linkage between f and **c**. The step **d** that solves the QP subproblem is not a descent direction for f when $\Delta q \leq 0$, so we cannot use any analogue of the Goldstein or Wolfe–Powell tests from unconstrained optimization. We feel that our proposals are noteworthy in that they enable a convergence proof to be made in such a way that the linkages between f and h are small and the impact on practical performance is negligible.

The authors of this paper have also contributed to other papers that suggest filter-type algorithms for which global convergence can be proved. One paper uses ideas akin to those suggested by Fletcher and Sainz de la Maza [7], in which an LP trust region subproblem is solved in order to obtain an estimate of the active set, which can then be used in an equality QP calculation to determine a trial step. The theoretical and practical properties of this approach have been investigated by a student, C. M. Chin, and are reported in Chin and Fletcher [2], [3]. Another approach, suggested by Fletcher, Gould, Leyffer, and Toint [4], is a trust region SQP algorithm using a filter but which allows the use of an approximate solution **d** to the QP subproblem. The algorithm is based on a decomposition of the step **d** into its normal and tangential components. A proof of global convergence to a first order critical point is given in that report. The proof is significantly different from that in this paper and provides a different outlook on the problem, more related to the familiar Cauchy point decrease condition that appears elsewhere in the trust region literature (see, for example, Conn, Gould, and Toint [1]). It is an advantage that the proof allows an approximate solution to the QP subproblem, but there is also a disadvantage that it relies on certain conditions that may require an expensive projection calculation to verify. Also, the filter envelope (2.5) is used rather than the slanting filter envelope (2.6) used in this paper. No practical experience with the Cauchy-type of algorithm is as yet available.

Global convergence proofs for other filter-related algorithms that do not use merit functions have also been set out in recent papers. Ulbrich, Ulbrich, and Vicente [14] use a decomposition into normal and tangential components of a primal-dual interior point step, as well as a filter to decide on acceptability. The work of Ulbrich and Ulbrich [13] uses nonmonotonic improvement conditions on both the normal and tangential steps and obtains global convergence using an acceptance test based on comparing the normal and tangential predicted reductions with a suitably chosen weighting parameter. Encouraging numerical results of a preliminary MATLAB program on a range of CUTE test problems are presented. Wächter and Biegler [15] describe a line search method in which the NLP problem is converted into equations and simple bounds, and a filter is used to balance the contributions of a barrier function for the simple bounds and a constraint violation function for the equations. Both [14] and [15] present additional results relating to second order convergence.

REFERENCES

[1] A.R. Conn, N.I.M. Gould, and Ph.L. Toint, *Trust Region Methods*, MPS/SIAM Ser. Optim. MP01, SIAM, Philadelphia, 2000.

[2] C.M. Chin and R. Fletcher, *On the Global Convergence of an SLP–Filter Algorithm That Takes EQP Steps*, Report NA/199, Department of Mathematics, Dundee University, Dundee, Scotland, 2001.

[3] C.M. Chin and R. Fletcher, *Numerical Results of SLPSQP, FilterSQP and LANCELOT on Selected Cute Test Problems*, Report NA/203, Department of Mathematics, Dundee University, Dundee, Scotland, 2001.

[4] R. Fletcher, N.I.M. Gould, S. Leyffer, and Ph-.L. Toint, *Global Convergence of Trust-Region SQP-Filter Algorithms for Nonlinear Programming*, Technical report 99/03, Department of Mathematics, University of Namur, Namur, Belgium, 1999 (revised, 2001).

[5] R. Fletcher and S. Leyffer, *Nonlinear programming without a penalty function*, Math. Program., 91 (2002), pp. 239–269.

[6] R. Fletcher, S. Leyffer, and Ph-.L. Toint, *On the Global Convergence of an SLP–Filter Algorithm*, Report NA/183, Department of Mathematics, Dundee University, Dundee, Scotland, 1999.

[7] R. Fletcher and E. Sainz de la Maza, *Nonlinear programming and nonsmooth optimization by successive linear programming*, Math. Programming, 43 (1989), pp. 235–256.

[8] J. Gauvin, *A necessary and sufficient regularity condition to have bounded multipliers in nonconvex programming*, Math. Programming, 12 (1977), pp. 136–138.

[9] O.L. Mangasarian, *Nonlinear Programming*, McGraw-Hill, New York, 1969.

[10] H. Jiang, M. Fukushima, L. Qi, and D. Sun, *A trust region method for solving generalized complementarity problems*, SIAM J. Optim., 8 (1998), pp. 140–157.

[11] Ch. Kanzow and M. Zupke, *Inexact trust-region methods for nonlinear complementarity problems*, in Reformulation—Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods, M. Fukushima and L. Qi, eds., Keuwer Academic Publishers, Dordrecht, The Netherlands, 1999, pp. 211–233.

[12] K. Madsen, *An algorithm for minimax solution of overdetermined systems of nonlinear equations*, J. Inst. Math. Appl., 16 (1975), pp. 321–328.

[13] M. Ulbrich and S. Ulbrich, *Non-monotone Trust Region Methods for Nonlinear Equality Constrained Optimization without a Penalty Function*, Technical report, Zentrum Mathematik, Technische Universität München, Munich, Germany, 2000.

[14] M. Ulbrich, S. Ulbrich, and L.N. Vicente, *A Globally Convergent Primal-Dual Interior-Point Filter Method for Nonconvex Nonlinear Programming*, Technical report TR00-12, Department of Computational and Applied Mathematics, Rice University, Houston, TX, 2000.

[15] A. Wächter and L.T. Biegler, *Global and Local Convergence of Line Search Filter Methods for Nonlinear Programming*, CAPD Technical report B-01-09, Department of Chemical Engineering, Carnegie Mellon University, Pittsburgh, PA, 2001.

SIAM J. OPTIM.
Vol. 13, No. 1, pp. 60–78

DUAL STOCHASTIC DOMINANCE AND RELATED MEAN-RISK MODELS*

WŁODZIMIERZ OGRYCZAK† AND ANDRZEJ RUSZCZYŃSKI‡

Abstract. We consider the problem of constructing mean-risk models which are consistent with the second degree stochastic dominance relation. By exploiting duality relations of convex analysis we develop the quantile model of stochastic dominance for general distributions. This allows us to show that several models using quantiles and tail characteristics of the distribution are in harmony with the stochastic dominance relation. We also provide stochastic linear programming formulations of these models.

Key words. decisions under uncertainty, stochastic dominance, Fenchel duality, mean-risk analysis, quantile risk measures, stochastic programming

AMS subject classifications. Primary, 90A05, 90A46, 52A41; Secondary, 90A09, 90C15

PII. S1052623400375075

1. Introduction.

The relation of *stochastic dominance* is one of the fundamental concepts of decision theory (cf. [32, 14]). It introduces a partial order in the space of real random variables. While theoretically attractive, stochastic dominance order is computationally very difficult, as it involves a multiobjective model with a continuum of objectives.

The practice of decision making under uncertainty frequently resorts to *mean-risk models* (cf. [18]). The mean-risk approach uses only two criteria: the *mean*, representing the expected outcome, and the *risk*, a scalar measure of the variability of outcomes. This allows a simple trade-off analysis, analytical or geometrical. However, for typical dispersion statistics used as risk measures, the mean-risk approach may lead to inferior conclusions; that is, some efficient (in the mean-risk sense) solutions may be stochastically dominated by other feasible solutions. It is of primary importance to construct mean-risk models which are in harmony with stochastic dominance relations.

The classical Markowitz model [17] uses the variance as the risk measure in the mean-risk analysis. Since its introduction, many authors have pointed out that the mean-variance model is, in general, not consistent with stochastic dominance rules. In our preceding paper [22] we have proved that the standard semideviation (square root of the semivariance) or the mean absolute deviation (from the mean) as risk measures make the corresponding mean-risk models consistent with the second degree stochastic dominance, provided that the trade-off coefficient is bounded by a certain constant. These results were further generalized in [7, 23], where it was shown that mean-risk models using higher order central semideviations as risk measures are in harmony with the stochastic dominance relations of the corresponding degree.

When applied to portfolio selection or similar optimization problems with polyhedral feasible sets, the mean-variance approach results in a quadratic programming problem. Following Sharpe's [31] work on a linear programming (LP) approxima-

*Received by the editors July 10, 2000; accepted for publication (in revised form) November 22, 2001; published electronically May 15, 2002.
http://www.siam.org/journals/siopt/13-1/37507.html
†Institute of Control and Computation Engineering, Warsaw University of Technology, 00–665 Warsaw, Poland (W.Ogryczak@ia.pw.edu.pl).
‡Department of Management Science and Information Systems and RUTCOR, Rutgers University, Piscataway, NJ 08854 (rusz@rutcor.rutgers.edu).

tion to the mean-variance model, many attempts have been made to linearize the portfolio optimization problem. This resulted in the consideration of various risk measures which were LP computable in the case of finite discrete random variables. Yitzhaki [33] introduced a mean-risk model using the Gini mean (absolute) difference as a risk measure. Konno and Yamazaki [12] analyzed a model in which risk is measured by the (mean) absolute deviation. Young [34] considered the minimax approach (the worst case performances) to measure risk. If the rates of return are multivariate and normally distributed, then most of these models are equivalent to the Markowitz mean-variance model. However, they do not require any specific type of return distributions, and, as opposed to the mean-variance approach, they can be applied to general (possibly nonsymmetric) random variables. In the case of finite discrete random variables, all these mean-risk models have LP formulations and are special cases of the multiple criteria LP model [21] based on majorization theory [10, 19] and Lorenz-type orders [16, 1].

In this paper we analyze a *dual* model of stochastic dominance by exploiting duality relations of convex analysis (see, e.g., [27]). These transformations allow us to show the consistency with stochastic dominance of mean-risk models, using quantiles and tail characteristics of the distribution as risk measures. We also show that these models are equivalent to certain stochastic LP problems, thus opening a new area of applications for stochastic programming.

The paper is organized as follows. In section 2 we formally define stochastic dominance relations and the concept of the consistency of mean-risk models with these relations. Section 3 introduces dual formulations of stochastic dominance and exploits Fenchel duality to characterize dominance in terms of quantile performance functions. In section 4 we consider several risk measures based on quantiles and tail characteristics of the distribution, and we analyze their relation to stochastic dominance. Section 5 is devoted to the analysis of mean-risk models using these risk measures. In section 6 we present stochastic LP formulations of these models. Finally, we draw some conclusions in section 7.

We use $(\Omega, \mathcal{B}, \mathbb{P})$ to denote an abstract probability space. For a random variable $X : \Omega \to \mathbb{R}$, we denote by P_X the measure induced by it on the real line. For a convex function $F : \mathbb{R} \to \bar{\mathbb{R}}$, we denote by F^* its convex conjugate (see [27]), $F^*(p) = \sup_\xi \{p\xi - F(\xi)\}$.

2. Stochastic dominance and mean-risk models. Stochastic dominance is based on an axiomatic model of risk-averse preferences [5]. It originated in the majorization theory [10, 19] for the discrete case and was later extended to general distributions [25, 8, 9, 29]. Since that time it has been widely used in economics and finance (see [3, 14] for numerous references).

In the stochastic dominance approach, random variables are compared by pointwise comparison of some performance functions constructed from their distribution functions. For a real random variable X, its first performance function is defined as the right-continuous cumulative distribution function itself:

$$F_X(\eta) = \mathbb{P}\{X \leq \eta\} \quad \text{for } \eta \in \mathbb{R}.$$

In the definition below, and elsewhere in this paper, we assume that larger outcomes are preferred to smaller.

The weak relation of the *first degree stochastic dominance* (FSD) is defined as follows (see [13, 25]):

$$X \succeq_{FSD} Y \quad \Leftrightarrow \quad F_X(\eta) \leq F_Y(\eta) \quad \text{for all } \eta \in \mathbb{R}.$$

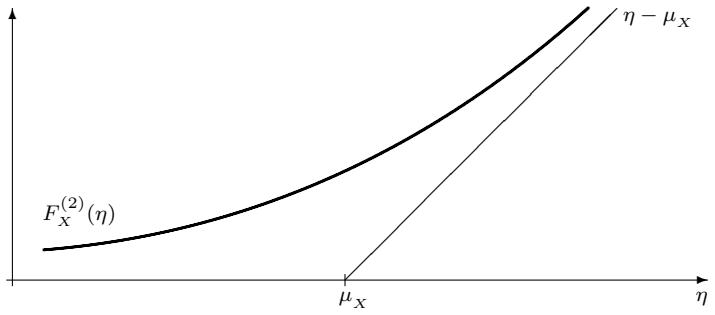


FIG. 2.1. *The O-R diagram.*

The second performance function $F_X^{(2)} : \mathbb{R} \to \mathbb{R}_+$ is given by areas below the distribution function F_X,

$$F_X^{(2)}(\eta) = \int_{-\infty}^{\eta} F_X(\xi)\, d\xi \quad \text{for } \eta \in \mathbb{R}, \tag{2.1}$$

and defines the weak relation of the *second degree stochastic dominance* (SSD):

$$X \succeq_{SSD} Y \quad \Leftrightarrow \quad F_X^{(2)}(\eta) \le F_Y^{(2)}(\eta) \quad \text{for all } \eta \in \mathbb{R} \tag{2.2}$$

(see [8, 9]). The corresponding strict dominance relations $\succ_{FSD}$ and $\succ_{SSD}$ are defined by the standard rule

$$X \succ Y \quad \Leftrightarrow \quad X \succeq Y \quad \text{and} \quad Y \not\succeq X. \tag{2.3}$$

Thus, we say that X *dominates* Y *under the FSD rules* ($X \succ_{FSD} Y$) if $F_X(\eta) \le F_Y(\eta)$ for all $\eta \in \mathbb{R}$, where at least one strict inequality holds. Similarly, we say that X *dominates* Y *under the SSD rules* ($X \succ_{SSD} Y$) if $F_X^{(2)}(\eta) \le F_Y^{(2)}(\eta)$ for all $\eta \in \mathbb{R}$, with at least one inequality strict.

Stochastic dominance relations are of crucial importance for decision theory. It is known that $X \succeq_{FSD} Y$ if and only if $\mathbb{E}U(X) \ge \mathbb{E}U(Y)$ for any nondecreasing function $U(\cdot)$ for which these expected values are finite. Also, $X \succeq_{SSD} Y$ if and only if $\mathbb{E}U(X) \ge \mathbb{E}U(Y)$ for every nondecreasing and concave $U(\cdot)$ for which these expected values are finite (see, e.g., [14]).

For a set Q of random variables, a variable $X \in Q$ is called *SSD-efficient* (or *FSD-efficient*) in Q if there is no $Y \in Q$ such that $Y \succ_{SSD} X$ (or $Y \succ_{FSD} X$).

The SSD relation is crucial for decision making under risk. As mentioned above, if $X \succ_{SSD} Y$, then X is preferred to Y within all risk-averse preference models that prefer larger outcomes. The function $F_X^{(2)}$ can also be expressed as the expected shortfall (see [22]): for each target value η we have

$$\begin{aligned} F_X^{(2)}(\eta) &= \int_{-\infty}^{\eta} (\eta - \xi)\, P_X(d\xi) \\ &= \mathbb{E}\{\max(\eta - X, 0)\} = \mathbb{P}\{X \le \eta\}\, \mathbb{E}\{\eta - X | X \le \eta\}. \end{aligned} \tag{2.4}$$

The function $F_X^{(2)}$ is continuous, convex, nonnegative, and nondecreasing. Its graph, referred to as the Outcome-Risk (O-R) diagram and illustrated in Figure 2.1, has two asymptotes which intersect at the point $(\mu_X, 0)$: the horizontal axis and the line $\eta - \mu_X$. In the case of a deterministic outcome ($X = \mu_X$), the graph of

$F_X^{(2)}$ coincides with the asymptotes, whereas any uncertain outcome with the same expected value μ_X yields a graph above (precisely, not below) the asymptotes. Hence, the space between the curve $(\eta, F_X^{(2)}(\eta))$, $\eta \in \mathbb{R}$, and its asymptotes represents the dispersion (and thereby the riskiness) of X in comparison to the deterministic outcome of μ_X. It is referred to as the *primal dispersion space*.

It is convenient to introduce also the vertical distance to the right asymptote,

$$\overline{F}_X^{(2)}(\eta) = F_X^{(2)}(\eta) - (\eta - \mu_X), \tag{2.5}$$

which can be rewritten as

$$\begin{aligned} \overline{F}_X^{(2)}(\eta) &= \int_\eta^\infty (\xi - \eta)\ P_X(d\xi) \\ &= \mathbb{E}\{\max(X - \eta, 0)\} = \mathbb{P}\{X \geq \eta\}\, \mathbb{E}\{X - \eta | X \geq \eta\}, \end{aligned} \tag{2.6}$$

thus expressing the expected surplus for each target outcome η (see [22]). The vertical diameter of the primal dispersion space at a point η is given as

$$d_X(\eta) = \min(F_X^{(2)}(\eta),\ \overline{F}_X^{(2)}(\eta)). \tag{2.7}$$

While SSD is a sound theoretical concept, its application to real world decision problems is difficult, because it requires a pairwise comparison of all possible outcome distributions. We would prefer to use simple mean-risk models and deduce from them whether a particular outcome distribution is dominated or not.

In general, considering a mean-risk model with the risk of a random outcome X measured by some nonnegative functional r_X, we can introduce the following definition.

DEFINITION 2.1. *We say that the mean-risk model* (μ_X, r_X) *is* consistent *with SSD if the following relation holds:*

$$X \succeq_{SSD} Y \quad \Rightarrow \quad \mu_X \geq \mu_Y \text{ and } r_X \leq r_Y.$$

It is known that the first inequality on the right-hand side is true: $X \succeq_{SSD} Y \Rightarrow \mu_X \geq \mu_Y$ (see [14]). The inequality for the risk term, though, is not true for some popular risk measures, like the variance or absolute deviation.

Directly from (2.4) we see that the mean-risk model with the risk functional defined as the expected shortfall below some fixed target t,

$$r_X^t = \mathbb{E}\{\max(t - X, 0)\},$$

is consistent with the SSD. Integrating the inequality $r_X^t \leq r_Y^t$ with respect to some probability measure P_T, we conclude that the expected shortfall from a *random* target T distributed according to P_T,

$$r_X = \int \mathbb{E}\{\max(t - X, 0)\}\ P_T(dt) = \mathbb{E}\{\max(T - X, 0)\}, \tag{2.8}$$

is consistent with the SSD.

While the use of consistent mean-risk models is quite straightforward, there are some reasonable risk measures which do not enjoy the consistency property of Definition 2.1. Therefore, following [23], we relax it by considering a scalarization of the

partial order in the (μ_X, r_X) space. This will allow us to derive new necessary conditions of dominance, which will make searching for an SSD-efficient solution a more tractable task.

DEFINITION 2.2. *We say that the mean-risk model* (μ_X, r_X) *is* α-consistent *with SSD, where* $\alpha > 0$*, if the following relation is true:*

$$X \succeq_{SSD} Y \quad \Rightarrow \quad \mu_X - \alpha r_X \geq \mu_Y - \alpha r_Y .$$

It is clear that α-consistency implies λ-consistency for all $0 \leq \lambda \leq \alpha$.

The concept of α-consistency turned out to be fruitful. In [22] we have proved that the mean-risk model in which the risk is defined as the *absolute semideviation*,

$$\bar{\delta}_X = r_X^{\mu_X} = \mathbb{E}\{\max(\mu_X - X, 0)\} = \int_{-\infty}^{\mu_X} (\mu_X - \xi)\ P_X(d\xi), \tag{2.9}$$

is 1-consistent with SSD. An identical result (under the condition of finite second moments) has been obtained in [22] for the *standard semideviation*,

$$\bar{\sigma}_X = \left(\mathbb{E}\left\{(\max(\mu_X - X, 0))^2\right\}\right)^{1/2} = \left(\int_{-\infty}^{\mu_X} (\mu_X - \xi)^2\ P_X(d\xi)\right)^{1/2}. \tag{2.10}$$

These results have been further extended in [23] to central semideviations of higher orders and stochastic dominance relations of higher degrees.

Remark 1. In [2] a class of *coherent risk measures* has been defined by means of several axioms. In our terms, these measures correspond to composite objectives of the form $\rho(X) = -\mu_X + \alpha r_X$ (note the opposite scalarization via the sign change), where $\alpha > 0$. The axioms are translation invariance, positive homogeneity, subadditivity, "monotonicity" ($X \geq Y$ a.s. $\Rightarrow \rho(X) \leq \rho(Y)$), and "relevance" ($X \leq 0, X \neq 0 \Rightarrow \rho(X) < 0$).

Both $\bar{\delta}_X$ and $\bar{\sigma}_X$, as seminorms in $\mathcal{L}_1$ and $\mathcal{L}_2$, are convex and positively homogeneous. Therefore the composite objectives $-\mu_X + \alpha\bar{\delta}_X$ and $-\mu_X + \alpha\bar{\sigma}_X$ do satisfy the first three axioms (contrary to the statement in [2, Rem. 2.10]). For $\alpha \in (0, 1]$, owing to the consistency with stochastic dominance in the sense of Definition 2.2, they also satisfy monotonicity and relevance, because $X \geq Y$ a.s. $\Rightarrow X \succeq_{SSD} Y$.

Our objective is to analyze risk measures using the quantiles of the distribution of X which are consistent with stochastic dominance.

3. Quantile dominance and the Lorenz curve. Let us consider the *quantile model* of stochastic dominance [15]. The first quantile function $F_X^{(-1)} : (0, 1] \to \overline{\mathbb{R}}$, corresponding to a real random variable X, is defined as the left-continuous inverse of the cumulative distribution function F_X (see [6]):

$$F_X^{(-1)}(p) = \inf\ \{\eta : F_X(\eta) \geq p\} \quad \text{for } 0 < p \leq 1.$$

Given $p \in [0, 1]$, the number $q = q_X(p)$ is called a *p-quantile* of the random variable X if

$$\mathbb{P}\{X < q\} \leq p \leq \mathbb{P}\{X \leq q\}.$$

For $p \in (0, 1)$ the set of such p-quantiles is a closed interval, and $F_X^{(-1)}(p)$ represents its left end [4].

Directly from the definition of FSD we see that

$$X \succeq_{FSD} Y \quad \Leftrightarrow \quad F_X^{(-1)}(p) \geq F_Y^{(-1)}(p) \quad \text{for all } 0 < p \leq 1. \tag{3.1}$$

Thus, the function $F^{(-1)}$ can be considered as a continuum-dimensional safety measure (negative of a risk measure) within the FSD; using any specific (left) p-quantile as a scalar safety measure is consistent with the FSD. It is not, however, consistent with the SSD, because it may happen that $X \succeq_{SSD} Y$ but $F_X^{(-1)}(p) < F_Y^{(-1)}(p)$ for some p.

Remark 2. *Value-at-risk* (VaR), defined as the maximum loss at a specified confidence level p, is a widely used quantile risk measure [26]. It corresponds to the right p-quantile of the random variable X representing gains [2], whereas our dual stochastic dominance model uses the left p-quantile. Nevertheless, the FSD consistency results can be also shown for the right quantile $q_X^r(p) = \sup\{\eta : F_X(\eta) \leq p\}$ (where $p \in [0, 1)$), thus justifying the VaR measures.

To obtain quantile measures consistent with the SSD, we introduce the *second quantile function* $F_X^{(-2)} : \mathbb{R} \to \bar{\mathbb{R}}$, defined as

$$F_X^{(-2)}(p) = \int_0^p F_X^{(-1)}(\alpha) d\alpha \quad \text{for} \quad 0 < p \leq 1, \tag{3.2}$$

$F_X^{(-2)}(0) = 0$. For completeness, we also set $F_X^{(-2)}(p) = +\infty$ for $p \notin [0, 1]$.

Similarly to $F_X^{(2)}$, the function $F_X^{(-2)}$ is well defined for any random variable X satisfying the condition $\mathbb{E}\,|X| < \infty$. By construction, it is convex. The graph of $F_X^{(-2)}$ is called the *absolute Lorenz curve (ALC) diagram*.

Remark 3. The Lorenz curves are used for inequality ordering [1, 6, 20] of positive random variables, relative to their (positive) expectations. Such a Lorenz curve, $L_X(p) = F_X^{(-2)}(p)/\mu_X$, is convex and increasing. The ALCs, though, are not monotone when negative outcomes occur.

Directly from (2.4), using the right-continuity of $F(\cdot)$, we obtain

$$\partial F_X^{(2)}(\eta) = [\mathbb{P}\{X < \eta\}, \mathbb{P}\{X \leq \eta\}]. \tag{3.3}$$

This allows us to develop a Fenchel duality relation between the second quantile function $F_X^{(-2)}$ and the second performance function $F_X^{(2)}$.

THEOREM 3.1. *For every random variable X with $\mathbb{E}\,|X| < \infty$ we have*

(i) $F_X^{(-2)} = [F_X^{(2)}]^*$ *and*

(ii) $F_X^{(2)} = [F_X^{(-2)}]^*$.

Proof. By the definition of the conjugate function, for every $p \in [0, 1]$,

$$[F_X^{(2)}]^*(p) = \sup_\eta \{\eta p - F_X^{(2)}(\eta)\}. \tag{3.4}$$

From (2.4) it is evident that $[F_X^{(2)}]^*(0) = 0$ and $[F_X^{(2)}]^*(1) = \mu_X$. For $p \in (0, 1)$ the supremum in (3.4) is attained at any η for which $p \in \partial F_X^{(2)}(\eta)$. By (3.3), η is a p-quantile of X, and we can choose $\eta = F_X^{(-1)}(p)$. Therefore, by [27, Thm. 23.5(iv)],

$$F_X^{(-1)}(p) \in \partial [F_X^{(2)}]^*(p).$$

This yields the representation

$$\left[F_X^{(2)}\right]^*(p) = \int_0^p F_X^{(-1)}(\alpha)\, d\alpha \quad \text{for} \quad p \in (0, 1].$$

If $p = 0$, then (3.4) yields 0, and for $p \notin [0,1]$ we obtain $+\infty$, as can be seen from Figure 2.1. This proves (i). Assertion (ii) is the consequence of the closedness of $F_X^{(2)}$ and [27, Thm. 12.2]. □

While the above result can also be obtained from the Young inequality ([35] and later generalizations), we hope that connections to convex analysis may prove fruitful.

It follows from Theorem 3.1 that we may fully characterize the SSD relation by using the conjugate function $F_X^{(-2)}$, similarly to the relation (3.1) for FSD.

THEOREM 3.2. $X \succeq_{SSD} Y \Leftrightarrow F_X^{(-2)}(p) \geq F_Y^{(-2)}(p)$ *for all* $0 \leq p \leq 1$.

Therefore, the properties of $F^{(-2)}$ are of profound importance for stochastic dominance relations.

COROLLARY 3.3. *The following statements are equivalent:*

(i) η *is a p-quantile of* X*;*
(ii) $\sup_\xi(\xi p - F_X^{(2)}(\xi))$ *is attained at* η*;*
(iii) $\sup_\alpha(\eta\alpha - F_X^{(-2)}(\alpha))$ *is attained at* p*;*
(iv) $F_X^{(-2)}(p) + F_X^{(2)}(\eta) = p\eta$.

Proof. Directly from definitions (2.4) and (3.2), assertion (i) is equivalent to

(v) $p \in \partial F_X^{(2)}(\eta)$ and
(vi) $\eta \in \partial F_X^{(-2)}(p)$.

The equivalence of (ii)–(vi) follows from Theorem 3.1 and [27, Thm. 23.5]. □

We can now provide another representation of the second quantile function. Let $p \in (0,1)$, and suppose that η is such that $\mathbb{P}\{X \leq \eta\} = p$. Then by Corollary 3.3(iv) and (2.4),

$$\begin{aligned} F_X^{(-2)}(p) &= p\eta - F_X^{(2)}(\eta) \\ &= p\eta + p\,\mathbb{E}\{X - \eta | X \leq \eta\} = p\,\mathbb{E}\{X | X \leq \eta\}. \end{aligned} \tag{3.5}$$

The last relation facilitates the understanding of the nature of the second quantile function, but cannot serve as a definition because η such that $\mathbb{P}\{X \leq \eta\} = p$ need not exist; (3.2) and Theorem 3.1(i) are precise descriptions.

Graphical interpretation provides an additional insight into the properties of the second quantile function. For any uncertain outcome X, its ALC $F_X^{(-2)}$ is a continuous convex curve connecting points $(0,0)$ and $(1,\mu_X)$, whereas a deterministic outcome with the same expected value μ_X corresponds to the chord connecting these points. Hence, the space between the curve $(p, F_X^{(-2)}(p))$, $0 \leq p \leq 1$, and its chord is related to the riskiness of X in comparison to the deterministic outcome of μ_X (Figure 3.1). We shall call it the *dual dispersion space*.

Both the size and the shape of the dual dispersion space are important for complete description of the riskiness of X. We shall use its size parameters as summary characteristics of riskiness.

Let us start from the vertical diameter of the dual dispersion space, defined as

$$h_X(p) = \mu_X p - F_X^{(-2)}(p). \tag{3.6}$$

LEMMA 3.4. *For every* $p \in (0,1)$

$$h_X(p) = \min_{\xi \in \mathbb{R}} \mathbb{E}\{\max(p(X-\xi), (1-p)(\xi - X))\}, \tag{3.7}$$

and the minimum in the expression above is attained at any p-quantile.

Proof. By Theorem 3.1(i),

$$h_X(p) = \inf_\xi((\mu_X - \xi)p + F_X^{(2)}(\xi)).$$

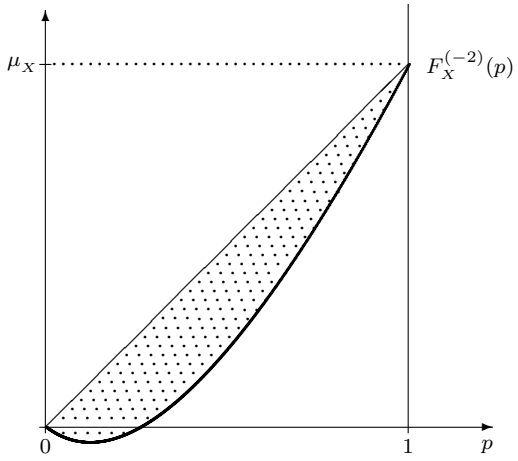


FIG. 3.1. *The ALC and the dual dispersion space.*

Subdifferentiating with respect to ξ and using (3.3), we see that the infimum is attained at any p-quantile. From (2.5) we obtain

$$h_X(p) = \min_{\xi} \left(p\overline{F}_X^{(2)}(\xi) + (1-p)F_X^{(2)}(\xi) \right).$$

With a view to (2.4) and (2.6),

$$h_X(p) = \min_{\xi} (p\,\mathbb{E}\{\max(0, X-\xi)\} + (1-p)\,\mathbb{E}\{\max(0, \xi - X)\}),$$

which completes the proof. □

The above result reveals a close relation between the vertical dimension of the dual dispersion space and the absolute deviation from the median,

$$\Delta_X = \mathbb{E}\left| X - F_X^{(-1)}\left(\frac{1}{2}\right)\right|.$$

COROLLARY 3.5. $h_X(\frac{1}{2}) = \frac{1}{2}\Delta_X$.

The maximum vertical diameter of the dual dispersion space (which exists by compactness and continuity) turns out to be the absolute semideviation of X.

LEMMA 3.6. $\max_{p\in[0,1]} h_X(p) = \bar{\delta}_X$, *and the maximum is attained at any* p_X *for which* $\mathbb{P}\{X < \mu_X\} \le p_X \le \mathbb{P}\{X \le \mu_X\}$.

Proof. By Theorem 3.1(ii),

$$\max_{p\in[0,1]} h_X(p) = \max_{p\in[0,1]} (\mu_X p - F_X^{(-2)}(p)) = F_X^{(2)}(\mu_X),$$

and the first assertion follows from (2.4) and (2.9). Now by Corollary 3.3, μ_X is a p_X-quantile. □

If the distribution is symmetric, then $p_X = 1/2$ is a maximizer, and we have $\max_{p\in[0,1]} h_X(p) = h_X(\frac{1}{2}) = \frac{1}{2}\Delta_X = \bar{\delta}_X$.

It is known that the doubled area of the dual dispersion space,

$$\Gamma_X = 2\int_0^1 (\mu_X p - F_X^{(-2)}(p))\; dp, \tag{3.8}$$

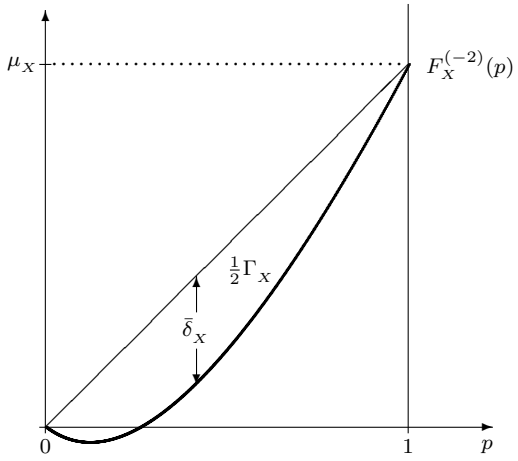


FIG. 3.2. *The ALC and risk measures.*

is equal to the Gini mean difference (see [20]):

$$\Gamma_X = \frac{1}{2} \iint |\eta - \xi| \; P_X(d\xi) \; P_X(d\eta). \tag{3.9}$$

The Gini mean difference (3.9) may be also expressed as the integral of $F_X^{(2)}$ with respect to the probability measure P_X:

$$\Gamma_X = \iint\limits_{\xi \le \eta} (\eta - \xi) \; P_X(d\xi) \; P_X(d\eta) = \int \mathbb{E}\{\max(\eta - X, 0)\} \; P_X(d\eta).$$

Thus, similar to (2.8), it represents the expected shortfall from a random target distributed according to P_X, but this distribution is a function of X. Therefore, the corresponding SSD-consistency results (cf. (2.8)) cannot be applied directly to the Gini mean difference.

Both Γ and $\bar{\delta}$ are well-defined size characteristics of the dual dispersion space (Figure 3.2). However, the absolute semideviation is a rather rough measure compared to the Gini mean difference. Note that $\bar{\delta}_X/2$ may be also interpreted in the ALC diagram as the area of the triangle given by vertices: $(0,0)$, $(1,\mu_X)$, and $(p_X, F_X^{(-2)}(p_X))$, where $\mathbb{P}\{X < \mu_X\} \le p_X \le \mathbb{P}\{X \le \mu_X\}$ (see Lemma 3.6). In fact, $\bar{\delta}_X$ is the Gini mean difference of a two-point distribution approximating X in such a way that μ_X and δ_X remain unchanged.

Dual risk characteristics can also be presented in the (primal) O-R diagram (Figure 3.3). Recall that $F^{(-2)}$ is the conjugate function of $F^{(2)}$, and therefore $F^{(-2)}$ describes the affine functions majorized by $F^{(2)}$ [27]. For any $p \in (0.1)$, the line with slope p supports the graph of $F^{(2)}$ at every p-quantile (Corollary 3.3(i),(ii)). It is given analytically as

$$S_X^p(\eta) = p(\eta - q_X(p)) + F_X^{(2)}(q_X(p)),$$

where $q_X(p)$ denotes a p-quantile of X.

From Corollary 3.3(iv) it follows that $F_X^{(-2)}(p) = -S_X^p(0)$, and thus the value of the ALC is given by the intersection of the tangent line S_X^p with the vertical (risk)

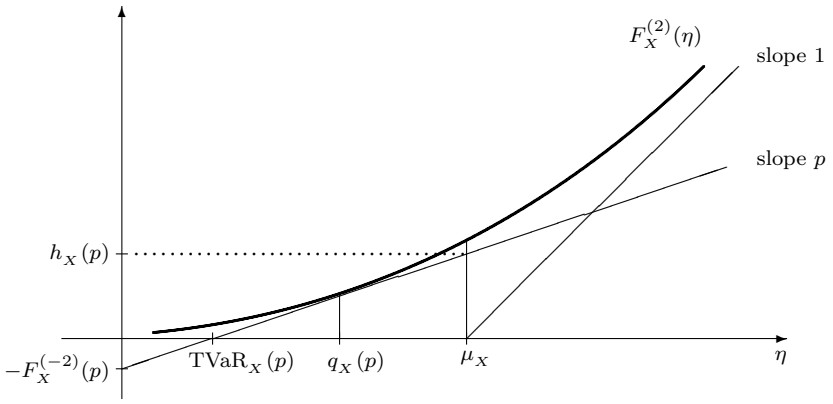


FIG. 3.3. *Dual quantities in the O-R diagram.*

axis. For any $p \in (0,1)$, the tangent line intersects the outcome axis at the point $\eta = F_X^{(-2)}(p)/p = \mu_X - h_X(p)/p$ (see (3.6)). In Figure 3.3 this point is marked as $\mathrm{TVaR}_X(p)$ due to its interpretation discussed in the next section.

Figure 3.3 also provides an interesting interpretation of Lemma 3.6. By elementary geometry, the tangent line S_X^p intersects the vertical line at $\eta = \mu_X$ at the value $S_X^p(\mu_X) = h_X(p)$, thus defining the vertical diameter of the dual dispersion space at p. This justifies $\bar{\delta}_X = F^{(2)}(\mu_X)$ as the *maximum* vertical diameter.

4. Dual risk measures. From the ALC diagram one can easily derive the following, commonly known, necessary condition for the SSD relation (cf. [14]):

$$X \succeq_{SSD} Y \quad \Rightarrow \quad \mu_X \geq \mu_Y . \tag{4.1}$$

But we can get much more.

Consider two random variables X and Y such that $X \succeq_{SSD} Y$ in the common ALC diagram (Figure 4.1). Since $\bar{\delta}_Y$ represents the maximal vertical diameter of the

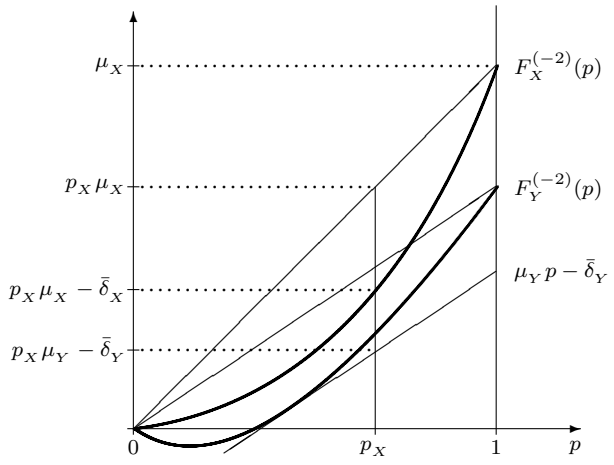


FIG. 4.1. $X \succeq_{SSD} Y \;\Rightarrow\; p_X\mu_X - \bar{\delta}_X \geq p_X\mu_Y - \bar{\delta}_Y$, *where* $p_X = \mathbb{P}\{X < \mu_X\} < 1$.

dual dispersion space for the variable Y, its ALC $F_Y^{(-2)}(p)$ is bounded from below by the straight line $\mu_Y p - \bar{\delta}_Y$. At the point $p_X = \mathbb{P}\{X < \mu_X\}$ at which $h_X(p_X) = \bar{\delta}_X$ (cf. Lemma 3.6), one gets

$$\mu_X p_X - \bar{\delta}_X = F_X^{(-2)}(p_X) \geq F_Y^{(-2)}(p_X) \geq \mu_Y p_X - \bar{\delta}_Y .$$

This simple analysis of the ALC diagram allows us to derive the following necessary condition for the SSD.

PROPOSITION 4.1. *If $X \succeq_{SSD} Y$, then $\mu_X \geq \mu_Y$ and $\mu_X - \bar{\delta}_X \geq \mu_Y - \bar{\delta}_Y$, where the second inequality is strict whenever $\mu_X > \mu_Y$.*

Proposition 4.1 was first shown in [22] with the use of an O-R diagram. Here, by placing the considerations within the (dual) ALC diagram, we make it transparent that the result is based on the comparison of the ALCs at only one point, p_X. For symmetric random variables we have $p_X \leq 1/2$, and the coefficient in front of $\bar{\delta}$ in Proposition 4.1 can be increased to 2.

The main application of the ALC diagram, though, is the analysis of risk and safety measures using quantiles of the distribution of the random outcome.

Tail VaR. The relation in Theorem 3.2 can be rewritten in the form

$$X \succeq_{SSD} Y \quad \Leftrightarrow \quad F_X^{(-2)}(p)/p \geq F_Y^{(-2)}(p)/p \quad \text{for all } 0 < p \leq 1, \tag{4.2}$$

thus justifying the safety measure

$$\mathrm{TVaR}_X(p) = F_X^{(-2)}(p)/p. \tag{4.3}$$

From Theorem 3.2 we immediately obtain the following observation.

PROPOSITION 4.2. *The mean-risk model $(\mu_X, -\mathrm{TVaR}_X)$ is consistent with the SSD relation.*

In light of (3.5), the quantity $\mathrm{TVaR}_X(p)$ may be interpreted as the expected (or tail) VaR measure (see [2, Def. 5.1] and [28]):

$$\mathrm{TVaR}_X(F_X(\eta)) = \mathbb{E}\{X | X \leq \eta\}.$$

By the convexity of $F^{(-2)}$, the function $\mathrm{TVaR}_X : (0,1] \to \mathbb{R}$ is nondecreasing, continuous, and $\mathrm{TVaR}_X(1) = \mu_X$. In the case of a lower bounded random variable, the value of $\mathrm{TVaR}_X(p)$ tends to the infimum of the outcomes when $p \to 0_+$. Hence, the max-min selection rule of [34] is a limiting case of the $(\mu_X, -\mathrm{TVaR}_X)$ model.

It follows from Lemma 3.4 that for every $p \in (0,1)$ the corresponding value $\mathrm{TVaR}_X(p)$ can be computed as

$$\mathrm{TVaR}_X(p) = \mathbb{E}\{X\} - \min_{\xi \in \mathbb{R}} \mathbb{E}\left\{\max\left(X - \xi, \frac{1-p}{p}(\xi - X)\right)\right\}. \tag{4.4}$$

This formula may be transformed into

$$\mathrm{TVaR}_X(p) = \max_{\xi \in \mathbb{R}} \left(\xi - \frac{1}{p}\mathbb{E}\{\max(0, \xi - X)\}\right), \tag{4.5}$$

which corresponds to the direct representation of $F^{(-2)}$ as the conjugate function to $F^{(2)}$ (c.f. (3.4)). By Corollary 3.3, the maximum above is attained at any p-quantile.

Interestingly, (4.5) also appears in [28] in so-called *conditional VaR* models; our analysis puts them into the context of dual stochastic dominance. An alternative proof of the consistency of conditional VaR with SSD has been given in [24].

Mean absolute deviation from a quantile. Proposition 4.2 allows us to identify an interesting α-consistent risk measure, following from the dual characterization of the SSD. Recalling the vertical diameter $h_X(p)$ of the dual dispersion space, we have the following result.

PROPOSITION 4.3. *For any $p \in (0,1)$, the mean-risk model $(\mu_X, h_X(p)/p)$ is 1-consistent with the SSD relation.*

Proof. By Proposition 4.2 and (3.6) we have

$$X \succeq_{SSD} Y \Rightarrow \mu_X \geq \mu_Y \quad \text{and} \quad \mu_X - h_X(p)/p \geq \mu_Y - h_Y(p)/p,$$

as required. □

Because of Lemma 3.4, we may interpret the risk measure $h_X(p)/p$ as the weighted mean absolute deviation from the p-quantile.

For $p = 1/2$, recalling Corollary 3.5, we obtain the following observation (illustrated graphically in Figure 4.2).

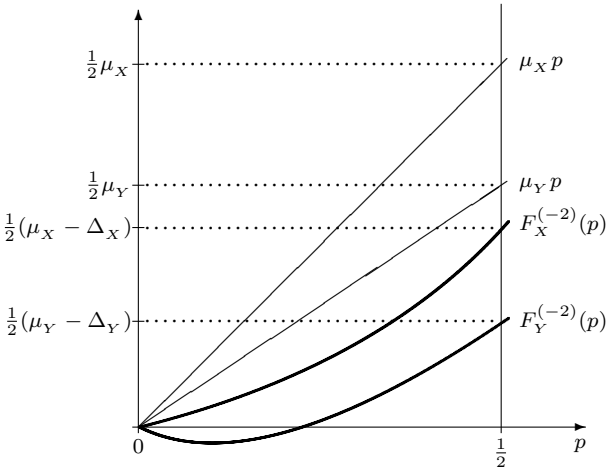


FIG. 4.2. *Median case:* $X \succeq_{SSD} Y \Rightarrow \frac{1}{2}(\mu_X - \Delta_X) \geq \frac{1}{2}(\mu_Y - \Delta_Y)$.

COROLLARY 4.4. *The mean-risk model (μ_X, Δ_X) is 1-consistent with the SSD relation.*

Comparing this to Proposition 4.1, we see that we are able to cover both the general and the symmetric case with a higher weight put on the risk term. Indeed, in the symmetric case we have $\Delta_X = 2\bar{\delta}_X$.

Tail Gini mean difference. Let us now pass to risk measures based on area characteristics of the dual dispersion space. Consider two random variables X and Y such that $X \succeq_{SSD} Y$ in the common ALC diagram (Figure 4.3). If $X \succeq_{SSD} Y$, then, due to Theorem 3.2, $F_X^{(-2)}$ is bounded from below by $F_Y^{(-2)}$, and $\mu_X \geq \mu_Y$ from (4.1). Thus the area of the dual dispersion space for X is (upper) bounded by the area of the dual dispersion space for Y plus the area of the triangle between the chords (with vertices: $(0,0)$, $(1,\mu_X)$, and $(1,\mu_Y)$). Hence, $\frac{1}{2}\Gamma_X \leq \frac{1}{2}\Gamma_Y + \frac{1}{2}(\mu_X - \mu_Y)$, and, due to the continuity of the Lorenz curves, this inequality becomes strict whenever $X \succ_{SSD} Y$. This allows us to derive the following necessary conditions for the SSD.

PROPOSITION 4.5. *For integrable random variables X and Y the following im-*

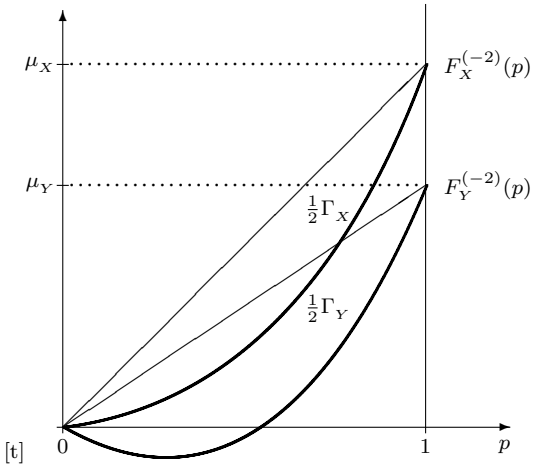


FIG. 4.3. $X \succeq_{SSD} Y \;\Rightarrow\; \frac{1}{2}\Gamma_X \le \frac{1}{2}\Gamma_Y + \frac{1}{2}(\mu_X - \mu_Y)$.

plications hold:

$$X \succeq_{SSD} Y \quad \Rightarrow \quad \mu_X - \Gamma_X \ge \mu_Y - \Gamma_Y, \tag{4.6}$$

$$X \succ_{SSD} Y \quad \Rightarrow \quad \mu_X - \Gamma_X > \mu_Y - \Gamma_Y. \tag{4.7}$$

Condition (4.6) was first shown by Yitzhaki [33] for bounded distributions.

Similarly, for $p \in (0,1]$ one may consider the *tail Gini* measure:

$$G_X(p) = \frac{2}{p^2}\int_0^p (\mu_X \alpha - F_X^{(-2)}(\alpha))d\alpha. \tag{4.8}$$

The next result is an obvious extension of Proposition 4.5.

PROPOSITION 4.6. *For every* $p \in (0,1]$,

$$X \succeq_{SSD} Y \quad \Rightarrow \quad \mu_X - G_X(p) \ge \mu_Y - G_Y(p). \tag{4.9}$$

In other words, the mean-risk model $(\mu_X, G_X(p))$ is 1-consistent with the SSD.

By convexity, $G_X(p) \ge h_X(p)/p$ for all $p \in (0,1]$, so Proposition 4.6 is stronger than Proposition 4.3.

The coefficient 1 in front of $G_X(p)$ (and $G_Y(p)$) cannot be increased for general distributions, but it can be doubled in the case of *symmetric* random variables (and $p = 1$). Indeed, for a symmetric random variable X one has $h_X(p) = h_X(1-p)$, and thus $G_X(\frac{1}{2}) = 2\Gamma_X$, which leads to the following result.

PROPOSITION 4.7. *For symmetric random variables* X *and* Y *the following implications hold:*

$$X \succeq_{SSD} Y \quad \Rightarrow \quad \mu_X - 2\Gamma_X \ge \mu_Y - 2\Gamma_Y,$$
$$X \succ_{SSD} Y \quad \Rightarrow \quad \mu_X - 2\Gamma_X > \mu_Y - 2\Gamma_Y.$$

5. Mean-risk models with dual risk measures. Given a certain set Q of integrable random variables X, let us analyze in more detail the mean-risk optimization problems of form

$$\max_{X \in Q} (\mu_X - \lambda r_X), \tag{5.1}$$

with $\lambda > 0$ and with risk functional r_X defined as one of our dual (quantile) measures. We assume that the set Q is convex, closed, and bounded in $\mathcal{L}_q$ for some $q > 1$.

The first issue that needs to be clarified is the convexity of problem (5.1). This will help to establish the existence of solutions and to formulate computationally tractable models.

LEMMA 5.1. *For every $p \in [0,1]$ the functional $X \to h_X(p)$ given by (3.6) is convex and positively homogeneous on $\mathcal{L}_1$.*

Proof. Let $\beta \in (0,1)$, $X, Y \in Q$, and let m_X and m_Y be the p-quantiles of X and Y. By Lemma 3.4,

$$h_{\beta X+(1-\beta)Y}(p) = \min_t \mathbb{E} \max\{p(\beta X + (1-\beta)Y - t), (1-p)(t - \beta X - (1-\beta)Y)\}$$
$$\leq \mathbb{E} \max\{p(\beta(X - m_X) + (1-\beta)(Y - m_Y)), (1-p)(\beta(m_X - X) + (1-\beta)(m_Y - Y))\}.$$

Using the inequality $\max(a+b, c+d) \leq \max(a,c) + \max(b,d)$ and Lemma 3.4 again, we obtain

$$\begin{aligned} h_{\beta X+(1-\beta)Y}(p) &\leq \beta \mathbb{E} \max\{p(X - m_X), (1-p)(m_X - X)\} \\ &\quad + (1-\beta)\mathbb{E} \max\{p(Y - m_Y), (1-p)(m_Y - Y)\} \\ &= \beta h_X(p) + (1-\beta) h_Y(p), \end{aligned}$$

because m_X and m_Y are p-quantiles. This proves the convexity. The positive homogeneity follows directly from (3.7). □

For the tail Gini mean difference used as a risk measure, we have a similar result.

LEMMA 5.2. *For every $p \in (0,1]$ the functional $X \to G_X(p)$ given by (4.8) is convex and positively homogeneous on $\mathcal{L}_1$.*

Proof. We have

$$G_X(p) = \frac{2}{p^2} \int_0^p h_X(\alpha)\, d\alpha,$$

and the result follows from Lemma 5.1. □

Remark 4. Again, the composite objectives of form $\rho(X) = -\mu_X + \alpha r_X$, where $\alpha \in (0,1]$ and r_X is defined as $h_X(p)/p$ or $G_X(p)$, satisfy all axioms of the so-called *coherent risk measures* discussed in [2] (cf. Remark 1). The convexity and positive homogeneity have just been proved, the translation invariance is trivial, and the monotonicity follows from Propositions 4.3 and 4.6, respectively. Indeed, as in Remark 1, $X \geq Y$ a.s. $\Rightarrow$ $X \succeq_{SSD} Y$, and these propositions apply.

Having established convexity, we can pass now to the analysis of the SSD-efficiency of the solutions to problem (5.1). We start from the case of the Gini mean difference $\Gamma_X = G_X(1)$.

THEOREM 5.3. *Assume that the set Q is convex, bounded, and closed in $\mathcal{L}_q$ for some $q > 1$, and $r_X = \Gamma_X$. Then for every $\lambda \in (0,1]$ the set of optimal solutions of (5.1) is nonempty, and each of its elements is SSD-efficient in Q.*

Proof. Let us show that the optimal set of (5.1) is nonempty. By Lemma 5.2 the objective functional is concave. In the reflexive Banach space $\mathcal{L}_q$, the set Q is weakly compact (as convex, bounded, and closed [11, Thm. 6, p. 179]), and the functional $\mu_X - \lambda\Gamma_X$ is weakly upper semicontinuous (as concave and bounded). Therefore the set of optimal solutions of (5.1) is nonempty.

Let $X \in Q$ be an optimal solution, and suppose that X is not SSD-efficient. Then there exists $Z \in Q$ such that $Z \succ_{SSD} X$. From (4.1) and (4.7) we obtain

$$\mu_Z \geq \mu_X \quad \text{and} \quad \mu_Z - \Gamma_Z > \mu_X - \Gamma_X.$$

Adding these inequalities, multiplied by $(1-\lambda)$ and λ, respectively, we obtain the sharp ($\lambda > 0$) inequality $\mu_Z - \lambda\Gamma_Z > \mu_X - \lambda\Gamma_X$. This contradicts the maximality of $\mu_X - \lambda\Gamma_X$. □

Let us now consider the risk measure $r_X = h_X(p)/p$. Recall that, owing to (3.6) and (4.3), the objective in (5.1) can be equivalently expressed as

$$\mu_X - \lambda h_X(p)/p = (1-\lambda)\mu_X + \lambda\,\mathrm{TVaR}_X(p).$$

THEOREM 5.4. *Assume that the set Q is convex, bounded, and closed in $\mathcal{L}_q$ for some $q > 1$, and $r_X = h_X(p)/p$ with $p \in (0,1)$. Then for every $\lambda \in (0,1]$ the set Q^* of optimal solutions of* (5.1) *is nonempty, and for each $X \in Q^*$ there exists a point $X^* \in Q^*$ which is SSD-efficient in Q and with $\mu_{X^*} = \mu_X$ and $h_{X^*}(p) = h_X(p)$.*

Proof. The proof that the optimal set Q^* of (5.1) is nonempty is the same as that in Theorem 5.3. By the convexity of the set Q and the concavity of the objective functional, the set Q^* is convex, closed, and bounded.

Suppose that $X \in Q^*$ is not SSD-efficient. Then there exists $Z \in Q$ such that $Z \succ_{SSD} X$. From (4.1) and Proposition 4.3 we obtain

$$\mu_Z \geq \mu_X \quad \text{and} \quad \mu_Z - h_Z(p)/p \geq \mu_X - h_X(p)/p.$$

Adding these inequalities, multiplied by $(1-\lambda)$ and λ, respectively, we obtain

$$\mu_Z - \lambda h_Z(p)/p \geq \mu_X - \lambda h_X(p)/p.$$

Since $Z \in Q$, we must have $Z \in Q^*$ and an equality above. Thus $\mu_Z = \mu_X$ and $h_Z(p) = h_X(p)$.

Define the set $Q^*(X) = \{Z \in Q^* : \mu_Z = \mu_X\}$, and consider the problem

$$\min_{Z \in Q^*(X)} \Gamma_Z. \tag{5.2}$$

The set $Q^*(X)$ is convex, closed, and bounded, and (5.2) is equivalent to maximizing $\mu_Z - \lambda\Gamma_Z$. By Theorem 5.3, a solution X^* of (5.2) exists and is SSD-efficient in $Q^*(X)$. It is also SSD-efficient in Q, because we have proved in the preceding paragraph that it cannot be dominated by a point $Z \in Q \setminus Q^*(X)$. By construction, $\mu_{X^*} = \mu_X$ and $h_{X^*}(p) = h_X(p)$, as required. □

Let us now consider the risk measure in the form of the tail Gini mean difference. Analogously to Theorem 5.4 we obtain the following result.

THEOREM 5.5. *Assume that the set Q is convex, bounded, and closed in $\mathcal{L}_q$ for some $q > 1$, and let $r_X = G_X(p)$ with $p \in (0,1)$. Then for every $\lambda \in (0,1]$ the set Q^* of optimal solutions of* (5.1) *is nonempty, and for each $X \in Q^*$ there exists an SSD-efficient point $X^* \in Q^*$ with $\mu_{X^*} = \mu_X$ and $G_{X^*}(p) = G_X(p)$.*

Remark 5. For symmetric random variables and $p \geq 1/2$, since $h_X(p) = h_X(1-p)$, all optimal solutions are SSD-efficient, as follows from Theorem 5.3. Also, since $G_X(\frac{1}{2}) = 2\Gamma_X$, the coefficient λ in (5.1) can be chosen from $(0,2]$.

6. Stochastic programming formulations. Let us formulate a more explicit convex optimization problem which is equivalent to (5.1) with $r_X = h_X(p)/p$:

$$\max \quad \mathbb{E}\,X - \frac{\lambda}{p}\mathbb{E}\,V \tag{6.1a}$$

$$\text{subject to} \quad V(\omega) \geq p(X(\omega) - t), \qquad \text{a.s.}, \tag{6.1b}$$

$$V(\omega) \geq (1-p)(t - X(\omega)), \quad \text{a.s.}, \tag{6.1c}$$

$$X \in Q, \quad V \in \mathcal{L}_1(\Omega), \quad t \in \mathbb{R}. \tag{6.1d}$$

The next result follows from Lemma 3.4.

PROPOSITION 6.1. *Problem* (6.1) *is equivalent to problem* (5.1) *with* $r_X = h_X(p)/p$ *in the following sense:*

(i) *for every solution* $\hat{X}$ *of* (5.1), *the triple*

$$\hat{X}, \quad \hat{t} = F^{(-1)}_{\hat{X}}(p), \quad \hat{V}(\omega) = \max(p(\hat{X}(\omega) - \hat{t}), (1-p)(\hat{t} - \hat{X}(\omega)))$$

is an optimal solution of (6.1);

(ii) *for every optimal solution* $(\hat{X}, \hat{t}, \hat{V})$ *of* (6.1), $\hat{X}$ *is an optimal solution of* (5.1), $\hat{t}$ *is a p-quantile of* $\hat{X}$, *and* $\mathbb{E}\hat{V}(\omega) = h_{\hat{X}}(p)$.

In particular, if

$$Q = \left\{ \sum_{i=1}^{n} d_i X_i \,:\, (d_1, \ldots, d_n) \in D \right\}, \tag{6.2}$$

where D is a convex closed polyhedron in $\mathbb{R}^n$ and $X_1, \ldots, X_n$ are integrable random variables, we recognize a linear two-stage problem of stochastic programming. In this problem $d \in D$ and $t \in \mathbb{R}$ are first-stage variables, while V is the second-stage variable. In the case of finitely many realizations $(x_1^j, \ldots, x_n^j)$, $j = 1, \ldots, N$, of $(X_1, \ldots, X_n)$, attained with probabilities $\pi_1, \ldots, \pi_N$, we obtain the problem

$$\begin{aligned}
\max \quad & \sum_{j=1}^{N} \pi_j \left(\sum_{i=1}^{n} d_i x_i^j - \frac{\lambda}{p} v^j \right) \\
\text{subject to} \quad & v^j \geq p \left(\sum_{i=1}^{n} d_i x_i^j - t \right), \qquad j = 1, \ldots, N, \\
& v^j \geq (1-p) \left(t - \sum_{i=1}^{n} d_i x_i^j \right), \quad j = 1, \ldots, N, \\
& d \in D, \quad v \in \mathbb{R}^N, \quad t \in \mathbb{R}.
\end{aligned}$$

Representing $\sum_{i=1}^{n} d_i x_i^j - t$ as a difference of its positive part u_j and its negative part w_j and eliminating the expectation from the objective, we can transform the last problem to a simple recourse formulation:

$$\begin{aligned}
\max \quad & \left[t + \sum_{j=1}^{N} \pi_j \left((1-\lambda) u_j - \left(1 - \lambda + \frac{\lambda}{p} \right) w_j \right) \right] \\
\text{subject to} \quad & \sum_{i=1}^{n} d_i x_i^j - t = u_j - w_j, \quad j = 1, \ldots, N, \\
& d \in D, \quad u \in \mathbb{R}^N_+, \quad w \in \mathbb{R}^N_+, \quad t \in \mathbb{R}.
\end{aligned}$$

Let us now formulate a stochastic programming problem which is equivalent to (5.1) with $r_X = G_X(p)$:

$$\max \quad \mathbb{E}\, X - \frac{2\lambda}{p^2} \int_0^p \int V(\alpha, \omega)\ \mathbb{P}(d\omega)\ d\alpha \tag{6.3a}$$

$$\text{subject to} \quad V(\alpha, \omega) \geq \alpha(X(\omega) - t(\alpha)), \qquad \text{a.s. in } [0,p] \times \Omega, \tag{6.3b}$$

$$V(\alpha, \omega) \geq (1-\alpha)(t(\alpha) - X(\omega)), \quad \text{a.s. in } [0,p] \times \Omega, \tag{6.3c}$$

$$X \in Q, \quad V \in \mathcal{L}_1([0,p] \times \Omega), \quad t \in \mathcal{L}_1([0,p]). \tag{6.3d}$$

The product space $[0,p]\times\Omega$ is assumed to be equipped with the product measure of the Lebesgue measure and $\mathbb{P}$.

PROPOSITION 6.2. *Problem* (6.3) *is equivalent to problem* (5.1) *with* $r_X = G_X(p)$ *in the following sense:*

(i) *for every solution* $\hat{X}$ *of* (5.1)*, the triple*

$$\hat{X},\quad \hat{t}(\alpha)=F^{(-1)}_{\hat{X}}(\alpha),\quad \hat{V}(\alpha,\omega)=\max(\alpha(\hat{X}(\omega)-\hat{t}(\alpha)),(1-\alpha)(\hat{t}(\alpha)-\hat{X}(\omega)))$$

is an optimal solution of (6.3)*;*

(ii) *for every optimal solution* $(\hat{X},\hat{t},\hat{V})$ *of* (6.3)*,* $\hat{X}$ *is an optimal solution of* (5.1)*,* $\hat{t}(\alpha)$ *is an* α*-quantile of* $\hat{X}$ *for almost all* $\alpha\in(0,p]$*, and* $\mathbb{E}\hat{V}(\alpha,\omega)=h_{\hat{X}}(\alpha)$ *for almost all* $\alpha\in(0,p]$*.*

Proof. For $X\in Q$ the quantile $F^{(-1)}_X(\cdot)$ is integrable in $(0,p]$, so restricting t to $\mathcal{L}_1([0,p])$ is allowed. The rest of the proof follows from Lemma 3.4, as in Proposition 6.1. □

In particular, if Q is defined by (6.2) and $(X_1,\dots,X_n)$ is a discrete random vector with N *equally probable* realizations $(x_1^j,\dots,x_n^j)$, $j=1,\dots,N$, we can further simplify this problem. We notice first that $h_X(\alpha)$ is a piecewise linear concave function with break points at k/N, $k=0,\dots,N$. Thus the inequalities (6.3b)–(6.3c) need to be enforced only at the break points. Moreover, the integral in the objective of (6.3) can be calculated exactly by using the values at the break points, by the method of trapezoids.

To be more specific, let m be the smallest integer for which $m/N\ge p$, and let $\alpha_k=k/N$, $k=0,\dots,m-1$; $\alpha_m=p$. We obtain the following two-stage stochastic program:

$$\begin{aligned}
\max\quad & \sum_{j=1}^{N}\pi_j\left(\sum_{i=1}^{n}d_ix_i^j-\frac{\lambda}{p^2}\sum_{k=0}^{m}(\alpha_{k+1}-\alpha_k)(v_{k+1}^j+v_k^j)\right)\\
\text{subject to}\quad & v_k^j\ge\alpha_k\left(\sum_{i=1}^{n}d_ix_i^j-t_k\right), \qquad j=1,\dots,N,\quad k=0,\dots,m,\\
& v_k^j\ge(1-\alpha_k)\left(t_k-\sum_{i=1}^{n}d_ix_i^j\right),\quad j=1,\dots,N,\quad k=0,\dots,m,\\
& d\in D,\quad v\in\mathbb{R}^N\times\mathbb{R}^{m+1},\quad t\in\mathbb{R}^{m+1}.
\end{aligned}$$

In the above problem, v_k^j represents the value of $V(\alpha_k)$ in the jth realization, and $t_k=t(\alpha_k)$. Similarly to problem (6.1), the last problem can also be transformed to a simple recourse formulation.

If the probabilities π_j of the realizations of $(X_1,\dots,X_n)$ are *not* equal, however, the break points may depend on our decisions, and the reduction to the finite dimensional case is harder. One way around this difficulty is to repeat the outcomes (as many times as needed) to ensure this property (in the case of rational probabilities). This, however, may dramatically increase the size of the problem. Another possibility is to introduce such a grid that contains *all* possible break points, but it may be unnecessarily large. Yet another possibility is to resort to an approximation with some reasonably chosen grid α_k, $k=1,\dots,m$. This would be a relaxation because $h(\cdot)$ is a concave function.

For $p = 1$ all these complications disappear, because the alternative definition (3.9) of Γ_X has an obvious LP representation:

$$\begin{aligned}
\max \quad & \left[\sum_{j=1}^{N} \pi_j \sum_{i=1}^{n} d_i x_i^j - \lambda \sum_{j=1}^{N} \sum_{l=j+1}^{N} \pi_j \pi_l v^{jl}\right] \\
\text{subject to} \quad & v^{jl} \geq \sum_{i=1}^{n} d_i (x_i^j - x_i^l), \quad j = 1, \ldots, N, \quad l = j+1, \ldots, N, \\
& v^{jl} \geq \sum_{i=1}^{n} d_i (x_i^l - x_i^j), \quad j = 1, \ldots, N, \quad l = j+1, \ldots, N, \\
& d \in D, \quad v \in \mathbb{R}^{N(N-1)/2}.
\end{aligned}$$

This has a much larger number of variables and constraints, however.

All finite dimensional stochastic programming models of this section can be solved by specialized decomposition methods [30].

7. Conclusions. We have defined dual relations of stochastic dominance for arbitrary random variables with finite expectations. The SSD can be expressed as a relation of conjugate functions to second order performance functions.

By using the concepts and methods of convex analysis and optimization theory, we have identified several security and risk measures which can be employed in mean-risk decision models: *tail Value-at-Risk*,

$$\mathrm{TVaR}_X(p) = q_X(p) - \frac{1}{p}\mathbb{E}\{\max(0, q_X(p) - X)\},$$

where $q_X(p)$ is a p-quantile; *weighted mean deviation from a quantile*,

$$h_X(p) = \mathbb{E}\{\max(p(X - q_X(p)), (1-p)(q_X(p) - X))\};$$

and *tail Gini mean difference*,

$$G_X(p) = \frac{2}{p^2}\int_0^p h_X(\alpha)\, d\alpha.$$

We have shown that the mean-risk models using these measures—$(\mu_X, -\mathrm{TVaR}_X(p))$, $(\mu_X, h_X(p))$, and $(\mu_X, G_X(p))$—are consistent with the SSD relation (in the sense of Definition 2.1 for $\mathrm{TVaR}_X(p)$, and Definition 2.2 for the other two measures). In particular, the optimal solutions of the corresponding mean-risk models, if unique, are efficient under the SSD relation.

Finally, we have found stochastic LP formulations of these models. This opens a new area of applications of the theory and methods for stochastic programming.

REFERENCES

[1] B. C. Arnold, *Majorization and the Lorenz Order: A Brief Introduction*, Lecture Notes in Statist. 43, Springer-Verlag, Berlin, 1980.

[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, *Stochastic dominance: A research bibliography*, Management Science, 28 (1982), pp. 698–712.

[4] P. Embrechts, C. Klüppelberg, and T. Mikosch, *Modelling Extremal Events for Insurance and Finance*, Springer-Verlag, New York, 1997.
[5] P. C. Fishburn, *Decision and Value Theory*, John Wiley & Sons, New York, 1964.
[6] J. L. Gastwirth, *A general definition of the Lorenz curve*, Econometrica, 39 (1971), pp. 1037–1039.
[7] J. Gotoh and H. Konno, *Third degree stochastic dominance and mean-risk analysis*, Management Science, 46 (2000), pp. 289–301.
[8] J. Hadar and W. Russell, *Rules for ordering uncertain prospects*, American Economic Review, 59 (1969), pp. 25–34.
[9] G. Hanoch and H. Levy, *The efficiency analysis of choices involving risk*, Rev. Econom. Stud., 36 (1969), pp. 335–346.
[10] G. H. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*, Cambridge University Press, Cambridge, MA, 1934.
[11] L. V. Kantorovich and G. P. Akilov, *Functional Analysis*, Pergamon Press, Oxford, UK, 1982.
[12] H. Konno and H. Yamazaki, *Mean-absolute deviation portfolio optimization model and its application to Tokyo stock market*, Management Science, 37 (1991), pp. 519–531.
[13] E. Lehmann, *Ordered families of distributions*, Annals of Mathematical Statistics, 26 (1955), pp. 399–419.
[14] H. Levy, *Stochastic dominance and expected utility: survey and analysis*, Management Science, 38 (1992), pp. 555–593.
[15] H. Levy and Y. Kroll, *Ordering uncertain options with borrowing and lending*, J. Finance, 33 (1978), pp. 553–573.
[16] M. O. Lorenz, *Methods of measuring concentration of wealth*, J. Amer. Statist. Assoc., 9 (1905), pp. 209–219.
[17] H. M. Markowitz, *Portfolio selection*, J. Finance, 7 (1952), pp. 77–91.
[18] H. M. Markowitz, *Mean-Variance Analysis in Portfolio Choice and Capital Markets*, Blackwell, Oxford, UK, 1987.
[19] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications*, Academic Press, San Diego, 1979.
[20] P. Muliere and M. Scarsini, *A note on stochastic dominance and inequality measures*, J. Econom. Theory, 49 (1989), pp. 314–323.
[21] W. Ogryczak, *Multiple criteria linear programming model for portfolio selection*, Ann. Oper. Res., 97 (2000), pp. 143–162.
[22] W. Ogryczak and A. Ruszczyński, *From stochastic dominance to mean-risk models: Semideviations as risk measures*, Eur. J. Oper. Res., 116 (1999), pp. 33–50.
[23] W. Ogryczak and A. Ruszczyński, *On consistency of stochastic dominance and mean-semideviation models*, Math. Program., 89 (2001), pp. 217–232.
[24] G. C. Pflug, *Some remarks on the value-at-risk and the conditional value-at-risk*, in Probabilistic Constrained Optimization: Methodology and Applications, S. Uryasev, ed., Kluwer Academic Publishers, Norwell, MA, 2000, pp. 278–287.
[25] J. P. Quirk and R. Saposnik, *Admissibility and measurable utility functions*, Rev. Econom. Stud., 29 (1962), pp. 140–146.
[26] *RiskMetrics*, Technical document, J. P. Morgan, New York, 1996.
[27] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[28] R. T. Rockafellar and S. Uryasev, *Optimization of conditional value-at-risk*, J. Risk, 2 (2000), pp. 21–41.
[29] M. Rothschild and J. E. Stiglitz, *Increasing risk:* I. *A definition*, J. Econom. Theory, 2 (1969), pp. 225–243.
[30] A. Ruszczyński, *Decomposition methods in stochastic programming*, Math. Program., 79 (1997), pp. 333–353.
[31] W. F. Sharpe, *A linear programming approximation for the general portfolio analysis problem*, J. Financial and Quantitative Analysis, 6 (1971), pp. 1263–1275.
[32] G. A. Whitmore and M. C. Findlay, eds., *Stochastic Dominance: An Approach to Decision-Making Under Risk*, D. C. Heath, Lexington, MA, 1978.
[33] S. Yitzhaki, *Stochastic dominance, mean variance, and Gini's mean difference*, American Economic Review, 72 (1982), pp. 178–185.
[34] M. R. Young, *A minimax portfolio selection rule with linear programming solution*, Management Science, 44 (1998), pp. 673–683.
[35] W. H. Young, *On classes of summable functions and their Fourier series*, Proc. Roy. Soc. London, A87 (1912), pp. 235–239.

SIAM J. OPTIM.
Vol. 13, No. 1, pp. 79–96

NEW SEQUENTIAL AND PARALLEL DERIVATIVE-FREE ALGORITHMS FOR UNCONSTRAINED MINIMIZATION*

U. M. GARCÍA-PALOMARES† AND J. F. RODRÍGUEZ‡

Abstract. This paper presents sequential and parallel derivative-free algorithms for finding a local minimum of smooth and nonsmooth functions of practical interest. It is proved that, under mild assumptions, a sufficient decrease condition holds for a nonsmooth function. Based on this property, the algorithms explore a set of search directions and move to a point with a sufficiently lower functional value. If the function is strictly differentiable at its limit points, a (sub)sequence of points generated by the algorithm converges to a first-order stationary point ($\nabla f(x) = 0$). If the function is convex around its limit points, convergence (of a subsequence) to a point with nonnegative directional derivatives on a set of search directions is ensured. Preliminary numerical results on sequential algorithms show that they compare favorably with the recently introduced pattern search methods.

Key words. nonsmooth function, unconstrained minimization, derivative-free algorithm, parallel algorithms, necessary and sufficient conditions

AMS subject classifications. 49D30, 65K05

PII. S1052623400370606

1. Introduction.

We are concerned with the problem of obtaining an unconstrained *local* minimizer and a *local* minimum of a nonsmooth functional $f(\cdot) : \mathbb{R}^n \to \mathbb{R}$. More specifically, we look for the values of the variables $x^1, \ldots, x^n$ in the whole Euclidean space $\mathbb{R}^n$, where $f(x^1, \ldots, x^n)$ attains a *local* minimum value. We recall that penalization and Lagrange techniques are usually applied to transform a constrained minimization problem into an unconstrained and/or box constrained problem. Hence, the efficient solution of unconstrained problems is of broad interest.

The algorithms proposed here use only function values, but we are aware that, when first and/or second derivatives are available, Newton-related methods are highly efficient. Nonetheless, real world applications in many cases preclude the use of derivatives, because the functional values may arise either from a complex simulation package or from inaccurate sample values. Furthermore, a numerical approximation of the derivatives is not always a reliable approach. Therefore, practitioners require efficient derivative-free methods. Old algorithms, developed mainly in the late '60s and early '70s, had a strong intuitive approach and often lacked a convergence theory. (The interested reader can examine these methods in many optimization books, such as [4, 17, 44].) Other recent approaches are reported in [8, 11, 43]. The simplex method *(a simplex in $\mathbb{R}^n$ is the convex hull of $n+1$ points $x_0, \ldots, x_n$)* [27] has become, due perhaps to its simplicity and success in the solution of practical problems with a small number of variables, the most widely used and cited in the literature of unconstrained minimization. Nonetheless, it can fail on small problems, and convergence to a nonstationary point may occur [25, 39]. We are just starting to understand the properties of the simplex method [20], which has triggered active research on derivative-free meth-

*Received by the editors October 27, 2000; accepted for publication (in revised form) November 30, 2001; published electronically June 5, 2002. This research was partially supported by our respective Departments and the *Decanato de Investigación y Desarrollo* of the Universidad Simón Bolívar.
http://www.siam.org/journals/siopt/13-1/37060.html

†Universidad Simón Bolívar, Departamento Procesos y Sistemas, Apdo 89000, Caracas 1080-A, Venezuela (garciap@usb.ve).

‡Departamento Mecánica, Apdo 89000, Caracas 1080-A, Venezuela (jrodri@usb.ve).

ods in the last decade. Related *simplicial* methods with a formal convergence theory have appeared in [9, 18, 36, 39, 40, 42]. Convergence of well-known derivative-free methods and *pattern search methods* have been analyzed in [3, 21, 41]. Essentially, a pattern search method examines the function values on a *nondegenerate simplex*. An iteration starts with a new simplex which satisfies a form of descent condition for the function values. Under standard assumptions, convergence (of a subsequence) to a point satisfying the first-order necessary condition ($\nabla f(x) = 0$) of general smooth functions is ensured. A possible drawback of these methods is that function values must be obtained infinitely often at all vertices of a simplex before a new iteration starts, which implies at least n function evaluations per iteration. To circumvent this difficulty, it is desirable to establish at each iteration a decrease of the function value sufficient to guarantee convergence. An old effort in this direction is found in [12], and more recent attempts in [18, 24, 42]. Additional material can be found in [19, Chapters 6,7] and references therein.

While this paper was under review, the referees brought [3] to our attention, in which the convergence of *generalized pattern search* methods is ensured without assuming global continuity. Convergence relies on the differentiability properties of limit points. All other works on derivative-free methods cited above assume $f(\cdot) \in C^1$ and often ask for Lipschitz continuity of the gradients to ensure convergence. Therefore, the convergence theory cannot be applied in certain cases of practical interest:

(i) Some industrial problems often require the minimization of functions which arise from a complex simulation process or from sample values. Smoothness of the function cannot be guaranteed.

(ii) Common functions, like the norm function $f(x) = \|f_1(x)\|$ and the Max function $f(x) = \text{Max}(f_1(x), \ldots, f_m(x))$, may not be everywhere differentiable, even in the convex case.

(iii) Most exact penalty functions are not everywhere differentiable.

This paper has a twofold objective: (a) to define a practical necessary condition for a class of nonsmooth functions, which should be valid as well for smooth functions and readily allow (b) the implementation of converging algorithms. The rest of the paper is organized as follows. The next section states the assumptions C1–C3, needed to ensure that the algorithm is well defined, and the nonsmooth necessary condition (NSNC) (2.4). Section 3 introduces the sufficient decrease criterion (3.1)–(3.2), proposes sequential and parallel algorithms, and develops the convergence theory. It is shown that the algorithms are well defined under conditions C1–C3. Furthermore, convergence to a point x satisfying NSNC is shown if $f(\cdot)$ is strictly differentiable at x or convex in a neighborhood of x. Section 4 presents extensions to smooth functions and the box constrained problem. It also analyzes useful features that improve the algorithm notably. Section 5 shows preliminary numerical results with examples from the CUTE collection and the Rosenbrock function with two, three, five, or ten variables. The number of function evaluations of our sequential algorithms are in general lower than those needed by the pattern search methods (PSM) with the same termination criteria. Finally, we state our conclusions and final remarks in section 6. It is pointed out that no PSM ensures convergence to a minimum of a convex nonsmooth function.

We end this introduction with a note on notation: A sequence is denoted by $\{x_i\}_1^\infty$, and a subsequence by $\{x_{i_k}\}_{k=1}^\infty$. Sometimes we just denote a sequence by $\{x_i\}$ and use $y \to x$ to denote that $\{y_i\} \to x$. $\mathbb{R}^n$ is the Euclidean n-dimensional space.

Greek lowercase letters are scalars. Latin lowercase letters $i, \ldots, q$ denote integers; f is reserved to denote the functional $f(\cdot) : \mathbb{R}^n \to \mathbb{R}$; and $o(\cdot) : \mathbb{R}_+ \to \mathbb{R}$ is a scalar function such that $\lim_{\eta \downarrow 0} \frac{o(\eta)}{\eta} = 0$. All other Latin lowercase letters represent vectors (points) in $\mathbb{R}^n$. Subindices represent different entities, and superindices components; for instance, y_i^k is the kth component of the vector y_i. The standard inner product in $\mathbb{R}^n$ is denoted by $x^T y \doteq \sum_{k=1}^n x^k y^k$, and xy^T is an $n \times n$ matrix with elements $x^i y^j$. The rest of the notation is standard.

2. Necessary condition for nonsmooth functions. B-differentiable functions, which were introduced in [34], have a directional derivative (B-derivative) $f'(\cdot, \cdot) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ that satisfies

$$[\eta > 0] \Rightarrow [f'(x, \eta d) = \eta f'(x, d)], \tag{2.1}$$

$$f(x + d) - f(x) - f'(x, d) = o(||d||). \tag{2.2}$$

In general, these functions are not Fréchet differentiable but appear naturally in many optimization problems. Some difficult smooth problems can be reformulated as nonsmooth problems with a simpler structure, which can be efficiently solved by suitably adapting Newton-related methods [29, 30, 33]. Moreover, necessary and sufficient conditions for nonlinear programming have been established for this kind of function [14, 16, 46]. Nonetheless, to the authors' knowledge there exists no direct search algorithm with guaranteed convergence to a local minimum of a B-differentiable function. We partially answer this question in this paper. We propose an algorithm that generates a subsequence $\{x_{i_k}\}_{k=1}^{\infty}$ that converges to a point x satisfying the NSNC given below, but if $f(\cdot)$ happens to be differentiable, then $\nabla f(x) = 0$.

In what follows we will assume the following conditions:

C1. $f(\cdot)$ is bounded below.

C2. For any $x, d \in \mathbb{R}^n$ there exists $f'(x, d) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ such that

$$[\eta > 0] \Rightarrow \left[\begin{array}{l} f'(x, \eta d) = \eta f'(x, d), \\ f(x + \eta d) - f(x) - \eta f'(x, d) = o(\eta) \end{array} \right]. \tag{2.3}$$

Note that $f'(x, d) = \lim_{\eta \downarrow 0} (f(x + \eta d) - f(x))/\eta$.

C3. The sequence $\{x_i\}_1^{\infty}$ remains in a compact set.

Condition C3 will be needed to merely ensure the existence of accumulation points of $\{x_i\}_1^{\infty}$. Conditions that imply C3 are, for instance,

(i) $f(\cdot)$ is coercive, i.e., $[\{\|x_i\|\} \to \infty] \Rightarrow [\{f(x_i)\} \to \infty]$, or

(ii) the lower level set $\{x \in \mathbb{R}^n : f(x) \leq f(x_1)\}$ is compact.

The next lemma follows a standard proof. It shows that C2 implies a well-known property of a local minimizer.

LEMMA 2.1. *Let* C2 *hold. If x is a local minimizer of $f(\cdot)$, then $f'(x, d) \geq 0$ for all $d \in \mathbb{R}^n$.*

Proof. If $f'(x, d) < 0$ for some $d \in \mathbb{R}^n$, there exists $\bar{\eta} > 0$ such that $o(\eta)/\eta \leq -f'(x, d)/2$ for all $0 < \eta \leq \bar{\eta}$. We now obtain from C2 that

$$f(x + \eta d) - f(x) = \eta(f'(x, d) + o(\eta)/\eta) \leq \eta \left(f'(x, d) - f'(x, d)/2\right) = \eta f'(x, d)/2 < 0,$$

and x is not a minimizer. ∎

The conclusion of the previous lemma is in general hard to verify unless $f(\cdot)$ is (sub)differentiable. We now state a practical NSNC, which will be helpful as well in constrained minimization on subspaces.

Nonsmooth necessary condition (NSNC). Let $x \in \mathbb{R}^n$ be a local minimizer of $f(\cdot)$ on a subspace S, and let $\mathcal{D} = \{d_1, \ldots, d_m\}$ be a set of m bounded nonzero directions in $\mathbb{R}^n$ that spans S. If C2 holds, then x satisfies the NSNC

$$[d \in \mathcal{D}] \Rightarrow [f'(x,d) \geq 0, f'(x,-d) \geq 0]. \tag{2.4}$$

We point out that (2.4) is adequate for differentiable functions, for if $\mathcal{S} = \mathbb{R}^n$, if $f'(x,d) \stackrel{\text{def}}{=} \nabla f(x)^T d$, and if x satisfies (2.4), then $\nabla f(x) = 0$. (The proof is a simpler version of Theorem 3.4.) To end this section we recall a definition that we will use frequently in this paper: $f(\cdot)$ *is strictly differentiable at* x *if* $\nabla f(x)$ *exists and* $\lim_{y \to x, \eta \downarrow 0} \frac{f(y+\eta d)-f(y)}{\eta} = \nabla f(x)^T d$ *for all* $d \in \mathbb{R}^n$ (see [7] for further details).

3. Sequential and parallel algorithms.

3.1. Sequential algorithms. This subsection studies a prototype algorithm amenable to both single and multiprocessor environments. It will be shown that, under C1–C3, the algorithm generates a subsequence $\{x_{i_k}\}_{k=1}^{\infty}$ that converges to a point satisfying the NSNC (2.4).

We now describe the algorithm identified below as Prototype Algorithm 3.1. Given an estimate x_i, a bounded stepsize $h_i > 0$, and a *finite* set of search directions $\mathcal{D} = \{d_1, \ldots, d_m\}$, the algorithm explores the function values at the points $x_i + h_i^j d_j$, $x_i - h_i^j d_j$, $j = 1, \ldots, m$. If the function sufficiently decreases along a search direction, i.e., for some $d_j \in \mathcal{D}$,

$$\text{either} \quad f(x_i + h_i^j d_j) - f(x_i) \leq -|o^j(h_i^j)|, \tag{3.1}$$

$$\text{or} \quad f(x_i - h_i^j d_j) - f(x_i) \leq -|o^j(h_i^j)|, \tag{3.2}$$

a new estimate x_{i+1} is generated and the associate stepsize component h_i^j may be expanded as long as $h_{i+1}^j \leq \lambda_t \tau_i$, with $\lambda_t > 1$, and $\{\tau_i\} \to 0$. On the other hand, if neither (3.1) nor (3.2) holds for any $d_j \in \mathcal{D}$, we declare the point x_i to be *blocked* by the stepsize vector h_i. The algorithm reduces the upper bound τ_i ($\tau_{i+1} < \tau_i$) as an attempt to unblock x_i. In the prototype algorithms, τ_i represents an upper bound of the ∞-norm of the stepsize vector h at blocked points.

Prototype Algorithm 3.1 ($f'(x,d)$ exists).

Data: $0 < \mu < 1,\ 0 < \lambda_s < 1 < \lambda_t,$ with $\mu\lambda_t < 1,$
$\mathcal{D} := \{d_1, \ldots, d_m\},\ x_i,\ 0 < h_i,\ ||h_i||_\infty \leq \tau_i,\ o^j(h_i^j),\ j = 1, \ldots, m.$

1. Define the index set $\mathcal{J}_i$ of unblocked directions as

$$\mathcal{J}_i \doteq \{1 \leq j \leq m : \quad f(x_i + \beta h_i^j d_j) - f(x_i) \leq -|o^j(h_i^j)| \text{ for some } \beta \in \{-1, 1\}\}. \tag{3.3}$$

2. **If** $\mathcal{J}_i \neq \emptyset$, let $\tau_{i+1} = \tau_i$ and choose $j \in \mathcal{J}_i$, x_{i+1}, h_{i+1}^j such that

$$x_{i+1} \in \{x \in \mathbb{R}^n : f(x) \leq f(x_i + \beta h_i^j d_j)\}, \quad \lambda_s \tau_i \leq h_{i+1}^j \leq \lambda_t \tau_i; \tag{3.4}$$

else ($\mathcal{J}_i = \emptyset$)

$$\text{let } x_{i+1} = x_i,\ \tau_{i+1} = \mu||h_i||_\infty, \text{ and choose } h_{i+1}^j \text{ such that } \lambda_s \tau_{i+1} \leq h_{i+1}^j \leq \tau_{i+1}, \quad j = 1, \ldots, m. \tag{3.5}$$

end if

Repeat 1–2 **while** τ_i is not small enough.

When gradients are available, a sufficient decrease condition has been formally established [1, 45], and a descent direction $d \in \mathbb{R}^n$ at x is easily characterized, namely, $d^T \nabla f(x) < 0$. Convergence of the search methods is based on the fact that at least one of the directions of search satisfies this descent condition. Our proof of convergence departs from this idea because our algorithms are mainly addressed to nonsmooth functions. In order to ensure convergence of the algorithm, we introduce the sufficient decrease condition (3.1)–(3.2) for nonsmooth functions. A similar condition was first discussed in [12] and later analyzed in [24] for continuously differentiable functions. A related concept, denoted as the fortified descent condition, is given in [42].

The following lemma is useful because it ensures that the algorithm is well-defined, in the sense that there always exists an h_i^j such that (3.1) or (3.2) holds whenever $f'(x_i, d_j) < 0$ or $f'(x_i, -d_j) < 0$.

LEMMA 3.1. *Let $x, d \in \mathbb{R}^n$ be, respectively, a given point and a bounded direction of search. Let $f'(x, d) < 0$. There exists $\eta > 0$ such that $f(x + \eta d) - f(x) \leq -|o(\eta)|$.*

Proof. Assume that no such η exists; then $[\eta > 0] \Rightarrow [f(x + \eta d) - f(x) > -|o(\eta)|]$. Hence, $\frac{f(x+\eta d)-f(x)}{\eta} > -\frac{|o(\eta)|}{\eta}$, and in the limit we obtain $f'(x, d) \geq 0$, which contradicts the assumption. □

We now proceed with the theoretical justification of the algorithm. We assume $\mu\lambda_t < 1$, C1–C3, and that given any $\epsilon > 0$ there exists $\delta(\epsilon) > 0$ such that $[\{h_{i_k}^j\}_{k=1}^{\infty} \geq \epsilon] \Rightarrow [\{o^j(h_{i_k}^j)\}_{k=1}^{\infty} \geq \delta(\epsilon)]$. (See the remark after Corollary 3.6.) Theorem 3.2 ensures that $\{h_i\}_1^{\infty} \to 0$. Theorems 3.4, 3.5 and Corollary 3.6 state that the sequence of blocked points converges to a point satisfying (2.4).

THEOREM 3.2. $\{h_i\}_1^{\infty} \to 0$.

Proof. By construction, $\lambda_s \tau_i \leq h_i^j \leq \lambda_t \tau_i$, $j = 1, \ldots, m$, and $\{\tau_i\}$ is a nonincreasing sequence that reduces its values only at blocked points. Indeed, by (3.4) and (3.5) we have $\tau_{i+1} = \mu ||h_i||_{\infty} \leq \mu\lambda_t \tau_i < \tau_i$. Hence, if blocked points occur infinitely often, the proof follows trivially for $\{\tau_i\} \to 0$. We now assume that (3.5) occurs a finite number of times and will reach a contradiction.

Let $\tau_i = \tau_k > 0$ for all $i \geq k$, and let $\epsilon = \lambda_s \tau_k$. We assert that $h_i^j \geq \epsilon$ for any $j \in \mathcal{J}_i, i \geq k$. Therefore for $i \geq k$ we obtain that

$$f(x_{i+1}) \leq f(x_i) - |o^j(h_i^j)| \leq f(x_i) - \delta(\epsilon),$$

and $\{f(x_i)\}$ decreases without bound, contradicting C1. □

COROLLARY 3.3. *There is an infinite number of blocked points.*

THEOREM 3.4. *Let $Span(\mathcal{D}) = \mathbb{R}^n$. Let x be a limit point of a sequence of blocked points, and let $f(\cdot)$ be strictly differentiable at x. Under these assumptions $\nabla f(x) = 0$.*

Proof. With no loss of generality, let us assume that $\{x_i\}_1^{\infty}$ is the subsequence of blocked points, and $\{x_i\}_1^{\infty} \to x$. For any $d_j \in \mathcal{D}$ we have

$$f(x_i + h_i^j d_j) - f(x_i) > -|o^j(h_i^j)|;$$

hence,

$$\nabla f(x)^T d_j = \lim_{\{x_i\} \to x, \{h_i^j\} \downarrow 0} \frac{f(x_i + h_i^j d_j) - f(x_i)}{h_i^j} \geq \lim_{\{h_i^j\} \downarrow 0} \frac{-|o^j(h_i^j)|}{h_i^j} = 0.$$

Similarly, $\nabla f(x)^T (-d_j) \geq 0$. Therefore $\nabla f(x)^T d_j = 0$. Since this equation is valid for all $d_j \in \mathcal{D}$ and $Span(\mathcal{D}) = \mathbb{R}^n$, we conclude that $\nabla f(x) = 0$. □

The last theorem is useful when strict differentiability holds at limit points. Obviously, (NSNC) holds. Although $f(\cdot) \in C^1$ is not required everywhere, C2 plus strict differentiability implies that $f(\cdot)$ must be smooth in a neighborhood of the limit point. We now turn our attention to convergence conditions that ensure (NSNC) without assuming strict differentiability. It is straightforward to show that for $d \in \mathcal{D}$, $\limsup_{y \to x, \eta \downarrow 0}(f(y + \eta d) - f(y))/\eta \geq 0$ at limit points of blocked point sequences. However, this result is in general not useful. There are examples that show that negative directional derivatives may appear at x and along some directions $d \in \mathcal{D}$ [2]. Generally, local convexity must be assumed in smooth problems to make sure x is a local minimum. Theorem 3.5 below proves that (NSNC) holds with a local convexity assumption. However, the Dennis–Woods function (see section 6) reveals that thus far no known search method ensures convergence to a minimum of a nonsmooth convex function.

Let us recall that when $f(\cdot)$ is convex, the function $\varphi(\eta) \doteq (f(x + \eta d) - f(x))/\eta$ is a nondecreasing function of $\eta > 0$ for fixed $x, d \in \mathbb{R}^n$. Indeed $\varphi'(\eta) = \frac{1}{\eta^2}\left[f(x) - f(x + \eta d) + \eta d^T \nabla f(x + \eta d)\right] > 0$. A general result for nondifferentiable convex functions appears in [35, Theorem 23.1].

THEOREM 3.5. *Let* C1–C3 *hold. Let* $\{x_i\} \to x$ *be a (sub)sequence of blocked points generated by the Prototype Algorithm* 3.1*, and let* $f(\cdot)$ *be convex in a neighborhood of* x*. Under these assumptions* (NSNC) *holds at* x*.*

Proof. By assumption we have for any j that

$$\begin{gathered} \{h_i^j\} \to 0 \text{ and} \\ (f(x_i + h_i^j d_j) - f(x_i))/h_i^j > -|o^j(h_i^j)|/h_i^j. \end{gathered} \tag{3.6}$$

We will prove that $f'(x, d_j) \geq 0$. Assume on the contrary that $f'(x, d_j) = -\alpha < 0$. If so, there exists $\bar{\eta} > 0$ such that $(f(x + \bar{\eta} d_j) - f(x))/\bar{\eta} \leq -\alpha/2$. Hence, for any sequence $\{x_i\} \to x$ and large enough i, we have that $(f(x_i + \bar{\eta} d_j) - f(x_i))/\bar{\eta} \leq -\alpha/4$. By convexity, $(f(x_i + \eta d_j) - f(x_i))/\eta \leq -\alpha/4$ for all $0 < \eta \leq \bar{\eta}$, which contradicts (3.6). We prove similarly that $f'(x, -d_j) \geq 0$. Since j was arbitrary, we conclude that (NSNC) holds. ∎

COROLLARY 3.6. *If the number of points that satisfy* (2.4) *is finite, the sequence of blocked points converges.*

Proof. The proof is trivial. See [28, Theorem 14.1.5]. ∎

Remark. For a practical implementation, the sufficient decrease condition may be written as $f(x_i \pm h_i^j d_j) - f(x_i) \leq -(h_i^j)^2$. Note that $h_i^j > \epsilon > 0 \Rightarrow o^j(h_i^j) = (h_i^j)^2 > \epsilon^2 = \delta(\epsilon) > 0$.

Remark. Once you have chosen an index, $j \in \mathcal{J}_i, x_{i+1}$ can be obtained by any heuristic or by any *finite* procedure that fulfills (3.4).

3.2. Parallel algorithms. We assume that we have p processors that share x_i, the best estimate, and can compute function values. We associate processor k with the index set $\mathcal{K}_k$, and $\bigcup_{k=1}^p \mathcal{K}_k = \{1, \ldots, m\}$. We define $\mathcal{D}_k \doteq \{d_j \in \mathcal{D} : j \in \mathcal{K}_k\}$.

Table 3.1 presents two direct translations of the prototype algorithm to parallel implementations with a balanced load among processors. Version 1 assumes that a function evaluation is costly and time consuming, whereas version 2 assumes that the communication and synchronization load can override the computational work.

In the parallel version 1, all processors simultaneously perform one function evaluation, say $f(x + \beta h^j d_j)$, for some $\beta \in \{-1, 1\}$, $j \in \mathcal{K}_k$, $k = 1, \ldots, p$, and a global reduction is used to determine the minimum function value among all these values. The minimizer, its function value, and the new stepsize vector are broadcast to all

TABLE 3.1
Iteration of derivative-free algorithms.

Input: $x \in \mathbb{R}^n, \varphi = f(x), \mathcal{D} = \{d_1, \ldots, d_m\}, h \in \mathbb{R}^m, block, \mathcal{K}_1, \ldots, \mathcal{K}_p$		
Sequential	Parallel (version 1)	Parallel (version 2)
		do in parallel
		$k =$ processor_id
if $block = 2$, **then**	**if** $block = 2$, **then**	**if** $block = 2$, **then**
$block = 0, \tau = 0.2\|\|h\|\|_\infty$	$block = 0, \tau = 0.2\|\|h\|\|_\infty$	$block = 0, \tau = 0.2\|\|h\|\|_\infty$
endif	**endif**	**endif**
$block = block + 1$	$block = block + 1$	$y_k = x, g^k = \varphi, b^k =$ **true**
for j $= 1, \ldots,$m	**for** i $= 1, \ldots,$m/p	**for** $j \in \mathcal{K}_k$
	do in parallel	
	$k =$ processor_id	
	$j_k = i$th index of $\mathcal{K}_k$	
$z = x + h^j d_j$	$z_k = x + h^{j_k} d_{j_k}$	$z = y_k + h^j d_j$
$\theta = f(z)$	$\theta_k = f(z_k)$	$\theta = f(z)$
if $\theta - \varphi \le -(h^j)^2$, **then**	$g^k = \theta_k - \varphi + (h^j)^2$	**if** $\theta - g^k \le -(h^j)^2$, **then**
$h^j = \min(1.4h^j, 4.9\tau)$		$h^j = \min(1.4h^j, 4.9\tau)$
$x = z, \varphi = \theta, block = 0$		$y_k = z, g^k = \theta, b^k =$ **false**
else	**if** $g^k > 0$ **then**	**else**
$h^j = -h^j$	$h^{j_k} = -h^{j_k}$	$h^j = -h^j$
endif	**endif**	**endif**
end for		**end for**
	end do in parallel	**end do** in parallel
		if and$(b^1, \ldots, b^p)$, **then**
		$block = block + 1$
		else
	$k = \arg\min_{1 \le q \le p}(g^q)$	$j = \arg\min_{1 \le k \le p}(g^k)$
	if $g^k \le 0$, **then**	
	$h^{j_k} = \min(1.4h^{j_k}, 4.9\tau)$	
	$x = z_k, \varphi = \theta_k, block = 0$	$x = y_j, \varphi = g^j, block = 0$
	endif	**endif**
	end for	

processors. In the parallel version 2, all processors carry out several function evaluations on a subset of the directions of search and broadcast the best point found and its function value. The iteration is completed as in the previous version. Practical implementations of both versions are given in Table 3.1. Note that if $\mathcal{K}_k = \{k\}$, both versions generate the same sequence $\{x_i\}_1^\infty$.

Both implementations have a serious drawback. Function evaluations may stem from simulations of complex systems with indefinite response time, which renders useless any effort to balance the load among processors. Consequently, both parallel versions in Table 3.1 may become highly inefficient. Fortunately, our prototype algorithm can be naturally adapted to the asynchronous parallel implementation proposed in [15] and overcome this difficulty. A processor, say processor k, works with its associate index set $\mathcal{K}_k$ and broadcasts an estimate x_{i+1} that satisfies (3.4) with $j \in \mathcal{K}_k$. The asynchronous algorithm is basically as follows.

ASYNCHRONOUS ALGORITHM (kth processor).

1. Get $x_i, f(x_i), h_i$, the successful triplet broadcast by the other processor.
2. Perform the (appropriate) function evaluation required by Algorithm 3.1 ($\mathcal{D}$ is replaced by $\mathcal{D}_k \doteq \{d_j : j \in \mathcal{K}_k\}$).
3. **If** there is a new better broadcast point, go to step 1.
4. **If** a better point is found along a direction d_j, then set $h_{i+1}^j = 1.4h_i^j$ and

TABLE 3.2
Example of a fault tolerance for a parallel algorithm.

Direction	Processor		
	1	2	3
d_1	X		X
d_2	X	X	
d_3		X	X

broadcast a successful triplet $x_{i+1}, f(x_{i+1}), h_{i+1}^j$;
else reduce $h^j, j \in \mathcal{K}_k$, by (5.1).
5. **If** $||h_i|| > \epsilon$, go to step 2; **else** STOP.
end of algorithm

Convergence theory of the asynchronous algorithm along with numerical results for all parallel implementations will be given in a forthcoming paper.

Before we end this section we would like to point out that a certain degree of fault tolerance in any parallel version can be included from the onset. We simply force every index to appear in at least two index subsets. This in turn forces any direction to be searched by at least two processors. If a processor goes down, it can pass unnoticed until it again goes up. Let us illustrate this idea with a trivial example. Let $\mathcal{D} = \{d_1, d_2, d_3\}$, $p = 3$, and $\mathcal{D}_1 = \{d_1, d_2\}$, $\mathcal{D}_2 = \{d_2, d_3\}$, and $\mathcal{D}_3 = \{d_1, d_3\}$, as shown in Table 3.2. If any one processor goes down, the others still search the whole set $\mathcal{D} = \{d_1, d_2, d_3\}$.

4. Extensions and future research.

4.1. The searching set. We can use any static set of linearly independent *unit* directions: the coordinate axis, random generated directions, conjugate directions [31, 32], and so on. It is commonly accepted that occasional but judicious adjustments to the search set might improve the convergence of direct search methods. For instance, the *rank ordered* pattern search method suggested in [21] includes the direction of best decrease on the simplex vertices; the *implicit filtering algorithm* searches on the *simplex gradient* (see subsection 4.2), and in [13] a quasi-Newton direction is included at blocked points. There is as well a convergence proof for dynamic sets in the algorithm proposed in [24].

For the sake of completeness we now sketch the convergence for dynamic sets. We denote by $\mathcal{D}_i = \{d_{i1}, \ldots, d_{im}\}$ a set of m unit directions at the ith iteration.

THEOREM 4.1. *Assume that* $\{d_{ij}\}_{i=1}^{\infty} \to d_j, j = 1, \ldots, m$. *If* C2 *holds and* $f(\cdot)$ *is Lipschitzian near any limit point* x *of the sequence of blocked points* $\{x_i\}_1^{\infty}$, *then* $f'(x, d_j) \geq 0$.

Proof. Let κ be the Lipschitz constant. With no loss of generality, assume that the sequence of blocked points $\{x_i\}_1^{\infty}$ converges to x. For any $d_{ij} \in \mathcal{D}_i$ we have

$$\begin{aligned} & f(x_i + h_i^j d_j) - f(x_i) \\ & \quad = f(x_i + h_i^j d_{ij}) - f(x_i) + f(x_i + h_i^j d_j) - f(x_i + h_i^j d_{ij}) \\ & \quad > -|o^j(h_i^j)| + f(x_i + h_i^j d_j) - f(x_i + h_i^j d_{ij}) \\ & \quad > -|o^j(h_i^j)| - |f(x_i + h_i^j d_j) - f(x_i + h_i^j d_{ij})|; \end{aligned}$$

therefore,

$$\frac{f(x_i + h_i^j d_j) - f(x_i)}{h_i^j} > -\frac{|o^j(h_i^j)|}{h_i^j} - \kappa ||d_j - d_{ij}||.$$

By taking limits, we obtain that $f'(x, d_j) \geq 0$. □

We note that if $f(\cdot)$ is convex and bounded above near the limit point x, it fulfills the conditions of the previous theorem [7, Proposition 2.2.6]. Furthermore, strict differentiable functions at x also satisfy the conditions of the theorem [7, Proposition 2.2.1]. In this case, we obtain that $\nabla f(x)^T d_j \geq 0$. Therefore, it is trivial to conclude that if $Span(d_1, \ldots, d_m) = \mathbb{R}^n$, if x is a limit point of a sequence of blocked points, and if $f(\cdot)$ is strictly differentiable at x, then $\nabla f(x) = 0$.

Since the sequence of blocked points converges to a point that satisfies (2.4), it seems worthwhile to explore the direction determined by the last two blocked points. The set $\mathcal{D}_3$ given below (see also [13]) includes this direction. Furthermore, it has desirable characteristics: (i) it is completely determined by the vector u, which significantly reduces the communication load in a multiprocessing environment, and (ii) its associated $n \times n$ matrix $D_3 \doteq [d_1, \ldots, d_n]$ is easily invertible, which is a nice feature to be discussed below.

This paper investigates the performance of the algorithm on the searching sets $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$ given next:

$\mathcal{D}_1$: $d_j = e_j, j = 1, \ldots, n$, the coordinate axis.

$\mathcal{D}_2$: See [17, p. 80]; also suggested in [9].

$$d_j^k = \begin{cases} \alpha & \text{if } k \neq j, \\ \beta & \text{if } k = j, \end{cases} \qquad \text{where } \alpha = \frac{\sqrt{n+1} - 1}{\sqrt{2}n},\ \beta = \alpha + \frac{1}{\sqrt{2}},\ j = 1, \ldots, n.$$

$\mathcal{D}_3$: Let x_{q+1}, x_q be two consecutive blocked points, and let

$$s = \frac{x_{q+1} - x_q}{||x_{q+1} - x_q||}, \quad j = \arg\max_k(|s^k|),$$

$$u^j = +\sqrt{\frac{(1 + |s^j|)}{2}}, \quad u^k = sign(s^j)\ s^k / 2u^j \text{ for } k \neq j.$$

Choose $d_k = (I - 2uu^T)(e_j + e_k), k = 1, \ldots, n$.

The columns of D_3^{-T} are

$$D_3^{-T} e_j = \frac{1}{2}(I - 2uu^T)\left(e_j - \sum_{k \neq j} e_k\right),$$

$$D_3^{-T} e_k = (I - 2uu^T)e_k, \quad k \neq j.$$

We end this subsection with a remark on dynamic sets $\mathcal{D}_i = \{d_{i1}, \ldots, d_{im}\}$. Let $\{d_{ij}\}_{i=1}^{\infty} \to d_j$, $j = 1, \ldots, m$. It is important that d_j, $j = 1, \ldots, m$, be linearly independent. Consider, for example, the problem $\min(x^2 + y)$ and $\mathcal{D}_i = \{(1, 0), (1, h_i)\}$. Starting at $(x_0, y_0) = (0, 0)$, the algorithm stalls at $(0, 0)$, which is not a stationary point. Indeed, $f(0, 0) = 0$, and

$$f(h_i(\pm 1, 0)) = h_i^2 > 0, \quad f(h_i(1, h_i)) = 2h_i^2 > 0, \quad \text{and } f(h_i(-1, -h_i)) = 0.$$

4.2. Smooth functions. If we assume that $f(\cdot)$ is strictly differentiable at any limit point, Theorem 3.4 shows that the sequence of blocked points generated by Algorithm 3.1 (and its parallel counterparts) converges to a first-order stationary point. We show below that, under this differentiability condition, a blocked point can be detected with fewer function values per iteration, which seems to imply that an algorithm with this property should be more efficient.

It is known that a basis of $n+1$ *positively independent* directions that positively span $\mathbb{R}^n$ suffices to prove convergence in direct search methods [3, 21, 24]. We recall that a basis $\{d_1, \ldots, d_{n+1}\}$ positively spans $\mathbb{R}^n$ if

$$\forall (x \in \mathbb{R}^n)\ \exists (\nu_1 \geq 0, \ldots, \nu_{n+1} \geq 0) : x = \sum_{k=1}^{n+1} \nu_k d_k.$$

We remark that the set $\mathcal{D}$ can easily be constructed. Let *Span* $\{d_1, \ldots, d_n\} = \mathbb{R}^n$, and let $d_{n+1} = -\sum_{k=1}^{n} \alpha_k d_k, \alpha_k > 0$.

Let us establish the counterpart of Lemma 3.1 for differentiable functions and a direction set that positively spans $\mathbb{R}^n$.

LEMMA 4.2. *Let $\nabla f(x)$ exist, and let $f'(x,d) \stackrel{\text{def}}{=} \nabla f(x)^T d$ for all $d \in \mathbb{R}^n$. Let $\mathcal{D}$ be a basis set of $n+1$ search directions that positively spans $\mathbb{R}^n$. If $\nabla f(x) \neq 0$, then there exist $\eta > 0, d_j \in \mathcal{D}$ such that $f(x + \eta d_j) - f(x) \leq -|o(\eta)|$.*

Proof. If no such η, d_j exist, then $f(x+\eta d) - f(x) > -|o(\eta)|$ for all $\eta > 0, d \in \mathcal{D}$. In the limit we obtain that $\nabla f(x)^T d \geq 0$ for all $d \in \mathcal{D}$. But by assumption $-\nabla f(x) = \sum_{k=1}^{n+1} \nu_k d_k$ for some $\nu_k \geq 0$; therefore $-\nabla f(x)^T \nabla f(x) = \sum_{k=1}^{n+1} \nu_k \nabla f(x)^T d_k \geq 0$, which only holds for $\nabla f(x) = 0$. ∎

Based on Lemma 4.2, the following algorithm seems appropriate for differentiable functions.

PROTOTYPE ALGORITHM 4.1 ($f(\cdot) \in C^1$).

Data: $0 < \mu < 1,\ 0 < \lambda_s < 1 < \lambda_t,$ with $\mu\lambda_t < 1,$
$\mathcal{D} := \{d_1, \ldots, d_{n+1}\},\ x_i,\ 0 < h_i,\ ||h_i||_\infty \leq \tau_i, o^j(h_i^j),\ j = 1, \ldots, n+1.$

1. Define the index set $\mathcal{J}_i$ of unblocked directions as

$$\mathcal{J}_i \doteq \{1 \leq j \leq n+1 : f(x_i + h_i^j d_j) - f(x_i) \leq -|o^j(h_i^j)|\}.$$

2. **If** $\mathcal{J}_i \neq \emptyset$, let $\tau_{i+1} = \tau_i$ and choose $j \in \mathcal{J}_i$, x_{i+1}, h_{i+1}^j such that

$$x_{i+1} \in \{x \in \mathbb{R}^n : f(x) \leq f(x_i + h_i^j d_j)\}, \quad \lambda_s \tau_i \leq h_{i+1}^j \leq \lambda_t \tau_i;$$

else ($\mathcal{J}_i = \emptyset$) let $x_{i+1} = x_i$, $\tau_{i+1} = \mu ||h_i||_\infty$, and choose h_{i+1}^j such that

$$\lambda_s \tau_{i+1} \leq h_{i+1}^j \leq \tau_{i+1}, \quad j = 1, \ldots, n+1.$$

end if

Repeat 1–2 **while** τ_i is not small enough.

Theorem 3.4 and Lemma 4.2 lead to the following result: If $f(\cdot)$ is everywhere differentiable and strict differentiable at limit points, Algorithm 4.1 generates a (sub)sequence $\{x_i\}$ that converges to a point that satisfies the first-order necessary condition. If local convexity is assumed in place of strict differentiability, Theorem 3.5 can be used to prove that $f'(x,d) \geq 0$ for all $d \in \mathcal{D}$, but this does not seem to be a useful nonsmooth necessary condition. Section 5 reports some numerical results on differentiable functions with Algorithms 3.1 and 4.1.

Now, let us extract first-order information. It is well known that the vector $r = D^{-T}c$, where $d_k, k = 1, \ldots, n$, is the kth column of the matrix D, and $c^k = (f(x + \eta d_k) - f(x))/\eta$, $k = 1, \ldots, n$, is a good approximation to $\nabla f(x)$ for η small enough. The vector r computed for a given simplex was denoted as the *simplex gradient* in [5] and used as a possible direction of descent in the implicit filtering algorithm [19, Chapter 7]. First-order information is quite helpful for sufficiently smooth functions

because it allows quasi-Newton directions *(superlinear rate of convergence)* along the lines suggested in [6, 26, 37] and more recently in [38].

If we are certain that $f(\cdot) \in C^2$, the above approach is practical; otherwise, a lot of effort is being wasted. In [13] the gradient approximation r, with $c^j \doteq (f(x_i + h_i^j d_j) - f(x_i - h_i^j d_j))/2h_i^j$, is only computed at blocked points. The direction $d_{m+1} = -Hr$, where H is a variable metric, can be used to obtain x_{i+1} by (3.4).

4.3. Box constraints. There is a trivial way to adapt Algorithm 3.1 to the box constrained minimization problem $\min f(x)$, for $x \in \mathcal{B} := \{x \in \mathbb{R}^n : s \le x \le t\}$, where s, t are vectors in $\mathbb{R}^n$ and $s \le t$. We merely use the coordinate axes as the directions of search, i.e., $\mathcal{D} = \{e_1, \ldots, e_n\}$, and define a function $F(\cdot)$ as

$$F(x) \doteq \begin{cases} f(x), & x \in \mathcal{B}, \\ \max\{f(x), f(x_B)\}, & \text{otherwise}, \end{cases}$$

where $x_B^k = median\ (s^k, x^k, t^k)$ is the kth component of the projection of x onto the set $\mathcal{B}$. ($F(x) = \infty$, $x \notin B$, was suggested in [22, 23].) Obviously, $\min_{x\in\mathcal{B}} f(x)$ and $\min_{x\in\mathbb{R}^n} F(x)$ are equivalent minimization problems. It is as well immediate to observe that, starting at any $x_0 \in \mathcal{B}$, convergence is preserved when Algorithm 3.1 is used for solving the latter minimization problem. We remark that in a practical implementation no evaluation of $f(x)$ should be performed for $x \notin \mathcal{B}$.

More efficient algorithms can be suggested. This is the subject of a forthcoming paper that will be coupled with the more general linearly constrained optimization problem.

5. Numerical experiments. We implemented Algorithms 3.1 and 4.1, with $\tau_i = ||h_i||_\infty$ at blocked points, $\lambda_s = 0.01/n$, and $\lambda_t = 0.98/\mu$. Algorithm 3.1 detects a blocked point when $2n$ consecutive function evaluations fail to satisfy the sufficiency decrease condition (both side evaluations on n independent directions fail). This algorithm will be denoted as the nonsmooth directional search algorithm (NSDSA) because it is especially suited for nonsmooth functions. Algorithm 4.1 detects a blocked point when $n + 1$ consecutive function evaluations fail to satisfy the sufficiency decrease condition. This implementation is called the smooth directional search algorithm (SDSA) because it does not seem adequate for nonsmooth functions.

Implemented NSDSA Algorithm.

Input: An estimate x, its function value $\varphi = f(x)$, stepsizes $h^j = 1, j = 1, \ldots, n$,
descent index $j = 1$, index search $k = 1$, # of failures $fail = 0$,
direction generator u (or the set $\mathcal{D} : Span(\mathcal{D}) = \Re^n$),
$\tau = 1$, convergence precision $\epsilon = 10^{-6}$.

repeat
Generate $d_k = (I - 2uu^T)(e_j + e_k)$ (or obtain d_k from the set $\mathcal{D}$),
$z = x + h^k d_k, \theta = f(z)$.
if $(\theta - \varphi \le -(h^k)^2)$, **then**
$h^k = \min(\gamma h^k, (0.98/\mu)\tau), k = j$,
$x = z, \varphi = \theta, fail = 0$;
else
$h^k = -h^k, fail = fail + 1$,
$k = (k \bmod n) + 1$,
if $(fail = 2n)$, **then**
reduce $||h||_\infty$ by (5.1), $\tau = ||h||_\infty, fail = 0$.
Update the direction generator u and the indicator j,

$k = j$.
end if
end if
until ($||h||_\infty < \epsilon$) (or similar function values)

IMPLEMENTED SDSA ALGORITHM.

Input: An estimate x, its function value $\varphi = f(x)$, stepsizes $h^j = 1, j = 1, \ldots, n+1$,
descent index $j = 1$, index search $k = 1$, # failures $fail = 0$,
direction generator u (or the set $\mathcal{D}$ with $n+1$ positive basis),
$\tau = 1$, convergence precision $\epsilon = 10^{-6}$.

repeat
Generate $d_k = (I - 2uu^T)(e_j + e_k)$ (or obtain d_k from the set $\mathcal{D}$),
$z = x + h^k d_k$, $\theta = f(z)$.
if ($\theta - \varphi \leq -(h^k)^2$), **then**
$h^k = \min(\gamma h^k, (0.98/\mu)\tau), k = j$,
$x = z, \varphi = \theta, fail = 0$;
else
$fail = fail + 1$,
$k = (k \bmod (n+1)) + 1$,
if ($fail = n + 1$), **then**
reduce $||h||_\infty$ by (5.1), $\tau = ||h||_\infty, fail = 0$.
Update the direction generator u and the indicator j,
$k = j$.
end if
end if
until ($||h||_\infty < \epsilon$) (or similar function values)

The direction indicator given by j means that d_j is the descent direction determined by the last two blocked points. As described above, d_j is explored after any successful iteration, and it is also the first direction in the set $\mathcal{D}$ that the algorithm explores. The numerical results reported below with the adaptive direction $\mathcal{D}_3$ also include a heuristic that improved the performance of the algorithm notably: d_k was explored before d_p only if $h_i^k \geq h_i^p$.

To get an initial insight into the performance of the sequential algorithms, we implemented both versions in C. The results obtained were compared with those from the rank ordered pattern search (ROPS) algorithm described in [21] ($n+1$ directions) and from the multidirectional search (MDS) algorithm from [40] ($2n$ directions). We report the number of function evaluations needed to obtain a solution.

In most direct search methods, the choice of parameters seems to be crucial for the quality of convergence of the algorithm. We tried different choices of γ and μ. Intuitively, the jth stepsize component in (3.4) should not increase significantly, for convergence is determined when $||h_i|| < \epsilon$ for a small positive ϵ. We report results with $h_{i+1}^j = (1 + \frac{1}{q})h_i^j$, where q is the number of contractions. In some experiments not reported here we observed that a uniform reduction in all components of h_i could now and then cause unnecessary small steps in later iterations. The stepsize vector in the SDSA and NSDSA was reduced according to the following rule:

$$h_{i+1}^j := \begin{cases} \mu h_i^j & \text{if } h_i^j > 0.01||h_i||_\infty/n, \\ 0.01||h_i||_\infty/n & \text{otherwise.} \end{cases} \tag{5.1}$$

The initial stepsize was $h = 1$, and the stopping criteria were $||h||_\infty \leq 10^{-6}$ or $\max(|f(x \pm h^j d_j) - f(x)|) \leq 10^{-6}(|f(x)| + 1)$ at blocked points. The latter criterion

TABLE 5.1
Number of function evaluations for the Rosenbrock function ($\gamma = 1.4, \mu = 0.6$).

$x_o = 3$								
n	MDS	ROPS	SDSA			NSDSA		
			$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$
2	3563	14261	90071	1320	473	14105	3298	466
3	21196	19530	F	143	1011	23914	10883	1054
5	26366	26030	719820	203	1689	53347	37912	1874
10	126051	82337	F	860	5018	144749	185823	5705
x_o standard$(-1.2, 1, \ldots)$								
n	MDS	ROPS	SDSA			NSDSA		
			$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$
2	6287	8810	7246	F	247	4674	4666	406
3	16666	15578	10340	F	772	13618	6416	878
5	27426	24158	39637	1211	759	22413	21390	1437
10	89051	57158	12076	120918	3821	42050	68420	3287

attempts to terminate the algorithm when it detects that no significant improvement of the functional values will take place, or when the function value decreases too slowly. This forced premature termination in some problems, probably due to very shallow function level sets. We should also point out that when the function values are imprecise, ϵ need not be too small. Its value can be determined from the engineering process.

For the MDS and ROPS algorithms, the initial polytope from [17, p. 80] was taken, as suggested in [9]. (See searching set $\mathcal{D}_2$ in section 3 above.) For the SDSA algorithm, the searching sets $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$ were augmented with the unit direction along $-\sum_{k=1}^{n} d_k$.

A generalized Rosenbrock function (5.2) of n variables ($n = 2, 3, 5, 10$) was used to study the influence of the parameters γ and μ and searching sets $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$ on the performance of the algorithm. Note that this function possesses multiple stationary points for $n > 2$.

$$f(x) = \sum_{k=1}^{n-1} \left[(x^k - 1)^2 + 100 \left(x^{k+1} - (x^k)^2 \right)^2 \right]. \tag{5.2}$$

Tables 5.1–5.3 show the results for the ROPS, SDSA, MDS, and NSDSA algorithms on two starting points $x = 3$ and $x = (-1.2, 1, -1.2, 1, \ldots)$. In these tables, F stands for a solution which differs by more than 20% from the optimum value. This situation always occurred when the algorithm stopped due to a small relative change in function value (i.e, $< 10^{-6}$). Had the algorithm continued, it would have taken a significant number of function evaluations to generate the solution. Also, in Table 5.3, γ was taken as $\gamma = 1 + 1/q$, with q being the number of contractions performed by the algorithm. In this way, the expansion parameter of the stepsize gets smaller when the algorithm is converging to the solution.

MDS and ROPS always called for more function evaluations on fixed direction sets and sometimes failed for $\gamma = 1 + 1/q$, $\mu = 0.2$, while NSDSA always found the optimal solution for the given termination criteria. For a specific combination of the parameters γ and μ, fewer than 1000 function evaluations were needed by NSDSA to obtain a stationary point of the Rosenbrock function of 10 variables.

TABLE 5.2
Number of function evaluations for the Rosenbrock function ($\gamma = 1.4, \mu = 0.2$).

$x_o = 3$								
n	MDS	ROPS	SDSA			NSDSA		
			$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$
2	4887	19427	F	1418	482	24094	4205	495
3	21988	17398	F	84	793	33953	19119	830
5	106726	40226	F	84	1523	72636	54570	1694
10	113811	88585	F	1375	4365	201939	186207	4134
x_o standard$(-1.2, 1, \ldots)$								
n	MDS	ROPS	SDSA			NSDSA		
			$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$
2	6255	8246	5783	F	350	7780	6373	346
3	15376	14410	14756	F	918	14916	8311	758
5	7126	14000	71017	9536	1180	6527	23684	822
10	101911	83074	1656	72541	1452	50293	107927	909

TABLE 5.3
Number of function evaluations for the Rosenbrock function ($\gamma = 1 + 1/q, \mu = 0.2$).

$x_o = 3$								
n	MDS	ROPS	SDSA			NSDSA		
			$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$
2	F	22838	F	65	1026	23842	4315	572
3	F	33162	F	100	1809	36731	18409	1376
5	48696	73892	552616	F	2977	62096	55127	2327
10	197251	F	F	2648	7611	125061	143467	5716
x_o standard$(-1.2, 1, \ldots)$								
n	MDS	ROPS	SDSA			NSDSA		
			$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$
2	F	20078	7771	F	659	6389	6699	528
3	16060	36834	24026	F	2058	14310	6814	1389
5	F	70778	83179	1261	1850	9824	15882	880
10	190811	F	33310	F	8277	38844	121122	3864

The results for the adaptive searching set $\mathcal{D}_3$ are certainly remarkable for both SDSA and NSDSA, but the performance of SDSA may be quite sensitive to the set of positive bases used. Table 5.4 shows the results for $\gamma = 1.4$, $\mu = 0.2$, and $\mathcal{D} = \{-e_1, \ldots, -e_n, \frac{1}{\sqrt{n}} \sum_{j=1}^{n} e_j\}$ (the negative of $\mathcal{D}_1$). These results seem to indicate that the adaptive searching set not only contributes to improving the efficiency of the algorithm but also makes it more robust and reliable. We might also conjecture that the extra computation of function values needed by NSDSA to detect a blocked point provides it with additional information that improves its overall performance.

To close this section and in order to get a better picture of the performance of the algorithms, we solved some test problems from the CUTE collection. For the SDSA and NSDSA algorithms, the adaptive searching set $\mathcal{D}_3$ was used. All algorithms were run with $\gamma = 1.4$, $\mu = 0.2$, and the termination criteria indicated above. Table 5.5 reports the number of function evaluations needed by the algorithms. Along with

TABLE 5.4
Number of function evaluations for the SDSA algorithm for the Rosenbrock function ($\gamma = 1.4, \mu = 0.2$) for the negative of the searching set $\mathcal{D}_1$.

n	$x_o = 3$	x_o standard$(-1.2, 1, \ldots)$
2	6956	3281
3	19271	9098
5	23196	36408
10	36746	185451

TABLE 5.5
Number of function evaluations (function value) for different algorithms in some problems from the CUTE collection.

Problem	ROPS	SDSA	MDS	NSDSA
HATFLDD, n=3 (6.6E-8)	2606 (3.8E-3)	177 (7.9E-4)	3652 (1.0E-3)	114 (2.9E-5)
MOREBV, n=10 (0.0)	1916 (5.1E-4)	2546 (5.3E-4)	3591 (3.2E-4)	3476 (5.7E-6)
FMINSURF, n=16 (1.0)	2467 (1.0)	11881 (1.0)	3697 (1.0)	17135 (1.0)
DIXMAANK, n=15 (1.0)	2882 (1.0)	251 (1.0)	2836 (1.0)	8601 (1.0)
EDENSCH, n=36 (219.3)	9844 (219.3)	20469 (219.3)	15733 (219.3)	5622 (219.3)
CRAGGLVY, n=50 (15.4)	94454 (F) (22.9)	38442 (15.4)	86351 (F) (21.5)	38993 (17.6)
ERRINROS, n=50 (39.9)	59570 (F) (40.7)	36810 (F) (45.3)	88851 (F) (40.7)	223668 (39.9)
CHAINWOO, n=100 (1.0)	>1E6 (F) (3.38E2)	258677 (1.0)	>1E6 (F) (10.06)	84061 (1.0)

the number of function evaluations, the value of the objective function attained by the algorithm at termination is given in parenthesis. An F indicates a solution which differs by more than 20% from the minimum function value or a final solution which is far away from the minimizer.

For all the test problems, SDSA and NSDSA were robust and always found optimum or near-optimum solutions. On the other hand, ROPS and MSD failed on the largest problems, although for two problems ROPS gave a solution with the lowest number of function evaluations. These results are particularly appealing because they show NSDSA to be competitive with derivative-free algorithms designed for smooth functions, which do not share the convergence property to a point satisfying (2.4) for a class of nonsmooth functions. Finally, we conjecture that an adaptive polytope would be an asset for any pattern search algorithm.

6. Conclusion and final remarks. This paper introduces the NSNC (2.4) and a sufficient decrease condition for nonsmooth functions. It also presents a detailed implementation of practical algorithms which, under mild conditions, converge to a stationary point of smooth and nonsmooth functions of practical interest. We visualize our algorithms as new direct search algorithms with the additional feature of allowing a sufficient decrease of function values that still ensure convergence. This is achieved

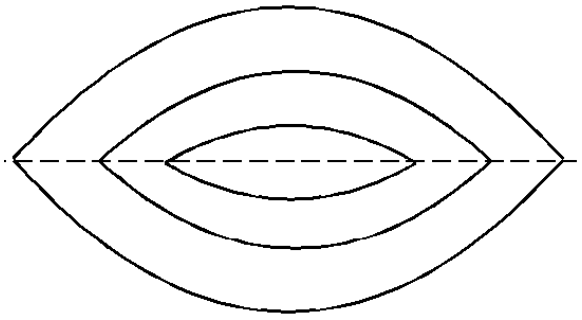


FIG. 6.1. *The Dennis–Woods function.*

by assuming that condition C2 holds globally. Our implementations can be thought of as a simplicial search, with "edges" defined by the directions of the searching set $\mathcal{D}$. This *simplex* is translated to the next iterate. The numerical results reported for sequential algorithms compare favorably with modern derivative-free algorithms recently introduced in the literature.

This paper complements recent work on generalized pattern search methods, while imposing a weaker set of conditions on the trial steps:

- Pattern search methods require a simple decrease. If $f(\cdot) \in C^1$, function values must be computed at all simplex vertices to ensure $\{\nabla f(x_i)\} \to 0$. Our algorithm can go from one "vertex" to the next as soon as it fulfills a sufficient decrease condition.
- Pattern search methods enforce a constant shrinkage/expansion factor for all edges, while ours allows independent shrinkage/expansion factors along the search directions.

Numerical results on the Rosenbrock function and some problems from the CUTE collection seem to indicate that adapting the searching set $\mathcal{D}$ to the direction of movement may have a remarkable effect on the quality of convergence. Other additional features of practical interest in actual implementations are (i) different stepsizes on the directions of search and (ii) the possibility of extracting first-order information that can be used to formulate variable metric algorithms [13].

Algorithms suitable to a multiprocessing environment were also suggested, and their computational performance will be reported in a forthcoming paper.

Either strict differentiability at limit points or convexity in a neighborhood of limit points is required for convergence to an NSNC point. Convergence to a first-order stationary point is ensured in [3] when the function $f(\cdot)$ is strictly differentiable at limit points, or differentiable, Lipschitzian, and regular near limit points. Previous works assumed $f(\cdot) \in C^1$ [24, 41]. Theorem 3.5 is a novel theoretical contribution; it requires local convexity to ensure convergence to a point satisfying (2.4). Actually, there still exists an intriguing gap in theory. No known pattern search method ensures convergence to the minimum of a nonsmooth convex function. We illustrate this with an analysis of the algorithms' behavior on the nonsmooth convex 2-variable Dennis–Woods function [10] (see Figure 6.1):

$$f(x) = \frac{1}{2}\max\left\{||x - c_1||^2, ||x - c_2||^2\right\}, \quad c_1 = (0, 32), \quad c_2 = (0, -32).$$

The origin is the minimizer of this function. If the searching set is defined as

$\mathcal{D} \doteq \{(1,1),(1,-1)\}$, the nonsmooth necessary condition (2.4) is satisfied as well for all points in the set $\mathcal{S} \doteq \{x \in \mathbb{R}^2 : x = (\alpha, 0)\}$ and any α value. Regardless of the initial point, our algorithm always converges to some point in S, which is theoretically what we can hope for. To circumvent this "convergence failure" away from the minimizer, we can randomly generate a new set $\mathcal{D}$ of search directions (or a new polytope) at unspecified blocked points, or we can work with a searching set $\mathcal{D}$ with more than n directions. Extension of multidirectional search to nonsmooth functions was considered in [40, Theorem 7.1], and we might as well expect convergence to (2.4). Indeed, the MDS algorithm always converges to a point in S [40].

Acknowledgments. This paper has been improved with the help of many colleagues and the decisive referees' contribution. Dr. J. Judice, Dr. M. Solodov, Dr. T. Kolda, and one anonymous referee called the authors' attention to recent reports that had not appeared in the open literature while this paper was under review. We are indebted to Dr. C. Audet who meticulously read the paper; he pointed out an error in a previous version and provided us with the excellent example cited in [2]. Dr. E. Hernández helped with the description of the parallel algorithm.

REFERENCES

[1] L. Armijo, *Minimization of functions having Lipschitz continuous first partial derivatives*, Pacific J. Math., 16 (1966), pp. 1–3.

[2] C. Audet, *A Counter-Example for the Derivative-Free Algorithm of García-Palomares and Rodríguez*, personal communication, École Polytechnique de Montréal and GERAD, Département de Mathématiques et de Génie Industriel, Montreal, 2000.

[3] C. Audet and J.E. Dennis, *Analysis of Generalized Pattern Searches*, Technical report TR00-07, Department of Computational and Applied Mathematics, Rice University, Houston, TX, 2000.

[4] M. Avriel, *Nonlinear Programming: Analysis and Methods*, Prentice–Hall, Englewood Cliffs, NJ, 1976.

[5] D.M. Bortz and C.T. Kelley, *The simplex gradient and noisy optimization problems*, in Computational Methods in Optimal Design and Control, J.T. Borggaard et al., eds., Birkhaüser Boston, Cambridge, MA, 1998, pp. 77–90.

[6] J. Brauninger, *A variable metric algorithm for unconstrained minimization without evaluation of derivatives*, Numer. Math., 36 (1981), pp. 359–373.

[7] F.H. Clarke, *Optimization and Nonsmooth Analysis*, Classics Appl. Math. 5, SIAM, Philadelphia, 1990.

[8] A.R. Conn, K. Scheinberg, and Ph.L. Toint, *Recent progress in unconstrained nonlinear optimization without derivatives*, Math. Programming, 79 (1997), pp. 397–414.

[9] J.E. Dennis, Jr., and V. Torczon, *Direct search methods on parallel machines*, SIAM J. Optim., 1 (1991), pp. 448–474.

[10] J.E. Dennis and D.J. Woods, *Optimization on microcomputers: The Nelder-Mead simplex algorithm*, in New Computing Environments: Microcomputers in Large-Scale Computing, A. Wouk, ed., SIAM, Philadelphia, 1987, pp. 116–122.

[11] L.C.W. Dixon, *Neural networks and unconstrained optimization*, in Algorithms for Continuous Optimization: The State of the Art, E. Spedicato, ed., Kluwer Academic Publishers, Norwell, MA, 1994, pp. 513–530.

[12] U.M. García-Palomares, *Análisis y Teorema de Convergencia de un Algoritmo de Minimización sin el Cálculo de Derivadas*, Acta Cient. Venezolana, 27 (1976), pp. 187–189.

[13] U.M. García-Palomares and J.F. Rodríguez, *Second-order Information in the Adaptive Search Exploration Algorithm*, presented at the 8th AIAA/ USAF/NASA/ISSMO/ Symposium on Multidisciplinary Analysis and Optimization, Long Beach, CA, 2000, paper AIAA-2000-4765.

[14] F. Gianessi, *General optimality conditions via a separation scheme*, in Algorithms for Continuous Optimization: The State of the Art, E. Spedicato, ed., Kluwer Academic Publishers, Norwell, MA, 1994, pp. 1–23.

[15] 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.

[16] L.R. Huang and K.F. Ng, *Second-order necessary and sufficient conditions in nonsmooth*

optimization, Math. Programming, 66 (1994), pp. 379–402.
[17] S.L.S. Jacoby, J.S. Kowalik, and J.T. Pizzo, *Iterative Methods for Nonlinear Optimization Problems*, Prentice–Hall, Englewood Cliffs, NJ, 1972.
[18] 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.
[19] C.T. Kelley, *Iterative Methods for Optimization*, Frontiers Appl. Math., SIAM, Philadelphia, 1999.
[20] J.C. Lagarias, J.A. Reeds, M.H. Wright, and P.E. Wright, *Convergence properties of the Nelder–Mead simplex method in low dimensions*, SIAM J. Optim., 9 (1998), pp. 112–147.
[21] R.M. Lewis and V. Torczon, *Rank Ordering and Positive Basis in Pattern Search Algorithms*, Technical report TR96-71, ICASE, Langley Research Center, Hampton, VA, 1996.
[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] R.M. Lewis and V. Torczon, *Pattern search algorithms for bound constrained minimization*, SIAM J. Optim., 9 (1999), pp. 1082–1099.
[24] S. Lucidi and M. Sciandrone, *On the Global Convergence of Derivative Free Methods for Unconstrained Optimization*, Technical report, Università di Roma "La Sapienza", Dipartimento di Informatica e Sistemistica, Rome, 1997.
[25] K.I.M. McKinnon, *Convergence of the Nelder–Mead simplex method to a nonstationary point*, SIAM J. Optim., 9 (1998), pp. 148–158.
[26] R. Mifflin, *A superlinearly convergent algorithm for minimization without evaluating derivatives*, Math. Programming, 9 (1975), pp. 100–117.
[27] J.A. Nelder and R. Mead, *A simplex method for function minimization*, The Computer Journal, 7 (1965), pp. 308–313.
[28] J. Ortega and W.C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.
[29] J.-S. Pang, *Newton's methods for B-differentiable equations*, Math. Oper. Res., 15 (1990), pp. 311–341.
[30] J.-S. Pang, S.-P. Han, and N. Rangaraj, *Minimization of locally Lipschitzian functions*, SIAM J. Optim., 1 (1991), pp. 57–82.
[31] M.J.D. Powell, *An efficient method of finding the minimum of a function of several variables without calculating derivatives*, The Computer Journal, 7 (1964), pp. 155–162.
[32] B.N. Psenichny, *A method of minimizing functions without computing derivatives*, Dokl. Akad. SSSR, 235 (1977), pp. 1097–1100.
[33] L. Qi, A. Ruszczyński, and R. Womersley, eds., *Computational Nonsmooth Optimization*, Mathematical Programming Series B 3-1, Elsevier Science, New York, 1997.
[34] S.M. Robinson, *Local structure of feasible sets in nonlinear programming, Part* III*: Stability and sensitivity*, Mathematical Programming Study, 30 (1987), pp. 45–66.
[35] R.T. Rockafellar, *Convex Analysis*, Princeton Math. Ser. 28, 2nd ed., Princeton University Press, Princeton, NJ, 1972.
[36] A.S. Rykov, *Simplex algorithms for unconstrained optimization*, Probl. Control Inform. Theory, 12 (1983), pp. 195–208.
[37] G.W. Stewart, *A modification of Davidon's method to accept difference approximations of derivatives*, J. ACM, 14 (1967), pp. 72–83.
[38] D. Stoneking, G. Bilbro, R. Trew, P. Gilmore, and C.T. Kelley, *Yield optimization using a GaAs process simulator coupled to a physical device model*, IEEE Trans. Microwave Theory and Techniques, 40 (1992), pp. 1353–1363.
[39] V. Torczon, *Multi-Directional Search: A Direct Search Algorithm for Parallel Machines*, Ph.D. thesis, Department of Mathematical Sciences, Rice University, Houston, TX, 1989.
[40] V. Torczon, *On the convergence of the multidirectional search algorithm*, SIAM J. Optim., 1 (1991), pp. 123–145.
[41] V. Torczon, *On the convergence of pattern search algorithms*, SIAM J. Optim., 7 (1997), pp. 1–25.
[42] P. Tseng, *Fortified-descent simplicial search method: A general approach*, SIAM J. Optim., 10 (1999), pp. 269–288.
[43] M.N. Vrahatis, G.S. Androulakis, and G.E. Manoussakis, *A new unconstrained optimization method for imprecise function and gradient values*, J. Math. Anal. Appl., 197 (1996), pp. 586–607.
[44] G.R. Walsh, *Methods of Optimization*, Wiley and Sons, New York, 1975.
[45] P. Wolfe, *On the convergence of gradient methods under descent*, IBM J. Research and Development, 16 (1972), pp. 407–411.
[46] B. Xiao and P.T. Harker, *A nonsmooth Newton method for variational inequalities,* I*: Theory*, Math. Programming, 65 (1994), pp. 151–194.

SIAM J. OPTIM.
Vol. 13, No. 1, pp. 97–116

ON THE GLOBAL CONVERGENCE OF DERIVATIVE-FREE METHODS FOR UNCONSTRAINED OPTIMIZATION*

STEFANO LUCIDI† AND MARCO SCIANDRONE‡

Abstract. In this paper, starting from the study of the common elements that some globally convergent direct search methods share, a general convergence theory is established for unconstrained minimization methods employing only function values. The introduced convergence conditions are useful for developing and analyzing new derivative-free algorithms with guaranteed global convergence. As examples, we describe three new algorithms which combine pattern and line search approaches.

Key words. unconstrained minimization, derivative-free methods

AMS subject classifications. 90C56, 65K05

PII. S1052623497330392

1. Introduction.

In this paper, we consider the problem of the form

$$\min_{x \in R^n} \quad f(x),$$

where $f : R^n \to R$ is a continuously differentiable function and where the first order derivatives of f can be neither calculated nor approximated explicitly.

The interest in studying minimization algorithms for solving these optimization problems derives from the increasing demand from industrial and scientific applications for such tools. Many derivative-free methods have been proposed in the literature; descriptions of these methods can be found, for instance, in [13] and [19].

An important class of such methods is formed by the so-called *direct search methods*, which base the minimization procedure on the comparison of objective function values computed on suitable trial points. Two particular subclasses of globally convergent direct search methods are the following:

- *pattern search methods* (see, e.g., [2], [6], [16], [19]), which present the distinguishing feature of evaluating the objective function on specified geometric patterns;
- *line search methods* (see, e.g., [1], [4], [5], [8], [10], [11], [12], [17], [20]), which draw their inspiration from the gradient-based minimization methods and perform one dimensional minimizations along suitable directions.

These two classes of methods present different interesting features. In fact, the pattern search methods can accurately sample the objective function in a neighborhood of a point and, hence, can identify a "good" direction, namely, a direction along which the objective function decreases significantly. The line search algorithms can perform large steps along the search directions and, hence, can exploit to a large extent the possible goodness of the directions. Therefore it could be worthwhile to combine

*Received by the editors November 18, 1997; accepted for publication (in revised form) November 28, 2001; published electronically June 5, 2002. This work was supported by Agenzia Spaziale Italiana, Rome, Italy.
http://www.siam.org/journals/siopt/13-1/33039.html
†Dipartimento di Informatica e Sistemistica, Università di Roma "La Sapienza," Via Buonarroti 12 00185 Roma, Italy (lucidi@dis.uniroma1.it).
‡Istituto di Analisi dei Sistemi ed Informatica, CNR, Viale Manzoni 30 00185 Roma, Italy (sciandro@iasi.rm.cnr.it).

these approaches in order to define new classes of derivative-free algorithms that could exploit as much as possible their different features, namely, algorithms which are able to determine "good" directions and to perform "significant" steplengths along such directions. Some examples of methods combining different direct search approaches have already been proposed in [3], [10], [14], [15], [17], [18]. In this paper, on the basis of the convergence analyses reported in [5], [7], and [16] for pattern and line search methods, respectively, we give general sufficient conditions for ensuring the global convergence of a sequence of points. These conditions, which do not require any information on first order derivatives, can be used as the basis for developing new globally convergent derivative-free algorithms and, in particular, algorithms which can follow a mixed pattern-line search approach.

More specifically, in section 2, we start by identifying the common key features of the pattern and line search methods which are behind their global convergence properties. This analysis indicates that the global convergence of a derivative-free algorithm can be guaranteed by satisfying some minimal and quite natural requirements on the search directions used and on the sampling of the objective function along these directions. Then, in section 3, we analyze theoretical requirements regarding the search directions. In section 4, we define general conditions sufficient to ensure global convergence without gradient information. Finally, in section 5, we propose new globally convergent algorithms which combine pattern and line search approaches. The appendix contains the proofs of two technical results.

Notation. We indicate by $\|\cdot\|$ the Euclidean norm (on the appropriate space). A subsequence of $\{x_k\}$ corresponding to an infinite subset K will be denoted by $\{x_k\}_K$. Given two sequences of scalars $\{u_k\}$ and $\{v_k\}$ such that

$$\lim_{k\to\infty} u_k = 0 \qquad \text{and} \qquad \lim_{k\to\infty} v_k = 0,$$

we say that $u_k = o(v_k)$ if

$$\lim_{k\to\infty} \frac{u_k}{v_k} = 0.$$

As usual we say that a set of directions $\{p^1, p^2, \ldots, p^r\}$ positively span R^n if for every $x \in R^n$ there exist $\lambda_i \geq 0$, for $i = 1, \ldots, p$, such that

$$x = \sum_{i=1}^{r} \lambda_i p^i.$$

Finally, we denote by e^i, with $i = 1, \ldots, n$, the orthonormal set of the coordinate directions.

2. Preliminary remarks. It is well known that, when the gradient is available, to define a globally convergent algorithm for unconstrained problems is not a difficult task. In fact, at each iteration, the gradient allows us

- to compute and select a "good" descent direction, namely, a direction along which the objective function decreases with a suitable rate;
- to determine a "sufficiently" large steplength along a descent search direction, namely, a steplength which is able to exploit the descent property of the search direction by enforcing a significant decrease in the value of the objective function relative to the norm of the gradient.

When the gradient is not available, we lose information about the local behavior of the objective function. In fact, the ith component $\nabla_i f$ of the gradient is the directional derivative of the objective function along the vector e^i, and $-\nabla_i f$ is the directional derivative along the vector $-e^i$. Therefore, the whole gradient vector provides the rate of change of the objective function along the $2n$ directions $[e^1, e^2, \ldots, e^n, -e^1, -e^2, \ldots, -e^n]$. This fact guarantees that the gradient information characterizes quite accurately the local behavior of the objective function in a neighborhood of the point at which the derivatives are computed.

Most of the algorithms belonging to the class of direct search methods follow, more or less visibly, the same strategy to overcome the lack of first order information contained in the gradient. Their common approach is based on the idea of investigating the behavior of the objective function in a neighborhood of the generic point by sampling the objective function along a set of directions. Clearly each of these algorithms presents properties and features which depend on the particular choice of the sets of directions and on the particular way in which the samplings of the objective function are performed.

The directions to be used in a derivative-free algorithm should be such that the local behavior of the objective function along them is sufficiently indicative of the local behavior of the function in a neighborhood of a point. Roughly speaking, these directions should have the property that, performing finer and finer samplings of the objective function along them, it is possible either

(i) to realize that the current point is a good approximation of a stationary point of the objective function, or

(ii) to find a specific direction along which the objective function decreases.

The important point is to identify larger and larger classes of sets of search directions which can be used to define globally convergent derivative-free algorithms. To this end, in the next section, we propose a general condition which formally characterizes classes of sets of directions complying with the properties (i) and (ii).

In addition to contributing to the previous points (i) and (ii), the method of sampling has the task of guiding the choice of the new point so as to ensure that the sequence of points produced by the algorithm is globally convergent towards stationary points of the objective function. On the basis of the common features of the sampling techniques of the direct search methods proposed in [5], [7], and [16], in section 4 we define sufficient conditions on the samplings of the objective function along suitable directions for the global convergence of a derivative-free method. Similar conditions were given in [18]; however, the ones proposed in this work are more general.

3. Search directions. Before describing our analysis, we recall the following basic assumption on the objective function.

Assumption A1. The function $f : R^n \to R$ is continuously differentiable.

As said before, the first step in defining a direct search method is to associate a suitable set of search directions p_k^i, $i = 1, \ldots, r$, with each point x_k produced by the algorithm. This set of directions should have the property that the local behavior of the objective function along them provides sufficient information to overcome the lack of the gradient.

Here, we introduce a new condition which characterizes the sets of directions p_k^i, $i = 1, \ldots, r$, that satisfy this property. This condition requires that the distance

between the points generated by an algorithm and the set of stationary points of the objective function tends to zero if and only if the directional derivatives of the objective function along the directions p_k^i, $i = 1, \ldots, r$, tend to assume nonnegative values. Formally we have the following condition.

Condition C1. Given a sequence of points $\{x_k\}$, the sequence directions $\{p_k^i\}$, $i = 1, \ldots, r$, are bounded and such that

$$\lim_{k\to\infty} \|\nabla f(x_k)\| = 0 \quad \text{if and only if} \quad \lim_{k\to\infty} \sum_{i=1}^{r} \min\{0, \nabla f(x_k)^T p_k^i\} = 0.$$

By drawing our inspiration from some results established in [7] and [16], we state the following proposition, which points out a possible interest in the sets of directions satisfying Condition C1.

PROPOSITION 3.1. *Let $\{x_k\}$ be a bounded sequence of points and let $\{p_k^i\}$, $i = 1, \ldots, r$, be sequences of directions which satisfy Condition* C1. *For every $\eta > 0$, there exist $\gamma > 0$ and $\delta > 0$ such that, for all but finitely many k, if x_k satisfies $\|\nabla f(x_k)\| \geq \eta$, then there exists a direction $p_k^{i_k}$, with $i_k \in \{1, \ldots, r\}$, for which*

$$f(x_k + \alpha p_k^{i_k}) \leq f(x_k) - \gamma\alpha \|\nabla f(x_k)\| \|p_k^{i_k}\| \tag{3.1}$$

for all $\alpha \in (0, \delta]$.

Proof. For the proof, see the appendix. ∎

The previous proposition guarantees that, whenever the current point is not a stationary point, it is possible to enforce sufficient decrease of the objective function by using sets of directions satisfying Condition C1. In other words, this ensures that Condition C1 implies that the sets of directions are able to comply with the requirement (ii) discussed in section 2.

From a theoretical point of view, Proposition 4.1, given in the next section, shows that Condition C1 is a sufficient requirement for the search directions to ensure the global convergence of the sequence of iterates (or at least one subsequence) to a stationary point of f. Roughly speaking, the role of Condition C1 in the field of derivative-free methods can be similar to that of the *gradient-related condition* used in the field of gradient-based algorithms. In fact, Condition C1 can be considered a mild technical condition on the sets of search directions which can be either naturally satisfied or easily enforced in a derivative-free algorithm (see Algorithm 3 in section 5).

In order to show that Condition C1 is a viable requirement on the search directions, we report two classes of sets of directions satisfying Condition C1 and some examples of these classes. The classes introduced here generalize the ones proposed in [7].

Classes of sets of search directions.

(a) The sequences $\{p_k^i\}$, with $i = 1, \ldots, r$, are bounded, and every limit point $(\bar{p}^1, \ldots, \bar{p}^r)$ of the sequence $\{p_k^1, \ldots, p_k^r\}$ is such that the vectors $\bar{p}^i$, with $i = 1, \ldots, r$, positively span R^n.

(b) The sequences $\{p_k^i\}$, with $i = 1, \ldots, r$, are bounded; the vectors p_k^i, $i = 1, \ldots, n$, are uniformly linearly independent; and, for all k, there exists a

direction p_k^{n+j}, with $j \geq 1$, given by

$$p_k^{n+j} = \sum_{\ell=1}^{2n} \rho_k^\ell \frac{(v_k^1 - v_k^\ell)}{\tilde{\xi}_k^\ell}, \tag{3.2}$$

where

- the sequences $\{\rho_k^\ell\}$, $\ell = 1, \ldots, 2n$, are bounded and such that $\rho_k^\ell \geq 0$ with $\rho_k^{2n} \geq \bar{\rho} > 0$ for all k;
- $\{v_k^1, v_k^2, \ldots, v_k^{2n}\} = \{z_k^1, z_k^1 + \xi_k^1 p_k^1, z_k^2, z_k^2 + \xi_k^2 p_k^2, \ldots, z_k^n, z_k^n + \xi_k^n p_k^n\}$, with the points v_k^ℓ, for $\ell = 1, \ldots, 2n$, ordered (and possibly relabeled) so that

$$f(v_k^1) \leq f(v_k^2) \leq \cdots \leq f(v_k^{n-1}) \leq \cdots \leq f(v_k^{2n}), \tag{3.3}$$

and the sequences $\{\xi_k^i\}$ and $\{z_k^i\}$, for $i = 1, \ldots, n$, are such that, for all k,

$$\xi_k^i > 0, \tag{3.4}$$

$$\frac{\max\limits_{i=1,\ldots,n} \{\xi_k^i\}}{\min\limits_{i=1,\ldots,n} \{\xi_k^i\}} \leq c_1, \tag{3.5}$$

$$\|z_k^i - x_k\| \leq c_2 \xi_k^i, \tag{3.6}$$

where c_1,$c_2 > 0$, and such that

$$\lim_{k \to \infty} \xi_k^i = 0; \tag{3.7}$$

- the sequences $\{\tilde{\xi}_k^\ell\}$, $\ell = 1, \ldots, 2n$, are such that $\min_{i=1,\ldots,n}\{\xi_k^i\} \leq \tilde{\xi}_k^\ell \leq \max_{i=1,\ldots,n}\{\xi_k^i\}$.

For the classes of sets of search directions we can state the following proposition.

PROPOSITION 3.2. *Let $\{x_k\}$ be a bounded sequence of points, and let $\{p_k^i\}$, $i = 1, \ldots, r$, be sequences of directions belonging to class* (a) *or class* (b). *Then, Condition* C1 *is satisfied.*

Proof. For the proof, see the appendix. □

Two examples of sets of directions belonging to the classes (a) and (b) are described in [7]. These classes are defined starting from a set of n uniformly linearly independent search directions, for example,

$$p_k^1 = e^1, \qquad p_k^2 = e^2, \qquad \ldots, \qquad p_k^n = e^n. \tag{3.8}$$

Then, to obtain a set of class (a), it is sufficient to consider also the directions

$$p_k^{n+1} = -e^1, \qquad p_k^{n+2} = -e^2, \qquad \ldots, \qquad p_k^{2n} = -e^n$$

or just the direction

$$p_k^{n+1} = -\sum_{i=1}^{n} e^i.$$

A set of class (b) can be obtained by adding to (3.8) the direction

$$p_k^{n+1} = \frac{x_k - x_k^{max}}{\xi_k},$$

where $x_k^{max} = \arg\max_{i=1,\dots,n}\{f(x_k+\xi_k p_k^i)\}$ and $\xi_k \to 0$ for $k \to \infty$. This corresponds to setting

$$z_k^i = x_k, \qquad \xi_k^i = \xi_k \qquad \text{for } i = 1, \dots, n,$$

$$\xi_k^{2n} = \xi_k,$$

$$\rho_k^1 = \rho_k^2 = \cdots = \rho_k^{2n-1} = 0, \qquad \rho_k^{2n} = 1.$$

A new class of sets of search directions satisfying Condition C1 will be defined within Algorithm 3 proposed in section 5. In particular, this class is constructed during the minimization procedure so as to exploit as much as possible the information on the objective function obtained in the preceding iterations.

4. Global convergence conditions. In this section we show that the global convergence of an algorithm can be guaranteed by means of the existence of suitable sequences of points along search directions p_k^i, $i = 1, \dots, r$, satisfying Condition C1. In particular, by using Condition C1 we can characterize a stationary point of f with the fact that the objective function does not decrease locally along the directions p_k^i, $i = 1, \dots, r$, in points sufficiently close to the current point x_k. This leads to the possibility of defining new general conditions for the global convergence of derivative-free algorithms by means of the existence of sequences of points showing that the objective function does not decrease along the directions p_k^i, $i = 1, \dots, r$. These conditions, even if very simple and intuitive, allow us to identify some minimal requirements on acceptable samplings of the objective function along the directions p_k^i, $i = 1, \dots, r$, that guarantee the global convergence of the method.

In the remainder of the paper we suppose that the following standard assumption holds.

Assumption A2. The level set

$$\mathcal{L}_0 = \{x \in R^n : f(x) \le f(x_0)\}$$

is compact.

The following proposition describes a set of global convergence conditions.

PROPOSITION 4.1. *Let $\{x_k\}$ be a sequence of points; let $\{p_k^i\}$, $i = 1, \dots, r$, be sequences of directions; and suppose that the following conditions hold:*

(a) $f(x_{k+1}) \le f(x_k)$;
(b) $\{p_k^i\}$, $i = 1, \dots, r$, *satisfy Condition* C1;
(c) *there exist sequences of points $\{y_k^i\}$ and sequences of positive scalars $\{\xi_k^i\}$, for $i = 1, \dots, r$, such that*

$$f(y_k^i + \xi_k^i p_k^i) \ge f(y_k^i) - o(\xi_k^i), \tag{4.1}$$

$$\lim_{k\to\infty} \xi_k^i = 0, \tag{4.2}$$

$$\lim_{k\to\infty} \|x_k - y_k^i\| = 0. \tag{4.3}$$

Then,

$$\lim_{k\to\infty} \|\nabla f(x_k)\| = 0. \tag{4.4}$$

Proof. From (a) it follows that $\{f(x_k)\}$ is a nonincreasing sequence, so that $\{x_k\}$ belongs to the compact set $\mathcal{L}_0$ and admits at least one limit point. Let $\bar{x}$ be any limit point of $\{x_k\}$. Then, there exists a subset $K_1 \subseteq \{0, 1, \ldots\}$ such that

$$\lim_{k\to\infty, k\in K_1} x_k = \bar{x},$$

$$\lim_{k\to\infty, k\in K_1} p_k^i = \bar{p}^i, \quad i = 1, \ldots, r.$$

Using (4.3), it follows that

$$\lim_{k\to\infty, k\in K_1} y_k^i = \bar{x}, \qquad i = 1, \ldots, r.$$

Now, recalling (4.1) for all $k \geq 0$, we have

$$f(y_k^i + \xi_k^i p_k^i) - f(y_k^i) \geq -o(\xi_k^i), \quad i = 1, \ldots, r. \tag{4.5}$$

By the mean-value theorem, we can write

$$f(y_k^i + \xi_k^i p_k^i) - f(y_k^i) = \xi_k^i \nabla f(u_k^i)^T p_k^i, \quad i = 1, \ldots, r, \tag{4.6}$$

where $u_k^i = y_k^i + \lambda_k^i \xi_k^i p_k^i$, with $\lambda_k^i \in (0, 1)$. By substituting (4.6) into (4.5), we obtain

$$\nabla f(u_k^i)^T p_k^i \geq -o(\xi_k^i), \quad i = 1, \ldots, r. \tag{4.7}$$

Now, it is easily seen from (4.2), taking into account the boundedness of p_k^i, that $u_k^i \to \bar{x}$ as $k \to \infty$ and $k \in K_1$. Hence, by the continuity of ∇f, from (4.7) and recalling (4.2), we get

$$\lim_{k\to\infty, k\in K_1} \nabla f(u_k^i)^T p_k^i = \nabla f(\bar{x})^T \bar{p}_i \geq 0, \quad i = 1, \ldots, r.$$

Then, recalling (b) and Condition C1, we have that

$$\nabla f(\bar{x}) = 0.$$

As $\bar{x}$ is any limit point of $\{x_k\}$, we conclude that

$$\lim_{k\to\infty} \|\nabla f(x_k)\| = 0. \qquad \square$$

Roughly speaking, according to (c), for each search direction p_k^i, the existence of suitable points y_k^i and $y_k^i + \xi_k^i p_k^i$ related to the "current" point x_k is assumed (see (4.2) and (4.3)) whenever a "failure" of a (sufficient) strict decrease of f occurs (see (4.1)).

Then, also considering (4.2), we have that at the point y_k^i the directional derivative of f along p_k^i can be approximated by a quantity which tends to be nonnegative. Therefore, due to the property of the search directions expressed by Condition C1, the global convergence of the sequence $\{x_k\}$ can be ensured by requiring that the failure points "cluster" more and more around x_k (see (4.3)).

Similar conditions were given in [18]; however, those of Proposition 4.1 are more general in the requirements placed on both the search directions p_k^i and the trial steps ξ_k^i, $i = 1, \ldots, r$.

The use of directions satisfying Condition C1 and the result of producing sequences (or subsequences) of points that satisfy the hypothesis of Proposition 4.1 are the common elements of the globally convergent derivative-free algorithms proposed in [5] and [16], which consider the pattern and line search approaches, respectively. This point is discussed in more detail in [9], where the known global convergence results of different algorithms are reobtained by using Condition C1 and Proposition 4.1. With regard to (4.1) and (4.2) of Proposition 4.1(c), we note only that

- in the pattern search algorithms, the failures (4.1) (with $o(\xi_k^i) = 0$) occur "naturally" by requiring only a simple decrease of f, while (4.2) follows by imposing further restrictions on the search directions and on the steplengths;
- in the line search algorithms, (4.1) and (4.2) are satisfied by enforcing a "sufficient" decrease of f depending on ξ_k^i and without imposing further restrictions on the search directions.

5. New globally convergent algorithms. In this section we try to motivate further the possible practical interest of the analysis performed in sections 3 and 4, by showing that Condition C1 and Proposition 4.1 can play the role of guidelines for defining new derivative-free algorithms and for analyzing their convergence properties.

Since the conditions given in Proposition 4.1 capture some common theoretical features of pattern and line search approaches, they are suitable for defining algorithms which combine these two approaches. In particular, our aim is to propose algorithms which are able to

- get sufficient information on the local behavior of the objective function f, like in a pattern strategy;
- exploit the possible knowledge of a "good" direction, like in a line search strategy.

In this section, as examples, we describe three new algorithms (Algorithm 1, Algorithm 2, and Algorithm 3). The basic idea of these algorithms is to sample, at each iteration k, the objective function f along a set $\{p_k^i\}_{i=1}^r$ of search directions. This is performed with the aim of detecting a "promising" direction (like in a pattern strategy), that is, a direction along which the objective function decreases "sufficiently." Then, once such a direction has been detected, a "sufficiently" large step is performed along it. Both the "sufficient" decrease of the objective function and the "sufficient" steplength are evaluated by means of criteria derived from the line search approach. These criteria, requiring sufficient decrease of the objective function, are stronger than the ones used in the pattern search algorithms (where the simple reduction of f is allowed). However, as we said before, they allow us more freedom in the choice of search directions and in the steplengths used to sample the objective function.

In particular, in Algorithm 1 and Algorithm 2, we assume that the sets of search directions satisfying Condition C1 are given. Algorithm 1 is very simple, and its scheme is similar to that of a pattern search algorithm. For this algorithm we can prove that at least one accumulation point of the sequence produced is a stationary

point of f. In Algorithm 2 a line search technique is introduced to exploit as much as possible a promising direction identified by the algorithm. For this algorithm we prove that any convergent subsequence generated by the algorithm tends to a stationary point of f. The approach of Algorithm 3 is the same as that of Algorithm 2; the distinguishing feature of Algorithm 3 is that of using sets of $n+1$ directions, in which the first n are given and the last one is computed on the basis of the information iteratively obtained with the aim of identifying a "good" direction. For this algorithm we prove the same convergence result stated for Algorithm 2.

The first algorithm is the following.

ALGORITHM 1.

Data. $x_0 \in R^n$, $\tilde{\alpha}_0 > 0$, $\gamma > 0$, $\theta \in (0,1)$.

Step 0. Set $k = 0$.

Step 1. If there exists $y_k \in R^n$ such that

$$f(y_k) \leq f(x_k) - \gamma \tilde{\alpha}_k,$$

then go to Step 4.

Step 2. If there exists $i \in \{1, \dots r\}$ and an $\alpha_k \geq \tilde{\alpha}_k$ such that

$$f(x_k + \alpha_k p_k^i) \leq f(x_k) - \gamma (\alpha_k)^2,$$

then set $y_k = x_k + \alpha_k p_k^i$, $\tilde{\alpha}_{k+1} = \alpha_k$ and go to Step 4.

Step 3. Set $\tilde{\alpha}_{k+1} = \theta \tilde{\alpha}_k$ and $y_k = x_k$.

Step 4. Find x_{k+1} such that $f(x_{k+1}) \leq f(y_k)$, set $k = k + 1$, and go to Step 1.

Algorithm 1 follows an approach similar to that of a pattern search algorithm. In particular, at each iteration it is possible to accept any single point for which sufficient decrease of the objective function is realized (Step 1). The stepsize α_k is reduced only when it is not possible to locally enforce the sufficient reduction of f along the search directions p_k^i, for $i = 1, \dots, r$ (Steps 2–3). At Step 4 the algorithm can accept any point which produces an improvement of the objective function with respect to the selected point y_k.

We note that, at Step 2, any extrapolation technique can be attempted to determine a good stepsize α_k whenever a suitable direction has been detected. However, the use of an extrapolation technique is not necessary to guarantee global convergence. (In particular, it is enough to use $\alpha_k = \tilde{\alpha}_k$.) Furthermore, we point out that, even if a set of r search directions p_k^i, $i = 1, \dots, r$, is associated to the current point x_k, so long as a sufficient decrease condition has been satisfied along a direction $p_k^{\bar{i}}$, the remaining directions can be ignored. This is a feature that Algorithm 1 has in common with the weak form of pattern search algorithms (see [16]).

Finally, we observe also that Step 1 and Step 4 allow the possibility of using any approximation scheme for the objective function to produce a new better point.

The convergence properties of the algorithm are reported in the following proposition.

PROPOSITION 5.1. *Let $\{x_k\}$ be the sequence produced by Algorithm* 1. *Suppose*

that the sequences of directions $\{p_k^i\}_{i=1}^r$ *satisfy Condition* C1. *Then we have*

$$\liminf_{k\to\infty} \|\nabla f(x_k)\| = 0. \tag{5.1}$$

Proof. We prove (5.1) by showing that conditions (a), (b), and (c) of Proposition 4.1 are satisfied (at least) by a subsequence of $\{x_k\}$.

Condition (a) follows from the instructions of the algorithm. Condition (b) is obviously true. Therefore we concentrate on Condition (c).

We can split the iteration sequence $\{k\}$ into three parts, K_1, K_2, and K_3, namely, those iterations where the test at Step 1 is satisfied, those where the test at Step 2 is satisfied, and those where Step 3 is performed. In particular, if $k \in K_1$, we have

$$f(x_{k+1}) \le f(x_k) - \gamma\tilde{\alpha}_k; \tag{5.2}$$

if $k \in K_2$, we have

$$f(x_{k+1}) \le f(x_k) - \gamma(\alpha_k)^2 \le f(x_k) - \gamma(\tilde{\alpha}_k)^2; \tag{5.3}$$

and if $k \in K_3$, we have

$$f(x_k + \tilde{\alpha}_k p_k^i) > f(x_k) - \gamma(\tilde{\alpha}_k)^2 \qquad \text{for } i = 1, \dots, r. \tag{5.4}$$

If K_1 is an infinite subset, then (5.2), the compactness of the level set $\mathcal{L}_0$, the continuity assumption on f, and Condition (a) imply

$$\lim_{k\to\infty, k\in K_1} \tilde{\alpha}_k = 0. \tag{5.5}$$

Now, let us assume that K_2 is an infinite subset. From (5.3), by repeating the same reasoning, we obtain

$$\lim_{k\to\infty, k\in K_2} \tilde{\alpha}_k = 0. \tag{5.6}$$

Now for each $k \in K_3$ let m_k be the biggest index such that $m_k < k$ and $m_k \in K_1 \cup K_2$. Then we have

$$\tilde{\alpha}_{k+1} = \theta^{k-m_k}\tilde{\alpha}_{m_k}. \tag{5.7}$$

(We can assume that $m_k = 0$ if the index m_k does not exist; that is, K_1 and K_2 are empty.)

As $k \to \infty$ and $k \in K_3$, either $m_k \to \infty$ (if $K_1 \cup K_2$ is an infinite subset) or $(k - m_k) \to \infty$ (if $K_1 \cup K_2$ is finite). Therefore, (5.7) together with (5.5) and (5.6) or the fact that $\theta \in (0, 1)$ yields

$$\lim_{k\to\infty, k\in K_3} \tilde{\alpha}_k = 0. \tag{5.8}$$

Thus, by using (5.5), (5.6), and (5.8), we can write

$$\lim_{k\to\infty} \tilde{\alpha}_k = 0. \tag{5.9}$$

From (5.9) it follows that there exists an infinite subset $K \subseteq \{0, 1, \dots\}$ such that $\tilde{\alpha}_{k+1} < \tilde{\alpha}_k$ for all $k \in K$; namely, Step 3 is performed for all $k \in K$. Therefore, we

have $K \subseteq K_3$, and hence (5.4) holds for all $k \in K$. Now, with reference to condition (c) of Proposition 4.1, for each $k \in K$ we set

$$\xi_k^i = \tilde{\alpha}_k, \quad y_k^i = x_k, \qquad i = 1, \ldots, r. \tag{5.10}$$

Then we have

$$f(y_k^i + \xi_k^i p_k^i) \geq f(y_k^i) - \gamma \left(\xi_k^i\right)^2;$$

moreover, recalling (5.9), it follows that

$$\lim_{k \to \infty, k \in K} \xi_k^i = 0,$$

so that (4.1) and (4.2) hold. Finally, (4.3) follows directly from (5.10), and this concludes the proof. □

Now we define pattern-line search algorithms producing sequences of points with the stronger property that every limit point is a stationary point of f. This additional property can be obtained by investigating in more detail the behavior of the objective function along the search directions p_k^i, $i = 1, \ldots, r$, and by using a derivative-free line search technique to ensure sufficiently large movements along any "good" direction identified by the algorithm. The first of these algorithms is the following.

ALGORITHM 2.

Data. $x_0 \in R^n$, $\tilde{\alpha}_0^i > 0$, $i = 1, \ldots, r$, $\gamma > 0$, $\delta, \theta \in (0, 1)$.
Step 0. Set $k = 0$.
Step 1. Set $i = 1$ and $y_k^1 = x_k$.
Step 2. If $f(y_k^i + \tilde{\alpha}_k^i p_k^i) \leq f(y_k^i) - \gamma(\tilde{\alpha}_k^i)^2$, then
compute α_k^i by *LS Procedure*$(\tilde{\alpha}_k^i, y_k^i, p_k^i, \gamma, \delta)$
and set $\tilde{\alpha}_{k+1}^i = \alpha_k^i$;
else set $\alpha_k^i = 0$ and $\tilde{\alpha}_{k+1}^i = \theta \tilde{\alpha}_k^i$.
Set $y_k^{i+1} = y_k^i + \alpha_k^i p_k^i$.
Step 3. If $i < r$, set $i = i + 1$ and go to Step 2.
Step 4. Find x_{k+1} such that

$$f(x_{k+1}) \leq f(y_k^{r+1}),$$

set $k = k + 1$, and go to Step 1.

LS PROCEDURE$(\tilde{\alpha}_k^i, y_k^i, p_k^i, \gamma, \delta)$.

Compute $\alpha_k^i = \min\{\delta^{-j} \tilde{\alpha}_k : j = 0, 1, \ldots\}$ such that

$$f(y_k^i + \alpha_k^i p_k^i) \leq f(x_k) - \gamma\left(\alpha_k^i\right)^2, \tag{5.11}$$

$$f\left(y_k^i + \frac{\alpha_k^i}{\delta} p_k^i\right) \geq \max\left[f(y_k^i + \alpha_k^i p_k^i), f(y_k^i) - \gamma\left(\frac{\alpha_k^i}{\delta}\right)^2\right]. \tag{5.12}$$

At each iteration k the algorithm examines the behavior of the objective function along all the search directions p_k^i, $i = 1, \ldots, r$ (Steps 1–3). However, whenever it detects a direction p_k^i where the function is sufficiently decreased, the algorithm produces a new point by performing a "sufficiently" large movement along this direction. This point is determined by means of a suitable stepsize α_k^i computed by a line search technique (LS Procedure). At Step 4, similarly to Algorithm 1, the new point x_{k+1} can be the point y_k^{r+1} produced by Steps 1–3 or any point where the objective function is improved with respect to $f(y_k^{r+1})$. This fact, as said before, allows us to adopt any approximation scheme for the objective function to produce a new better point and hence to improve the efficiency of the algorithm without affecting its convergence properties.

Comparing Algorithms 1 and 2, it is easy to observe that Algorithm 2 requires stronger conditions to produce the new point. In fact, all the directions must be investigated at each iteration, and the use of a line search technique is necessary. However, in Algorithm 2 it is possible to associate to each direction p_k^i a different initial stepsize $\tilde{\alpha}_k^i$, which is updated on the basis of the behavior of the objective function along p_k^i observed in the current iteration. This feature can be useful when the search directions are the same for all iterations ($p_k^i = \bar{p}^i$, $i = 1, \ldots, r$, for all k). In fact, in this case, the instructions of the algorithm should guarantee that the initial stepsizes $\tilde{\alpha}_k^i$, $i = 1, \ldots, r$, take into account the different behavior of f along different search directions.

Finally, we note that Algorithm 2, similarly to the strong form of pattern search algorithms, is required to examine, at each iteration, the local behavior of f along all the r directions p_k^i, $i = 1, \ldots, r$. However, at each iteration the current point x_k is updated by means of intermediate points y_k^{i+1} whenever sufficient decrease of f is obtained along any of the search directions p_k^i, $i \in \{1, \ldots, r\}$.

From a theoretical point of view, it is possible to state the following convergence result, which is stronger than the one obtained for Algorithm 1.

PROPOSITION 5.2. *Let $\{x_k\}$ be the sequence produced by Algorithm* 2. *Suppose that the sequences of directions $\{p_k^i\}_{i=1}^r$ satisfy Condition* C1. *Then, Algorithm* 2 *is well defined and we have*

$$\lim_{k \to \infty} \|\nabla f(x_k)\| = 0. \tag{5.13}$$

Proof. In order to prove that Algorithm 2 is well defined, we must show that, given an integer $i \leq r$ such that the test of Step 2 is satisfied, there exists a finite integer j for which (5.11) and (5.12) hold with $\alpha_k^i = \delta^{-j} \tilde{\alpha}_k^i$. With this goal, we give a proof by contradiction. We assume that either

$$f(y_k^i + \delta^{-j} \tilde{\alpha}_k^i p_k^i) < f(y_k^i) - \gamma \big(\delta^{-j} \tilde{\alpha}_k^i\big)^2 \qquad \text{for all } j$$

or

$$f(y_k^i + \delta^{-j-1} \tilde{\alpha}_k^i p_k^i) < f(y_k^i + \delta^{-j} \tilde{\alpha}_k^i p_k^i) \leq f(y_k^i) - \gamma \big(\delta^{-j} \tilde{\alpha}_k^i\big)^2 \qquad \text{for all } j.$$

Then, taking the limits for $j \to \infty$, we obtain in both cases that f is unbounded below, which contradicts Assumption A2.

Now we prove (5.13) by showing that conditions (a), (b), and (c) of Proposition 4.1 are satisfied.

Condition (a) follows from the instructions of the algorithm. Condition (b) is obviously true. Then we must show that condition (c) holds.

We first prove that for $i = 1, \ldots, r$ we have

$$\lim_{k\to\infty} \alpha_k^i = 0 \tag{5.14}$$

and

$$\lim_{k\to\infty} \tilde{\alpha}_k^i = 0. \tag{5.15}$$

From the instructions of the algorithm we have

$$f(x_{k+1}) \le f(y_k^{r+1}) \le f(x_k) - \gamma \sum_{i=1}^{r} \left(\alpha_k^i\right)^2,$$

so that, since $\{x_k\}$ belongs to the compact set $\mathcal{L}_0$, $\{f(x_k)\} \to \bar{f}$ and hence $\alpha_k^i \to 0$, for $i = 1, \ldots r$. Given $i \in \{1, \ldots r\}$, we split the iteration sequence $\{k\}$ into two parts, K and $\bar{K}$, namely, those iterations where $\alpha_k^i > 0$ and those where $\alpha_k^i = 0$. For all $k \in K$ we have $\alpha_k^i \ge \tilde{\alpha}_k^i$, so that, if K is an infinite subset, it follows that

$$\lim_{k\to\infty, k\in K} \tilde{\alpha}_k^i = 0. \tag{5.16}$$

For each $k \in \bar{K}$, let m_k be the biggest index such that $m_k < k$ and $m_k \in K$. (We can assume $m_k = 0$ if the index m_k does not exist, that is, K is empty.) Then we have

$$\tilde{\alpha}_k^i = (\theta)^{k-m_k} \tilde{\alpha}_{m_k}^i.$$

As $k \to \infty$ and $k \in \bar{K}$, either $m_k \to \infty$ (if K is an infinite subset) or $(k - m_k) \to \infty$ (if K is finite). Therefore, (5.16) and the fact that $\theta \in (0, 1)$ imply $\lim_{k\to\infty, k\in\bar{K}} \tilde{\alpha}_k^i = 0$. Now, with reference to condition (c) of Proposition 4.1, we set

$$\xi_k^i = \begin{cases} \dfrac{\alpha_k^i}{\delta} & \text{if } k \in K, \\ \tilde{\alpha}_k^i & \text{if } k \in \bar{K}. \end{cases} \tag{5.17}$$

Then, we have $f(y_k^i + \xi_k^i p_k^i) \ge f(y_k^i) - (\xi_k^i)^2$; moreover, recalling (5.14) and (5.15), it follows that $\lim_{k\to\infty} \xi_k^i = 0$, so that (4.1) and (4.2) hold. Finally, since we have

$$\|y_k^i - x_k\| \le \sum_{j=1}^{i-1} \alpha_k^j \|p_k^j\|,$$

by again using (5.14) it follows that

$$\lim_{k\to\infty} \|x_k - y_k^i\| = 0, \tag{5.18}$$

so that even (4.3) is satisfied and this concludes the proof. □

Remark. By the proof of Proposition 5.2, in particular by (5.18), we note also that

$$\lim_{k\to\infty} \|\nabla f(y_k^i)\| = 0 \qquad \text{for } i = 1, \ldots, r+1. \qquad \square$$

We conclude this section by describing Algorithm 3. This algorithm and Algorithm 2 differ only in their search directions. In particular, we recall that in Algorithm

2 the sets of directions $\{p_k^i\}_{i=1}^r$ satisfying Condition C1 are given. In Algorithm 3 we instead assume that, at each iteration, only n linearly independent directions are given. Then the algorithm, on the basis of the behavior of the objective function along these directions, determines a further direction that should have a good descent property and that is able (with the other directions) to ensure the global convergence of the sequence produced.

ALGORITHM 3.

Data. $x_0 \in R^n$, $c > 0$, $\tilde{\alpha}_0^i > 0$, $i = 1, \ldots, n+1$, $\gamma > 0$, $\delta, \theta \in (0,1)$.

Step 0. Set $k = 0$.

Step 1. Set $i = 1$, $y_k^1 = x_k$, $V_k = \{y_k^1\}$, $S_k = \{\emptyset\}$.

Step 2. If $f(y_k^i + \tilde{\alpha}_k^i p_k^i) \le f(y_k^i) - \gamma(\tilde{\alpha}_k^i)^2$, then
compute α_k^i by *LS Procedure*$(\tilde{\alpha}_k^i, y_k^i, p_k^i, \gamma, \delta)$ and
set $\tilde{\alpha}_{k+1}^i = \alpha_k^i$, $V_k = V_k \cup \{y_k^i + \alpha_k^i p_k^i\}$,
$S_k = S_k \cup \{\alpha_k^i\}$;
else set $\alpha_k^i = 0$, $\tilde{\alpha}_{k+1}^i = \theta \tilde{\alpha}_k^i$, $V_k = V_k \cup \{y_k^i + \tilde{\alpha}_k^i p_k^i\}$,
$S_k = S_k \cup \{\tilde{\alpha}_k^i\}$.
Set $y_k^{i+1} = y_k^i + \alpha_k^i p_k^i$.

Step 3. If $i < n$, set $i = i + 1$ and go to Step 2.

Step 4. Compute $\alpha_k^{min} = \min_{\alpha \in S_k}\{\alpha\}$ and $\alpha_k^{max} = \max_{\alpha \in S_k}\{\alpha\}$.
If $\frac{\alpha_k^{max}}{\alpha_k^{min}} \le c$, then compute p_k^{n+1} such that

$$p_k^{n+1} = \frac{v_k^{max} - v_k^{min}}{\xi_k},$$

where $v_k^{max} = \arg\max_{v \in V_k}\{f(v)\}$,
$v_k^{min} = \arg\min_{v \in V_k}\{f(v)\}$, and
$\xi_k \in [\alpha_k^{min}, \alpha_k^{max}]$;
else set $p_k^{n+1} = -\sum_{i=1}^n p_k^i$.

Step 5. If $f(y_k^n + \tilde{\alpha}_k^{n+1} p_k^{n+1}) \le f(y_k^n) - \gamma(\tilde{\alpha}_k^{n+1})^2$, then
compute α_k^{n+1} by *LS Procedure*$(\tilde{\alpha}_k^i, y_k^i, p_k^{n+1}, \gamma, \delta)$
and set $\tilde{\alpha}_{k+1}^{n+1} = \alpha_k^{n+1}$;
else set $\alpha_k^{n+1} = 0$ and $\tilde{\alpha}_{k+1}^{n+1} = \theta \tilde{\alpha}_k^{n+1}$.
Set $y_k^{n+1} = y_k^n + \alpha_k^{n+1} p_k^{n+1}$.

Step 6. Find x_{k+1} such that

$$f(x_{k+1}) \le f(y_k^{n+1}),$$

set $k = k + 1$, and go to Step 1.

Steps 1–3 are essentially the same as those of Algorithm 2. In these steps the algorithm produces the points y_k^i, with $i = 1, \ldots, n$, by examining the behavior of the objective function along the linearly independent directions p_k^i, with $i = 1, \ldots, n$. At Step 4 we check whether the steplengths used to sample the objective function along the n directions have been "sufficiently regular," namely, whether the ratio between the biggest steplength and the smallest one is not too high. In this case, the objective function values corresponding to points generated along the n linearly

independent directions are sufficiently representative of the local behavior of f. Hence, the direction p_k^{n+1} is computed taking these values into account, and it is given by the direction (suitably scaled) from the point with the highest objective value to the point with the lowest objective value. The aim is to approximate the direction of steepest descent. Whenever the test on the ratio between the biggest steplength and the smallest one is not satisfied, the direction p_k^{n+1} is chosen in such a way that the set $\{p_k^1, \ldots, p_k^{n+1}\}$ is a positive basis for R^n. Roughly speaking, the test at Step 4 can be viewed as a derivative-free angle condition which, as for the usual angle condition adopted in gradient-based algorithms, allows us to define sets of search directions satisfying Condition C1 and hence to ensure the global convergence of the algorithm.

At Step 5, the point y_k^{n+1} is produced by essentially repeating the instructions of Step 2 for the computed direction p_k^{n+1}. Finally, according to Step 6, the algorithm can update the current point by any point which produces an improvement of the objective function value with respect to $f(y_k^{n+1})$. Now we prove the following convergence result.

PROPOSITION 5.3. *Let $\{x_k\}$ be the sequence produced by Algorithm* 3. *Suppose that the vectors $\{p_k^i\}$, with $i = 1, \ldots, n$, are bounded and uniformly linearly independent. Then Algorithm* 3 *is well defined and we have*

$$\lim_{k \to \infty} \|\nabla f(x_k)\| = 0. \tag{5.19}$$

Proof. In order to prove the thesis, since Algorithm 3 is an instance of Algorithm 2, we need only show that the sets of directions $\{p_k^i\}_{i=1}^{n+1}$ satisfy Condition C1. First let us suppose that there exists an index $\bar{k}$ such that for all $k \geq \bar{k}$ we have

$$p_k^{n+1} = -\sum_{i=1}^{n} p_k^i.$$

Then, the sets p_k^i, with $i = 1, \ldots, n+1$, belong to the class (a) of sets of search directions defined in section 3, and hence Condition C1 is satisfied.

Now, let us consider any subset $K \subseteq \{0, 1, \ldots\}$ such that, for all $k \in K$, p_k^{n+1} is given by

$$p_k^{n+1} = \frac{v_k^{max} - v_k^{min}}{\xi_k},$$

according to Step 4. The instructions of this step imply

$$\frac{\alpha_k^{max}}{\alpha_k^{min}} \leq c \qquad \text{for all } k \in K. \tag{5.20}$$

In this case, we can prove that the sets p_k^i, with $k \in K$ and $i = 1, \ldots, n+1$, belong to the class (b) of sets of search directions defined in section 3. In fact, we can define

$$z_k^i = y_k^i, \qquad \xi_k^i = \begin{cases} \alpha_k^i & \text{if } \alpha_k^i > 0, \\ \tilde{\alpha}_k^i & \text{otherwise,} \end{cases} \qquad \text{for } i = 1, \ldots, n,$$

and we can set

$$\rho_k^1 = \rho_k^2 = \cdots = \rho_k^{2n-1} = 0, \quad \rho_k^{2n} = 1,$$

$$\tilde{\xi}_k^{2n} = \xi_k,$$

so that (3.2) becomes

$$p_k^{n+1} = \frac{(v_k^1 - v_k^{2n})}{\tilde{\xi}_k^{2n}} = \frac{v_k^{max} - v_k^{min}}{\xi_k}.$$

The conditions on ρ_k^l, with $l = 1, \ldots, 2n$, are obviously satisfied. Recalling the definitions of ξ_k^i for $i = 1, \ldots, n$, we have that (3.4) holds; moreover, the test at Step 4 implies that (3.5) is satisfied with $c_1 = c$ (see (5.20)). Regarding (3.6), recalling the boundedness of $\{p_k^j\}$ with $j = 1, \ldots, n$, we can write for all $i \in \{1, \ldots, n\}$

$$\|z_k^i - x_k\| \leq \sum_{j=0}^{i-1} \xi_k^j \|p_k^j\| \leq \max_{l=1,\ldots,n} \xi_k^l \sum_{j=0}^{i-1} \|p_k^j\| \leq c\xi_k^i \sum_{j=0}^{i-1} \|p_k^j\| \leq \tilde{c}\xi_k^i,$$

so that (3.6) holds with $c_2 = \tilde{c}$. Finally, by repeating the same reasoning used in the proof of Proposition 5.2, we can prove (5.14), (5.15), and (5.18), so that (3.7) is satisfied. □

6. Conclusions. In this work we have tried to establish a general convergence theory for unconstrained optimization without derivatives. Toward that aim, we have stated a set of conditions by satisfying which a pattern search or a line search algorithm is guaranteed to enjoy global convergence. On the basis of the theoretical analysis, we have defined new derivative-free algorithms which combine pattern and line search approaches. Future work will be devoted to designing an efficient code and to performing computational experiments in order to thoroughly investigate the practical interest of the proposed approach.

7. Appendix.

Proof of Proposition 3.1. Assume, by contradiction, that the assertion of the proposition is false. Therefore, there exists a value $\eta > 0$ such that, for every pair γ_t, δ_t, we can find an index $k(t)$ and scalars $\alpha_{k(t)}^i$, with $i = 1, \ldots, r$, for which we have

$$\|\nabla f(x_{k(t)})\| \geq \eta,$$

$$f(x_{k(t)} + \alpha_{k(t)}^i p_{k(t)}^i) > f(x_{k(t)}) - \gamma_t \alpha_{k(t)}^i \|\nabla f(x_{k(t)})\| \|p_{k(t)}^i\|,$$

and

$$0 < \alpha_{k(t)}^i \leq \delta_t$$

for all $i \in \{1, \ldots, r\}$. Now, taking into account the boundedness of $\{x_k\}$, we have that there exist (by relabeling if necessary) sequences $\{x_k\}$, $\{\gamma_k\}$, $\{\delta_k\}$, $\{\alpha_k^i\}$, $\{p_k^i\}$, with $i = 1, \ldots, r$, such that

$$x_k \to \bar{x}, \tag{7.1}$$

$$\gamma_k \to 0, \tag{7.2}$$

$$\delta_k \to 0, \tag{7.3}$$

$$\alpha_k^i \leq \delta_k, \tag{7.4}$$

$$f(x_k + \alpha_k^i p_k^i) > f(x_k) - \gamma_k \alpha_k^i \|\nabla f(x_k)\| \|p_k^i\|. \tag{7.5}$$

By the continuity assumption, we have that $\|\nabla f(\bar{x})\| \geq \eta$; then, by using Condition C1, for k sufficiently large there exists an index $i \in \{1, \ldots, r\}$ such that

$$\nabla f(x_k)^T p_k^i \leq \rho < 0. \tag{7.6}$$

Now, by (7.3), (7.4), and the boundedness of $\{p_k^i\}$ for $i = 1, \ldots, r$, we have that

$$\lim_{k\to\infty} \alpha_k^i \|p_k^i\| = 0 \tag{7.7}$$

for all $i \in \{1, \ldots, r\}$. By (7.5) and the mean-value theorem, we can write

$$\nabla f(x_k)^T p_k^i + (\nabla f(x_k + \theta_k^i \alpha_k^i p_k^i) - \nabla f(x_k))^T p_k^i \geq -\gamma_k \|\nabla f(x_k)\| \|p_k^i\|, \tag{7.8}$$

where $\theta_k^i \in (0, 1)$. From (7.7), (7.8), (7.2) and recalling again the boundedness of $\{p_k^i\}$, we get a contradiction with (7.6) for k sufficiently large. □

Proof of Proposition 3.2. If $\lim_{k\to\infty} \|\nabla f(x_k)\| = 0$, then the boundedness of $\{p_k^i\}$ for $i = 1, \ldots, r$ implies that $\lim_{k\to\infty} \min\{0, \nabla f(x_k)^T p_k^i\} = 0$, $i = 1, \ldots, r$.

In order to prove that

$$\lim_{k\to\infty} \sum_{i=1}^{r} \min\{0, \nabla f(x_k)^T p_k^i\} = 0 \tag{7.9}$$

implies

$$\lim_{k\to\infty} \|\nabla f(x_k)\| = 0, \tag{7.10}$$

we assume, by contradiction, that the assertion is false. Therefore, taking into account the boundedness of $\{x_k\}$, there exist a subset $K_1 \subseteq \{0, 1, \ldots\}$ and a positive number η such that

$$\lim_{k\to\infty, k\in K_1} x_k = \bar{x}, \tag{7.11}$$

$$\|\nabla f(\bar{x})\| \geq \eta > 0. \tag{7.12}$$

Now we distinguish the two classes of sets of search directions.

Class (a). By recalling the assumptions on the sets of search directions of this class, we have that we can find a subset $K_2 \subseteq K_1$ such that we have

$$\lim_{k\to\infty, k\in K_2} p_k^i = \bar{p}^i, \qquad i = 1, \ldots, r,$$

where $\bar{p}^1, \ldots, \bar{p}^r$ positively span R^n. Therefore, we can write

$$-\nabla f(\bar{x}) = \sum_{i=1}^{r} \beta^i \bar{p}^i, \tag{7.13}$$

with $\beta^i \geq 0$ for $i = 1, \ldots, r$. Then, recalling (7.12), we obtain

$$-\eta^2 \geq \sum_{i=1}^{r} \beta^i \nabla f(\bar{x})^T \bar{p}^i. \tag{7.14}$$

From (7.14), recalling the continuity assumption on ∇f, it follows that

$$\lim_{k\to\infty, k\in K_2} \sum_{i=1}^{r} \min\{0, \nabla f(x_k)^T p_k^i\} = \sum_{i=1}^{r} \min\{0, \nabla f(\bar{x})^T \bar{p}^i\} < 0,$$

which contradicts (7.9).

Class (b). By the boundedness assumptions on the sequences $\{p_k^i\}$, with $i = 1, \ldots, r$, and $\{\rho_k^l\}$, with $l = 1, \ldots, 2n$, we have that there exists a subset $K_2 \subseteq K_1$ such that we have

$$\lim_{k\to\infty, k\in K_2} p_k^i = \bar{p}^i, \qquad i = 1, \ldots, r, \tag{7.15}$$

$$\lim_{k\to\infty, k\in K_2} \rho_k^l = \bar{\rho}^i, \qquad l = 1, \ldots, 2n, \tag{7.16}$$

where $\bar{\rho}^{2n} \geq \bar{\rho} > 0$.

From the definitions of $\tilde{\xi}_k^l$ and v_k^l with $l = 1, \ldots, 2n$, the boundedness of $\{p_k^i\}$ with $i = 1, \ldots, r$ (for the sake of simplicity, we assume $\|p_k^i\| = 1$), and (3.5), (3.7), (3.6), it follows that the vectors $(v_k^l - v_k^1)/\tilde{\xi}_k^l$ are bounded. In fact, from (3.6), for k sufficiently large and for each $l \in \{1, \ldots, 2n\}$ we can write

$$\|v_k^l - v_k^1\| \leq \|v_k^l - x_k\| + \|x_k - v_k^1\| \leq \sigma_1^l \xi_k^l + \sigma_2^l \xi_k^1$$

with $\sigma_1^l, \sigma_2^l > 0$. From the assumptions on $\tilde{\xi}_k^l$, by (3.5) we have

$$\frac{1}{c_1} \leq \frac{\xi_k^i}{\tilde{\xi}_k^l} \leq c_1 \tag{7.17}$$

for each $i \in \{1, \ldots, n\}$ and for each $l \in \{1, \ldots, 2n\}$. Then, the boundedness of $\{p_k^i\}$, with $i = 1, \ldots, r$, implies the boundedness of $(v_k^l - v_k^1)/\tilde{\xi}_k^l$ for $l = 1, \ldots, 2n$. Hence we have

$$\lim_{k\to\infty, k\in K_2} \frac{v_k^l - v_k^1}{\tilde{\xi}_k^l} = \bar{y}^l, \qquad l = 1, \ldots, 2n. \tag{7.18}$$

Furthermore, (3.7) and (7.15) imply

$$\lim_{k\to\infty, k\in K_2} v_k^l = \bar{x}, \qquad l = 1, \ldots, 2n. \tag{7.19}$$

From (3.3), for all $k \geq 0$ and for $l = 1, \ldots, 2n$, we can write

$$f(v_k^l) - f(v_k^1) \geq 0,$$

from which, by using the mean-value theorem, it follows that

$$\tilde{\xi}_k^l \nabla f\left(v_k^l + \theta_k^l \tilde{\xi}_k^l \frac{v_k^l - v_k^1}{\tilde{\xi}_k^l}\right)^T \left(\frac{v_k^l - v_k^1}{\tilde{\xi}_k^l}\right) \geq 0, \tag{7.20}$$

with $\theta_k^l \in (0, 1)$. Then, recalling (7.11) and (7.18), taking into account (3.7), and by using the continuity assumption on ∇f for $l = 1, \ldots, 2n$, we have

$$\lim_{k\to\infty, k\in K_2} \nabla f\left(v_k^l + \theta_k^l \tilde{\xi}_k^l \frac{v_k^l - v_k^1}{\tilde{\xi}_k^l}\right)^T \left(\frac{v_k^l - v_k^1}{\tilde{\xi}_k^l}\right) = \nabla f(\bar{x})^T \bar{y}^l \geq 0. \tag{7.21}$$

Now, from (3.2), we get

$$\nabla f(x_k)^T p_k^{n+j} = \sum_{l=1}^{2n} \rho_k^l \nabla f(x_k)^T \frac{(v_k^1 - v_k^l)}{\tilde{\xi}_k^l}. \tag{7.22}$$

On the other hand, from (7.9) and recalling the continuity assumption on ∇f, it follows that

$$\lim_{k\to\infty, k\in K_2} \nabla f(x_k)^T p_k^i = \nabla f(\bar{x})^T \bar{p}^i = l^i \geq 0, \quad i = 1, \ldots, r. \tag{7.23}$$

Therefore, from (7.22), taking the limits for $k \to \infty$ and $k \in K_2$, we obtain

$$\nabla f(\bar{x})^T \bar{p}^{n+j} = -\sum_{l=1}^{2n} \bar{\rho}_l \nabla f(\bar{x})^T \bar{y}^l \geq 0,$$

where $\bar{\rho}^l \geq 0$ and $\bar{\rho}^{2n} > 0$. Hence, recalling (7.21), it follows that

$$\nabla f(\bar{x})^T \bar{y}^{2n} = 0. \tag{7.24}$$

Now, from (3.3), we get

$$\frac{f(v_k^{2n}) - f(v_k^1)}{\tilde{\xi}_k^{2n}} \geq \frac{f(z_k^i + \xi_k^i p_k^i) - f(z_k^i)}{\tilde{\xi}_k^{2n}}, \qquad i = 1, \ldots, n. \tag{7.25}$$

By using the mean-value theorem, we have

$$\frac{f(v_k^{2n}) - f(v_k^1)}{\tilde{\xi}_k^{2n}} = \frac{\nabla f\left(v_k^1 + \theta_k \alpha_k \frac{v_k^{2n} - v_k^1}{\tilde{\xi}_k^{2n}}\right)^T (v_k^{2n} - v_k^1)}{\tilde{\xi}_k^{2n}}, \tag{7.26}$$

$$\frac{f(z_k^i + \xi_k^i p_k^i) - f(z_k^i)}{\tilde{\xi}_k^{2n}} = \nabla f(z_k^i + u_k^i \xi_k^i p_k^i)^T p_k^i \frac{\xi_k^i}{\tilde{\xi}_k^{2n}}, \tag{7.27}$$

with $\theta_k \in (0,1)$, $u_k^i \in (0,1)$, $i = 1, \ldots, n$.

By substituting (7.26) and (7.27) into (7.25), taking the limits for $k \to \infty$ and $k \in K_2$, and recalling (7.17) and the continuity assumption on ∇f, we obtain

$$\nabla f(\bar{x})^T \bar{y}^{2n} \geq \nabla f(\bar{x})^T \bar{p}^i \frac{1}{c_1}, \qquad i = 1, \ldots, n.$$

Then, from (7.23) and (7.24), it follows that

$$\nabla f(\bar{x})^T \bar{p}^i = 0, \qquad i = 1, \ldots, n.$$

The linear independence of $\bar{p}^i$, with $i = 1, \ldots, n$, implies

$$\nabla f(\bar{x}) = 0,$$

which contradicts (7.12). □

Acknowledgments. The authors are grateful to the referees for their helpful comments and suggestions, which led to significant improvements in the paper.

REFERENCES

[1] R. De Leone, M. Gaudioso, and L. Grippo, *Stopping criteria for linesearch methods without derivatives*, Math. Programming, 30 (1984), pp. 285–300.

[2] J. E. Dennis, Jr., and V. Torczon, *Direct search methods on parallel machines*, SIAM J. Optim., 1 (1991), pp. 448–474.
[3] T. Glad and A. Goldstein, *Optimization of functions whose values are subject to small errors*, BIT, 17 (1977), pp. 160–169.
[4] L. Grippo, *A class of unconstrained minimization methods for neural network training*, Optim. Methods Softw., 4 (1994), pp. 135–150.
[5] L. Grippo, F. Lampariello, and S. Lucidi, *Global convergence and stabilization of unconstrained minimization methods without derivatives*, J. Optim. Theory Appl., 56 (1988), pp. 385–406.
[6] R. Hooke and T. A. Jeeves, *Direct search solution of numerical and statistical problems*, J. ACM, 8 (1961), pp. 212–229.
[7] R. M. Lewis and V. Torczon, *Rank Ordering and Positive Bases in Pattern Search Algorithms*, Technical report TR 96-71, ICASE, NASA Langley Research Center, Hampton, VA, 1996.
[8] S. Lucidi and M. Sciandrone, *Numerical results for unconstrained optimization without derivatives*, in Nonlinear Optimization and Applications, G. Di Pillo and F. Giannessi, eds., Plenum Publishing, New York, 1995, pp. 261–270.
[9] S. Lucidi and M. Sciandrone, *On the Global Convergence of Derivative Free Methods for Unconstrained Optimization without Derivatives*, Technical report R. 18-96, DIS, Università di Roma "La Sapienza," Rome, 1996.
[10] R. Mifflin, *A superlinearly convergent algorithm for minimization without evaluating derivatives*, Math. Programming, 9 (1975), pp. 100–117.
[11] M. J. D. Powell, *An efficient method for finding the minimum of a function of several variables without calculating derivatives*, Comput. J., 7 (1964), pp. 155–163.
[12] M. J. D. Powell, *Unconstrained minimization algorithms without computation derivatives*, Boll. Unione Mat. Ital., 9 (1974), pp. 60–69.
[13] M. J. D. Powell, *Direct search algorithms for optimization calculations*, Acta Numer., 7 (1998), pp. 287–336.
[14] A. S. Rykov, *Simplex direct search algorithms*, Automat. Remote Control, 41 (1980), pp. 784–793.
[15] A. S. Rykov, *Simplex methods of direct search*, Engineering Cybernetics, 18 (1980), pp. 12–18.
[16] V. Torczon, *On the convergence of pattern search algorithms*, SIAM J. Optim., 7 (1997), pp. 1–25.
[17] P. Tseng, *Fortified-descent simplicial search method: A general approach*, SIAM J. Optim., 10 (1999), pp. 269–288.
[18] Yu Wen-ci, *Positive basis and a class of direct search techniques*, Scientia Sinica, Special Issue of Mathematics, 1 (1979), pp. 53–67.
[19] M. H. Wright, *Direct search methods: Once scorned, now respectable*, in Proceedings of the 1995 Dundee Biennial Conference in Numerical Analysis, D. F. Griffiths and G. A. Watson, eds., Addison-Wesley Longman, Harlow, United Kingdom, 1996, pp. 191–208.
[20] W. J. Zangwill, *Minimizing a function without calculating derivatives*, Comput. J., 10 (1967), pp. 293–296.

SIAM J. OPTIM.
Vol. 13, No. 1, pp. 117–156

GENERALIZED BUNDLE METHODS*

ANTONIO FRANGIONI†

Abstract. We study a class of generalized bundle methods for which the stabilizing term can be any closed convex function satisfying certain properties. This setting covers several algorithms from the literature that have been so far regarded as distinct. Under a different hypothesis on the stabilizing term and/or the function to be minimized, we prove finite termination, asymptotic convergence, and finite convergence to an optimal point, with or without limits on the number of serious steps and/or requiring the proximal parameter to go to infinity. The convergence proofs leave a high degree of freedom in the crucial implementative features of the algorithm, i.e., the management of the bundle of subgradients (β-strategy) and of the proximal parameter (t-strategy). We extensively exploit a dual view of bundle methods, which are shown to be a dual ascent approach to one nonlinear problem in an appropriate dual space, where nonlinear subproblems are approximately solved at each step with an inner linearization approach. This allows us to precisely characterize the changes in the subproblems during the serious steps, since the dual problem is not tied to the local concept of ε-subdifferential. For some of the proofs, a generalization of inf-compactness, called *-compactness, is required; this concept is related to that of asymptotically well-behaved functions.

Key words. nondifferentiable optimization, bundle methods

AMS subject classifications. 90C25

PII. S1052623498342186

Introduction. We are concerned with the numerical solution of the *primal problem*

$$(\Pi) \qquad \inf_x \{f(x) : x \in X\}, \tag{0.1}$$

where $f : \Re^n \to \Re$ is finite-valued and convex (hence continuous) and $X \subseteq \Re^n$ is closed convex. Here f is only known through an oracle ("black box") that, given any $x \in X$, returns the values $f(x)$ and $z \in \partial f(x)$. To simplify the treatment, we will assume $X = \Re^n$ until section 8, where the extension to the constrained case is studied.

We study a class of generalized bundle methods for the solution of (0.1), where a stabilizing term, which can be any closed convex function satisfying certain weak conditions, is added to (a model of) f. These methods sample f in a sequence of *tentative points* $\{x_i\}$ to gather the f-values $\{f(x_i)\}$ and the *bundle* of first-order information $\beta = \{z_i \in \partial f(x_i)\}$. A distinguished vector $\bar{x}$ is taken as the *current point*, and β is used to compute a *tentative descent direction* d^* along which the next tentative point is generated. After a "successful" step, the current point can be updated; otherwise, the new information is used to enhance β, hopefully obtaining a better direction at the next iteration.

Several bundle methods proposed in the literature follow this pattern; some of them can be shown to actually belong to our class. Also, our generalized bundle methods provide implementable forms for some penalty-based algorithms for structured convex optimization. All of these algorithms have been analyzed either from the above primal viewpoint—the minimization of f—or from an application-specific

*Received by the editors July 20, 1998; accepted for publication (in revised form) December 6, 2001; published electronically, June 5, 2002.
http://www.siam.org/journals/siopt/13-1/34218.html
†Department of Computer Science, University of Pisa, Corso Italia 40, 56125 Pisa, Italy (frangio@di.unipi.it).

dual viewpoint, when f itself is a dual function. A dual analysis of some general bundle methods exists—indeed, it motivated the development of the very first bundle methods—but it is related to the "local" concept of ε-subdifferential, and it does not easily extend to a wider class of methods. Instead, we extensively exploit a dual view of (0.1), where bundle methods are shown to be a penalty function approach to a "global" dual problem with approximate solution, via an inner linearization approach, of the penalized problem. The algorithms can be entirely described in terms of this dual problem; this is interesting for applications and helps in the convergence proofs.

We analyze in detail the features that are relevant for practical implementations, such as the management of the bundle (β-strategy) and of the proximal parameter (t-strategy). General rules are given which ensure convergence while leaving a large degree of freedom in practical implementations. For some variants of the algorithm, we require f to be *-*compact*, an assumption properly generalizing inf-compactness. *-compact functions are *asymptotically well behaved* [Au97], but our definition seems to be better suited for the case of bundle methods.

The structure of this paper is the following: section 1 is devoted to the derivation of the dual viewpoint of generalized bundle methods. Some useful properties of pairs of primal and dual solutions of the stabilized master problems are proved in section 2. In section 3, the conditions on the stabilizing term are presented and discussed. Section 4 is devoted to the description of the algorithms and to the discussion of the rules for the β-strategy and the t-strategy. Convergence proofs of several variants of the algorithm are given next: section 5 is dedicated to convergence of the null step sequences, section 6 is dedicated to convergence of the serious step sequences, and section 7 is dedicated to the "third level" that is necessary for some classes of stabilizing terms. Some extensions of generalized bundle methods, e.g., to constrained optimization, are discussed in section 8, the relationships with other algorithms from the literature are analyzed in section 9, and conclusions are drawn in section 10.

Throughout the paper the following notation is used. The scalar product between two vectors v and w is denoted by vw. $\|v\|_p$ stands for the L_p norm of the vector v, and the ball around 0 of radius δ in the L_p norm will be denoted by $B_p(\delta)$. Given a set X, $I_X(x) = 0$ if $x \in X$ (and $+\infty$ otherwise) is its *indicator function*, $\sigma_X(z) = \sup_x\{zx : x \in X\}$ is its *support function*, and $d_X(y) = \inf_x\{\|x - y\| : x \in X\}$ is the *distance* from y to X. Given a function f, $\partial_\varepsilon f(x)$ is its *ε-subdifferential* at x, $epi\, f = \{(v, x) : v \geq f(x)\}$ is its *epigraph*, $dom\ f = \{x : f(x) < \infty\}$ is its *domain*, and $S_\delta(f) = \{x : f(x) \leq \delta\}$ is its *level set* corresponding to the f-value δ. Given a problem

$$\text{(P)} \qquad \inf[\sup]_x\{f(x) : x \in X\},$$

v(P) denotes the optimal value of f over X; as usual, $X = \emptyset \Rightarrow v(\text{P}) = +\infty[-\infty]$.

1. Duality for generalized bundle methods. The dual description of generalized bundle methods relies on a well-established tool from convex analysis, the *conjugate* of f (see [HL93b, Chapter X]):

$$f^*(z) = \sup_x\{zx - f(x)\}. \tag{1.1}$$

f^* is a closed convex function and enjoys several properties; those useful in the paper are briefly recalled below.

(1.i) $(f^*)^* = f$ (duality of the conjugate operator),

(1.ii) $f_1 \le f_2 \Rightarrow f_1^* \ge f_2^*$ ("monotonicity" of the conjugate operator),
(1.iii) $(f(\cdot + x))^*(z) = f^*(z) - zx \quad \forall z, x$ (effect of a simple variable change),
(1.iv) $z \in \partial_\varepsilon f(x) \Leftrightarrow x \in \partial_\varepsilon f^*(z)$ (duality of the subdifferential mappings),
(1.v) $z \in \partial_\varepsilon f(x) \Leftrightarrow f(x) + f^*(z) \le zx + \varepsilon$ (characterization of the ε-subdifferentials),
(1.vi) $zx = f(x) + f^*(z) \Leftrightarrow z \in \partial f(x)$ (basic relation between the function values),
(1.vii) $zx \le f(x) + f^*(z) \quad \forall z, x$ (Fenchel's inequality).

A fundamental property of f^* is that it characterizes all the affine functions supporting *epi* f as

$$zx - \varepsilon \le f(x) \quad \forall x \Leftrightarrow \sup_x \{zx - f(x)\} = f^*(z) \le \varepsilon.$$

Note that, when the oracle is called at some point x returning $f(x)$ and $z \in \partial f(x)$, $f^*(z)$ can be calculated via (1.vi); that is, the f^*-values are available if the f-values are, and vice-versa.

We remark that the above properties hold for any closed convex function; in the following, we will often take the conjugate of other functions apart from f, most notably of the "stabilizing term" to be introduced shortly.

1.1. The dual problem. Since f^* is related with the minimization of f by

$$v(\Pi) = \inf_x \{f(x)\} = -\sup_x \{0x - f(x)\} = -f^*(0),$$

we propose the following (apparently weird) *dual problem* as the dual of (0.1):

$$(\Delta) \qquad \inf_z \{f^*(z) : z = 0\}. \tag{1.2}$$

Problem (1.2) is a reasonable dual, since $v(\Pi) = -v(\Delta)$ and it deals with dual objects: every vector z that is a subgradient of f at some point belongs to *dom* f^* (cf. (1.v)). Furthermore, consider the *Lagrangian relaxation* of (1.2) w.r.t. the constraints $z = 0$, using $\bar{x}$ as Lagrangian multipliers:

$$(\Delta_{\bar{x}}) \qquad \inf_z \{f^*(z) - z\bar{x}\}. \tag{1.3}$$

From (1.1) and (1.i), one has

$$-v(\Delta_{\bar{x}}) = \sup_z \{z\bar{x} - f^*(z)\} = (f^*)^*(\bar{x}) = f(\bar{x});$$

therefore, the *dual pricing problem* (1.3) can be seen as the problem that the oracle has to solve for computing $f(\bar{x})$. From the dual viewpoint, the oracle inputs $\bar{x}$ and returns a contact point $(f^*(z), z)$ between *epi* f^* and the affine function with slope $(1, -\bar{x})$ that supports the set. This notation reveals that (0.1) itself is the Lagrangian dual of (1.2) w.r.t. the constraints $z = 0$.

1.2. Approximations of f and bundle algorithms. Our aim is the construction of an algorithm that solves (0.1)—or, equivalently, (1.2)—given the oracle for f. A number of bundle algorithms have been proposed for this task, all based on the idea of using the bundle β for constructing a *model* f_β of the original function f. The

model is usually required to be a lower approximation of the function, i.e., $f_\beta \leq f$, so that the *primal master problem*

$$(\Pi_{\beta,\bar{x}}) \qquad \inf_d \{f_\beta(\bar{x}+d)\} \tag{1.4}$$

gives a lower bound on the primal problem (0.1). The optimal solution d^* of (1.4) is then used as a (tentative) descent direction, analogously to what is done in Newton methods. From the dual viewpoint, $f^*_\beta \geq f^*$ (cf. (1.ii)) implies that the *dual master problem*

$$(\Delta_{\beta,\bar{x}}) \qquad \inf_z \{f^*_\beta(z) - z\bar{x} : z = 0\} \tag{1.5}$$

is an upper approximation of the dual problem (1.2).

The most popular model of f is the *cutting plane model*

$$\hat{f}_\beta(x) = \max_z \{zx - f^*(z) : z \in \beta\}, \text{ for which} \tag{1.6}$$

$$\hat{f}^*_\beta(z) = \inf_\theta \left\{ \sum_{w\in\beta} f^*(w)\theta_w : \sum_{w\in\beta} w\theta_w = z, \quad \theta \in \Theta \right\}, \tag{1.7}$$

where $\Theta = \{\sum_{w\in\beta} \theta_w = 1, \theta \geq 0\}$ is the unitary simplex [HL93b, Proposition X.3.4.1]; note that $dom\ \hat{f}^*_\beta = conv(\beta)$. Using $\hat{f}_\beta$ in (1.4) gives the well-known cutting plane algorithm [HL93b, Algorithm XII.4.2.1], where the unknown f is replaced with its known polyhedral outer approximation $\hat{f}_\beta$. In the corresponding (1.5), the unknown f^* is replaced with its known polyhedral inner approximation $\hat{f}^*_\beta$ (a "pin-function").

1.3. Stabilized master problems. The cutting plane algorithm has some serious drawbacks, both in theory and in practice. First of all, the primal master problem (1.4) may be unbounded, that is, the dual master problem (1.5) may be infeasible; this is usually the case in the first iterations. Furthermore, two subsequent tentative points can be arbitrarily far apart; this is known as the "instability" of the cutting plane method. Most bundle methods try to alleviate this problem by introducing some "stabilizing device" into (1.4). Here, the *stabilizing term* D_t—a closed convex function—is added to f_β to discourage points "far away" from $\bar{x}$, where $t > 0$ is the *proximal parameter* dictating the "strength" of D_t. That is, at each step the *stabilized primal master problem*

$$(\Pi_{\beta,\bar{x},t}) \qquad \inf_d \{f_\beta(\bar{x}+d) + D_t(d)\} \tag{1.8}$$

is solved, and its optimal solution d^* is used as a (tentative) descent direction. By *Fenchel's duality* [HL93b, section XII.5.4], the dual of (1.8) is (using (1.iii)) the *stabilized dual master problem*

$$(\Delta_{\beta,\bar{x},t}) \qquad \inf_z \{f^*_\beta(z) - z\bar{x} + D^*_t(-z)\}. \tag{1.9}$$

Under proper assumptions (cf. Lemma 2.1 below), $v(\Delta_{\beta,\bar{x},t}) = -v(\Pi_{\beta,\bar{x},t})$. We see that the *primal stabilizing term* D_t corresponds to a *dual penalty function* D^*_t associated with the constraints $z = 0$; (1.9) is a (generalized) *augmented Lagrangian* of (1.5). The stabilizing term is a member of a family of functions parameterized in t;

in the bundle methods proposed so far, t is either a factor, like in $D_t = \frac{1}{2t}\|\cdot\|_2^2$, or the radius of a ball, like in $D_t = I_{B_\infty(t)}$. In general, we will not require the function $t \to D_t(d)$ to have any specific form.

Note that f-values or f^*-values must be stored in memory together with the subgradients; due to (1.vi) the two choices are equivalent. In the standard notation of bundle methods, for $z \in \partial f(x)$ the *linearization error* (cf. [HL93b, Definition XI.4.2.3])

$$\alpha = f^*(z) - z\bar{x} + f(\bar{x}) = f(\bar{x}) - f(x) - z(\bar{x} - x) \geq 0 \tag{1.10}$$

of z w.r.t. $\bar{x}$ is typically used in place of $f^*(z)$. This notation corresponds to defining the *translated function* $f_{\bar{x}}(d) = f(\bar{x} + d) - f(\bar{x})$ and its *translated model* $f_{\bar{x},\beta}$, and to considering a "local" form of (1.8) that uses $f_{\bar{x},\beta}$ [Fr98]. However, the corresponding dual problem is written in terms of $f^*_{\bar{x}}$, i.e., of a family of functions changing with $\bar{x}$, rather than in terms of the unique f^*. Furthermore, the notation based on linearization errors hides the dependency of some of the subproblem's data on the current point $\bar{x}$; that is why we use f^*-values.

1.4. Stabilization in the original problems. The above duality argument can also be applied to the original function f; the stabilized dual problem

$$(\Delta_{\bar{x},t}) \qquad \inf_z \{f^*(z) - z\bar{x} + D_t^*(-z)\} \tag{1.11}$$

is the (Fenchel) dual of the *stabilized primal problem*

$$(\Pi_{\bar{x},t}) \qquad \phi_t(\bar{x}) = \inf_d \{f(\bar{x} + d) + D_t(d)\}. \tag{1.12}$$

A primal analysis of generalized bundle methods would focus on (1.12), that is, the calculation of the *generalized Moreau–Yosida regularization* ϕ_t of f in $\bar{x}$. With a proper D_t [BPP91], ϕ_t has the same set of minima as f but enjoys additional properties, e.g., smoothness; hence, minimizing ϕ_t could be an advantageous alternative to the minimization of f. Unfortunately, solving (1.12) with the sole help of the black box for f is as difficult as solving (0.1); therefore, bundle methods resort to a two-level approach, repeatedly solving the approximation (1.8) until the accumulation of information in β makes f_β a "good enough" approximation of f, and only then changing $\bar{x}$. If t is properly managed, the whole process eventually solves (0.1).

But a dual analysis of generalized bundle methods is also possible, which focuses instead on (1.2) and its *generalized augmented Lagrangian* (1.11), where the constraints $z = 0$ are replaced with the linear term $-\bar{x}z$ (with Lagrangian multipliers $\bar{x}$) and the nonlinear term $D_t^*(-z)$ in the objective function. A classical ascent method would require repeatedly solving (1.11) and updating $\bar{x}$ using the corresponding first-order information; unfortunately, solving (1.11)—which is equivalent to (1.12)—is difficult. On the contrary, (1.9) may be efficiently solvable; furthermore, the oracle for f solves (1.3), and hence $v(\Delta_{\bar{x}+d})$ gives a lower bound on (1.11) if $-zd$ is a linear lower approximation of $D_t^*(-z)$. Hence, a viable approach is again a two-level one, where in the inner level a sequence of (1.9) and (1.3) is solved for fixed $\bar{x}$ in order to approximate (1.11), while in the outer level the Lagrangian multipliers $\bar{x}$ and the parameter t, dictating the "strength" of the penalty function, are updated.

This dual interpretation of bundle methods is related to—although independently obtained from—the general dual algorithmic scheme of [ACC93]; by taking their "perturbation function" $\varphi(x, \bar{x})$ as $f(x - \bar{x})$, the Lagrangian dual of (1.3), i.e., (1.2), is

obtained. However, in our case the relevant dual object is simply the conjugate f^*, and the whole process takes place in the graph space of f^*. This is confirmed by [Nu97], where a step in the same direction has been made using the graph of the $\varepsilon \to \partial_\varepsilon f(0)$ multifunction that is equivalent to *epi* f^* (cf. section 9.1).

2. Properties of subproblem solutions. The following two lemmas will be useful in the analysis of the algorithm.

LEMMA 2.1. *Let f_β and D_t be two closed convex functions such that dom $f_\beta(\bar{x} + \cdot) \cap$ int dom $D_t \neq \emptyset$, and assume that* (1.8) *and of* (1.9) *have optimal solutions d^* and z^*, respectively; then*

$$v(\Delta_{\beta,\bar{x},t}) = -v(\Pi_{\beta,\bar{x},t}), \tag{2.1}$$

$$-z^* \in \partial D_t(d^*) \qquad \text{and} \qquad d^* \in \partial D_t^*(-z^*), \tag{2.2}$$

$$z^* \in \partial f_\beta(\bar{x} + d^*) \qquad \text{and} \qquad \bar{x} + d^* \in \partial f_\beta^*(z^*), \tag{2.3}$$

$$f_\beta(\bar{x} + d^*) + f_\beta^*(z^*) = z^*(\bar{x} + d^*), \tag{2.4}$$

$$D_t(d^*) + D_t^*(-z^*) = -z^* d^*. \tag{2.5}$$

Proof. Equation (2.1) is [HL93b, (X.2.3.2)]. Apply [HL93b, Proposition XII.5.4.1] with the nonsymmetric assumption [HL93b, (X.2.3.Q.jj′)] to the pair (1.8)–(1.9) to show that any optimal solution d^* of (1.8) belongs to $\partial[f_\beta(\bar{x} + \cdot)]^*(z^*) \cap \partial D_t^*(-z^*)$; this gives $d^* \in \partial D_t^*(-z^*)$ and, via (1.iii), $\bar{x} + d^* \in \partial f_\beta^*(z^*)$. For the rest, apply (1.iv) and (1.vi). ∎

We remark that Lemma 2.1 works for any closed convex function f_β, even if it is not a model of f. We will always keep the requirement on f_β to the bare minimum, in the spirit of [CL93]; this will provide more general results, and it will be useful in section 8 where extensions of the method are discussed. Also, note that Lemma 2.1 with $f_\beta = f$ characterizes the properties of the solutions d^* and z^* of the primal and dual stabilized problems (1.12) and (1.11). When $f_\beta \leq f (\Rightarrow f_\beta^* \geq f^*$ by (1.ii)), the optimal solutions of the master problems allow us to derive information on those of the original problems.

LEMMA 2.2. *If $f_\beta \leq f$ and the hypothesis of Lemma* 2.1 *hold, then the optimal value of* (1.11) *can be bracketed using* (1.9) *and*

$$\Delta f = f(\bar{x} + d^*) - f_\beta(\bar{x} + d^*) \geq 0, \tag{2.6}$$

$$\textit{i.e.,} \qquad v(\Delta_{\bar{x},\beta,t}) - \Delta f \leq v(\Delta_{\bar{x},t}) \leq v(\Delta_{\bar{x},\beta,t}). \tag{2.7}$$

Proof. $v(\Delta_{\bar{x},t}) \leq v(\Delta_{\beta,\bar{x},t})$ comes from $f_\beta^* \geq f^*$. From (2.2),

$$D_t^*(-z) \geq D_t^*(-z^*) - d^*(z - z^*) \quad \forall z.$$

Add $f^*(z) - z\bar{x}$ to both sides, then add and remove $f_\beta^*(z^*) - z^*\bar{x}$ to the right-hand side to obtain

$$f^*(z) - z\bar{x} + D_t^*(-z) \geq v(\Delta_{\beta,\bar{x},t}) - [f_\beta^*(z^*) - f^*(z) + (\bar{x} + d^*)(z - z^*)] \quad \forall z.$$

Take the inf on z on both sides and recognize the stabilized dual problem (1.11) on the left and the dual pricing problem (1.3) at $\bar{x} + d^*$ plus $f_\beta(\bar{x} + d^*)$ (via (2.4)) on the right. ∎

For future reference, let us record here the alternative formula

$$\Delta f = f_\beta^*(z^*) - f^*(z) + (\bar{x} + d^*)(z - z^*), \tag{2.8}$$

where $z \in \partial f(\bar{x} + d^*)$. ($z$ is an optimal solution of (1.3) at $\bar{x} + d^*$.)

Let us briefly comment on the above lemmas. Equation (2.3) shows that the dual optimal solution z^* gives, in primal terms, a linear lower approximation of the model f_β which, by (2.4), is tight in $\bar{x} + d^*$. Conversely, by (2.2) the primal direction d^* gives, in dual terms, a subgradient of D_t^* at $-z^*$. Lemma 2.2 shows that the gap between the model and the original function in $\bar{x}+d^*$ is a measure of the gap between (1.9) and (1.11); thus, if $\Delta f = 0$, then z^* is optimal for (1.11) ($f_\beta^*(z^*) = f^*(z^*)$), and d^* is optimal for (1.12).

If $f_\beta^* \geq f^*$, a useful object in the analysis of the algorithms is

$$\alpha^* = f_\beta^*(z^*) - z^*\bar{x} + f(\bar{x}) \geq 0 \tag{2.9}$$

(use (1.vii)); using (1.v) in (2.9), one obtains

$$z^* \in \partial_{(\alpha^*)} f(\bar{x}). \tag{2.10}$$

Note that all of the above relations are independent of the choice of f_β and D_t; in the literature, analogous results have usually been obtained algebraically for specific choices, such as $D_t = \frac{1}{2t}\|\cdot\|_2^2$ and $f_\beta = \hat{f}_\beta$. However, not all the results for particular cases generalize; a relevant example is $d^* = -tz^*$, which is central in the analysis of proximal bundle methods but it is not true in general.

3. Conditions on D_t. Of course, the primal stabilizing term D_t has to satisfy some conditions. First of all, in order to be able to apply the results of the previous paragraph, D_t has to be a closed convex function $\forall t > 0$. Then, a set of weak properties that suffice for constructing a convergent algorithm is the following:

- (P1) $\forall t > 0$, $D_t(0) = 0$ and $0 \in \partial D_t(0)$ (D_t is *nonnegative*).
- (P2) $\forall t > 0$ and $\varepsilon > 0$, $S_\varepsilon(D_t)$ is *compact* and $0 \in$ int $S_\varepsilon(D_t)$ ($S_\varepsilon(D_t)$ is *full-dimensional*).
- (P3) $\forall t > 0$, $\lim_{\|d\|\to\infty} D_t(d)/\|d\| = +\infty$ (D_t is *strongly coercive*).
- (P4) $\forall t > 0$, $D_t \geq D_\tau$ for each $\tau \geq t$ (D_t is *nonincreasing* in t).
- (P5) $\lim_{t\to\infty} D_t(d) = 0$ $\forall d$ ($\{D_t\}$ *converges pointwise* to the constant zero function).

We will show that the above conditions on D_t are equivalent to the following conditions on D_t^*:

- (P*1) $\forall t > 0$, $D_t^*(0) = 0$ and $0 \in \partial D_t^*(0)$ (D_t^* is *nonnegative*).
- (P*2) $\forall t > 0$ and $\varepsilon > 0$, $S_\varepsilon(D_t^*)$ is *compact* and $0 \in$ int $S_\varepsilon(D_t^*)$ ($S_\varepsilon(D_t^*)$ is *full-dimensional*).
- (P*3) $\forall t > 0$, D_t^* is *finite everywhere*.
- (P*4) $\forall t > 0$, $D_t^* \leq D_\tau^*$ for each $\tau \geq t$ (D_t^* is *nondecreasing* in t).
- (P*5) $\forall \varepsilon > 0$, $\lim_{t\to\infty} \inf_z\{D_t^*(z) : \|z\| \geq \varepsilon\} = +\infty$ ($\{D_t^*\}$ *converges* "*uniformly*" to $I_{\{0\}}$).

The following remarks about (P1)–(P5) are useful:

- Having a minimum in 0 where they evaluate to 0, both D_t and D_t^* are nonnegative functions $\forall t > 0$.
- As a consequence of (P1) and (P*1), D_t and D_t^* are *radially nondecreasing*, i.e.,

$$\forall \alpha \geq 1 \qquad D_t(\alpha d) \geq D_t(d) \quad \forall d \qquad \text{and} \qquad D_t^*(\alpha z) \geq D_t^*(z) \quad \forall z, \tag{3.1}$$

since, e.g., $d = (1/\alpha)\alpha d + (1-1/\alpha)0$ and, by convexity, $D_t(d) \le (1/\alpha)D_t(\alpha d) + (1-1/\alpha)D_t(0) \le D_t(\alpha d)$ as $\alpha \ge 1$ and $D_t(\alpha d) \ge 0$.

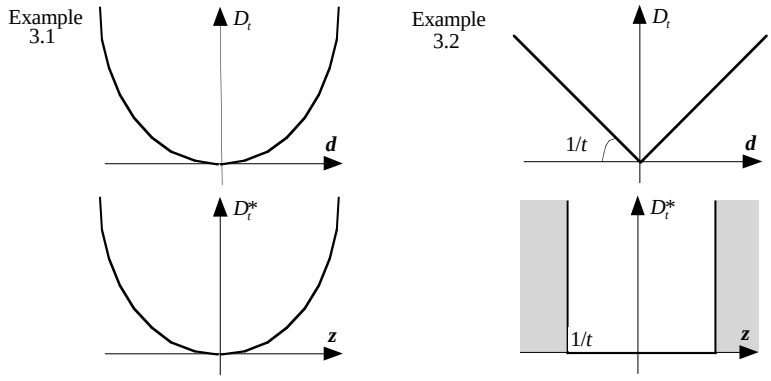
– Another consequence of (P1) and (1.v) is

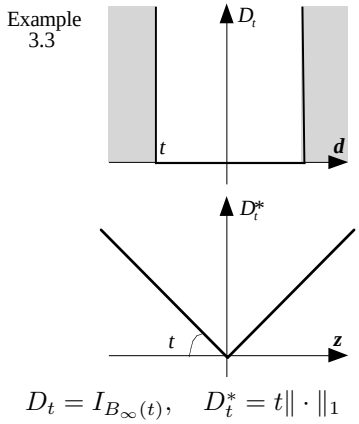
$$S_\varepsilon(D_t) = \partial_\varepsilon D_t^*(0) \qquad \text{and} \qquad S_\varepsilon(D_t^*) = \partial_\varepsilon D_t(0); \tag{3.2}$$

a rephrasing of (P2) is therefore that *both* the level sets of D_t *and* its ε-subdifferentials at 0 must be compact, and the same holds for D_t^*.

– (P2) guarantees that the hypothesis of Lemma 2.1 is true, as $0 \in \text{int } dom\ D_t$ and $0 \in dom\ f_\beta(\bar{x} + \cdot)$. (This is true even in the constrained case, cf. section 8.1, assuming of course that $\bar{x} \in X$.)

– (P2) and (P*2) are stated for $\varepsilon > 0$: $S_0(D_t)$ and $S_0(D_t^*)$ may or may not be full-dimensional, as in the following examples.





– It is intuitive why (P2) and (P*2) are necessary. The noncompactness of $S_\varepsilon(D_t)$ for some $\varepsilon > 0$ means that D_t is constantly 0 along some direction d and therefore cannot "stabilize" f along d. In fact, all the nonempty level sets of a closed convex function have the same *asymptotic cone* [HL93a, Proposition IV.3.2.5], so that $S_0(D_t)$ is also noncompact. On the other hand, if 0 belongs to the frontier of $dom\ D_t$, then some d is a "forbidden" direction, i.e., $D_t(\alpha d) = +\infty\ \forall \alpha > 0$.

– Strongly coercive (or 1-coercive) functions increase faster than any linear

function at infinity; (P3) guarantees that (1.8) has a bounded nonempty set of optimal solutions.
- Concerning (P4) and (P*4), note the role of t in Examples 3.1–3.3 above.
- The need for (P4) and (P5) is also intuitively clear: t must make D_t "weaker" as it grows, and it must be possible to make D_t as weak as desired in order to avoid "blocking" promising directions. Dually, a penalty term must increase as the penalty parameter does (see (P*4)), and it must be equivalent to the constraints it replaced, at least in the limit (see (P*5)).

D_t need not be "norm-like" [KCL95, Be96] or a *Bregman distance* [CT93]; in particular, it is not necessary that $D_t(0) = 0 \Leftrightarrow d = 0$ [IST94, Ki99]. Also, $t \to D_t(d)$ need not have the $1/t$ form.

THEOREM 3.1. (P1)–(P5) *are equivalent to* (P*1)–(P*5).

Proof. For the first four properties, the equivalence is pairwise.

1. The equivalence between (P1) and (P*1) is an easy consequence of (1.iv) and (1.vi).
2. The equivalence between (P2) and (P*2) can be obtained as a consequence of the following little-known result: for any proper convex function D, $\bar{d} \in \text{int } dom\ D \Leftrightarrow \bar{d} \in \text{int } S_\delta(D)\ \forall \delta > D(\bar{d})$. One of the implications is obvious; for the other, $\bar{d} \in \text{int } dom\ D$ means that there exists a ball $B(\bar{d}, \varepsilon)$ with $\varepsilon > 0$ such that $B(\bar{d}, \varepsilon) \subseteq \text{int } dom\ D$. By [HL93a, Theorem IV.3.1.2], D is Lipschitz over the ball, i.e., $|D(d) - D(\bar{d})| \le L\|d - \bar{d}\|\ \forall d \in B(\bar{d}, \varepsilon)$ for some constant $L > 0$; hence, $S_\delta(D) \supseteq B(\bar{d}, \min\{\varepsilon, (\delta - D(\bar{d}))/L\})$ as desired.
   Using this result, [HL93b, Theorem XI.1.1.4], and (3.2), one has
   $$0 \in \text{int} S_\varepsilon(D_t) \Leftrightarrow 0 \in \text{int } dom\ D_t \Leftrightarrow \partial_\varepsilon D_t(0) \text{ compact} \Leftrightarrow S_\varepsilon(D_t^*) \text{ compact}.$$
   To complete the proof of the equivalence, simply exchange D_t with D_t^*.
3. The equivalence between (P3) and (P*3) is [HL93b, Remark X.1.3.10].
4. The equivalence between (P4) and (P*4) is (1.ii).
5. For the last step, we will show that [(P1) + (P4) + (P5)] ⇒ (P*5) and [(P*1) + (P*4) + (P*5)] ⇒ (P5).

[(P1) + (P4) + (P5)] ⇒ (P*5). Due to (3.2) (which requires (P1) ≡ (P*1)), (P*5) can be rewritten as

$$\forall \varepsilon > 0 \lim_{t \to \infty} \inf_z \{D_t^*(z) : \|z\| = \varepsilon\} = +\infty.$$

Now, assume by contradiction that $\varepsilon > 0$ exists such that the limit is not $+\infty$; since the feasible set is compact and D_t^* is closed, for each t there exists a z_t achieving the inf, and we can write

$$\lim_{t \to \infty} D_t^*(z_t) \le M < +\infty.$$

From (1.vii) $\forall t\ \forall d\ \forall z,\ D_t(d) + D_t^*(z) \ge zd$; choosing $z = z_t$ and using $D_t^*(z_t) \le M$, one obtains

$$\forall t\ \forall d \qquad D_t(d) \ge z_t d - M.$$

But all the z_t belong to a compact set, and therefore some cluster point z^* exists with $\|z^*\| = \varepsilon$; plugging $d^* = (2M/\varepsilon^2)z^*$ into the above inequality and taking the limit for $t \to \infty$, one gets

$$\lim_{t \to \infty} D_t(d^*) \ge \lim_{t \to \infty} \left(\frac{2M}{\varepsilon^2} \right) z_t z^* - M = 2M - M > 0,$$

which contradicts (P5).

[(P*1) + (P*4) + (P*5)] ⇒ (P5). As a preliminary, we must show that for every d there exists a sufficiently large $\bar{t}$ such that $D_t(d) < +\infty$ for each $t \geq \bar{t}$; due to (P4) ≡ (P*4), it is only necessary to show that this happens for at least one t. Assume by contradiction that one $\bar{d}$ exists such that $\bar{d} \notin dom\ D_t\ \forall t$. Using [HL93a, Theorem V.2.2.2], one has

$$\forall t\ \exists z_t : \|z_t\| = 1 \quad \sup_d \{z_t d : d \in dom\ D_t\} \leq z_t \bar{d}.$$

Now

$$\begin{aligned} D_t^*(z_t) &= \sup_d\{z_t d - D_t(d)\} = \sup_d\{z_t d - D_t(d) : d \in dom\ D_t\} \\ &\leq \sup_d\{z_t d : d \in dom\ D_t\} \leq z_t \bar{d} \qquad (D_t \geq 0). \end{aligned}$$

Using $\|z_t\| = 1$, this finally gives $\forall t,\ D_t^*(z_t) \leq \|\bar{d}\|_2$, which contradicts (P*5). Hence, each d is in $dom\ D_t$ for a sufficiently large t.

We now want to prove that (P5) holds, so assume by contradiction that one $\bar{d}$ exists such that $D_t(\bar{d}) \geq \varepsilon > 0\ \forall t > 0$. (It must be $\bar{d} \neq 0$ due to (P1) ≡ (P*1), and note that we are using (P4) ≡ (P*4).) Since $D_t(\bar{d}) > D_t(0) = 0$, $0 \notin \partial D_t(\bar{d})$. In fact, from the subgradient inequality

$$D_t(d) \geq D_t(\bar{d}) + z(d - \bar{d}) \qquad \forall d\ \forall z \in \partial D_t(\bar{d})$$

one gets for $d = 0$, using (P1),

$$z\bar{d} \geq D_t(\bar{d}) \geq \varepsilon \quad \forall z \in \partial D_t(\bar{d}) \Rightarrow \|z\| \geq \varepsilon' = \varepsilon/\|\bar{d}\| \quad \forall z \in \partial D_t(\bar{d}).$$

Now, for each t choose any $z_t \in \partial D_t(\bar{d})$. Using (1.vi), $\|z_t\| \geq \varepsilon'$, and (P1), we obtain

$$\lim_{t\to\infty} \inf_z \{D_t^*(z) : \|z\| = \varepsilon'\} \leq \liminf_{t\to\infty} D_t^*(z_t) \leq \liminf_{t\to\infty} z_t\bar{d} - D_t(\bar{d}) \leq \liminf_{t\to\infty} z_t\bar{d}.$$

There exists a large enough $\bar{t}$ such that $2d \in dom\ D_{\bar{t}}$; hence, by (3.1), $d \in dom\ D_{\bar{t}}$ also. Again using the subgradient inequality, (P4) (which is implied by (P*4)), and (P1), we have

$$\forall t \geq \bar{t} \qquad D_{\bar{t}}(2\bar{d}) \geq D_t(2\bar{d}) \geq D_t(\bar{d}) + z_t(2\bar{d} - \bar{d}) \geq z_t\bar{d},$$

which finally gives

$$\lim_{t\to\infty} \inf_z \{D_t^*(z) : \|z\| = \varepsilon'\} \leq \liminf_{t\to\infty} z_t\bar{d} \leq D_{\bar{t}}(2\bar{d}) < \infty,$$

contradicting (P*5) and therefore finishing the proof of the theorem. □

Condition (P*5) may be a bit clumsy to check. The following result gives a handy sufficient condition that should work in most cases.

THEOREM 3.2. *If* (P*4) *holds,* $\{D_t^*\}$ *converges pointwise to* $I_{\{0\}}$, *i.e.,*

$$\lim_{t\to\infty} D_t^*(z) = +\infty \quad \forall z \neq 0,$$

and for any two sequences $\{t_i\} \to +\infty$ *and* $\{z_i\} \to \bar{z}$, *where* $z_i \in dom\ D_{t_i}^*$, *one has*

$$\liminf_{t\to\infty} D_{t_i}^*(z_i) \geq \lim_{i\to\infty} D_{t_i}^*(\bar{z}),$$

then (P*5) *holds.*

Proof. The thesis is obvious if for any fixed ε there exists a t such that *dom* $D_t^* \subseteq B_2(\varepsilon)$, since by (P*4) the domain of D_t^* can only shrink as t increases. Hence, we can assume that *dom* $D_t^* \backslash B_2(\varepsilon)$ is nonempty $\forall t$. Assume by contradiction that for some $\varepsilon > 0$ and $\{t_i\} \to +\infty$ there exist one $\delta < \infty$ and a sequence $\{z_i\}$ of points outside $B_2(\varepsilon)$ such that $D_{t_i}^*(z_i) \leq \delta \ \forall i$. Let $\bar{z}_i = (\varepsilon/\|z_i\|_2)z_i$ (the projection of z_i on $B_2(\varepsilon)$). By (3.1), $D_{t_i}^*(\bar{z}_i) \leq D_{t_i}^*(z_i)$. Now, $B_2(\varepsilon)$ is a compact set; hence we can assume $\{\bar{z}_i\} \to \bar{z}$ with $\|\bar{z}\|_2 = \varepsilon > 0$. Using the hypothesis,

$$\infty = \lim_{i\to\infty} D_{t_i}^*(\bar{z}) \leq \liminf_{i\to\infty} D_{t_i}^*(\bar{z}_i) \leq \delta < \infty. \qquad \square$$

All the D_t^* proposed so far satisfy (P*5); they are either continuous in both z and t (cf. Examples 3.1, 3.3) or indicator functions of balls shrinking as t increases (cf. Example 3.2). It is clear from the proof of Theorem 3.2 that these two possibilities—which in our setting can be mixed—have two distinct ways of ensuring that (P*5) holds. Bundle methods using these two different types of stabilizing term, i.e., *penalty* and *trust-region*, have so far been viewed as distinct [HL93b, sections XV.2.1 and XV.2.2].

It is possible to avoid the strong coercivity assumption (P3) (cf. Example 3.2), provided that other assumptions guarantee that (1.8) is bounded below.

(P3′) f is *bounded below*, a finite f_* such that $f_* \leq v(\Pi)$ is *known* and $f_\beta \geq f_* \ \forall \beta$.

(P*3′) (1.2) is *nonempty*, a finite f_* such that $f^*(0) \leq -f_*$ is *known* and $f_\beta^*(0) \leq -f_* \ \forall \beta$.

Note that there are three separate conditions in (P3′): a suitable f_* must *exist*, must *be known*, and the corresponding "flat" subgradient must be *explicitly kept* in the bundle. From the dual viewpoint, (P*3′) guarantees that 0 is a feasible solution for (1.9). A more general condition would be requiring f_β to be always bounded below; with such a model, the cutting plane algorithm could be directly applied without stabilization. However, the constant zero function is not a valid stabilizing term, even if (P3) is not enforced, due to the first part of (P2) (compactness).

Two other variants of the above properties allow us to obtain stronger convergence results:

(P3″) $\forall t$ D_t is *strongly coercive* and *strictly convex.*

(P*3″) $\forall t$ D_t^* is *finite everywhere* and *differentiable.*

(P5′) $\forall t$ $\partial D_t(0) = \{0\}$ (D_t is *differentiable* in 0, i.e., $\nabla D_t(0) = 0$).

(P*5′) $\forall t$ $S_0(D_t^*) = \{0\}$ (D_t^* is *strictly convex* in 0, i.e., 0 is the *unique minimum* of D_t^*).

(P3″) is a strengthening of (P3) that allows us to keep the size of β bounded. The equivalence between (P3″) and (P*3″) is [HL93b, Theorem X.4.1.1]. Under (P5′), 0 is a stationary point of $f(\bar{x} + \cdot) + D_t$ if and only if $\bar{x}$ is a stationary point of f; with (P5′) replacing (P5), it is possible to prove convergence without requiring $t \to \infty$. The equivalence between (P5′) and (P*5′) is a consequence of (3.2). (P5′) implies the second part of (P2) (full dimensionality); this is easily seen in the dual, as (P*5′) implies the first part of (P*2) (compactness), since all the level sets of D_t^* share the same asymptotic cone of $S_0(D_t^*) = \{0\}$.

So far, nothing has been required about the form of the $t \to D_t(d)$ functions; in this very general setting, D_t and $D_{t'}$ for $t \neq t'$ may be two almost completely unrelated functions. In some cases, stronger results can be obtained under the following (pretty

```
⟨ let μ ≥ 1 and ε ≥ 0 be fixed; choose the initial x̄, t, and β ⟩        // initialization
do
    ⟨ solve (Π_{β,x̄,t}) and (Δ_{β,x̄,t}) for d* and z*, respectively ⟩;   // find a direction
    ⟨ move along d*, generating some new z and a trial point x ⟩;        // probe f along d*
    if (a large enough improvement has been obtained)                    // NS/SS decision
        then x̄ = x;                                                     // a serious step
    ⟨ add some new z to β, delete some old z from β ⟩;                   // the β-strategy
    ⟨ update t, depending on the previous history ⟩;                     // the t-strategy
while (α* + μ D*_t(−z*) > ε);                                           // stopping condition
```

FIG. 1. *The "two-level" bundle algorithm.*

```
⟨ choose the initial x̄, t̲ > 0, ε > 0, and β ⟩;
do forever
    ⟨ run the algorithm of Figure 1, ensuring that t ≥ t̲
      and using ε in the stopping criteria ⟩
    ⟨ increase t̲ and decrease ε ⟩;
enddo
```

FIG. 2. *The "three-level" bundle algorithm.*

reasonable) assumptions:

$$D_t = \frac{1}{t} D \Rightarrow D_t^* = \frac{1}{t} D^*(t\cdot), \tag{3.3}$$

where D satisfies (P2) and (P2) and is finite everywhere ($\Rightarrow$ (P5));

$$D_t^* = tD^* \Rightarrow D_t = tD\left(\frac{1}{t}\cdot\right), \tag{3.4}$$

where D^* satisfies (P*1) and (P*2) and is strictly convex in 0 ($\Rightarrow$ (P*5) + (P*5′)).

Of course, conditions equivalent to (P3)/(P*3) (D strongly coercive/D^* finite everywhere) or (P3′)/(P*3′) will also be required, whereas (P4)/(P*4) come directly from the nonnegativity of D/D^*.

Finally, let us record for future use two useful consequences of (P1)–(P5), the second being just that a penalty method using D_t works.

LEMMA 3.3. *$\forall \varepsilon > 0\ \forall \delta > 0$ there exists a $\underline{t}$ such that $S_\varepsilon(D_t) \supseteq B_2(\delta) \forall t \geq \underline{t}$.*

Proof. $\{D_t\}$ converges *uniformly* to $0(\cdot)$ on every compact set C, i.e., $\forall \varepsilon > 0$ there exists a $\underline{t}$ such that $D_t(d) \leq \varepsilon\ \forall d \in C, \forall t \geq \underline{t}$: use (P5), [HL93a, Theorem IV.3.1.5], and the fact that *ri dom* $0(\cdot) = \Re^n$. The result follows, using $C = B_2(\delta)$, since, due to (P4), $s_\varepsilon(D_t)$ are nondecreasing in t. □

LEMMA 3.4. *For any fixed $\bar{x}$, $\lim_{t\to\infty} v(\Pi_{\bar{x},t}) = v(\Pi)$.*

Proof. Note that $v(\Pi_{\bar{x},t})$ is nonincreasing in t by (P4). Assume by contradiction $\lim_{t\to\infty} v(\Pi_{\bar{x},t}) = \underline{v} > v(\Pi)$, i.e., one $\bar{d}$ exists such that $f(\bar{x} + \bar{d}) < \underline{v}$: using (P5), we get

$$\underline{v} = \lim_{t\to\infty} v(\Pi_{\bar{x},t}) \leq \lim_{t\to\infty} [f(\bar{x} + d) + D_t(\bar{d})] = f(\bar{x} + \bar{d}) < \underline{v}. \qquad \square$$

4. The bundle algorithm. We will analyze two main variants of the generalized bundle algorithm, described, respectively, in Figures 1 and 2.

The "two-level" bundle algorithm of Figure 1 implements the standard ideas of a bundle approach: the generalized Moreau–Yosida regularization ϕ_t of f (cf. section 1.4) is minimized (2nd level), with sequences of consecutive null steps performing

the approximate computation of $\phi_t(\bar{x})$ (1st level). The algorithm of Figure 2 adds another level, where t is forced to increase, possibly to $+\infty$; this is useful for those cases in which, due to properties of D_t, the standard two-level approach is not able to guarantee convergence unless t is "large enough."

In order to obtain a convergent algorithm, assumptions are needed about the following:

- properties of the stabilizing term D_t,
- choice of the model f_β,
- properties of the function f,
- handling of the t parameter (t-strategy) and the NS/SS decision,
- handling of the bundle (β-strategy).

The required properties for D_t have been described in the previous section. We will always assume f_β to be a closed convex function such that $f_\beta \le f$; for some results, f_β will be required to be the cutting plane model $\hat{f}_\beta$ (1.6). The assumptions on the last three points will be discussed in the following.

4.1. Assumptions on f. For some variants of the algorithm, we will require f to be a *-*compact* function, i.e., such that

$$e(l,L) := \sup_x\{d_{S_l(f)}(x) : x \in S_L(f)\} < \infty \qquad \forall L \ge l > v(\Pi) \ge -\infty.$$

Here f is *-compact if the excess of any level set $S_L(f)$ over $S_l(f)$ is finite; that is, f never becomes "infinitely flat."

Let us now briefly present some properties of *-compact functions that are useful in our treatment. (The interested reader is referred to [Fr98] for a more detailed study.) Recall that a nonempty closed convex set C is compact if and only if its *asymptotic cone*

$$C_\infty = \{d : x + \alpha d \in C \ \forall x \in C, \forall \alpha \ge 0\}$$

is the set $\{0\}$ (see [HL93a, Proposition III.2.2.3]); all the nonempty level sets of a closed convex function f have the same asymptotic cone (see [HL93a, Proposition IV.3.2.5]), denoted by f_∞.

THEOREM 4.1. *If $\forall L > v(\Pi)$ there exists a compact set C_L such that $S_L(f) \subseteq C_L + f_\infty$, then f is *-compact.*

Proof. Select $L \ge l > v(\Pi)$ and choose any $x_l \in S_l(f)$ (there must be at least one) to be kept fixed. From the hypothesis, for any $\bar{x} \in S_L(f)$ there exists an $x_L \in C_L$ and a $d \in f_\infty$ such that $\bar{x} = x_L + d$. Since $x_l + d \in S_l(f)$, we obtain

$$\inf_x\{\|x - \bar{x}\| : x \in S_l(f)\} \le \|(x_l + d) - (x_L + d)\| = \|x_l - x_L\|.$$

Therefore,

$$\sup_x \left\{\inf_x\{\|x - \bar{x}\| : x \in S_l(f)\} : \bar{x} \in S_L(f)\right\}$$
$$\le \sup_x\{\|x_l - x\| : x \in C_L\} < \infty, \qquad \text{since } C_L \text{ is compact.} \qquad \square$$

Note that C_L is not required to be convex, just compact.

COROLLARY 4.2. *All polyhedral functions are *-compact.*

Proof. The level sets of polyhedral functions are obviously polyhedra. Any polyhedron has a minimal representation as the sum of a (compact) polytope and a polyhedral cone [Ro70, Theorem 19.1]. The cone appearing in the minimal representation of each level set can only be f_∞. $\square$

Note that the hypothesis of Theorem 4.1 is obviously true if $f_\infty = \{0\}$, i.e., all inf-compact functions are *-compact. The converse is not true, however, since by Corollary 4.2 there are *-compact functions that are not inf-compact; *-compactness properly generalizes inf-compactness. It is easy to prove that many other functions are *-compact, such as the quadratic ones. *-compactness is a powerful assumption, since it allows us to prove the following result.

LEMMA 4.3. *If f is *-compact, then for any $\infty > L \geq l > v(\Pi)$ and $\varepsilon > 0$ there exists a $\underline{t} > 0$ such that $v(\Pi_{\bar{x},t}) \leq l + \varepsilon \ \forall \bar{x} \in S_L(f)$ and $t \geq \underline{t}$.*

Proof. Given any $\bar{x} \in S_L(f)$, call $\hat{x}$ the projection of $\bar{x}$ over $S_l(f)$; since $f(\bar{x}) \leq L$, using *-compactness, one has $\|\hat{x} - \bar{x}\| \leq e(l, L) = \delta < \infty$. By Lemma 3.3, there exists $\underline{t}$ such that $S_\varepsilon(D_t) \supseteq B_2(\delta) \ \forall t \geq \underline{t}$, and therefore $v(\Pi_{\bar{x},t}) \leq f(\hat{x}) + D_t(\hat{x} - \bar{x}) \leq l + \varepsilon \ \forall t \geq \underline{t}$. ∎

Using the above property, we can supplement Lemma 3.4, proving "convergence" for the optimal value of (1.12) for every "reasonable" choice of the sequences $\{\bar{x}_i\}$ and $\{t_i\}$.

LEMMA 4.4. *If f is *-compact, then for any sequence $\{\bar{x}_t\}$ such that $f(\bar{x}_t) \leq L < \infty$,*

$$\underline{v} := \liminf_{t \to \infty} v(\Pi_{\bar{x}_t,t}) = v(\Pi).$$

Proof. Assume by contradiction that $v(\Pi) < l = \underline{v} - 3\varepsilon$ for some $\varepsilon > 0$. Applying Lemma 4.3, we obtain that, for large enough t, $v(\Pi_{\bar{x}_t,t}) \leq \underline{v} - 3\varepsilon + \varepsilon$. Furthermore, from the definition of $\underline{v}$, there exists a large enough t such that $\underline{v} \leq v(\Pi_{\bar{x}_t,t}) + \varepsilon$. Hence, for this (large enough) t,

$$\underline{v} \leq v(\Pi_{\bar{x}_t,t}) + \varepsilon \leq \underline{v} - 2\varepsilon + \varepsilon + \varepsilon < \underline{v}. \qquad ∎$$

A final observation has to be made about polyhedral functions. In order to prove finite convergence results, a natural (but in principle nontrivial) assumption about the black box is required: as f is characterized by a finite set of vectors and their f^*-values, (cf. (1.6)), the black box has to return as subgradients only those "extreme" vectors characterizing f. More generally, one could require

(4.1)
only *finitely many different* pairs $(f^*(z), z)$ can be returned by the black box.

4.2. Assumptions on the *t*-strategy and the NS/SS decision. In order to leave a large degree of freedom in the implementation of the algorithm, we prove convergence under four general rules; several different t-strategies, with different performances in practice, can be designed following these guidelines [Fr97, Chapter I.5]. Since these rules measure improvements w.r.t. the current value $f(\bar{x})$, let us introduce the following notation:

$$\delta_{\bar{x}}(d) = f(\bar{x} + d) - f(\bar{x}) \qquad \text{is the } \textit{actual} \text{ improvement and} \tag{4.2}$$

$$\delta_{\beta,\bar{x}}(d) = f_\beta(\bar{x} + d) - f(\bar{x}) \qquad \text{is the } \textit{predicted} \text{ improvement} \tag{4.3}$$

for a step at $\bar{x}+d$. Note that $\delta_{\bar{x}}(d) - \delta_{\beta,\bar{x}}(d) = \Delta f$, and that $\delta_{\beta,\bar{x}}(d^*) \leq 0$. (Otherwise, $d = 0$ would be a better solution of (1.8) than d^*.) In the following, we will use "SS" as a shorthand for "serious step," i.e., an iteration of the algorithm where the current point $\bar{x}$ is changed. Analogously, "NS" will stand for "null step," i.e., an iteration of the algorithm where $\bar{x}$ is not changed.

(4.i) If an SS is performed, then

$$\delta_{\bar{x}}(d^*) \le m\delta_{\beta,\bar{x}}(d^*) \tag{4.4}$$

for a fixed $m \in (0,1)$; the converse is *not* required, i.e., an SS may not be done even if a "considerable" improvement has been obtained, except for what is required by (4.iii) below.

(4.ii) During a sequence of *consecutive* NS, t can *increase* only *finitely many times*.

(4.iii) During a sequence of *consecutive* NS, (4.4) can happen only *finitely many times*; that is, after finitely many NS, *any* step such that

$$\delta_{\bar{x}}(d^*) > m\delta_{\beta,\bar{x}}(d^*) \tag{4.5}$$

must be accepted.

(4.iv) During a sequence of *consecutive* NS, at all iterations (but possibly a finite number) f must be evaluated in $\bar{x}+d^*$, and the model f_+ of the following iteration must take into account the corresponding $z \in \partial f(\bar{x}+d^*)$, in the sense that $f_+^*(z) \le f^*(z)$.

Let us briefly discuss the above rules. By (4.i), an SS is performed only if a consistent improvement is obtained. Changing the current point is not mandatory if some alternative strategy—typically increasing t—appears to be preferable, but, by (4.iii), this must not happen forever. A reasonable answer to a "bad" step is to decrease t; increasing t is also possible, but it must be properly limited, e.g., by (4.ii). Finally, inserting the newly obtained subgradient into β is not mandatory if some alternative strategy—typically decreasing t—appears to be preferable, but, by (4.iv), this must not happen forever. Using $f_+^* \ge f^*$, (4.iv) is equivalent to $f_+^*(z) = f^*(z)$; from the primal viewpoint, it says that

$$f(\bar{x}+d^*) = f_+(\bar{x}+d^*) \quad \text{and} \quad z \in \partial f_+(\bar{x}+d^*). \tag{4.6}$$

In some cases, a strengthened form of rule (4.ii) is useful, as follows.

(4.ii′) During a sequence of *consecutive* NS, t can *change* only *finitely many times*.

A consequence of rules (4.ii) (or (4.ii′)), (4.iii), and (4.iv) is that, for any sequence of consecutive NS, there exists an iteration index h such that for all the subsequent iterations in the sequence, t is nonincreasing (fixed), $\delta_{\bar{x}}(d^*) > m\delta_{\beta,\bar{x}}(d^*)$, and z is added to β. In the following, we will often refer to this h.

Inhibiting serious steps allows us to drop the *-compactness assumption in some variants of the algorithm; thus, we will sometimes use the following rule.

(4.iii′) Only *finitely many* SS are done; after the last one, the stopping condition becomes $\Delta f \le \varepsilon$.

This rule is rather abstract, but several practical implementations can be imagined. For instance, the current point can just be kept fixed. Alternatively, if $v(\Pi)$ is finite, one could choose some $\varepsilon > 0$ and inhibit SS if $\delta_{\beta,\bar{x}}(d^*) \ge -\underline{\varepsilon}$ (a "negligible" step), as long as the t-strategy is properly managed.

With the three-level algorithm of Figure 2, sometimes the following weakened form of (4.iii′), which allows any *total* number of SS to be performed, suffices.

(4.iii″) For *each run* of the two-level bundle algorithm, only finitely many SS are done; after the last one, the stopping condition becomes $\Delta f \le \varepsilon$.

At the end of this section, let us remark that the very concept of SS, although apparently primal in nature, has a noteworthy "dual interpretation." From the dual viewpoint, a bundle method is an approximated ascent approach to $\sup_x\{v(\Delta_{x,t})\}$,

where an ascent in the value of the stabilized dual problem (1.11), i.e., $v(\Delta_{\bar{x}+d^*,t}) \geq v(\Delta_{\bar{x},t})$, is desired. Unfortunately, the values of $v(\Delta_{\bar{x}+d^*,t})$ and $v(\Delta_{\bar{x},t})$ are unknown, and therefore the condition cannot be checked; however, they can be estimated, using the dual pricing problem (1.3) ($v(\Delta_x) = -f(x)$) and the stabilized dual master problem (1.9), as

$$v(\Delta_{\bar{x}+d^*,t}) \geq v(\Delta_{\bar{x}+d^*}) \quad \text{and} \quad v(\Delta_{\beta,\bar{x},t}) \geq v(\Delta_{\bar{x},t}) \geq v(\Delta_{\bar{x}}).$$

(Remember Lemma 2.1: $(\Delta_{\bar{x}+d^*})$ is a linearization of $(\Delta_{\bar{x}+d^*,t})$ in $-z^*$ using the subgradient d^*.) Now, (4.4) is equivalent, via (2.6), to

$$v(\Delta_{\bar{x}+d^*}) \geq mv(\Delta_{\beta,\bar{x},t}) + (1-m)v(\Delta_{\bar{x}}).$$

Therefore, $v(\Delta_{\bar{x}+d^*})$ and $mv(\Delta_{\beta,\bar{x},t})+(1-m)v(\Delta_{\bar{x}})$ are taken as estimates of $v(\Delta_{\bar{x}+d^*,t})$ and $v(\Delta_{\bar{x},t})$, respectively, and used to decide whether $\bar{x}+d^*$ are better multipliers than $\bar{x}$. Note that there is a safeguard against "wild" decisions: at least, $v(\Delta_{\bar{x}+d^*}) \geq v(\Delta_{\bar{x}})$. Hence, even if $v(\Delta_{\bar{x},t})$ does not actually improve moving to $\bar{x}+d^*$, at least its lower approximation $v(\Delta_{\bar{x}})$ does.

4.3. Assumptions on the β-strategy. An important detail of any implementable bundle method is the *β-strategy*, i.e., how the information in β is managed to keep the computational cost of the solution of (1.8)/(1.9) reasonably low. Removing subgradients from β is important in practice, but heedless removals can impair convergence of the algorithm. A "minimal" requirement for any β-strategy is the following.

DEFINITION 4.5. *A β-strategy is* weakly monotone *if, during a sequence of consecutive NS, for each $i \geq h$ the optimal value of* (1.9) *is monotonically nonincreasing or, equivalently, the optimal value of* (1.8) *is monotonically nondecreasing.*

The equivalence between the two conditions in Definition 4.5 is (2.1). A weakly monotone β-strategy ensures at least convergence (to some value) of the optimal value of (1.8)/(1.9) during a sequence of consecutive NS. The definition does not specify how that monotonicity is obtained; a pretty minimal assumption on f_β is the following.

DEFINITION 4.6. *A β-strategy is* monotone *if, during a sequence of consecutive NS, for each $i \geq h$*

$$f^*_{\beta_{i+1}}(z^*_i) \leq f^*_{\beta_i}(z^*_i), \tag{4.7}$$

or, equivalently,

$$f_{\beta_{i+1}}(\bar{x}+d) \geq f_{\beta_i}(\bar{x}+d^*_i) + z^*_i(d-d^*_i) \quad \forall d. \tag{4.8}$$

The equivalence between (4.7) and (4.8) can be easily proved using (2.4) and (1.1). A monotone β-strategy is weakly monotone; since $t_{i+1} \leq t_i \Rightarrow D^*_{t_{i+1}} \leq D^*_{t_i}$ for $i \geq h, v(\Delta_{\bar{x},\beta_{i+1},t_{i+1}}) \leq f^*_{\beta_{i+1}}(z^*_i) - z^*_i\bar{x} + D^*_{t_{i+1}}(-z^*_i) \leq f^*_{\beta_i}(z^*_i) - z^*_i\bar{x} + D^*_{t_i}(-z^*_i) = v(\Delta_{\bar{x},\beta_i,t_i})$.

The practical implementation of a monotone β-strategy depends on the model. For the cutting plane model $\hat{f}_\beta$, at each iteration the following two *moves* can be considered:

- remove some z from β (removal),
- add z^* to β (*aggregation*), with f^*-value $\hat{f}^*_\beta(z^*)$.

Rewriting (1.9) with $f_\beta = \hat{f}_\beta$ in the following equivalent form (cf. (1.7))

$$\inf_\theta \{\Sigma_{z\in\beta}(f^*(z) - z\bar{x})\theta_z + D^*_t(-\Sigma_{z\in\beta}z\theta_z) : \Sigma_{z\in\beta}\theta_z = 1, \theta \geq 0\}, \tag{4.9}$$

it is clear that aggregation offers one way for implementing a monotone β-strategy. If—as often happens—(1.9) is actually solved via (4.9), an alternative is to just avoid discarding all the $z \in \beta$ whose corresponding optimal multiplier θ_z^* is strictly positive, as

$$\hat{f}_\beta^*(z^*) = \sum_{z \in \beta} f^*(z)\theta_z^* \quad \text{and} \quad z^* = \sum_{z \in \beta} z\theta_z^*.$$

In principle, no more than $n+1$ of the optimal multipliers need to be strictly positive, although in practice whether or not such a minimal solution is obtained depends on the actual solver; even for $D_t^* = \frac{1}{2}t\|\cdot\|_2^2$, active-set algorithms [Ki89, Fr96] would guarantee it, while interior-point algorithms may not. The above discussion justifies the following result.

LEMMA 4.7. *If $f_\beta = \hat{f}_\beta$ and, during a sequence of consecutive NS, for each iteration after h either all the z such that $\theta_z^* > 0$ are kept in β or z^* is added to β with $\hat{f}_\beta^*(z^*)$ as the corresponding f^*-value, then the β-strategy is monotone.*

A monotone β-strategy allows us to keep the size of β bounded (down to 2); if (P3″) does not hold, however, it is not sufficient to guarantee convergence [Fr97, section I.4.2]. A stronger property has to be used, which essentially inhibits all removals at length.

DEFINITION 4.8. *A β-strategy is* strictly monotone *if it is monotone and, if some z has been removed from β, no other removal is permitted until $v(\Pi_{\beta,\bar{x},t})$ increases by a fixed $\mu > 0$.*

A strictly monotone β-strategy guarantees convergence for every choice of D_t; although it does not give any finite bound on the size of β, it can still be practical. Furthermore, there is a trade-off between the size of β—hence the computational cost of (1.9)—and the speed of convergence of the overall process [HL93b, section XIV.4.5]; a small β is a good choice only in some cases [CFG01].

Finally, if f is a polyhedral function, *finite* termination to an optimal solution can be proved, provided that aggregation is properly limited.

DEFINITION 4.9. *A β-strategy is* safe *if only finitely many aggregations are done.*

5. Convergence of NS sequences (1st level). The convergence proof is divided into three parts. In this section we assume that no SS occurs, i.e., we examine infinite sequences of consecutive NS; we shall show that these sequences allow us to compute the generalized Moreau–Yosida regularization with any finite precision. Therefore, in the next section we will be allowed to disregard what happens between two consecutive SS, i.e., focus on the convergence of the minimization process of the generalized Moreau–Yosida regularization (2nd level). Finally, in section 7 we will discuss the convergence of the 3rd level.

In this section, the iteration index i denotes the ith NS of the (only) infinite sequence of consecutive NS that the algorithm is supposed to perform, and therefore the current point $\bar{x}$ is fixed. The iteration index h is the one implied by the rules (4.ii) (or (4.ii′)), (4.iii), and (4.iv). To simplify the notation, let (Δ_i) and (Π_i) denote, respectively, the dual and primal stabilized master problems (1.9) and (1.8) solved in that iteration; z_i^* and d_i^* their solutions; $f_i(\hat{f}_i)$ the corresponding (cutting plane) model; z_i the subgradient reported by the evaluation of $f(\bar{x} + d_i^*)$; δ_i the predicted improvement; and so on. Also, we will use the shorthand index "+" for "$i+1$."

In the following, we will always assume that (P1), (P2), (P4), and (P5) hold; additional assumptions will be explicitly listed. The first step in the convergence proof is to show that the algorithm is well defined, i.e., that the primal and dual

stabilized master problems have optimal solutions. This requires (P3) or (P3′), as well as minimal assumptions on f_β.

LEMMA 5.1. *Under the hypothesis of Lemma* 2.1*, if either* (P3′) *or* (P3) *hold, then* (Δ_i) *and* (Π_i) *attain finite optimal solutions* z_i^* *and* d_i^**, respectively.*

Proof. If (P3′) holds, then from $f_i \geq f_*$ we have that $f_i(\bar{x}+d)+D_{t_i}(d) > f(\bar{x})\ \forall d \notin S_\delta(D_{t_i})$, where $\delta = f(\bar{x}) - f_*$. Since $S_\delta(D_{t_i})$ is compact by (P2), the infimum must be finitely attained. Otherwise (P3) holds, i.e., D_t is strongly coercive; hence $f_\beta(\bar{x} + \cdot) + D_t(\cdot)$ is strongly coercive too. (Strongly coercive functions increase faster than any linear function at infinity, and any convex function is minorized by an affine function [HL93a, Proposition IV.1.2.1].) Therefore, (Π_i) has a bounded nonempty set of optimal solutions [HL93a, Remark IV.3.2.8]. Finally, [HL93a, Theorem X.2.3.2] shows that an optimal z_i^* exists for (Δ_i) whenever an optimal d_i^* exists for (Π_i). ∎

We will now focus on proving the boundedness of the sequences $\{d_i^*\}$ and/or $\{z_i^*\}$, under a set of different assumptions.

LEMMA 5.2. *Under the hypothesis of Lemma* 2.2*, if* (P3′) *holds, then* $\{d_i^*\}$ *and* $\{z_i^*\}$ *are bounded.*

Proof. Boundedness of $\{d_i^*\}$ was in fact established in Lemma 5.1, as $D_{t_i}(d_i^*) \leq \delta = f(\bar{x}) - f_*$ and, by (P4), $S_\delta(D_{t_i}) \subseteq S_\delta(D_{t_h})$; the latter is compact by (P2). By (2.3) and $f_i \leq f$, z_i^* is an ε_i-subgradient of f at $\bar{x}+d_i^*$ for $\varepsilon_i = \Delta f_i = f(\bar{x}+d_i^*) - f_i(\bar{x}+d_i^*) \geq 0$; since f is finite everywhere, and therefore bounded over any compact set, and $f_i \geq f_*$, $\varepsilon_i \leq \bar{\varepsilon} < \infty$. Hence, $z_i^* \in \partial_{\bar{\varepsilon}} f(\bar{x} + d_i^*)\ \forall i \geq h$. The image of a compact set in int *dom* $f = \Re^n$ under the $\bar{\varepsilon}$-subdifferential mapping (see [HL93b, Proposition XI.4.1.2]) is compact. ∎

Thus, (P3′)/(P*3′) guarantee the boundedness of both solution sequences. In all the development, the first part of (P2) (compactness) is only used in Lemma 5.2, and therefore it could be dropped if (P3) holds; however, strong coercivity implies the boundedness of the level sets [HL93a, Proposition IV.3.2.5.(ii)], and hence there is no loss of generality—and a gain in symmetry—in requiring it to hold in general.

In Lemma 5.2, the boundedness of $\{z_i^*\}$ is obtained as a consequence of the boundedness of $\{d_i^*\}$, finiteness of f, and $f_i \geq f_*$; with basically the same argument, it is possible to prove the boundedness of $\{d_i^*\}$, *given* the boundedness of $\{z_i^*\}$ and (P3)/(P*3).

LEMMA 5.3. *Under the hypothesis of Lemma* 2.2*, if* (P*3) *holds and* $\{z_i^*\}$ *is bounded, then* $\{d_i^*\}$ *is bounded.*

Proof. By (2.2), $d_i^* \in \partial D_{t_i}^*(-z_i^*)$. From $t_i \leq t_h (\Rightarrow D_{t_h}^* \geq D_{t_i}^*$ by (P*4)), it is clear that d_i^* must also be an ε-subgradient of $D_{t_h}^*$ for a proper ε; indeed, it is easy to check that d_i^* is a ε_i-subgradient of $D_{t_i}^*$ at $-z_i^*$ for $\varepsilon_i = D_{t_h}^*(-z_i^*) \geq D_{t_h}^*(-z_i^*) - D_{t_i}^*(-z_i^*) \geq 0$. Since, by (P*3), $D_{t_h}^*$ is finite convex (hence continuous) and $\{z_i^*\}$ is bounded, $\varepsilon_i \leq \bar{\varepsilon} < \infty$. Hence, $d_i^* \in \partial_{\bar{\varepsilon}} D_{t_h}^*(-z_i^*)\ \forall i \geq h$; reasoning as in Lemma 5.2, we obtain that $\{d_i^*\}$ is bounded. ∎

Boundedness of $\{z_i^*\}$ under (P3)/(P*3) is not easy to establish in general; however, when obtained, it allows us to prove the boundedness of all the relevant sequences, as the same argument proves boundedness of $\{z_i\}$, given the boundedness of $\{d_i^*\}$.

LEMMA 5.4. *Under the hypothesis of Lemma* 2.1*, if* $\{d_i^*\}$ *is bounded, then the sequences* $\{f^*(z_i)\}$ *and* $\{z_i\}$ *obtained by evaluating* $f(\bar{x} + d_i^*)$ *are bounded.*

Proof. Since f is finite everywhere and $z_i \in \partial f(\bar{x} + d_i^*)$, we can invoke [HL93a, Remark VI.6.2.3] to conclude that all the z_i belong to a compact set. Also, $-f(0) \leq f^*(z_i) = z_i(\bar{x} + d_i^*) - f(\bar{x} + d_i^*)$, which is bounded above since both $\{d_i^*\}$ and $\{z_i\}$ are bounded and f is bounded below over any compact set. ∎

Note that the above results do not depend on the β-strategy. Indeed, there are several situations in which the boundedness of $\{d_i^*\}$ is "free"; among them, let us mention the following:

- $dom(D_{t_h})$ is compact (D_t is a "trust region," cf. Example 3.3), as $dom(D_{t_i}) \subseteq dom(D_{t_h})$ by (P4);
- (P3) holds and $dom(f^*)$ is compact (f is globally Lipschitz, e.g., polyhedral), as $dom(f_i^*) \subseteq dom(f^*)$ and Lemma 5.3 gives boundedness of $\{d_i^*\}$.

Conversely, let us mention that the boundedness of $\{d_i^*\}$ implies the boundedness of $\{z_i^*\}$ whenever the cutting plane model $\hat{f}_\beta$ is used, as from (1.7) every z_i^* belongs to the convex hull of $\{z_i\}$ and, from Lemma 5.4, the latter set is bounded whenever $\{d_i^*\}$ is.

5.1. Results with a weakly monotone β-strategy. We will now prove some results which only require a weakly monotone β-strategy and $f_\beta \le f$. The basic observation is that, with a weakly monotone β-strategy, by Definition 4.5 we have, $\forall i \ge h$,

(5.1)
$$D_{t_i}^*(-z_i^*) \le f_i^*(z_i^*) - z_i^*\bar{x} + f(\bar{x}) + D_{t_i}^*(-z_i^*) = v(\Delta_i) + f(\bar{x}) \le v(\Delta_h) + f(\bar{x}) < \infty$$

(use (1.vii) and $f_i^* \ge f^*$). In the proximal bundle case ($D_t^* = \frac{1}{2}t\|\cdot\|_2^2$), where $d_i^* = -t_i z_i^*$, (5.1) proves the boundedness of $\{d_i^*\}$; this is also true in the more general case, provided that D_t has the form (3.3). The proof relies on the following "primal view" of (5.1),

$$D_{t_i}(0) - D_{t_i}(d_i^*) - (-z_i^*)(0 - d_i^*) \le v(\Delta_h) + f(\bar{x}) < \infty \tag{5.2}$$

(use (2.5) and (P1)); (5.2) can be expressed by saying that the *linearization error* (cf. (1.10)) in 0, made by approximating D_{t_i} with its linearization in d_i^* using slope $-z_i^*$, is bounded.

LEMMA 5.5. *Under the hypothesis of Lemma* 2.2, *if* (P3) *holds, a weakly monotone β-strategy is used, and D_t has the form* (3.3), *then $\{d_i^*\}$ is bounded.*

Proof. From (2.2), $-z_i^* \in \partial D_{t_i}(d_i^*)$; since D_t has the form (3.3), defining $\bar{z}_i^* := t_i z_i^*$, we have that $-\bar{z}_i^* \in \partial D(d_i^*)$. Hence, (5.2) can be written as

$$\frac{1}{t_i}D(0) - \frac{1}{t_i}D(d_i^*) - \frac{1}{t_i}(-\bar{z}_i^*)(0 - d_i^*) \le \varepsilon < \infty,$$

whence $D(0) - D(d_i^*) - (-\bar{z}_i^*)(0 - d_i^*) \le \varepsilon t_i \le \varepsilon t_h < \infty$.

Now, call $\bar{V}_\varepsilon(\bar{d})$ the set of all d such that the linearization error in $\bar{d}$, made by approximating D with its linearization in d using any $z \in \partial D(d)$, is smaller than ε. Since D is strongly coercive, $\bar{V}_\varepsilon(\bar{d})$ is compact for any $\bar{d}$ and any fixed ε [HL93b, Proposition XI.4.2.6.(i)]. □

An alternative result does not require (3.3) but rather that t_i remain bounded away from zero; actually, in this case the boundedness of $\{z_i^*\}$ is obtained, which, in view of Lemma 5.3, is a stronger result.

LEMMA 5.6. *Under the hypothesis of Lemma* 2.2, *if* (P3) *holds, a weakly monotone β-strategy is used, and $t_i \ge \underline{t} > 0$ (t is bounded away from 0), then the sequences $\{f_i^*(z_i^*)\}$ and $\{z_i^*\}$ are bounded.*

Proof. From (5.1), (P*4), and $\underline{t} \le t_i$ we have $D_{\underline{t}}^*(-z_i^*) \le v(\Delta_h) + f(\bar{x}) < \infty$; the level sets of $D_{\underline{t}}^*$ are compact from (P*2), and hence $\{z_i^*\}$ is bounded. Looking again

at (5.1), we notice that $f_i^*(z_i^*)$ is bracketed between bounded quantities; hence it is also bounded. □

In practice, t should not become too small anyway, so the condition in the above lemma is not really binding; yet, in many cases it can simply be dropped.

5.2. Convergence with a monotone β-strategy. The above boundedness results are instrumental for proving the actual convergence of a sequence of NS, that is, the fact that the stabilized master problems (1.8) and (1.9) can be used to approximate the stabilized problems (1.12) and (1.11) within any required degree of accuracy. Due to (2.7), it is only necessary to prove that $\{\Delta f_i\} \to 0$. With a monotone β-strategy, this requires (P*3″).

Consider the (convex) function

$$r_i(z) := f_i^*(z) - z\bar{x} + f(\bar{x}),$$

so that $r_i(z) + D_{t_i}^*(z)$ is, but for the constant $f(\bar{x})$, the objective function of (Δ_i); from $f_i^* \geq f^*$ and (1.vii), $r_i \geq 0$. Now define

$$\zeta_i := z_i - z_i^* \quad \text{and} \quad z_i(\gamma) := z_i^* + \gamma\zeta_i.$$

From Definition 4.6 ($f_+^*(z_i^*) \leq f_i^*(z_i^*)$) and (4.iv) ($f_+^*(z_i) \leq f^*(z_i)$) we have that, $\forall \gamma \in [0,1]$,

$$\begin{aligned} h_i(\gamma) &:= [f_i^*(z_i^*)(1-\gamma) + f^*(z_i)\gamma - z_i(\gamma)\bar{x} + f(\bar{x})] + D_{t_i}^*(-z_i(\gamma)) \\ &\geq [f_+^*(z_i^*) - z_i^*\bar{x} + f(\bar{x})](1-\gamma) + [f_+^*(z_i) - z_i\bar{x} + f(\bar{x})]\gamma + D_{t_i}^*(-z_i(\gamma)) \\ &\geq r_+(z_i^*)(1-\gamma) + r_+(z_i)\gamma + D_{t_+}^*(-z_i(\gamma)) \geq r_+(z_i(\gamma)) + D_{t_+}^*(-z_i(\gamma)). \end{aligned}$$

(We have also used $t_+ \leq t_i \Rightarrow D_{t_+}^* \leq D_{t_i}^*$ and the convexity of r_+.) Therefore, defining

$$(\vartheta_i) \qquad \min_\gamma \{h_i(\gamma) : \gamma \in [0,1]\},$$

$$v_i(z) := \begin{cases} r_i(z) + D_{t_i}^*(-z) & \text{if } z = z_i(\gamma) \text{ for some } \gamma \in [0,1], \\ +\infty & \text{otherwise,} \end{cases}$$

one clearly has

$$v(\vartheta_i) \geq \min_z \{v_+(z)\} \geq v(\Delta_+) + f(\bar{x}).$$

We will study the behavior of $v(\vartheta_i)$ during sequences of consecutive NS to estimate the convergence speed of $v(\Delta_i)$. In practice, this corresponds to the "aggressive" monotone β-strategy, where aggregation is performed at every step and all subgradients but z_i^* and z_i are discarded.

Due to (P*3″), $D_{t_i}^*$ is differentiable, and hence so is h_i; from (2.2), we have $d_i^* = \nabla D_{t_i}^*(-z_i(0))$; hence by (2.6)

$$h_i'(0) = f^*(z_i) - f_i^*(z_i^*) - (z_i - z_i^*)(\bar{x} + d_i^*) = -\Delta f_i. \tag{5.3}$$

Using (1.vi) and (2.4), the NS condition (4.5) can be written as

$$z_i(\bar{x} + d_i^*) - f^*(z_i) - f(\bar{x}) > m[z_i^*(\bar{x} + d_i^*) - f_i^*(z_i^*) - f(\bar{x})];$$

hence

$$h_i'(0) < (1-m)[z_i^*(\bar{x}+d_i^*) - f_i^*(z_i^*) - f(\bar{x})].$$

From (1.vi), $-z_i^* d_i^* = D_{t_i}(d_i^*) + D_{t_i}^*(-z_i^*) \geq D_{t_i}^*(-z_i^*)$; hence

$$h_i'(0) < (1-m)[-f_i^*(z_i^*) + z_i^*\bar{x} - D_{t_i}^*(-z_i^*) - f(\bar{x})] = -(1-m)h_i(0). \tag{5.4}$$

Using (5.4) it is possible to show, adapting standard results from smooth optimization [OR70], that $\{-h_i'(0) = \Delta f_i\} \to 0$ if $\{z_i^*\}$ is bounded and (P*3″) holds. In the general case, this requires some assumptions on the behavior of t_i, the simplest one being rule (4.ii′).

THEOREM 5.7. *Under the hypothesis of Lemma* 2.2, *if* (P*3″) *holds, rule* (4.ii′) *is in force, a monotone* β*-strategy is used, and* $\{z_i^*\}$ *is bounded, then* $\{\Delta f_i\} \to 0$.

Proof. Wait until the iteration h implied by rules (4.ii′), (4.iii), and (4.iv): $t_i = t\ \forall i \geq h$. The boundedness of $\{z_i^*\}$, together with Lemma 5.3 ((P*3″) ⇒ (P*3)) and Lemma 5.4, implies that $\{z_i\}$ is also bounded; therefore, the set

$$Z := conv(\{z_i^*\} \cup \{z_i\})$$

is compact and contains all the segments $[z_i^*, z_i]$. From (P*3″), D_t^* is differentiable and therefore *continuously differentiable* [HL93a, Remark VI.6.2.6] on Z; that is, ∇D_t^* is continuous, and therefore *uniformly continuous*, on the compact set Z. Note that if $\nabla D_t^*(z_i^*) = d_i^* = d_+^* = \nabla D_t^*(z_+^*)$, then $\Delta f_+ = 0$ as, from (4.6), $f(\bar{x}+d_i^*) = f_+(\bar{x}+d_i^*) = f_+(\bar{x}+d_+^*)$. Hence

$$\sup\{\|\nabla D_t^*(z') - \nabla D_t^*(z'')\| : z', z'' \in Z\} > 0.$$

The *reverse modulus of continuity* of ∇D_t^* over Z

$$\kappa(v) := \inf\{\|z'-z''\| : \|\nabla D_t^*(-z') - \nabla D_t^*(-z')\| \geq v, z', z'' \in Z\}$$

is an F-function, i.e., nondecreasing and such that $\kappa(0) = 0$ and $\kappa(v) > 0$ for $v > 0$ [OR70, Definition 14.2.6]. (Our definition of κ is nonstandard, in that *dom* κ may not be the whole $\Re_+$, but we will always evaluate κ at points of its domain.)

We claim the existence of an F-function ρ such that

$$v(\Delta_i) - v(\Delta_+) \geq h_i(0) - v(\vartheta_i) \geq \rho(-h_i'(0)) > \rho((1-m)h_i(0)),$$

which clearly implies that $\{-h_i'(0) = \Delta f_i\} \to 0$ and $\{h_i(0)\} \to 0$ and therefore proves the theorem. (Note that $\{v(\Delta_i)\}$ is bounded below, as $v(\Delta_i) \geq -f(\bar{x})$.) The function ρ estimates how much of the decrease "promised" by $h_i'(0)$ is actually attained in the optimal solution of (ϑ_i).

A special case for which the estimate is easy is $z_i = z_i^*$, i.e., $\zeta_i = 0$: the corresponding $h_i(\gamma)$ is linear, the optimal solution of (ϑ_i) is $\gamma = 1$, and $h_i(1) = h_i(0) + h_i'(0)$; hence $\rho \equiv 1$.

Otherwise, for the reverse modulus of continuity of h_i' over $[0,1]$ one has

$$\begin{aligned}
\sigma_i(v) &:= \inf\{|\gamma'-\gamma''| : |h_i'(\gamma') - h_i(\gamma'')| \geq v, \gamma', \gamma'' \in [0,1]\} \\
&\geq \frac{1}{\|\zeta_i\|} \inf\{\|(\gamma'-\gamma'')\zeta_i\| : \|(\nabla D_t^*(-z_i(\gamma')) - \nabla D_t^*(-z_i(\gamma')))\zeta_i\| \\
&\qquad\qquad \geq v, \gamma', \gamma'' \in [0,1]\} \\
&\geq \frac{1}{\|\zeta_i\|} \inf\left\{\|z'-z''\| : \|\nabla D_t^*(z') - \nabla D_t^*(z'')\| \geq \frac{v}{\|\zeta_i\|}, z', z'' \in Z\right\}
\end{aligned}$$

(note that $\zeta_i \neq 0$), and therefore

$$\sigma_i(v) \geq \frac{1}{\|\zeta_i\|}\kappa\left(\frac{v}{\|\zeta_i\|}\right). \tag{5.5}$$

Now, define

$$\gamma^* := \inf_{\gamma}\left\{\gamma \geq 0 : h_i'(\gamma) \geq \frac{1}{2}h_i'(0)\right\} \geq \sigma_i\left(-\frac{1}{2}h_i'(0)\right). \tag{5.6}$$

(1/2 is arbitrary; any strictly positive number would do.) By (5.3), if $\gamma^* = 0$, then $\Delta f_i = 0$ and the theorem is proved. Otherwise, the following two cases may arise.

If $\gamma^* \geq 1$, then $\gamma = 1$ is the optimal solution of (ϑ_i) and $h_i'(\gamma) \leq \frac{1}{2}h_i'(0)\ \forall\gamma \in [0,1]$. ($h_i'$ is nondecreasing since h_i is convex.) In particular, $h_i'(1) \leq \frac{1}{2}h_i'(0) < 0$, and therefore

$$h_i(1) + h_i'(1)(0-1) \leq h_i(0) \Rightarrow h_i(1) \leq h_i(0) + \frac{1}{2}h_i'(0).$$

If $\gamma^* < 1$, then by the mean-value theorem there exists some $\bar\gamma \in (0,\gamma^*)$ such that

$$h_i(\gamma^*) = h_i(0) + h_i'(\bar\gamma)\gamma^* \Rightarrow h_i(\gamma^*) \leq h_i(0) + \frac{1}{2}h_i'(0)\gamma^*.$$

Hence, using (5.5) and (5.6),

$$v(\Delta_i) - v(\Delta_+) \geq h_i(0) - h_i(\gamma^*) \geq -\frac{1}{2\|\zeta_i\|}h_i'(0)\kappa\left(-\frac{1}{2\|\zeta_i\|}h_i'(0)\right).$$

Thus, the claim is proved with $\rho(v) = \frac{v}{2}\min\left\{1, \frac{1}{diam(Z)}\kappa\left(\frac{v}{2diam(Z)}\right)\right\}$, where $diam(Z) < \infty$ is the maximum distance of any two points in Z. □

In the above proof, rule (4.ii′) is needed because $t \to D_t^*(z)$ may be almost any function; rule (4.ii) suffices, thereby allowing $\{t_i\} \to 0$, if this function is "simple."

THEOREM 5.8. *Under the hypothesis of Lemma* 2.2, *if* (P*3″) *holds, a monotone* β*-strategy is used,* $\{z_i^*\}$ *is bounded, and* D_t^* *has either the form* (3.3) *or the form* (3.4), *then* $\{\Delta f_i\} \to 0$.

Proof. With the notations of Theorem 5.7, let $D_t^* = tD^*$ and call κ_i and κ the reverse modulus of continuity of $\nabla D_{t_i}^*$ and of ∇D^*, respectively, on Z. It is easy to check that

$$\sigma_i(v) \geq \frac{1}{\|\zeta_i\|}\kappa_i\left(\frac{v}{\|\zeta_i\|}\right) \geq \frac{1}{\|\zeta_i\|}\kappa\left(\frac{v}{t_i\|\zeta_i\|}\right) \geq \frac{1}{\|\zeta_i\|}\kappa\left(\frac{v}{t_h\|\zeta_i\|}\right),$$

as $t_i \leq t_h$ and κ is nondecreasing. If $D_t = \frac{1}{t}D$ instead, one has $D_t^*(z) = \frac{1}{t}D^*(tz)$ and therefore $\nabla D_t^*(z) = \nabla D^*(tz)$; simple calculations yield

$$\sigma_i(v) \geq \frac{1}{\|\zeta_i\|}\kappa_i\left(\frac{v}{\|\zeta_i\|}\right) \geq \frac{1}{t_i\|\zeta_i\|}\kappa\left(\frac{v}{\|\zeta_i\|}\right) \geq \frac{1}{t_h\|\zeta_i\|}\kappa\left(\frac{v}{\|\zeta_i\|}\right),$$

where κ is the reverse modulus of continuity of ∇D^* on the (compact) set $\{tz : z \in Z, t \in [0,t_h]\}$. In both cases, the proof of Theorem 5.7 can be easily adapted by using the above functions in place of the reverse modulus of continuity of D_t^* for the fixed t provided by rule (4.ii′). □

The similar theorem [Au87, Theorem 2.3] (with a different proof) is proved for a *fixed* t and D_t of the form (3.3), differentiable and satisfying (P3″). Note that differentiability—which would at first appear to be a natural assumption—is necessary *in the dual* rather than in the primal, the critical property of D_t being strict convexity. It is also interesting to note that in [Au87] a primal notation is used, but (1.9) is developed—only for the simple case $\beta_+ = \{z_i^*, z_i\}$—as a tool for proving [Au87, Theorem 2.3].

The above theorems rely on the compactness of Z, which is a consequence of the boundedness of $\{z_i^*\}$. The latter is, in several cases, either free or a consequence of the boundedness of $\{d_i^*\}$, which may not require t_i bounded away from 0 (cf. Lemma 5.5). Thus, these results generalize those available for the proximal bundle method. Indeed, applying Theorem 5.7 to $D_t^* = \frac{1}{2}t\|\cdot\|_2^2$, whose reverse modulus of continuity is $\kappa(v) = v/t$ (that does *not* depend on Z), one obtains an estimate that is only a 1/2 factor—due to the arbitrary 1/2 in the proof—away from the tightest possible one, if aggregation is allowed:

$$v(\Delta_i) - v(\Delta_+) \geq \frac{(1-m)(v(\Delta_i) + f(\bar{x}))}{2} \min\left\{1, \frac{(1-m)(v(\Delta_i) + f(\bar{x}))}{t_i \|z_i - z_i^*\|_2^2}\right\}.$$

The above estimate was obtained in [Fr97, Theorem I.2.2.2] (apart from a minor error) with basically the same arguments of Theorem 5.7, only using ad hoc relations.

Finally, note that all the results until now do not require f_β to be the cutting plane model, and therefore they can be used in the analysis of "nonstandard" bundle methods [GM91].

5.3. Convergence with a strictly monotone β-strategy. When (P3′) does not hold, a monotone β-strategy does not guarantee convergence [Fr97, section I.4.2], and strict monotonicity is required. Furthermore, the following strengthened form of the rule in (4.iv)

$$\exists \text{ an index } h \text{ such that}, \forall i > j \geq h, f^*(z_j) = f_i^*(z_j) \tag{5.7}$$

is required; that is, at length the "accuracy" of f_β^* as a model of f^* cannot "deteriorate" once it has become "exact" in the dual points z_i.

THEOREM 5.9. *Under the hypothesis of Lemma* 2.2*, if* (5.7) *holds, a strictly monotone β-strategy is used, and $\{d_i^*\}$ is bounded, then $\{\Delta f_i\} \to 0$.*

Proof. By (2.3), $\bar{x} + d_i^* \in \partial f_i^*(z_i^*)$; hence, using (5.7),

$$f^*(z_j) - (\bar{x} + d_i^*)z_j \geq f_i^*(z_j) - (\bar{x} + d_i^*)z_j \geq f_i^*(z_i^*) - (\bar{x} + d_i^*)z_i^* \quad \forall i > j. \tag{5.8}$$

From $z_i \in \partial f(\bar{x} + d_i^*)$ and $z_i^* \in \partial f_i(\bar{x} + d_i^*)$ (cf. (2.3)), using (1.vi), one has

$$\Delta f_i = f(\bar{x} + d_i^*) - f_i(\bar{x} + d_i^*) = f_i^*(z_i^*) - f^*(z_i) + (\bar{x} + d_i^*)(z_i - z_i^*). \tag{5.9}$$

Using (5.8) in (5.9) to eliminate z_i^*, one obtains

$$\Delta f_i \leq \min_{i>j}\{f^*(z_j) - f^*(z_i) + (\bar{x} + d_i^*)(z_i - z_j)\}. \tag{5.10}$$

Sending $j \to \infty$, the min in (5.10) goes to 0 since $\{d_i^*\}$ is bounded, and hence, by Lemma 5.4, $\{z_i\}$ and $\{f^*(z_i)\}$ are also bounded. (Extract a subsubsequence such that both the z-values and the f^*-values converge to a cluster point.) □

The proof of Theorem 5.9 is essentially that of [HL93b, Theorem XII.4.2.3] for the cutting plane method, working *in the dual space* rather than in the primal space.

This shows the usefulness of our dual treatment, as the primal proofs of convergence of proximal and trust-region bundle methods are not easy to unify. Also, note that t is not even mentioned in the proof, and hence nothing prevents $t_i \to 0$.

It is easy to verify that the cutting plane model $\hat{f}_\beta$ with a strictly monotone β-strategy guarantees (5.7). A strictly monotone β-strategy is weakly monotone, i.e., $v(\Pi_i)$ is nondecreasing; since it is also upper bounded by $f(\bar{x})$, it is clear that it can increase by any fixed quantity $\mu > 0$ only finitely many times. Hence, after some iteration h, no information is removed from β. Now, from (1.7) one has that $\hat{f}_i^*(z_j) \leq f^*(z_j)\ \forall i > j (z_j \in \beta_i)$, but $\hat{f}_i^* \geq f^*$.

Finally, note that this theorem requires the boundedness of $\{d_i^*\}$ (which implies that of $\{z_i\}$) but *not* of $\{z_i^*\}$. With $f_\beta = \hat{f}_\beta$, however, this is actually not an advantage, since, as we noted previously, in this case the boundedness of $\{d_i^*\}$ implies that of $\{z_i^*\}$.

5.4. Overall NS convergence result. We have shown that, under a number of different assumptions on the function, the model, and the stabilizing term, $\{\Delta f_i\} \to 0$ during infinite sequences of NS. In view of (2.7), this means that NS can be used to approximate (1.12) and (1.11) as closely as desired. This is more easily seen if t is fixed at length (e.g., rule (4.ii′) is in effect); then, an infinite sequence of NS solves (1.12) and (1.11) for the current point $\bar{x}$ and the fixed t. Compactness of $\{d_i^*\}$ is typically required, and that of $\{z_i^*\}$ is usually available as well; hence, by the lower semicontinuity of the objective functions, subsequences of $\{d_i^*\}$ and $\{z_i^*\}$ can be extracted which converge to finite optimal solutions, respectively, for (1.12) and (1.11).

From the algorithmic viewpoint, $\{\Delta f_i\} \to 0$ implies the finite termination of the sequences of NS for $\varepsilon > 0$; this uses the following basic relation about the predicted improvement:

$$-\delta_i \geq -\delta_i - D_{t_i}(d_i^*) = -f_i(\bar{x} + d_i^*) + f(\bar{x}) - D_{t_i}(d_i^*) = \alpha_i^* + D_{t_i}^*(-z_i^*) = 0 \tag{5.11}$$

(use $D_{t_i} \geq 0$, (2.4), (2.5), and (2.9)).

THEOREM 5.10. *Assume that $\{\Delta f_i\} \to 0$; if $\varepsilon > 0$, then after finitely many consecutive NS either the stopping condition of the algorithm in Figure* 1 *holds or an SS is done; otherwise ($\varepsilon = 0$), $\{D_{t_i}^*(-z_i^*)\} \to 0$, and $\{\alpha_i^*\} \to 0$.*

Proof. Assume that infinitely many NS are done; (4.5) can be rewritten, using (4.2) and (4.3) first and then (5.11), as

$$\Delta f_i > -(1-m)\delta_i \geq (1-m)[\alpha_i^* + D_{t_i}^*(-z_i^*)]. \tag{5.12}$$

Since $\{\Delta f_i\} \to 0$, both $\{\alpha_i^*\}$ and $\{D_{t_i}^*(-z_i^*)\}$ must go to zero; if $\varepsilon > 0$, then the stopping condition of the algorithm in Figure 1 eventually holds. ∎

Note that, in general, $\{D_{t_i}^*(-z_i^*)\} \to 0$ does not imply $\{z_i^*\} \to 0$; consider the case where D_t^* is a "trust region" (cf. Example 3.2) and/or $\{t_i\} \to 0$.

The above development can be extended to the case where δ_i in (4.4)/(4.5) is replaced with $\underline{\delta}_i = \delta_i + D_{t_i}(d_i^*)$; this corresponds to checking $f(\bar{x}+d_i^*)$ against $v(\Pi_i) = f_i(\bar{x} + d_i^*) + D_{t_i}(d_i^*)$ rather than against $f_i(\bar{x} + d_i^*)$. In fact, it is easy to check that (5.12) can be obtained as well from (5.11) and the modified from of (4.5) using $\underline{\delta}_i$. As observed in [HL93b, section XV.3], [Ki99, section 5], this descent test is weaker than (4.4) ($\underline{\delta}_i \geq \delta_i$), and therefore it may reduce the number of NS.

5.5. The polyhedral case. Finite termination of NS sequences requires $\varepsilon > 0$ and $m < 1$; in general, there is no chance of solving $(\Delta_{\bar{x},t})$ to optimality, i.e., of

obtaining $f_i(\bar{x}+d_i^*) = f(\bar{x}+d_i^*)$, unless f is polyhedral. A finite convergence theorem for NS sequences can be proved for two different sets of assumptions, basically corresponding to those of Theorem 5.7 (with a *safe* β-strategy) and those of Theorem 5.9 (with $f_\beta = \hat{f}_\beta$).

THEOREM 5.11. *Assume that* $\{\Delta f_i\} \to 0$, *f is polyhedral, and* (4.1) *is satisfied. If either* (P*3″) *holds, rule* (4.ii′) *is in effect, and a safe β-strategy is used, or $f_\beta = \hat{f}_\beta$ and a strictly monotone β-strategy is used, then after finitely many consecutive NS either the stopping condition of the algorithm in Figure* 1 *holds or an SS is done, even if $\varepsilon = 0$ and $m = 1$.*

Proof. Assume by contradiction that the stopping condition does not hold and that $\Delta f_i = \delta_{\bar{x}}(d_i^*) - \delta_i \geq \delta_{\bar{x}}(d_i^*) - m\delta_i > 0$ for infinitely many i.

If (P*3″) holds, (5.3) gives $h_i'(0) = -\Delta f_i < 0$, and therefore $v(\Delta_+) > v(\Delta_i)$; we can conclude that the set $\{v(\Delta_i)\}$ must be infinite. But the assumptions on f and the safe β-strategy ensure that there are only finitely many different possible sets β; after the iteration h implied by rule (4.ii′), $v(\Delta_i)$ can have only finitely many different values ($\bar{x}$ and t are fixed).

For the other case ($f_\beta = \hat{f}_\beta$ and a strictly monotone β-strategy), note that if the pair $(f^*(z_i), z_i)$ already belongs to β_i, then $\Delta f_i = 0$. (Use (1.6) and $\hat{f}_i \leq f$.) Now, Definition 4.8 and $\{\Delta f_i\} \to 0$ ensure that, at length, removals are inhibited; by (4.1), only finitely many "new" pairs $(f^*(z_i), z_i)$ can ever be generated, which yields the contradiction. □

6. Convergence of SS sequences (2nd level). Having proved convergence of the NS sequences, in the following we disregard what happens between two consecutive SS. However, we are not allowed to entirely disregard NS; in fact, it may happen that only finitely many SS are done, so that a "tail" of (possibly infinitely many) consecutive NS is done after the last SS. In order to deal with the two different cases—finitely many and infinitely many serious steps—in a unified way, in this section we will use the following notation: the index i denotes the ith serious step if at least i SS are performed; otherwise it denotes the $(i-k)$th NS of the only infinite sequence of NS that starts right after that the last SS (the kth) is performed. With this notation, $\bar{x}_i, d_i^*, z_i^*, \delta_i, \ldots$ refer to the status of the algorithm just *before* the change of the current point occurring at step i, if any.

The standing assumption for all the results in this section is

conditions sufficient to guarantee $\{\Delta f_i\} \to 0$ during an infinite sequence of NS hold.

Several different such conditions exist, as we have shown in the previous sections. About the model, without further notice, we will require only $f_i \leq f$.

The first step for proving the convergence of the SS sequences consists in bounding the decrease that each step obtains. From (4.4), $f(\bar{x}_+) - f(\bar{x}_i) \leq m\delta_i$; hence this boils down to bounding the predicted improvement δ_i, for which one can use (5.11) and the stopping condition:

$$-\delta_i \geq \frac{1}{\mu}\alpha_i^* + D_{t_i}^*(-z_i^*) = \frac{1}{\mu}[\alpha_i^* + \mu D_{t_i}^*(-z_i^*)] > \frac{\varepsilon}{\mu}$$

(use $\mu \geq 1$ and $\alpha_i^* \geq 0$). Hence

$$f(\bar{x}_i) \leq f(\bar{x}_0) + m\left(\sum_{j<i}\delta_j\right) \leq f(\bar{x}_0) - \frac{m}{\mu}\left(\sum_{j<i}\alpha_j^* + \mu D_{t_j}^*(-z_j^*)\right). \tag{6.1}$$

Note that (6.1) holds even if δ_i in (4.4) is replaced by $\underline{\delta}_i = \delta_i + D_{t_i}(d_i^*)$, as discussed in section 5.4. Finite termination, at least, is at hand whenever $\underline{f} := \lim_{i\to\infty} f(\bar{x}_i) > -\infty$.

LEMMA 6.1. *If $\underline{f} > -\infty$ and $\varepsilon > 0$, then only finitely many iterations can be done.*

Proof. By Theorem 5.10, only finitely many NS can be done between two consecutive SS; from (6.1), $-\infty < \underline{f} \le f(\bar{x}_0) - m\varepsilon i/\mu$, and therefore only finitely many SS can be done. ∎

Lemma 6.1 gives no information about how "good" the obtained solution is when the algorithm stops. Without qualification, nothing can be said; if D_t is nonsmooth in 0—(P5′) does not hold—the fact that 0 is optimal for (1.12) does not imply that $\bar{x}$ is optimal for (0.1); i.e., a minimum of the generalized Moreau–Yosida regularization ϕ_t may not be a minimum of f.

6.1. Convergence under (P5′)/(P*5′). The immediate effect of assumption (P5′)/(P*5′) is to guarantee convergence of the dual iterates to 0, provided that $\underline{f} > -\infty$ and t remains bounded away from zero.

THEOREM 6.2. *If $\underline{f} > -\infty$, (P*5′) holds, $t_i \ge \underline{t} > 0$, and $\varepsilon = 0$, then $\{\alpha_i^*\} \to 0$ and $\{z_i^*\} \to 0$.*

Proof. Since $\varepsilon = 0$, $\{D_{t_i}^*(-z_i^*)\} \to 0$ and $\{\alpha_i^*\} \to 0$. This is guaranteed by the stopping condition if the algorithm terminates finitely, by Theorem 5.10 if finitely many SS are done, and by (6.1) and $\underline{f} > -\infty$ if infinitely many SS are done.

Under (P5′), $\{D_{t_i}^*(-z_i^*)\} \to 0$ and $t_i \ge \underline{t}$ imply that $\{z_i^*\} \to 0$; in fact, from (P*4) $\{D_{\underline{t}}^*(-z_i^*)\} \to 0$, so that from (P*2) all the z_i^* belong to a compact set (a proper level set of $D_{\underline{t}}^*$) and, extracting a subsequence if necessary, $\{z_i^*\} \to z^*$. $D_{\underline{t}}^*$ is lower semicontinuous; hence

$$0 = \liminf_{i\to\infty} D_{\underline{t}}^*(-z_i^*) \ge D_{\underline{t}}^*(z^*) \ge 0.$$

Due to (P*5′), $D_{\underline{t}}^*(z^*) = 0 \Leftrightarrow z^* = 0$. ∎

The requirement on t can be weakened if D_t^* has the special form (3.4) (which implies (P*5′)). In fact, by (6.1) and $\underline{f} > -\infty$,

$$\infty > \frac{m}{\mu}\left(\sum_{i\to\infty} \alpha_i^* + \mu D_{t_i}^*(-z_i^*)\right) \ge m\left(\sum_{i\to\infty} t_i D^*(-z_i^*)\right).$$

Hence, we can replace $t_i \ge \underline{t}$ with the milder condition

$$\text{if } \textit{infinitely many} \text{ SS are done, then } \sum_{i\to\infty} t_i = \infty$$

and still be guaranteed that $\{D^*(-z_i^*)\} \to 0$. In turn, this implies $\{z_i^*\} \to 0$, since all the z_i^* belong to a proper level set of D^*, which is compact by (P*2), and D^* is strictly convex in 0. Note that, by Theorem 5.8, under proper conditions (3.4) allows us to drop $t_i \ge \underline{t} > 0$ for sequences of NS also.

Therefore, in the following we will assume that

$$\begin{aligned} &\textit{either } t \text{ is bounded away from zero} \\ &\textit{or } D_t^* \text{ has the form (3.4) and } \sum_{i\to\infty} t_i = \infty. \end{aligned} \tag{6.2}$$

Yet, without qualification, convergence of the dual iterates does not imply convergence of the function values; a possibility is the usual "asymptotic complementary slackness" condition.

THEOREM 6.3. *If* (P*5′) *and* (6.2) *hold,* $\varepsilon = 0$, *and* $\liminf_{i\to\infty} z_i^* \bar{x}_i = 0$, *then* $\{f(\bar{x}_i)\} \to v(\Pi)$.

Proof. If $\underline{f} = -\infty$, then $\{\bar{x}_i\}$ is a minimizing sequence, so assume $\underline{f} > -\infty$; the hypotheses of Theorem 6.2 are satisfied. From (2.9) and $f_\beta^* \geq f^*$,

$$z_i^* \bar{x}_i = f_i^*(z_i^*) + f(\bar{x}_i) - \alpha_i^* \geq f^*(z_i^*) + f(\bar{x}_i) - \alpha_i^*.$$

Taking the lim inf on both sides and using the hypothesis, we obtain

$$\begin{aligned} 0 = \liminf_{i\to\infty} z_i^* \bar{x}_i &\geq \liminf_{i\to\infty} [f^*(z_i^*) + f(\bar{x}_i) - \alpha_i^*] \\ &\geq \liminf_{i\to\infty} f^*(z_i^*) + \liminf_{i\to\infty} f(\bar{x}_i) + \liminf_{i\to\infty} -\alpha_i^*. \end{aligned}$$

Now, use $\{z_i^*\} \to 0$ and $\{\alpha_i^*\} \to 0$ (by Theorem 6.2), $\{f(\bar{x}_i)\} \to \underline{f}$, and the lower semicontinuity of f^* to obtain $0 \geq f^*(0) + \underline{f}$, i.e., $v(\Pi) = -f^*(0) \geq \underline{f}$; since $\underline{f} \geq v(\Pi)$, the thesis is proved. □

Under (P*5′), if $\{\bar{x}_i\}$ has a cluster point x^*—which happens, for instance, if only finitely many serious steps are done—then x^* is optimal for (0.1); in fact, Theorem 6.2 applies, and therefore $\{z_i^* \bar{x}_i\} \to 0$ as $\{z_i^*\} \to 0$. This could have been directly proved in primal notation using (2.10), the fact that $\{\alpha_i^*\} \to 0$, and [HL93b, Proposition XI.4.1.1]. Hence, the bundle algorithm converges at least if f is inf-compact; however, something better can be done.

THEOREM 6.4. *If* (P*5′) *and* (6.2) *hold,* $\varepsilon = 0$, *and* f *is* *-*compact, then* $\{f(\bar{x}_i)\} \to v(\Pi)$.

Proof. Assume by contradiction that $v(\Pi) < l = \underline{f} - \lambda$ for $\lambda > 0$, and let $\hat{x}_i$ be the projection of $\bar{x}_i$ over $S_l(f)$. Since $f(\bar{x}_i)$ is nonincreasing, $f(\bar{x}_i) \leq f(\bar{x}_0) = L \ \forall i$; therefore, by *-compactness, $\|\hat{x}_i - \bar{x}_i\| \leq e(l, L) < \infty \ \forall i$. From (2.10), $f(\bar{x}_i) \geq \underline{f}$ and the ε-subgradient inequality

$$\underline{f} - \lambda = f(\hat{x}_i) \geq f(\bar{x}_i) + z_i^*(\hat{x}_i - \bar{x}_i) - \alpha_i^* \geq \underline{f} - \|z_i^*\| \cdot \|\hat{x}_i - \bar{x}_i\| - \alpha_i^*$$

that yield the desired contradiction since, from Theorem 6.2, $\{z_i^*\} \to 0$ and $\{\alpha_i^*\} \to 0$. □

Theorem 6.4 in fact proves that a *-compact f is *asymptotically well-behaved* (a.w.b.) [Au97]. A sequence $\{x_i\}$ is a *stationary sequence* for the function f if two sequences $\{z_i\} \to 0$ and $\{\varepsilon_i\} \to 0$ exist such that z_i is an ε_i-subgradient of f at x_i; f is a.w.b. if every stationary sequence is a minimizing sequence. In [Au97] it is proved that f is a.w.b. if and only if all the following three functions

$$\begin{aligned} r(l) &= \inf_x \left\{ \inf_z \{\|z\| : z \in \partial f(x)\} : f(x) = l \right\}, \\ k(l) &= \inf_x \left\{ \inf_z \left\{ f'\left(x, \frac{z}{\|z\|}\right) : z \in \partial f(x) \right\} : f(x) = l \right\}, \\ l(l) &= \inf_x \left\{ \frac{(f(x) - l)}{d_{S_l(f)}(x)} : f(x) > l \right\} \end{aligned}$$

are strictly positive for each $l > v(\Pi)$; by Theorem 6.4, *-compactness is another sufficient condition for "well-behavedness." A result quite similar to Lemma 4.4, in

a more general setting, can be found in [Au97], but it requires *weak coercivity* of $f(0 \in ri\ dom\ f^*)$. Clearly, *-compact functions need not be weakly coercive (take an affine function). On the other hand, weak coercivity ensures convergence of the primal iterates as well as of the function values [Au97, Theorem 6], and therefore it can be a convenient alternative to *-compactness when stronger convergence properties are required.

Note that, even when $\{\bar{x}_i\}$ is guaranteed to be a minimizing sequence, stopping as soon as $\bar{x}_i$ is ε-optimal for some fixed $\varepsilon > 0$ is not straightforward. Indeed, an estimate of the quality of $\bar{x}_i$ is available only if $z_i^* = 0$, since then $\bar{x}_i$ is α_i^*-optimal (use (2.10)). In practice, the stopping condition has to require that z_i^* is "small enough"; this is the meaning of the extra stopping parameter $\mu \geq 1$. For D_t^* in the special form (3.4), for instance, μ makes the stopping condition be that of $t^* = t\mu \geq t$; in our experience, guessing a value of μ that produces a true ε-optimal solution is usually fairly easy.

6.2. The polyhedral case. If f is polyhedral ($\Rightarrow^*$-compact), one can prove *finite* convergence for $\varepsilon = 0$; of course, this first requires finite convergence of the 1st level. The basic result is that, at length, the primal stabilized master problem (1.8) is equivalent to its nonstabilized version; this follows from the next technical lemma.

LEMMA 6.5. *Assume that f is polyhedral,* (4.1) *is satisfied, $f_\beta = \hat{f}_\beta$, and a safe β-strategy is used; for any function h^* satisfying* (P*1) *and* (P*5$'$) *there exists a constant $\Psi_f > 0$ such that, however fixed β, if a $z \in \partial \hat{f}_\beta(x)$ exists such that $h^*(z) < \psi_f$, then $0 \in \partial \hat{f}_\beta(x)$.*

Proof. From (4.1) and the safe β-strategy, there is only a *finite* number of different possible β. Since each $\hat{f}_\beta$ has only a *finite* set of possible different subdifferentials [HL93a, Corollary VI.4.3.2], there is a *finite set* Γ_f containing all possible subdifferentials of some $\hat{f}_\beta$ at some point x. Let $\psi(Z) = \inf_{z \in Z}\{h^*(z)\}$ (≥ 0 due to (P*1)) and $\psi_f = \min\{\psi(Z) : Z \in \Gamma_f, \psi(Z) > 0\} > 0$; $\psi(Z) = 0$ for any Z such that $h^*(z) < \psi_f$ for some $z \in Z$. Closedness of the subdifferentials and $h^*(z) = 0 \Leftrightarrow z = 0$ (via (P*5$'$)) do the rest. ∎

Note that, when f itself is polyhedral, there exists one finite β such that $f = \hat{f}_\beta$; hence, a fortiori for each h^* there exists a $\psi_f > 0$ such that $z \in \partial f(x)$ and $h^*(z) < \psi_f \Rightarrow x$ is optimal for (Π).

THEOREM 6.6. *Under the hypotheses of Theorem* 5.11 *and Lemma* 6.5*, if f is bounded below, $t_i = \underline{t} > 0$,* (P*5$'$) *holds, $\varepsilon = 0$, and $m = 1$, then the two-level bundle algorithm finitely solves* (Π).

Proof. Setting $m = 1$ and $\varepsilon = 0$ is allowed by Theorem 5.11; after finitely many consecutive NS, either the algorithm stops or $\hat{f}_i(\bar{x}_i + d_i^*) = f(\bar{x}_i + d_i^*)$ and an SS is done. If the algorithm stops, by $\varepsilon = 0$ one has $\alpha_i^* = 0$ and, from (P5$'$), $z_i^* = 0$; therefore, $\bar{x}_i$ is optimal (cf. (2.10)). Hence, assume by contradiction that infinitely many SS are done; by (6.1) and the boundedness of f, as in the proof of Theorem 6.2, we get $\{D_{\underline{t}}^*(-z_i^*)\} \to 0$. Since $z_i^* \in \partial \hat{f}_i(\bar{x}_i + d_i^*)$, applying Lemma 6.5 with $h^* = D_{\underline{t}}^*$ shows that, for large enough i, $0 \in \partial \hat{f}_i(\bar{x}_i + d_i^*)$; i.e., $\bar{x}_i + d_i^*$ is a minimum of $\hat{f}_i$. Hence, at length every $f(\bar{x}_i)$ is a minimum of some $\hat{f}_\beta$; but from the hypotheses there are only finitely many different sets β, which contradicts $f(\bar{x}_+) > f(\bar{x}_i)$. ∎

Note that, as for Theorem 6.2, the requirement over t can be weakened if $D_{\underline{t}}^*$ has the form (3.4).

Let us mention that setting $m = 1$ all along is only the simplest possibility; what is really required is that only finitely many "inexact" SS (with $\Delta f > 0$) be performed between two "exact" SS (with $\Delta f = 0$). Hence, m can be reset to any value < 1 after

each exact SS, provided that it is set to 1 after finitely many consecutive (inexact) SS.

7. Convergence of the 3rd level. If (P*5′) does not hold, convergence requires $t \to \infty$ and therefore the "three-level" bundle algorithm of Figure 2. Hence, let us once again change our notation: from now on, the index i refers to the end of the ith call to the algorithm of Figure 1, with $\underline{t} = \underline{t}_i$ and $\varepsilon = \varepsilon_i > 0$, from within the cycle of the "three-level" bundle algorithm. Therefore, the standing assumption is now

> conditions sufficient to guarantee finite termination
>
> of the two-level bundle algorithm hold.

We also assume that $\{\underline{t}_i\} \to \infty$ and $\{\varepsilon_i\} \to 0$.

7.1. Primal convergence. It is instructive to compare Lemma 5.4 with Theorem 6.4. In the former—where $\bar{x}$ is fixed—the optimal values of (1.12) converge to that of (0.1) without the *-compactness assumption, while in the latter—where SS are allowed—it is required. The same happens with the bundle algorithm.

THEOREM 7.1. *If f is *-compact, then* $\lim_{i\to\infty} f(\bar{x}_i) = v(\Pi)$.

Proof. Since $t_i \geq \underline{t}_i$, $\{t_i\} \to \infty$. The stopping condition implies $v(\Delta_i) \leq \varepsilon_i - f(\bar{x}_i)$, i.e., $v(\Pi_i) + \varepsilon_i \geq f(\bar{x}_i)$, and since $v(\Pi_{\bar{x}_i,t_i}) \geq v(\Pi_i)$, we obtain $v(\Pi) \leq f(\bar{x}_i) \leq v(\Pi_{\bar{x}_i,t_i}) + \varepsilon_i$; now apply Lemma 4.4. □

Note that *-compactness is used in Lemma 4.4 ⇒ Theorem 7.1 without any reference to a stationary sequence; hence, unlike Theorem 6.4, a.w.b.-ness could not be used here. Furthermore, $\{t_i\} \to \infty$ is required in order to solve (Π) with "infinite accuracy"; a suitably large t suffices for obtaining any finite accuracy (of course, f must be bounded below). In fact, using Lemma 4.3, it is easy to show that, if f is bounded below, then for any starting point $\bar{x}_0$ and any fixed $\varepsilon > 0$ there exists a $\bar{t}$ such that $v(\Pi_{\bar{x}_i,\bar{t}}) \leq v(\Pi) + \varepsilon$ (use $f(\bar{x}_i) \leq f(\bar{x}_0)$). Given a suitable estimate of $\bar{t}$, the two-level bundle algorithm can directly solve (Π) with any finite accuracy.

Eliminating the *-compactness assumption is possible, at the cost of inhibiting—at length—the serious steps, i.e., using rule (4.iii′). In this case, t needs to go all the way up to ∞.

THEOREM 7.2. *With rule* (4.iii′) *in force,* $\liminf_{i\to\infty} f(\bar{x}_i + d_i^*) = v(\Pi)$.

Proof. Wait for the last SS to be performed, and call $\bar{x}(= \bar{x}_i)$ the fixed current point. Assume by contradiction that $\liminf_{i\to\infty} f(\bar{x} + d_i^*) - 2\delta > v(\Pi)$ for some $\delta > 0$; hence, there exists a $\bar{d}$ such that $f(\bar{x} + \bar{d}) \leq f(\bar{x} + d_i^*) - 2\delta \ \forall i$. Due to (P5) and $\{t_i\} \to \infty$, $D_{t_i}(\bar{d}) \leq \delta$ for a large enough i; therefore

$$v(\Pi_{\bar{x},t_i}) \leq f(\bar{x} + \bar{d}) + D_{t_i}(\bar{d}) \leq f(\bar{x} + d_i^*) - \delta \leq f(\bar{x} + d_i^*) + D_{t_i}(d_i^*) - \delta.$$

When the inner loop terminates, $\Delta f_i \leq \varepsilon_i$; hence, using (2.6) and $v(\Pi_{\bar{x},t_i}) \geq v(\Pi_i)$,

$$\varepsilon_i + v(\Pi_{\bar{x},t_i}) \geq \varepsilon_i + v(\Pi_i) \geq \Delta f_i + v(\Pi_i) = f(\bar{x} + d_i^*) + D_{t_i}(d_i^*),$$

which leads to $\varepsilon_i \geq f(\bar{x} + d_i^*) + D_{t_i}(d_i^*) - v(\Pi_{\bar{x},t_i}) \geq \delta$, contradicting $\{\varepsilon_i\} \to 0$. □

7.2. Dual convergence. From the dual viewpoint, $\{\bar{x}_i + d_i^*\}$ is a maximizing sequence for the Lagrangian dual of (1.2) w.r.t. the constraints $z = 0$ (cf. section 1), and $\{z_i\}$ are the optimal solutions of the corresponding dual pricing problems (1.3), with $\bar{x} = \bar{x}_i + d_i^*$. Further, from $f^* \leq f_i^*$ and (2.8), the alternative stopping condition of (4.iii′) ($\Delta f_i \leq \varepsilon_i$) gives

$$f^*(z_i^*) - z_i^*(\bar{x}_i + d_i^*) \leq f_i^*(z_i^*) - z_i^*(\bar{x}_i + d_i^*) \leq f^*(z_i) - z_i(\bar{x}_i + d_i^*) + \varepsilon_i,$$

i.e., z_i^* is an ε_i-optimal solution for (1.3) with $\bar{x} = \bar{x}_i + d_i^*$. Using (1.v) in the above relation gives

$$z_i^* \in \partial_{\varepsilon_i} f(\bar{x} + d_i^*),$$

i.e., ε_i-optimal solutions for (1.3) are ε_i-*subgradients* of f; this is of particular interest when f itself is a dual function (cf. section 9). Thus, if $\{\bar{x}_i + d_i^*\} \to x^*$ and $\{z_i^*\} \to z^*$, then $z^* \in \partial f(x^*)$ [HL93b, Proposition XI.4.1.1]; one would like to show that $\{z_i^*\} \to 0$ whenever f is bounded. This is possible, and it does not require *-compactness.

THEOREM 7.3. *If f is bounded below and rule* (4.iii′) *is in force, then* $\{z_i^*\} \to 0$.

Proof. Using (2.1) and $D_{t_i} \geq 0$, we obtain

$$-v(\Delta_i) = v(\Pi_i) = f_i(\bar{x}_i + d_i^*) + D_{t_i}(-z_i^*) \geq f_i(\bar{x}_i + d_i^*).$$

Using the previous relation with the stopping condition of (4.iii′), $\Delta f_i = f(\bar{x}_i + d_i^*) - f_i(\bar{x}_i + d_i^*) \leq \varepsilon_i$, gives, together with boundedness of f and monotonicity of $\{\varepsilon_i\}$,

$$v(\Delta_i) \leq -f_i(\bar{x}_i + d_i^*) \leq \varepsilon_i - f(\bar{x}_i + d_i^*) \leq \varepsilon_0 - v(\Pi) < \infty.$$

Now, using (2.9) and $f_i^* \geq f^*$, one obtains

$$\infty > v(\Delta_i) = f_i^*(z_i^*) - z_i^* \bar{x}_i + D_{t_i}^*(-z_i^*) \geq D_{t_i}^*(-z_i^*) - f(\bar{x}_i).$$

By rule (4.iii′), only finitely many serious steps are done, hence at length, $f(\bar{x}_i) = f(\bar{x})$ for a fixed $\bar{x}$; by (P*5), $\|z_i^*\|_2 \geq \varepsilon > 0$ for infinitely many i and $\{t_i\} \to \infty$ imply $\{D_{t_i}^*(-z_i^*)\} \to \infty$. □

Note that, if f is bounded below, a dual proof of Theorem 7.1 exists, using Theorem 7.3 ($\{z_i^*\} \to 0$) and the fact that $\{\bar{x}_i\}$ is a stationary sequence; however, the case of f unbounded below would need a separate treatment (a.w.b.-ness is tailored over bounded functions with unbounded level sets).

7.3. The polyhedral case. The three-level bundle method allows us to drop assumption (P5′) from the finite termination proofs in the polyhedral ($\Rightarrow^*$-compact) case. Indeed, for bounded polyhedral functions one can prove the following strengthened form of Lemma 4.4, where z_t and d_t denote the optimal solutions of $(\Delta_{\bar{x},t})$ and $(\Pi_{\bar{x},t})$, respectively.

LEMMA 7.4. *If f is polyhedral and bounded below, then for each $L < \infty$ there exists a $\underline{t} > 0$ such that $\bar{x} + d_t$ is an optimal solution of* (Π) $\forall t > \underline{t}$ *and $\bar{x}$ such that* $f(\bar{x}) \leq L$.

Proof. Fix any $\bar{x}$ such that $f(\bar{x}) \leq L$; it is easy to show, mirroring Theorem 7.3, that $\{z_t\} \to 0$ for $t \to \infty$ (use $D_t^*(-z_t) - L \leq D_t^*(-z_t) - f(\bar{x}) \leq f^*(z_t) - z_t\bar{x} + D_t^*(-z_t) = v(\Delta_{\bar{x},t}) \leq -v(\Pi) < \infty$ and (P*5)). Then, using $z_t \in \partial f(\bar{x} + d_t)$ (cf. (2.3)) and Lemma 6.5 with $h^* = \|\cdot\|$, we obtain that, for large enough t, $0 \in \partial f(\bar{x} + d_t)$; i.e., $\bar{x} + d_t$ is a minimum of f. □

This result allows us to derive a finite convergence proof; since f is polyhedral, we can directly fix $\varepsilon_i = 0$ and use rule (4.iii″).

THEOREM 7.5. *Under the hypotheses of Theorem* 5.11 *and Lemma* 6.5, *if $\varepsilon_i = 0$ $\forall i$ and rule* (4.iii″) *is in force, then the three-level bundle algorithm finitely solves* (Π).

Proof. From Theorem 5.11, we know that only finitely many consecutive NS can be done: either the normal stopping rule fires or an SS is performed. However, from rule (4.iii″), only finitely many SS can be done; hence, either the stopping rule fires, or a sequence of consecutive NS is started. Theorem 5.11 tells us that such a sequence

finitely produces $\Delta f = 0$; hence the two-level bundle algorithm finitely terminates with either $\alpha^* + D_t^*(-z^*) = 0$ or $\Delta f = 0$.

If $\alpha_i^* + D_{t_i}^*(-z_i^*) = 0$ happens infinitely many times, $\alpha_i^* = 0$ and (2.10) tell us that $z_i^* \in \partial f(\bar{x}_i)$. Theorem 7.3 shows that $\|z_i^*\| \to 0$ as $\{t_i\} \to \infty$; hence, applying Lemma 6.5 with $h^* = \|\cdot\|$ shows that, for large enough i, $0 \in \partial f(\bar{x}_i)$, i.e., $\bar{x}_i$ is optimal for (Π). If $\Delta f_i = 0$ happens infinitely many times, recall from (2.7) that this means that d_i^* is optimal for $(\Pi_{\bar{x}_i,t_i})$ and use Lemma 7.4. □

Let us remark that the three-level bundle algorithm applied to a polyhedral f lacks a convenient stopping criterion; either $\bar{x}_i$ or $\bar{x}_i + d_i^*$ at some point becomes optimal, but there is no easy way to tell when this happens. In order to be able to stop, either the solver of $(\Delta_{\beta,\bar{x},t})$ should always return $z^* = 0$ whenever it can, or an estimate of $\underline{t}$ of Lemma 7.4 should be available.

8. Extensions. The generalized bundle algorithm presented in the previous paragraphs can incorporate a number of important algorithmic variants. For instance, (4.iv) allows us to seamlessly add a line search on d^*, provided only that, at length, the unit step is always probed. (4.i)–(4.iii) allow us to adapt the curved search approach of [SZ92] to our more general setting; other t-strategies, originally devised for $D_t = \frac{1}{2t}\|\cdot\|_2^2$, can be adapted as well [Fr97, section I.5]. Multiple [ε-]subgradients can be added to β at each call of the oracle if the latter is—as happens in some applications—capable of providing them. Finally, it should not be hard to extend the proofs of convergence to the case in which f is not computed exactly, following what is done in [GV97, Ki99]. More complex extensions are discussed in the following section.

8.1. The constrained case. Generalized bundle methods can cope with constraints $x \in X$ if X is a closed convex set. Basically, all that is needed is to insert full knowledge about X into (1.8), i.e., to solve at each iteration

$$(\Pi_{\beta,\bar{x},t}) \qquad \inf_d \{f_\beta(\bar{x}+d) + D_t(d) : (\bar{x}+d) \in X\}. \tag{8.1}$$

Problem (8.1) can be viewed as (1.8) using the *restricted model* $f_{X,\beta} = f_\beta + I_X$, which is a model of the actual function to be minimized, the *restricted function* $f_X = f + I_X$. Under the natural assumption that $\bar{x} \in dom\ f_\beta \cap X$, the dual of (8.1) is just (1.9) with f_β^* replaced by

$$f_{X,\beta}^*(z) = (f_\beta + I_X)^*(z) = \inf_w \{f_\beta^*(z-w) + \sigma_X(w)\} \tag{8.2}$$

(see [HL93b, Theorem X.2.3.1]) as $(I_X)^* = \sigma_X$. The problem can be written in a "direct" form, avoiding the complicated-looking infimal convolution (8.2), by means of the simple variable change $z = \bar{z} + w$:

$$(\Delta_{\beta,\bar{x},t}) \qquad \inf_{\bar{z},w} \{f_\beta^*(\bar{z}) + \sigma_X(w) - \bar{x}(\bar{z}+w) + D_t^*(-\bar{z}-w)\}. \tag{8.3}$$

The extension of the theory is not completely straightforward: f_X is *not* finite everywhere, and $f_{X,\beta}$ is a model of f_X rather than of f. Hence, (2.3)/(2.4) are valid with $f_{X,\beta}$ replacing f_β; in particular, we have that $z^* \in \partial f_{X,\beta}(\bar{x}+d^*)$. On the other hand, the black box produces subgradients of f rather than of f_X, i.e., $z \in \partial f(\bar{x}+d^*)(\bar{x}+d^* \in X)$; there is an "asymmetry" that has to be taken into account.

The main observation is that some of the properties of z^* have now to be referred to $\bar{z}^*$. In fact, from $f^*_{X,\beta}(z^*) = f^*_\beta(\bar{z}^*) + \sigma_X(w^*)$ and $z^* \in \partial f_{X,\beta}(\bar{x}+d^*)$, using (1.vi), we obtain

$$[f^*_\beta(\bar{z}^*) - \bar{z}^*(\bar{x}+d^*) + f_\beta(\bar{x}+d^*)] + [\sigma_X(w^*) - w^*(\bar{x}+d^*) + I_X(\bar{x}+d^*)] = 0.$$

By (1.vii) both quantities in square brackets are nonnegative, and therefore both must be zero; hence, by (1.vi) we get $\bar{z}^* \in \partial f_\beta(\bar{x}+d^*)$ (and $w^* \in \partial I_X(\bar{x}+d^*)$). Thus, in the constrained case one has to carefully distinguish $\bar{z}^*$ from z^*. For instance, when aggregation is done, it is $\bar{z}^*$, together with its f^*-value $f^*_i(\bar{z}^*_i)$, that is added to β instead of z^*; the inequality in Definition 5.6 becomes

$$f^*_i(\bar{z}^*_i) \geq f^*_+(\bar{z}^*_i). \tag{8.4}$$

In this setting, Lemma 5.2 proves that $\{\bar{z}^*_i\}$, rather than $\{z^*_i\}$, is bounded; however, Lemma 5.1 and Lemmas 5.3–5.6 do not change. The boundedness of $\{\bar{z}^*_i\}$ also implies that of $\{z^*_i\}$ under certain assumptions, as the following lemma shows.

LEMMA 8.1. *If* (P*3) *holds,* D^*_t *has the form* (3.4)*, and* $\{\bar{z}^*_i\}$ *is bounded, then* $\{z^*_i\}$ *is bounded.*

Proof. Since $(\bar{z}^*_i, w^*_i)$ is the optimal solution of (8.3) and $\sigma_X(w^*_i) - \bar{x}w^*_i \geq 0$ as $\bar{x} \in X$, we have

$$\begin{aligned} &f^*_i(\bar{z}^*_i) - \bar{x}\bar{z}^*_i + D^*_i(-\bar{z}^*_i - w^*_i) \\ &\quad \leq f^*_i(\bar{z}^*_i) + \sigma_X(w^*_i) - \bar{x}(\bar{z}^*_i + w^*_i) + D^*_i(-\bar{z}^*_i - w^*_i) \leq f^*_i(\bar{z}^*_i) - \bar{x}\bar{z}^*_i + D^*_i(-\bar{z}^*_i), \end{aligned}$$

and therefore

$$D^*_i(-\bar{z}^*_i - w^*_i) \leq D^*_i(-\bar{z}^*_i). \tag{8.5}$$

Since D^*_t has the form (3.4), we can divide both sides of (8.5) by t_i to obtain

$$D^*(-\bar{z}^*_i - w^*_i) \leq D^*(-\bar{z}^*_i).$$

Since, by (P*3), D^* is finite everywhere and $\{\bar{z}^*_i\}$ is bounded, the left-hand side is finite; therefore, all $z^*_i = -\bar{z}^*_i - w^*_i$ belong to a level set of D^*, which is compact by (P*2). □

In order to extend the proof of Theorem 5.7, "asymmetric" definitions of h_i and r_i,

$$\begin{aligned} h_i(\gamma) &:= [f^*_{X,i}(z^*_i)(1-\gamma) + f^*(z_i)\gamma - z_i(\gamma)\bar{x} + f(\bar{x})] + D^*_{t_i}(-z_i(\gamma)), \\ r_i(z) &:= f^*_{X,i}(z) - z\bar{x} + f(\bar{x}) = f^*_{X,i}(z) - z\bar{x} + f_X(\bar{x}) \geq 0, \end{aligned}$$

are required. Using (8.4), one obtains

$$f^*_{X,i}(z^*_i) = f^*_i(\bar{z}^*_i) + \sigma_X(w^*_i) \geq f^*_+(\bar{z}^*_i) + \sigma_X(w^*_i) \geq \inf_w \{f^*_+(z^*_i - w) + \sigma_X(w)\} = f^*_{X,+}(z^*_i),$$

while from (4.iv) and $\sigma_X(0) = 0$,

$$f^*(z_i) \geq f^*_+(z_i) = f^*_+(z_i) + \sigma_X(0) \geq \inf_w \{f^*_+(z_i - w) + \sigma_X(w)\} = f^*_{X,+}(z_i);$$

now, proceeding as in section 5.2 $v(\vartheta_i) = v(\Delta_+) + f(\bar{x})$ is readily obtained. Furthermore, (2.6)/(2.8) can be written (in an asymmetric fashion) as

$$\Delta f = f(\bar{x}+d^*) - f_{X,\beta}(\bar{x}+d^*) = f^*_{X,\beta}(z^*) - f^*(z) + (\bar{x}+d^*)(z-z^*), \tag{8.6}$$

which easily gives the equivalent to (5.4),

$$h_i'(0) = -\Delta f_i < -(1-m)[f_{X,i}^*(z_i^*) - z_i^*\bar{x} + D_{t_i}^*(-z_i^*) + f(\bar{x})] = -(1-m)h_i(0),$$

which allows us immediately to extend the proofs of Theorems 5.7 and 5.8 to the constrained case. Note that $D_t^* = \frac{1}{2}t\|\cdot\|_2^2$ has *both* the forms (3.3) and (3.4); therefore, exploiting Lemma 8.1, our convergence results for $f_\beta = \hat{f}_\beta$ generalize the best ones known for the proximal bundle case.

The only difficulty in extending the proof of Theorem 5.9 comes from the fact that (5.7) does *not* guarantee $f_{X,i}^*(z_h) = f_X^*(z_h)\ \forall i \geq h$. However, $f_{X,i} \geq f_i$ and (5.7) give $f_{X,i}^*(z_h) \leq f_i^*(z_h) = f^*(z_h)$; thus, operating as in Theorem 5.9, one obtains the equivalent to (5.8):

$$f^*(z_h) - (\bar{x} + d_i^*)z_h \geq f_{X,i}^*(z_i^*) - (\bar{x} + d_i^*)z_i^* \qquad \forall i > h.$$

Combined with the "asymmetric" definition (8.6) of Δf_i (with $z = z_i$), this gives (5.10). All the other results in section 5 plainly extend to the constrained case.

It is then easy to check that almost all other results in section 6 and section 7 remain valid, with the only provision being that we look at f_X, rather than at f, as the actual function to be minimized. In particular, note that, by (8.1), $\bar{x} + d^*$ is always feasible and rule (4.iv) can be satisfied. The only exceptions are the results about polyhedral functions, which also require X to be a *polyhedral* set. In fact, it is easy to prove that Lemma 6.5 fails if X is not polyhedral, as f_X may have infinitely many different subdifferentials (take f affine and $X = B_2(\delta)$). However, if f satisfies condition (4.1) and X is polyhedral, then f_X has finitely many different subdifferentials; this allows us to extend Lemma 6.5 and all the subsequent results.

Finally, let us mention that, when X is a polyhedron $Hx \leq h$, (8.3) boils down to

$$(\Delta_{\beta,\bar{x},t}) \qquad \inf_{z,\omega}\{f_\beta^*(z) + \omega h - \bar{x}(z + \omega H) + D_t^*(-z - \omega H) : \omega \geq 0\}$$

(ω being the "dual" variables). In this case, it is not even required that all the defining inequalities of X be known in advance; when an unfeasible x is probed, the black box should just return $+\infty$ and some "extremal" violated inequality. (Assumption (4.1) on the black box must be satisfied.) Clearly, only finitely many steps are required to eventually acquire a complete description of X.

8.2. Decomposable functions. Another important extension is a different treatment of *decomposable* functions,

$$f(x) = \sum_{h \in K} f^h(x),$$

where $1 < |K| = k < \infty$; examples are cost-decomposition approaches to block-structured convex problems [PZ92, GK95, CFG01]. Here, the computation of each $f^h(\bar{x})$ gives a $z^h \in \partial f^h(\bar{x})$; rather than aggregating this information into the unique $z = \sum_{h \in K} z^h$, one may keep it in a disaggregated form [Ki95, GV97], where β is partitioned into k disjoint subsets β^h and there is one model f_β^h for each f^h. The

disaggregated subproblems

$$(\Pi_{\beta,\bar{x},t}) \qquad \inf_d \left\{ \sum_{h\in K} f^h_\beta(\bar{x}+d) + D_t(d) \right\},$$

$$(\Delta_{\beta,\bar{x},t}) \qquad \inf_z \left\{ \sum_{h\in K} (f^h_\beta)^*(z^h) - \left(\sum_{h\in K} z^h \right) \bar{x} + D^*_t \left(- \sum_{h\in K} z^h \right) \right\}$$

are then solved instead of the aggregated versions. Using the disaggregated model $f^K_\beta = \sum_{h\in K} f^h_\beta$ is well known to be potentially beneficial: in the polyhedral case, for instance, $\hat{f}^K_\beta$ is a (much) better description of f than the ordinary aggregate model $\hat{f}_\beta$.

It is easy to show that the "critical" properties are inherited by the disaggregated model f^K_β if they hold for all the f^h_β individually. For (4.iv), for instance, one has that $(f^h_+)^*(z^h) \leq (f^h)^*(z^h)\ \forall h$ implies

$$f^*(z) = \sum_{h\in K} (f^h)^*(z^h) \geq \sum_{h\in K} (f^h_+)^*(z^h) \geq \inf_{\bar{z}} \left\{ \sum_{h\in K} (f^h_+)^*(\bar{z}^h) : \sum_{h\in K} \bar{z}^h = z \right\} = (f^K_+)^*(z).$$

Analogously, it is possible to show that if (4.7)/(5.7) hold for all the f^h_β, then they hold for f^K_β. Thus, the analysis of the previous paragraphs immediately extends to the "disaggregated" variant of generalized bundle methods, independently on the stabilizing term D_t. Of course, these results can be used together with those of section 8.1 to construct a disaggregated constrained generalized bundle method.

9. Comparisons. A number of algorithms that have been proposed in the literature can be shown to be special cases of, or closely related to, the generalized bundle algorithm.

9.1. Other bundle approaches. The algorithm in Figure 1 covers the proximal bundle method [HL93b, Algorithm XV.3.3.4], where $D_t = \frac{1}{2t}\|\cdot\|^2_2$ and $D^*_t = \frac{1}{2}t\|\cdot\|^2_2$. A dual interpretation of this method is well known [HL93b, section XV.2.4]: (1.9) is a Lagrangian relaxation of the problem of finding the *steepest ε-descent direction* for $\hat{f}_\beta$ in $\bar{x}$. Historically, this dual interpretation motivated the development of the first bundle methods; however, it has drawbacks in that (1.9) (resp., (1.11)) is described in terms of a "local" object, the ε-subdifferential of f_β (resp., f) in $\bar{x}$, so that it is difficult to relate two problems corresponding to different current points. Conceptual descent methods have been proposed, based on this dual interpretation, where the L_2-norm in the dual is replaced with any norm $|||\cdot|||$ (see [HL93a, Algorithm VIII.2.1.5]); however, this does not readily extend to other forms of bundle methods, where D_t is

- $\frac{1}{t}\|\cdot\|_p$ for $p \geq 1$ (in practice, the L_1- and L_∞-norms) [KCL95];
- $\frac{1}{t}h(|||d|||)$, where $|||\cdot|||$ is any norm and h is a convex continuous and differentiable function with invertible derivative such that $h(0) = h'(0) = 0$ [Be96];
- $\frac{1}{t}D(d)$ for D strictly convex, strongly coercive, differentiable, and finite everywhere [Au87];
- the *indicator function* of the ball of radius t under some norm $|||\cdot|||$; this amounts to restricting the next trial point inside a *trust region* [HL93b, Algorithm XV.2.1.1].

It is easy to see that conditions (P1)–(P5) are less restrictive than all those above. Remarkably, the convergence proofs for the first three cases, where $D_t(0) = 0 \Leftrightarrow d = 0$, are quite different from those used in the fourth case, where $D_t(d) = 0$ in some ball around the origin. Our analysis is the first that covers both situations in a uniform way. Furthermore, our analysis is the first that fully exploits duality. In [Be96] it was noted that using a norm $|||\cdot|||$ in the primal leads to some dual problem involving the *conjugate norm* $|||\cdot|||^*$, much in the spirit of [HL93a, Algorithm VIII.2.1.5], but this was not extended to a dual interpretation of the algorithm. In [Au87], (1.9) is only used to prove [Au87, Theorem 2.3]. In other cases duality was completely overlooked, even when linear duality could have been used [KCL95]. A first step towards this development was done in [Fr97], where $D_t^* = \frac{1}{t}\|\cdot\|_p$ with $p \in \{1, 2, \infty\}$ was studied; due to the interpretation of (1.9) in terms of ε-subgradients, those bundle variants had an interest on their own, as a bundle algorithm with a dual trust region was one of the open questions in [HL93b, Remark XV.2.5.1].

Other approaches directly related to generalized bundle methods are proximal-type algorithms; there, the stabilized problem (1.12) is solved with a "nonuniform" stabilizing term, which depends on $\bar{x}$ as well as on t. This is used to incorporate constraints in the stabilizing term, which also serves as a barrier function to keep the iterates feasible. Stabilizing terms studied in the literature are either *D-functions* [Ec93, CT93],

$$D_{\bar{x},t}(d) = \frac{1}{t}(\psi(\bar{x}+d) - \psi(\bar{x}) - \nabla\psi(\bar{x})d),$$

where ψ is a fixed strictly convex and differentiable function such that the level sets of $D_{\bar{x},t}$ are compact, or *φ-divergences* [IST94, IT95, Te97],

$$D_{\bar{x},t}(d) = \frac{1}{t}\sum_{i=1,\dots,n} \bar{x}_i\varphi\left(\frac{\bar{x}_i + d_i}{\bar{x}_i}\right),$$

where φ is a fixed univariate function that is (among other things) continuously differentiable, strictly convex, and such that $\varphi(1) = \varphi'(1) = 0$. These stabilizing terms satisfy (P1), (P4), and (P5), and they have bounded level sets [IST94] which contain 0 in the interior if $\bar{x}$ lies in the zone of $D_{\bar{x},t}$ (int *dom* ψ in the first case and $\Re^n_{++}$ in the second), where proximal-type algorithms work. Conditions parallel to (P3) and (P3′) are also required: boundedness of f, that corresponds to (P3′), is widely used, but in [CT93] the requirement is rather *im* $\nabla\psi = \Re^n$, i.e., *dom* $\psi^* = \Re^n$, i.e., (P*3) as

$$D^*_{\bar{x},t}(z) = \frac{1}{t}\psi^*(tz + \nabla\psi(\bar{x})) - \bar{x}(tz + \nabla\psi(\bar{x})) + \psi(\bar{x}).$$

In both cases, $D_{\bar{x},t}$ is differentiable and $D_{\bar{x},t}(d) = 0 \Leftrightarrow d = 0$; this is not required in our approach, even though both differentiability (in 0) and strict convexity help to enhance (different parts of) the convergence proofs. Also, all of the above methods require the exact solution of (1.12), which is a rather strong condition. Finally, our dual viewpoint extends the one that has been developed for proximal-type algorithms, which is limited to the case in which (0.1) is itself a Lagrangian dual (cf. section 9.2).

The differentiability of $D_{\bar{x},t}$, but not strict convexity, is dropped in [Ki98], where *B-functions* are introduced; there, the compactness requirement is also different. (There is no need for "local" compactness, as the solution of (1.12) is assumed to be given.) An implementable version of the proximal method using B-functions, the

bundle Bregman proximal method, is then proposed in [Ki99]. The analysis provides strong convergence results, for instance allowing inexact solution of the stabilized master problem and avoiding the *-compactness assumption. However, it does not subsume the results of the present article, which do not require the stabilizing term to be a B-function. Furthermore, our "more technical" (cf. [CL93, Remark 4.6]) dual proof of Theorem 5.7 provides estimates on the rate of convergence during NS sequences, and we don't require f_β to be the cutting plane model, thereby allowing easy extensions, e.g., to the disaggregated case (cf. section 8.2).

Finally, a related but different approach can be found in [Nu97]. There, the dual object is the graph of the $\varepsilon \to \partial_\varepsilon f(0)$ mapping, which is equivalent (modulo a rotation) to *epi* f^*. The approach in [Nu97] can be summarized, in our notation, as follows: at each step i, find a separating hyperplane between *epi* $\hat{f}_i^*$ and the point $(-\underline{f}_i, 0)$, where $\underline{f}_i$ is the best f-value found so far. The hyperplane must be nonvertical, i.e., in the form $(1, -x_i)$; it is easy to check that $(1, -x_i)$ is a separating hyperplane if and only if $\hat{f}_i(x_i) \leq \underline{f}_i$. Condition (P*3$'$) is required in order to ensure that $\hat{f}_i^*(0) < \infty$. Not all choices of separating hyperplanes give a convergent algorithm; in [Nu97], an abstract rule is given, and an implementation is proposed under the form of the min-problem

$$\inf_{\sigma,z}\{|||(-\underline{f}_i, 0) - (\sigma, z)||| : (\sigma, z) \in \mathit{epi}\ \hat{f}_i^*\}, \tag{9.1}$$

where $|||\cdot|||$ is any norm whose dual optimal solution provides x_i. Problem (9.1) is clearly related to (1.9) (cf. [Fr98]), but with a decidedly different flavor. On one hand, in (9.1) the cost function of the $\hat{f}^*$-values need not be linear, but, on the other hand, D_t^* in (1.9) need not be norm-like. Furthermore, the treatment in [Nu97] ignores the concept of current point and the updating of the proximal parameter t.

To conclude this section, let us mention that there are important classes of bundle methods that are *not* covered by our analysis: such are *proximal level* methods [LNN95], [HL93b, Algorithm XV.2.3.1], *analytic center cutting plane* methods [Ne95, GV97], *dual ε-descent algorithms* [HL93b, Algorithm XIV.3.4.2], algorithms based on a biobjective view of the direction finding problem [Fu98], and Newton-type bundle methods [LS98, LV98, MSQ98]. The extension of our theory to some of the above algorithms might be possible and is currently under research.

9.2. Algorithms for structured convex problems. It is well known that, under proper assumptions [HL93b, Chap. XII], the convex problem

$$(\mathrm{P}) \qquad \sup_u\{c(u) : h(u) = 0, u \in U\} \tag{9.2}$$

is equivalent to its Lagrangian dual (0.1), where

$$(\mathrm{D}_{\bar{x}}) \qquad f(\bar{x}) = \sup_u\{c(u) + \bar{x}h(u) : u \in U\}. \tag{9.3}$$

Here,

$$f^*(z) = \sup_x \left\{ -\sup_u\{c(u) + x(h(u) - z) : u \in U\} \right\} \tag{9.4}$$

$$= -\sup_u\{c(u) : h(u) = z, u \in U\}. \tag{9.5}$$

((9.4) is the Lagrangian dual of (9.5), whence the identity.) Thus, $-f^*$ is the *value function* of (9.2) w.r.t. the constraints $h(u)$; plugging (9.5) into (1.11), one obtains

$$(\mathrm{D}_{\bar{x},t}) \qquad \sup_u\{c(u) + \bar{x}h(u) - D_t^*(-h(u)) : u \in U\}. \tag{9.6}$$

Hence, generalized bundle methods applied to a Lagrangian dual are approximated generalized augmented Lagrangian approaches to the solution of (9.2). If c and h are affine, and the cutting plane model $\hat{f}_\beta$ is used, in view of (1.7) the stabilized dual master problem (1.9) becomes

(9.7)

$$(\mathrm{D}_{\beta,\bar{x},t}) \qquad \inf_z \left\{ \inf_\theta \left\{ \sum_{u\in\beta} -c(u)\theta_u : \sum_{u\in\beta} h(u)\theta_u = z, \theta \in \Theta \right\} - z\bar{x} + D_t^*(-z) \right\}$$
$$= \sup_u \{ c(u) + \bar{x}h(u) - D_t^*(-h(u)) : u \in Conv(\beta) = U_\beta \},$$

where β is now considered a set of optimal solutions $u_i \in U$ of the dual pricing problem (1.3) such that $z_i = h(u_i)$. Thus, the generalized bundle method uses an *inner linearization* approach, where U is substituted with its inner linearization U_β, to approximately solve $(\mathrm{D}_{\bar{x},t})$. In fact, let u^* be the optimal solution of (9.7); from (4.7), the sequence $\{z_i^* = h(u_i^*)\}$ of optimal solutions of (1.9) corresponds to a sequence $\{u_i^*\}$ of α_i^*-*optimal* solutions for (9.3) (cf. section 6.1). If $\{z_i^*\} \to 0$ and $\{\alpha_i^*\} \to 0$, any cluster point of $\{u_i^*\}$ is optimal for (9.2). Similar results hold for inequality constraints $h(u) \le 0$.

Hence, generalized bundle methods are related to nonquadratic penalty methods. For instance, in [PZ92, PZ94], (9.6) is considered with $\bar{x} = 0$ and $D_t^*(z) = t\sum_i \Phi_\varepsilon^*(z_i) \Rightarrow D_t(z) = t\sum_i \Phi_\varepsilon(\frac{1}{t}d_i)$ for some $\varepsilon > 0$ and

$$\Phi_\varepsilon^*(z_i) = \begin{cases} \dfrac{z_i^2}{2\varepsilon} & \text{if } -\varepsilon \le z_i \le \varepsilon, \\ |z_i| - \dfrac{\varepsilon}{2} & \text{otherwise,} \end{cases} \qquad \Phi_\varepsilon(d_i) = \begin{cases} \dfrac{\varepsilon}{2}d_i^2 & \text{if } -1 \le d_i \le 1, \\ +\infty & \text{otherwise.} \end{cases}$$

Here Φ_ε is a smooth approximation of the nonsmooth exact penalty function $t\|z\|_1$. The algorithm of [PZ94] requires us to compute an exact optimal solution u^* of (9.6) for given t and ε, and then either increases t if $\|h(u^*)\|_\infty > \varepsilon$ (u^* is not ε-feasible), or decreases ε otherwise. The suggested procedure for solving (9.6), used in [PZ92], is *simplicial decomposition*, i.e., inner linearization. Hence, in the affine case the algorithm in [PZ94] is very similar to a three-level bundle algorithm that never performs SS. The only difference is that ε is not decreased to improve the approximation of (9.6) (which is assumed to be exactly solved, although this may not be practical) but rather to force D_t to behave more and more like $t\|\cdot\|_1$; however, this is permitted by our theory. Thus, a generalized bundle method with the above D_t^* offers an alternative to the algorithm of [PZ94], which may be more efficient because (9.6) is only approximately solved and changes of $\bar{x}$ are allowed. Furthermore, we remark that, although Φ_ε is mentioned in [PZ94], the corresponding stabilized primal master problem is not described there; however, the corresponding (1.9) is a box-constrained quadratic problem that could be solved with specialized codes (see [Ki89, Fr96]) more efficiently than the nonlinear problem (9.7).

A similar idea has been used to develop ε-approximation algorithms for (block-)structured convex problems [GK95]. In order to solve (9.2) (with $h(u) \le 0$), (9.6) (with $\bar{x} = 0$) is considered, where

$$D_t^*(z) = \ln \sum_i \mathrm{e}^{tz_i}. \tag{9.8}$$

This D_t^* is a smooth approximation of $t\|z\|_\infty$, i.e., $(\ln n) + t\|z\|_\infty \geq D_t^*(z) \geq t\|z\|_\infty$. Problem (9.6) is then approximately solved with an inner linearization approach, i.e., solving (9.7) and using the gradient of D_t^* in z^* (resp., u^*) to generate a new point z (resp., u). At each step, only the "minimal" bundle $\{u^*, u\}$ is kept. (D_t^* satisfies (P*3$''$)). This approach is not exactly a generalized bundle method, as D_t^* is not zero in the feasible region. However, generalized bundle methods could use slightly modified forms of the above exponential penalty function while allowing changes in $\bar{x}$.

10. Conclusions. We have proved convergence of several variants of generalized bundle methods; different convergence properties can be obtained according to the characteristics of the function to be minimized and of the stabilizing term employed. The statements of the properties needed for convergence allow great flexibility in the implementation of the algorithm; several different t-strategies and β-strategies, which are well known to be crucial in practice, can be fitted within this framework.

Our conditions on D_t are less restrictive than those in [Au87, KCL95, Be96], are different from those in [Ki99], and allow a unified treatment of "penalty-like" and "trust-region-like" stabilizing terms [HL93b, sections XV.2.1 and XV.2.2], which have so far been considered as distinct. Very little regularity is required for $D_t(d)$ as a function of t. Weak requirements on f, such as *-compactness, avoid stronger requirements on D_t. A distinguishing feature of our analysis is the extensive exploitation of a new dual viewpoint of bundle methods. Some algorithms that have been proposed outside the bundle framework [PZ94, GK95] can be shown to be closely related to our class.

Our results suggest that practical implementations of generalized bundle algorithms are possible with several different nonquadratic stabilizing terms; examples are primal and/or dual trust regions based on "linear" (L_1- or L_∞-)norms, which require the solution of just a linear program at each step. Preliminary computational experiences [Be96] seem to confirm the effectiveness of these approaches. Other stabilizing terms, e.g., exponential or linear-quadratic, may exhibit better convergence in practice than the L_2-norm, and thus compensate for the more difficult subproblem to be solved.

Finally, it may be possible to extend these results to an even larger class of bundle algorithms.

Acknowledgments. I'm deeply indebted to Claude Lemaréchal for his precious advice, which considerably improved both the presentation and the contents of this paper; in particular, his contribution was fundamental in correcting an error in a previous version of Theorem 3.1 and its consequences. I'm also grateful to a referee for his numerous and detailed comments and for pointing out several glitches, among which was an error in the proof of Theorem 7.3.

REFERENCES

[Au87] A. Auslender, *Numerical methods for nondifferentiable convex optimization*, Math. Program. Study, 30 (1987), pp. 102–126.

[Au97] A. Auslender, *How to deal with the unbounded in optimization: Theory and algorithms*, Math. Programming, 79 (1997), pp. 3–18.

[ACC93] A. Auslender, R. Cominetti, and J.-P. Crouzeix, *Convex functions with unbounded level sets and applications to duality theory*, SIAM J. Optim., 3 (1993), pp. 669–687.

[Be96] C. Berger, *Contribution à l'Optimisation Non-Différentiable et à la Décomposition en Programmation Mathématique*, Ph.D. Thesis, Département de Mathématiques et de

Génie Industriel, École Polytechnique de Montréal, Montreal, QC, Canada, 1996.

[BPP91] M. Bougeard, J.-P. Penot, and A. Pommelet, *Towards minimal assumptions for the infimal convolution regularization*, J. Approx. Theory, 64 (1991), pp. 245–270.

[CL93] R. Correa and C. Lemaréchal, *Convergence of some algorithms for convex minimization*, Math. Programming, 62 (1993), pp. 261–275.

[CT93] G. Chen and M. Teboulle, *Convergence analysis of a proximal-like minimization algorithm using Bregman functions*, SIAM J. Optim., 3 (1993), pp. 538–543.

[CFG01] T. G. Crainic, A. Frangioni, and B. Gendron, *Bundle-based relaxation methods for multicommodity capacitated fixed charge network design problems*, Discrete Appl. Math., 112 (2001), pp. 73–99.

[Ec93] J. Eckstein, *Nonlinear proximal point algorithms using Bregman functions with applications to convex programming*, Math. Oper. Res., 18 (1993), pp. 292–226.

[Fr96] A. Frangioni, *Solving semidefinite quadratic problems within nonsmooth optimization algorithms*, Comput. Oper. Res., 23 (1996), pp. 1099–1118.

[Fr97] A. Frangioni, *Dual Ascent Methods and Multicommodity Flow Problems*, Ph.D. Dissertation, TD 5/97, Dipartimento di Informatica, Università di Pisa, Pisa, Italy, 1997.

[Fr98] A. Frangioni, *Generalized Bundle Methods*, Technical report TR 04/98, Dipartimento di Informatica, Università di Pisa, Pisa, Italy, 1998.

[Fu98] A. Fuduli, *Metodi Numerici per la Minimizzazione di Funzioni Convesse NonDifferenziabili*, Ph.D. Thesis, DEIS, Università della Calabria, Calabria, Italy, 1998.

[GK95] M. D. Grigoriadis and L. G. Kahchiyan, *An exponential-function reduction method for block-angular convex programs*, Networks, 26 (1995), pp. 59–68.

[GM91] M. Gaudioso and M. F. Monaco, *Quadratic approximations in convex nondifferentiable optimization*, SIAM J. Control Optim., 29 (1991), pp. 58–70.

[GV97] J. Gondzio and J.-P. Vial, *Warm Start and ε-Subgradients in Cutting Plane Scheme for Block-angular Linear Programs*, Logilab Technical report, 1997.1, Paris, France, 1997.

[HL93a] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms* I—*Fundamentals*, Grundlehren Math. Wiss. 305, Springer-Verlag, New York, 1993.

[HL93b] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms* II—*Advanced Theory and Bundle Methods*, Grundlehren Math. Wiss. 306, Springer-Verlag, New York, 1993.

[IST94] A. N. Iusem, B. F. Svaiter, and M. Teboulle, *Entropy-like proximal methods in convex programming*, Math. Oper. Res., 19 (1994), pp. 790–814.

[IT95] A. N. Iusem and M. Teboulle, *Convergence rate analysis of nonquadratic proximal methods for convex and linear programming*, Math. Oper. Res., 20 (1994), pp. 657–677.

[Ki89] K. C. Kiwiel, *A dual method for certain positive semidefinite quadratic programming problems*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 175–186.

[Ki95] K. C. Kiwiel, *Approximations in proximal bundle methods and decomposition of convex programs*, J. Optim. Theory Appl., 84 (1997), pp. 529–548.

[Ki98] K. Kiwiel, *Proximal minimization methods with generalized Bregman functions*, SIAM J. Control Optim., 35 (1997), pp. 1142–1168.

[Ki99] K. Kiwiel, *A bundle Bregman proximal method for convex nondifferentiable optimization*, Math. Program., 85 (1999), pp. 241–258.

[KCL95] S. Kim, K. N. Chang, and J. Y. Lee, *A descent method with linear programming subproblems for nondifferentiable convex optimization*, Math. Programming, 71 (1995), pp. 17–28.

[LNN95] C. Lemaréchal, A. Nemirovskii, and Y. Nesterov, *New variants of bundle methods*, Math. Programming, 69 (1995), pp. 111–147.

[LS98] C. Lemaréchal and C. Sagastizàbal, *Variable metric bundle methods: From conceptual to implementable forms*, Math. Programming, 16 (1997), pp. 393–410.

[LV98] L. Luksan and J. Vlcek, *A bundle-Newton method for nonsmooth unconstrained optimization*, Math. Programming, 83 (1998), pp. 373–391.

[MSQ98] R. Mifflin, D. Sun, and L. Qi, *Quasi-Newton bundle-type methods for nondifferentiable convex optimization*, SIAM J. Optim., 8 (1998), pp. 583–603.

[Ne95] Y. Nesterov, *Complexity estimates of some cutting plane methods based on the analytic barrier*, Math. Programming, 69 (1995), pp. 149–176.

[Nu97] E. A. Nurminski, *Separating plane algorithms for convex optimization*, Math. Programming, 76 (1997), pp. 373–391.

[OR70] J. M. Ortega and W. C. Rheinboldt, *Iterative Solutions of Nonlinear Equations in*

Several Variables, Academic Press, New York, 1970.

[PZ92] M. C. Pinar and S. A. Zenios, *Parallel decomposition of multicommodity network flows using a linear-quadratic penalty algorithm*, ORSA J. Comput., 4 (1992), pp. 235–248.

[PZ94] M. C. Pinar and S. A. Zenios, *On smoothing exact penalty functions for convex constrained optimization*, SIAM J. Optim., 4 (1994), pp. 486–511.

[Ro70] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[SZ92] H. Schramm and J. Zowe, *A version of the bundle idea for minimizing a nonsmooth function: Conceptual idea, convergence analysis, numerical results*, SIAM J. Optim., 2 (1992), pp. 121–152.

[Te97] M. Teboulle, *Convergence of proximal-like algorithms*, SIAM J. Optim., 7 (1997), pp. 1069–1083.

SIAM J. OPTIM.
Vol. 13, No. 1, pp. 157–167

ON THE PROBLEM OF OPTIMAL CUTTING*

DORIN BUCUR†, GIUSEPPE BUTTAZZO‡, AND NICOLAS VARCHON†

Abstract. This paper deals with the existence of an optimal cutting in a membrane satisfying the following assumption: it has to connect two given points in order to leave the membrane the strongest possible. We prove the existence of a solution for this problem in a rather general setting, and we present some open questions related to the regularity of the optimum or to possible extensions for plates in the elasticity framework.

Key words. optimal cut, membrane, variation of a crack

AMS subject classifications. 49Q10, 35J20, 35B20

PII. S1052623401387118

1. Introduction.

The main purpose of this paper is to raise the question of existence and regularity of an optimal "cut" in a membrane with the only constraint being that the cut has to connect two or more a priori given points and to leave the membrane the strongest possible. More precisely, the problem, illustrated in Figure 1.1, can be modeled as follows.

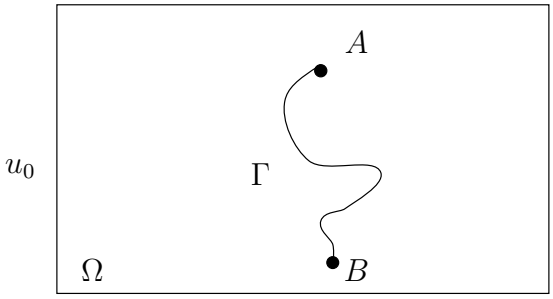


FIG. 1.1. *An admissible cut* Γ *in the membrane* Ω.

Notice that, in general, admissible cuts Γ do not need to be curves; for instance, if we want to connect three points we have to expect that the optimal cut has a triple junction shape.

Let Ω be a two-dimensional bounded open set with a smooth boundary (say, the rectangle in the figure above), $u_0 \in H^1(\Omega)$, and A, B two given points in Ω. An admissible cut in Ω will simply be a compact connected subset Γ of $\overline{\Omega}$ containing the

*Received by the editors April 2, 2001; accepted for publication (in revised form) January 9, 2002; published electronically June 26, 2002.
http://www.siam.org/journals/siopt/13-1/38711.html

†CNRS-Département de Mathématiques, Université de Franche-Comté, 16 route de Gray, 25030 Besançon, France (bucur@math.univ-fcomte.fr, varchon@math.univ-fcomte.fr). The research of the first author was completed during a stay at the University of Pisa as an INDAM visitor; he thanks the Dipartimento di Matematica for their kind hospitality.

‡Dipartimento di Matematica, Università di Pisa, Via Buonarroti 2, 56127 Pisa, Italy (buttazzo@dm.unipi.it). The research of this author is part of the European Research Training Network "Homogenization and Multiple Scales" under contract HPRN-2000-00109.

points A and B, and we denote by $\mathcal{U}_{ad}$ the class of admissible cuts, that is,

$$\mathcal{U}_{ad} = \big\{\Gamma \subseteq \overline{\Omega} \ : \ A, B \in \Gamma, \ \Gamma \text{ is compact and connected}\big\}.$$

For every $\Gamma \in \mathcal{U}_{ad}$, the energy $\mathcal{E}(\Gamma)$ associated with Γ will be

$$\mathcal{E}(\Gamma) = \min\left\{\int_{\Omega\setminus\Gamma} |\nabla u|^2\,dx \ : \ u \in H^1_{loc}(\Omega\setminus\Gamma),\ u = u_0 \text{ on } \partial\Omega\right\},$$

so that the optimization problem we deal with can be written as

$$\max\big\{\mathcal{E}(\Gamma) \ : \ \Gamma \in \mathcal{U}_{ad}\big\}. \tag{1.1}$$

For fixed $\Gamma \in \mathcal{U}_{ad}$, the function $u_\Gamma \in H^1_{loc}(\Omega\setminus\Gamma)$ minimizing the energy of the membrane is the weak variational solution of the following problem:

$$\begin{cases} -\Delta u_\Gamma = 0 \ \text{ in } \Omega\setminus\Gamma, \\ \frac{\partial u_\Gamma}{\partial n} = 0 \ \text{ on } \partial\Gamma, \quad u_\Gamma = u_0 \ \text{ on } \partial\Omega\setminus\Gamma. \end{cases} \tag{1.2}$$

A similar situation occurs in the so-called image segmentation problem, where, given a function $g \in L^2(\Omega)$, the energy of a segmentation Γ is

$$\mathcal{E}(\Gamma) = \min\left\{\int_{\Omega\setminus\Gamma} |\nabla u|^2\,dx + \int_{\Omega\setminus\Gamma} (u-g)^2\,dx \ : \ u \in H^1(\Omega\setminus\Gamma)\right\}.$$

The optimal segmentation of the image g is then obtained by minimizing the Mumford–Shah functional, i.e., by solving the optimization problem

$$\min\big\{\mathcal{E}(\Gamma) + \mathcal{H}^1(\Gamma) \ : \ \Gamma \in \mathcal{U}_{ad}\big\}, \tag{1.3}$$

where $\mathcal{U}_{ad}$ is the family of compact sets contained in $\overline{\Omega}$ and $\mathcal{H}^1(\Gamma)$ is the one-dimensional Hausdorff measure of Γ. For a given Γ, the minimizer u_Γ of the energy satisfies in this case the following equation:

$$\begin{cases} -\Delta u_\Gamma + u_\Gamma = g \text{ in } \Omega\setminus\Gamma, \\ \frac{\partial u_\Gamma}{\partial n} = 0 \text{ on } \Gamma\cup\partial\Omega. \end{cases} \tag{1.4}$$

Let us observe that problems (1.1) and (1.3) are somehow similar in the sense that for a given Γ the minimizer of the energy solves an elliptic equation with homogeneous Neumann boundary conditions on Γ. From this point of view, one has to study in both problems the dependence of the solution of a Neumann problem on the geometric variation of Γ.

Problems (1.1) and (1.3) are nevertheless deeply different. First of all, in problem (1.1) one has to maximize the energy, while in problem (1.3) one has to minimize it. This is the main reason why problem (1.3) can be seen as a minimum problem in the space SBV (see [2]), the "crack" Γ being seen as the "jump set" of the SBV function u.

A second difference is that the elliptic equation (1.4) has a zero order term; therefore the solution belongs to the Sobolev space $H^1(\Omega\setminus\Gamma)$. The solution of problem (1.2) belongs only to the Dirichlet space (see [15])

$$L^{1,2}(\Omega\setminus\Gamma) = \big\{u \in H^1_{loc}(\Omega\setminus\Gamma) \ : \ \nabla u \in L^2(\Omega\setminus\Gamma)\big\},$$

which coincides with $H^1(\Omega \setminus \Gamma)$ only if Γ is smooth enough, for instance, Lipschitz continuous and connected (see [17, Corollary 2.2, p. 21]).

Another main difference between problems (1.1) and (1.3) is the presence of the penalty term given by the Hausdorff measure. Without this term the minimizer of the Mumford–Shah functional would not exist in general, the infimum being equal to zero. On the other hand, the presence of this term in functional (1.1) is not necessary (for the existence of a solution). From an intuitive point of view, since one looks for the strongest membrane, the length of the crack should not be too big. One could add in problem (1.1) a penalty term given by the Hausdorff measure by considering the following:

$$\max_{\Gamma\in\mathcal{U}_{ad}} \min\left\{\int_{\Omega\setminus\Gamma} |\nabla u|^2\,dx - \alpha\mathcal{H}^1(\Gamma) : u \in H^1_{loc}(\Omega\setminus\Gamma), u = u_0 \text{ on } \partial\Omega\setminus\Gamma\right\}, \tag{1.5}$$

where $\alpha > 0$ is fixed. In this case, the existence of an optimal crack could be derived as a consequence of the result of Chambolle and Doveri [11].

An approach by duality was followed by Dal Maso and Toader in [12] where they studied a model for the quasi-static growth of brittle fracture. They dealt with an optimization problem of the type

$$\min_{\Gamma\in\mathcal{U}_{ad}} \min\left\{\int_{\Omega\setminus\Gamma} |\nabla u|^2\,dx + \alpha\mathcal{H}^1(\Gamma) : u \in H^1_{loc}(\Omega\setminus\Gamma), u = u_0 \text{ on } \partial\Omega\setminus\Gamma\right\},$$

for which they proved the existence of an optimum.

The question of slitting a membrane and keeping a certain energy functional unchanged has already been studied in the literature. We refer the reader to the pioneering papers of Weinberger [21] and Hersch [16]. For preserving the first eigenvalue of the Laplacian, the cut is done along curves where the normal derivative of the eigenfunction vanishes. This idea is discussed in a different manner in section 3. The question of extremizing harmonic measures over various slit placements was studied in [4, 5, 13]. Certain motivations and extensions arising in more physical problems can be found in [18, 19].

In this paper, we prove the existence of a solution for problem (1.1) in a general setting: an anisotropic membrane subject to an external force. The main technical difficulty of this setting is that, for a given Γ, the function which minimizes the energy does not have, in general, a conjugate; therefore a direct approach by duality is not possible. We refer to [8, 9] for some results in shape optimization of cracks based on duality. Here, we use an approach based on the Mosco convergence (see [3] for the exact definition) of $L^{1,2}$-spaces along with an adjustment procedure for the traces on $\partial\Omega$ in order to prove the existence of a solution to problem (1.1).

Section 2 is devoted to the proof of the existence result. The reader who is not interested in this proof, which is rather technical, can skip section 2 and go directly to section 3, which can be read independently. There we give some intuitive examples and find the exact solution in some particular situations.

2. Existence of an optimal cutting. Let Ω be a two-dimensional bounded open connected set. For simplicity, we suppose that the boundary of Ω is Lipschitz (see [14]). Consequently the number of the connected components of Ω^c is finite.

For $i = 1, \dots, l$ let K_i be l compact sets contained in $\overline{\Omega}$, and $K \subseteq \overline{\Omega}$ be a compact set such that $\cup_{i=1}^{l} K_i \subseteq K$. Let $f \in L^2(\Omega)$ such that supp $f \cap K = \emptyset$.

Remark 2.1. The assumption that supp $f \cap K = \emptyset$ is made for technical reasons that will be clear in the proof of Theorem 2.2. However, we want to stress the fact that the most interesting case is when $f \equiv 0$, so that the only datum of the problem is the boundary condition u_0.

We also notice that when the datum K is regular enough (for instance, with a Lipschitz boundary), then, thanks to the equality $L^{1,2}(\Omega \setminus K) = H^1(\Omega \setminus K)$, the assumption supp $f \cap K = \emptyset$ can be relaxed into the weaker one $f = 0$ a.e. on K.

Notice also that the optimization criterion (1.1) rules out the admissible Γ with $\mathcal{E}(\Gamma) = -\infty$. This automatically implies that the optimization (1.1) is performed on the class of cuts Γ such that the integral of f vanishes on every connected component of Γ which does not touch the boundary $\partial\Omega$ on a set of positive capacity.

In what follows, we denote by $\mathcal{U}_{ad}$ the following admissible class of "cuts" which is supposed to be nonempty:

$$\mathcal{U}_{ad} = \big\{\Gamma \ : \quad \Gamma = \cup_{i=1}^{l} \Gamma^i, \\ \forall i = 1, \dots, l, \quad K_i \subseteq \Gamma^i \subseteq K,\ \Gamma^i \text{ compact connected}\big\}.$$

For every $\Gamma \in \mathcal{U}_{ad}$ we consider the energy

$$\mathcal{E}(\Gamma) = \min\big\{E(u,\Gamma) \ : \ u \in H^1_{loc}(\Omega \setminus \Gamma),\ u = u_0 \text{ on } \partial\Omega\big\}, \tag{2.1}$$

where

$$E(u,\Gamma) = \frac{1}{2}\int_{\Omega\setminus\Gamma} \langle A\nabla u.\nabla u\rangle\, dx - \int_\Omega f u\, dx.$$

Here $u_0 \in H^1(\Omega)$ is a given function and $A \in L^\infty(\Omega, \mathbb{R}^4)$ is a given symmetric matrix satisfying for some $\alpha > 0$ the ellipticity condition

$$\langle A\xi.\xi\rangle \geq \alpha|\xi|^2 \quad \text{for every } \ \xi \in \mathbb{R}^2.$$

If $u \in H^1_{loc}(\Omega \setminus \Gamma)$, the trace of u on $\partial\Omega$ does not exist in general, even if $\partial\Omega$ is smooth. Nevertheless, in our case $\nabla u \in L^2(\Omega \setminus \Gamma)$; hence u belongs to the Dirichlet space $L^{1,2}(\Omega \setminus \Gamma)$ (see [15]). In that case, the trace of u on $\partial\Omega \setminus \Gamma$ is well defined, since $\partial\Omega$ is supposed to be Lipschitz. A second equivalent way to give sense to the equality $u = u_0$ on $\partial\Omega \setminus \Gamma$ is as follows. Let us fix an extension u_0^* of u_0 outside Ω, say in $\Omega^* \setminus \Omega$, where Ω^* is a Lipschitz bounded open set such that $\overline{\Omega} \subseteq \Omega^*$. The trace of u is equal to u_0 on $\partial\Omega \setminus \Gamma$ if and only if the function

$$u^* = \begin{cases} u(x) \text{ if } x \in \Omega \setminus \Gamma, \\ u_0^*(x) \text{ if } x \in \Omega^* \setminus (\Omega \cup \Gamma) \end{cases} \tag{2.2}$$

belongs to $L^{1,2}(\Omega^* \setminus \Gamma)$.

For every fixed $\Gamma \in \mathcal{U}_{ad}$, problem (2.1) has a solution. This is an immediate consequence of the fact that the support of f is compactly embedded in $\overline{\Omega} \setminus K$ and that, thanks to Remark 2.1, the integral of f vanishes on the connected sets of $\overline{\Omega} \setminus K$ not touching $\partial\Omega$ on a set of positive capacity. In fact, if a connected component of $\overline{\Omega} \setminus \Gamma$ contains a part of the support of f and does not touch $\partial\Omega$ on a set of positive capacity, in this region the solution is defined up to a constant, the gradient being fixed. With this remark, the solution is unique (more precisely its gradient is unique) and belongs to the Dirichlet space $L^{1,2}(\Omega \setminus \Gamma)$.

The main result of this section is contained in the following theorem.

THEOREM 2.2. *The optimization problem*

$$\max\left\{\mathcal{E}(\Gamma) \ : \ \Gamma \in \mathcal{U}_{ad}\right\} \tag{2.3}$$

has at least one solution.

Proof. In order to prove the existence of a solution for problem (2.3), we follow the direct method of the calculus of variations. Let $\{\Gamma_n\}_n \subseteq \mathcal{U}_{ad}$ be a maximizing sequence for (2.3). Without loss of generality, we can suppose that for every $i = 1, \dots, l$

$$\Gamma_n^i \xrightarrow{H} \Gamma^i,$$

the convergence being understood in the Hausdorff metric (see, for instance, [11, 20]). We denote $\Gamma = \cup_{i=1}^{l} \Gamma^i$, the Hausdorff limit of Γ_n. Our purpose is to prove that Γ is a solution for problem (2.3). Notice that for every $i = 1, \dots, l$ the set Γ^i is compact and connected and $K_i \subseteq \Gamma_i \subseteq K$; hence $\Gamma \in \mathcal{U}_{ad}$.

It remains to prove that for every $u \in L^{1,2}(\Omega \setminus \Gamma)$ with $u = u_0$ on $\partial\Omega$ there exists a sequence $\{u_n\}_n$ such that $u_n \in L^{1,2}(\Omega \setminus \Gamma_n)$ with $u_n = u_0$ on $\partial\Omega \setminus \Gamma_n$ and

$$E(\Gamma, u) \geq \limsup_{n\to\infty} E(\Gamma_n, u_n). \tag{2.4}$$

The construction of the sequence $\{u_n\}_n$ is strongly related to the Mosco convergence of the spaces $L^{1,2}(\Omega \setminus \Gamma_n)$ (see [8] for the construction of the sequence). We observe that if $u \notin L^{1,2}(\Omega \setminus \Gamma)$, then $E(u, \Gamma) = +\infty$ and inequality (2.4) holds trivially. For $u \in L^{1,2}(\Omega \setminus \Gamma)$ we construct a sequence $u_n \in L^{1,2}(\Omega \setminus \Gamma_n)$ with $u = u_0$ on $\partial\Omega \setminus \Gamma_n$ such that

$$\tilde{\nabla} u_n \to \tilde{\nabla} u \qquad \text{strongly in } L^2(\Omega) \tag{2.5}$$

and

$$\int_\Omega u_n f dx \to \int_\Omega u f dx.$$

In relation (2.5) we denoted by $\tilde{\nabla} u_n$ the extension by zero of ∇u_n on Γ_n, since ∇u_n is a priori defined only on $\Omega \setminus \Gamma_n$. Of course, the function $\tilde{\nabla} u_n$ is not anymore a gradient on Ω.

In order to construct the sequence $\{u_n\}_n$ we recall the following lemma, which is a consequence of [8, Theorem 4.1].

LEMMA 2.3. *Let $\Gamma_n, \Gamma \in \mathcal{U}_{ad}$ be such that $\Gamma_n \xrightarrow{H} \Gamma$. Then for every $u \in L^{1,2}(\Omega \setminus \Gamma)$ there exists $u_n \in \{L^{1,2}(\Omega \setminus \Gamma_n)\}_n$ such that $\tilde{\nabla} u_n \longrightarrow \tilde{\nabla} u$ strongly in $L^2(\Omega)$.*

Notice that the sequence $\{u_n\}_n$ given by Lemma 2.3 is not sufficient to conclude the proof of Theorem 2.2, since the trace of u_n on $\partial\Omega \setminus \Gamma_n$ is not equal to u_0. Nevertheless, Lemma 2.3 can be used to prove the following proposition.

PROPOSITION 2.4. *Let $\Gamma_n, \Gamma \in \mathcal{U}_{ad}$ be such that $\Gamma_n \xrightarrow{H} \Gamma$. Then for every $u \in L^{1,2}(\Omega \setminus \Gamma)$ such that $u_{|\partial\Omega\setminus\Gamma} = u_0$ there exists a sequence $u_n \in L^{1,2}(\Omega \setminus \Gamma_n)$ such that $\tilde{\nabla} u_n \longrightarrow \tilde{\nabla} u$ strongly in $L^2(\Omega)$ and $u_{n|\partial\Omega\setminus\Gamma_n} = u_0$.*

Proof. Let us denote by u^* the extension of u by u_0^* on $\Omega^* \setminus \Omega$. Then we apply Lemma 2.3 to Ω^* and Γ_n, Γ, and we find a sequence $u_n^* \in L^{1,2}(\Omega^* \setminus \Gamma_n)$ such that $\tilde{\nabla} u_n^* \longrightarrow \tilde{\nabla} u^*$ strongly in $L^2(\Omega^*)$.

For every $n \in \mathbb{N}$, let us denote by u_n the solution of the minimization problem

$$\min\left\{\int_{\Omega^*\setminus\Gamma_n} |\nabla\phi - \nabla u_n^*|^2\,dx : \phi \in L^{1,2}(\Omega^*\setminus\Gamma_n), \phi = u_0^* \text{ on } \Omega^*\setminus\Omega\right\}. \tag{2.6}$$

Since $u_n - u_n^* \in L^{1,2}(\Omega^*\setminus\Gamma_n)$ and since $\Omega^*\setminus\Omega$ is Lipschitz, we get that $u_n - u_n^* \in H^1(\Omega^*\setminus\Omega)$. Moreover, there exists a bounded continuous linear extension operator T from $H^1(\Omega^*\setminus\Omega)$ to $H^1(\Omega^*)$. Taking as a test function in (2.6) the function $\phi = u_n^* + T(u_n - u_n^*)$, we get

$$\min\left\{\int_{\Omega^*\setminus\Gamma_n} |\nabla\phi - \nabla u_n^*|^2\,dx : \phi \in L^{1,2}(\Omega^*\setminus\Gamma_n), \phi = u_0^* \text{ on } \Omega^*\setminus\Omega\right\}$$

$$\leq \int_{\Omega^*\setminus\Gamma_n} |\nabla T(u_n - u_n^*)|^2 dx.$$

Using the Poincaré inequality on the space $\{u \in H^1(\Omega^*) : \int_{\partial\Omega^*} u dx = 0\}$ and the boundedness of the extension operator T, we get

$$\int_{\Omega^*\setminus\Gamma_n} |\nabla T((u_n - u_n^*)_{|\Omega^*\setminus\Omega})|^2 dx \leq C\int_{\Omega^*\setminus\Omega} |\nabla(u_n - u_n^*)|^2 dx$$

$$= C\int_{\Omega^*\setminus\Omega} |\nabla(u_0^* - u_n^*)|^2 dx.$$

This last term converges to zero from Lemma 2.3.

Taking the restrictions of u_n to $\Omega\setminus\Gamma_n$, all the requirements are satisfied and the proof is concluded. □

Proof of Theorem 2.2 *(continuation).* Returning to the proof of Theorem 2.2, we observe that the sequence $\{u_n\}_n$ defined in Proposition 2.4 satisfies relation (2.4). Indeed, the gradients extended by zero converge strongly in L^2 by construction; hence, using the boundedness of A, we have

$$\int_\Omega \langle A\tilde{\nabla}u_n, \tilde{\nabla}u_n\rangle dx \to \int_\Omega \langle A\tilde{\nabla}u, \tilde{\nabla}u\rangle dx.$$

It remains to prove that

$$\int_\Omega u_n f dx \to \int_\Omega u f dx.$$

Fix a connected component U of $\overline{\Omega}\setminus K$ containing a part of the support of f. Two possibilities may occur.

Suppose first that $\mathrm{cap}(\mathrm{U}\cap\partial\Omega) > 0$. Since $\partial\Omega$ is Lipschitz and Γ is closed, the set $\partial\Omega\setminus\Gamma$ is relatively open; hence there exists an open Lipschitz set V such that $\mathrm{supp}\, f \subseteq V \subseteq U$ and $\mathrm{cap}(\overline{\mathrm{V}}\cap\partial\Omega) > 0$. Then, the Poincaré inequality stands true in $H^1((\Omega^*\setminus\overline{\Omega})\cup V)$, so that $u_n \longrightarrow u$ strongly in $L^2(V)$, which implies

$$\int_V u_n f dx \to \int_V u f dx.$$

Suppose now that $\mathrm{cap}(\mathrm{U} \cap \partial\Omega) = 0$. In this case there exists an open Lipschitz set V such that $\mathrm{supp}\, f \subseteq V \subseteq U$ and $\overline{V} \cap \partial\Omega = \emptyset$. By hypothesis, we have that $\int_V f dx = 0$; hence the Poincaré inequality holds in $H^1(V)/\mathbb{R}$. Consequently

$$\int_V u_n f dx \to \int_V u f dx.$$

The support of f being compactly contained in $\overline{\Omega} \setminus K$, the proof is concluded. □

If K does not touch $\partial\Omega$, one could drop the hypothesis on the regularity of Ω by simply imposing a constraint of the type $(u - u_0)\varphi \in H_0^1(\Omega \setminus \Gamma)$, where $\varphi \in C^\infty(\mathbb{R}^2)$ is a fixed function such that $\varphi = 0$ on Γ, and use a partition of unity.

Remark 2.5. Let $c \geq 0$. If we denote

$$\mathcal{U}_{ad}^c = \{\Gamma \in \mathcal{U}_{ad}, |\Gamma| \geq c\},$$

the existence result of Theorem 2.2 still holds in $\mathcal{U}_{ad}^c$. Indeed, the proof does not change; the only point to be verified is that $\mathcal{U}_{ad}^c$ is closed for the Hausdorff convergence. This is a direct consequence of the upper semicontinuity of the Lebesgue measure for the Hausdorff convergence.

Remark 2.6. If in the previous remark we take $c > 0$ and $K_i = \emptyset$ for $i = 1, \ldots, l$, the optimal cutting problem becomes, roughly speaking, the following: *find the "strongest" membrane attached on the boundary, of measure less than or equal to* $|\Omega| - c$ *and with at most* l *holes.*

A vector version of this problem, set into the elasticity frame, is the so-called cantilever problem (see [1] and [10]). One of the main difficulties is to manage the fact that the Korn inequality fails to be true on nonsmooth domains. We refer the reader to the recent paper of Chambolle [10] for an existence result for this problem.

3. Additional remarks and open problems. The uniqueness of the optimal cut does not hold in general. Trivially, let $u_0 \equiv 0$, $f \equiv 0$, $K_1 = \{A, B\}$, $K = \overline{\Omega}$. Then any compact connected set containing A and B solves problem (2.3).

In some particular situations one can produce at least one solution of the problem. In a symmetric setting, there exists an optimal cut which is also symmetric. Indeed, let $f \equiv 0$ and Ω be a rectangle; let d be a symmetry line of the rectangle. Suppose that $K_1 = \{A, B\}$ are two points on d and that u_0 is also symmetric with respect to d. It can be easily seen that a solution of problem (2.3) (with $K = \overline{\Omega}$) is the segment AB. This follows by simply observing that the harmonic function in the rectangle which is equal to u_0 on the boundary of the rectangle has zero normal derivatives on the segment AB.

There are some other situations when the position of the optimal cut can be determined. Let Ω be a simply connected bounded open subset of $\mathbb{R}^2$ with Lipschitz boundary and let $f \equiv 0$. Let $K_1 = \{A, B\}$, the points A and B being placed on a connected level set of the conjugate function of the harmonic function v in Ω which is equal to u_0 on $\partial\Omega$, i.e.,

$$\begin{cases} -\Delta v = 0 \text{ in } \Omega, \\ v = u_0 \text{ on } \partial\Omega. \end{cases} \tag{3.1}$$

Since Ω is simply connected, we recall that v has a harmonic conjugate, denoted ϕ and determined by the Cauchy–Riemann relations

$$\frac{\partial \phi}{\partial x} = \frac{\partial v}{\partial y} \text{ and } \frac{\partial \phi}{\partial y} = -\frac{\partial v}{\partial x}.$$

Suppose that $\phi(A) = \phi(B)$, A, B belonging to a connected component of $\phi^{-1}(\phi(A))$. Let set Γ be a compact connected subset of $\phi^{-1}(\phi(A))$ containing A, B. In order to simplify the following proof, we assume that Γ does not touch the boundary of Ω. All results remain true if Γ touches $\partial\Omega$, but a few technical difficulties appear if $\Omega \setminus \Gamma$ is not connected.

Under the previous assumptions, we observe that

$$\min\left\{\int_\Omega |\nabla u|^2\,dx : u \in H^1(\Omega), u = u_0 \text{ on } \partial\Omega\right\} \tag{3.2}$$

$$= \min\left\{\int_{\Omega\setminus\Gamma} |\nabla u|^2\,dx : u \in H^1_{loc}(\Omega\setminus\Gamma), u = u_0 \text{ on } \partial\Omega\right\}.$$

Indeed, on the one hand, we obviously have

$$\min\left\{\int_\Omega |\nabla u|^2\,dx : u \in H^1(\Omega), u = u_0 \text{ on } \partial\Omega\right\}$$

$$\geq \min\left\{\int_{\Omega\setminus\Gamma} |\nabla u|^2\,dx : u \in H^1_{loc}(\Omega\setminus\Gamma), u = u_0 \text{ on } \partial\Omega\right\}.$$

On the other hand, let $v^* \in H^1_{loc}(\Omega\setminus\Gamma)$ be the minimizer in the right-hand side of the previous relation. (Observe that all connected components of $\Omega\setminus\Gamma$ touch $\partial\Omega$ since v^* is harmonic.)

We give the following.

LEMMA 3.1. *There exists a function $\phi^* \in H^1_0(\Omega)$ and a constant $c^* \in \mathbb{R}$ such that $\nabla v^* = \text{curl } \phi^*$ in $\Omega\setminus\Gamma$ and*

$$\begin{cases} -\Delta\phi^* = 0 & \text{in} \quad \Omega\setminus\Gamma, \\ \phi^* = c^* & \text{on} \quad \Gamma, \\ \frac{\partial\phi^*}{\partial n} = \partial_t u_0 & \text{on} \quad \partial\Omega, \end{cases} \tag{3.3}$$

where $\partial_t u_0$ is the weak tangential derivative in the sense of distribution on $\partial\Omega$.

Since Γ is not smooth, the meaning of the equality $\phi^* = c^*$ on Γ is understood in the sense of traces on nonsmooth sets, that is $\tilde\phi^* = c^*$ quasi-everywhere on Γ (i.e., up to a set of capacity zero). Here $\tilde\phi^*$ is a quasi-continuous representative of ϕ^* (see [14] or [15]).

Proof. The function ϕ^* is the harmonic conjugate of v^* in $\Omega\setminus\Gamma$. The proof is a consequence of the fact that $\partial\Omega$ is connected and the following equalities hold:

$$\int_{\partial\Omega} \frac{\partial v^*}{\partial n} d\mathcal{H}^1 = 0,$$

$$\frac{\partial v^*}{\partial n} = 0 \text{ on } \Gamma.$$

The existence of the conjugate is a consequence of a result of [17, Theorem 3.1, p. 37] through an approximation procedure of the nonsmooth boundary Γ. We refer to [8] where this kind of result was proved in a slightly different setting. □

If ϕ is the harmonic conjugate of v in Ω, then we have

$$\begin{cases} -\Delta\phi = 0 \text{ in } \Omega, \\ \frac{\partial\phi}{\partial n} = \partial_t u_0 \text{ on } \partial\Omega. \end{cases} \tag{3.4}$$

By hypothesis, we also have $\phi = c$ on Γ. Consequently, by subtraction we get

$$\left\{\begin{array}{rcl} -\Delta(\phi-\phi^*)=0 & \text{in} & \Omega\setminus\Gamma,\\ \phi-\phi^*=c-c^* & \text{on} & \Gamma,\\ \frac{\partial(\phi-\phi^*)}{\partial n}=0 & \text{on} & \partial\Omega. \end{array}\right. \tag{3.5}$$

If $c - c^* \neq 0$, then $\phi - \phi^*$ attains its minimum (or maximum, or both) on $\partial\Omega$, and in that point we would have, following the Hopf maximum principle, $\frac{\partial\phi^*}{\partial n} \neq 0$, which is a contradiction with (3.5). Therefore $c - c^* = 0$; hence $\phi = \phi^*$ in $\Omega \setminus \Gamma$ and $v = v^*$ in $\Omega \setminus \Gamma$. Since the level set of a nonconstant harmonic function cannot have a positive Lebesgue measure, the set Γ, which is contained in the level set $\{\phi = c\}$, has measure zero. Then we get

$$\int_{\Omega\setminus\Gamma} |\nabla v^*|^2\,dx = \int_\Omega |\nabla v|^2\,dx$$

$$\geq \min\left\{\int_\Omega |\nabla u|^2\,dx : u\in H^1(\Omega), u = u_0 \text{ on } \partial\Omega\right\}.$$

Consequently, equality (3.2) holds, and Γ is a solution for problem (2.3).

A natural question is whether the optimal cut Γ touches the boundary of Ω or not. There are indeed some situations in which the optimal cut Γ necessarily touches the boundary. We give the following example. Let Ω be the rectangle

$$\Omega = \{(x,y)\in\mathbb{R}^2 \ : \ x\in[0,1],\ y\in[-1,1]\}.$$

Let $u(x,y) = 2(x+1)y$ and $u_0 \in C(\partial\Omega)$ be given by the trace of u on $\partial\Omega$, i.e., $u_0 = u\big|_{\partial\Omega}$. We take $f \equiv 0$ and $K_1 = \{A, B\}$ with $A = (\frac{\sqrt{5}}{2} - 1, \frac{1}{2})$, $B = (\frac{\sqrt{5}}{2} - 1, -\frac{1}{2})$, and we take $K = \overline{\Omega}$. The optimal cut solving problem (2.3) is given by the curve

$$\Gamma = \left\{(x,y)\in\mathbb{R}^2 \ : \ x = \sqrt{y^2+1} - 1,\ y\in\left[-\frac{1}{2},\frac{1}{2}\right]\right\},$$

which obviously touches the boundary of Ω at the origin.

The proof of the optimality of Γ follows the same scheme as presented in the beginning of this section, where we studied the possibility of identifying the optimal cut. In fact, one can observe that the points A and B belong to a level set of the function $v(x,y) = (x+1)^2 - y^2$ which is the harmonic conjugate of u. One can, moreover, observe that all optimal cuts in $\overline{\Omega}$ must touch the boundary. This is an easy consequence of the uniqueness of the minimizer of $E(u,\Gamma)$ over $u \in L^{1,2}(\Omega\setminus\Gamma)$, with $u = u_0$ on $\partial\Omega$.

Remark 3.2. When the dimension n is greater than 2, the formulation of the optimal cutting problem (1.1) becomes trivial. Indeed, since curves have zero capacity in $\mathbb{R}^n$ when $n > 2$, any curve Γ joining the points A, B would be optimal. We do not know how to formulate in a natural way the optimal cutting problem (1.1) in higher dimension in order to obtain connected sets Γ with positive capacity.

Another possible variant to investigate could be to consider the optimal cutting problem (1.1) for the p-Laplacian:

$$\mathcal{E}_p(\Gamma) = \min\left\{\int_{\Omega\setminus\Gamma} |\nabla u|^p dx : u = u_0 \text{ on } \partial\Omega\right\}.$$

It is known (see [6]) that if $p > n-1$, curves have positive p-capacity and classes of compact connected sets Γ, like $\mathcal{U}_{ad}$, are closed for a suitable γ_p-convergence. Nevertheless, the duality argument of [8], used in Theorem 2.2, does not work for p-harmonic functions, which leaves the problem open.

We finally present some open problems which seem to be of interest.

Problem 1. The first question is concerned with the regularity of optimal cuts Γ. As shown at the beginning of this section, the optimal cut is not unique. The question is to prove that, among all the minimizers, there exists at least one "smooth" one. This might be of interest even in the simple case of a rectangle and two points $\{A, B\}$ with u_0 smooth and no force on the membrane.

Since the regularity of the optimal cut seems rather difficult to prove, a first question would be to identify its Hausdorff dimension, in particular to prove (or disprove) that its Lebesgue measure is zero.

The following simple example shows that in general we should not expect C^1 solutions Γ, even if $f \equiv 0$ and K reduces to two points. Indeed, take Ω the unit ball and the boundary datum $u_0(x,y) = 2xy$; according to the argument at the beginning of this section, if we consider two points A and B on a level line of the harmonic conjugate function $u + 0^*(x,y) = x^2 - y^2$, then this level line is an optimal cut for problem (1.1). Now, if we take $A = (-\frac{1}{2}, \frac{1}{2})$ and $B = (\frac{1}{2}, \frac{1}{2})$, the optimal cut Γ is the Cartesian curve $y = |x|$ with $x \in [-\frac{1}{2}, \frac{1}{2}]$, which has a singular point at the origin.

Problem 2. Instead of an optimal cut in a membrane, we may study the existence of an optimal cut in a plate. Given a smooth bounded set $\Omega \subseteq \mathbb{R}^2$ and a constant $\nu \in (0,1)$, for every compact connected set $\Gamma \subseteq \Omega$ we define

$$\mathcal{E}(\Gamma) = \min \left\{ \int_{\Omega\setminus\Gamma} \left[\nu(\Delta u)^2 + (1-\nu) \sum_{1\le i,j\le 2} (\partial_i\partial_j u)^2 \right] dx \right.$$
$$\left. : u \in H^2_{loc}(\Omega\setminus\Gamma), u = u_0 \text{ on } \partial\Omega\setminus\Gamma, \frac{\partial u}{\partial n} = u_1 \text{ on } \partial\Omega\setminus\Gamma \right\},$$

where u_0 and u_1 are both prescribed. Given two points $A, B \in \Omega$, the question is to prove the existence of a solution for the problem

$$\max_{\Gamma\in\mathcal{U}_{ad}} \mathcal{E}(\Gamma)$$

in the same frame of Theorem 2.2. The difficulty comes from the fact that a result similar to the one of Proposition 2.4 should be proved in H^2-spaces, and this, to our knowledge, is an open problem.

REFERENCES

[1] G. Allaire, E. Bonnetier, G. Francfort, and F. Jouve, *Shape optimization by the homogenization method*, Numer. Math., 76 (1997), pp. 27–68.

[2] L. Ambrosio, *Existence theory for a new class of variational problems*, Arch. Ration. Mech. Anal., 111 (1990), pp. 291–322.

[3] H. Attouch, *Variational Convergence for Functions and Operators*, Pitman, Boston, 1984.

[4] A. Baernstein, *On the harmonic measure of slit domains*, Complex Variables Theory Appl., 9 (1987), pp. 131–142.

[5] A. Baernstein, *Dubinin's symmetrization theorem*, in Complex Analysis, I (College Park, MD, 1985–86), Lecture Notes in Math. 1275, Springer-Verlag, Berlin, 1987, pp. 23–30.

[6] D. Bucur and P. Trebeschi, *Shape optimization problems governed by nonlinear state equation*, Proc. Roy. Soc. Edinburgh Sect. A, 128 (1998), pp. 945–963.
[7] D. Bucur and N. Varchon, *Boundary variation for the Neumann problem*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 24 (2000), pp. 807–821.
[8] D. Bucur and N. Varchon, *A Duality Approach for the Boundary Variations of Neumann Problems*, preprint 00/14, Université de Franche-Comté, Besançon, France, 2000.
[9] D. Bucur and N. Varchon, *Stabilité de la solution d'un problème de Neumann pour des variations de frontière*, C. R. Acad. Sci. Paris Ser. I Math., 331 (2000), pp. 371–374.
[10] A. Chambolle, *A Density Result in Two-Dimensional Linearized Elasticity and Applications*, preprint, Ceremade, Paris, 2001.
[11] A. Chambolle and F. Doveri, *Continuity of Neumann linear elliptic problems on varying two-dimensional bounded open sets*, Comm. Partial Differential Equations, 22 (1997), pp. 811–840.
[12] G. Dal Maso and R. Toader, *A Model for the Quasi-Static Growth of a Brittle Fracture: Existence and Approximation Results*, preprint, SISSA, Trieste, Italy, 2001.
[13] V. N. Dubinin, *Change of harmonic measure in symmetrization*, Mat. Sb. (N.S.), 124 (1984), pp. 272–279 (in Russian).
[14] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*, Stud. Adv. Math., CRC Press, Boca Raton, FL, 1992.
[15] J. Heinonen, T. Kilpelainen, and O. Martio, *Nonlinear Potential Theory of Degenerate Elliptic Equations*, Clarendon Press, Oxford, New York, Tokyo, 1993.
[16] J. Hersch, *The method of interior parallels applied to polygonal or multiply connected membranes*, Pacific J. Math., 13 (1963), pp. 1229–1238.
[17] V. Girault and P. A. Raviart, *Finite Element Methods for Navier-Stokes Equations*, Springer-Verlag, Berlin, 1986.
[18] R. V. Gol'dshteĭn, A. V. Marchenko, and A. Yu. Semënov, *Boundary waves in a fluid under an elastic membrane with a crack*, Dokl. Akad. Nauk, 339 (1994), pp. 331–334 (in Russian); Phys. Dokl., 39 (1994), pp. 813–815 (in English).
[19] R. Lipton, *The second Stekloff eigenvalue and energy dissipation inequalities for functionals with surface energy*, SIAM J. Math. Anal., 29 (1998), pp. 673–680.
[20] V. Sverak, *On optimal shape design*, J. Math. Pures Appl., 72 (1993), pp. 537–551.
[21] H. F. Weinberger, *An effectless cutting of a vibrating membrane*, Pacific J. Math., 13 (1963), pp. 1239–1240.

SIAM J. OPTIM.
Vol. 13, No. 1, pp. 168–178

A MULTIPLIER RULE FOR MULTIOBJECTIVE PROGRAMMING PROBLEMS WITH CONTINUOUS DATA*

DINH THE LUC†

Abstract. In this note we present a new multiplier rule for a constrained multiobjective programming problem with continuous data by using the concept of unbounded approximate Jacobians recently developed by Jeyakumar and Luc [*SIAM J. Control Optim.*, 36 (1998), pp. 1815–1832].

Key words. efficient solution, multiplier rule, approximate Jacobian, multiobjective programming

AMS subject classifications. 90C29, 90C30

PII. S1052623400378286

1. Introduction. Let $C \subset R^m$ be a convex and closed cone with apex at the origin and with $\text{int}C \neq \emptyset$, where $\text{int}C$ stands for the interior of the set C. We recall that C is said to be pointed if $C \cap (-C) = \{0\}$. The positive polar cone of C, which is denoted by C', is defined by $C' := \{\xi \in R^m : \langle \xi, c\rangle \geq 0, c \in C\}$.

For a and $b \in R^m$, we shall write $a \geq_C b$ iff $a - b \in C$, and $a \gg_C b$ iff $a - b \in \text{int}C$. The notation $a \leq_C b$ means $b \geq_C a$, and $a \ll_C b$ means $b \gg_C a$. Throughout this paper, C and K denote convex, closed, and pointed cones with apex at the origin and with nonempty interior in R^m and R^k, respectively. For a nonempty set $A \subseteq R^m$, the notation coA, $\overline{co}A$, and $\overline{A}$ will stand for its convex hull, closed convex hull, and its closure, respectively.

Let f, g, and h be vector functions from R^n to R^m, R^k, and R^l, respectively. Consider the following constrained multiobjective programming problem (VP):

$$\text{VMin } f(x),$$

$$g(x) \leq_K 0, \tag{1.1}$$

$$h(x) = 0, \tag{1.2}$$

which amounts to finding a point $x_0 \in R^m$ (called a weakly efficient solution) that satisfies the constraints (1.1) and (1.2) such that there is no $x \in R^n$ satisfying these constraints and $f(x) \ll_C f(x_0)$. If this is true for $x \in U$, where U is some neighborhood of x_0, then we call x_0 a local weakly efficient solution. We refer the interested reader to [4, 6, 13, 20] for a complete list of definitions of optimal solutions in multiobjective optimization and their relationships. To our knowledge, [12] is the first paper that deals with optimality conditions of multiobjective programming. Nowadays, there exists a very rich literature on this topic (see [3, 4, 6, 13, 14, 15, 20] and the references given therein). Most existing results concern the case in which the functions f, g, and h are differentiable [6, 12, 20] or locally Lipschitz [3]. The more general case of problems with inequality constraints when the data are not locally Lipschitz, or even set-valued, has been treated in [1, 13, 14, 16] and some others. The

*Received by the editors September 18, 2000; accepted for publication (in revised form) December 14, 2001; published electronically June 26, 2002.
http://www.siam.org/journals/siopt/13-1/37828.html

†Department of Mathematiques, Universite d'Avignon, 33 Rue Louis Pasteur, 8400 Avignon, France (dtluc@univ-avignon.fr). A portion of the research for this paper was completed at the Institute of Mathematics, Hanoi, Vietnam.

main tool used in the nonsmooth case is Clarke's generalized subdifferential for locally Lipschitz data and the contingent derivative or its modifications for set-valued data.

On the other hand, quite recently the authors of [7] have introduced the notion of an approximate Jacobian, which is a kind of generalized derivative for continuous vector functions, and applied it to the study of nonsmooth problems. It turns out from a series of papers [7, 8, 9, 10, 11, 18] that an approximate Jacobian provides a very useful tool to treat problems involving continuous, not necessarily locally Lipschitz, functions. As we shall see in the next section, the approximate Jacobian is defined in a flexible way by using a directional Dini derivative, so that several calculus rules of differentiable functions can be extended to continuous functions, including elementary rules (sum, product, composition, ...), the mean value theorem, open mapping theorem, etc. More importantly, many known generalized derivatives of vector functions such as Clarke's generalized Jacobian [2], Ioffe's prederivative (when it is given by linear operators) [5], Mordukhovich's coderivative [17], Warga's unbounded derivate containers [19], and others are examples of approximate Jacobians. Therefore, results expressed in terms of approximate Jacobians are also true when applied to the generalized derivatives above. As was noted in [7], a locally Lipschitz function may admit an approximate Jacobian whose closed convex hull is strictly contained in Clarke's generalized Jacobian or in Mordukhovich's coderivative. Therefore, even for locally Lipschitz problems, optimality conditions obtained by using approximate Jacobians sometimes yield sharp results. (See Example 3.1 of [18] for this situation.)

The purpose of this note is to use an approximate Jacobian to derive a multiplier rule for problem (VP) when the data f, g, and h are continuous, not necessarily locally Lipschitz. The paper is organized as follows. In the next section we recall the definition of an approximate Jacobian and the elementary calculus rules that will be needed in what follows. In section 3 we prove the main result of the paper about the existence of multipliers for local weakly efficient solutions. This result is concretized to the class of problems with Gâteaux differentiable data. An example is given in the final section to illustrate our approach.

2. Approximate Jacobians. Let f be a continuous vector function from R^n to R^m. A closed set of $(m\times n)$-matrices $\partial f(x) \subseteq L(R^n, R^m)$ is said to be an approximate Jacobian of f at x if for every $u \in R^n$ and $v \in R^m$ one has

$$(vf)^+(x,u) \le \sup_{M\in\partial f(x)} \langle v, M(u)\rangle,$$

where vf is the real function $\sum_{i=1}^n v_i f_i$. Here $v_1, \dots, v_m$ are components of v, $f_1, \dots, f_m$ are components of f, and $(vf)^+(x,u)$ is the upper Dini directional derivative of the function vf at x in the direction u; that is,

$$(vf)^+(x,u) := \limsup_{t\downarrow 0} \frac{(vf)(x+tu) - (vf)(x)}{t}.$$

If, for every $x \in R^n$, $\partial f(x)$ is an approximate Jacobian of f at x, then the set-valued map $\partial f : R^n \Rightarrow L(R^n, R^m)$ is called an approximate Jacobian map of f. When $m = 1$, the approximate Jacobian is also called the generalized subdifferential.

It follows from the definition that if a function is Gâteaux differentiable at a point, then its Gâteaux derivative is an approximate Jacobian, and any other approximate Jacobian contains it in its convex hull. Conversely, if a function admits a singleton approximate Jacobian, then the function is Gâteaux differentiable, and its Gâteaux

derivative coincides with that singleton element. Approximate Jacobians were introduced and studied in [7]. Further developments and applications of this concept were given in [7, 8, 9, 10, 11]; we refer the interested reader to those papers for more details on the approximate Jacobian. For our purposes we shall need the following elementary calculus rules of approximate Jacobians, which were already established in [7, 8, 9, 10, 11]:

(a) Suppose that $f : R^n \to R$ is continuous. If f admits an approximate Jacobian $\partial f(x)$ at x and attains its minimum at x, then $0 \in \overline{co}\partial f(x)$.

(b) Suppose that f_1 and $f_2 : R^n \to R^m$ are continuous. If $\partial f_1(x)$ and $\partial f_2(x)$ are approximate Jacobians of f_1 and f_2, respectively, at x, then $\partial f_1(x) + \partial f_2(x)$ is an approximate Jacobian of $f_1 + f_2$ at x.

(c) Suppose that $f_1 : R^n \to R^m$ and $f_2 : R^n \to R^l$ are continuous. If $\partial f_1(x)$ and $\partial f_2(x)$ are approximate Jacobians of f_1 and f_2 at x, then the set $(\partial f_1(x), \partial f_2(x))$ is an approximate Jacobian of the function $(f_1, f_2) : R^n \to R^m \times R^l$ at x.

(d) *Mean value theorem*: Let f be a continuous function from R^n to R^m. Let $a, b \in R^n$ and let $\partial f(x)$ be an approximate Jacobian of f at $x \in [a, b]$. Then $f(b) - f(a) \in \overline{co}(\partial f[a, b](b - a))$.

Some more terminologies are in order. Let $A \subset R^n$ be a nonempty set. The recession cone of A, denoted by A_∞, consists of all limits $\lim_{i\to\infty} t_i a_i$, where $a_i \in A$ and $\{t_i\}$ is a sequence of positive numbers converging to 0. It is important to notice that when A is closed and convex, $A + A_\infty = A$. Let $F : R^n \Rightarrow R^m$ be a set-valued map. It is said to be upper semicontinuous at x_0 if for every $\epsilon > 0$ there is some $\delta > 0$ such that $F(x_0 + \delta B_n) \subset F(x_0) + \epsilon B_m$, where B_n and B_m denote the closed unit balls in R^n and R^m, respectively.

3. A multiplier rule. Let us consider problem (VP) described in the introduction. We define $H := (f, g, h)$, a continuous function from R^n to $R^m \times R^k \times R^l$. The product space $R^m \times R^k \times R^l$ is equipped with the Euclidean norm. The space $L(R^n, R^m \times R^k \times R^l)$ is equipped with the norm of linear operators; i.e., for an $(m + k + l) \times n$-matrix M,

$$\|M\| = \max_{x\in R^n, \|x\|\le 1} \|M(x)\|.$$

The closed unit ball of this space is denoted by B. We also denote by T the set of all vectors $\lambda \in (C, K, \{0\})'$ with $\|\lambda\| = 1$. The following lemma will be needed.

LEMMA 3.1. *Let $\omega_0 \in R^m \times R^k \times R^l$ be a nonzero vector with*

$$\max_{\lambda\in T} \langle \lambda, \omega_0 \rangle > 0.$$

Then there exists a unique point $\lambda_0 \in T$ such that

$$\langle \lambda_0, \omega_0 \rangle = \max_{\lambda\in T} \langle \lambda, \omega_0 \rangle.$$

Moreover, for every $\epsilon > 0$, there is some $\delta > 0$ such that

$$\max_{\lambda\in T} \langle \lambda, \omega \rangle = \max_{\lambda\in T, \|\lambda-\lambda_0\|\le\epsilon} \langle \lambda, \omega \rangle$$

for all ω with $\|\omega - \omega_0\| \le \delta$.

Proof. That the function $\langle \lambda, \omega_0 \rangle$ attains its maximum on T is obvious because T is compact. Suppose to the contrary that there are two distinct points λ_0 and λ_1

that maximize this function on T. It follows from the hypothesis that $\lambda_1 \neq -\lambda_0$. Let $\lambda_2 := (\lambda_0 + \lambda_1)/\|\lambda_0 + \lambda_1\|$. Then $\lambda_2 \in T$ and

$$\langle \lambda_2, \omega_0 \rangle = \frac{2}{\|\lambda_0 + \lambda_1\|} \langle \lambda_0, \omega_0 \rangle.$$

The Euclidean norm being strictly convex, we have

$$\|\lambda_0 + \lambda_1\| < \|\lambda_0\| + \|\lambda_1\| = 2,$$

which yields a contradiction:

$$\langle \lambda_2, \omega_0 \rangle > \langle \lambda_0, \omega_0 \rangle.$$

To prove the second part, suppose to the contrary that there is some $\epsilon_0 > 0$ such that for each $\delta = 1/i, i \geq 1$, one can find a vector ω_i, with $\|\omega_i - \omega_0\| \leq 1/i$, verifying

$$\max_{\lambda \in T} \langle \lambda, \omega_i \rangle \neq \max_{\lambda \in T, \|\lambda - \lambda_0\| \leq \epsilon} \langle \lambda, \omega_i \rangle.$$

Let $\lambda_i \in T$ be a maximizing point of the function $\langle \lambda, \omega_i \rangle$ on T. Then $\|\lambda_i - \lambda_0\| > \epsilon_0$. We may assume that the sequence $\{\lambda_i\}$ converges to some $\lambda_* \in T$. It follows, on one hand, that $\|\lambda_* - \lambda_0\| \geq \epsilon_0$. On the other hand, as T is compact, one has

$$\langle \lambda_*, \omega_0 \rangle = \lim_{i \to 0} \langle \lambda_i, \omega_i \rangle = \max_{\lambda \in T} \langle \lambda, \omega_0 \rangle,$$

which shows that λ_* is a maximizing point of the function $\langle \lambda, \omega_0 \rangle$ on T. This contradicts the uniqueness of λ_0 by the first part. The proof is complete. □

We now formulate and prove the main result of the paper.

THEOREM 3.2. *Assume that ∂H is an approximate Jacobian map of H which is upper semicontinuous at x_0. If x_0 is a local weakly efficient solution of* (VP)*, then there is a vector $\lambda_0 = (\xi_0, \theta_0, \gamma_0) \in T$ such that*

$$0 \in \lambda_0(\overline{co}\partial H(x_0) \cup co[(\partial H(x_0))_\infty \setminus \{0\}]),$$
$$\theta_0 g(x_0) = 0.$$

Proof. Let us choose a vector $e \in \text{int}C$ so that

$$\max_{\xi \in C', \|\xi\| \leq 1} \langle \xi, e \rangle = 1.$$

For each $\epsilon > 0$, define functions $H_\epsilon : R^n \to R^m \times R^k \times R^l$ and $P_\epsilon : R^n \to R$ as follows:

$$H_\epsilon(x) := (f(x) - f(x_0) + \epsilon e, g(x), h(x)),$$
$$P_\epsilon(x) := \max_{\lambda \in T} \langle \lambda, H_\epsilon(x) \rangle.$$

It is clear that these functions are continuous. Let $U \subset R^n$ be a neighborhood that exists by the definition of the local weakly efficient solution x_0. We claim that

$$P_\epsilon(x) > 0 \quad \text{for all } x \in U.$$

Indeed, suppose that there is some $x \in U$ such that $P_\epsilon(x) \leq 0$. Setting $\lambda = (0, 0, \beta) \neq 0$, we obtain $\beta h(x) \leq 0$ for all $\beta \in R^l \setminus \{0\}$ and hence $h(x) = 0$. Taking $\lambda =$

$(0,\gamma,0), \gamma \in K' \setminus \{0\}$, we obtain $\gamma(g(x)) \leq 0$ for all $\gamma \in K' \setminus \{0\}$, which implies $g(x) \in -K$. By a similar argument, choosing $\lambda = (\xi, 0, 0)$, we have $\xi(f(x) - f(x_0) + \epsilon e) \leq 0$ for all $\xi \in C' \setminus \{0\}$. Since $e \in \text{int}C$, we derive $f(x) - f(x_0) \in \text{int}C$. This contradicts the fact that x_0 is a local weakly efficient solution of (VP).

Furthermore, since $P_\epsilon(x_0) = \epsilon < \inf P_\epsilon + \epsilon$, by Ekeland's variational principle, there is an x_ϵ such that $\|x_0 - x_\epsilon\| < \sqrt{\epsilon}$ and

$$P_\epsilon(x_\epsilon) < P_\epsilon(x) + \sqrt{\epsilon}\|x - x_\epsilon\| \quad \text{for all } x \neq x_\epsilon.$$

In particular, the net $\{x_\epsilon\}$ converges to x_0 as ϵ tends to 0, and x_ϵ provides a minimum of the function

$$Q_\epsilon(x) := P_\epsilon(x) + \sqrt{\epsilon}\|x - x_\epsilon\|.$$

According to rule (a), if $\partial Q_\epsilon(x_\epsilon)$ is an approximate Jacobian of Q_ϵ at x_ϵ, then

$$0 \in \overline{co}\partial Q_\epsilon(x_\epsilon). \tag{3.1}$$

Our aim at the moment is to find a suitable approximate Jacobian of Q_ϵ. This can be done if we are able to find a suitable approximate Jacobian $\partial P_\epsilon(x_\epsilon)$ of P_ϵ, because the set $\sqrt{\epsilon}B_n$ is an approximate Jacobian of the function $x \mapsto \|x - x_\epsilon\|$ at x_ϵ. By rule (b), the set $\partial P_\epsilon(x_\epsilon) + \sqrt{\epsilon}B_n$ is an approximate Jacobian of Q_ϵ at x_ϵ. We observe that $\partial H(x_\epsilon)$ is an approximate Jacobian of H_ϵ at x_ϵ, because the function H_ϵ is the sum of H and of the constant function $x \mapsto (-f(x_0) + \epsilon e, 0, 0)$.

Moreover, for $\epsilon > 0$, let λ_ϵ be the unique vector that maximizes the function $\langle \lambda, H_\epsilon(x_\epsilon)\rangle$ on T (by Lemma 3.1). We claim that for each integer $r \geq 1$ there is some $\epsilon(r) > 0$ such that for every $\epsilon \in (0, \epsilon(r)]$ the set

$$L_\epsilon := \left\{ \lambda \left(M + \frac{1}{r}N \right) : \lambda \in T, \|\lambda - \lambda_\epsilon\| \leq \epsilon, M \in \partial H(x_0), N \in B \right\}$$

is an approximate Jacobian of P_ϵ at x_ϵ. Indeed, let $\delta > 0$ be a positive number that exists by virtue of Lemma 3.1. Since H_ϵ is continuous, there is some $t_0 > 0$ such that

$$\|H_\epsilon(x_\epsilon) - H_\epsilon(x)\| < \delta \quad \text{for all } x \in U \text{ with } \|x - x_\epsilon\| \leq t_0.$$

For every $u \in R^n$, we deduce from Lemma 3.1 that

$$\begin{aligned} P_\epsilon(x_\epsilon + tu) - P_\epsilon(x_\epsilon) &= \max_{\lambda \in T} \langle \lambda, H_\epsilon(x_\epsilon + tu)\rangle - \max_{\lambda \in T} \langle \lambda, H_\epsilon(x_\epsilon)\rangle \\ &= \max_{\lambda \in T, \|\lambda - \lambda_\epsilon\| \leq \epsilon} \langle \lambda, H_\epsilon(x_\epsilon + tu)\rangle - \max_{\lambda \in T, \|\lambda - \lambda_\epsilon\| \leq \epsilon} \langle \lambda, H_\epsilon(x_\epsilon)\rangle \\ &\leq \max_{\lambda \in T, \|\lambda - \lambda_\epsilon\| \leq \epsilon} \langle \lambda, H_\epsilon(x_\epsilon + tu) - H_\epsilon(x_\epsilon)\rangle \end{aligned}$$

for every $t \geq 0$ with $\|tu\| \leq t_0$. Applying the mean value theorem, we find for each such t a matrix $M_t \in \overline{co}\partial H[x_\epsilon, x_\epsilon + tu] + (1/2r)B$ such that

$$H_\epsilon(x_\epsilon + tu) - H_\epsilon(x_\epsilon) = M_t(tu).$$

Since ∂H is upper semicontinuous at x_0 and $\lim_{\epsilon \to 0} x_\epsilon = x_0$, for each $r \geq 1$ there is some $\epsilon(r) > 0$ such that for every $\epsilon \in (0, \epsilon(r)]$ one has

$$\overline{co}\partial H[x_\epsilon, x_\epsilon + tu] \subset \overline{co}\partial H(x_0) + \frac{1}{2r}B$$

for t sufficiently small. It follows that

$$\begin{aligned} P_\epsilon^+(x_\epsilon, u) &\le \limsup_{t\downarrow 0} \max_{\lambda\in T, \|\lambda-\lambda_\epsilon\|\le\epsilon} \langle \lambda, M_t(u)\rangle \\ &\le \sup_{M\in\overline{co}\partial H(x_0), N\in B, \lambda\in T, \|\lambda-\lambda_\epsilon\|\le\epsilon} \left\langle \lambda, \left(M + \frac{1}{r}N\right)(u)\right\rangle \\ &\le \sup_{\xi\in L_\epsilon} \langle \xi, u\rangle. \end{aligned}$$

Similarly,

$$(-P_\epsilon)^+(x_\epsilon, u) \le \sup_{\xi\in L_\epsilon} (-\langle \xi, u\rangle).$$

Consequently, L_ϵ is an approximate Jacobian of P_ϵ at x_ϵ. Summing up the above, we conclude that for each $r \ge 1$ there is an $\epsilon(r) > 0$ such that for $0 < \epsilon \le \epsilon(r)$ the set

$$\partial Q_\epsilon(x_\epsilon) := L_\epsilon + \sqrt{\epsilon} B_n$$

is an approximate Jacobian of Q_ϵ at x_ϵ. We may choose $\epsilon(r) \downarrow 0$ as $r \to \infty$. Relation (3.1) becomes

$$\begin{aligned} 0 \in \overline{co}\partial Q_\epsilon(x_\epsilon) &\subset \overline{co} L_\epsilon + \sqrt{\epsilon} B_n \\ &\subset co\{\lambda M : \lambda \in T, \|\lambda - \lambda_\epsilon\| \le \epsilon, M \in \partial H(x_0)\} \\ &\quad + co\left\{\frac{1}{r}\lambda N : \lambda \in T, \|\lambda - \lambda_\epsilon\| \le \epsilon, N \in B\right\} + 2\sqrt{\epsilon} B_n. \end{aligned}$$

Taking into account the fact that B, B_n, and T are all compacts, we derive the existence of vectors

$$\xi_r \in co\{\lambda M : \lambda \in T, \|\lambda - \lambda_\epsilon(r)\| \le \epsilon(r), M \in \partial H(x_0)\}$$

such that

$$\lim_{r\to\infty} \xi_r = 0.$$

We apply Caratheodory's theorem to express the vectors ξ_r as

$$\xi_r = \sum_{j=1}^{n+1} a_{rj}\lambda_{rj} M_{rj},$$

where $\sum_{j=1}^{n+1} a_{rj} = 1$, $a_{rj} \ge 0$, $\lambda_{rj} \in T$ with $\|\lambda_{rj} - \lambda_{\epsilon(r)}\| \le \epsilon(r)$, and $M_{rj} \in \partial H(x_0)$, $j = 1, \ldots, n+1$.

Since T is compact, without loss of generality we may assume that the sequence $\{\lambda_{\epsilon(r)}\}$ converges to some $\lambda_0 \in T$. Then

$$\lim_{r\to\infty} \lambda_{rj} = \lambda_0 \quad \text{for all } j = 1, \ldots, n+1.$$

Moreover, by taking a subsequence if necessary, we also may assume that the sequences $\{a_{rj}\}_r$ converge to $a_{0j}, j = 1, \ldots, n+1$, and that

$$\xi_r = \sum_{j\in I_1} a_{rj}\lambda_{rj} M_{rj} + \sum_{j\in I_2} a_{rj}\lambda_{rj} M_{rj} + \sum_{j\in I_3} a_{rj}\lambda_{rj} M_{rj},$$

where the above sums have the following properties:

(1) for each $j \in I_1$, the sequence $\{M_{rj}\}_r$ is bounded and converges to some $M_{0j} \in \partial H(x_0)$;
(2) for each $j \in I_2$, the sequence $\{M_{rj}\}_r$ is unbounded, but the sequence $\{a_{rj}M_{rj}\}_r$ is bounded and converges to some M_{*j};
(3) for each $j \in I_3$, the sequence $\{a_{rj}M_{rj}\}_r$ is unbounded, and there is some $j_0 \in I_3$ such that the sequences $\{a_{rj}M_{rj}/\|a_{rj_0}M_{rj_0}\|\}_r$ converge to some $M_{\infty j}$, $j \in I_3$.

Let us first consider the case in which I_3 is nonempty. By dividing ξ_r by $\|a_{rj_0}M_{rj_0}\|$ and passing to the limit when r tends to ∞, we obtain

$$0 = \lim_{r\to\infty} \frac{\xi_r}{\|a_{rj_0}M_{rj_0}\|} = \lim_{r\to\infty} \sum_{j\in I_3} \lambda_{rj} \frac{a_{rj}M_{rj}}{\|a_{rj_0}M_{rj_0}\|} = \lambda_0 \sum_{j\in I_3} M_{\infty j}.$$

In the latter sum, we have $M_{\infty j} \in [\partial H(x_0)]_\infty$ and $M_{\infty j_0} \neq 0$. Hence

$$0 \in \lambda_0 co([\partial H(x_0)]_\infty \setminus \{0\}). \tag{3.2}$$

It remains to consider the case in which I_3 is empty. For $j \in I_2$, one has $a_{0j} = 0$, which implies that $\sum_{j\in I_1} a_{0j} = 1$ and $M_{*j} \in [\partial H(x_0)]_\infty$. Thus,

$$\begin{aligned} 0 &= \lim_{r\to\infty} \xi_r \\ &= \lambda_0 \left(\sum_{i\in I_1} a_{0j}M_{0j} + \sum_{j\in I_2} M_{*j} \right) \in \lambda_0(co[\partial H(x_0)] + co[(\partial H(x_0))_\infty]) \subset \lambda_0 \overline{co}\partial H(x_0). \end{aligned}$$

This and (3.2) establish the multiplier rule. As to the complementary slackness $\theta_0 g(x_0) = 0$, we observe that if $g_i(x_0) < 0$, then the vector λ_ϵ must have the corresponding component $\theta_{\epsilon i} = 0$, and when passing to limit we obtain $\theta_{0i} = 0$ as requested.

The following modified version of Theorem 3.2 is useful in those situations in which some of the components of the data admit bounded approximate Jacobians.

THEOREM 3.3. *Assume that $H = (H_1, H_2)$ and ∂H_i, $i = 1, 2$, are approximate Jacobian maps of H that are upper semicontinuous at x_0. If x_0 is a local weakly efficient solution of* (VP)*, then there is a vector $\lambda_0 = (\xi_0, \theta_0, \gamma_0) \in T$ such that $\theta_0 g(x_0) = 0$ and*

$$0 \in \lambda_0(\overline{co}\partial H_1(x_0) \cup co[(\partial H_1(x_0))_\infty \setminus \{0\}], \overline{co}\partial H_2(x_0) \cup co[(\partial H_2(x_0))_\infty \setminus \{0\}]).$$

Proof. Use rule (c) and the proof of Theorem 3.2. □

We notice that when the data of the problem are locally Lipschitz, one can use Clarke's generalized Jacobian as an approximate Jacobian. In this case the recession part of the multiplier rule disappears, and Theorem 3.2 gives the multiplier rule of [3]. In a private communication, N. V. Hung of the Hanoi Institute of Mathematics has observed a similar result for the case in which H admits a bounded approximate Jacobian. His result, however, is limited to the case of locally Lipschitz functions, because a function that has a bounded upper semicontinuous approximate Jacobian map is locally Lipschitz (see [11]). The recession part in the conclusion of Theorem 3.2 is a very characteristic feature of those problems that have continuous, but not locally Lipschitz continuous, data. (See the example in the next section.)

Let us now apply Theorem 3.2 to a particular problem in which the data are Gâteaux differentiable but not necessarily locally Lipschitz. To this purpose let us define for a Gâteaux differentiable function $\phi : R^n \to R^m$ the following sets:

$$\tilde{\nabla}\phi(x) = \{\lim \nabla\phi(x_i) : x_i \to x\},$$
$$\nabla^\infty\phi(x) = \{\lim t_i\nabla\phi(x_i) : x_i \to x, t_i \downarrow 0\}.$$

Actually $\tilde{\nabla}\phi(x)$ is the upper limit of the sets $\{\nabla\phi(x')\}$ when $x' \to x$ in the sense of Kuratowski–Painleve, and $\nabla^\infty\phi(x)$ is the outer horizon limit of $\{\nabla\phi(x')\}$ when $x' \to x$. It follows that $\tilde{\nabla}\phi(x)$ is a closed set, and $\nabla^\infty\phi(x)$ is a nonempty closed cone. When ϕ has a locally bounded derivative around x, one has $\nabla^\infty\phi(x) = \{0\}$, and $\tilde{\nabla}\phi(x)$ is a compact set. This is the case when ϕ is locally Lipschitz. When $m = 1$ and ϕ is locally Lipschitz, the set $\tilde{\nabla}\phi(x)$ is also called the B-subdifferential of f at x, and $co\tilde{\nabla}\phi(x)$ is exactly the Clarke generalized subdifferential.

COROLLARY 3.4. *Assume that x_0 is a local weakly efficient solution of* (VP) *and the functions f, g, and h are Gâteaux differentiable in a neighborhood of x_0. Then there exists a vector $\lambda_0 = (\xi_0, \theta_0, \gamma_0) \in T$ such that $\theta_0 g(x_0) = 0$ and*

$$0 \in \lambda_0(\overline{co}\tilde{\nabla}H(x_0) \cup co[\nabla^\infty H(x_0) \setminus \{0\}]).$$

Proof. We may assume without loss of generality that $H = (f, g, h)$ is differentiable at every $x \in R^n$ with $\|x - x_0\| \le 1$. For every $k \ge 1$ let us construct an approximate Jacobian of H as follows:

$$\partial H(x) = \begin{cases} L(R^n, R^m) & \text{if} \quad \|x - x_0\| \ge 1/k, \\ \{\nabla H(x)\} & \text{if} \quad 0 < \|x - x_0\| < 1/k, \\ \overline{\{\nabla H(x') : \|x - x_0\| < 1/k\}} & \text{if} \quad x = x_0. \end{cases}$$

It is evident that the set-valued map $x \mapsto \partial H(x)$ is an approximate Jacobian map of H which is upper semicontinuous at x_0. According to Theorem 3.2, there is a vector $\lambda_k = (\xi_k, \theta_k, \gamma_k) \in T$ such that

$$0 \in \lambda_k(\overline{co}\partial H(x_0) \cup co[(\partial H(x_0))_\infty \setminus \{0\}]),$$
$$\theta_k g(x_0) = 0.$$

By taking a subsequence if necessary, we need consider only two cases:

(A) There exist $\alpha_{kj} \ge 0, x_{kj} \in R^n, j = 1, \ldots, mn+1$, and $m \times n$-matrices b_k with

$$\sum_{j=1}^{mn+1} \alpha_{kj} = 1, \quad \|x_{kj} - x_0\| < \frac{1}{k}, \quad j = 1, \ldots, mn+1, \quad \|b_k\| \le 1$$

such that

$$0 = \lambda_k \left\{ \sum_{j=1}^{mn+1} \alpha_{kj}\nabla H(x_{kj}) + \left(\frac{1}{k}\right) b_k \right\};$$

(B) There exist $\alpha_{kj} \ge 0, \beta_{kj} \ge 0, x_{kj} \in R^n, j = 1, \ldots, mn+1$, and $m \times n$-matrices b_k with

$$\sum_{j=1}^{mn+1} \alpha_{kj} = 1, \quad \|x_{kj} - x_0\| < \frac{1}{k}, \quad \|\nabla H(x_{kj}\| \ge k, \quad j = 1, \ldots, mn+1, \quad \|b_k\| \le 1$$

such that

$$0 = \lambda_k \left\{ \sum_{j=1}^{mn+1} \alpha_{kj}\beta_{kj}\nabla H(x_{kj}) + \left(\frac{1}{k}\right) b_k \right\}.$$

We may assume that $\{\lambda_k\}$ converges to some $\lambda_0 \in T$ because T is compact. By using an argument similar to that of the proof of Theorem 3.2, we derive from (A) that either

$$0 \in \lambda_0 \overline{co}\tilde{\nabla} H(x_0) \quad \text{or} \quad 0 \in \lambda_0 co[\nabla^\infty H(x) \setminus \{0\}],$$

and from (B) that

$$0 \in \lambda_0 co[\nabla^\infty H(x) \setminus \{0\}].$$

This completes the proof. □

We mention that Gâteaux differentiable functions are not necessarily locally Lipschitz or continuously differentiable. Therefore the existing multiplier rules of [3, 6] do not apply to problems with Gâteaux differentiable data. Moreover, the method of constructing approximate Jacobians that we have presented above can be extended to the class of almost everywhere Gâteaux differentiable functions, and a similar multiplier rule can be obtained for problems with data of this class. We leave this extension to the interested reader.

4. Example. Consider the following biobjective problem in R^5:

$$\begin{aligned} \text{VMin } & (-x_2 + x_3 + (x_5)^2, x_2 + (x_4)^2)x_5 \geq 0, \\ & (x_1)^{2/3} sgn(x_1) + (x_2)^4 - x_3 = 0, \\ & (x_1)^{1/3} + (x_2)^2 - x_4 = 0, \end{aligned}$$

where the ordering cone of R^2 is the positive orthant R^2_+. The function $H = (f, g, h)$, where $f(x) := (-x_2+x_3+(x_5)^2, x_2+(x_4)^2)$, $g(x) := x_5$, and $h(x) := ((x_1)^{2/3}sgn(x_1)+(x_2)^4-x_3, (x_1)^{1/3}+(x_2)^2-x_4)$, is not locally Lipschitz at $x = (x_1, \ldots, x_5)$ with $x_1 = 0$. It is not hard to see that the set

$$\partial H(x) := \left\{ \begin{pmatrix} 0 & -1 & 1 & 0 & 2x_5 \\ 0 & 1 & 0 & 2x_4 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ \frac{2}{3}(x_1)^{-1/3}sgn(x_1) & 4(x_2)^3 & -1 & 0 & 0 \\ \frac{1}{3}(x_1)^{-2/3} & 2x_2 & 0 & -1 & 0 \end{pmatrix} \right\}$$

is an approximate Jacobian of H at $x = (x_1, \ldots, x_5)$ with $x_1 \neq 0$, and the set

$$\partial H(x) := \left\{ \begin{pmatrix} 0 & -1 & 1 & 0 & 2x_5 \\ 0 & 1 & 0 & 2x_4 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ \alpha & 4(x_2)^3 & -1 & 0 & 0 \\ \alpha^2 & 2x_2 & 0 & -1 & 0 \end{pmatrix} : \alpha \geq 1 \right\}$$

is an approximate Jacobian of H at x with $x_1 = 0$. Moreover, the set-valued map $x \mapsto \partial H(x)$ is upper semicontinuous.

Let us first consider $x \in R^5$ with $x_1 \neq 0$. Observe that H is continuously differentiable with $\partial H(x) = \{\nabla H(x)\}$, and the multiplier rule is written as

$$0 = \lambda_0 \nabla H(x).$$

In particular, we derive the following equation that a local weakly efficient solution must satisfy:

$$2(x_1)^{-1/3} sgn(x_1)(1 - 4x_2x_4) + (x_1)^{-2/3}(1 - 4(x_2)^3) = 0.$$

This result can evidently be obtained by the classical necessary optimality condition (see [6]), because the problem is continuously differentiable in a small neighborhood of x.

Now we consider the case in which $x \in R^5$ has $x_1 = 0$. Set $H_1 = (f, g)$ and $H_2 = h$. The function H_1 is continuously differentiable, and the map $x' \mapsto \{\nabla H_1(x')\}$ is an upper semicontinuous approximate Jacobian map of H_1. The function H_2 is neither differentiable nor locally Lipschitz at x. By defining

$$\partial H_2(x) := \left\{ \begin{pmatrix} \alpha & 4(x_2)^3 & -1 & 0 & 0 \\ \alpha^2 & 2x_2 & 0 & -1 & 0 \end{pmatrix} : \alpha \geq 1 \right\},$$

we see that the set-valued map $x' \mapsto \nabla H_2(x')$ for x' having the first component nonzero, and $x \mapsto \partial H_2(x)$ for the other x, is an upper semicontinuous approximate Jacobian map of H_2. The recession cone of $\partial H_2(x)$ is given by

$$(\partial H_2(x))_\infty = \left\{ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ \alpha & 0 & 0 & 0 & 0 \end{pmatrix} : \alpha \geq 0 \right\}.$$

According to Theorem 3.3, a local weakly efficient solution must satisfy either of the following conditions:

(i) $0 = (\xi_0, \theta_0)\nabla H_1(x)$ and $0 \in \gamma_0 \partial H_2(x)$;

(ii) $0 = (\xi_0, \theta_0)\nabla H_1(x)$ and $0 \in \gamma_0[(\partial H_2(x))_\infty \setminus \{0\}]$.

Let us look, for instance, at $x = 0$. Condition (i) implies $\xi_0 = (0, 0)$, $\theta_0 = 0$, and $\gamma_0 = (0, 0)$. In other words, at $x = 0$ there is no multiplier $\lambda_0 \in T$ that verifies this condition. However, the multiplier λ_0 with $\xi_0 = (0, 0)$, $\theta_0 = 0$, and $\gamma_0 = (1, 0)$ verifies (ii), which means that $x = 0$ is a candidate to be a local weakly efficient solution. Actually a scalarization technique [15] confirms that it is.

REFERENCES

[1] G. Y. Chen and J. Jahn, *Optimality conditions for set-valued optimization problems*, Math. Meth. Oper. Res., 48 (1998), pp. 187–200.

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

[3] B. D. Craven, *Nonsmooth multiobjective programming,* Numer. Funct. Anal. Optim., 10 (1989), pp. 49–64.

[4] F. Giannessi, G. Maestroeni, and L. Pellegrini, *On the theory of vector optimization and variational inequalities. Image space analysis and separation,* in Vector Variational Inequalities and Vector Equilibria, F. Giannessi, ed., Kluwer Academic Publishers, London, 2000, pp. 153–216.

[5] A. D. Ioffe, *Nonsmooth analysis: Differential calculus of non-differentiable mappings,* Trans. Amer. Math. Soc., 266 (1981), pp. 1–56.

[6] J. Jahn, *Mathematical Theory of Vector Optimization in Partially Ordered Spaces*, Peter Lang, Germany, 1985.

[7] V. Jeyakumar and D. T. Luc, *Approximate Jacobian matrices for nonsmooth continuous maps and C^1-optimization,* SIAM J. Control Optim., 36 (1998), pp. 1815–1832.

[8] V. Jeyakumar, D. T. Luc, and S. Schaible, *Characterizations of generalized monotone nonsmooth continuous maps using approximate Jacobians,* J. Convex Anal., 5 (1998), pp. 119–132.

[9] V. Jeyakumar and D. T. Luc, *Nonsmooth calculus, minimality and monotonicity of convexificators,* J. Optim. Theory Appl., 101 (1999), pp. 599–621.

[10] V. Jeyakumar and D. T. Luc, *Open mapping theorem using unbounded generalized Jacobians,* J. Math. Anal. Appl., 2002 (to appear).

[11] V. Jeyakumar, D. T. Luc, and Y. Wang, *Lagrange Multipliers for Equality Constraints without Lipschitz Continuity,* Applied Mathematics report AMR00/1, University of New South Wales, Sydney, Australia.

[12] H. W. Kuhn and F. H. Tucker, *Nonlinear programming*, in Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, 1951, pp. 481–492.

[13] D. T. Luc, *Theory of Vector Optimization*, Lecture Notes in Econom. and Math. Systems 319, Springer-Verlag, Germany, 1989.

[14] D. T. Luc, *Contingent derivative of set-valued maps and applications to vector optimization,* Math. Programming, 50 (1991), pp. 99–111.

[15] D. T. Luc, *Vector Optimization,* Lecture delivered at the summer school "Generalized convexity and monotonicity," Samos, Greece, 1999.

[16] D. T. Luc and C. Malivert, *Invex optimization problems,* Bull. Austral. Math. Soc., 46 (1992), pp. 47–66.

[17] B. S. Mordukhovich, *Generalized differential calculus for nonsmooth and set-valued mappings,* J. Math. Anal. Appl., 183 (1994), pp. 250–288.

[18] X. Wang and V. Jeyakumar, *A sharp Lagrange multiplier rule for nonsmooth mathematical programming problems involving equality constraints*, SIAM J. Optim., 10 (2000), pp. 1136–1148.

[19] J. Warga, *Fat homeomorphisms and unbounded derivate containers*, J. Math. Anal. Appl., 81 (1981), pp. 545–560.

[20] P. L. Yu, *Multi-criteria Decision Making: Concepts, Techniques and Extensions*, Plenum Press, New York, London, 1985.

SIAM J. OPTIM.
Vol. 13, No. 1, pp. 179–203

PRIMAL-DUAL INTERIOR-POINT METHODS FOR SECOND-ORDER CONIC OPTIMIZATION BASED ON SELF-REGULAR PROXIMITIES*

JIMING PENG†, CORNELIS ROOS‡, AND TAMÁS TERLAKY†

Abstract. Recently the authors introduced the notions of *self-regular* functions and *self-regular* proximity functions and used them in the design and analysis of interior-point methods (IPMs) for linear and semidefinite optimization (LO and SDO). In this paper, we consider an extension of these concepts to second-order conic optimization (SOCO). This nontrivial extension requires the development of various new tools. Versatile properties of general analytical functions associated with the second-order cone are exploited. Based on the so-called self-regular proximity functions, new primal-dual Newton methods for solving SOCO problems are proposed. It will be shown that these new large-update IPMs for SOCO enjoy polynomial $\mathcal{O}(\max\{p,q\}N^{(q+1)/2q}\log\frac{N}{\varepsilon})$ iteration bounds analogous to those of their LO and SDO cousins, where N is the number of constraining cones and p,q are constants, the so-called growth degree and barrier degree of the corresponding proximity. Our analysis allows us to choose not only a constant q but also a q as large as $\log N$. In this case, our new algorithm has the best known $\mathcal{O}(N^{1/2}\log N\log\frac{N}{\varepsilon})$ iteration bound for large-update IPMs.

Key words. second-order conic optimization, primal-dual interior-point method, self-regular proximity function, polynomial complexity

AMS subject classification. 90C05

PII. S1052623401383236

1. Introduction.

Mathematically, a typical second-order cone can be defined by

$$K = \left\{ (x_1, x_2, \ldots, x_n) \in \Re^n : x_1^2 - \sum_{i=2}^{n} x_i^2 \geq 0, x_1 \geq 0 \right\}.$$

Second-order conic optimization (SOCO) is the problem of minimizing a linear objective function subject to the intersection of an affine set and the direct product of several second-order cones. SOCO can be viewed as a direct generalization of linear optimization (LO). Several important types of problems can be modelled as SOCO problems. For example, a general convex quadratic optimization problem with convex quadratic constraints can be cast as a SOCO problem [14]. SOCO also includes robust LO, robust least-squares, matrix-fractional problems, and problems involving sums and maxima of norms, etc., as specific cases. For more details about various applications of SOCO, we refer to the survey paper [12] and the references therein.

An alternative way to describe the second-order cone is via matrix representation.

*Received by the editors January 5, 2001; accepted for publication (in revised form) January 9, 2002; published electronically June 26, 2002. The research of the first two authors is mainly supported by the project *High Performance Methods for Mathematical Optimization* under the Dutch SWON-grant 613-304-200. Both the first author and the third author were partially supported by the National Science and Engineering Research Council of Canada, grant 227650-00, and an FPP grant from IBM T.J. Watson Research Lab.

http://www.siam.org/journals/siopt/13-1/38323.html

†Department of Computing and Software, McMaster University, Hamilton, ON, Canada, L8S 4L7 (pengj@mcmaster.ca, terlaky@mcmaster.ca).

‡Faculty of Information Technology and Systems, Delft University of Technology, P.O. Box 5031, 2600 GA Delft, The Netherlands (C.Roos@its.tudelft.nl).

For any $x = (x_1, \ldots, x_n)^T \in \Re^n$ let us define the matrix

$$\mathrm{mat}(x) = \begin{pmatrix} x_1 & x_{2:n} \\ x_{2:n}^T & x_1 E_{n-1} \end{pmatrix}, \tag{1.1}$$

where $x_{2:n} = (x_2, x_3, \ldots, x_n)$ and E_{n-1} denotes the identity matrix in $\Re^{(n-1)\times(n-1)}$. Using the above definition, one can easily prove that the vector $x \in K$ if and only if the matrix $\mathrm{mat}(x)$ is positive semidefinite, i.e., $\mathrm{mat}(x) \succeq 0$. This observation means that SOCO is essentially a specific case of semidefinite optimization (SDO). This delicate circumstance partially explains why SOCO did not attract as much attention as its counterparts LO and SDO.

We consider in this paper the standard SOCO problem, which takes the form

$$\begin{aligned} (\mathrm{SOCO}) \quad & \min c^T x \\ & \text{subject to (s.t.) } Ax = b, \quad x \succeq_K 0, \end{aligned}$$

and its dual

$$\begin{aligned} (\mathrm{SOCD}) \quad & \max b^T y \\ & \text{s.t. } A^T y + s = c, \quad s \succeq_K 0, \end{aligned}$$

where K is the product of several second-order cones, i.e., $K = K^1 \times K^2 \times \cdots \times K^N$ with

$$K^j = \left\{ (x_1^j, \ldots, x_{n_j}^j)^T \in \Re^{n_j} : \left(x_1^j\right)^2 \geq \sum_{i=2}^{n_j} \left(x_i^j\right)^2, x_1^j \geq 0 \right\},$$

$A \in \Re^{m\times n}$ with $n = \sum_{j=1}^N n_j$, and

$$x = \begin{pmatrix} x^1 \\ x^2 \\ \vdots \\ x^N \end{pmatrix}, \quad x^j \in \Re^{n_i},\ j = 1, 2, \ldots, N, \quad x \in \Re^n.$$

Further, K_+ denotes the interior of K. As is standard, the notation $x \succeq_K s$ (or $x \succ_K s$) means that $x - s \in K$ (or $x - s \in K_+$). In this paper the matrix A is further assumed to be of full row rank, i.e., $\mathrm{rank}\, A = m$.

An efficient approach to SOCO problems is to solve them using interior-point methods (IPMs) (see [10]). To be more specific, let us go into more details. Throughout this paper, we assume that both (SOCO) and (SOCD) satisfy the interior-point condition (IPC), i.e., there exists (x^0, y^0, s^0) such that

$$Ax^0 = b,\ x^0 \succ_K 0, \qquad A^T y^0 + s^0 = c,\ s^0 \succ_K 0.$$

It is known that the IPC is a rather mild assumption in the study of SOCO, since by using the homogeneous self-dual model described in [4] we could cast the original problem as a slightly larger SOCO problem such that a strictly feasible point for this new problem could be easily obtained. For this and other properties of SOCO, we

refer to [4], the recent book [27], and the references therein. Under the IPC, finding an optimal solution of SOCO is equivalent to solving the following system:

$$\begin{aligned} Ax &= b, \quad x \succeq_K 0, \\ A^T y + s &= c, \quad s \succeq_K 0, \\ \mathrm{mat}(x)s &= 0. \end{aligned} \tag{1.2}$$

The basic idea of primal-dual IPMs is to replace the third equation in (1.2), the so-called *complementarity condition* for (SOCO) and (SOCD), by the parameterized equation $\mathrm{mat}(x)s = \mu\tilde{e}$, with $\mu > 0$, where

$$\tilde{e} = \begin{pmatrix} \tilde{e}^1 \\ \tilde{e}^2 \\ \vdots \\ \tilde{e}^N \end{pmatrix}, \quad \tilde{e}^j = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \Re^{n_j}, \quad j = 1, 2, \ldots, N.$$

Thus we consider the system

$$\begin{aligned} Ax &= b, \quad x \succeq_K 0, \\ A^T y + s &= c, \quad s \succeq_K 0, \\ \mathrm{mat}(x)s &= \mu\tilde{e}. \end{aligned} \tag{1.3}$$

If the IPC holds, then for each $\mu > 0$ the parameterized system (1.3) has a unique solution. This solution is denoted by $(x(\mu), y(\mu), s(\mu))$, and we call $x(\mu)$ the μ-center of (SOCO) and $(y(\mu), s(\mu))$ the μ-center of (SOCD). The set of μ-centers (with μ running through all positive real numbers) gives a homotopy path, which is called *the central path*. The central path converges to the solution set of SOCO as μ reduces to zero [5, 15, 24]. IPMs trace the central path appropriately and find an approximate solution to the underlying SOCO problem as μ goes to zero.

To trace the central path efficiently, various strategies have been introduced to keep the iterative sequence in a certain neighborhood of the central path as well as to reduce the parameter μ. These strategies have played an important role in both the analysis and practice of IPMs. It is worth pointing out that two general strategies are widely used in IPMs with respect to the update of the parameter μ. These are the so-called small-update and large-update IPMs. It has been proven and generally accepted that the worst-case iteration bound of small-update IPMs is better than that for large-update IPMs, while the latter are much more efficient in practice (see the discussion in the introduction of [20]). This is a gap between the theory and practice of IPMs.

Recently, we introduced the concept of an univariate *self-regular* function and showed that any such function can be naturally extended to a proximity function on the positive orthant and the cone of positive definite matrices [20]. The *self-regular* proximities obtained in this way can be used in IPMs to keep control on the distance of an iterative sequence from the central path as well as to define the corresponding search directions. By using some new analysis tools developed in [19, 20] and employing the new search directions, we were able to show that the resulting new large-update IPMs for LO and SDO have polynomial $\mathcal{O}(n^{(q+1)/2q} \log \frac{n}{\epsilon})$ iteration bounds, where q is a constant, the so-called barrier degree of the proximity. This improves the previously known $\mathcal{O}(n \log \frac{n}{\epsilon})$ iteration bound of large-update IPMs.

The present work aims to extend the results of [20] to SOCO. We first point out that because of the relations among LO, SDO, and SOCO, many IPMs for SOCO can be cast as straightforward extensions of their counterparts for LO and SDO, and the polynomial convergence of those IPMs for SOCO can also be obtained by directly applying the known results of IPMs for SDO to SOCO. However, as pointed out in the book [15], although a SOCO problem can be solved via using an SDO approach, IPMs solving SOCO problems directly have iteration bounds depending on the number N of cones, which is lower than those of IPMs applied to the semidefinite formulation of the SOCO problem, where the complexity of an IPM is dependent on the number n of variables. In the case of SOCO, n might be much larger than N. This observation led to some works on IPMs for solving SOCO directly [13, 25]. It should also be noted that by using the notion of Jordan algebra, Faybusovich [7] and Alizadeh and Schmieta [3] discussed the complexity of IPMs for symmetric cones, which include SOCO as a specific case. The approach based on Jordan algebra was extensively investigated later by Schmieta and Alizadeh [22, 23], who proposed a way to transfer the Jordan algebra associated with the second-order cone into the so-called Clifford algebra in the cone of matrices and then carried out a unified analysis of the analysis for many IPMs in symmetric cones. It is worth noting that the complexity of IPMs for SOCO in [22] matches those presented in [2, 13, 16, 17, 25].

A direct and interesting question here is whether we can extend our approach in [20] to SOCO by using the above-mentioned techniques and bound the number of iterations by a function of N. As we will see later, this is far from an easy task. Let us explain why. First we point out that the proof of the global convergence of IPMs based on the *self-regular* approach involves the analysis of general analytic functions which are defined on the associated cone and their derivatives. For instance, a key in establishing the complexity of the algorithm for SDO in [20] is the following inequality:

$$\Psi(XS) \leq \frac{1}{2}\big(\Psi(X^2) + \Psi(S^2)\big) \qquad \forall X, S \succ 0,$$

where $\Psi(\cdot)$ is a so-called self-regular function in the cone of positive definite matrices. The proof of this inequality refers to the singular value decomposition of a matrix. This prevents a direct extension of the proof in [20] to SOCO. As we will see in section 2, this inequality does not hold in general for the case of SOCO. Another important step in the analysis of [20] is to estimate the second-order derivative of a matrix-valued mapping that involves the derivatives of each element of the underlying matrix function. However, the approach suggested in [22] transfers a vector $x \in \Re^n$ in the second-order cone into a matrix $X \in \Re^{2^n \times 2^n}$. In this situation, it is too difficult to get any reasonable estimation about the second-order derivative of the resulting matrix-valued mapping. The above-mentioned issues indicate that in order to extend the approach of [20] to SOCO, a new and separate treatment is necessary.

This work follows the same main steps as those in [20]. The paper is organized as follows. In section 2, we introduce the definition of general analytic functions defined on K and discuss versatile properties of these functions. Section 3 is devoted to introducing the new search direction and the scaling technique for SOCO. The notions of *self-regular* functions and *self-regular* proximities in the second-order cone K are discussed as well. In section 4 we describe our new algorithm based on a self-regular proximity. Then we establish the polynomial complexity of the algorithm and finally close this paper with some concluding remarks.

We mention that in the rest of this paper we denote by $\Re_+$ the nonnegative axis, i.e., $\Re_+ = [0, \infty)$, and by $\Re_{++}$ the positive axis, i.e., $\Re_{++} = (0, \infty)$.

2. Preliminary results on functions associated with second-order cones. In the design and analysis of IPMs, we always resort to some functions defined in a suitable space. In this section we will present some fundamental results about general functions defined in the second-order cone. As we discussed in the introduction, Jordan algebra [6, 7, 22, 23] has played an important role in extending IPMs to symmetric cones. However, these existing known results on Jordan algebra are not enough to generalize the approach presented in [20] to SOCO.

We first consider general functions on the second-order cone via Jordan algebra. To ease the discussion, in this technical section we assume that the cone K is defined with $N = 1$. First we observe that closely associated with the cone K is a matrix

$$Q = \operatorname{diag}(1, -1, \ldots, -1).$$

We refer to Q as the representation matrix of the cone K since there holds simply

$$K = \{x \in \Re^n : x^T Q x \geq 0,\ x_1 \geq 0\}.$$

Obviously one has $Q^2 = E$.

2.1. Jordan algebra. The Euclidean Jordan algebra for the second-order cone K is defined by the bilinear operator

$$x \circ s = (x^T s, x_1 s_2 + s_1 x_2, \ldots, x_1 s_n + s_1 x_n)^T = (x^T s, x_1 s_{2:n} + s_1 x_{2:n})^T, \tag{2.1}$$

where $x, s \in \Re^n$. Obviously, the Jordan product $\circ$ is commutative, i.e., $x \circ s = s \circ x$. It is also easy to verify that for any $x, s \in \Re^n$ one has

$$x \circ s = \operatorname{mat}(x)s.$$

It may be worthwhile to point out that the cone K is not closed under the Jordan product. For example, if $n = 3$, then $x = (1.5, 1, 1)^T \in K$ and $s = (1.5, 1, -1)^T \in K$, but $x \circ s = (2.25, 3, 0)^T \notin K$.

2.2. Eigenvalues, trace, and determinant associated with second-order cones. We denote by $\lambda_{\max}(x)$ and $\lambda_{\min}(x)$ the maximal and minimal eigenvalues of the matrix $\operatorname{mat}(x)$, respectively. Namely,

$$\lambda_{\max}(x) = x_1 + \|x_{2:n}\|, \qquad \lambda_{\min}(x) = x_1 - \|x_{2:n}\|. \tag{2.2}$$

The trace and the determinant of a vector $x \in \Re^n$ associated with K can be defined as follows [6].

DEFINITION 2.1.[1] *For any $x \in \Re^n$, the trace of x associated with K is defined by*

$$\mathbf{Tr}(x) = \lambda_{\max}(x) + \lambda_{\min}(x) = 2x_1, \tag{2.3}$$

and the determinant of x associated with K is given by

$$\det(x) = \lambda_{\max}(x)\lambda_{\min}(x) = x_1^2 - \|x_{2:n}\|^2. \tag{2.4}$$

From the above definitions, one can easily see that for any $x, s \in \Re^n$ one has

$$\mathbf{Tr}(x \circ s) = 2x^T s, \qquad \mathbf{Tr}(x \circ x) = 2\|x\|^2.$$

[1] These definitions can be viewed as variants of the trace and determinant of general matrices.

Our next lemma collects several elementary results about the behavior of the trace and determinant of the Jordan product of two vectors. These results demonstrate the differences between the determinant and trace for elements of the second-order cone K and those notions as usually defined for matrices.[2]

LEMMA 2.2. *Suppose that x and s are two vectors in K. Then we have*

$$\lambda_{\max}(x)\lambda_{\min}(s)+\lambda_{\min}(x)\lambda_{\max}(s)\leq \mathbf{Tr}(x\circ s)\leq \lambda_{\max}(x)\lambda_{\max}(s)+\lambda_{\min}(x)\lambda_{\min}(s) \tag{2.5}$$

and

$$\det(x\circ s)\leq \det(x)\det(s). \tag{2.6}$$

Furthermore, equality holds in (2.6) *if and only if there exist two constants $\beta_1,\beta_2\in\Re$ with $|\beta_1|+|\beta_2|>0$ such that $\beta_1 x_{2:n}=\beta_2 s_{2:n}$.*

Proof. We first consider the relation (2.5). Using the notations $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ given by (2.2) and the well-known Cauchy–Schwartz inequality, since both x and s belong to K, one has

$$\begin{aligned}
0 &\leq \lambda_{\max}(x)\lambda_{\min}(s)+\lambda_{\min}(x)\lambda_{\max}(s) = 2(x_1s_1-\|x_{2:n}\|\,\|s_{2:n}\|)\\
&\leq 2x^Ts=\mathbf{Tr}(x\circ s)\leq 2(x_1s_1+\|x_{2:n}\|\,\|s_{2:n}\|)\\
&=\lambda_{\max}(x)\lambda_{\max}(s)+\lambda_{\min}(x)\lambda_{\min}(s),
\end{aligned}$$

which gives (2.5).

To prove (2.6), we note that, by making use of the definition (2.4), one gets

$$\begin{aligned}
\det(x\circ s) &= (x^Ts)^2-\|x_1s_{2:n}+s_1x_{2:n}\|^2\\
&=(x_1s_1)^2+(x_{2:n}s_{2:n}^T)^2-(x_1)^2\|s_{2:n}\|^2-(s_1)^2\|x_{2:n}\|^2\\
&\leq (x_1s_1)^2+\|x_{2:n}\|^2\|s_{2:n}\|^2-(x_1)^2\|s_{2:n}\|^2-(s_1)^2\|x_{2:n}\|^2\\
&=((x_1)^2-\|x_{2:n}\|^2)((s_1)^2-\|s_{2:n}\|^2)=\det(x)\det(s),
\end{aligned}$$

and equality holds if and only if $|x_{2:n}s_{2:n}^T|=\|x_{2:n}\|\,\|s_{2:n}\|$. This means equality holds only when the vectors $x_{2:n}$ and $s_{2:n}$ are linearly dependent. The proof of the lemma is complete. ∎

2.3. Functions associated with second-order cone and their derivatives. Note that if $n=1$, then $K=\Re_+$, and if $n\geq 1$, then $\Re_+\subseteq K$. Our aim in this section is to show that any function mapping $\Re_+$ into $\Re_+$ can be naturally extended to a function that maps K into itself. First we observe that for every $x\in\Re^n$ we have the so-called *spectral decomposition*

$$x=\lambda_{\max}(x)z_1+\lambda_{\min}(x)z_2,$$

where

$$z^1=\frac{1}{2}\left(1,\frac{x_{2:n}}{\|x_{2:n}\|}\right)^T \quad\text{and}\quad z^2=\frac{1}{2}\left(1,\frac{-x_{2:n}}{\|x_{2:n}\|}\right)^T.$$

[2]This distinction can be anticipated by noticing that $\mathbf{Tr}(\mathrm{mat}(x))=nx_1=n/2\mathbf{Tr}(x)$ and that

$$\det(\mathrm{mat}(x))=x_1^{n-2}(x_1^2-\|x_{2:n}\|^2)=x_1^{n-2}\det(x).$$

Here by convention $\frac{x_{2:n}}{\|x_{2:n}\|} = 0$ if $x_{2:n} = 0$. It is easy to see that

$$z^1 \circ z^2 = 0. \tag{2.7}$$

Now we are ready to give the definition of general analytical functions associated with the second-order cone K.

DEFINITION 2.3. *Suppose that $\psi(t)$ is a function from $\Re$ to $\Re$ and $x \in \Re^n$. Then the function $\psi(x) : \Re^n \to \Re^n$ associated with the second-order cone K is defined as follows:* [3]

$$\psi(x) = \psi(\lambda_{\max}(x))z^1 + \psi(\lambda_{\min}(x))z^2. \tag{2.8}$$

The function $\psi(t)$ is called the kernel function of $\psi(x)$. It can easily be verified that if $\psi(t) \geq 0$ for any $t \geq 0$ and $x \in K$, then the above definition implies that $\psi(x) \in K$. Thus it becomes clear that every nonnegative (positive) function on the nonnegative (positive) axis naturally extends to a function that maps (the interior of) K into itself. Likewise for the LO and SDO cases the function $\psi(t)$ is called the kernel function of $\psi(x)$. As a consequence of the above definition we have a big source of functions mapping K into itself. For instance, we may write x^p, where p is any number in $\Re$ and $x \in K$. Let us consider some special cases, for example, $p = -1$. In this case Definition 2.3 yields

$$x^{-1} = \frac{1}{\det(x)}(x_1, -x_2, -x_3, \ldots, -x_n)^T \qquad \forall x \succ_K 0,$$

and one may easily see that $x \circ x^{-1} = \tilde{e}$. Moreover, it is also clear that any analytic function like $\exp(x)$ is now well defined. Similarly we can define the function $\psi'(x)$ by (2.8), whose kernel function is $\psi'(t)$.

The following result concerning the behavior of composing functions with respect to the Jordan product follows directly from Definition 2.3 and (2.7).

LEMMA 2.4. *Suppose that $\psi_1(t)$ and $\psi_2(t)$ are two functions from $\Re$ into $\Re$ and that $\psi_1(x)$ and $\psi_2(x)$ are two associated functions defined by* (2.8). *If $\psi_0(t) = \psi_1(t)\psi_2(t)$, then $\psi_0(x) = \psi_1(x) \circ \psi_2(x)$ holds for any $x \in \Re^n$.*

It is trivial to verify the following result about general functions associated with the second-order cone defined by Definition 2.3.

LEMMA 2.5. *Suppose that the function $\psi(x) : \Re^n \to \Re^n$ is defined by Definition* 2.3. *Then*

$$\begin{aligned} \|\psi(x)\| &= \frac{\sqrt{2}}{2}\sqrt{\psi^2(\lambda_{\max}(x)) + \psi^2(\lambda_{\min}(x))}, \\ \mathbf{Tr}(\psi(x)) &= \psi(\lambda_{\max}(x)) + \psi(\lambda_{\min}(x)), \\ \det(\psi(x)) &= \psi(\lambda_{\max}(x))\psi(\lambda_{\min}(x)). \end{aligned}$$

Let us elaborate a little more on the relations among the eigenvalues of two vectors in the second-order cone and those of their Jordan product. Note that from Lemma 2.2 we know that for any $x, s \succ_K 0$, the determinant $\det(x \circ s) = \det(x)\det(s)$ if and only if the vectors $x_{2:n}$ and $s_{2:n}$ are linearly dependent. Without loss of generality, we may assume that $x_{2:n} \neq 0$ and $s_{2:n} = \beta x_{2:n}$ for some $\beta \in \Re$. In the following discussion we

[3] We note that recently Fukushima, Luo, and Tseng [9] also defined functions associated with a second-order cone. Their definition is slightly different from our definition (2.8). However, the definition given by (2.8) is clearer and more direct.

will show that this further implies that the vector s can be represented as a function of x and that the vector $x \circ s \in K_+$.

LEMMA 2.6. *Suppose that x and s are two vectors in K_+ with $x_{2:n} \neq 0$. If*

$$\det(x \circ s) = \det(x)\det(s),$$

then there exists a function $\psi(t) : \Re_+ \to \Re_+$ such that $s = \psi(x)$ and the Jordan product $x \circ s \in K_+$. Moreover,

$$\lambda_{\min}(x)\lambda_{\min}(s) \leq \lambda_{\min}(x \circ s) \leq \lambda_{\max}(x \circ s) \leq \lambda_{\max}(x)\lambda_{\max}(s). \tag{2.9}$$

Proof. By making use of the last conclusion of Lemma 2.2, since $\det(x \circ s) = \det(x)\det(s)$, one can claim that

$$s_{2:n} = \frac{\beta \|s_{2:n}\|}{\|x_{2:n}\|} x_{2:n},$$

where β equals 1 or -1. Let $\psi(t)$ be a univariate function from $\Re_{++}$ into $\Re_{++}$ satisfying the following equalities:

$$\frac{1}{2}(\psi(\lambda_{\max}(x)) + \psi(\lambda_{\min}(x))) = s_1, \qquad \frac{1}{2}(\psi(\lambda_{\max}(x)) - \psi(\lambda_{\min}(x))) = \beta \|s_{2:n}\|.$$

Such a function exists since $s \in K_+$. Thus, by definition (2.8) we can claim $s = \psi(x)$. Now invoking Lemma 2.4, we can write $x \circ s = \psi_1(x)$, where the kernel function $\psi_1(t) = t\psi(t)$. Therefore, since $\psi_1(t) > 0$ for any $t > 0$ and $x_{2:n} \neq 0$, from its basic definition (2.8) it follows that $x \circ s = \psi_1(x) \succ_K 0$.

It remains to prove (2.9). Since $x \circ s = \psi_1(x)$, we thus have

$$\begin{aligned}\lambda_{\max}(x \circ s) &= \max\{\psi_1(\lambda_{\max}(x)), \psi_1(\lambda_{\min}(x))\} \\ &\leq \lambda_{\max}(x) \max\{\psi(\lambda_{\max}(x)), \psi(\lambda_{\min}(x))\} = \lambda_{\max}(x)\lambda_{\max}(s),\end{aligned}$$

which, together with the assumption in the lemma that $\det(x \circ s) = \det(x)\det(s)$, further yields

$$\lambda_{\min}(x \circ s) \geq \lambda_{\min}(x)\lambda_{\min}(s).$$

This completes the proof of the lemma. □

It is worthwhile to compare the above lemma with its analogue in matrix theory. Suppose that X and S are both symmetric and positive definite. Let $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ denote the maximal and minimal eigenvalues of the corresponding matrix. Then we have

$$\lambda_{\min}(X)\lambda_{\min}(S) \leq \lambda_{\min}(XS) \leq \lambda_{\max}(XS) \leq \lambda_{\max}(X)\lambda_{\max}(S). \tag{2.10}$$

Note that for any $X, S \in \Re^{n \times n}$ the relation $\det(XS) = \det(X)\det(S)$ holds trivially. The results presented in Lemma 2.6 are very helpful in our later discussion about the features of *self-regular* functions associated with the second-order cone. These functions can be viewed as a direct extension of univariate *self-regular* functions introduced in [20].

DEFINITION 2.7. *A function $\psi(t) \in \mathcal{C}^2 : \Re_{++} \to \Re_+$ is self-regular if it satisfies the following conditions:*

C.1 *$\psi(t)$ is strictly convex with respect to $t > 0$ and vanishes at its global minimal point $t = 1$; i.e., $\psi(1) = \psi'(1) = 0$. Further, there exist positive constants $\nu_2 \geq \nu_1 > 0$ and $p \geq 1$, $q \geq 1$ such that*

$$\nu_1(t^{p-1} + t^{-1-q}) \leq \psi''(t) \leq \nu_2(t^{p-1} + t^{-1-q}) \quad \forall t \in (0, \infty). \tag{2.11}$$

C.2 *For any $t_1, t_2 > 0$,*

$$\psi(t_1^r t_2^{1-r}) \leq r\psi(t_1) + (1-r)\psi(t_2) \quad \forall r \in [0, 1]. \tag{2.12}$$

We call parameter q the *barrier degree* and p the *growth degree* of the function $\psi(t)$ if it is *self-regular*. A typical family of *self-regular* functions is given by

$$\Upsilon_{p,q}(t) = \frac{1}{p(p+1)}\left(t^{p+1} - 1\right) + \frac{1}{q(q-1)}\left(t^{1-q} - 1\right) + \frac{p-q}{pq}(t-1), \quad p, q \geq 1. \tag{2.13}$$

It is worth mentioning that the function $\Upsilon_{p,q}(t)$ satisfies condition C.1 with $\nu_1 = \nu_2 = 1$.

The definition of a *self-regular* function for the second-order cone K is recorded as follows.

DEFINITION 2.8. *A function $\psi(x)$ associated with the second-order cone K given by* (2.8) *is said to be self-regular if its kernel function $\psi(t)$ is self-regular.*

We denote by $\Psi(x)$ the trace of the function $\psi(x)$, i.e.,

$$\Psi(x) = \mathbf{Tr}(\psi(x)) = \psi(\lambda_{\max}(x)) + \psi(\lambda_{\min}(x)). \tag{2.14}$$

Our next proposition characterizes several important properties of a *self-regular* function for the second-order cone K.

PROPOSITION 2.9. *Let the functions $\psi(x) : K_+ \to K$ and $\Psi(x) : K_+ \to \Re_+$ be defined by* (2.8) *and* (2.14)*, respectively. If the function $\psi(x)$ is self-regular, then the following statements hold:*

(i) *$\Psi(x)$ is strictly convex with respect to $x \in K_+$ and vanishes at its global minimal point $x = \tilde{e}$; i.e., $\Psi(\tilde{e}) = 0$, $\psi(\tilde{e}) = \psi'(\tilde{e}) = 0$. Further, there exist positive constants $\nu_1, \nu_2 > 0$ and $p, q \geq 1$ such that*

$$\nu_1(x^{p-1} + x^{-1-q}) \preceq_K \psi''(x) \preceq_K \nu_2(x^{p-1} + x^{-1-q}). \tag{2.15}$$

(ii) *Suppose that x and s are two vectors in K_+. If $v \in K_+$ satisfies*

$$\det\left(v^2\right) = \det(x)\det(s), \quad \mathbf{Tr}\left(v^2\right) = \mathbf{Tr}(x \circ s),$$

then

$$\Psi(v) \leq \frac{1}{2}(\Psi(x) + \Psi(s)). \tag{2.16}$$

Proof. To show that the first claim of the proposition is true, we need to show that $\Psi(x)$ is strictly convex for $x \succ_K 0$, that is, for any $x, s \succ_K 0$ and $x \neq s$ there holds

$$\Psi\left(\frac{x+s}{2}\right) < \frac{1}{2}(\Psi(x) + \Psi(s)).$$

Since $x, s \in K_+$, through simple calculus one can prove that

$$\lambda_{\max}\left(\frac{x+s}{2}\right) = \frac{x_1 + s_1}{2} + \frac{1}{2}\left\|x_{2:n} + s_{2:n}\right\| \leq \frac{1}{2}(\lambda_{\max}(x) + \lambda_{\max}(s))$$

and similarly

$$\lambda_{\min}\left(\frac{x+s}{2}\right) = \frac{x_1+s_1}{2} - \frac{1}{2}\left\|x_{2:n}+s_{2:n}\right\| \geq \frac{1}{2}(\lambda_{\min}(x)+\lambda_{\min}(s)).$$

Recalling the definitions of $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$, it follows trivially that

$$\lambda_{\max}\left(\frac{x+s}{2}\right)+\lambda_{\min}\left(\frac{x+s}{2}\right) = x_1+s_1 = \frac{1}{2}(\lambda_{\max}(x)+\lambda_{\min}(x)+\lambda_{\max}(s)+\lambda_{\min}(s)).$$

Thus, from the above three relations we can conclude that there exist two constants $\beta_1 \geq 0$ and $\beta_2 \geq 0$ with $\beta_1+\beta_2 = 1$ such that

$$\begin{aligned}
\lambda_{\max}\left(\frac{x+s}{2}\right) &= \frac{\beta_1}{2}(\lambda_{\min}(x)+\lambda_{\min}(s)) + \frac{\beta_2}{2}(\lambda_{\max}(x)+\lambda_{\max}(s)),\\
\lambda_{\min}\left(\frac{x+s}{2}\right) &= \frac{\beta_2}{2}(\lambda_{\min}(x)+\lambda_{\min}(s)) + \frac{\beta_1}{2}(\lambda_{\max}(x)+\lambda_{\max}(s)).
\end{aligned}$$

Now, making use of the strict convexity of the function $\psi(t)$ twice, one obtains

$$\begin{aligned}
\Psi\left(\frac{x+s}{2}\right) &= \psi\left(\lambda_{\max}\left(\frac{x+s}{2}\right)\right) + \psi\left(\lambda_{\min}\left(\frac{x+s}{2}\right)\right)\\
&= \psi\left(\frac{\beta_1}{2}(\lambda_{\min}(x)+\lambda_{\min}(s)) + \frac{\beta_2}{2}(\lambda_{\max}(x)+\lambda_{\max}(s))\right)\\
&\quad + \psi\left(\frac{\beta_2}{2}(\lambda_{\min}(x)+\lambda_{\min}(s)) + \frac{\beta_1}{2}(\lambda_{\max}(x)+\lambda_{\max}(s))\right)\\
&\leq \psi\left(\frac{\lambda_{\min}(x)+\lambda_{\min}(s)}{2}\right) + \psi\left(\frac{\lambda_{\max}(x)+\lambda_{\max}(s)}{2}\right)\\
&\leq \frac{1}{2}(\psi(\lambda_{\max}(x)) + \psi(\lambda_{\min}(x)) + \psi(\lambda_{\max}(s)) + \psi(\lambda_{\min}(s)))\\
&= \frac{1}{2}(\Psi(x)+\Psi(s)).
\end{aligned}$$

Note that since $x \neq s$, at least one of the two inequalities in the above proof holds strictly. This proves the strict convexity of $\Psi(x)$. The remaining terms in the first statement can be verified through direct calculus.

It remains to prove the second statement of the proposition. For this we first observe that, since $v \in K_+$, there hold

$$\det(v) = \det\left(v^2\right)^{\frac{1}{2}} = \det(x)^{\frac{1}{2}}\det(s)^{\frac{1}{2}} = (\lambda_{\min}(x)\lambda_{\min}(s))^{\frac{1}{2}}(\lambda_{\max}(x)\lambda_{\max}(s))^{\frac{1}{2}} \tag{2.17}$$

and

$$\begin{aligned}
\mathbf{Tr}(v) &= \lambda_{\max}(v)+\lambda_{\min}(v) = \left(\lambda_{\max}(v)^2+\lambda_{\min}(v)^2+2\lambda_{\max}(v)\lambda_{\min}(v)\right)^{\frac{1}{2}}\\
&= \left(\mathbf{Tr}(v^2)+2\det(v)\right)^{\frac{1}{2}} = \left(\mathbf{Tr}(x\circ s)+2(\det(x)\det(s))^{\frac{1}{2}}\right)^{\frac{1}{2}}\\
&\leq \left(\lambda_{\min}(x)\lambda_{\min}(s)+\lambda_{\max}(x)\lambda_{\max}(s)+2(\lambda_{\min}(x)\lambda_{\min}(s)\lambda_{\max}(x)\lambda_{\max}(s))^{\frac{1}{2}}\right)^{\frac{1}{2}}\\
&= \sqrt{\lambda_{\min}(x)\lambda_{\min}(s)}+\sqrt{\lambda_{\max}(x)\lambda_{\max}(s)}, \qquad (2.18)
\end{aligned}$$

where the inequality follows from (2.5). Therefore, by making use of (2.17) and (2.18), we obtain

$$\begin{aligned}\lambda_{\max}(v) &= \frac{1}{2}(\mathbf{Tr}(v) + \lambda_{\max}(v) - \lambda_{\min}(v)) = \frac{1}{2}\mathbf{Tr}(v) + \frac{1}{2}\sqrt{\mathbf{Tr}(v)^2 - 4\det(v^2)} \\ &\leq \frac{1}{2}\mathbf{Tr}(v) + \frac{1}{2}\Big(\sqrt{\lambda_{\max}(x)\lambda_{\max}(s)} - \sqrt{\lambda_{\min}(x)\lambda_{\min}(s)}\Big) \\ &\leq \lambda_{\max}(x)^{\frac{1}{2}}\lambda_{\max}(s)^{\frac{1}{2}},\end{aligned}$$

where all inequalities follow from (2.17) and (2.18). Now, invoking (2.17), we can further claim

$$\lambda_{\min}(v) \geq \lambda_{\min}(x)^{\frac{1}{2}}\lambda_{\min}(s)^{\frac{1}{2}}.$$

From the above discussions we can easily verify that there exists a constant $r \in [\frac{1}{2}, 1)$ such that

$$\begin{aligned}\lambda_{\min}(v) &= \lambda_{\min}(x)^{\frac{r}{2}}\lambda_{\min}(s)^{\frac{r}{2}}\lambda_{\max}(x)^{\frac{1-r}{2}}\lambda_{\max}(s)^{\frac{1-r}{2}}, \\ \lambda_{\max}(v) &= \lambda_{\max}(x)^{\frac{r}{2}}\lambda_{\max}(s)^{\frac{r}{2}}\lambda_{\min}(x)^{\frac{1-r}{2}}\lambda_{\min}(s)^{\frac{1-r}{2}}.\end{aligned}$$

By applying condition C.2 twice, we deduce

$$\begin{aligned}\Psi(v) &= \psi(\lambda_{\min}(v)) + \psi(\lambda_{\max}(v)) \\ &= \psi\Big(\lambda_{\min}(x)^{\frac{r}{2}}\lambda_{\min}(s)^{\frac{r}{2}}\lambda_{\max}(x)^{\frac{1-r}{2}}\lambda_{\max}(s)^{\frac{1-r}{2}}\Big) \\ &\qquad + \psi\Big(\lambda_{\max}(x)^{\frac{r}{2}}\lambda_{\max}(s)^{\frac{r}{2}}\lambda_{\min}(x)^{\frac{1-r}{2}}\lambda_{\min}(s)^{\frac{1-r}{2}}\Big) \\ &\leq \psi\Big(\lambda_{\min}(x)^{\frac{1}{2}}\lambda_{\min}(s)^{\frac{1}{2}}\Big) + \psi\Big(\lambda_{\max}(x)^{\frac{1}{2}}\lambda_{\max}(s)^{\frac{1}{2}}\Big) \\ &\leq \frac{1}{2}(\psi(\lambda_{\min}(x)) + \psi(\lambda_{\max}(x)) + \psi(\lambda_{\max}(s)) + \psi(\lambda_{\min}(s))) \\ &= \frac{1}{2}(\Psi(x) + \Psi(s)).\end{aligned}$$

This completes the proof of the proposition. □

It is of interest to compare Proposition 2.9 with its SDO analogue, Proposition 4.4 in [20]. First we find that statement (ii) in the present paper is slightly different from condition C.4 which is required in [20]. Actually, one can easily see that the matrix used in condition C.4 of [20] satisfies certain conditions similar to the one posed in Proposition 2.9. However, the choice of the vector v allowing such conditions in the second-order cone is much more restricted than in the case of SDO. One possible reason for this phenomenon is that, for general $x, s \in K_+$, the Jordan product $x \circ s$ might not belong to K. For instance, let us consider an example in $K \in \Re^3$ with $x = (2+t, 1, 1)^T$, $s = (2+t, 1, -1)^T \in K_+$, where t is some small positive number. Obviously one has $x \circ s = ((2+t)^2, 4+2t, 0)^T \in K_+$. Moreover, it is trivial to see that $\lambda_{\min}(x \circ s) = 2t + t^2$. Thus when t reduces to zero, the function $\Psi(x \circ s)$ goes to infinity. However, one can readily verify that for sufficiently small $t > 0$ both of the functions $\Psi(x)$ and $\Psi(s)$ are bounded above. This example shows that, for $x, s \in K_+$, if $\det(x \circ s) \neq \det(x)\det(s)$, then the relation

$$\Psi(x \circ s) \leq \frac{1}{2}\big(\Psi(x^2) + \Psi(s^2)\big)$$

might fail. We also mention that, in the SDO case, for any positive definite matrices X and S, since the matrix XS is diagonalizable and has positive eigenvalues, the function $\Psi(XS)$ is well defined (see [20]).

Likewise, what we have observed in the LO and SDO cases [20] is that, in order to establish the complexity of the algorithm, we need to bound the derivatives of certain proximity functions in suitable spaces. With specification to SOCO, this requires us to discuss the derivatives of the functions $\psi(x(t))$ and $\Psi(x(t))$, where

$$x(t) = (x_1(t), \ldots, x_n(t))^T$$

is a mapping from $\Re$ into $\Re^n$. Our next result for functions in the second-order cone K resembles Lemma 4.9 for matrix function in [20]. However, our proof for the SOCO case here is direct and much simpler. First, for simplicity, let us denote by $x'(t)$ the derivative of $x(t)$ with respect to t such that

$$x'(t) = (x_1'(t), \ldots, x_n'(t))^T.$$

Our following results provide means to measure the first-order directional derivative of a general function $\Psi(x(t))$ and bound its second-order derivative with respect to the variable t. Recall that, by (2.8), we can define the function $\psi'(x)$ as the function whose kernel function is $\psi'(t)$.

LEMMA 2.10. *Suppose that $x(t)$ is a mapping from $\Re$ into $\Re^n$. If $x(t)$ is twice differentiable with respect to t for all $t \in (l_t, u_t)$, and $\psi(t)$ is also a twice continuously differentiable function in a suitable domain which contains $\lambda_{\min}(x(t))$ and $\lambda_{\max}(x(t))$, then*

$$\frac{d}{dt}\mathbf{Tr}(\psi(x(t))) = \mathbf{Tr}(\psi'(x(t)) \circ x'(t)) \qquad \forall t \in (l_t, u_t)$$

and

$$\frac{d^2}{dt^2}\mathbf{Tr}(\psi(x(t))) \leq \varpi \mathbf{Tr}(x'(t) \circ x'(t)) + \mathbf{Tr}(\psi'(x(t)) \circ x''(t)), \tag{2.19}$$

where

$$\varpi = \max\left\{ |\psi''(\lambda_{\max}(x(t)))|, |\psi''(\lambda_{\min}(x(t)))|, \frac{|\psi'(\lambda_{\max}(x(t))) - \psi'(\lambda_{\min}(x(t)))|}{2\|x_{2:n}(t)\|} \right\}.$$

Proof. Without loss of generality, we assume that $\|x_{2:n}\| > 0$. From Lemma 2.5 we obtain

$$\mathbf{Tr}(\psi(x(t))) = \psi(\lambda_{\max}(x(t))) + \psi(\lambda_{\min}(x(t))).$$

It follows that

$$\begin{aligned}\frac{d}{dt}\mathbf{Tr}(\psi(x(t))) = {} & \psi'(\lambda_{\max}(x(t)))\left(x_1'(t) + \frac{1}{\|x_{2:n}(t)\|}\sum_{i=2}^n x_i(t)x_i'(t)\right) \\ & + \psi'(\lambda_{\min}(x(t)))\left(x_1'(t) - \frac{1}{\|x_{2:n}(t)\|}\sum_{i=2}^n x_i(t)x_i'(t)\right).\end{aligned}$$

Now recalling definition (2.8), we obtain

$$\psi'(x(t)) = \frac{1}{2}\left(\psi'(\lambda_{\max}(x(t))) + \psi'(\lambda_{\min}(x(t))), \frac{\psi'(\lambda_{\max}(x(t))) - \psi'(\lambda_{\min}(x(t)))}{\|x_{2:n}(t)\|}x_{2:n}(t)\right)^T.$$

By simple calculus, from the above two equalities one can readily check that

$$\frac{d}{dt}\mathbf{Tr}(\psi(x(t))) = 2\psi'(x(t))^T x'(t) = \mathbf{Tr}(\psi'(x(t)) \circ x'(t)).$$

This proves the first statement of the lemma.

To prove the second statement of the lemma, we first observe that

$$\frac{d^2}{dt^2}\Psi(x(t)) = \mathbf{Tr}\left(\frac{d}{dt}\psi'(x(t)) \circ x'(t)\right) + \mathbf{Tr}(\psi'(t) \circ x''(t)).$$

It is straightforward to check that

$$\frac{d}{dt}\psi'(x(t)) = v1 + v2 + v3,$$

where

$$v1 = \frac{\psi''(\lambda_{\max}(x(t)))\left(x_1'(t) + \frac{\sum_{i=2}^n x_i(t)x_i'(t)}{\|x_{2:n}(t)\|}\right)}{2}\left(1, \frac{x_{2:n}(t)}{\|x_{2:n}(t)\|}\right)^T,$$

$$v2 = \frac{\psi''(\lambda_{\min}(x(t)))\left(x_1'(t) - \frac{\sum_{i=2}^n x_i(t)x_i'(t)}{\|x_{2:n}(t)\|}\right)}{2}\left(1, -\frac{x_{2:n}(t)}{\|x_{2:n}(t)\|}\right)^T,$$

$$v3 = \frac{\psi'(\lambda_{\max}(x(t))) - \psi'(\lambda_{\min}(x(t)))}{2\|x_{2:n}(t)\|}\left(0, x_{2:n}'(t) - \frac{\sum_{i=2}^n x_i(t)x_i'(t)}{\|x_{2:n}(t)\|^2}x_{2:n}(t)\right)^T.$$

By using the well-known Cauchy–Schwarz inequality, we deduce

$$\frac{\sum_{i=2}^n x_i(t)x_i'(t)}{\|x_{2:n}(t)\|} \le \|x_{2:n}'(t)\|.$$

This relation, together with the definition of ϖ, further implies

$$\begin{aligned}\mathbf{Tr}((v1+v2)\circ x'(t)) &= \psi''(\lambda_{\max}(x(t)))\left(x_1'(t) + \frac{\sum_{i=2}^n x_i(t)x_i'(t)}{\|x_{2:n}(t)\|}\right)^2 \\ &\quad + \psi''(\lambda_{\min}(x(t)))\left(x_1'(t) - \frac{\sum_{i=2}^n x_i(t)x_i'(t)}{\|x_{2:n}(t)\|}\right)^2 \\ &\le 2\varpi\left((x_1'(t))^2 + \left(\frac{\sum_{i=2}^n x_i(t)x_i'(t)}{\|x_{2:n}(t)\|}\right)^2\right).\end{aligned}$$

On the other hand, through simple calculus, one has

$$\begin{aligned}\mathbf{Tr}(v3\circ x'(t)) &= \frac{\psi'(\lambda_{\max}(x(t))) - \psi'(\lambda_{\min}(x(t)))}{\|x_{2:n}(t)\|}\left(\|x_{2:n}'(t)\|^2 - \left(\frac{\sum_{i=2}^n x_i(t)x_i'(t)}{\|x_{2:n}(t)\|}\right)^2\right) \\ &\le 2\varpi\left(\|x_{2:n}'(t)\|^2 - \left(\frac{\sum_{i=2}^n x_i(t)x_i'(t)}{\|x_{2:n}(t)\|}\right)^2\right).\end{aligned}$$

Finally, by summing up the above two inequalities, we obtain the desired relation (2.19), which completes the proof of the lemma. □

It is worthwhile to consider the special case $K \subset \Re^2$, where we can also cast a SOCO problem as an SDO problem. Note that for the SDO case one has

$$X(t) = \begin{pmatrix} x_1(t) & x_2(t) \\ x_2(t) & x_1(t) \end{pmatrix}.$$

In this situation, the equalities

$$\|X'(t)\|^2 = 2\|x'(t)\|^2 = \mathbf{Tr}(x'(t) \circ x'(t))$$

hold trivially. Recalling the difference between the definitions of $\Psi(x)$ by (2.14) and $\Psi(X)$ in [20], one can easily verify that the estimations given in Lemma 2.10 are precisely the same as the ones presented in its SDO analogue Lemma 4.9 of [20].

3. Self-regular proximity functions and new search directions for SOCO.

3.1. Scaling schemes. In the present section we consider diverse search directions used in IPMs for solving SOCO and introduce some new search directions based on *self-regular* proximity functions in the second-order cone.

Most IPMs for solving SOCO employ different search directions together with suitable strategies for following the central path appropriately. As in the SDO case, the search directions for SOCO are usually derived from certain Newton systems in various scaled spaces. Note that the standard linearized Newton system for (1.3) can be written as

$$\begin{pmatrix} A & 0 & 0 \\ 0 & E_n & A^T \\ \mathrm{mat}(s) & \mathrm{mat}(x) & 0 \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta s \\ \Delta y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \mu\tilde{e} - \mathrm{mat}(x)s \end{pmatrix}, \quad x, s \succ_K 0. \tag{3.1}$$

This system might not be well defined if its Jacobian matrix is singular. To obtain a Newton-type system that has a unique solution, people usually refer to some scaling schemes. In what follows we will introduce certain variants of such scaling schemes for SOCO, as first proposed and studied by Tsuchiya [25, 26]. In the rest of this section, we consider the more general case of $N > 1$. In this situation, the definitions $\psi(x)$, $\Psi(x)$ and the Jordan algebra should be modified accordingly as follows:

$$\psi(x) = (\psi(x^1), \psi(x^2), \ldots, \psi(x^N))^T, \qquad \Psi(x) = \sum_{i=1}^{N} \Psi(x^i), \tag{3.2}$$

$$x \circ s = \left(\left(x^1 \circ s^1\right)^T, \left(x^2 \circ x^2\right)^T, \ldots, \left(x^N \circ s^N\right)^T \right)^T. \tag{3.3}$$

Before we discuss these scaling techniques, recall from section 2 that closely associated with each cone K^j are the matrices

$$E_{n_j} := \mathrm{diag}\,(1, 1, \ldots, 1) \quad \text{and} \quad Q^j = \mathrm{diag}\,(1, -1, \ldots, -1),$$

where E_{n_j} denotes the identity matrix in space $\Re^{n_j \times n_j}$ and Q^j is the representation matrix of the cone K^j, since

$$K^j = \{x^j \in \Re^{n_j} : (x^j)^T Q^j x^j \geq 0,\ x_1^j \geq 0\}.$$

It is trivial to see that $(Q^j)^2 = E_{n_j}$.

Now we are ready to give the definition of a scaling matrix for general second-order cones.

DEFINITION 3.1. *A matrix $W^j \in \Re^{n_j \times n_j}$ is a scaling matrix for the cone K^j if it satisfies the following condition:*

$$W^j Q^j W^j = Q^j, \quad W^j \succ 0.$$

We remind the reader that here $W^j \succ 0$ means that W^j is positive definite and symmetric. In view of this definition, if W^j is a scaling matrix for the cone K^j, so is $\left(W^j\right)^{-1}$.

A scaled pair $(\tilde{x}, \tilde{s})$ is obtained by the transformation

$$\tilde{x} = UWx, \qquad \tilde{s} := (UW)^{-1} s,$$

where

$$W := \operatorname{diag}\left(W^1, W^2, \ldots, W^N\right), \qquad U := \operatorname{diag}\left(u_1 E_{n_1}, u_2 E_{n_2}, \ldots, u_N E_{n_N}\right),$$

$$u_1, \ldots, u_N > 0.$$

Several elementary properties of such a transformation are summarized in the following proposition.

PROPOSITION 3.2. *For any $j \in \{1, 2, \cdots, N\}$, we have*

(i) $\mathbf{Tr}\left(x^j \circ s^j\right) = \mathbf{Tr}\left(\tilde{x}^j \circ \tilde{s}^j\right)$;

(ii) $u_j^2 \det\left(x^j\right) = \det\left(\tilde{x}^j\right)$, $u_j^{-2} \det\left(s^j\right) = \det\left(\tilde{s}^j\right)$;

(iii) $x \succ_K 0$ *(or* $x \succeq_K 0$*) if and only if* $\tilde{x} \succ_K 0$ *(or* $\tilde{x} \succeq_K 0$*).*

Proof. The proof follows directly from the definition of scaling matrices. For details, we refer to [4, 25]. □

Let us define

$$\widetilde{A} = A(UW)^{-1}, \qquad \tilde{c} = (UW)^{-T} c.$$

One can rewrite system (1.3) in the scaled space as

$$\begin{cases} \widetilde{A}\tilde{x} = b, \\ \widetilde{A}^T y + \tilde{s} = \tilde{c}, \\ \operatorname{mat}(\tilde{x})\tilde{s} = \mu\tilde{e}, \qquad \tilde{x}, \tilde{s} \succeq_K 0. \end{cases} \tag{3.4}$$

If both x and s are strictly feasible for (SOCO) and (SOCD), so are the vectors $\tilde{x}$ and $\tilde{s}$ for the new SOCO problem in the scaled space. In this case, the linearized Newton system for (3.4) amounts to solving the following equation system:

$$\begin{cases} \widetilde{A}\tilde{d}_x = 0, \\ \widetilde{A}^T d_y + \tilde{d}_s = 0, \\ \operatorname{mat}(\tilde{x})\tilde{d}_s + \operatorname{mat}(\tilde{s})\tilde{d}_x = \mu\tilde{e} - \operatorname{mat}(\tilde{x})\tilde{s}, \qquad \tilde{x}, \tilde{s} \succeq_K 0. \end{cases} \tag{3.5}$$

There are several popular choices for the scaling matrices W^j and the constants u_j. For instance, if UW is the identity matrix and system (3.5) is well defined, then

the solution of (3.5) yields the so-called A.H.O. search direction [1]; if UW is chosen such that $\tilde{s} = \tilde{e}$ (or $\tilde{x} = \tilde{e}$), then we obtain the primal (or dual) H.K.M. direction; if UW is chosen such that $\tilde{x} = \tilde{s}$, then we have the NT search direction [25, 26]. Note that when the NT-scaling is used, since $\tilde{x} = \tilde{s}$, the matrix in the scaled space $\widetilde{A}\mathrm{mat}(\tilde{s})^{-1}\mathrm{mat}(\tilde{s})\widetilde{A}^T = \widetilde{A}\widetilde{A}^T$ is positive definite; thus in this case system (3.5) is well defined. In [13], Monteiro and Tsuchiya studied some other search directions for SOCO as well.

Analogous to [20], in this paper we choose the NT-scaling scheme to define the corresponding proximity and thus the search direction. In what follows we present a variational principle that exhibits an interesting relation between a self-regular proximity function and the NT-scaling. This result is similar to that for SDO in [18].

Let us consider the primal SOCO problem in the scaled space,

$$\begin{aligned} \text{(Scaled SOCO)} \quad & \min \ \tilde{c}^T\tilde{x} \\ & \text{s.t. } \widetilde{A}\tilde{x} = b, \qquad \tilde{x} \succeq_K 0, \end{aligned}$$

and its dual problem

$$\begin{aligned} \text{(Scaled SOCD)} \quad & \max \ b^T y \\ & \text{s.t. } \widetilde{A}^T y + \tilde{s} = \tilde{c}, \quad \tilde{s} \succeq_K 0. \end{aligned}$$

We assume that a certain barrier method is employed to solve both the scaled primal and dual problems; namely, we minimize a specific potential function $\tilde{c}^T\tilde{x} + \Psi(\tilde{x})$ and maximize $b^T y - \Psi(\tilde{s})$, where $\Psi(\cdot)$ is a barrier function for the second-order cone K. The question arises: for which kinds of scaling matrix W and matrix U does the function

$$\tilde{c}^T\tilde{x} - b^T y + \Psi(\tilde{x}) + \Psi(\tilde{s})$$

attain its global minimal value? We have the following.

PROPOSITION 3.3. *Suppose that the functions $\psi(x)$ and $\Psi(\cdot)$ are defined by* (2.8) *and* (3.2). *If the function $\psi(t)$ satisfies condition* C.2 *strictly, then the function $\tilde{c}^T\tilde{x} - b^T y + \Psi(\tilde{x}) + \Psi(\tilde{s})$ attains its global minimal value with matrices W and U such that $\tilde{x} = \tilde{s}$.*

Proof. First we observe that the inner product $\tilde{c}^T\tilde{x}$ is invariant for any nonsingular matrices W and U. Thus we need only to prove that $\Psi(\tilde{x}) + \Psi(\tilde{s})$ has a global minimizer when the matrices W and U are chosen so that $\tilde{x} = \tilde{s}$. The existence of such matrices W and U follows from the choice of the NT-scaling. For any $x, s \succ_K 0$ and $j = 1, 2, \ldots, N$, let us denote by u_j, W^j_{NT} the scaling vector and scaling matrix such that

$$u_j := \left(\frac{\det\left(x^j\right)}{\det\left(s^j\right)}\right)^{\frac{1}{4}}, \tag{3.6}$$

$$w^j := \frac{u_j^{-1}s^j + u_j Q^j x^j}{\sqrt{2}\sqrt{\mathbf{Tr}\left(x^j \circ s^j\right) + \sqrt{\det\left(x^j\right)\det\left(s^j\right)}}}, \tag{3.7}$$

and

$$W^j_{NT} = \begin{pmatrix} w^j_1 & (w^j_{2:n_j})^T \\ w^j_{2:n_j} & E_{n_j-1} + \frac{1}{1+w^j_1} w^j_{2:n_j}(w^j_{2:n_j})^T \end{pmatrix} = -Q^j + \frac{1}{1+w^j_1}(\tilde{e}^j + w^j)(\tilde{e}^j + w^j)^T.$$

Note that when $x, s \succ_K 0$, the denominator in the expression of w^j is positive. For the above choices there holds $\tilde{x}^j = u_j W_{NT}^j x^j = u_j^{-1}(W_{NT}^j)^{-1} s^j = \tilde{s}^j$ (see [4, 25, 26]). For notational convenience, we also define $W_{NT} = \operatorname{diag}(W_{NT}^j)$ and $U_{NT} = \operatorname{diag}(u_1 E_{n_1}, \ldots, u_N E_{n_N})$. Hence it remains to show that for these specific choices of W_{NT} and U_{NT} the value of the function $\Psi(\tilde{x}) + \Psi(\tilde{s})$ is optimal. To distinguish the NT-scaling scheme from many other scaling schemes, we denote by

$$v = \begin{pmatrix} v^1 \\ v^2 \\ \vdots \\ v^N \end{pmatrix} = \tilde{x}_{NT} = \tilde{s}_{NT} = \begin{pmatrix} u_1 W_{NT}^1 x^1 \\ u_2 W_{NT}^2 x^2 \\ \vdots \\ u_N W_{NT}^N x^N \end{pmatrix} = \begin{pmatrix} u_1^{-1}\big(W_{NT}^1\big)^{-1} s^1 \\ u_2^{-1}\big(W_{NT}^2\big)^{-1} s^2 \\ \vdots \\ u_N^{-1}\big(W_{NT}^N\big)^{-1} s^N \end{pmatrix}$$

the scaled vector based on the NT-scaling, while $\tilde{x}$ and $\tilde{s}$ denote the scaled vectors using general scaling techniques. It follows that

$$\Psi(v) = \frac{1}{2}\sum_{j=1}^{N}\big(\psi\big(\lambda_{\min}(v^j)\big) + \psi\big(\lambda_{\min}(v^j)\big)\big).$$

Thus the proof will be finished if we can show that for any $j = 1, \ldots, N$,

$$\begin{aligned}\psi\big(\lambda_{\min}(v^j)\big) + \psi\big(\lambda_{\min}(v^j)\big) \leq\ & \frac{1}{2}\big(\psi\big(\lambda_{\min}(\tilde{x}^j)\big) + \psi\big(\lambda_{\min}(\tilde{x}^j)\big)\big) \\ & + \frac{1}{2}\big(\psi\big(\lambda_{\min}(\tilde{s}^j)\big) + \psi\big(\lambda_{\min}(\tilde{s}^j)\big)\big),\end{aligned} \tag{3.8}$$

and the equality is true if and only if $\tilde{x}^j = \tilde{s}^j$. Now by recalling the definitions of the scaling matrices W_{NT} and U_{NT}, we can conclude that for any $j = 1, \ldots, N$ there holds

$$\mathbf{Tr}\big(x^j \circ s^j\big) = \mathbf{Tr}\big(\tilde{x}^j \circ \tilde{s}^j\big) = \mathbf{Tr}\big(v^j \circ v^j\big) = \mathbf{Tr}\big([v^j]^2\big), \tag{3.9}$$

$$\det\big(v^j\big) = u_j^2 \det\big(x^j\big) = \sqrt{\det\big(x^j\big)\det\big(s^j\big)} = \sqrt{\det\big(\tilde{x}^j\big)\det\big(\tilde{s}^j\big)}. \tag{3.10}$$

Thus the vector v^j satisfies the requirements in the second statement of Proposition 2.9, where x and s are replaced by $\tilde{x}^j$ and $\tilde{s}^j$, respectively. Progressing in a similar vein as we have done in the proof of the second statement of Proposition 2.9, we can obtain the desired relation (3.8), which concludes the proof of the proposition. □

We remark that, as observed by Tsuchiya [26], a large-update IPM for SOCO based on the NT search direction always has a theoretically lower iteration bound than the large-update IPMs relying on other search directions.

3.2. New proximity functions and search directions. To describe our new search direction, we need more notation. Let us denote

$$\bar{A} := \frac{1}{\sqrt{\mu}} A(U_{NT} W_{NT})^{-1}, \qquad v = \frac{1}{\sqrt{\mu}} U_{NT} W_{NT} x, \tag{3.11}$$

$$d_x = \frac{1}{\sqrt{\mu}} U_{NT} W_{NT} \Delta x, \qquad d_s = \frac{1}{\sqrt{\mu}} (U_{NT} W_{NT})^{-1} \Delta s. \tag{3.12}$$

Before introducing the new search direction for SOCO, let us first define the proximity function used in our new IPM for SOCO. Analogously to the LO and SDO cases, the new proximity function for SOCO is given by

$$\Psi(x,s,\mu) := \Psi(v) = \mathbf{Tr}(\psi(v)), \tag{3.13}$$

where $\psi(\cdot)$ is a univariate *self-regular* function.

The new search direction that we propose for SOCO is a slight modification of the NT direction defined by the solution of the following system:

$$\begin{aligned} \bar{A}d_x &= 0, \\ \bar{A}^T\Delta y + d_s &= 0, \\ d_x + d_s &= -\psi'(v). \end{aligned} \tag{3.14}$$

Once we get d_x and d_s, we can compute Δx and Δs via (3.12). In view of the orthogonality of Δx and Δs, one can easily verify that

$$d_x^T d_s = 0. \tag{3.15}$$

We proceed to discuss versatile properties of these *self-regular* proximities for SOCO. For this we need more notation. Let us define

$$\sigma = \sqrt{2}\,\|\psi'(v)\| \tag{3.16}$$

and

$$\begin{aligned} \lambda_{\max}(v) &= \max\{\lambda_{\max}(v^j) : j = 1,\ldots,N\}, \\ \lambda_{\min}(v) &= \min\{\lambda_{\min}(v^j) : j = 1,\ldots,N\}. \end{aligned} \tag{3.17}$$

From Lemma 2.5 we obtain

$$\sigma^2 = \sum_{j=1}^{N}\Big(\big(\psi'(\lambda_{\max}(v^j))\big)^2 + \big(\psi'(\lambda_{\min}(v^j))\big)^2\Big).$$

By using this relation and taking a similar chain of reasoning as in the proof of Proposition 3.3 of [20], we can prove the following results, which include several features of the proximity. These properties are naturally shared by general *self-regular* functions in the second-order cone K.

PROPOSITION 3.4. *Let the proximity $\Psi(v)$ be defined by* (3.13)*, and σ by* (3.16)*. If the kernel function $\psi(\cdot)$ used in the proximity satisfies condition* C.1*, then we have*

$$\Psi(v) \le \frac{\sigma^2}{2\nu_1}, \tag{3.18}$$

$$\lambda_{\min}(v) \ge \left(1 + \frac{q\sigma}{\nu_1}\right)^{-\frac{1}{q}}, \tag{3.19}$$

and

$$\lambda_{\max}(v) \le \left(1 + \frac{p\sigma}{\nu_1}\right)^{\frac{1}{p}}. \tag{3.20}$$

For any $\vartheta > 1$,

$$\Psi(\vartheta v) \leq \frac{\nu_2}{\nu_1}\left(\vartheta^{p+1}\Psi(v) + 2\vartheta\Upsilon'_{p,q}(\vartheta)\sqrt{N\nu_1\Psi(v)} + 2N\nu_1\Upsilon_{p,q}(\vartheta)\right). \tag{3.21}$$

In particular, there exist two constant ν_3 *and* ν_4 *depending only on the kernel function* $\psi(t)$ *such that for any* $\vartheta \in (1, 1+\nu_3]$ *we have*

$$\Psi(\vartheta v) \leq \frac{\nu_2\nu_4}{\nu_1}\Psi(v) + \frac{2\nu_2\nu_4\sqrt{N\nu_1\Psi(v)}}{\nu_1}(\vartheta - 1) + 2N\nu_2\nu_4(\vartheta - 1)^2. \tag{3.22}$$

We close this section by discussing the relations between the duality gap and the proximity. By following a similar chain of reasoning as in the proof of Lemma 2.12 in [20], we can easily deduce

$$\begin{aligned}\frac{\Psi(v)}{\nu_1} &\geq \frac{1}{2}\sum_{j=1}^{N}\left(\left(\lambda_{\max}(v^j) - 1\right)^2 + \left(\lambda_{\min}(v^j) - 1\right)^2\right)\\ &= \|v\|^2 - \sum_{j=1}^{N}\left(\lambda_{\max}(v^j) + \lambda_{\min}(v^j)\right) + N \geq \|v\|^2 - 2\sqrt{N}\,\|v\| + N.\end{aligned}$$

The above relation means that

$$\|v\| \leq \sqrt{N} + \sqrt{\frac{\Psi(v)}{\nu_1}}.$$

It readily follows that

$$\mathbf{Tr}(x \circ s) = 2\mu\,\|v\|^2 \leq 2N\mu + 4\mu\sqrt{\frac{N\Psi(v)}{\nu_1}} + \frac{2\Psi(v)}{\nu_1}\mu. \tag{3.23}$$

Therefore, $\mathbf{Tr}(x \circ s) \leq \mathcal{O}(N\mu)$ holds whenever $\Psi(v) \leq \mathcal{O}(N)$. In such a situation, the proximity plays the role of a potential function for minimizing the duality gap.

4. Polynomial primal-dual algorithms for SOCO.

4.1. The algorithm. The present section describes the new primal-dual algorithm for solving SOCO. First we mention that, by using the NT-scaling, one can rewrite the centrality condition for SOCO as $v = \tilde{e}$. Consequently, the neighborhood of the central path used in our new algorithm is also dependent on v. Define

$$\mathcal{F}_{SOCO} = \{(x, s) \in K \times K : Ax = b;\ A^T y + s = c\}.$$

We define the neighborhood of the central path as follows:

$$\mathcal{N}(\tau, \mu) = \{(x, s) : (x, s) \in \mathcal{F}_{SOCO}, \Psi(x, s, \mu) = \Psi(v) \leq \tau\}. \tag{4.1}$$

Assuming that an initial point in a certain neighborhood of the central path is available (actually, by using the so-called self-dual embedding model, one can further get as an initial point the point on the central path corresponding to $\mu = 1$; see [11, 27]), we can start from this point. By reducing μ properly and solving system (3.14), we obtain a search direction. Then the iterate can be updated by means of line search. If the current iterate goes outside the neighborhood of the targeted point $\mu\tilde{e}$, then we

will utilize the inner iterations to get a new iterate in the neighborhood. Otherwise we progress with the outer iteration and update μ by a fixed factor. The algorithm will stop when the duality gap, bounded by a multiple of μ, is sufficiently small, and hence an approximate solution of the underlying problem is presented. The procedure of the new algorithm is outlined as follows.

PRIMAL-DUAL ALGORITHM FOR SOCO.

Input:
a proximity parameter $\tau \geq \nu_1^{-1}$,
an accuracy parameter $\varepsilon > 0$,
a fixed barrier update parameter $\theta \in (0, 1)$,
a strictly feasible (x, s) and $\mu = 1$ such that $\Psi(x, s, \mu) \leq \tau$.

begin
  while $N\mu \geq \varepsilon$, **do**
  begin
    $\mu := (1 - \theta)\mu$;
    while $\Psi(x, s, \mu) \geq \tau$, **do**
    begin
      Solve the system (3.14);
      compute a step size α;
      $x := x + \alpha\Delta x$,
      $s := s + \alpha\Delta s$,
      $y := y + \alpha\Delta y$.
    end
  end
end

Remark 4.1. The algorithm will stop when an iterate satisfies $N\mu \leq \varepsilon$ and $\Psi(x, s, \mu) \leq \tau$. By recalling (3.23), we can claim

$$x^T s = \frac{1}{2}\mathbf{Tr}(x \circ s) \leq N\mu + 2\mu\sqrt{\frac{N\tau}{\nu_1}} + \mu\frac{\tau}{\nu_1}.$$

For instance, let us choose the parameter $\tau = N$ and the proximity-satisfying condition C.1 with $\nu_1 = 1$. In such a case, the algorithm indeed works in a large neighborhood of the central path. One can easily verify that the algorithm will finally report a solution satisfying $x^T s \leq 4\varepsilon$.

4.2. Complexity of the algorithm. Having stated the algorithm in the previous section, we are going to establish the polynomial complexity of the algorithm in the present section. As we have already observed in [20] for LO and SDO, a crucial step in the estimate of the algorithm's complexity is to evaluate how fast we can reduce the value of the proximity for a feasible step size along the search direction.

Note that once the search direction $(\Delta x, \Delta s)$ is obtained, we need to decide how far we can go along this direction while staying in the feasible region; this amounts to estimating the maximal feasible step size. It should be noticed that for any step size α, the primal-dual pair $(x + \alpha\Delta x, s + \alpha\Delta s)$ is feasible if and only if the scaled primal-dual pair $(v + \alpha d_x, v + \alpha d_s)$ (see Proposition 3.2(iii)) is feasible. In what follows, we give a certain sufficient condition for a step size to be strictly feasible and thus provide a lower bound for the maximal step size. To facilitate the analysis, for any $x^j \in \Re^{n_j}$, $j = 1, \ldots, N$, we define

$$\lambda_{\max}(|x^j|) = |x_1^j| + \|x_{2:n_j}^j\|, \qquad \lambda_{\min}(|\,|x^j|\,|) = |x_1^j| - \|x_{2:n_j}^j\|$$

and

$$\lambda_{\max}(|x|) = \max\{\lambda_{\max}(|x^j|) : j = 1, \ldots, N\},$$

$$\lambda_{\min}(|x|) = \min\{\lambda_{\min}(|x^j|) : j = 1, \ldots, N\}.$$

A direct consequence of the above definitions is

$$\frac{1}{2}\sum_{j=1}^{N}\left(\lambda_{\max}(|x^j|)^2 + \lambda_{\min}(|x^j|)^2\right) = \|x\|^2, \quad x = (x^1, \ldots, x^N)^T. \tag{4.2}$$

Now we have the following result.

LEMMA 4.2. *Let $\alpha_{\max}$ be the maximal feasible step size and*

$$\bar{\alpha} = \lambda_{\min}(v)\sigma^{-1}. \tag{4.3}$$

Then we have

$$\alpha_{\max} \geq \bar{\alpha} \geq \sigma^{-1}\left(1 + \frac{q\sigma}{\nu_1}\right)^{-\frac{1}{q}}.$$

Proof. The proof is similar to that of Lemma 4.7 in [20] for SDO, and thus the details are omitted here. □

In view of Lemma 4.2, it is clear that we can use any $\alpha \in (0, \bar{\alpha})$ as a step size. Note that, after such a step, we get a new primal-dual pair $(x + \alpha\Delta x, s + \alpha\Delta s)$ or the scaled pair $(v + \alpha d_x, v + \alpha d_s)$, and then we need to use the NT-scaling scheme to transform the primal and dual vectors to the same vector, which we denote by v^+. On the other hand, according to (3.13), the proximity after this step is defined as $\Psi(v^+)$. Let us denote the gap between the proximity before and after one step as a function of the step size, that is,

$$g(\alpha) = \Psi(v^+) - \Psi(v). \tag{4.4}$$

The main task in the rest of this section is to study the decreasing behavior of $g(\alpha)$ for $\alpha \in [0, \bar{\alpha})$.

Since v^+ is the vector scaled by using the NT-scaling, from Proposition 3.2 we conclude that for every $j \in \{1, \ldots, N\}$

$$\det\left(((v^+)^j)^2\right) = \det\left(x + \alpha\Delta x\right)\det\left(s + \alpha\Delta s\right) = \det\left(v^j + \alpha d_x^j\right)\det\left(v^j + \alpha d_s^j\right)$$

and

$$\mathbf{Tr}\left(((v^+)^j)^2\right) = \mathbf{Tr}((x + \alpha\Delta x) \circ (s + \alpha\Delta s)) = \mathbf{Tr}\left((v^j + \alpha d_x^j) \circ (v^j + \alpha d_s^j)\right).$$

Thus for any $j = 1, \ldots, N$ the vectors $(v^+)^j, v^j + \alpha d_x^j$, and $v^j + \alpha d_x^j$ satisfy the requirement in the second statement of Proposition 2.9. Therefore, when the kernel function $\psi(\cdot)$ in (3.13) is self-regular, it follows readily from the second statement of Proposition 2.9 that

$$g(\alpha) = \Psi(v^+) - \Psi(v) \leq \frac{1}{2}(\Psi(v + \alpha d_x) + \Psi(v + \alpha d_s)) - \Psi(v) =: g_1(\alpha).$$

In what follows, we estimate the decrement of the function $g_1(\alpha)$ for $\alpha \in [0, \bar{\alpha})$. For our specific purpose, we will first estimate the first and second derivatives of $g_1(\alpha)$. From Lemma 2.10 it follows that

$$g_1'(\alpha) = \frac{1}{2}\mathbf{Tr}(\psi'(v+\alpha d_x)\circ d_x + \psi'(v+\alpha d_s)\circ d_s) \tag{4.5}$$

and

$$g_1''(\alpha) = \frac{1}{2}\frac{d^2}{d\alpha^2}\mathbf{Tr}(\psi(v+\alpha d_x) + \psi(v+\alpha d_s)). \tag{4.6}$$

The next result presents an upper bound for the second-order derivatives of $g_1(\alpha)$. This result is also similar to Lemma 4.8 for SDO in [20], and thus we omit its detailed proof.

LEMMA 4.3. *Suppose that the kernel function $\psi(\cdot)$ used in* (3.13) *is self-regular. Then*

$$g_1''(\alpha) \le \frac{1}{2}\nu_2\sigma^2\big((\lambda_{\max}(v)+\alpha\sigma)^{p-1} + (\lambda_{\min}(v)-\alpha\sigma)^{-q-1}\big) \quad \forall \alpha \in (0, \bar{\alpha}).$$

The remaining discussions in this section follow a very similar procedure as in the LO and SDO cases. First we observe that, by applying Lemma 2.10 to the function $g(\alpha)$, we readily claim

$$g'(0) = g_1'(0) = -\frac{\sigma^2}{2}.$$

From Lemma 4.3 it follows that

$$g(\alpha) \le g_1(\alpha) \le -\frac{\sigma^2\alpha}{2} + \frac{1}{2}\nu_2\sigma^2\int_0^\alpha\int_0^\xi \big((\lambda_{\max}(v)+\zeta\sigma)^{p-1} + (\lambda_{\min}(v)-\zeta\sigma)^{-q-1}\big)d\zeta d\xi,$$

which is essentially the same as its LO analogue (the relation (45) in [20]), where the variables $v_{\max}, v_{\min}$ are replaced by $\lambda_{\max}(v)$ and $\lambda_{\min}(v)$, respectively. Let us define

$$g_2(\alpha) := -\frac{\sigma^2\alpha}{2} + \frac{1}{2}\nu_2\sigma^2\int_0^\alpha\int_0^\xi \big((\lambda_{\max}(v)+\zeta\sigma)^{p-1} + (\lambda_{\min}(v)-\zeta\sigma)^{-q-1}\big)d\zeta d\xi.$$

It is straightforward to verify that $g_2(\alpha)$ is strictly convex and twice differentiable for all $\alpha \in [0, \bar{\alpha})$. Let α^* be the unique global minimizer of $g_2(\alpha)$ in the interval $[0, \bar{\alpha})$, namely,

$$\alpha^* = \arg\min_{0\le\alpha<\bar{\alpha}} g_2(\alpha), \tag{4.7}$$

or, equivalently, α^* is the unique solution of the following equation:

$$-\sigma + \frac{\nu_2}{p}((\lambda_{\max}(v)+\alpha^*\sigma)^p - \lambda_{\max}(v)^p) + \frac{\nu_2}{q}\big((\lambda_{\min}(v)-\alpha^*\sigma)^{-q} - \lambda_{\min}(v)^{-q}\big) = 0. \tag{4.8}$$

For this choice of α^*, by applying Lemma 3.4 of [20], we can readily claim that

$$g(\alpha^*) \le g_2(\alpha^*) \le \frac{1}{2}g_2'(0)\alpha^* = \frac{1}{2}g'(0)\alpha^*. \tag{4.9}$$

Thus it remains to estimate the value of α^*.

LEMMA 4.4. *Let the constant α^* be defined by (4.7). Suppose that $\Psi(v) \geq \nu_1^{-1}$ and $v_{\max} > 1$, and let*

$$\nu_5 = \min\left\{\frac{\nu_1}{2\nu_1\nu_2 + p(\nu_1+2\nu_2)}, \frac{\nu_1^2}{(1+\nu_1)(2\nu_2(\nu_1+q)+\nu_1 q)}\right\}. \tag{4.10}$$

Then

$$\alpha^* \geq \nu_5 \sigma^{-\frac{q+1}{q}} \tag{4.11}$$

holds. In the special case in which $\psi(t) = \Upsilon_{p,q}(t)$ is given by (2.13) with $\nu_1 = \nu_2 = 1$, the above bound simplifies to

$$\alpha^* \geq \min\left(\frac{1}{3p+2}, \frac{1}{4+6q}\right)\sigma^{-\frac{q+1}{q}}. \tag{4.12}$$

Proof. See the proof of Theorem 3.6 of [20]. □

Our next result estimates the decreasing value of the proximity in the case in which the step size α is given by α^* (4.7) or $\alpha = \nu_5\sigma^{(q-1)/q}$. The proof of the theorem is analogous to that of its LO counterpart; thus the details are omitted here.

THEOREM 4.5. *Let the function $g(\alpha)$ be defined by (4.4) with $\Psi(v) \geq \nu_1^{-1}$. Then the step size $\alpha = \alpha^*$ given by (4.7) or $\alpha = \nu_5\sigma^{(q-1)/q}$ is strictly feasible. Moreover, we have*

$$g(\alpha) \leq \frac{1}{2}g'(0)\alpha \leq -\frac{\nu_5\nu_1^{\frac{q-1}{2q}}}{4}\Psi(v)^{\frac{q-1}{2q}}.$$

In the special case of $\psi(t) = \Upsilon_{p,q}(t)$ with $\nu_1 = \nu_2 = 1$, the above bound simplifies to

$$g(\alpha) \leq -\min\left(\frac{1}{12p+8}, \frac{1}{24q+16}\right)\Psi(v)^{\frac{q-1}{2q}}.$$

To get the total complexity result of the algorithm, we still need to describe the growth behavior of the proximity $\Psi(v)$. Suppose that the current point is in the neighborhood $\mathcal{N}(\mu,\tau)$ given by (4.1) and thus the inequality $\Psi(v) \leq \tau$ holds at the present iterate; then we update μ to $(1-\theta)\mu$ for some $\theta \in (0,1)$. By making use of relation (3.21) in Proposition 3.4, one can show that after the update of μ, the proximity is still bounded above by the number $\psi_0(\theta,\tau,2N)$. Here $\psi_0(\theta,\tau,2N)$ denotes the expression at the right-hand side of (3.21), where $\psi(v)$ and ϑ are replaced by τ and $\frac{1}{\sqrt{1-\theta}}$, respectively; i.e.,

$$\psi_0(\theta,\tau,2N) := \frac{\nu_2\tau}{\nu_1(1-\theta)^{\frac{p+1}{2}}} + 2\nu_2\Upsilon'_{p,q}((1-\theta)^{-\frac{1}{2}})\sqrt{\frac{N\tau}{\nu_1(1-\theta)}} + 2N\nu_2\Upsilon_{p,q}((1-\theta)^{-\frac{1}{2}}).$$

The following lemma is an immediate consequence of Lemma 3.9 of [20].

LEMMA 4.6. *Let $\Psi(x,s,\mu) \leq \tau$ and $\tau \geq \nu_1^{-1}$. Then, after an update of the barrier parameter, no more than*

$$\left\lceil \frac{8q\nu_1^{-\frac{q-1}{2q}}}{\nu_5(q+1)}(\psi_0(\theta,\tau,2N))^{\frac{q+1}{2q}} \right\rceil$$

iterations are needed to recenter. In the special case in which $\psi(t) = \Upsilon_{p,q}(t)$ *is given by* (2.13) *with* $\nu_1 = \nu_2 = 1$*, at most*

$$\left\lceil \frac{8q \max(3p+2, 6q+4)}{q+1} (\psi_0(\theta, \tau, 2N))^{\frac{q+1}{2q}} \right\rceil$$

inner iterations are needed to recenter.

Thus the total complexity of the algorithm can be estimated as follows.

THEOREM 4.7. *If* $\tau \geq \nu_1^{-1}$*, the total number of iterations required by the primal-dual Newton algorithm is not more than*

$$\left\lceil \frac{8q\nu_1^{-\frac{q-1}{2q}}}{\nu_5(q+1)} (\psi_0(\theta, \tau, 2N))^{\frac{1+q}{2q}} \right\rceil \left\lceil \frac{1}{\theta} \log \frac{N}{\varepsilon} \right\rceil.$$

In the special case in which $\psi(t) = \Upsilon_{p,q}(t)$ *is given by* (2.13) *with* $\nu_1 = \nu_2 = 1$*, the total number of iterations required by the primal-dual Newton algorithm is less than or equal to*

$$\left\lceil \frac{8q \max(3p+2, 6q+4)}{q+1} (\psi_0(\theta, \tau, 2N))^{\frac{q+1}{2q}} \right\rceil \left\lceil \frac{1}{\theta} \log \frac{N}{\varepsilon} \right\rceil.$$

Neglecting the influence of the constants in the expression in Theorem 4.7, one can safely conclude that for any fixed $\theta \in (0,1)$ with constants $p, q \geq 1$, our large-update algorithm for SOCO in the present section has an $\mathcal{O}(\max\{p,q\} N^{(q+1)/2q} \log \frac{N}{\varepsilon})$ iterations bound, while the algorithm with small-update ($\theta = \mathcal{O}(1/\sqrt{N})$) still stays with the complexity of the $\mathcal{O}(\sqrt{N} \log \frac{N}{\varepsilon})$ iterations bound. Furthermore, from Theorem 4.7 one can readily see that if p is a constant and $q = \log N$, then the new large-update algorithm has an $\mathcal{O}(\sqrt{N} \log N \log \frac{N}{\varepsilon})$ iterations bound.

5. Conclusions. In this paper, we extended the notion of self-regular proximity functions to second-order cones. New IPMs for SOCO based on self-regular proximity functions were introduced, and the complexity results of these algorithms were established. The complexity results of the algorithms in the present paper matched those of their counterparts for LO and SDO in [20]. Therefore, everything relevant to self-regularity, including algorithm, main claims, and complexity analysis, extends to optimization over direct products of nonnegative orthant, second-order, and semidefinite cones in a "componentwise" fashion. However, as demonstrated in section 2, to make such an extension, new analytical tools had to be developed, and this was a nontrivial task.

In closing this paper we would like to point out that, although theoretically our new large-update IPMs have better iteration bound than classical large-update IPMs, much work will be required to test the practical efficiency of the new approach. Regarding this point, the considerable flexibility in choosing the parameters p and q might help us to find new IPMs that are efficient in both theory and practice.

REFERENCES

[1] I. ADLER AND F. ALIZADEH, *Primal-Dual Interior Point Algorithms for Convex Quadratically Constrained and Semidefinite Optimization Problems,* Technical report RRR-111-95, Rutgers Center for Operations Research, New Brunswick, NJ, 1995.

[2] F. Alizadeh, J.P. Haeberly, and M.L. Overton, *A new primal-dual interior-point method for semidefinite programming*, in Proceedings of the 5th SIAM Conference on Applied Linear Algebra, Snowbird, UT, J.G. Lewis, ed., Proc. Appl. Math. 72, SIAM, Philadelphia, 1994, pp. 113–117.

[3] F. Alizadeh and S. Schmieta, *Symmetric cones, potential reduction methods and word-by-word extensions,* in Handbook of Semidefinite Programming (Theory, Algorithms and Applications), H. Wolkowicz, R. Saigal, and L. Vandenberghe, eds., Kluwer Academic Publishers, Boston, 2000, pp. 195–234.

[4] E.D. Andersen, C. Roos, and T. Terlaky, *On Implementing a Primal-Dual Interior-Point Method for Conic Quadratic Optimization,* Technical report, Faculty of Information Technology and System, Delft University of Technology, Delft, The Netherlands, 2000.

[5] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization. Analysis, Algorithms, and Engineering Applications,* MPS-SIAM Ser. Optim. MP02, SIAM, Philadelphia, 2001.

[6] J. Faraut and A. Korányi, *Analysis on Symmetric Cones*, Oxford University Press, London, Oxford, UK, 1994.

[7] L. Faybusovich, *Euclidean Jordan algebras and interior-point algorithms*, Positivity, 1 (1997), pp. 331–357.

[8] L. Faybusovich, *A Jordan-Algebraic Approach to Potential-Reduction Algorithms*, Technical report, Department of Mathematics, University of Notre Dame, Notre Dame, IN, 1998.

[9] M. Fukushima, Z.Q. Luo, and P. Tseng, *Smoothing functions for second-order-cone complementarity problems,* SIAM J. Optim., 12 (2001), pp. 436–460.

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

[11] E. de Klerk, *Interior Point Methods for Semidefinite Programming*, Ph.D. Thesis, Faculty of ITS/TWI, Delft University of Technology, Delft, The Netherlands, 1997.

[12] M.S. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret, *Applications of second-order cone programming*, Linear Algebra Appl., 284 (1998), pp. 193–228.

[13] R.D.C. Monteiro and T. Tsuchiya, *Polynomial convergence of primal-dual algorithms for the second-order cone program based on the MZ-family of directions*, Math. Program., 88 (2000), pp. 61–83.

[14] A.S. Nemirovskii and K. Scheinberg, *Extension of Karmarkar's algorithm onto convex quadratically constrained quadratic programming,* Math. Programming, 72 (1996), pp. 273–289.

[15] Y. Nesterov and A. Nemirovskii, *Interior Point Polynomial Algorithms in Convex Programming,* SIAM Stud. Appl. Math. 13, SIAM, Philadelphia, 1994.

[16] Y.E. Nesterov and M.J. Todd, *Self-scaled barriers and interior-point methods for convex programming,* Math. Oper. Res., 22 (1997), pp. 1–42.

[17] Y.E. Nesterov and M.J. Todd, *Primal-dual interior-point methods for self-scaled cones,* SIAM J. Optim., 8 (1998), pp. 324–364.

[18] J. Peng, *New Design and Analysis of Interior-Point Methods,* Ph.D. thesis, Faculty of Technical Mathematics and Informatics, Delft University of Technology, Delft, The Netherlands, 2001.

[19] J. Peng, C. Roos, and T. Terlaky, *A new class of polynomial primal-dual methods for linear and semidefinite optimization,* European J. Oper. Res., to appear.

[20] J. Peng, C. Roos, and T. Terlaky, *Self-regular proximities and new search directions for linear and semidefinite optimization,* Math. Programming, to appear.

[21] W. Rudin, *Principles of Mathematical Analysis*, MacGraw–Hill, New York, 1978.

[22] S. Schmieta and F. Alizadeh, *Associative and Jordan algebras, and polynomial time interior-point algorithms for symmetric cones*, Math. Oper. Res., 26 (2001), pp. 543–564.

[23] S. Schmieta and F. Alizadeh, *Extension of Primal-Dual Interior-Point Algorithms to Symmetric Cones*, Technical report RRR 13-99, Rutgers Center for Operations Research, Rutgers University, Piscataway, NJ, 1999.

[24] J. Sturm, *Theory and algorithms of semidefinite programming,* in High Performance Optimization, H. Frenk, C. Roos, T. Terlaky, and S. Zhang, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999, pp. 3–194.

[25] T. Tsuchiya, *A Polynomial Primal-Dual Path-Following Algorithm for Second-Order Cone Programming,* Technical report 649, The Institute of Statistical Mathematics, Tokyo, 1997.

[26] T. Tsuchiya, *A convergent analysis of the scaling-invariant primal-dual path-following algorithms for second-order cone programming,* Optim. Methods Softw., 11/12 (1999), pp. 141–182.

[27] H. Wolkowicz, R. Saigal, and L. Vandenberghe, eds., *Handbook of Semidefinite Programming (Theory, Algorithms and Applications)*, Kluwer Academic Publishers, Boston, 2000.

SIAM J. OPTIM.
Vol. 13, No. 1, pp. 204–211

CONJUGATE SCALING ALGORITHM FOR FENCHEL-TYPE DUALITY IN DISCRETE CONVEX OPTIMIZATION*

SATORU IWATA† AND MAIKO SHIGENO‡

Abstract. This paper presents a polynomial time algorithm for solving submodular flow problems with a class of discrete convex cost functions. This class of problems is a common generalization of the submodular flow and valuated matroid intersection problems. The algorithm adopts a new scaling technique that scales the discrete convex cost functions via the conjugacy relation. The algorithm can be used to find a pair of optima in the form of the Fenchel-type duality theorem in discrete convex analysis.

Key words. discrete convexity, submodular flow, scaling algorithm

AMS subject classifications. 49M45, 90C25, 90C35

PII. S1052623499352012

1. Introduction.

The Fenchel-type duality concerning M- and L-convex/concave functions is of fundamental importance in the theory of discrete convex analysis [16, 18, 19]. This paper aims at an algorithmic approach to this duality framework.

Let V be a finite set, and χ_v denote the characteristic vector of $v \in V$. The characteristic vector of $X \subseteq V$ is denoted by χ_X. We write $\mathrm{supp}^+(z) = \{v \mid v \in V, z(v) > 0\}$ and $\mathrm{supp}^-(z) = \{v \mid v \in V, z(v) < 0\}$ for a vector $z \in \mathbf{Z}^V$. For functions $g : \mathbf{Z}^V \to \mathbf{Z} \cup \{+\infty\}$ and $h : \mathbf{Z}^V \to \mathbf{Z} \cup \{-\infty\}$, we denote by $\mathrm{dom}_{\mathbf{Z}}\, g$ and $\mathrm{dom}_{\mathbf{Z}}\, h$ their effective domains, i.e., $\mathrm{dom}_{\mathbf{Z}}\, g = \{x \mid x \in \mathbf{Z}^V, g(x) < +\infty\}$ and $\mathrm{dom}_{\mathbf{Z}}\, h = \{x \mid x \in \mathbf{Z}^V, h(x) > -\infty\}$.

A function $g : \mathbf{Z}^V \to \mathbf{Z} \cup \{+\infty\}$ with nonempty effective domain is said to be M-convex [16, 18, 20] if it satisfies the following:

- $\forall x, y \in \mathrm{dom}_{\mathbf{Z}}\, g$, $\forall u \in \mathrm{supp}^+(x - y)$, $\exists v \in \mathrm{supp}^-(x - y)$ such that

$$g(x) + g(y) \geq g(x - \chi_u + \chi_v) + g(y + \chi_u - \chi_v).$$

It is not difficult to see that the effective domain of an M-convex function forms the set of integral points in a base polyhedron with an integral rank function. A function $h : \mathbf{Z}^V \to \mathbf{Z} \cup \{-\infty\}$ is called M-concave if $-h$ is an M-convex function. M-concave functions generalize matroid valuations invented by Dress and Wenzel [2].

Let $\langle \cdot, \cdot \rangle$ designate the inner product of vectors, i.e., $\langle p, x \rangle = \sum\{p(v)x(v) \mid v \in V\}$. In particular, we denote $x(X) = \langle \chi_X, x \rangle$ for $X \subseteq V$. If g is an M-convex function, $x(V)$ is constant for every $x \in \mathrm{dom}_{\mathbf{Z}} g$.

For a pair of vectors $p, q \in \mathbf{Z}^V$, let $p \vee q$ and $p \wedge q$ denote the vectors defined by $(p \vee q)(v) = \max\{p(v), q(v)\}$ and $(p \wedge q)(v) = \min\{p(v), q(v)\}$, respectively. We also denote by $\mathbf{1}$ the vector in $\mathbf{Z}^V$ with all of its components being equal to one, i.e., the characteristic vector of V.

*Received by the editors February 10, 1999; accepted for publication (in revised form) January 3, 2002; published electronically July 16, 2002. This research was supported in part by a Grant-in-Aid for Scientific Research of the Ministry of Education, Science, Sports and Culture of Japan.
http://www.siam.org/journals/siopt/13-1/35201.html

†Department of Mathematical Engineering and Information Physics, University of Tokyo, Tokyo 113-8656, Japan (iwata@sflab.sys.es.osaka-u.ac.jp).

‡Institute of Policy and Planning Sciences, University of Tsukuba, Tsukuba, Ibaraki 305-8573, Japan (maiko@shako.sk.tsukuba.ac.jp).

A function $f : \mathbf{Z}^V \to \mathbf{Z} \cup \{+\infty\}$ with nonempty effective domain is said to be L-convex [18] if it satisfies the following:

- $\exists r \in \mathbf{Z}\ \forall p \in \mathbf{Z}^V$ such that $f(p + \mathbf{1}) = f(p) + r$;
- $\forall p, q \in \mathbf{Z}^V$, $f(p) + f(q) \geq f(p \vee q) + f(p \wedge q)$.

L-convex functions generalize the Lovász extensions of submodular functions [13]. They are in a close relation to the submodular integrally convex functions of Favati and Tardella [4]; see Fujishige and Murota [8] for this connection. A function $h : \mathbf{Z}^V \to \mathbf{Z} \cup \{-\infty\}$ is called L-concave if $-h$ is an L-convex function.

These two notions of discrete convexity are conjugated to each other. For a function $g : \mathbf{Z}^V \to \mathbf{Z} \cup \{+\infty\}$, we denote by $g^\bullet$ the convex conjugate function defined by

$$g^\bullet(p) = \sup\{\langle p, x\rangle - g(x) \mid x \in \mathbf{Z}^V\} \qquad (p \in \mathbf{Z}^V).$$

The convex conjugate function of an M-convex function is L-convex, and vice versa [18]. The concave conjugate function h° of $h : \mathbf{Z}^V \to \mathbf{Z} \cup \{-\infty\}$ is similarly defined by

$$h^\circ(p) = \inf\{\langle p, x\rangle - h(x) \mid x \in \mathbf{Z}^V\} \qquad (p \in \mathbf{Z}^V).$$

The concave conjugate function of an M-concave function is L-concave, and vice versa. This conjugacy framework is a discrete counterpart of the well-known conjugate duality in convex analysis [24].

Analogously to the Fenchel duality theorem in convex analysis, Murota [16] shows that any pair of an M-convex function g and an M-concave function h satisfies

$$\sup\{h(x) - g(x) \mid x \in \mathbf{Z}^V\} = \inf\{g^\bullet(p) - h^\circ(p) \mid p \in \mathbf{Z}^V\} \tag{1.1}$$

if $g(x) - h(x) \neq +\infty$ for some $x \in \mathbf{Z}^V$, or $g^\bullet(p) - h^\circ(p) \neq +\infty$ for some $p \in \mathbf{Z}^V$. Throughout this paper, we assume M-convex/concave functions to have bounded effective domains, and accordingly, L-convex/concave functions to be bounded. The equality (1.1) always holds in this situation. The original proof by Murota [20] is based on an algorithm that solves submodular flow problems with M-convex cost functions, which we call discrete convex submodular flow problems. The time complexity of this algorithm is pseudopolynomial, i.e., polynomial in the input values but not in the input size. See Fujishige and Murota [8] for an alternative shorter proof of this Fenchel-type duality theorem.

In this paper, we present a polynomial time algorithm for solving the discrete convex submodular flow problem. The new algorithm naturally provides an efficient method for finding both optima in (1.1).

In order to obtain a polynomial time bound, it is now standard to apply the scaling approach. However, a straightforward scaling scheme does not work for M-convex cost functions. For example, a function g', defined by $g'(x) = \lceil g(x)/\alpha \rceil$ for an M-convex function g and a positive integer α, is not necessarily M-convex. Instead, we scale M-convex functions via the conjugacy relation, exploiting the fact that if f is L-convex, then so is f', defined by $f'(p) = f(\alpha p)$.

The outline of this paper is as follows. In section 2, we describe the discrete convex submodular flow problem and its connection to Fenchel-type duality. Section 3 is devoted to a primal-dual algorithm for solving the problem and its continuous version. In section 4, we present the conjugate scaling method, which performs the primal-dual algorithm in each scaling phase.

2. The discrete convex submodular flow problem. Let $G = (V, A)$ be a directed graph with a vertex set V and an arc set A. The initial and terminal vertices of an arc a are denoted by $\partial^+ a$ and $\partial^- a$. For a vertex $v \in V$, we denote by $\delta^+ v$ and $\delta^- v$ the sets of arcs leaving v and entering v, respectively. The boundary $\partial\varphi$ of a function φ on the arc set A is defined by

$$\partial\varphi(v) = \sum_{a \in \delta^+ v} \varphi(a) - \sum_{a \in \delta^- v} \varphi(a) \qquad (v \in V).$$

We denote by n the cardinality of the vertex set V.

With the directed graph $G = (V, A)$ are associated functions $\overline{c} : A \to \mathbf{Z} \cup \{+\infty\}$ and $\underline{c} : A \to \mathbf{Z} \cup \{-\infty\}$ as upper and lower capacities. Let $\gamma : A \to \mathbf{Z}$ be a cost function on the arc set and $g : \mathbf{Z}^V \to \mathbf{Z} \cup \{+\infty\}$ an M-convex cost function such that $x(V) = 0$ for $x \in \mathrm{dom}_{\mathbf{Z}} g$. As a common generalization of the submodular flow problem [3, 6, 7] and the valuated matroid intersection problem [14, 15, 17, 21], Murota [18, 20] addresses the following generalized submodular flow problem with a nonseparable discrete convex cost function, which we call the discrete convex submodular flow problem:

$$\begin{array}{lll} \text{(DCSF)} & \text{minimize} & g(\partial\varphi) + \displaystyle\sum_{a \in A} \gamma(a)\varphi(a) \\ & \text{subject to} & \underline{c}(a) \le \varphi(a) \le \overline{c}(a) \quad (a \in A), \\ & & \partial\varphi \in \mathrm{dom}_{\mathbf{Z}} g, \\ & & \varphi(a) \in \mathbf{Z} \quad (a \in A). \end{array}$$

This is nothing but the submodular flow problem if the M-convex cost function g is constant. Thus there are efficient algorithms [5, 9, 27] for finding a feasible solution, which will be referred to as a feasible flow.

For a vector $p \in \mathbf{Z}^V$, we denote by γ_p the reduced cost function, i.e.,

$$\gamma_p(a) = \gamma(a) + p(\partial^+ a) - p(\partial^- a) \qquad (a \in A).$$

Partition A into $A_p^+ = \{a \mid a \in A,\ \gamma_p(a) > 0\}$, $A_p^\circ = \{a \mid a \in A,\ \gamma_p(a) = 0\}$, and $A_p^- = \{a \mid a \in A,\ \gamma_p(a) < 0\}$. The following theorem of Murota [18, 20] characterizes the optimality for the discrete convex submodular flow problem.

THEOREM 2.1. *A feasible flow $\varphi : A \to \mathbf{Z}$ is optimal if and only if there exists a vector $p \in \mathbf{Z}^V$ that satisfies the following:*

(i) $\forall a \in A_p^-,\ \varphi(a) = \overline{c}(a)$.
(ii) $\forall a \in A_p^+,\ \varphi(a) = \underline{c}(a)$.
(iii) $\partial\varphi \in \arg\min\{g(x) - \langle p, x\rangle \mid x \in \mathbf{Z}^V\}$.

In the rest of this section, we examine the connection between the Fenchel-type duality theorem and the problem (DCSF).

Let V' be a copy of V, and set $W = V \cup V'$. For each $v \in V$, we denote its copy by $v' \in V'$. A vector $\widetilde{y} \in \mathbf{Z}^W$ can be regarded as the direct sum of $y \in \mathbf{Z}^V$ and $y' \in \mathbf{Z}^{V'}$. Given a pair of an M-convex function $g : \mathbf{Z}^V \to \mathbf{Z} \cup \{+\infty\}$ and an M-concave function $h : \mathbf{Z}^V \to \mathbf{Z} \cup \{-\infty\}$, consider an M-convex function $d : \mathbf{Z}^W \to \mathbf{Z} \cup \{+\infty\}$ defined by $d(\widetilde{y}) = g(y) - h(-y')$.

Let $G = (W, A)$ be a directed graph with the arc set $A = \{(v, v') \mid v \in V\}$. The upper and lower capacity functions and the cost functions are given as $\overline{c}(a) = +\infty$, $\underline{c}(a) = -\infty$, and $\gamma(a) = 0$ for each $a \in A$. Then we consider the discrete convex submodular flow problem with the cost function $d(\partial\varphi)$.

Let φ be an optimal flow, and determine $x \in \mathbf{Z}^V$ by restricting $\partial\varphi$ to V. Then the resulting x attains the maximum in the left-hand side of (1.1). Let $\widetilde{p} \in \mathbf{Z}^W$ be a vector that proves the optimality of φ via Theorem 2.1. Then $\widetilde{p}(v) = \widetilde{p}(v')$ holds for each $v \in V$. Restricting $\widetilde{p}$ to V, we obtain a vector $p \in \mathbf{Z}^V$ that achieves the minimum in the right-hand side of (1.1) as well.

As an extension of M-convex functions, Murota and Shioura [22] have introduced $\mathrm{M}^\natural$-convex functions. The effective domain of an $\mathrm{M}^\natural$-convex function is the set of integer points in an integral generalized polymatroid. Fujishige and Murota [8] have extended L-convex functions to $\mathrm{L}^\natural$-convex functions and established the conjugacy relation between $\mathrm{M}^\natural$- and $\mathrm{L}^\natural$-convex functions. The Fenchel-type duality theorem naturally extends to this framework, and the pair of optima can be obtained by solving the discrete convex submodular flow problem constructed in the same way.

3. A primal-dual algorithm. This section introduces a continuous version of problem (DCSF) and presents an algorithm for solving it. The algorithm extends the primal-dual submodular flow algorithm in [1, 12].

We first extend the concept of L-convexity by saying that a function $f : \mathbf{Z}^V \to \mathbf{R} \cup \{+\infty\}$ is L-convex if it satisfies the following:

- $\exists r \in \mathbf{R}$ $\forall p \in \mathbf{Z}^V$ such that $f(p + \mathbf{1}) = f(p) + r$;
- $\forall p, q \in \mathbf{Z}^V$, $f(p) + f(q) \geq f(p \vee q) + f(p \wedge q)$.

We also extend the definition of convex conjugate functions. For a function $f : \mathbf{Z}^V \to \mathbf{R} \cup \{+\infty\}$, the convex conjugate function $f^\bullet : \mathbf{R}^V \to \mathbf{R} \cup \{+\infty\}$ is now defined by

$$f^\bullet(x) = \sup\{\langle p, x\rangle - f(p) \mid p \in \mathbf{Z}^V\} \qquad (x \in \mathbf{R}^V).$$

We denote by $\mathrm{dom}_{\mathbf{R}} f^\bullet$ the effective domain of $f^\bullet$ in $\mathbf{R}^V$. Then $x(V) = r$ for any $x \in \mathrm{dom}_{\mathbf{R}} f^\bullet$ if f is an L-convex function that satisfies $f(p + \mathbf{1}) = f(p) + r$ for every $p \in \mathbf{Z}^V$. See Murota and Shioura [23] for a more general framework of discrete convex functions with continuous variables.

With a directed graph $G = (V, A)$, associate lower and upper capacity functions $\underline{c} : A \to \mathbf{R} \cup \{-\infty\}$ and $\overline{c} : A \to \mathbf{R} \cup \{+\infty\}$ as well as an integral arc cost function $\gamma : A \to \mathbf{Z}$. Let f be an L-convex function that satisfies $f(p + \mathbf{1}) = f(p) + r$ for every $p \in \mathbf{Z}^V$. The following continuous version of (DCSF) will be referred to as $\mathrm{CSF}(f, \gamma)$:

$$\begin{aligned} \text{minimize} \quad & f^\bullet(\partial\varphi) + \sum_{a \in A} \gamma(a)\varphi(a) \\ \text{subject to} \quad & \underline{c}(a) \leq \varphi(a) \leq \overline{c}(a) \quad (a \in A), \\ & \partial\varphi \in \mathrm{dom}_{\mathbf{R}} f^\bullet, \\ & \varphi(a) \in \mathbf{R} \quad (a \in A). \end{aligned}$$

For an integral vector $p \in \mathbf{Z}^V$, let $B_p(f)$ denote a polyhedron defined by

$$B_p(f) = \{x \mid x \in \mathbf{R}^V,\, x(V) = 0,\, \forall X \subseteq V : x(X) \leq f(p + \chi_X) - f(p)\}.$$

Recall the partition of A into $A_p^+ = \{a \mid a \in A,\, \gamma_p(a) > 0\}$, $A_p^\circ = \{a \mid a \in A,\, \gamma_p(a) = 0\}$, and $A_p^- = \{a \mid a \in A,\, \gamma_p(a) < 0\}$. An optimality criterion for $\mathrm{CSF}(f, \gamma)$ is given by the following continuous version of Theorem 2.1.

THEOREM 3.1. *A feasible flow $\varphi : A \to \mathbf{R}$ is optimal if and only if there exists a function $p : V \to \mathbf{Z}$ that satisfies the following:*

(i) $\forall a \in A_p^-,\ \varphi(a) = \overline{c}(a)$.
(ii) $\forall a \in A_p^+,\ \varphi(a) = \underline{c}(a)$.

(iii) $\partial\varphi \in B_p(f)$.

Proof. Note that the "if" part is rather trivial. The "only if" part follows from the validity of the primal-dual algorithm described below.

The primal-dual algorithm repeats the following process for a feasible flow φ and a potential p with $\partial\varphi \in B_p(f)$. Given such φ and p, we denote $D^+_\varphi(v) = \{a \mid v = \partial^+ a,\ a \in A^-_p,\ \varphi(a) < \overline{c}(a)\}$, $D^-_\varphi(v) = \{a \mid v = \partial^- a,\ a \in A^+_p,\ \varphi(a) > \underline{c}(a)\}$, and $D_\varphi(v) = D^+_\varphi(v) \cup D^-_\varphi(v)$ for $v \in V$.

The algorithm picks up a vertex v^* with nonempty $D_\varphi(v^*)$. If no such vertex exists, the current φ and p are optimal. Otherwise, with reference to the new lower and upper capacities defined by

$$\underline{c}^*(a) = \begin{cases} \varphi(a) & (a \in A^-_p), \\ \underline{c}(a) & (a \in A^\circ_p \cup A^+_p), \end{cases} \qquad \overline{c}^*(a) = \begin{cases} \varphi(a) & (a \in A^+_p), \\ \overline{c}(a) & (a \in A^\circ_p \cup A^-_p), \end{cases}$$

the algorithm solves the following maximum submodular flow problem:

$$\begin{array}{lll} \text{(MSF)} & \text{maximize} & \psi(D^+_\varphi(v^*)) - \psi(D^-_\varphi(v^*)) \\ & \text{subject to} & \underline{c}^*(a) \le \psi(a) \le \overline{c}^*(a) \quad (a \in A), \\ & & \partial\psi \in B_p(f). \end{array}$$

A cut for (MSF) means a vertex subset that contains v^*. For each cut S, let $\Delta^+ S$ and $\Delta^- S$, respectively, denote the sets of arcs leaving S and entering S. We now consider the cut capacity

$$\begin{aligned} \kappa_\varphi(S) = {} & \overline{c}^*(\Delta^- S \backslash D^-_\varphi(v^*)) - \underline{c}^*(\Delta^+ S \backslash D^+_\varphi(v^*)) \\ & + \overline{c}^*(D^+_\varphi(v^*) \backslash \Delta^+ S) - \underline{c}^*(D^-_\varphi(v^*) \backslash \Delta^- S) + f(p + \chi_S) - f(p). \end{aligned}$$

Then it follows from [7, Theorem 5.11] that the optimal objective value of (MSF) is equal to the minimum cut capacity $\min\{\kappa_\varphi(S) \mid v^* \in S \subseteq V\}$ unless $D_\psi(v^*)$ becomes empty.

If $D_\psi(v^*)$ is empty, the algorithm updates φ to ψ without changing p. Otherwise, it finds a minimum capacity cut S containing v^*. Since $\psi(S) = f(p + \chi_S) - f(p)$, it follows from $\partial\psi \in B_p(f)$ and Lemma 3.2 below that every $X \subseteq V$ satisfies

$$\begin{aligned} \partial\psi(X) &= \partial\psi(X \cup S) + \partial\psi(X \cap S) - \partial\psi(S) \\ &\le f(p + \chi_{S \cup X}) + f(p + \chi_{S \cap X}) - f(p) - f(p + \chi_S) \\ &\le f(p + \chi_S + \chi_X) - f(p + \chi_S), \end{aligned}$$

which means $\partial\psi \in B_{p+\chi_S}(f)$. Thus the algorithm updates p to $p + \chi_S$, as well as φ to ψ, without violating Theorem 3.1(iii).

The primal-dual algorithm repeats this process until Theorem 3.1(i) and (ii) are satisfied. Note that one iteration reduces by at least one the sum of $\max\{|\gamma_p(a)| \mid a \in D_\varphi(v)\}$ for those vertices with nonempty $D_\varphi(v)$. Since γ_p is integral, the algorithm eventually terminates after a finite number of iterations. Thus Theorem 3.1 has been proved. □

The following easy lemma, to which we have referred in the above argument, will also be used later in section 4. Although it is immediate from the local projected submodularity in [8, Theorem 3], we describe a direct proof here.

LEMMA 3.2. *If f is an L-convex function, then $Y \subseteq Z \subseteq V$ implies*

$$f(p) + f(p + \chi_Y + \chi_Z) \ge f(p + \chi_Y) + f(p + \chi_Z)$$

for any $p \in \mathbf{Z}^V$.

Proof. For $q = p + \chi_Y - \chi_{\overline{Z}}$, we have $p \vee q = p + \chi_Y$ and $p \wedge q = p - \chi_{\overline{Z}}$. Then it follows from the L-convexity of f that

$$\begin{aligned} f(p) + f(p + \chi_Y + \chi_Z) &= f(p) + f(p + \chi_Y - \chi_{\overline{Z}}) + r \\ &\geq f(p + \chi_Y) + f(p - \chi_{\overline{Z}}) + r \\ &= f(p + \chi_Y) + f(p + \chi_Z) \end{aligned}$$

holds for $p \in \mathbf{Z}^V$. □

4. Conjugate scaling. This section presents a cost-scaling framework to solve the discrete convex submodular flow problem (DCSF).

Given a vector $y \in \mathrm{dom}_{\mathbf{Z}} g$, we can efficiently find an integer subgradient of g at y, i.e., a vector $p \in \mathbf{Z}^V$ such that $g(x) - g(y) \geq \langle p, x - y \rangle$ holds for $x \in \mathbf{Z}^V$, by solving a shortest path problem [18, Theorem 4.7]. Thus we henceforth assume without loss of generality that an initial submodular flow φ satisfies in Theorem 2.1(iii) for $p = \mathbf{0}$ by replacing g appropriately.

Let $f_\alpha : \mathbf{Z}^V \to \mathbf{R} \cup \{+\infty\}$ with $\alpha \in \mathbf{Z}$ be an L-convex function defined by

$$f_\alpha(p) = \frac{g^\bullet(\alpha p)}{\alpha} \qquad (p \in \mathbf{Z}^V).$$

Recall here that $g^\bullet$ denotes the convex conjugate function of the M-convex function g. Our cost-scaling algorithm repeatedly applies the primal-dual algorithm concerning f_α with an integer parameter α as follows.

Algorithm Conjugate Scaling.

Step 0: Let φ be an initial feasible flow satisfying $\partial\varphi \in \arg\min\{g(x) \mid x \in \mathbf{Z}^V\}$. Put $p^* \leftarrow \mathbf{0}$, $K \leftarrow \max\{|\gamma(a)| \mid a \in A\}$, and $\alpha \leftarrow 2^{\lceil \log_2 K \rceil}$.

Step 1: Repeat the following (1-1)–(1-4) while $\alpha \geq 1$.

- **(1-1)** $\xi(a) \leftarrow \lceil \gamma(a)/\alpha \rceil$ for $a \in A$.
- **(1-2)** Find an integer vector $p \in \mathbf{Z}^V$ that maximizes $\langle p, \partial\varphi \rangle - f_\alpha(p)$ subject to $2p^* \leq p \leq 2p^* + n\mathbf{1}$.
- **(1-3)** Solve CSF(f_α, ξ) by the primal-dual algorithm starting from φ and p to obtain an optimal flow φ^* and an optimal potential p^*.
- **(1-4)** $\varphi \leftarrow \varphi^*$, $\alpha \leftarrow \alpha/2$.

Recall that the primal-dual algorithm requires initial φ and p with $\partial\varphi \in B_p(f_\alpha)$. We now intend to verify that the integer vector p obtained in Step 1 (1-2) satisfies this condition.

Let q be a minimal integer vector that maximizes $\langle q, \partial\varphi \rangle - f_\alpha(q)$ subject to $q \geq 2p^*$. Denote by d the minimum positive integer that is not equal to $q(v) - 2p^*(v)$ for any $v \in V$, and consider a vertex subset $U = \{u \mid q(u) - 2p^*(u) > d\}$. Note that $q(v) = 2p^*(v)$ holds for some $v \in V$ because $f_\alpha(p) = f_\alpha(p + \mathbf{1})$ for any $p \in \mathbf{Z}^V$.

Lemma 4.1. *The vertex subset U is empty.*

Proof. We first claim

$$f_\alpha(2p^* + 2\chi_U) - f_\alpha(2p^* + \chi_U) \leq f_\alpha(q) - f_\alpha(q - \chi_U). \tag{4.1}$$

Put $\ell = \max\{q(v) - 2p^*(v)\}$, and consider $Y_i = \{v \mid q(v) - 2p^*(v) \geq i\}$ for $i = 1, \ldots, \ell$. We also denote $q_j = 2p^* + \sum_{i=1}^{j} \chi_{Y_i}$ for $j = 0, 1, \ldots, \ell$. Note that $q_0 = 2p^*$, $q_\ell = q$, and $Y_d = Y_{d+1} = U$ hold. Since $Y_i \supseteq U$ for $i = 1, \ldots, d$, Lemma 3.2 implies that $f_\alpha(q_{j-1} + 2\chi_U) - f_\alpha(q_{j-1} + \chi_U) \leq f_\alpha(q_j + 2\chi_U) - f_\alpha(q_j + \chi_U)$ for $j = 1, \ldots, d-1$.

Since $Y_i \subseteq U$ for $i = d, \ldots, \ell$, Lemma 3.2 also implies that $f_\alpha(q_j) - f_\alpha(q_j - \chi_U) \leq f_\alpha(q_{j+1}) - f_\alpha(q_{j+1} - \chi_U)$ for $j = d+1, \ldots, \ell-1$. Thus, by $q_{d-1} + \chi_U = q_d = q_{d+1} - \chi_U$, we obtain (4.1).

The current φ, obtained by the primal-dual algorithm in the previous scaling phase, satisfies $\partial\varphi(U) \leq f_{2\alpha}(p^* + \chi_U) - f_{2\alpha}(p^*) = \{f_\alpha(2p^* + 2\chi_U) - f_\alpha(2p^*)\}/2$. The L-convexity of f_α implies $f_\alpha(2p^* + \chi_U) - f_\alpha(2p^*) = f_\alpha(2p^* + \chi_U + \mathbf{1}) - f_\alpha(2p^* + \mathbf{1}) \leq f_\alpha(2p^* + 2\chi_U) - f_\alpha(2p^* + \chi_U)$. Therefore, $\partial\varphi(U) \leq f_\alpha(2p^* + 2\chi_U) - f_\alpha(2p^* + \chi_U) \leq f_\alpha(q) - f_\alpha(q - \chi_U)$, where the last inequality follows from (4.1). Hence $\langle q, \partial\varphi\rangle - f_\alpha(q) \leq \langle q - \chi_U, \partial\varphi\rangle - f_\alpha(q - \chi_U)$, which contradicts the definition of q unless U is empty. □

As a consequence of Lemma 4.1, we have $q \leq 2p^* + n\mathbf{1}$. Hence the integer vector p obtained in Step 1 (1-2) in fact maximizes $\langle p, \partial\varphi\rangle - f_\alpha(p)$ over $\mathbf{Z}^V$. In particular, $\langle p + \chi_X, \partial\varphi\rangle - f_\alpha(p + \chi_X) \leq \langle p, \partial\varphi\rangle - f_\alpha(p)$ holds for every $X \subseteq V$, which implies $\partial\varphi \in B_p(f_\alpha)$.

We now discuss the time complexity, provided that an evaluation oracle for the M-convex function g is available. The algorithm performs $O(\log K)$ scaling phases. In each scaling phase, f_α is computed in polynomial time by an M-convex function minimization algorithm of Shioura [26]. The maximization problem in Step 1 (1-2) is in fact submodular function minimization over a distributive lattice. Hence it is solvable in polynomial time by the ellipsoid method [10] or recently developed combinatorial algorithms [11, 25]. The number of iterations in the primal-dual algorithm in Step 1 (1-3) is at most $\sum_v \max\{|\gamma_p(a)| \mid a \in D_\varphi(v)\}$, where the summation is taken over those vertices adjacent to arcs violating Theorem 3.1(i) or (ii), and hence bounded by $O(n^2)$. Each iteration solves one maximum submodular flow problem in polynomial time. Thus we have the following theorem.

THEOREM 4.2. *The algorithm* CONJUGATE SCALING *solves the discrete convex submodular flow problem* (DCSF) *in polynomial time.*

5. Conclusion. We have devised a polynomial time algorithm for the discrete convex submodular flow problem by scaling the convex cost function via the conjugacy relation. The resulting algorithm is an extension of the primal-dual algorithm [1, 12]. It may be interesting to know whether other polynomial time submodular flow algorithms extend to this general framework.

Acknowledgments. The authors are grateful to Kazuo Murota for suggesting the scaling approach via the conjugacy relation and for careful reading of the manuscript. Thanks are also due to Tom McCormick and Akiyoshi Shioura for helpful comments on the manuscript.

REFERENCES

[1] W. H. CUNNINGHAM AND A. FRANK, *A primal-dual algorithm for submodular flows*, Math. Oper. Res., 10 (1985), pp. 251–262.
[2] A. W. M. DRESS AND W. WENZEL, *Valuated matroids*, Adv. Math., 93 (1992), pp. 214–250.
[3] J. EDMONDS AND R. GILES, *A min-max relation for submodular functions on graphs*, Ann. Discrete Math., 1 (1977), pp. 185–204.
[4] P. FAVATI AND F. TARDELLA, *Convexity in nonlinear integer programming*, Ricerca Operativa, 53 (1990), pp. 3–44.
[5] A. FRANK, *Finding feasible vectors of Edmonds–Giles polyhedra*, J. Combin. Theory Ser. B, 36 (1984), pp. 221–239.
[6] A. FRANK AND É. TARDOS, *Generalized polymatroids and submodular flows*, Math. Programming, 42 (1988), pp. 489–563.
[7] S. FUJISHIGE, *Submodular Functions and Optimization*, North–Holland, Amsterdam, 1991.

[8] S. Fujishige and K. Murota, *Notes on L-/M-convex functions and the separation theorems*, Math. Program., 88 (2000), pp. 129–146.
[9] S. Fujishige and X. Zhang, *New algorithms for the intersection problem of submodular systems*, Japan J. Indust. Appl. Math., 9 (1992), pp. 369–382.
[10] M. Grötschel, L. Lovász, and A. Schrijver, *The ellipsoid method and its consequences in combinatorial optimization*, Combinatorica, 1 (1981), pp. 169–197.
[11] S. Iwata, L. Fleischer, and S. Fujishige, *A combinatorial strongly polynomial algorithm for minimizing submodular functions*, J. ACM, 48 (2001), pp. 761–777.
[12] S. Iwata, S. T. McCormick, and M. Shigeno, *A fast cost scaling algorithm for submodular flow*, Inform. Process. Lett., 74 (2000), pp. 123–128.
[13] L. Lovász, *Submodular functions and convexity*, Mathematical Programming—The State of the Art, A. Bachem, M. Grötschel, and B. Korte, eds., Springer-Verlag, Berlin, 1983, pp. 235–257.
[14] K. Murota, *Valuated matroid intersection* I*: Optimality criteria*, SIAM J. Discrete Math., 9 (1996), pp. 545–561.
[15] K. Murota, *Valuated matroid intersection* II*: Algorithms*, SIAM J. Discrete Math., 9 (1996), pp. 562–576.
[16] K. Murota, *Convexity and Steinitz's exchange property*, Adv. Math., 124 (1996), pp. 272–311.
[17] K. Murota, *Fenchel-type duality for matroid valuations*, Math. Programming, 82 (1998), pp. 357–375.
[18] K. Murota, *Discrete convex analysis*, Math. Programming, 83 (1998), pp. 313–371.
[19] K. Murota, *Discrete convex analysis* (in Japanese), Discrete Structures and Algorithms V, S. Fujishige, ed., Kindaikagaku-sha, Tokyo, 1998, pp. 51–100.
[20] K. Murota, *Submodular flow problem with a nonseparable cost function*, Combinatorica, 19 (1999), pp. 87–109.
[21] K. Murota, *Matrices and Matroids for Systems Analysis*, Springer-Verlag, Berlin, 2000.
[22] K. Murota and A. Shioura, *M-convex function on generalized polymatroid*, Math. Oper. Res., 24 (1999), pp. 95–105.
[23] K. Murota and A. Shioura, *Extension of M-convexity and L-convexity to polyhedral convex functions*, Adv. Appl. Math., 25 (2000), pp. 352–427.
[24] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[25] A. Schrijver, *A combinatorial algorithm minimizing submodular functions in strongly polynomial time*, J. Combin. Theory Ser. B, 80 (2000), pp. 346–355.
[26] A. Shioura, *Minimization of an M-convex function*, Discrete Appl. Math., 84 (1998), pp. 215–220.
[27] É. Tardos, C. A. Tovey, and M. A. Trick, *Layered augmenting path algorithms*, Math. Oper. Res., 11 (1986), pp. 362–370.

SIAM J. OPTIM.
Vol. 13, No. 1, pp. 212–227

TUBULARITY AND ASYMPTOTIC CONVERGENCE OF PENALTY TRAJECTORIES IN CONVEX PROGRAMMING*

T. CHAMPION†

Abstract. In this paper, we give a sufficient condition for the asymptotic convergence of penalty trajectories in convex programming with multiple solutions. We show that, for a wide class of penalty methods, the associated optimal trajectory converges to a particular solution of the original problem, characterized through a minimization selection principle. Our main assumption for this convergence result is that all the functions involved in the convex program are tubular. This new notion of regularity, weaker than that of quasianalyticity, is defined and studied in detail.

Key words. convex programming, penalty methods, asymptotic analysis, tubular functions

AMS subject classifications. 49M37, 65K10, 90C25, 90C30

PII. S1052623401384771

1. Introduction.

Let us consider a general convex program

$$(\mathrm{CP}_0) \qquad \inf\{\Phi_0(x) \,:\, x \in C\}\,,$$

where Φ_0 is convex and the constraint C is a convex subset of $\mathbb{R}^N$ which can be written in the form

$$C := \left\{x \in \mathbb{R}^N \,:\, \Phi_i(x) \le 0, \quad i = 1, \dots, M\right\}$$

with continuous convex functions Φ_i. In order to handle this kind of constraint, for numerical computations or theoretical study, it has become standard to approximate this problem by means of a penalization method. Given a penalty function $\theta : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$, we associate with (CP_0) a family $(\mathrm{CP}_r)_{r>0}$ of approximating problems given by

$$(\mathrm{CP}_r) \qquad \inf\left\{\Phi_0(x) + \alpha(r)\sum_{i=1}^{M}\theta\left(\frac{\Phi_i(x)}{r}\right) \,:\, x \in \mathbb{R}^N\right\},$$

where $\alpha :]0, +\infty[\to]0, +\infty[$ is a rescaling function. Suitable assumptions (see section 2) on the functions Φ_i, θ, and α guarantee that the optimal values $v(\mathrm{CP}_r)$ converge to $v(\mathrm{CP}_0)$ as r goes to 0. Our work addresses the asymptotic behavior of the net of optimal solutions $(x_r)_{r>0}$ of the approximating problems (CP_r) as r goes to 0. One of the fundamental properties of usual penalty methods is that the penalty trajectory $(x_r)_{r>0}$ is bounded as r goes to 0 and that every cluster point of this net is an optimal solution of the initial problem (CP_0). Here, we are particularly interested in the case in which (CP_0) has more than one solution (so it has infinitely many, since the optimal set is convex). In this case, the penalty trajectory may have several cluster points as r goes to 0; see section 5.3.

*Received by the editors January 9, 2001; accepted for publication (in revised form) December 6, 2001; published electronically July 16, 2002.
http://www.siam.org/journals/siopt/13-1/38477.html

†Laboratoire d'Analyse, de Calcul Scientifique Industriel et d'Optimisation de Montpellier, Département de Mathématiques, case courrier 051, Université Montpellier II, Place Eugène Bataillon, 34095 Montpellier, France (champion@math.univ-montp2.fr).

The convergence of the whole trajectory to a single solution may be of practical interest in numerical computations: when the trajectory does not converge, it may have a bad oscillating behavior as r tends to 0. In linear programming, it is known that the penalty trajectory converges to a particular solution (related to the penalty function used) for some penalty methods. For example, this particular solution is called the analytic center for the logarithmic barrier method (see MacLinden [11], Sonnevend [15], or Auslender, Cominetti, and Haddou [3] for comments) and is also called the absolute minimizer for the exponential penalty (see Cominetti and San Martin [6]). In [3], the authors present a general analysis for the asymptotic convergence of penalty trajectories in linear programming. For more general convex programs, asymptotic convergence to the analytic center for the logarithmic barrier was obtained for analytic functions Φ_i by Monteiro and Zhou [13], while convergence to the absolute minimizer for the exponential penalty was proven for quasianalytic functions Φ_i by Alvarez [1]. Recently, Cominetti [5] proposed a unified approach to this problem of asymptotic convergence for a wide class of penalty functions.

In this work, we use the notion of nonlinear averages introduced in this context in [5]. It allows us to isolate particular solutions of (CP_0), the θ-centers, which are local solutions (in the sense of Dal Maso and Modica [7]) of an auxiliary optimization problem. This notion of a particular solution generalizes those discussed above in linear programming and that defined in [5]. We provide sufficient conditions which ensure that there is a unique θ-center and that the penalty trajectory converges to this optimal solution. The main assumption we make to show the convergence of the penalty trajectory is the tubularity of the functions Φ_i. The notion of tubularity for a function Φ, introduced in section 4, is a regularity condition on the behavior of Φ in the neighborhood of any nontrivial segment on which it is constant. The tubularity condition generalizes that of quasianalyticity and may be of independent interest for the study of convex (or nonconvex) problems with multiple solutions.

In section 2, we recall the basic results and state the assumptions needed in the rest of the paper. In section 3, we define the θ-center of the convex mathematical program (CP_0), while section 4 is devoted to the definition and the study of tubular functions. This notion then allows us to show our main convergence result (Theorem 5.1). Finally, we discuss the possible extension of this convergence result to more general penalty methods and show that when the hypotheses of Theorem 5.1 are not fulfilled, the approximating net (x_r) may fail to converge.

2. Penalty methods in convex programming. Let us consider the convex programming problem (CP_0) given as before by

$$(\mathrm{CP}_0) \qquad \inf\{\Phi_0(x) \,:\, x \in C\}\,,$$

where the feasible set C is a nonempty closed convex subset of $\mathbb{R}^N$ of the form

$$C := \left\{x \in \mathbb{R}^N \,:\, \Phi_i(x) \le 0\,,\ \ i = 1, \dots, M\right\}.$$

In what follows, we will make the following assumptions on the functions Φ_i:

$$(\mathrm{H}_0) \qquad \begin{cases} \Phi_0 : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\} \text{ is a closed proper convex function;} \\ \text{for } 1 \le i \le M,\ \Phi_i : \mathbb{R}^N \to \mathbb{R} \text{ are continuous convex functions;} \\ \text{the set } S(\mathrm{CP}_0) \text{ of the optimal solutions of } (\mathrm{CP}_0) \text{ is } \textit{nonempty} \\ \qquad \text{and } \textit{compact}. \end{cases}$$

We follow [3] and [5] and consider the class of penalty methods for (CP_0) which consist of approximating (CP_0) by the family $(\mathrm{CP}_r)_{r>0}$ of optimization problems

given above, where the positive parameter r is intended to go to 0. We shall assume that the functions $\alpha :]0, +\infty[\to]0, +\infty[$ and $\theta : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ satisfy

$$(\mathrm{H}_1) \qquad \begin{cases} \theta \text{ is increasing and convex on } \operatorname{dom}(\theta) =]-\infty, \eta[\,,\ \eta \in [0, +\infty]; \\ \lim_{t \to \eta^-} \theta(t) = +\infty, \quad \theta_\infty(-1) = 0, \quad \theta_\infty(1) > 0; \\ \lim_{r \to 0^+} \alpha(r) = 0, \quad \theta_\infty(1) \liminf_{r \to 0^+} \alpha(r)/r = +\infty\,. \end{cases}$$

As noted in [3], many penalty methods of the type $(\mathrm{CP}_r)_{r>0}$ with a function θ satisfying (H_1) appear in the literature. We refer to [3] for an extensive list.

Examples. The logarithmic barrier method is obtained for the choice $\theta_1(t) = -\log(-t)$, the inverse barrier method for $\theta_2(t) = -1/t$ (both with $\eta = 0$ and $\alpha(r) = r$), while the exponential penalty method is obtained with $\theta_1(t) = \exp(t)$, $\eta = +\infty$, and $\alpha(r) = r$.

In the case of *interior* penalty methods (when $\eta = 0$), we shall assume that Slater's condition holds; that is, there exists x in $\operatorname{dom}(\Phi_0)$ such that $\Phi_i(x) < 0$ for all i in $\{1, \dots, M\}$.

We recall the following result from [5], which states that the penalty method defined above is a good approximation scheme for solving (CP_0).

THEOREM 2.1. *Suppose that* (H_0) *and* (H_1) *hold. Then for* $r > 0$ *sufficiently small, the optimal set* $S(\mathrm{CP}_r)$ *is nonempty and compact. This holds for any positive* r *if one has* $\theta_\infty(1) = +\infty$. *Moreover, each selection* $x_r \in S(\mathrm{CP}_r)$ *stays bounded as* r *tends to* 0, *any cluster point* x_0 *of such a net* $(x_r)_{r>0}$ *belongs to* $S(\mathrm{CP}_0)$, *and*

$$\lim_{r \to 0} \left[\Phi_0(x_r) + \alpha(r) \sum_{i=1}^{M} \theta\left(\frac{\Phi_i(x_r)}{r} \right) \right] = \inf\{\Phi_0(x) \,:\, x \in C\}\,.$$

Remark. This result also holds when the constraints Φ_i (for $i \in \{1, \dots, M\}$) are assumed to be l.s.c. only on $\mathbb{R}^N$. The continuity assumption for these functions will be revealed only as necessary in the proof of the selection property (see Theorem 5.2).

The following easy lemma ensures the uniqueness of the optimal solution to (CP_r) under a further condition on the functions Φ_i.

LEMMA 2.2. *Suppose that* (H_0), (H_1) *hold and that* θ *is strictly convex. Assume that the functions* Φ_i *are such that*

$$(\mathrm{H}_2) \qquad \text{if the function } z \mapsto (\Phi_0(z), \dots, \Phi_M(z)) \text{ is constant on } [x, y], \text{ then } x = y.$$

Then for every positive r, (CP_r) *has at most one solution.*

Remark. In the case of linear programming, (H_2) is satisfied as soon as the kernel of the linear part of the affine operator $(\Phi_0, \dots, \Phi_M)$ is reduced to $\{0\}$, which is obviously the case when $S(\mathrm{CP}_0)$ is compact.

3. Nonlinear averages and θ-centers. We now want to identify those optimal solutions $x_0 \in S(\mathrm{CP}_0)$ which can be obtained as cluster points (when $r \to 0$) of a selection $x_r \in S(\mathrm{CP}_r)$. Following Cominetti [5], we are thus led to introduce a notion of a particular (or viscosity) solution of (CP_0), the notion of a θ-center. To this end, we first recall the notion of θ-average (see [5]), which can be viewed as an asymptotic nonlinear average for vectors in $\mathbb{R}^m_-$.

PROPOSITION AND DEFINITION 3.1. *Let* $\theta :]-\infty, 0[\to \mathbb{R}$ *be increasing and convex. Then for any* $m \geq 1$ *there exists a unique continuous function* $A^m_\theta : \mathbb{R}^m_- \to \mathbb{R}$, *which we call the* θ-average, *such that for any* $y \in]-\infty, 0[^m$ *one has*

$$A^m_\theta(y) = \limsup_{r \to 0^+} r\theta^{-1}\left(\frac{1}{m} \sum_{i=1}^{m} \theta\left(\frac{y_i}{r} \right) \right).$$

Moreover, A_θ^m is positively homogeneous, convex, symmetric, componentwise nondecreasing, and satisfies

$$\forall y \in]-\infty, 0]^m \qquad \frac{1}{m}\sum_{i=1}^m y_i \leq A_\theta^m(y) \leq \max_{1\leq i\leq m} y_i.$$

We refer to [5] for the proof of this proposition and for further properties of the θ-average.

Examples. The computation of the θ-averages for the examples of the preceding section leads to the following:

$$\begin{array}{lllll}
(1)\,\text{(logarithmic barrier)} & \forall y \in \mathbb{R}^m_- & A^m_{\theta_1}(y) = & -\left(\displaystyle\prod_{1\leq i\leq m}(-y_i)\right)^{\frac{1}{m}}, \\
(2)\,\text{(inverse barrier)} & \forall y \in]-\infty, 0[^m & A^m_{\theta_2}(y) = & \left(\dfrac{1}{m}\displaystyle\sum_{1\leq i\leq m}\frac{1}{y_i}\right)^{-1}, \\
(3)\,\text{(exponential penalty)} & \forall y \in \mathbb{R}^m_- & A^m_{\theta_3}(y) = & \displaystyle\max_{1\leq i\leq m} y_i\,.
\end{array}$$

Notice that the limsup in the definition of the θ-average over $]-\infty, 0[^m$ need not be a limit: it is still an open problem whether it is a limit or not in the general case. However, it is a limit for the above examples, as well as for the other examples of penalty functions θ given in [3]. In the proof of our main result, we will have to assume that this also holds, i.e.,

$$(\mathrm{H}_3) \qquad \forall m \geq 1, \quad \forall y \in]-\infty, 0[^m \qquad A^m_\theta(y) = \lim_{r\to 0^+} r\theta^{-1}\left(\frac{1}{m}\sum_{i=1}^m \theta\left(\frac{y_i}{r}\right)\right).$$

We next define the notion of a θ-center of (CP_0): as Theorem 5.2 shows, this is the viscosity solution (in the sense of [2]) associated with the penalty function θ.

Definition 3.2. *An optimal solution $x_* \in S(\mathrm{CP}_0)$ is a* θ-center *of (CP_0) if for any $J \subset \{1, \ldots, M\}$, x_* is an optimal solution of*

$$(\mathrm{CP}_{0,\theta,J}) \quad \inf\left\{A_\theta^{|J|}((\Phi_j(x))_{j\in J}) \,:\, x \in S(CP_0) \text{ such that } \forall i \notin J\,,\ \Phi_i(x) = \Phi_i(x_*)\right\}.$$

Examples. We give two examples in the setting of linear programming: for every $i \in \{0, \ldots, M\}$, Φ_i is affine and $x \mapsto (\Phi_0(x), \ldots, \Phi_M(x))$ is injective (otherwise $S(\mathrm{CP}_0)$ is not compact).

(4) In the case $\theta = \theta_1$, i.e., for the logarithmic barrier method, there exists a unique θ-center, usually called the analytic center. Indeed, set $I = \{i \,:\, 1 \leq i \leq M\,, \forall x \in S(\mathrm{CP}_0),\ \Phi_i(x) = 0\}$; either $I = \{1, \ldots, M\}$, in which case $S(\mathrm{CP}_0)$ is a singleton (and thus is reduced to the unique θ-center), or $I \neq \{1, \ldots, M\}$, in which case, since the nonlinear average $A_{\theta_1}^{M-|I|}$ is strictly convex on $]-\infty, 0[^{M-|I|}$, the analytic center of (CP_0) is the unique optimal solution of the auxiliary problem

$$\inf\left\{A_{\theta_1}^{M-|I|}((\Phi_i(x))_{i\notin I}) = -\left(\prod_{i\notin I}(-\Phi_i(x))\right)^{\frac{1}{M-|I|}} \,:\, x \in S(\mathrm{CP}_0)\right\}.$$

(5) When $\theta = \theta_3$, i.e., for the exponential penalty method, the θ-center is also called the centroid, or absolute minimizer. Its existence and uniqueness is proven in [6]. In this case, the above definition reads: $x_* \in S(\mathrm{CP}_0)$ is the centroid if, for any $J \subset \{1, \dots, M\}$, x_* is an optimal solution of

$$\inf \left\{ \max_{j \in J} \Phi_j(x) \ : \ x \in S(\mathrm{CP}_0) \text{ such that } \forall i \notin J, \quad \Phi_i(x) = \Phi_i(x_*) \right\}.$$

The notion of θ-center defined above is equivalent to that given in [5] under the more restrictive assumptions therein. With the hypotheses made in [5], the θ-center is shown to exist and be unique, while our hypotheses don't a priori imply either the existence or the uniqueness of a θ-center. However, Theorems 5.1 and 5.2 below ensure the existence of at least one θ-center when the functions Φ_i are tubular. We now give a condition on the functions A_θ^m (for $m \geq 1$) under which the θ-center is uniquely defined.

PROPOSITION 3.3. *Assume that* (H_0), (H_1), (H_2), *and the following* (H_4) *hold for* $m \geq 1$*:*

$$(\mathrm{H}_4) \quad \forall x, y \in \mathbb{R}^m_- \qquad \max_{1 \leq i \leq m} x_i \neq \max_{1 \leq i \leq m} y_i \Rightarrow A_\theta^m \left(\frac{x+y}{2} \right) < \max\{ A_\theta^m(x), A_\theta^m(y) \}.$$

Then (CP_0) *has at most one* θ*-center.*

Proof. By contradiction, suppose that x and y are two distinct θ-centers of (CP_0). Then the set $I \subset \{1, \dots, M\}$ of indices i for which Φ_i is not constant over $[x, y]$ is nonempty; otherwise (H_2) implies that $x = y$. Since x and y are both θ-centers of (CP_0), they are both optimal solutions of

$$(\mathrm{CP}_{0,\theta,I}) \quad \inf \left\{ A_\theta^{|I|}((\Phi_i(z))_{i \in I}) \ : \ z \in S(\mathrm{CP}_0) \text{ such that } \forall j \notin I, \quad \Phi_j(z) = \Phi_j(x) \right\}.$$

Let $z = (x+y)/2$ be the middle of $[x, y]$; then we claim that for any i in I

$$\Phi_i(z) < \max\{ \Phi_i(x), \Phi_i(y) \}. \tag{3.1}$$

By contradiction, assume that (3.1) is false; then, since Φ_i is convex, we obtain

$$\Phi_i(z) = \max\{ \Phi_i(x), \Phi_i(y) \}.$$

We infer from the definition of z that Φ_i is constant on $[x, y]$, which contradicts the definition of I. Hence, one either has $\max_{i \in I} \Phi_i(z) < \max_{i \in I} \Phi_i(x)$ or $\max_{i \in I} \Phi_i(z) < \max_{i \in I} \Phi_i(y)$. Without loss of generality, we assume that the first inequality holds. Then (H_4) yields

$$A_\theta^{|I|}((\Phi_i((x+y)/4))_{i \in I}) < \max\{ A_\theta^{|I|}((\Phi_i(x))_{i \in I}), A_\theta^{|I|}((\Phi_i(z))_{i \in I}) \}.$$

Since $A_\theta^{|I|}$ is convex and $A_\theta^{|I|}((\Phi_i(x))_{i \in I}) = A_\theta^{|I|}((\Phi_i(y))_{i \in I})$, this implies

$$A_\theta^{|I|}((\Phi_i((x+y)/4))_{i \in I}) < A_\theta^{|I|}((\Phi_i(x))_{i \in I}).$$

But for $j \notin I$, one has $\Phi_j((x+y)/4) = \Phi_j(x)$, so the above inequality contradicts the optimality of x for $(\mathrm{CP}_{0,\theta,I})$. ∎

Notice that for the usual penalty functions (e.g., for θ_1, θ_2, and θ_3), hypothesis (H_4) is satisfied, so that under condition (H_2), the θ-center is uniquely determined by Definition 3.2.

4. Tubularity and related notions. In [5], Cominetti shows that when all the functions Φ_i are quasianalytic, and under conditions (H_0)–(H_4) and some more hypotheses on the nonlinear averages A_θ^m, the penalty trajectory $(x_r)_{r>0}$ converges towards the unique θ-center of (CP_0) as r goes to 0. We recall that quasianalyticity is defined as follows.

DEFINITION 4.1. *A function $\Phi : E \to \mathbb{R} \cup \{+\infty\}$ is* quasianalytic *if whenever $x \neq y$ are such that Φ is finite and constant on $[x, y]$, then Φ is constant on the whole line passing through x and y.*

For example, every convex analytic function or strictly convex function is quasianalytic. However, simple functions such as convex piecewise affine functions or finite suprema of quadratic forms are not in general quasianalytic. This motivates the introduction of a weaker property than quasianalyticity for which the convergence of the penalty trajectory towards the θ-center still holds. Before giving the definition of this weaker property, we recall the notion of tubularity for subsets of $\mathbb{R}^N$. In the study of the l^∞-projection on a closed convex subset of $\mathbb{R}^N$, Huotari and Marano were led to define the notion of total tubularity for a convex set (also called property P; see [9], [10] and [12]) as a sufficient condition for the convergence of the Polya algorithm. Here we shall use the term tubular instead of totally tubular.

DEFINITION 4.2. *Let d belong to $\mathbb{R}^N \setminus \{0\}$. A closed convex set C is* d-tubular *if for all x in C such that $x + td$ belongs to C for some positive t there exists a neighborhood V of x in C and a positive ε such that $z + \varepsilon d \in C \ \forall\ z$ in $V \cap C$.*

A closed convex subset C of $\mathbb{R}^N$ is tubular *if it is d-tubular for any d in $\mathbb{R}^N \setminus \{0\}$.*

In terms of local recession vectors (see Definition 6.33 in [14]), the above definition reads as follows: C is d-tubular if d is a local recession vector for C at x whenever $x \in C$ is such that $x + td$ belongs to C for some positive t.

Examples. Any convex polyhedron and any cylinder of convex base in $\mathbb{R}^3$ is tubular. One can also prove that any convex subset of $\mathbb{R}^2$ is tubular (see Proposition 4.5). Notice that simple convex sets may fail to be tubular, as shown in Proposition 4.4 for the convex cone of $\mathbb{R}^{N+1}$, obtained as the epigraph of $x \mapsto \|x\|_{N,2} = (\sum_{i=1}^N x_i^2)^{\frac{1}{2}}$ (for $N \geq 2$). We refer to [10] for further comments and examples.

Let us introduce the sufficient condition for the convergence Theorem 5.1, which is a generalization of the notion of tubularity to functions.

DEFINITION 4.3. *A closed proper function $\Phi : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ is* d-tubular *(with $d \neq 0$) if whenever $x \in \mathbb{R}^N$ and $t > 0$ are such that Φ is finite and constant on $[x, x + td]$, there exists a neighborhood V of x and a positive ε such that for any z in V the function $s \mapsto \Phi(z + sd)$ is nonincreasing on $[0, \varepsilon]$ whenever $\Phi(z) \leq \Phi(x) + \varepsilon$.*

The function Φ is tubular *if it is d-tubular for any $d \neq 0$.*

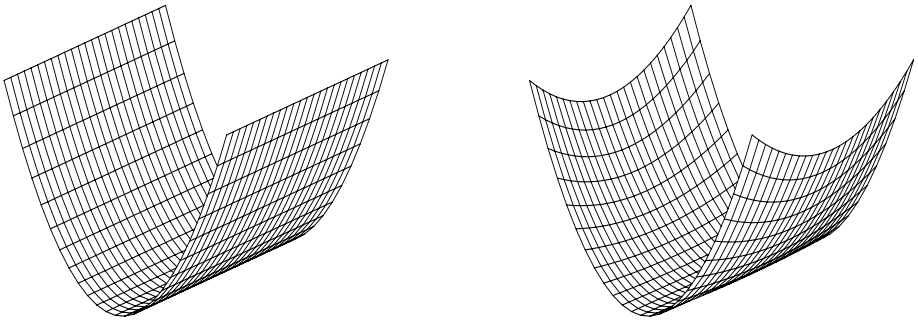
Example. The tubularity property is satisfied by $\Psi_1(x, y) = y^2$ since it is always constant in the direction $(1, 0)$, whereas $\Psi_2(x, y) = (x^2 + 12)y^2$ (which is convex on $[-2, 2] \times \mathbb{R}$) is constant on $[-2, 2] \times \{0\}$ but is increasing on $]0, 2[\times\mathbb{R}^*$ in the direction $(1, 0)$. See Figure 4.1 for an illustration.

The following proposition establishes some links between general tubularity, quasianalyticity, and the tubularity of the epigraph.

PROPOSITION 4.4. i. *If Φ is convex and quasianalytic, then it is tubular.*

ii. *A closed convex set C is d-tubular if and only if its indicator function δ_C is d-tubular.*

iii. *If a continuous convex function Φ is d-tubular, then its epigraph is $(d, 0)$-tubular. Conversely, if the epigraph of a closed convex function Φ is $(d, 0)$-tubular, then Φ is d-tubular.*



$\Psi_1(x,y) = y^2$ is tubular. $\Psi_2(x,y) = (x^2+12)y^2$ is not tubular.

FIG. 4.1.

iv. *A closed proper convex function $\Phi : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ is tubular if and only if whenever Φ is constant on $[x,y]$, there exists a neighborhood V of x and a positive ε such that*

$$\forall z \in \operatorname{dom}(\partial\Phi) \cap V \text{ such that } \Phi(z) \le \Phi(x) + \varepsilon \quad \forall \xi \in \partial\Phi(z), \qquad \langle \xi, y - x \rangle \le 0.$$

v. *For any $N \ge 2$, the continuous convex function $\Psi : x \mapsto \|x\|_{N,2}$ is tubular, but its epigraph is not tubular.*

Proof. i. Suppose that Φ is quasianalytic and that $x \in \mathbb{R}^N$ and $t > 0$ are such that Φ is finite and constant on $[x, x+td]$ for some $d \neq 0$. Then Φ is constant on the line $(x, x+d)$, and since it is convex, Φ is constant on any line $(z, z+d)$ for z in the domain of Φ.

ii. This is straightforward from the definitions.

iii. Let Φ be continuous and d-tubular. Let (x,r) in $\operatorname{epi}(\Phi)$ and $t > 0$ be such that $(x,r) + t(v,0)$ belongs to $\operatorname{epi}(\Phi)$.

Suppose first that there exists a positive s such that $(x,r)+s(d,0)$ is in the interior of $\operatorname{epi}(\Phi)$. Then there exists an open subset V of $\mathbb{R}^{N+1}$ such that $(x,r) + s(d,0) \in V \subset \operatorname{epi}(\Phi)$. Therefore, for any (z,r') in the neighborhood $(V - s(d,0)) \cap \operatorname{epi}(\Phi)$ of (x,r) in $\operatorname{epi}(\Phi)$, one has $(z,r') + s(d,0) \in V \subset \operatorname{epi}(\Phi)$. This shows that $\operatorname{epi}(\Phi)$ is $(d,0)$-tubular in this case.

On the other hand, suppose that $[(x,r),(x,r)+t(d,0)]$ is included in the boundary of $\operatorname{epi}(\Phi)$. This means that $r = \Phi(x)$ and that Φ is constant on the segment $[x, x+td]$. Since Φ is d-tubular, there exist a neighborhood U of x and a positive ε such that, for any z in U for which $\Phi(z) \le \Phi(x) + \varepsilon$, the function $s \mapsto \Phi(z+sd)$ is nonincreasing on $[0,\varepsilon]$. Let U' be a neighborhood of x such that $U' + sd \subset U$ for some positive s. Then for any (z,r') in the neighborhood $U' \times]\Phi(x) - \varepsilon, \Phi(x) + \varepsilon[\cap \operatorname{epi}(\Phi)$ of (x,r) in $\operatorname{epi}(\Phi)$, $(z,r') + s(d,0)$ belongs to $\operatorname{epi}(\Phi)$. This concludes the proof of the first part of the claim.

Suppose now that the epigraph of Φ is $(d,0)$-tubular for some $d \neq 0$. Let x and $t > 0$ be such that Φ is finite and constant on $[x, x+td]$. Then $(x, \Phi(x))$ and $(x+td, \Phi(x))$ belong to $\operatorname{epi}(\Phi)$, so there exists a neighborhood V of x and a positive ε such that

$$\forall z \in V \text{ such that } \Phi(z) \in]\Phi(x) - \varepsilon, \Phi(x) + \varepsilon[, \qquad (z + \varepsilon d, \Phi(z)) \in \operatorname{epi}(\Phi).$$

Let $\eta > 0$ be such that $B(x,\eta) \subset V \cap \{y : \Phi(y) > \Phi(x) - \varepsilon\}$. Then it is easy to check

that for every z in $B(x,\eta)$ such that $\Phi(z) \le \Phi(x) + \varepsilon$, the function $s \mapsto \Phi(z+sd)$ is nonincreasing on $[0,\varepsilon]$.

iv. This part is a straightforward consequence of Proposition 8.50 (equivalence between (d) and (f)) of [14].

v. The function Ψ is tubular since it is never constant on a nontrivial segment. To show that its epigraph is not tubular, we use characterization iv above to prove that its indicator function is not tubular. We are then led to show that there exist x and y in $\mathrm{epi}(\Psi)$ such that for any neighborhood V of x one has

$$\exists z \in \mathrm{epi}(\Psi) \cap V, \quad \exists \xi \in N_{\mathrm{epi}(\Psi)}(z), \quad \langle \xi, y-x\rangle > 0.$$

Suppose that $N \ge 2$ and $x \ne 0$. We notice that $N_{\mathrm{epi}(\Psi)}(x,\Psi(x)) = \mathbb{R}_+\big(\frac{x}{\|x\|_2}, -1\big)$. As the segment $[(x,\Psi(x)),(0,0)]$ is included in $\mathrm{epi}(\Psi)$, one is led to consider the expression

$$\left\langle \left(\frac{z}{\|z\|_2}, -1\right), (-x, -\Psi(x))\right\rangle = -\frac{1}{\|z\|_2}\langle z, x\rangle + \|x\|_{N,2}$$

for $z \ne 0$ in a neighborhood of x. The Cauchy–Schwarz inequality implies that this is positive for any z which is not colinear to x. Such a z exists in any neighborhood of x since $N \ge 2$; thus the epigraph of Ψ is not tubular. □

Remark. We deduce from Proposition 4.4ii and iv the following characterization of tubular sets: a closed convex set C is tubular if and only if whenever x and y belong to C, there exists a neighborhood V of x such that

$$\forall z \in C \cap V, \quad \forall \xi \in N_C(z) \qquad \langle \xi, y-x\rangle \le 0.$$

We now prove that, in low dimension, every closed convex subset of $\mathbb{R}^N$ is tubular. Notice that the result below is sharp since Proposition 4.4 provides an example of a nontubular convex subset of $\mathbb{R}^N$ for any $N \geqslant 3$.

PROPOSITION 4.5. *Every closed convex subset of $\mathbb{R}$ and $\mathbb{R}^2$ (or of any vector space of dimension less than two) is tubular. As a consequence, any l.s.c. proper convex function $\Phi : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ is tubular.*

Proof. For $N = 1$, the proof is straightforward since a convex subset of $\mathbb{R}$ is an interval.

Let C be a closed convex subset of $\mathbb{R}^2$, and let x, y belong to C. Without loss of generality, we may assume that $x = (0,0)$ and $y = (1,0)$. We first show that $C \cap \mathbb{R} \times [0,+\infty[$ is tubular.

If $C \cap \mathbb{R}\times\,]0,+\infty[$ is empty, there is nothing to prove (it reduces to the one-dimensional case). On the contrary, assume that some $z = (z_1, z_2)$ belongs to this set. Let δ denote the distance from x to the line (y,z), and set $\varepsilon = \min(\delta, z_2)/2$. We claim that for any v in $C \cap \mathbb{R} \times [0,+\infty[\cap B(x,\varepsilon)$, $v + \varepsilon(y-x)$ belongs to C. Indeed, we infer from the definition of ε that there exists $t_0 > \varepsilon$ such that $v + t_0(y-x)$ belongs to $[z,y]$. Since C is convex, $v + \varepsilon(y-x)$ belongs to C. This proves our claim.

The same arguments yield a positive η such that for any v in $C \cap \mathbb{R}\times]-\infty,0] \cap B(x,\eta)$, $v + \eta(y-x)$ belongs to C. We now set $\gamma = \min(\varepsilon,\eta)$; then the neighborhood $V = B(x,\gamma)$ of x and the positive γ are such that for any $z \in C \cap B(x,\gamma)$, $z + t(y-x)$ belongs to C. This completes the proof.

We may apply characterization iii of Proposition 4.4 to get that any closed proper convex function $\Phi : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ is tubular, but we can also deduce this directly from the definition. Indeed, assume that Φ is constant and finite on $[x, x+td]$ (with

$d \neq 0$ and $t > 0$). Without loss of generality, we may assume that $d > 0$; then, since Φ is convex, one has

$$\forall s < r \leq t \qquad \frac{\Phi(x+rd) - \Phi(x+sd)}{(r-s)d} \leq \frac{\Phi(x+td) - \Phi(x)}{td} = 0,$$

so that Φ is nonincreasing on $]-\infty, x+td]$. □

Remark. Notice that in the proof of the tubularity of a closed convex subset C of $\mathbb{R}^2$, the trick is to write C as the union of its intersection with the two half planes $\mathbb{R}\times[0,+\infty[$ and $\mathbb{R}\times]-\infty,0]$. In $\mathbb{R}^3$ one would need infinitely many half planes, which is the reason why this proof can't be adapted to dimensions higher than 3.

It is noticed in Proposition 4.4v that the Euclidean norm $x \mapsto \|x\|_2$ is tubular. This may be checked using calculus, but it is also a consequence of the strict convexity of its sublevels, as the next proposition shows. We also prove below that the set of proper tubular convex functions is stable under composition by an increasing convex function.

PROPOSITION 4.6. i. *If $\Phi : \mathbb{R}^N \to \mathbb{R} \cup +\infty$ is a closed proper convex function whose sublevel sets are (void or) strictly convex, then it is tubular.*

ii. *If $\Phi : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ is a closed proper tubular convex function and $\theta : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ is an increasing convex function such that $\theta \circ \Phi \not\equiv +\infty$, then $\theta \circ \Phi$ is tubular.*

Proof. i. Suppose that $\Phi : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ is a closed proper convex function whose level sets are strictly convex, and assume that Φ is finite and constant over $[x, y]$, with $x \neq y$. As the set $\{z \in \mathbb{R}^N : \Phi(z) \leq \Phi(x) = \Phi(y)\}$ is strictly convex, there exists an open ball B centered at the middle $(x+y)/2$ of $[x, y]$ included in this set. We claim that the neighborhood $B - (y-x)/2$ of x and $\varepsilon = 1/2$ have the desired property. To show this, we first notice that

$$\Phi(x) = \Phi(y) = \Phi\left(\frac{x+y}{2}\right) = \min\{\Phi(z) : z \in \mathbb{R}^N\}.$$

Otherwise, there exists $z \in \mathbb{R}^N$ such that $\Phi(z) < \Phi(x)$. If we set $h(t) = \Phi(tz + (1-t)(x+y)/2)$, then h is convex and we have

$$h(1) < h(0) \qquad \text{and} \qquad h(t) \leq h(0) = \Phi(x)$$

for any negative t such that $tz + (1-t)(x+y)/2$ belongs to B. This obviously contradicts the convexity of h.

Let z belong to $B-(y-x)/2$; then the function $s \mapsto \Phi(z+s(y-x))$ is convex on $[0, \varepsilon]$ and attains its minimum at ε, so that it is nonincreasing on $[0, \varepsilon]$. This concludes the proof.

ii. Assume that $\Phi : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ is a closed proper tubular convex function and $\theta : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ is an increasing convex function. If $\theta \circ \Phi$ is finite and constant on a segment $[x, x+td]$, then since θ is increasing, Φ is also constant on $[x, x+td]$. As a consequence, there exists a neighborhood V of x and a positive ε such that for any $z \in V$ the function $s \mapsto \Phi(z+sd)$ is nonincreasing on $[0, \varepsilon]$ whenever $\Phi(z) \leq \Phi(x) + \varepsilon$. Since θ is increasing, $s \mapsto \theta \circ \Phi(z+sd)$ is also nonincreasing on $[0, \varepsilon]$ for any z in V such that $\theta \circ \Phi(z) \leq \theta(\varepsilon + \Phi(x))$. Therefore $\theta \circ \Phi$ is tubular. □

The following proposition provides different ways to build tubular sets and functions.

PROPOSITION 4.7. i. *If $C_1, \ldots, C_M$ are closed tubular convex subsets of $\mathbb{R}^N$, then $\bigcap_{i=1}^M C_i$ is tubular.*

ii. *If $\Phi_1, \dots, \Phi_M$ are continuous tubular convex functions, then* $\sup_{1\le i\le M} \Phi_i$ *is tubular.*

iii. *Let $\Phi_1, \dots, \Phi_M : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ be l.s.c. proper convex functions. Then the l.s.c. proper convex function $(x_1, \dots, x_M) \longmapsto \sum_{i=1}^M \Phi_i(x_i)$ is tubular on $\mathbb{R}^M$.*

Proof. i. This is a simple consequence of Definition 4.2.

ii. Let $\Phi_1, \dots, \Phi_M$ be continuous convex functions, and set $\Phi = \sup_i \Phi_i$; then $\mathrm{epi}(\Phi) = \bigcap_i \mathrm{epi}(\Phi_i)$. From Proposition 4.4iii, we know that each $\mathrm{epi}(\Phi_i)$ is $(d, 0)$-tubular for every $d \neq 0$. It is easily checked that this also holds true for the finite intersection $\bigcap_i \mathrm{epi}(\Phi_i)$. Applying Proposition 4.4 again yields that Φ is tubular.

iii. We limit ourselves to the case $M = 2$, the proof being easily adapted to the general case. We infer from Proposition 4.5 that, as Φ_1 and Φ_2 are proper l.s.c. convex functions on $\mathbb{R}$, they are tubular.

We set $\Psi(x, y) = \Phi_1(x) + \Phi_2(y)$. Assume that Ψ is constant on $[(x, y), (x, y) + t(d_1, d_2)]$ for some positive t. We must find a neighborhood V of (x, y) and a positive ε such that for any $z \in V$ the function $s \mapsto \Psi(z + s(d_1, d_2))$ is nonincreasing on $[0, \varepsilon]$ whenever $\Psi(z) \le \Psi(x, y) + \varepsilon$. If $d_1 = d_2 = 0$, there is nothing to prove.

Suppose first that $d_1 = 0$ or $d_2 = 0$. Without loss of generality, we assume that the first holds. Then Φ_2 is constant on $[y, y + td_2]$, so that there exists a neighborhood V of y and a positive ε such that $s \mapsto \Phi_2(z + sd_2)$ is nonincreasing on $[0, \varepsilon]$ whenever $z \in V$ and $\Phi_2(z) \le \Phi_2(y) + \varepsilon$. Recalling the definition of Ψ, it is easy to check that for any (z_1, z_2) in the neighborhood $\{x' : \Phi_1(x') > \Phi_1(x) - \varepsilon/2\} \times V$ of (x, y) and such that $\Psi(z_1, z_2) \le \Psi(x, y) + \varepsilon/2$, the function $s \mapsto \Psi(z + s(0, d_2))$ is nonincreasing on $[0, \varepsilon/2]$. This concludes the proof in this case.

Now, suppose that $d_1 \neq 0$ and $d_2 \neq 0$. Without loss of generality, we may assume that both d_1 and d_2 are positive. Since $\Phi_1(x + std_1) = \Psi(x, y) - \Phi_2(y + std_2)$ for s in $[0, 1]$, we deduce that Φ_1 is linear on $[x, x + td_1]$. By symmetry, Φ_2 is also linear on $[y, y + td_2]$. As a consequence, Ψ is constant on any segment included in $[x, x + td_1] \times [y, y + td_2]$ with direction (d_1, d_2). We claim that the neighborhood $V =]x - \frac{d_1}{4}, x + \frac{d_1}{4}[\times]y - \frac{d_2}{4}, y + \frac{d_2}{4}[$ of (x, y) and $\varepsilon = 1/2$ have the desired property. Indeed, let (z_1, z_2) belong to V; then the convex function $s \mapsto \Psi((z_1, z_2) + s(d_1, d_2))$ is constant on $[1/4, 3/4]$. Reasoning as in the proof of Proposition 4.5, we infer that this function is nonincreasing on $]-\infty, 1/2]$. This completes the proof. □

COROLLARY 4.8. *Any closed convex polyhedron as well as any convex piecewise affine function is tubular. Moreover, every finite* sup *of convex analytic functions on $\mathbb{R}^N$ is tubular.*

Remark. The notion of tubularity is not stable with respect to the addition. Indeed, the functions $\Phi_1(x, y) = (x^2 + 12)y^2 + y$ (which is never constant on a nontrivial segment of $\mathbb{R}^2$) and $\Phi_2(x, y) = -y$ are convex and tubular on $[-2, 2] \times \mathbb{R}$, but their sum is no longer tubular (see the example after Definition 4.3).

5. Asymptotic convergence to the θ-center.

5.1. The convergence result. We turn to the main result of this paper, which is a generalization of previous similar results in [3] and [5] on the selection of a particular solution of (CP_0) by a penalty method.

THEOREM 5.1. *Assume that* (H_0)–(H_4) *hold and that the function Φ_i is tubular for any i in $\{0, \dots, M\}$. Then the net $(x_r)_{r>0}$ of the optimal solutions of the penalized problems* (CP_r) *converges as r tends to 0 towards the unique θ-center of* (CP_0).

Remark. To use this selection result, it would be of practical interest to be able to associate with any given function A^m (having the properties of nonlinear averages

described in Definition 3.1) a penalty function θ such that $A^m = A^m_\theta$. It is still an open problem whether this is possible in the general case. For example, for the function $A^m(x) := -(\frac{1}{m}\sum_{i=1}^m (-x_i)^p)^{\frac{1}{p}}$ (where $p \in]0,1[$), one may choose $\theta_p(t) := -(-t)^p$.

Before proving Theorem 5.1, we illustrate it with two examples. In the case of the logarithmic penalty method, MacLinden [11] shows the convergence of the penalty trajectory towards the analytic center (i.e., the θ-center when $\theta = \theta_1$ is the log penalty) under the strict complementarity assumption. In the following example, this assumption does not hold, while Theorem 5.1 ensures the convergence of the penalty trajectory.

Example. Set $\Phi_0(x,y) = x^2 + \max\{0, -1-y\}$, $\Phi_1(x,y) = x$, and $\Phi_2(x,y) = y$. Then (H_0)–(H_4) are fulfilled (with $\theta = \theta_1$ being the log penalty function), and the functions Φ_i are tubular (see Proposition 4.7), so that the penalty trajectory converges to the unique θ-center $(0,-1)$. However, every optimal solution of problem (CP_0) is of the form $(0,\alpha)$ for α in $[-1,0]$, and the unique solution of the dual problem is $(0,0)$, so that the strict complementarity assumption does not hold.

When (H_0)–(H_4) hold, the penalty trajectory may converge to an optimal solution different from the θ-center if the functions Φ_i are not tubular (as the following example shows), or may not converge at all (see section 5.3).

Example. Set $\Phi_0(x,y) = (x^2+3)y^2 + \delta_C(x,y)$ with $C = [-1,1]\times\mathbb{R}$, $\Phi_1(x,y) = x-6$, and $\Phi_2(x,y) = y$. Then hypotheses (H_0)–(H_4) are fulfilled (with $\theta = \theta_1$), but Φ_0 is not tubular on C. Let (x_r, y_r) denote the unique solution of (CP_r); then for $r > 0$ sufficiently small, (x_r, y_r) belongs to the interior of C, and the optimality conditions read

$$2x_r y_r^2 = \frac{r}{x_r - 6} \qquad \text{and} \qquad 2y_r(x_r^2+3) = \frac{r}{y_r}.$$

As a consequence, $(x_r, y_r)_r$ converges to $(-\frac{1}{2}, 0)$, whereas the θ-center is $(-1,0)$.

Theorem 5.1 is, in fact, a straightforward consequence of the following selection result. Theorem 5.2 below characterizes those optimal solutions of (CP_0) which can be obtained as limit points (as $r \to 0$) of nets $(x_r)_r$ of optimal approximate solutions. Notice that when (H_4) does not hold, there may be several θ-centers, so the following theorem does not imply the convergence of the penalty trajectories.

THEOREM 5.2. *Suppose that* (H_0), (H_1), *and* (H_3) *hold. Also assume that the function* Φ_i *is tubular for any* i *in* $\{0,\dots,M\}$. *Let* $x_0 \in S(CP_0)$ *be a cluster point of* $(x_r)_{r>0}$ *as* r *goes to* 0, *where* $x_r \in S(CP_r)$ $\forall$ r. *Then* x_0 *is a* θ*-center of* (CP_0).

We shall need the following lemma in the proof of Theorem 5.2.

LEMMA 5.3. *Let* x_0 *and* x^* *belong to* $S(CP_0)$. *Then there exists* $t \in]0,1[$ *such that for any* i *in* $I = \{1 \le i \le M : \Phi_i(x_0) = \Phi_i(x^t)\}$ *the function* Φ_i *is constant on* $[x_0, x^t]$, *where we have set* $x^t := tx_0 + (1-t)x^*$.

Proof. For $1 \le i \le M$, we set $s_i = \max\{s \in [0,1/2] : \Phi_i(x_0) = \Phi_i(x^s)\}$, where x^s denotes $sx_0 + (1-s)x^*$. Notice that if $s \in]0, s_i[$ is such that $\Phi_i(x^s) = \Phi_i(x_0)$, then Φ_i is constant on $]0, s_i[$.

If for every $1 \le i \le M$ one has $s_i = 0$, then $t = 1/4$ has the desired property. Indeed, $I = \{1 \le i \le M : \Phi_i(x_0) = \Phi_i(x^t)\}$ is empty for this choice of t.

Otherwise, we set $t = 1/2\min\{s_i : s_i > 0\}$. Then if i belongs to $I = \{1 \le i \le M : \Phi_i(x_0) = \Phi_i(x^t)\}$, the function Φ_i is constant on $]x_0, x^t[$ since $t \in]0, s_i[$. This concludes the proof. □

Proof of Theorem 5.2. Let $x_0 \in S(CP_0)$ be a cluster point of $(x_r)_{r>0}$ as r goes to 0, where $x_r \in S(CP_r)$ $\forall$ r. To simplify the notations, we assume that the whole net

converges to x_0 as r tends to 0. Let x^* belong to $S(\mathrm{CP}_0)$ and $I \subset \{1,\dots,M\}$ such that $\forall\ i$ in I, $\Phi_i(x_0) = \Phi_i(x^*)$. Then we must check that

$$A_\theta^{M-|I|}((\Phi_j(x_0))_{j\notin I}) \le A_\theta^{M-|I|}((\Phi_j(x^*))_{j\notin I}).$$

We apply Lemma 5.3 to get a real $t \in]0,1[$ for which the function Φ_j is constant on $[x_0, x^t]$ for every $j \in J = \{1 \le i \le M\ :\ \Phi_i(x_0) = \Phi_i(x^t)\}$, where we have set $x^t = tx_0 + (1-t)x^*$. Notice that as x_0 and x^t both belong to $S(\mathrm{CP}_0)$, the function Φ_0 is constant on the segment $[x_0, x^t]$. The optimality condition for x_r reads

$$0 \in \partial\Phi_0(x_r) + \alpha(r)\sum_{i=1}^{M} \partial\left(\theta\left(\frac{\Phi_i(.)}{r}\right)\right)(x_r).$$

Since the functions Φ_0 and $\theta(\Phi_j(.)/r)$ are tubular (see Lemma 4.6) and constant over $[x_0, x^t]$ for j in J, we infer from Proposition 4.4iv that for r small enough,

$$\sum_{i\notin J}\langle \xi_i^r, x^t - x_0\rangle \ge 0$$

for some vectors ξ_i^r in $\partial(\theta(\Phi_i(.)/r))(x_r)$. We deduce from the previous inequality and the convexity of the functions $\theta(\Phi_i(.)/r)$ that

$$\sum_{i\notin J}\theta\left(\frac{\Phi_i(x_r)}{r}\right) \le \sum_{i\notin J}\theta\left(\frac{\Phi_i(x_r + x^t - x_0)}{r}\right).$$

We notice that for i in $I \setminus J$ we have $\Phi_i(x^t) < \Phi_i(x^*) = \Phi_i(x_0)$. As a consequence, for r small enough, we deduce from the strict monotonicity of θ and the continuity of the functions Φ_i that

$$\sum_{i\notin J, i\notin I}\theta\left(\frac{\Phi_i(x_r)}{r}\right) \le \sum_{i\notin J, i\notin I}\theta\left(\frac{\Phi_i(x_r + x^t - x_0)}{r}\right).$$

We can now choose real numbers δ_i and δ_i^t such that $\delta_i < \Phi(x_0) \le 0$ and $\Phi_i(x^t) < \delta_i^t < 0$ for $i \notin (I \cup J)$. Indeed, for $i \notin (I\cup J)$, $\Phi_i(x^t)$ belongs to $]-\infty, \max\{\Phi_i(x_0), \Phi_i(x^*)\}[$, so that $\Phi_i(x^t) < 0$. For $i \in J \setminus I$, we notice that $\Phi_i(x_0) = \Phi_i(x^t) \le \Phi_i(x^*)$, and we choose $\delta_i = \delta_i^t < \Phi_i(x_0)$. With these notations, we conclude from the monotonicity of θ that

$$\sum_{i\notin I}\theta\left(\frac{\delta_i}{r}\right) \le \sum_{i\notin I}\theta\left(\frac{\delta_i^t}{r}\right).$$

Notice that δ and δ^t both belong to $]-\infty, 0[^{M-|I|}$. We thus divide the above inequality by $M - |I|$, compose it by the increasing function θ^{-1}, then divide by r and let r go to 0 (that's where (H_3) is needed). This leads to

$$A_\theta^{M-|I|}(\delta) \le A_\theta^{M-|I|}(\delta^t).$$

Then, letting δ_i (resp., δ_i^t) go to $\Phi_i(x_0)$ (resp., $\Phi_i(x^t)$), we get

$$A_\theta^{M-|I|}((\Phi_j(x_0))_{j\notin I}) \le A_\theta^{M-|I|}((\Phi_j(x^t))_{j\notin I}).$$

As $A_\theta^{M-|I|}$ is convex and componentwise nondecreasing, and since $x^t = tx_0 + (1-t)x^*$ with $0 < t < 1$, this proves our claim. □

5.2. Extensions to other penalty methods. In the proof of the selection result Theorem 5.2, the hypotheses (H_0) and (H_1) are mainly assumed in order to ensure that the conclusion of Theorem 2.1 holds. To be more precise, in Theorem 5.2 we can take the following hypotheses, (H_0') and (H_1'), instead of (H_0) and (H_1), with

$$(H_0') \qquad \begin{cases} \Phi_0 : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\} \text{ is a closed proper convex function;} \\ \text{for } 1 \le i \le M,\ \Phi_i : \mathbb{R}^N \to \mathbb{R} \text{ are continuous convex functions.} \end{cases}$$

$$(H_1') \qquad \begin{cases} \theta \text{ is increasing and convex on } \operatorname{dom}(\theta) =]-\infty, \eta[\,,\ \eta \in [0, +\infty]; \\ \alpha \text{ is positive on }]0, +\infty[, \text{ and the conclusions of Theorem 2.1 hold.} \end{cases}$$

Notice that the boundedness of a selection $(x_r)_r$ of approximate solutions $x_r \in S(\mathrm{CP}_r)$ as well as the optimality of any cluster point x_0 of such a family is contained in (H_1'), so that $S(\mathrm{CP}_0)$ is nonempty. We can then get the following convergence result, which is a simple extension of Theorem 5.1 to this setting.

THEOREM 5.4. *Assume that* (H_0'), (H_1'), (H_2)–(H_4) *hold and that the function* Φ_i *is tubular for any* i *in* $\{0, \dots, M\}$. *Then as* r *tends to* 0 *the net* $(x_r)_{r>0}$ *of the optimal solutions of the penalized problems* (CP_r) *converges towards the unique* θ*-center of* (CP_0).

For example, the above result applies to the nonlinear algorithm studied in [4]. This algorithm is based on the penalty scheme (CP_r) and generates a bounded sequence of approximate optimal solutions (x_r) whose cluster points are optimal solutions of (CP_0). For this algorithm, hypothesis (H_1) is not satisfied (because it is associated with a function α such that $\alpha(r)/r$ is bounded near 0), whereas hypothesis (H_1') follows from Lemma 6 therein.

5.3. A nonconvergence result. It is possible, under some further hypotheses on the penalty function θ, to build a continuous convex function $\Phi_0 : \mathbb{R}^2 \to \mathbb{R}$ such that (H_0) holds with the affine constraints $\Phi_1(x) = x_1 - 1$ and $\Phi_2(x) = x_2 - 1$ and for which the net $(x_r)_{r>0}$ of the optimal solutions of $(\mathrm{CP}_r)_{r>0}$ does not converge as r goes to 0. Notice that for Φ_1 and Φ_2 defined as above, the hypothesis (H_2) is clearly satisfied. Our example is defined as follows: Φ_0 is given as the supremum of a denumerable family of affine functions ϕ_n, which corresponds to defining its epigraph as a denumerable intersection of half spaces. The difficulty is to define ϕ_n so that it is associated with a point x^n and a real number r_n for which $x^n = x_{r_n}$ (the optimal solution of (CP_{r_n})), $\phi_n = \Phi_0$ in a neighborhood of x^n, the net x^n has at least two cluster points, and r_n goes to 0 as $n \to +\infty$. Our construction requires the following hypothesis on θ:

$$(H_5) \qquad \begin{cases} \theta \text{ is differentiable on }]-\infty, K[\text{ for some negative } K, \\ \text{and } t \mapsto |t|\theta'(t) \text{ is nondecreasing on }]-\infty, K[. \end{cases}$$

The hypothesis of differentiability on θ is not really restrictive since most penalty functions studied in the literature (and, in particular, any such function cited in [3]) are at least of class C^1 on their domain. The monotonicity of $t \mapsto |t|\theta'(t)$ is more technical, but simple calculations show that the penalty functions θ_1, θ_2, and θ_3 of the preceding sections have this property. As a consequence, Theorem 5.5 applies to the exponential penalty method and the logarithmic barrier method. We also refer to a recent work by Gilbert, Gonzaga, and Karas [8], where the particular case of the logarithmic barrier method is considered: they give an example of a $\mathcal{C}^\infty$-smooth

function Φ_0 such that, for the constraint $\Phi_1(x_1, x_2) := x_2 \geq 0$, the penalty trajectory does not converge as r goes to 0. Notice that such behavior is impossible for an analytic function Φ_0, since analyticity implies tubularity.

THEOREM 5.5. *Assume that* (H_5) *holds and that* (H_1) *is valid with* $\alpha(r) = r$. *Then there exists a continuous convex function* $\Phi : \mathbb{R}^2 \to \mathbb{R}$ *such that the net* $(x_r)_{r>0}$ *of the optimal solutions of the family of problems* $(CP'_r)_{r>0}$ *with*

$$(CP'_r) \qquad \inf \left\{ \Phi(x_1, x_2) + r\theta\left(\frac{x_1 - 1}{r}\right) + r\theta\left(\frac{x_2 - 1}{r}\right) \,:\, (x_1, x_2) \in \mathbb{R}^2 \right\}$$

has at least two cluster points as r *goes to* 0.

Proof. We shall define Φ through its epigraph $\mathrm{epi}(\Phi)$. We first assume that there exist two sequences $(a^n)_{n\geq 1}$ and $(r_n)_{n\geq 1}$ of elements of $]0, 1/3] \times]0, 1[\times]1, +\infty[$ and $]0, +\infty[$, respectively, such that

(i) $(r_n)_n$ is decreasing and $\lim_{n\to+\infty} r_n = 0$,

(ii) $\forall n \geq 1$, $a_1^{2n} = 1/3 - 1/4n$ and $a_1^{2n+1} = 1/4n$, $a_2^n < a_2^{n+1} < 1$, $(a_2^n - 1)/r_n \leq 3K/2$, and $a_3^n > a_3^{n+1} > 1$,

(iii) $\forall n \geq 1$, $\forall k \neq n$, $a^k \in D(a^n, r_n)$,

where $D(x, r)$ is the open subset of $\mathbb{R}^3$ given by

$$D(x, r) = \left\{ z \in \mathbb{R}^3 \,:\, z_3 > x_3 - \theta'\left(\frac{x_1 - 1}{r}\right)(z_1 - x_1) - \theta'\left(\frac{x_2 - 1}{r}\right)(z_2 - x_2) \right\}$$

when (x_1, x_2) belongs to $]-\infty, 1[^2$ and $r > 0$ are such that $\max\left\{\frac{x_1-1}{r}, \frac{x_2-1}{r}\right\} < K$.

Let $\mathcal{C} := \bigcap_{n\geq 1} \overline{D(a^n, r_n)}$; then $\mathcal{C}$ is a closed convex subset of $\mathbb{R}^3$ and is the epigraph of a continuous convex function $\phi : \mathbb{R}^2 \longrightarrow \mathbb{R}$. Indeed, let $x = (x_1, x_2) \in \mathbb{R}^2$. We must show that there exists $\phi(x)$ in $\mathbb{R}$ such that $\{(x_1, x_2, x_3) : x_3 \geq \phi(x)\} = \mathcal{C} \cap \{(x_1, x_2, x_3) : x_3 \in \mathbb{R}\}$. From the definition of $\mathcal{C}$ we infer that if $(x_1, x_2, x_3) \in \mathcal{C}$, then $\{(x_1, x_2, y) : y \geq x_3\} \subset \mathcal{C}$. Moreover, we deduce from (ii) and (H_5) that $\forall\ n \geq 1$

$$a_3^n - \sum_{i=1,2} \theta'\left(\frac{a_i^n - 1}{r_n}\right)(x_i - a_i^n) \leq |a_3^n| + \theta'\left(\frac{3K}{2}\right) \sum_{i=1,2} |x_i - a_i^n|,$$

and as $(a^n)_n$ is bounded, there exists x_3 in $\mathbb{R}$ such that (x_1, x_2, x_3) belongs to $\mathcal{C}$. We can thus define $\phi(x) := \min\{x_3 : (x_1, x_2, x_3) \in \mathcal{C}\}$.

From the definition of $\mathcal{C}$, we have that for any $n \geq 1$, $\phi(a_1^n, a_2^n) = a_3^n$ and

$$\left(-\theta'\left(\frac{a_1^n - 1}{r}\right), -\theta'\left(\frac{a_2^n - 1}{r}\right)\right) \in \partial\phi(a_1^n, a_2^n).$$

Let $m = \min\{\phi(x) : x \in [-1, 2]^2\}$; then we claim that the function $\Phi : (x_1, x_2) \mapsto \max\{\phi(x_1, x_2), x_1^2 + x_2^2 - 5 + m\}$ has the desired properties. Indeed, Φ is continuous, convex, and coercive on $\mathbb{R}^2$, so that (H_0) is satisfied. Moreover, Φ is equal to ϕ on $[-1, 2]^2$, so we infer from the previous remarks that for any $n \geq 1$ the optimal solution x_{r_n} of (CP_{r_n}) is (a_1^n, a_2^n). As a consequence, the net $(x_r)_{r>0}$ has at least two cluster points as r goes to 0, namely, $(0, l)$ and $(\frac{1}{3}, l)$, where $l = \lim_{n\to+\infty} a_2^n$. This proves the claim.

It remains to prove that there exist two sequences $(a^n)_{n\geq 1}$ and $(r_n)_{n\geq 1}$ for which (i)–(iii) hold. We build such sequences by induction on n.

Since θ is convex and increasing, θ' is positive on $]-\infty, K[$. We set $a^1 = (1/3, 1/2, 1+\varepsilon)$, with $\varepsilon = 1/9\theta'(2K)$ (notice that ε is positive) and $r_1 = -1/3K$. Then $(0,1,1)$ and $(1,1,1)$ belong to $D(a^1, r_1)$.

Now fix $n \in \mathbb{N}\backslash\{0\}$ and assume that $a^1, \dots, a^n$ belong to $]0,1/3] \times]0,1[\times]1, +\infty[$ and that $0 < r_n < \cdots < r_1$ are such that (ii) and (iii) hold up to n. Then we claim that for $r \in]0, \min\{\frac{1}{n+1}, r_n\}[$ small enough, one has

1. $a_2^n < a(r)_2 < 1$, $\frac{a(r)_2 - 1}{r} \leq \frac{3K}{2}$, and $a_3^n > a(r)_3 > 1$,
2. $\forall i \in \{1, \dots, n\}$, $a(r) \in D(a^i, r_i)$ and $a^i \in D(a(r), r)$,
3. $(0,1,1) \in D(a(r), r)$ and $(1,1,1) \in D(a(r), r)$,

where $a(r) = (\beta, 1-\sqrt{r}, 1+\delta(r))$, with $\beta := 1/3 - 1/(2(n+1))$ if n is odd, $\beta := 1/(2n)$ otherwise, and $\delta(r) = 1/12\theta'(-1/2r)$. Notice that since $\theta_\infty(-1) = 0$, the limit of θ' at $-\infty$ is equal to 0, so that $\delta(r)$ tends to 0 as r decreases to 0. Let us check that if $r > 0$ is small enough, then $a = a(r)$ and r satisfy properties 1, 2, and 3.

1. It is easily seen that $a(r)$ belongs to $]0,1/3[\times]0,1[\times]1, +\infty[$ for r small enough. Moreover, $a(r)$ and $\frac{a(r)_2 - 1}{r}$, respectively, tend to $(\beta, 1, 1)$ and $-\infty$ as r goes to 0, so that for r small enough, condition 1 is fulfilled for $a = a(r)$.

2. The set $\bigcap_{i=1}^n D(a^i, r_i)$ is a convex open subset of $\mathbb{R}^3$ containing the segment $[(0,1,1), (1,1,1)]$. This implies that $a(r)$ belongs to $\bigcap_{i=1}^n D(a^i, r_i)$ for r small enough, which is the first part of condition 2. The second part of condition 2 for $i \in \{1, \dots, n\}$ reads

$$a_3^i > 1 + \delta(r) - \theta'\left(\frac{\beta - 1}{r}\right)(a_1^i - \beta) - \theta'\left(-\frac{1}{\sqrt{r}}\right)(a_2^i - 1 + \sqrt{r})$$

$$\iff f(r) = a_3^i - 1 - \delta(r) + \theta'\left(\frac{\beta - 1}{r}\right)(a_1^i - \beta) + \theta'\left(-\frac{1}{\sqrt{r}}\right)(a_2^i - 1 + \sqrt{r}) > 0.$$

Since $\beta < 1$ and θ' tends to 0 at $-\infty$, we have $\lim_{r \to 0^+} f(r) = a_3^i - 1 > 0$. As a consequence, condition 2 is satisfied for $r > 0$ small enough.

3. It is sufficient to check that $(0,1,1) \in D(a(r), r)$. This amounts to

$$-\delta(r) - \beta\theta'\left(\frac{\beta - 1}{r}\right) + \theta'\left(-\frac{1}{\sqrt{r}}\right)\sqrt{r} > 0.$$

Since θ' is nondecreasing on $]-\infty, K[$ and β belongs to $]0,1/3[$, $4\delta(r)$ is greater than $\beta\theta'(\frac{\beta-1}{r})$. It is therefore sufficient to check that

$$-5\delta(r) + \theta'\left(-\frac{1}{\sqrt{r}}\right)\sqrt{r} = -\frac{5}{12}\theta'\left(-\frac{1}{2r}\right) + \theta'\left(-\frac{1}{\sqrt{r}}\right)\sqrt{r} > 0.$$

We infer from (H_5) that

$$\frac{1}{\sqrt{r}}\theta'\left(-\frac{1}{\sqrt{r}}\right) \geq \frac{1}{2r}\theta'\left(-\frac{1}{2r}\right) > \frac{5}{12r}\theta'\left(-\frac{1}{2r}\right)$$

as soon as $0 < r < \min\{1/4, 1/K^2\}$.

As a consequence, if we set $a^{n+1} := a(r)$ and $r_{n+1} = r$ for r small enough, the families $(a^k)_{1 \leq k \leq n+1}$ and $(r_k)_{1 \leq k \leq n+1}$ satisfy (ii) and (iii) and $r_{n+1} \leq \frac{1}{n+1}$.

The induction on n thus yields two sequences $(a^n)_n$ and $(r_n)_n$ which satisfy (i)–(iii). This concludes the proof. □

Acknowledgments. The author wishes to express his deep gratitude to H. Attouch for numerous interesting talks on this subject and to thank the anonymous referees for helping improve the paper. Part of this work was performed during post-doctoral fellowships at the Laboratoire de Mathématiques Appliquées of the Ecole Nationale Supérieure des Techniques Avancées (Paris) as well as at the Centro de Modelamiento Matematico of the Universidad de Chile (Santiago), and the author wishes to thank them both for the warm hospitality and support.

REFERENCES

[1] F. Alvarez, *Absolute minimizer in convex programming by exponential penalty*, J. Convex. Anal., 7 (2000), pp. 197–202.

[2] H. Attouch, *Viscosity solutions of minimization problems*, SIAM J. Optim., 6 (1996), pp. 769–806.

[3] A. Auslender, R. Cominetti, and M. Haddou, *Asymptotic analysis for penalty and barrier methods in convex and linear programming*, Math. Oper. Res., 22 (1997), pp. 43–62.

[4] R. Castillo and C. Gonzaga, *A nonlinear programming algorithm based on non-coercive penalty functions*, Math. Program., to appear.

[5] R. Cominetti, *Nonlinear averages and convergence of penalty trajectories in convex programming*, in Proceeding of the Workshop on Ill-posed Variational Problems and Regularization Techniques, Trier, Germany, 1998, Lecture Notes in Econom. and Math. Systems 477, M. Thera and R. Tichatschke, eds., Springer, Berlin, 1999, pp. 65–78.

[6] R. Cominetti and J. San Martin, *Asymptotic analysis for the exponential penalty trajectory in linear programming*, Math. Programming, 67 (1994), pp. 169–187.

[7] G. Dal Maso and L. Modica, *On the convergence of local minima*, Boll. Unione Mat. Ital. Sez. A (6), 1 (1982), pp. 55–61.

[8] J. Ch. Gilbert, C. Gonzaga, and E. Karas, *Examples of ill-behaved central paths in convex optimization*, Rapport de Recherche INRIA, Le Chesney, France, 2000.

[9] R. Huotari, *Tubular sets and multivariate Polya Algorithm*, J. Austral. Math. Soc. Ser. A, 55 (1993), pp. 232–237.

[10] R. Huotari and M. Marano, *The Polya algorithm on tubular sets*, J. Comput. Appl. Math., 54 (1994), pp. 151–157.

[11] L. MacLinden, *An analogue of Moreau's proximation theorem with applications to the nonlinear complementary problem*, Pacific J. Math., 88 (1980), pp. 101–161.

[12] M. Marano, *Strict approximation on closed convex sets*, Approx. Theory Appl., 6 (1990), pp. 99–109.

[13] R. Monteiro and F. Zhou, *On the existence and convergence of the central path for convex programming and some duality results*, Comput. Optim. Appl., 10 (1998), pp. 51–77.

[14] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.

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

SIAM J. OPTIM.
Vol. 13, No. 1, pp. 228–239

NONLINEARLY CONSTRAINED BEST APPROXIMATION IN HILBERT SPACES: THE STRONG CHIP AND THE BASIC CONSTRAINT QUALIFICATION*

CHONG LI† AND XIAO-QING JIN‡

Abstract. We study best approximation problems with nonlinear constraints in Hilbert spaces. The strong "conical hull intersection property" (CHIP) and the "basic constraint qualification" (BCQ) condition are discussed. Best approximations with differentiable constraints and convex constraints are characterized. The analysis generalizes some linearly constrained results of recent works [F. Deutsch, W. Li, and J. Ward, *J. Approx. Theory*, 90 (1997), pp. 385–444; F. Deutsch, W. Li, and J. D. Ward, *SIAM J. Optim.*, 10 (1999), pp. 252–268].

Key words. best approximation, strong CHIP, BCQ condition, differentiable constraint, convex constraint

AMS subject classifications. 41A65, 41A29

PII. S1052623401385600

1. Introduction.

In recent years, a lot of attention has been focused on constrained best approximation problems in Hilbert spaces; see, e.g., [5, 6, 9, 10, 11, 16, 17]. These problems find applications (cf. [2]) in statistics, mathematical modeling, curve fitting, and surface fitting. The setting is as follows. Let X be a Hilbert space, C a nonempty closed convex subset of X, and A a bounded linear operator from X to a finite-dimensional Hilbert space Y. Given "data" $b \in Y$, the problem consists of finding the best approximation $P_K(x)$ to any $x \in X$ from the set

$$K := C \cap A^{-1}(b) = C \cap \{x \in X : Ax = b\}.$$

Generally, it is easier to compute the best approximation from C than from K. Therefore, the interest of several papers [5, 6, 9, 11, 16, 17] was centered on the following problem: for any $x \in X$, does there exist a $y \in Y$ such that $P_K(x) = P_C(x + A^*y)$? It was proved in [9] that a sufficient and necessary condition for an affirmative answer to this question is that the pair $\{C, A^{-1}(b)\}$ satisfy the strong "conical hull intersection property" (CHIP).

Very recently, Deutsch, Li, and Ward in [10] considered a more general problem of finding the best approximation $P_K(x)$ to any $x \in X$ from the set

$$K = C \cap \{x \in X : Ax \leq b\} \tag{1.1}$$

and established a result similar to that of [9]. More precisely, they proved the following theorem (see Theorem 3.2 and Lemma 3.1 in [10]).

THEOREM DLW. *Let A be defined on X by*

$$Ax := (\langle x, h_1\rangle, \langle x, h_2\rangle, \ldots, \langle x, h_m\rangle)$$

*Received by the editors February 26, 2001; accepted for publication (in revised form) February 11, 2002; published electronically July 16, 2002.
http://www.siam.org/journals/siopt/13-1/38560.html

†Department of Mathematics, Zhejiang University, Hangzhou 310027, P. R. China (cli@seu.edu.cn). The research of this author is supported by the National (grant 19971013) and Jiangsu Provincial (grant BK99001) Natural Science Foundations of China.

‡Faculty of Science and Technology, University of Macau, Macau, P. R. China (xqjin@umac.mo). The research of this author is supported by research grants RG010/99-00S/JXQ/FST and RG026/00-01S/JXQ/FST from the University of Macau.

for some $h_i \in X \setminus \{0\}$ *for* $i = 1, 2, \ldots, m$. *Let* $b \in \mathbb{R}^m$ *and* $x^* \in K = C \cap \{x \in X : Ax \leq b\}$. *Then the following two statements are equivalent:*

(i) *For any* $x \in X$, $x^* = P_K(x) \Longleftrightarrow x^* = P_C(x - \sum_{i=1}^m \lambda_i h_i)$ *for some* $\lambda_i \geq 0$ *with* $\lambda_i(\langle x, h_i\rangle - b_i) = 0$ *for all* i.

(ii) $\{C, H_1, \ldots, H_m\}$ *has the strong CHIP at* x^*, *where* $H_i := \{x \in X : \langle x, h_i\rangle \leq b_i\}$ *for all* i.

Theorem DLW gives an unconstrained reformulation for the linearly constrained system, for which a complete theory has been established. The importance of such a theory was described in detail in [10, 11], etc. One natural problem is: can one extend such a theory to a nonlinearly constrained system? Admittedly, this problem for a general nonlinearly constrained system is quite difficult. In this paper, we shall relax the linearity assumption made on the operator A in the constraint (1.1) in two ways. First, we study the case in which A is assumed to be Fréchet differentiable, and second, we examine the case in which A is convex (i.e., each component is convex).

In the Fréchet differentiable case, we will give a theorem (Theorem 4.1) that is similar to Theorem DLW, where h_i in Theorem DLW is replaced by the Fréchet derivative $\bigtriangledown A_i(x^*)$ of A_i at x^*, for $i = 1, 2, \ldots, m$. Note that, when A is nonlinear, the approximating set K is, in general, nonconvex (see Example 4.1). Thus Theorem DLW does not work in this case, since K can not be re-expressed as the intersection of C and a polyhedron. In addition, the nonconvexity of the set K makes the original problem very complicated. In fact, there is no successful way to characterize the best approximation from general nonconvex sets. The merit of the present results lies in converting a nonconvex constrained problem into a convex unconstrained one.

In the convex case, the sets H_i, $i = 1, 2, \ldots, m$, may not be well defined, although K is convex and, in general, Theorem DLW does not work either (see Example 5.1). To establish a similar unconstrained reformulation result, we introduce the concept of the "basic constraint qualification" (BCQ) relative to C, which is a generalization of the BCQ considered in [12, 13]. We prove that the BCQ relative to C is a sufficient and necessary condition to ensure the following "perturbation property": for any $x \in X$, $P_K(x) = x^*$ if and only if $P_C(x - \sum_1^m \lambda_i h_i) = x^*$ for some $h_i \in \partial A_i(x^*)$ and $\lambda_i \geq 0$ with $\lambda_i(A_i(x^*) - b_i) = 0$. Clearly, in either case, the present results generalize the main results in [10].

The paper is organized as follows. We describe some notations and a useful proposition in section 2. To deal with the problem with differentiable constraints, we need to linearize the constraints in section 3. Unconstrained reformulation results for differentiable constraints and convex constraints are established in sections 4 and 5, respectively. Finally, a concluding remark is given in section 6.

2. Preliminaries. Let X be a Hilbert space. For a nonempty subset S of X, the convex hull (resp., conical hull) of S, denoted by conv$S$ (resp., cone$S$), is the intersection of all convex sets (resp., convex cones) including S, while the dual cone S° of S is defined by

$$S^\circ = \{x \in X : \ \langle x, y\rangle \leq 0 \ \ \forall y \in S\}.$$

Then the normal cone of S at x is defined by $N_S(x) = (S - x)^\circ$. The closure (resp., interior, relative interior) of any set S is denoted by $\overline{S}$ (resp., int$S$, ri$S$).

For a function f from X to $\mathbb{R}$, the subdifferential of f at $x \in X$, denoted by $\partial f(x)$, is defined by

$$\partial f(x) = \{z \in X : \ \ f(x) + \langle z, y - x\rangle \leq f(y) \ \ \forall y \in X\}.$$

It is well known that $\partial f(x) \neq \emptyset$ for all $x \in X$ if f is a continuous convex function.

Let G be a nonempty closed convex set in X. Then for any $x \in X$, there exists a unique best approximation $P_G(x)$ from G to x. We define

$$\tau(x, y) = \lim_{t \to +0} \frac{\|x + ty\| - \|x\|}{t}.$$

Since $x/\|x\|$ is the unique supporting functional of x, we have

$$\tau(x, y) = \frac{\langle x, y \rangle}{\|x\|}.$$

The following well-known characterization of the best approximation is useful; see [9, 10].

PROPOSITION 2.1. *Let G be a convex subset of X, $x \in X$, and $g_0 \in G$. Then $P_G(x) = g_0 \Longleftrightarrow \langle x - g_0, g_0 - g \rangle \geq 0$ for any $g \in G \Longleftrightarrow x - g_0 \in (G - g_0)^\circ$.*

3. Linearization of the constraints. In the remainder of the paper, we always assume that $C \neq \emptyset$ is a closed convex subset of X. Suppose that

$$A(\cdot) = (A_1(\cdot), \ldots, A_m(\cdot))$$

is Fréchet differentiable from X to $\mathbb{R}^m$ and $b = (b_1, \ldots, b_m) \in \mathbb{R}^m$. Let $m_e \in \{1, 2, \ldots, m\}$ be a fixed integer. Define

$$K_0 = \{x \in X : \ A_i(x) = b_i, \ i \in E\} \cap \{x \in X : \ A_i(x) \leq b_i, \ i \in I\}$$

and

$$K = C \cap K_0,$$

where

$$E = \{1, 2, \ldots, m_e\}, \qquad I = \{m_e + 1, \ldots, m\}.$$

Furthermore, let

$$I(x^*) = \{i \in I : A_i(x^*) = b_i\} \quad \forall x^* \in K.$$

The following concepts can be easily found in any book on constrained optimization; see, e.g., [14, 20].

DEFINITION 3.1. *Let $x^* \in K$. A vector $d \neq 0$ is called a feasible direction of K at x^* if there exists $\delta > 0$ such that*

$$x^* + td \in K \quad \forall t \in [0, \delta].$$

The set of all feasible directions of K at x^* is denoted by $\mathrm{FD}(x^*)$.

DEFINITION 3.2. *Let $x^* \in K$. A vector d is called a linearized feasible direction of K at x^* if*

$$\langle d, \bigtriangledown A_i(x^*) \rangle = 0 \quad \forall i \in E$$

and

$$\langle d, \bigtriangledown A_i(x^*) \rangle \leq 0 \quad \forall i \in I(x^*),$$

where $\bigtriangledown A_i(x^*)$ is the Fréchet derivative of A_i at x^*. The set of all linearized feasible directions of K at x^* is denoted by $\mathrm{LFD}(x^*)$.

DEFINITION 3.3. *Let $x^* \in K$. A vector d is called a sequentially feasible direction of K at x^* if there exist a sequence $\{d_k\} \subset X$ and a sequence $\{\delta_k\}$ of real positive numbers such that*

$$d_k \to d, \qquad \delta_k \to 0, \qquad x^* + \delta_k d_k \in K, \qquad k = 1, 2, \ldots.$$

The set of all sequentially feasible directions of K at x^* is denoted by $\mathrm{SFD}(x^*)$.

Obviously, we have the following inclusion relationship for various feasible directions.

PROPOSITION 3.1. *Let $x^* \in K$. Then*

$$\mathrm{FD}(x^*) \subseteq \mathrm{SFD}(x^*) \subseteq \mathrm{LFD}(x^*).$$

For convenience, let

$$K_S(x^*) = \overline{\mathrm{conv}(x^* + \mathrm{SFD}(x^*))} \cap C$$

and

$$K_L(x^*) = (x^* + \mathrm{LFD}(x^*)) \cap C.$$

Then $K_S(x^*)$ and $K_L(x^*)$ are closed convex cones.

The following two theorems describe the equivalence of the best approximation from K and from $K_S(x^*)$, which plays an important role in our study.

THEOREM 3.1. *Let $x^* \in K$. Then, for any $x \in X$, if $P_K(x) \ni x^*$, we have*

$$P_{K_S(x^*)}(x) = x^*.$$

Proof. For any $\bar{x} \in x^* + \mathrm{SFD}(x^*)$, $d = \bar{x} - x^* \in \mathrm{SFD}(x^*)$ there exist $d_k \in X$ with $d_k \to d$ and $\delta_k > 0$ with $\delta_k \to 0$ such that $x^* + \delta_k d_k \in K$. It follows from $P_K(x) \ni x^*$ that

$$\|x - x^*\| \le \|x - x^* - \delta_k d_k\|, \qquad k = 1, 2, \ldots.$$

Since

$$\begin{aligned} \tau(x - x^*, x^* - \bar{x}) &= \lim_k \frac{\|x - x^* - \delta_k d\| - \|x - x^*\|}{\delta_k} \\ &\ge \lim_k \frac{\|x - x^* - \delta_k d\| - \|x - x^* - \delta_k d_k\|}{\delta_k} \end{aligned}$$

and

$$\left| \frac{\|x - x^* - \delta_k d\| - \|x - x^* - \delta_k d_k\|}{\delta_k} \right| \le \|d_k - d\|,$$

it follows that

$$\langle x - x^*, x^* - \bar{x} \rangle \ge 0 \quad \forall \bar{x} \in x^* + \mathrm{SFD}(x^*).$$

Since $K_S(x^*) \subseteq \overline{\mathrm{conv}(x^* + \mathrm{SFD}(x^*))}$, we have

$$\langle x - x^*, x^* - \bar{x} \rangle \ge 0 \quad \forall \bar{x} \in K_S(x^*).$$

This, with Proposition 2.1, implies that $P_{K_S(x^*)}(x) = x^*$, and the theorem follows. ∎

THEOREM 3.2. *Let $x^* \in K$. Then the following two statements are equivalent:*

(i) $K \subseteq K_S(x^*)$.
(ii) *For any* $x \in X$, $P_{K_S(x^*)}(x) = x^* \Longrightarrow P_K(x) = x^*$.

Proof. It suffices to prove that (ii) $\Longrightarrow$ (i). Let $G = \overline{\text{conv}(x^* + \text{SFD}(x^*))}$. Suppose on the contrary that $K \not\subseteq K_S(x^*)$. Then $K \not\subseteq G$, so that there is $\bar{x} \in K$ but $\bar{x} \notin G$. Let $g_0 = P_G(\bar{x})$ and $x = \bar{x} - g_0 + x^*$. Then $P_G(x) = x^*$. In fact, since $G = x^* + \overline{\text{conv}(\text{SFD}(x^*))}$, for any $g \in G$, there exist $\bar{g}_0$, $\bar{g} \in \overline{\text{conv}(\text{SFD}(x^*))}$ such that

$$g_0 = x^* + \bar{g}_0 \qquad \text{and} \qquad g = x^* + \bar{g}.$$

Note that G is a cone with vertex x^*. It follows that

$$g + g_0 - x^* = x^* + \bar{g} + \bar{g}_0 \in G,$$

which, by Proposition 2.1, implies that

$$\langle \bar{x} - g_0, g_0 - (g + g_0 - x^*) \rangle \geq 0$$

as $g_0 = P_G(\bar{x})$. Thus we have

$$\langle x - x^*, x^* - g \rangle = \langle \bar{x} - g_0, g_0 - (g + g_0 - x^*) \rangle \geq 0,$$

which proves that $P_G(x) = x^*$. Now define

$$x_t = x^* + t(x - x^*) \quad \forall t > 0.$$

From

$$\|x_t - x^*\| = t\|x - x^*\| \leq t \left\| x - \left[\left(1 - \frac{1}{t}\right) x^* + \left(\frac{1}{t}\right) g \right] \right\| = \|x_t - g\| \quad \forall g \in G,\ t > 1,$$

it follows that $P_G(x_t) = x^*$ for $t > 1$. Therefore, from (ii) and $K_S(x^*) \subseteq G$, we have $P_K(x_t) = x^*$ for $t > 1$.

On the other hand, for $t > 1$ we obtain

$$\begin{aligned} \|x_t - \bar{x}\|^2 &= \|x^* + t(\bar{x} - g_0) - \bar{x}\|^2 = \|x^* - g_0 + (t-1)(\bar{x} - g_0)\|^2 \\ &= (t-1)^2\|\bar{x} - g_0\|^2 + 2(t-1)\langle \bar{x} - g_0, x^* - g_0 \rangle + \|x^* - g_0\|^2. \end{aligned}$$

Since $g_0 = P_G(\bar{x})$, it follows from Proposition 2.1 that $\langle \bar{x} - g_0, x^* - g_0 \rangle \leq 0$, and hence

$$\begin{aligned} \|x_t - \bar{x}\|^2 &\leq (t-1)^2\|\bar{x} - g_0\|^2 + \|x^* - g_0\|^2 \\ &= t^2\|\bar{x} - g_0\|^2 - 2t\|\bar{x} - g_0\|^2 + \|\bar{x} - g_0\|^2 + \|x^* - g_0\|^2 \\ &< t^2\|\bar{x} - g_0\|^2 = \|x_t - x^*\|^2 \end{aligned}$$

for all $t > 1$ large enough. This means that $x^* \notin P_K(x_t)$, which is a contradiction. The proof is complete. □

Similarly, we have the following result for $K_L(x^*)$.

THEOREM 3.3. *Let* $x^* \in K$. *Then the following statements are equivalent:*
(i) $K \subseteq K_L(x^*)$.
(ii) *For any* $x \in X$, $P_{K_L(x^*)}(x) = x^* \Longleftrightarrow P_K(x) \ni x^*$.

COROLLARY 3.1. *Let* $x^* \in K$. *Consider the following statements:*

(i) $K \subseteq K_L(x^*)$ *and* $K_S(x^*) = K_L(x^*)$.
(ii) *For any* $x \in X$, $P_K(x) \ni x^* \Longleftrightarrow P_{K_L(x^*)}(x) = x^*$.
(iii) *For any* $x \in X$, $P_K(x) \ni x^* \Longrightarrow P_{K_L(x^*)}(x) = x^*$.

Then (i) $\Longrightarrow$ (ii) $\Longrightarrow$ (iii). *Furthermore, if* $K \subseteq K_S(x^*)$, *then* (i) $\Longleftrightarrow$ (ii) $\Longleftrightarrow$ (iii).

Proof. If (i) holds, by Theorems 3.1 and 3.2, we have

$$P_K(x) \ni x^* \Longleftrightarrow P_{K_S(x^*)}(x) = x^* \Longleftrightarrow P_{K_L(x^*)}(x) = x^*.$$

Therefore (ii) holds. The implication (ii) $\Longrightarrow$ (iii) is trivial. Now assume that $K \subseteq K_S(x^*)$. If (iii) holds, then, for any $x \in X$,

$$P_{K_S(x^*)}(x) = x^* \Longrightarrow P_K(x) \ni x^* \Longrightarrow P_{K_L(x^*)}(x) = x^*.$$

Thus, with almost the same arguments as in the proof of Theorem 3.2, we have $K_L(x^*) \subseteq K_S(x^*)$. By Proposition 3.1, $K_L(x^*) = K_S(x^*)$ and so $K \subseteq K_L(x^*)$; i.e., (i) holds. □

It should be noted that if K is convex (e.g., $A_1, \ldots, A_{m_e}$ are linear and $A_{m_e+1}, \ldots, A_m$ are convex), $K \subseteq K_S(x^*)$ holds. We therefore have the following corollary.

COROLLARY 3.2. *Let* $x^* \in K$. *If* K *is convex, then the following statements are equivalent:*

(i) $K_S(x^*) = K_L(x^*)$.
(ii) *For any* $x \in X$, $P_K(x) = x^* \Longleftrightarrow P_{K_L(x^*)}(x) = x^*$.
(iii) *For any* $x \in X$, $P_K(x) = x^* \Longrightarrow P_{K_L(x^*)}(x) = x^*$.

4. Reformulations of differentiable constraints. The following notation of the strong CHIP, taken from [9, 10], plays an important role in optimization theory; see, e.g., [7, 8, 12, 18].

DEFINITION 4.1. *Let* $\{C_0, \ldots, C_m\}$ *be a collection of closed convex sets and* $x \in \cap_{j=0}^m C_j$. *Then* $\{C_0, \ldots, C_m\}$ *is said to have the strong CHIP at* x *if*

$$\left(\bigcap_{j=0}^m C_j - x\right)^\circ = \sum_{j=0}^m (C_j - x)^\circ.$$

Now, for convenience, we write

$$\bigtriangledown A_{i+m}(x^*) = -\bigtriangledown A_i(x^*), \quad i = 1, 2, \ldots, m_e,$$

$$\bar{b}_i = b_i - A(x^*) + \langle x^*, \bigtriangledown A_i(x^*)\rangle, \quad i = 1, 2, \ldots, m + m_e,$$

$$H_i = \{d \in X : \langle d, \bigtriangledown A_i(x^*)\rangle \le \bar{b}_i\}, \quad i = 1, 2, \ldots, m + m_e,$$

and

$$E_0 = E \cup I(x^*) \cup \{m+1, \ldots, m+m_e\}, \qquad E_1 = I \setminus I(x^*).$$

We define the bounded linear mapping $\bigtriangledown A(x^*)|$ from X to $\mathbb{R}^{m_e}$ by

$$\bigtriangledown A(x^*)|x = (\langle x, \bigtriangledown A_1(x^*)\rangle, \ldots, \langle x, \bigtriangledown A_{m_e}(x^*)\rangle) \in \mathbb{R}^{m_e} \quad \forall x \in X.$$

The inverse of $\bigtriangledown A(x^*)|$, which is generally a set-valued mapping, is denoted by $A(x^*)|^{-1}$. Let

$$\bar{b} = (\bar{b}_1, \ldots, \bar{b}_{m_e}).$$

Then we are ready to give the main result of this section.

THEOREM 4.1. *Let $x^* \in K$. Suppose that $K \subseteq K_L(x^*)$ and $K_S(x^*) = K_L(x^*)$. Then the following statements are equivalent:*

(i) $\{C,\ \bigtriangledown A(x^*)|^{-1}(\bar{b}),\ H_i,\ i \in I(x^*)\}$ *has the strong CHIP at* x^*.

(ii) $\{C,\ \bigtriangledown A(x^*)|^{-1}(\bar{b}),\ H_i,\ i \in I\}$ *has the strong CHIP at* x^*.

(iii) *For any* $x \in X$,

$$P_K(x) \ni x^* \Longleftrightarrow P_C\left(x - \sum_1^m \lambda_i \bigtriangledown A_i(x^*)\right) = x^*$$

for some $\lambda_i,\ i = 1, \ldots, m$, *with* $\lambda_i \geq 0$ *for all* $i \in I$, *and* $\lambda_i = 0$ *for all* $i \notin E \cup I(x^*)$.

Proof. We first assume that (i) holds. Since $x^* \in \operatorname{int} \cap_{i \in E_1} H_i$, it follows from Proposition 2.3 of [10] that $\{C \cap (\cap_{i \in E_0} H_i), H_i, i \in E_1\}$ has the strong CHIP at x^*. Thus (i) implies that $\{C,\ \bigtriangledown A(x^*)|^{-1}(\bar{b}),\ H_i,\ i = 1, \ldots, m\}$ has the strong CHIP at x^*. Therefore, (ii) holds.

Now suppose that (ii) holds. By Corollary 3.1, we have that, for any $x \in X$, $P_K(x) \ni x^* \Longleftrightarrow P_{K_L(x^*)}(x) = x^*$. We will show that $P_{K_L(x^*)}(x) = x^*$ if and only if $P_{K_L^0(x^*)}(x) = x^*$, where $K_L^0(x^*) = C \cap (\cap_{i=1}^{m+m_e} H_i)$. In fact, it is clear that $P_{K_L(x^*)}(x) = x^*$ implies $P_{K_L^0(x^*)}(x) = x^*$. Conversely, assume that $P_{K_L^0(x^*)}(x) = x^*$. Since $K_L(x^*) \bigcap U(x^*, r) \subseteq K_L^0(x^*)$ for some $r > 0$, where $U(x^*, r)$ denotes the open ball with center x^* and radius $r > 0$, x^* is a best approximation to x from $K_L(x^*) \bigcap U(x^*, r)$, that is, x^* is a local best approximation to x from $K_L(x^*)$, and hence $P_{K_L(x^*)}(x) = x^*$ by [3]. Note that any finite collection of half-spaces has the strong CHIP [9]. It follows that $\{C,\ \bigtriangledown A(x^*)^{-1}(\bar{b}),\ H_i :\ i \in I\}$ has the strong CHIP at $x^* \Longleftrightarrow \{C,\ H_i :\ i = 1, 2, \ldots, m + m_e\}$ has the strong CHIP at x^*. Thus, using Theorem DLW, we have

$$P_{K_L^0(x^*)}(x) = x^* \Longleftrightarrow P_C\left(x - \sum_{i=1}^{m+m_e} \lambda_i \bigtriangledown A_i(x^*)\right) = x^*$$

for some $\lambda_i \geq 0$, $i = 1, \ldots, m + m_e$, with $\lambda_i(\langle x^*, \bigtriangledown A_i(x^*)\rangle - \bar{b}_i) = 0$. Consequently, (iii) holds.

Finally, if (iii) holds, it follows from Corollary 3.1 that, for any $x \in X$,

$$P_{K_L(x^*)}(x) = x^* \Longleftrightarrow P_K(x) \ni x^* \Longleftrightarrow P_C\left(x - \sum_1^m \lambda_i \bigtriangledown A_i(x^*)\right) = x^*$$

for some λ_i, $i = 1, \ldots, m$, with $\lambda_i \geq 0$ for all $i \in I$, and $\lambda_i = 0$ for all $i \notin E \cup I(x^*)$. Consequently,

$$P_{K_L(x^*)}(x) = x^* \Longleftrightarrow P_C\left(x - \sum_{i \in E \cup I(x^*)} \lambda_i \bigtriangledown A_i(x^*)\right) = x^*$$

for some λ_i, $i \in E \cup I(x^*)$, satisfy $\lambda_i \geq 0$ for all $i \in I(x^*)$, or equivalently,

$$P_{K_L(x^*)}(x) = x^* \Longleftrightarrow P_C\left(x - \sum_{i \in E_0} \lambda_i \bigtriangledown A_i(x^*)\right) = x^*$$

for some $\lambda_i \geq 0,\ i \in E_0$. Thus, using Theorem DLW again, we know that $\{C,\ H_i,\ i \in E_0\}$ has the strong CHIP at x^*, and so does $\{C,\ \bigtriangledown A(x^*)|^{-1}(\bar{b}),\ H_i,\ i \in I(x^*)\}$; i.e., (i) holds. The proof of the theorem is complete. □

Remark 4.1. Recall that the constraint qualification condition on $\text{span}(C - x^*)$ is satisfied at x^* if $\text{SFD}(x^*) \cap \text{span}(C - x^*) = \text{LFD}(x^*) \cap \text{span}(C - x^*)$, which plays an important role in nonlinear optimization theory; see [1, 14]. Clearly, if the constraint qualification condition is satisfied at x^* (indeed, it does if each $A_i,\ i \in I(x^*)$, is linear or the Mangasarian–Fromovitz constraint qualification on $\text{span}(C - x^*)$ (see [15]) is satisfied, with $x^* \in \text{ri}C$), then $K_S(x^*) = K_L(x^*)$.

The following proposition shows that the conditions $K \subseteq K_L(x^*)$ and $K_S(x^*) = K_L(x^*)$ are "almost" necessary.

PROPOSITION 4.1. *Suppose that the conclusion of Theorem* 4.1 *is valid. Suppose in addition that one of conditions* (i)–(iii) *holds; then* $K \subseteq K_L(x^*)$. *Moreover, if* $K \subseteq K_S(x^*)$, *in particular if* K *is convex, then* $K_S(x^*) = K_L(x^*)$.

Proof. Under the assumption of Proposition 4.1, we have that, for any $x \in X$, $P_K(x) \ni x^* \Longleftrightarrow P_{K_L(x^*)}(x) = x^*$. Thus, by Theorem 3.1, $K \subseteq K_L(x^*)$. Moreover, if $K \subseteq K_S(x^*)$, we have $K_S(x^*) = K_L(x^*)$ from Corollary 3.1. □

Now we give an example to illustrate the main theorem of this section.

Example 4.1. Let $X = \mathbb{R}^2$, $C = \{(x_1, x_2) :\ (x_1 - 4)^2 + x_2^2 \leq 16\}$, and

$$A_1(x) = x_2 - \sin x_1, \quad A_2(x) = -x_1 - x_2 \qquad \forall x = (x_1, x_2) \in X.$$

For $x^* = (0, 0)$ we have

$$\bigtriangledown A_1(x^*) = (-1, 1), \qquad \bigtriangledown A_2(x^*) = (-1, -1),$$

and

$$K_L(x^*) = K_S(x^*) = \{(x_1, x_2) :\ x_2 \leq x_1, -x_1 \leq x_2\}.$$

Let $m_e = 0$. Clearly, $K \subset K_L(x^*)$. Since $\text{int}C \cap H_1 \cap H_2 \neq \emptyset$, it follows from Proposition 2.3 of [10] that $\{C, H_1, H_2\}$ has the strong CHIP. Then, by Theorem 4.1, for any $x = (x_1, x_2) \in X$, $P_K(x) \ni x^*$ if and only if there exist $\lambda_1,\ \lambda_2 \geq 0$ such that $P_C(x - \lambda_1(-1, 1) - \lambda_2(-1, -1)) = x^*$. Observe that, for any $y = (y_1, y_2)$, $P_C(y) = x^*$ if and only if $y_1 \leq 0,\ y_2 = 0$. It follows that $P_K(x) \ni x^*$ if and only if $x = (x_1, x_2)$ satisfies that $x_1 + x_2 \leq 0$ and $x_1 - x_2 \leq 0$. We remark that this result can not be deduced from Theorem DLW.

5. Reformulations of convex constraints. Throughout this section, we always assume that $A_i,\ i = 1, \ldots, m$, are convex continuous functions. Without loss of generality, let

$$C_i = \{x \in X : A_i(x) \leq 0\}, \quad i = 1, \ldots, m,$$

and

$$K = C \cap \left(\bigcap_{i=1}^{m} C_i \right).$$

We first introduce the concept of the BCQ relative to C. For convenience, in what follows, $\text{cone}\{\partial A_i(x) : A_i(x) = 0\}$ is understood to be 0 when $A_i(x) < 0$ for all i.

DEFINITION 5.1. *Let $x \in K$. The system of convex inequalities*

$$A_1(x) \leq 0, \ldots, A_m(x) \leq 0 \tag{5.1}$$

is said to satisfy the BCQ relative to C at x if

$$N_K(x) = N_C(x) + \text{cone}\{\partial A_i(x) : A_i(x) = 0\}.$$

The system of convex inequalities (5.1) *is said to satisfy the BCQ relative to C if it satisfies the BCQ relative to C at any $x \in K$.*

Remark 5.1. When $C = X$, the BCQ relative to C at x is just the BCQ at x considered in [12, 13]. Note that if $x \in K$ and $A_i(x) = 0$, then $\text{cone}(\partial A_i(x)) \subseteq N_{C_i}(x)$, and the equality holds if x is not a minimizer of A_i; see [4, Corollary 1, p. 50].

Similar to the general BCQ, we also have the following properties about the BCQ relative to C.

PROPOSITION 5.1. *Let $x \in K$. The system* (5.1) *satisfies the BCQ relative to C at x if and only if*

$$N_K(x) \subseteq N_C(x) + \text{cone}\{\partial A_i(x) : A_i(x) = 0\}.$$

Proof. Note that

$$N_C(x) + \text{cone}\{\partial A_i(x) : i \in I(x)\} \subseteq N_C(x) + \textstyle\sum_{i\in I(x)} N_{C_i}(x)$$

$$\subseteq N_C(x) + \textstyle\sum_{i=1}^m N_{C_i}(x) \subseteq N_K(x).$$

The result follows. □

PROPOSITION 5.2. *Let $x \in K$. Suppose that the system* (5.1) *satisfies the BCQ relative to C at x. Then $\{C, C_1, \ldots, C_m\}$ has the strong CHIP at x.*

DEFINITION 5.2. *The system* (5.1) *is said to satisfy the weak Slater condition on C if there exists some $\bar{x} \in (\text{ri}C) \cap K$, called a weak Slater point, such that for any i, A_i is affine or $A_i(\bar{x}) < 0$.*

Remark 5.2. When $C = X$, the weak Slater condition on C is just the weak Slater condition studied in [12, 13].

The following proposition is a generalization of Corollary 7 of [12].

PROPOSITION 5.3. *Suppose that the system* (5.1) *satisfies the weak Slater condition on C. Then it satisfies the BCQ relative to C.*

Proof. Let $I_0 = \{i \in I : A_i \text{ is affine}\}$, $H_0 = \cap_{i\notin I_0} C_i$, and $H = \cap_{i\in I_0} C_i$. From Theorem 5.1 of [10] and Proposition 2.3 of [10], it follows that $\{C, H\}$ and $\{C \cap H, H_0\}$ have the strong CHIP. Thus, for any $x \in K$, we have

$$N_K(x) = N_{C\cap H}(x) + N_{H_0}(x) = N_C(x) + N_H(x) + N_{H_0}(x).$$

Observe that the system (5.1) satisfies the weak Slater condition [12]. Then Remark 5.1 implies that the system (5.1) satisfies the BCQ. Hence

$$N_H(x) + N_{H_0}(x) = \text{cone}\{\partial A_i(x) : \ \ i \in I(x)\}$$

for $\{H, H_0\}$ has the strong CHIP by Proposition 2.3 of [10]. Therefore, the system (5.1) satisfies the BCQ relative to C. The proof is complete. □

The following lemma isolates a condition that does not depend upon the BCQ but also still allows the computation of $P_K(x)$ via a perturbation technique.

LEMMA 5.1. *Let* $x^* = P_C(x - \sum_1^m \lambda_i h_i) \in K$ *for some* $h_i \in \partial A_i(x^*)$ *and* $\lambda_i \geq 0$ *with* $\lambda_i = 0$ *for* $i \notin I(x^*)$. *Then* $x^* = P_K(x)$.

Proof. Since $\lambda_i = 0$ for all $i \notin I(x^*)$ and $x^* = P_C(x - \sum_{i \in I(x^*)} \lambda_i h_i)$, it follows from Proposition 2.1 that

$$x - \sum_{i \in I(x^*)} \lambda_i h_i - x^* \in (C - x^*)^\circ.$$

Hence

$$x - x^* \in (C - x^*)^\circ + \sum_{i \in I(x^*)} \lambda_i h_i \subseteq (C - x^*)^\circ + \text{cone}\{\partial A_i(x^*) : \ \ i \in I(x^*)\} \subseteq (K - x^*)^\circ.$$

Using Proposition 2.1 again, we have $x^* = P_K(x)$. ∎

The main theorem of this section is stated as follows.

THEOREM 5.1. *Let* $x^* \in K$. *Then the following two statements are equivalent:*

(i) *The system* (5.1) *satisfies the BCQ relative to* C *at* x^*.

(ii) *For any* $x \in X$,

$$P_K(x) = x^* \Longleftrightarrow x^* = P_C\left(x - \sum_1^m \lambda_i h_i\right)$$

for some $h_i \in \partial A_i(x^*)$ *and* $\lambda_i \geq 0$ *with* $\lambda_i = 0$ *for* $i \notin I(x^*)$.

Proof. Assume that (i) holds. To show (ii), by Lemma 5.1, we need only to prove that, for any $x \in X$, $P_K(x) = x^*$ implies that $x^* = P_C(x - \sum_1^m \lambda_i h_i)$ for some $h_i \in \partial A_i(x^*)$ and $\lambda_i \geq 0$, with $\lambda_i = 0$ for $i \notin I(x^*)$. From Proposition 2.1 and (i), we have

$$x - x^* \in (K - x^*)^\circ \subseteq (C - x^*)^\circ + \text{cone}\{\partial A_i(x^*) : \ \ i \in I(x^*)\}.$$

Therefore, there exist $h_i \in \partial A_i(x^*)$ and $\lambda_i \geq 0$ for $i \in I(x^*)$ such that

$$x - x^* \in (C - x^*)^\circ + \sum_{i \in I(x^*)} \lambda_i h_i.$$

That is,

$$x - \sum_{i \in I(x^*)} \lambda_i h_i - x^* \in (C - x^*)^\circ.$$

It follows from Proposition 2.1 that $x^* = P_C(x - \sum_1^m \lambda_i h_i)$ and (ii) holds.

Conversely, assume that (ii) holds. For $z \in (K - x^*)^\circ$, let $x = z + x^*$. Observe that $x - x^* \in (K - x^*)^\circ$ implies that $P_K(x) = x^*$. It follows from (ii) that $x^* = P_C(x - \sum_1^m \lambda_i h_i)$ for some $h_i \in \partial A_i(x^*)$ and $\lambda_i \geq 0$, with $\lambda_i = 0$ for $i \notin I(x^*)$. Using Proposition 2.1, we have

$$z = x - \sum_1^m \lambda_i h_i - x^* + \sum_1^m \lambda_i h_i \in (C - x^*)^\circ + \text{cone}\{\partial A_i(x^*) : \ \ i \in I(x^*)\}.$$

Hence

$$(K - x^*)^\circ \subseteq (C - x^*)^\circ + \text{cone}\{\partial A_i(x^*) : \ \ i \in I(x^*)\}.$$

From Proposition 5.1, (i) holds. The proof is complete. ∎

COROLLARY 5.1. *The following two statements are equivalent:*

(i) *The system of convex inequalities* (5.1) *satisfies the BCQ relative to* C.
(ii) *For any* $x \in X$, $x^* \in K$, $P_K(x) = x^* \iff x^* = P_C(x - \sum_1^m \lambda_i h_i)$ *for some* $h_i \in \partial A_i(x^*)$ *and* $\lambda_i \geq 0$, *with* $\lambda_i = 0$ *for* $i \notin I(x^*)$.

The following corollary describes the relationship between the BCQ and the strong CHIP.

COROLLARY 5.2. *Let* $x^* \in K$. *Suppose that* A_i, $i = 1, \ldots, m$, *are, in addition, differentiable at* x^*. *Let* $K_S(x^*)$, $K_L(x^*)$, *and* H_i, $i \in I(x^*)$, *be defined as in the previous sections. Then the following statements are equivalent:*

(i) *The system of convex inequalities* (5.1) *satisfies the BCQ relative to* C *at* x^*.
(ii) $\{C, H_1, H_2, \ldots, H_m\}$ *has the strong CHIP at* x^* *and* $K_S(x^*) = K_L(x^*)$.
(iii) *For any* $x \in X$, $P_K(x) = x^* \iff x^* = P_C(x - \sum_1^m \lambda_i \bigtriangledown A_i(x^*))$ *for some* $\lambda_i \geq 0$, *with* $\lambda_i = 0$ *for all* $i \notin I(x^*)$.

Proof. The equivalence of (i) and (iii) is a direct consequence of Theorem 5.1; hence we need only to prove that (ii) is equivalent to (iii). Since K is convex, Theorem 4.1 gives the implication (iii) $\Longrightarrow$ (ii). Conversely, assume that (iii) holds. By Lemma 3.1 of [10], we have that

$$P_K(x) = x^* \Longrightarrow P_{K_L(x^*)}(x) = x^*.$$

Then from Corollary 3.2 it follows that $K_S(x^*) = K_L(x^*)$. Again, using Theorem 4.1, we have that $\{C, H_1, H_2, \ldots, H_m\}$ has the strong CHIP at x^*. The proof is complete. □

Finally, we give an example with nondifferentiable convex constraints.

Example 5.1. Let $X = l^2$ and C be the half-space defined by

$$C = \{x = (x_1, x_2, \ldots) \in l^2 : \ x_1 \leq 0\}.$$

Define

$$A(x) = \sum_{k=1}^{\infty} |x_k| - 1 \qquad \forall x = (x_1, x_2, \ldots) \in l^2,$$

and take $x^* = (x_k^*) \in K$, where

$$x_k^* = \begin{cases} 0 & \text{for } k = 2n+1, \quad n = 0, 1, 2, \ldots, \\ \frac{1}{2^n} & \text{for } k = 2n, \qquad n = 1, 2, \ldots. \end{cases}$$

Then $x^* \in K$, $A(x^*) = 0$, and

$$\partial A(x^*) = \{z = (z_1, z_2, \ldots) : \ z_{2n} = 1, \ z_{2n+1} \in [-1, 1], \ n = 0, 1, 2, \ldots\}.$$

Since the system of convex inequalities $A(x) \leq 0$ satisfies the weak Slater condition on C, it satisfies the BCQ relative to C. Thus, using Theorem 5.1, we get that, for any $x = (x_1, x_2, \ldots) \in l^2$, $P_K(x) = x^*$ if and only if there exists $b \geq 0$ such that

$$x_1 \geq -b, \quad x_{2n} = \frac{1}{2^n} + b, \quad x_{2n+1} \in [-b, b], \qquad n = 1, 2, \ldots.$$

In fact, for any $x \in l^2$, $P_C(x) = x^*$ if and only if $x_1 \geq 0$, $x_k = x_k^*$, $k > 1$. By Theorem 5.1, $P_K(x) = x^*$ if and only if there exist $\lambda \geq 0$ and $t_{2n+1} \in [-1, 1]$ such that $P_C(x - \lambda(t_1, 1, t_3, 1, \ldots)) = x^*$. From this we can deduce our desired result.

6. Concluding remark. Nonlinear best approximation problems in Hilbert spaces have been studied in this paper. As in the case of linear constraints, the strong CHIP is used to characterize the "perturbation property" of best approximations in the case of differentiable constraints. However, this is the first time that the "perturbation property" has been characterized using the generalized BCQ for convex constraints. Our main results are Theorems 4.1 and 5.1. In particular, for both differentiable and convex constraints, the equivalence of the generalized BCQ, the "perturbation property," and the strong CHIP with the constraint qualification condition $K_L(x^*) = K_S(x^*)$ has been obtained. Moreover, some examples with nonlinear constraints have been given to show that our main results genuinely generalize some recent work obtained in [9, 10] on best approximations with linear constraints.

Acknowledgments. We wish to thank the referees for their valuable comments and suggestions. We wish to express our gratitude to Dr. K. F. Ng and Dr. W. Li for their careful reading of drafts of the present paper and for their helpful remarks.

REFERENCES

[1] M. Bazaraa, J. Goode, and C. Shetty, *Constraint qualifications revisited*, Management Sci., 18 (1972), pp. 567–573.
[2] C. De Boor, *On "best" interpolation*, J. Approx. Theory, 16 (1976), pp. 28–48.
[3] B. Brosowski and F. Deutsch, *On some geometric properties of suns*, J. Approx. Theory, 10 (1974), pp. 245–267.
[4] F. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley & Sons, New York, 1983.
[5] C. Chui, F. Deutsch, and J. Ward, *Constrained best approximation in Hilbert space*, Constr. Approx., 6 (1990), pp. 35–64.
[6] C. Chui, F. Deutsch, and J. Ward, *Constrained best approximation in Hilbert space* II, J. Approx. Theory, 71 (1992), pp. 231–238.
[7] 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.
[8] F. Deutsch, W. Li, and J. Swetits, *Fenchel duality and the strong conical intersection property*, J. Optim. Theory Appl., 102 (1997), pp. 681–695.
[9] F. Deutsch, W. Li, and J. Ward, *A dual approach to constrained interpolation from a convex subset of Hilbert space*, J. Approx. Theory, 90 (1997), pp. 385–444.
[10] 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.
[11] F. Deutsch, V. Ubhaya, J. Ward, and Y. Xu, *Constrained best approximation in Hilbert space* III: *Application to n-convex functions*, Constr. Approx., 12 (1996), pp. 361–384.
[12] 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., 86 (1999), pp. 135–160.
[13] J. Hiriart-Urruty and C. Lemarechal, *Convex Analysis and Minimization Algorithms* I, Grundlehren Math. Wiss. 305, Springer, New York, 1993.
[14] O. Mangasarian, *Nonlinear Programming*, McGraw–Hill, New York, 1969.
[15] O. L. Mangasarian and S. Fromovitz, *The Fritz John necessary optimality conditions in the presence of equality constraints*, J. Math. Anal. Appl., 17 (1967), pp. 37–47.
[16] C. Micchelli, P. Smith, J. Swetits, and J. Ward, *Constrained L_p-approximation*, Constr. Approx., 1 (1985), pp. 93–102.
[17] C. A. Micchelli and F. I. Utreras, *Smoothing and interpolation in a convex subset of a Hilbert space*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 728–746.
[18] I. Singer, *Duality for optimization and best approximation over finite intersection*, Numer. Funct. Anal. Optim., 19 (1998), pp. 903–915.
[19] R. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[20] Y. Yuan and W. Sun, *Optimization Theory and Methods*, Science Press, Beijing, 1997 (in Chinese).

SIAM J. OPTIM.
Vol. 13, No. 1, pp. 240–263

RECURSIVE ALGORITHMS FOR STOCK LIQUIDATION: A STOCHASTIC OPTIMIZATION APPROACH*

G. YIN†, R. H. LIU‡, AND Q. ZHANG‡

Abstract. This work develops stochastic optimization algorithms for a class of stock liquidation problems. The stock liquidation rules are based on hybrid geometric Brownian motion models allowing regime changes that are modulated by a continuous-time finite-state Markov chain. The optimal selling policy is of threshold type and can be obtained by solving a set of two-point boundary value problems. The total number of equations to be solved is the same as the number of states of the underlying Markov chain. To reduce the computational burden, using a stochastic optimization approach, recursive algorithms are constructed to approximate the optimal threshold values. Convergence and rates of convergence of the algorithm are studied. Simulation examples are presented, and the computation results are compared with the analytic solutions. Finally, the algorithms are tested using real market data.

Key words. geometric Brownian motion, regime change, stochastic optimization, recursive algorithm

AMS subject classifications. 90A09, 60J10, 60J27, 62L20

PII. S1052623401392901

1. Introduction.

This work is concerned with decision making in stock liquidation, which is crucial in successful trading. In finance literature, the celebrated Black–Scholes model based on geometric Brownian motion (GBM) is widely used in the analysis of options pricing and portfolio management; see Merton [13], among others. This model uses a stochastic differential equation with deterministic expected returns and nonrandom volatility and gives a reasonably good description of the market in a short period of time. However, it has limitations due to its insensitivity to random parameter changes. In fact, one of the reservations from "Wall Street" about using the traditional GBM is that stock price movements are far from being a "random walk" in a longer time horizon. Thus, various modifications to the model have been made. For example, to characterize price movements, Merton [14] considered diffusions with pure jumps, Clark [3] studied time-changed Brownian motions, and Praetz [16] proposed a hyperbolic model in lieu of the traditional log-normal distribution. More recently, Fouque, Papanicolaou, and Sircar [6], Hull [8], and Musiela and Rutkowski [15] have studied stochastic volatility that is dictated by an additional stochastic differential equation. For a complete review of the literature, we refer the reader to the books [4, 8, 9, 10, 15] and the references therein.

One of the main factors that affects decision making in a marketplace is the trend of the stock market. It is necessary to incorporate such trends into models to capture detailed stock price movements. In a recent paper of Zhang [20], a hybrid switching GBM model, i.e., a number of GBMs modulated by a finite-state Markov chain, is

*Received by the editors July 26, 2001; accepted for publication (in revised form) January 24, 2002; published electronically July 16, 2002.

http://www.siam.org/journals/siopt/13-1/39290.html

†Department of Mathematics, Wayne State University, Detroit, MI 48202 (gyin@math.wayne.edu). The research of this author was supported in part by the National Science Foundation under grant DMS-9877090.

‡Department of Mathematics, University of Georgia, Athens, GA 30602 (rliu@math.uga.edu, qingz@math.uga.edu). The research of the second and third authors was supported in part by USAF grant F30602-99-2-0548.

proposed and developed. Such switching processes can be used to represent market trends or the trends of an individual stock. In addition, various economic factors such as interest rates, business cycles, etc. can also be incorporated into the model; see [7, 19] for related references.

In liquidation decision making for a single nondividend stock, a selling rule is determined by two threshold levels, a target price and a stop-loss limit. One makes a selling decision whenever the price reaches either the target price or the stop-loss limit. The objective is to choose these threshold levels so as to maximize an expected return function. In [20], such optimal threshold levels are obtained by solving a set of two-point boundary value problems. In particular, if the underlying Markov chain has only two states, then the corresponding two-point boundary value problem has an analytic solution, and the optimal solution can be obtained in a closed form. However, more detailed market study that requires the underlying Markov chain to have more than two states ($m > 2$) is often necessary. In this case, the computation becomes much more involved because a set of two-boundary value problems must be solved, and a closed-form solution is difficult to obtain, although the existence of solutions was proved in [20]. It is thus of practical interest to find feasible algorithms yielding good approximations to the optimal policy.

With the goal of reducing computational effort, we develop an alternative approach in this work. Focusing our attention on threshold selling rules, in lieu of solving a set of boundary value problems, we formulate the problem as a stochastic optimization procedure and propose a class of stochastic recursive algorithms for resolution. The essential feature of our approach is the use of stochastic approximation methods; see Kushner and Yin [12] for up-to-date development of stochastic approximation algorithms. Recent references on stochastic approximation can also be found in [2].

The rest of the paper is arranged as follows. Section 2 offers a precise formulation of the problem. We introduce the model and present the stochastic optimization algorithms. To maximize the expected return as a function of the threshold values, gradient estimates of the objective function are provided via finite difference methods. One of the advantages of our approach, which is particularly useful for on-line computation, is the simple form and systematic nature of the algorithms. Section 3 describes the asymptotic properties of the algorithm. By virtue of weak convergence methods, we obtain the convergence of the algorithm and ascertain the convergence rates. Section 4 presents modifications and variations of the algorithms, which include projection procedures, gradient estimates made without averaging of the observations, and random direction finite difference gradient estimates. To demonstrate the utility of the algorithms, simulations and numerical experiments are given in section 5. Our simulation study demonstrates that the stochastic optimization algorithms proposed indeed produce good approximation results. For the hybrid GBM model when $m = 2$, a closed-form solution is available, which enables us to make comparisons of the analytic solution with that of the approximation. Next, we proceed to use real market data to test our algorithms, showing that one need not estimate the generator of the "hidden" Markov chain. Only observed data are used in the recursive calculation, which provides an opportunity for on-line implementation. Finally, the paper concludes with some further remarks in section 6.

Before we proceed, some notes on the notation are in order. Throughout the paper, we use K to denote a generic positive constant whose values may be different for different usages. For a suitable function $g(\cdot)$, $g_x(\cdot)$ and $g_{xx}(\cdot)$ denote the gradient and Hessian of g, respectively. For any $z \in \mathbb{R}^{\ell_1 \times \ell_2}$ with some $\ell_1, \ell_2 \geq 1$, z' denotes its transpose. For a vector v, v^i denotes its ith component.

2. Formulation.

2.1. Hybrid GBM model. Suppose that $\alpha(t)$ is a finite-state continuous-time Markov chain with state space $\mathcal{M} = \{1, \ldots, m\}$, which represents market trends and other economic factors. For example, when $m = 2$, $\alpha(t) = 1$ stands for an upward trend, and $\alpha(t) = 2$ a downward trend. We may also consider, for instance, $\alpha(t) = (\alpha_1(t), \alpha_2(t))$, where $\alpha_1(t)$ models the market trends and $\alpha_2(t)$ represents the interest rates at time t. To account for more complex situations and finer distinctions, we assume that the chain has more than two states, i.e., $m \geq 2$ in general. Let $S(t)$ be the price of the stock. We consider a hybrid GBM model, in which $S(t)$ satisfies the stochastic differential equation

$$\frac{dS(t)}{S(t)} = \mu(\alpha(t))dt + \sigma(\alpha(t))dw(t), \quad S(0) = S_0 \ \text{ initial price}, \tag{2.1}$$

where $w(\cdot)$ is a real-valued standard Brownian motion that is independent of $\alpha(\cdot)$. The model can be viewed as a hybrid or switching Black–Scholes model.

Note that in (2.1), both the drift (appreciation rate) and the diffusion coefficient (volatility) depend on the Markov chain $\alpha(t)$. Define another process

$$X(t) = \int_0^t r(\alpha(s))ds + \int_0^t \sigma(\alpha(s))dw(s), \tag{2.2}$$

where

$$r(i) = \mu(i) - \frac{\sigma^2(i)}{2} \quad \text{for each} \ \ i = 1, \ldots, m. \tag{2.3}$$

Using $X(t)$, we can write the solution of (2.1) as $S(t) = S_0 \exp(X(t))$.

Let $z = (z^1, z^2)' \in (0, \infty) \times (0, \infty)$. We consider two barriers or two boundaries of the threshold, a lower boundary $z^1 > 0$ and an upper boundary $z^2 > 0$ such that whenever the stock price reaches the upper bound $S_0 \exp(z^2)$ or the lower bound $S_0 \exp(-z^1)$, we sell the stock to take profit or to prevent further loss.

In what follows, we formulate the task of finding optimal threshold values as an optimization problem. Let τ be a stopping time defined by $\tau = \inf\{t > 0 : S(t) \notin (S_0 \exp(-z^1), S_0 \exp(z^2))\}$, or equivalently, $\tau = \inf\{t > 0 : X(t) \notin (-z^1, z^2)\}$. Noting that τ is independent of S_0, we aim to find the optimal threshold level $z_* = (z_*^1, z_*^2)' \in (0, \infty) \times (0, \infty)$, so that a suitable objective function (the expected return) is maximized. The problem can be rewritten as:

$$\text{Problem } \mathcal{P}: \ \begin{cases} \text{Find argmax } \varphi(z) = E[\Phi(X(\tau)) \exp(-\widetilde{\varrho}\tau)], \\ z = (z^1, z^2)' \in (0, \infty) \times (0, \infty), \end{cases} \tag{2.4}$$

where $\Phi(\cdot)$ is a suitable real-valued function (for example, $\Phi(x) = e^x - 1$), and $\widetilde{\varrho} > 0$ is the discount rate.

Although in simple cases such as $m = 2$ an analytic solution may be available, in general a closed-form solution may be virtually impossible to obtain. Even in the case of $m = 2$, the computation for obtaining the closed-form solution is not simple. Our contribution is to devise a numerical approximation procedure that estimates the optimal lower and upper bounds in a systematic way. We will use a stochastic optimization procedure to resolve the issue by constructing a sequence of estimates of the optimal threshold value z_*, using $z_{n+1} = z_n + \{\text{step size}\} \cdot \{\text{gradient estimate of } \varphi(z)\}$,

where the step size can be either a decreasing sequence of real numbers or a small positive constant.

Remark 2.1. In accordance with (2.4), we need $z_n \in (0,\infty)\times(0,\infty)$. Nevertheless, for ease of presentation, to obtain asymptotic properties of the recursive algorithm, we first work out the details under no constraints ($z_n \in \mathbb{R}^2$). That is, we solve Problem $\mathcal{P}$ defined in (2.4) with the constraint removed. Then, in the subsequent sections, we modify the algorithm by adding the constraints via the use of a projection method. The proofs of the untruncated algorithms are then easily adapted to those of the constraint algorithms.

2.2. Gradient estimates and recursive algorithm. The approximation procedures will depend on how the gradient estimates of $\varphi_z(z)$ are constructed. Let us begin with a simple noisy finite difference scheme. Several of its variants will be discussed in section 4.

Using (2.1), generate a sample path of $X(t)$ that is the solution of (2.2). At time n (n being the iteration number), with the threshold value fixed at $z_n = (z_n^1, z_n^2)' \in \mathbb{R}^2$ (see Remark 2.1), we compute τ_n, the first exit time of $X(t)$ from $I_{z_n} = (-z_n^1, z_n^2)'$ (the interval with the lower and upper boundaries set at $-z_n^1$ and z_n^2, respectively), by

$$\tau_n = \inf\{t > 0 : X(t) \notin I_{z_n}\}. \tag{2.5}$$

2.3. Method 1: Gradient estimates using averaged samples (FDEA). Define a combined process ξ_n that includes the random effects from $X(t)$ and the stopping time τ_n as

$$\xi_n = (X(\tau_n), \tau_n)', \tag{2.6}$$

where $X(\tau_n)$ denotes the random process $X(t)$ stopped at τ_n. In what follows, we call $\{\xi_n\}$ the sequence of collective noise. Let $\widetilde{\varphi}(\cdot,\cdot)$ and $\widehat{\varphi}(\cdot,\cdot)$ be real-valued functions defined on $\mathbb{R}^2 \times \mathbb{R}^1$. When the threshold value is set at z, take random samples of size n_0 with random noise sequences $\{\xi_{n,\ell}^{\pm}\}_{\ell=1}^{n_0}$ such that

$$\widehat{\varphi}(z, \xi_n^{\pm}) \stackrel{\text{def}}{=} \frac{\widetilde{\varphi}(z, \xi_{n,1}^{\pm}) + \cdots + \widetilde{\varphi}(z, \xi_{n,n_0}^{\pm})}{n_0}. \tag{2.7}$$

We assume that

$$E\widehat{\varphi}(z, \xi_n^{\pm}) = \varphi(z) \quad \text{for each } z. \tag{2.8}$$

Then, for each z, $\widehat{\varphi}(z, \xi_n^{\pm})$ is an estimator of $\varphi(z)$. In our simulation study, we can use independent random samples to estimate the mean of $\Phi(X(\tau_n))\exp(-\widetilde{\varrho}\tau_n)$. By the law of large numbers, $\widehat{\varphi}(z, \xi_n)$ converges to $\varphi(z)$ w.p.1 as $n_0 \to \infty$. To allow more flexibility, we will not assume the independence in the proof of convergence theorem, which is useful for dealing with real data. In what follows, in lieu of using (2.7) with $\widetilde{\varphi}(z, \xi_{n,\ell}^{\pm})$, we will work with the form $\widehat{\varphi}(z, \xi_n)$, give conditions needed for obtaining convergence and rate of convergence, and derive the asymptotic properties of the underlying algorithms.

Consider a stochastic approximation procedure with finite difference-type gradient estimates. Use $Y_n^{\pm} = (Y_n^{\pm,1}, Y_n^{\pm,2})' \in \mathbb{R}^2$ to denote the outcomes of two simulation runs or two observations from real data taken at the nth iteration, where $Y_n^{\pm,\iota} = Y^{\pm,\iota}(z_n, \xi_n^{\pm})$ with

$$Y_n^{\pm,\iota}(z, \xi_n^{\pm}) = \widehat{\varphi}(z \pm \delta_n e_\iota, \xi_n^{\pm}) \quad \text{for } \iota = 1, 2, \tag{2.9}$$

e_ι being the standard unit vectors $e_1 = (1,0)'$ and $e_2 = (0,1)'$, and $\xi_n^\pm$ being *two* different (collective) noises taken at the threshold values $z \pm \delta_n e_\iota$, respectively. For notational simplicity, here and hereafter, we often use ξ_n to represent both ξ_n^+ and ξ_n^- whenever there is no confusion. The gradient estimate at time n is given by $D\widehat{\varphi}(z_n, \xi_n) \stackrel{\text{def}}{=} (Y_n^+ - Y_n^-)/(2\delta_n)$. A stochastic optimization algorithm then takes the form

$$z_{n+1} = z_n + \varepsilon_n D\widehat{\varphi}(z_n, \xi_n), \tag{2.10}$$

where $\{\varepsilon_n\}$ is a sequence of real numbers known as step sizes.

To proceed, define

$$\begin{aligned} \rho_n &= (Y_n^+ - Y_n^-) - E_n(Y_n^+ - Y_n^-), \\ \chi_n^\iota &= [E_n Y_n^{+,\iota} - \varphi(z_n + \delta_n e_\iota)] - [E_n Y_n^{-,\iota} - \varphi(z_n - \delta_n e_\iota)], \quad \iota = 1,2, \\ b_n^\iota &= \frac{\varphi(z_n + \delta_n e_\iota) - \varphi(z_n - \delta_n e_\iota)}{2\delta_n} - \varphi_{z^\iota}(z_n), \quad \iota = 1,2, \end{aligned} \tag{2.11}$$

where E_n denotes the conditional expectation with respect to $\mathcal{F}_n$, the σ-algebra generated by $\{z_1, \xi_j^\pm : j < n\}$, $\varphi_{z^\iota}(z) = (\partial/\partial z^\iota)\varphi(z)$, and $\varphi_z(\cdot) = (\varphi_{z^1}(\cdot), \varphi_{z^2}(\cdot))'$ denotes the gradient of $\varphi(\cdot)$. Note that the σ-algebras generated by $\{z_1, \xi_j^\pm : j < n\}$ and $\{z_j, \xi_j^\pm : j < n\}$ are the same (see also [12]). In the above, χ_n^ι and b_n^ι for $\iota = 1,2$ represent the noise and bias, and $\{\rho_n\}$ is a martingale difference sequence. Note that it is reasonable to assume that, after taking conditional expectation, the resulting function is smooth. Thus we separate the noise into two parts, uncorrelated noise $\{\rho_n\}$ and correlated noise $\{\chi_n\}$.

In what follows, whenever we wish to emphasize the dependence on (z, ξ), we spell it out, for example, using the notation $\chi^\iota(z, \xi)$, similar to that of $Y_n^{\pm,\iota}(z, \xi^\pm)$ defined in (2.9). Write $\chi_n = (\chi_n^1, \chi_n^2)'$ and $b_n = (b_n^1, b_n^2)'$, and note that $\chi_n = \chi_n(z_n, \xi_n)$, which will be used in what follows. With the noise $\chi_n(z_n, \xi_n)$ and the bias b_n defined above, algorithm (2.10) becomes

$$z_{n+1} = z_n + \varepsilon_n \varphi_z(z_n) + \varepsilon_n \frac{\rho_n}{2\delta_n} + \varepsilon_n b_n + \varepsilon_n \frac{\chi_n(z_n, \xi_n)}{2\delta_n}. \tag{2.12}$$

3. Asymptotic properties of the recursive algorithm. This section is devoted to the study of asymptotic properties of the recursive stochastic approximation algorithm (2.10). We begin with the study of the convergence of the underlying algorithm. In lieu of dealing with the discrete iterations, we take a continuous-time interpolation leading to a limit ordinary differential equation (ODE). The stationary points of the ODE are the threshold values that we are searching for. Then the rate of convergence is studied via an appropriate scaling. We show that a suitably scaled sequence of the estimation errors converges weakly to a diffusion process. The scaling factor, together with the asymptotic covariance of the limit diffusion, gives us the desired rates of convergence.

3.1. Conditions. To carry out the asymptotic analysis, we define the following:

$$\begin{cases} t_n = \displaystyle\sum_{i=1}^{n-1} \varepsilon_i, \qquad m(t) = \max\{n : t_n \le t\}, \\ N_n = \min\{i : t_{n+i} - t_n \ge T\} \ \text{ for an arbitrary } \ T > 0, \\ z^0(t) = z_n \ \text{ for } \ t \in [t_n, t_{n+1}), \qquad z^n(t) = z^0(t + t_n). \end{cases} \tag{3.1}$$

Note that $z^0(\cdot)$ is a piecewise constant process and $z^n(\cdot)$ is its shift, whose purpose is to bring the asymptotics to the foreground. It follows that the interpolated process $z^n(\cdot)$ can be rewritten as

$$z^n(t) = z_n + \sum_{j=n}^{m(t_n+t)-1} \varepsilon_j \varphi_z(z_j) + \sum_{j=n}^{m(t_n+t)-1} \varepsilon_j \frac{\rho_j}{2\delta_j} + \sum_{j=n}^{m(t_n+t)-1} \varepsilon_j b_j + \sum_{j=n}^{m(t_n+t)-1} \frac{\varepsilon_j}{2\delta_j} \chi_j(z_j, \xi_j). \tag{3.2}$$

Note that $z^n(\cdot) \in D^2[0,\infty)$, the space of $\mathbb{R}^2$-valued functions that are right continuous and have left-hand limits, endowed with the Skorohod topology [5]. To proceed, we need the following conditions.

(A0) The sequences $\{\varepsilon_n\}$ and $\{\delta_n\}$ satisfy $0 < \varepsilon_n \to 0$, $\sum_n \varepsilon_n = \infty$, $0 < \delta_n \to 0$, and $\varepsilon_n/\delta_n^2 \to 0$ as $n \to \infty$. Moreover,

$$\limsup_n \sup_{0 \le i < N_n} \left(\frac{\varepsilon_{n+i}}{\varepsilon_n} \right) < \infty, \qquad \limsup_n \left(\frac{\delta_{n+i}}{\delta_n} \right) < \infty,$$

$$\limsup_n \left[\frac{(\varepsilon_{n+i}/\delta_{n+i}^2)}{(\varepsilon_n/\delta_n^2)} \right] < \infty.$$

(A1) The second derivative $\varphi_{zz}(\cdot)$ is continuous.

(A2) For each compact set G,

(a) $\sup_n E|Y_n^{\pm} I_{\{z_n \in G\}}|^2 < \infty$;

(b) for each z belonging to a bounded set,

$$\sup_n \sum_{j=n}^{n+N_n-1} E^{1/2} \left| E_n \chi_j(z, \xi_j) \right|^2 < \infty, \qquad \lim_n \sup_{0 \le i < N_n} E|\widetilde{\gamma}_i^n| = 0,$$

where

$$\widetilde{\gamma}_i^n = \left(\frac{1}{\varepsilon_{n+i}} \right) \sum_{j=n+i}^{n+N_n-1} \frac{\varepsilon_j}{2\delta_j} E_{n+i}[\chi_j(z_{n+i+1}, \xi_j) - \chi_j(z_{n+i}, \xi_j)], \quad i < N_n.$$

3.2. Remarks on conditions. Strictly speaking, since the step size $\{\varepsilon_n\}$ and the finite difference sequence $\{\delta_n\}$ are at our disposal, (A0) is not a condition or restriction. We put it here as a condition for convenience of presentation. A frequently used choice is $\varepsilon_n = O(1/n)$ and $\delta_n = O(1/n^{1/6})$. In such a case, (A0) is satisfied.

Using (2.11), condition (A2)(a) yields $\sup_n E|\chi_n(z_n, \xi_n) I_{\{z_n \in G\}}|^2 < \infty$. It then follows that for any $z_n \in G$, $\{\chi_n(z_n, \xi_n)\}$ is uniformly integrable. Due to the finite difference approximation, the noise $\chi_n/(2\delta_n)$ has a covariance that is inversely proportional to δ_n. As a result, the algorithm will have a rate of convergence slower than a root-finding stochastic approximation algorithm. One possible choice for diminishing the noise effect is to select a constant step size $\delta_n = \delta$. Then the noise covariance will be reduced at the expense of a nonzero bias.

Smoothness of $\varphi(\cdot)$. Note that our objective is to minimize the function $\varphi(z)$ given in (2.4). Condition (A1) is satisfied in the context of typical stock selling scenarios. For example, taking $\Phi(x) = (\exp(x) - 1)$, it can be shown that the function $\varphi \in C^\infty$. The smoothness of $\varphi(z)$ in (A1) is verifiable for various situations. Here, we give an explicit representation of $\varphi(\cdot)$ for the case $m = 2$ (i.e., the Markov chain

has two states). Let the generator of $\alpha(\cdot)$ be given by

$$Q = \begin{pmatrix} -\lambda_1 & \lambda_1 \\ \lambda_2 & -\lambda_2 \end{pmatrix},$$

with $\lambda_1 > 0$ and $\lambda_2 > 0$. As in [20, p. 74], it can be shown that $\varphi(z)$ is given by

$$\varphi(z) = \sum_{i=1}^{4} c^i(z)(P(\alpha(0)=1) + \kappa^i P(\alpha(0)=2))e^{\eta_i z^1} \tag{3.3}$$

such that η_i $(i = 1, 2, 3, 4)$ are the four real roots of $\psi(\eta) = 0$, with $\psi(\eta)$ given by

$$\psi(\eta) = \frac{\sigma^2(1)\sigma^2(2)}{4}\left\{\left(\eta^2 + \frac{2r(1)}{\sigma^2(1)}\eta - \frac{2(\tilde{\varrho}+\lambda_1)}{\sigma^2(1)}\right)\left(\eta^2 + \frac{2r(2)}{\sigma^2(2)}\eta - \frac{2(\tilde{\varrho}+\lambda_2)}{\sigma^2(2)}\right) - \frac{4\lambda_1\lambda_2}{\sigma^2(1)\sigma^2(2)}\right\};$$

that κ^i for $i = 1, 2, 3, 4$ are defined by $\kappa^i = \frac{1}{\lambda_1}\big(-\frac{\sigma^2(1)}{2}\,\eta_i^2 - r(1)\eta_i + \tilde{\varrho} + \lambda_1\big)$; and that $(c^1, c^2, c^3, c^4) = (c^1(z), c^2(z), c^3(z), c^4(z))$ (as a function of the threshold $z = (z^1, z^2)$) is the unique solution of

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ \kappa^1 & \kappa^2 & \kappa^3 & \kappa^4 \\ e^{\eta_1(z^1+z^2)} & e^{\eta_2(z^1+z^2)} & e^{\eta_3(z^1+z^2)} & e^{\eta_4(z^1+z^2)} \\ \kappa^1 e^{\eta_1(z^1+z^2)} & \kappa^2 e^{\eta_2(z^1+z^2)} & \kappa^3 e^{\eta_3(z^1+z^2)} & \kappa^4 e^{\eta_4(z^1+z^2)} \end{pmatrix}\begin{pmatrix} c^1 \\ c^2 \\ c^3 \\ c^4 \end{pmatrix} = \begin{pmatrix} \Phi(-z^1) \\ \Phi(-z^1) \\ \Phi(z^2) \\ \Phi(z^2) \end{pmatrix}. \tag{3.4}$$

It can be shown as in [20] that the 4×4 matrix on the left-hand side of (3.4) is invertible. Since $\Phi(x) = e^x - 1$ for $x \in \mathbb{R}$, and the inverse of the matrix involves combinations of exponential functions of z^1 and z^2, $c^i(z)$ are infinitely differentiable on $(0,\infty)\times(0,\infty)$ for $i = 1, 2, 3, 4$. Consequently, the infinite differentiability of $\varphi(\cdot)$ on $(0,\infty)\times(0,\infty)$ follows from the differentiability of $c^i(z)$'s and expression (3.3).

Noise conditions. Condition (A2)(b) is essentially a mixing condition. If $E\chi_j(x,\xi_j) = 0$ and $\{\chi_j(z,\xi_j)\}$ is a stationary φ-mixing sequence such that $E|\chi_j(z,\xi_j)|^{2+\Delta} < \infty$ for some $\Delta > 0$ and the mixing measure satisfies a certain summability condition (see also the conditions in (A3) given in the next section), then (A2)(b) is verified.

Take, for instance, $\widetilde{\varphi}(z,\xi) = \varphi(z) + f_0(z)\xi$, where $f_0(\cdot)$ is a bounded and continuous function. Suppose that for a positive integer m_0, $\{\xi^{\pm}_{n,\ell}\}$ are m_0-dependent sequences (see [1, p. 167]), for example, $\xi^{\pm}_{n,\ell} = \sum_{j=0}^{m_0} c_j\zeta^{\pm}_{n-j,\ell}$, where $\{\zeta^{\pm}_{n,\ell}\}$ are martingale difference sequences satisfying $E|\zeta^{\pm}_{n,\ell}|^2 < \infty$. Then $\{\xi^{\pm}_{n,\ell}\}$ are mixing processes, and the mixing measures satisfy $\varpi(j) = 0$ for all $j > m_0$. For each z belonging to a bounded set, it is easily verified that $E|\frac{1}{n_0}\sum_{\ell=1}^{n_0}[\widetilde{\varphi}(z+\delta_n e_\iota, \xi^{\pm}_{n,\ell})]|^2 < \infty$. Thus $E|Y^{\pm,\iota}_n|^2 < \infty$ and (A2)(a) is verified.In addition, for each $j > n$, $E_n\chi_j(z,\xi_j) =$

$E_n\{[E_jY_j^{+,\iota}(z,\xi_j)-\varphi(z+\delta_ne_\iota)]-[E_jY_j^{-,\iota}-\varphi(z-\delta_ne_\iota)]\}=0$. Thus (A2)(b) is also verified.

Next consider the case of $\xi_{n,\ell}^{\pm}=g_0(\zeta_{n,\ell}^{\pm})$, where $g_0(\cdot)$ is a real-valued function and $\{\zeta_{n,\ell}^{\pm}\}$ are homogeneous finite-state Markov chains whose transition matrices are irreducible and aperiodic. In this case, the noise is bounded since the Markov chain takes only finitely many values. Then $\{\xi_{n,\ell}^{\pm}\}$ are ϕ-mixing sequences with exponential mixing rates [1, p. 167], i.e., $\varpi(j)=c_0\varpi^j$ for some $c_0>0$ and some $0<\varpi<1$. Take either $\widetilde{\varphi}(z,\xi)=\varphi(z)+f_1(z)\xi$ with $f_1(\cdot)$ a continuous function and $E\xi_{n,\ell}^{\pm}=0$, or $\widetilde{\varphi}(z,\xi)=\varphi(z)+h_0(z,\xi)$ with $h_0(\cdot,\xi)$ a smooth function for each ξ satisfying $Eh_0(z,\xi_{n,\ell}^{\pm})=0$ for each z. Using the exponential mixing rates, again, it is easily verified that both (A2)(a) and (A2)(b) are satisfied.

3.3. Convergence. To obtain the desired convergence results, we need to prove the tightness of $\{z^n(\cdot)\}$ and then to characterize its weak limit. We proceed by using a truncation device [11, 12]. Let ν be a fixed but otherwise arbitrary positive real number, and $\pi_\nu(\cdot)$ be a smooth function with compact support satisfying $\pi_\nu(z)=1$ when $|z|\le\nu$, and $\pi_\nu(z)=0$ when $|z|\ge\nu+1$. Corresponding to (2.12), define $\{z_n^\nu\}$ recursively by $z_1^\nu=z_1$ and

$$z_{n+1}^\nu=z_n^\nu+\left[\varepsilon_n\varphi_z(z_n^\nu)+\frac{\varepsilon_n}{2\delta_n}\rho_n+\varepsilon_nb_n+\varepsilon_n\frac{\chi_n(z_n^\nu,\xi_n)}{2\delta_n}\right]\pi_\nu(z_n^\nu),\quad n\ge1. \tag{3.5}$$

Define the interpolation as $z^{0,\nu}(t)=z_n^\nu$ for $t\in[t_n,t_{n+1})$ and $z^{n,\nu}(t)=z^{0,\nu}(t_n+t)$. Then $z^{n,\nu}(t)=z^n(t)$ up until the first exit from the ν-sphere $S_\nu=\{z\in\mathbb{R}^2:\ |z|\le\nu\}$. Thus, $z^{n,\nu}(\cdot)$ is a ν-truncation of $z^n(\cdot)$ (see [11, p. 43] and [12, p. 278]).

In view of (A1), the continuity of $\varphi_{zz}(\cdot)$ implies the boundedness of $\varphi_{zz}(z)$ for z in a bounded set. Thus, for each $\iota=1,2$,

$$\begin{aligned}b_n^\iota\pi_\nu(z_n^\nu)&=\left[\frac{\varphi(z_n+\delta_ne_\iota)-\varphi(z_n-\delta_ne_\iota)}{2\delta_n}-\varphi_{z^\iota}(z_n)\right]\pi_\nu(z_n^\nu)\\&=O\left(\frac{|\varphi_{zz}(z_n^+)|\delta_n^2}{2\delta_n}\right)=O(\delta_n),\end{aligned} \tag{3.6}$$

where z_n^+ is on the line segment joining $z_n^\nu-\delta_ne_\iota$ and $z_n^\nu+\delta_ne_\iota$.

In what follows, we first show that the truncated process $\{z^{n,\nu}(\cdot)\}$ is tight in $D^2[0,\infty)$, the space of $\mathbb{R}^2$-valued functions that are right continuous, have left-hand limits, and are endowed with the Skorohod topology. Then we obtain the weak convergence of $z^{n,\nu}(\cdot)$ and characterize the limit as a solution of an ODE. Finally, letting $\nu\to\infty$, we conclude that the untruncated process $z^n(\cdot)$ also converges.

In view of (3.2) and (3.5),

$$z^{n,\nu}(t)=z_n^\nu+\sum_{j=n}^{m(t_n+t)-1}\left(\varepsilon_j\varphi_z(z_j^\nu)+\frac{\varepsilon_j}{2\delta_j}\rho_j+\varepsilon_jb_j+\frac{\varepsilon_j}{2\delta_j}\chi_j(z_j^\nu,\xi_j)\right)\pi_\nu(z_j^\nu).$$

One term that is difficult to deal with is $\sum_{j=n}^{m(t_n+t)-1}(\varepsilon_j/(2\delta_j))\chi_j(z_j^\nu,\xi_j)\pi_\nu(z_j^\nu)$ in the process of averaging. To proceed, we claim that this term has weak limit 0. Working

with $i < N_n$ (or, equivalently, $i \le N_n - 1$), define

$$\begin{aligned}
&\Delta_{n,i} = \sum_{j=n}^{n+i} \frac{\varepsilon_j}{2\delta_j} \chi_j(z_j^\nu, \xi_j)\pi_\nu(z_j^\nu), \quad i < N_n,\\
&\Delta^n(t) = \Delta_{n,i} \quad \text{for } \ t \in [t_{n+i}, t_{n+i+1}),\\
&\Gamma_i^n = \sum_{j=n+i}^{n+N_n-1} \frac{\varepsilon_j}{2\delta_j} E_{n+i}\chi_j(z_{n+i}^\nu, \xi_j)\pi_\nu(z_{n+i}^\nu), \quad i < N_n,\\
&\gamma_i^n = \frac{1}{\varepsilon_{n+i}} \sum_{j=n+i}^{n+N_n-1} \frac{\varepsilon_j}{2\delta_j} E_{n+i}\left[\chi_j(z_{n+i+1}^\nu, \xi_j)\pi_\nu(z_{n+i+1}^\nu) - \chi_j(z_{n+i}^\nu, \xi_j)\pi_\nu(z_{n+i}^\nu)\right],\\
&\qquad\qquad i < N_n.
\end{aligned} \tag{3.7}$$

Note that Γ_i^n and γ_i^n are introduced to facilitate the analysis. Their purpose is to add some perturbations so as to eliminate certain unwanted terms. This follows from the use of perturbed test function methods, which were first introduced to treat problems arising in partial differential equations and later successfully used in stochastic systems. (See [12] for applications in stochastic approximation.)

LEMMA 3.1. *Under* (A0)–(A2), $\Delta^n(\cdot)$ *converges weakly to* 0.

Proof. For each $\kappa > 0$, define $\ell_n^\kappa = n + \min\{i : |\Delta_{n,i}| > \kappa\}$. We first show that for each $\kappa > 0$ the truncated sequence $\{\Delta^{n,\kappa}(\cdot)\}$, defined by

$$\Delta^{n,\kappa}(t) = \sum_{j=n}^{(m(t_n+t)-1)\wedge \ell_n^\kappa} \frac{\varepsilon_j}{2\delta_j} \chi_j(z_j^\nu, \xi_j)\pi_\nu(z_j^\nu),$$

converges weakly to 0, where $(a \wedge b) = \min(a, b)$.

Define

$$\Delta_{n,i}^\kappa = \sum_{j=n}^{(n+i)\wedge \ell_\kappa^n} \frac{\varepsilon_j}{2\delta_j} \chi_j(z_j^\nu, \xi_j)\pi_\nu(z_j^\nu), \quad i < N_n. \tag{3.8}$$

Then

$$\sup_{0\le i<N_n} |\Delta_{n,i}^\kappa| \le \kappa + \sup_{1\le j<N_n} \frac{\varepsilon_{n+j}}{2\delta_{n+j}} |\chi_{n+j}(z_{n+j}^\nu, \xi_{n+j})\pi_\nu(z_{n+j}^\nu)|. \tag{3.9}$$

By virtue of (A0), an application of the Chebyshev inequality yields that for $0 \le j < N_n$ and for any $\mu > 0$,

$$\begin{aligned}
&P\left(\sup_{0\le j<N_n} \frac{\varepsilon_{n+j}}{2\delta_{n+j}} |\chi_{n+j}(z_{n+j}^\nu, \xi_{n+j})\pi_\nu(z_{n+j}^\nu)| \ge \mu\right)\\
&\le \sum_{j=0}^{N_n-1} P\left(\frac{\varepsilon_{n+j}}{2\delta_{n+j}} |\chi_{n+j}(z_{n+j}^\nu, \xi_{n+j})\pi_\nu(z_{n+j}^\nu)| \ge \mu\right)\\
&\le \frac{KT}{\mu^2} O\left(\frac{\varepsilon_n}{\delta_n^2}\right) \sum_{j=0}^{N_n-1} \varepsilon_{n+j} \limsup_n \left[\frac{(\varepsilon_{n+j}/\delta_{n+j}^2)}{(\varepsilon_n/\delta_n^2)}\right] \to 0 \quad \text{as } \ n \to \infty.
\end{aligned} \tag{3.10}$$

Thus for each κ, $\{\Delta^{n,\kappa}(\cdot)\}$ is bounded in probability by virtue of (3.9) and (3.10).

Next we apply the perturbed test function method of [12, Theorem 7.4.3]. Let $\psi(\cdot) \in C_0^2$ (the space of real-valued C^2 functions with compact support). Note that, owing to definition (3.8), $E_{n+i}[\psi(\Delta^\kappa_{n,i+1}) - \psi(\Delta^\kappa_{n,i})] = 0$ for $n+i \geq \ell^n_\kappa$. Thus in what follows, we need consider only $n+i < \ell^n_\kappa$. In this range,

$$E_{n+i}\left[\psi(\Delta^\kappa_{n,i+1}) - \psi(\Delta^\kappa_{n,i})\right] = \psi'_z(\Delta^\kappa_{n,i})\frac{\varepsilon_{n+i}}{2\delta_{n+i}} E_{n+i}\chi_{n+i}(z^\nu_{n+i}, \xi_{n+i})\pi_\nu(z^\nu_{n+i}) + O\left(\frac{\varepsilon^2_{n+i}}{\delta^2_{n+i}}\right)\left|E_{n+i}\chi_{n+i}(z^\nu_{n+i}, \xi_{n+i})\pi_\nu(z^\nu_{n+i})\right|^2.$$

Define the perturbed test function by

$$\psi^n_i = \psi(\Delta^\kappa_{n,i}) + \psi'_z(\Delta^\kappa_{n,i})\Gamma^n_i,$$

where Γ^n_i is given by (3.7). Note that

$$\begin{aligned} &E_{n+i}\left[\psi'_z(\Delta^\kappa_{n,i})\Gamma^n_{i+1} - \psi'_z(\Delta^\kappa_{n,i})\Gamma^n_i\right] \\ &\quad = E_{n+i}\left[\psi'_z(\Delta^\kappa_{n,i+1}) - \psi'_z(\Delta^\kappa_{n,i})\right]\Gamma^n_{i+1} + E_{n+i}\psi'_z(\Delta^\kappa_{n,i})[\Gamma^n_{i+1} - \Gamma^n_i], \end{aligned}$$

that

$$\begin{aligned} E_{n+i}[\psi_z(\Delta^\kappa_{n,i+1}) - \psi'_z(\Delta^\kappa_{n,i})]\Gamma^n_{n+i} &\leq K E_{n+i}|(\Delta^\kappa_{n,i+1} - \Delta^\kappa_{n,i})\Gamma^n_{i+1}| \\ &\leq K\frac{\varepsilon_{n+i}}{2\delta_{n+i}} E_{n+i}|\chi_{n+i}(z^\nu_{n+i}, \xi_{n+i})\pi_\nu(z^\nu_{n+i})||\Gamma^n_{i+1}|, \end{aligned}$$

and that

$$E_{n+i}[\Gamma^n_{i+1} - \Gamma^n_i] = \varepsilon_{n+i}\psi'_z(\Delta^\kappa_{n,i})\gamma^n_i - \frac{\varepsilon_{n+i}}{2\delta_{n+i}} E_{n+i}\chi_j(z^\nu_{n+i}, \xi_{n+i})\pi_\nu(z^\nu_{n+i}).$$

Note also that $\sup_{i<N_n}|\Gamma^n_i| \to 0$ in probability by (A0)–(A2). We write

$$\begin{aligned} (3.11)\qquad &E_{n+i}[\psi^n_{i+1} - \psi^n_i] \\ &= \varepsilon_{n+i} O\left(\frac{\varepsilon_{n+i}}{\delta^2_{n+i}}|\chi_{n+i}(z^\nu_{n+i}, \xi_{n+i})\pi_\nu(z^\nu_{n+i})|^2\right) \\ &\quad + \frac{\varepsilon_{n+i}}{\delta_{n+i}} O\left(\left|E_{n+i}\chi_{n+i}(z^\nu_{n+i}, \xi_{n+i})\pi_\nu(z^\nu_{n+i})\right| \right. \\ &\qquad\qquad \left. \cdot \left|\sum_{j=n+i+1}^{n+N_n} \frac{\varepsilon_j}{2\delta_j} E_{n+i+1}\chi_j(z^\nu_{n+i+1}, \xi_j)\pi_\nu(z^\nu_{n+i+1})\right|\right) \\ &\quad + \varepsilon_{n+i}\psi'_z(\Delta^\kappa_{n,i})\gamma^n_i\pi_\nu(z^\nu_{n+i}). \end{aligned}$$

The use of the truncation function $\pi_\nu(\cdot)$ and (A2) imply that $\{|\chi_n(z^\nu_n, \xi_n)|\pi_\nu(z^\nu_n)\}$ is uniformly integrable. This together with $\varepsilon_n/\delta^2_n \to 0$ then yields that the term on the second line of (3.11) goes to 0 in mean, uniformly in $0 \leq i < N_n$. Using (A2), the term on the third and fourth lines also tends to 0 in mean, uniformly in $0 \leq i < N_n$. Again using (A2), the expectation of the last term is bounded by $O(\varepsilon_n/\delta^2_n) \to 0$, uniformly in $0 \leq i < N_n$. Thus, [12, Theorem 7.4.3] implies that the κ-truncated sequence $\{\Delta^{n,\kappa}(\cdot)\}$ converges weakly to the zero process. Finally, by virtue of [12,

Theorem 7.3.6], the original untruncated sequence $\{\Delta^n(\cdot)\}$ also converges to the zero process. □

THEOREM 3.2. *Assume* (A0)–(A2) *and that* $\{z_n\}$ *is tight in* $\mathbb{R}^2$. *Suppose the differential equation*

$$\dot{z} = \varphi_z(z) \tag{3.12}$$

has a unique solution for each initial condition. Then $z^n(\cdot)$ *converges weakly to* $z(\cdot)$, *the solution of* (3.12). *In addition, suppose that* (3.12) *has a unique stationary point* z_*, *globally asymptotically stable in the sense of Liapunov, and that* $\{s_n\}$ *is a sequence of real numbers satisfying* $s_n \to \infty$. *Then* $z^n(s_n + \cdot)$ *converges weakly to* z_*.

Remark 3.3. In lieu of the tightness assumption of $\{z_n\}$, we could provide a set of sufficient conditions under which we can derive the tightness of $\{z_n\}$ by means of perturbed Liapunov function methods. The basic idea is that we use a Liapunov function $V(\cdot)$ for (3.12) and show via a stability argument that $EV(z_n)$ is bounded. In this process, we need to introduce a small perturbation of the Liapunov function, resulting in the desired cancellation. However, for convenience here, we simply assume that this condition holds.

If the collection of stationary points of (3.12) is not a singleton, we can consider the associated invariant sets of the ODE; further details can be found in [12]. The singleton-set assumption, however, is convenient for the rate of convergence study. A sufficient condition guarantees that the uniqueness of z_* is the convexity of the objective function. As far as applications are concerned, since we are interested in approximated optimal solutions, we will not worry even if the set of minimizers is nonunique.

Proof. The proof is divided into four steps.

Step 1: We will obtain the tightness of $\{z^{n,\nu}(\cdot)\}$. For any $\eta > 0$, $t > 0$, and $0 \le s \le \eta$, by use of (3.5), it is easily seen that

$$z^{n,\nu}(t+s) - z^{n,\nu}(t) = \widetilde{z}^{n,\nu}(t+s) - \widetilde{z}^{n,\nu}(t) + o_n(1), \tag{3.13}$$

where

$$\widetilde{z}^{n,\nu}(t+s) - \widetilde{z}^{n,\nu}(t) = \sum_{j=m(t_n+t)}^{m(t_n+t+s)-1} \varepsilon_j \left[\varphi_z(z_j^\nu) + \frac{\rho_j}{2\delta_j} + b_j\right] \pi_\nu(z_j^\nu), \tag{3.14}$$

and $o_n(1)$ converges weakly (or in probability) to 0 as $n \to \infty$. In accordance with [11, Lemma 5, p. 50], the tightness of $\{z^{n,\nu}(\cdot)\}$ will follow from that of $\{\widetilde{z}^{n,\nu}(\cdot)\}$.

Use E_t^n to denote the conditional expectation with respect to $\mathcal{F}_t^n = \sigma\{z^n(s) : s \le t\}$. Then

$$\begin{aligned}
&E_t^n \left|\widetilde{z}^{n,\nu}(t+s) - \widetilde{z}^{n,\nu}(t)\right|^2 \\
&\quad \le K E_t^n \left| \sum_{j=m(t_n+t)}^{m(t_n+t+s)-1} \varepsilon_j \varphi_z(z_j^\nu)\pi_\nu(z_j^\nu) \right|^2 + K E_t^n \left| \sum_{j=m(t_n+t)}^{m(t_n+t+s)-1} \varepsilon_j \frac{\rho_j}{2\delta_j} \pi_\nu(z_j^\nu) \right|^2 \\
&\qquad + K E_t^n \left| \sum_{j=m(t_n+t)}^{m(t_n+t+s)-1} \varepsilon_j b_j \pi_\nu(z_j^\nu) \right|^2 \\
&\quad \le O\left(\left(\sum_{j=m(t_n+t)}^{m(t_n+t+s)-1} \varepsilon_j \right)^2 + \sum_{j=m(t_n+t)}^{m(t_n+t+s)-1} \frac{\varepsilon_j^2}{4\delta_j^2} \right)^2 .
\end{aligned}$$

Thus $\lim_{\eta\to 0}\limsup_n EE_t^n|\widetilde{z}^{n,\nu}(t+s)-\widetilde{z}^{n,\nu}(t)|^2=0$. By the tightness criterion (see [5, section 3.8, p. 132] and [11, p. 47]), $\{\widetilde{z}^{n,\nu}(\cdot)\}$ is tight and so is $\{z^{n,\nu}(\cdot)\}$.

Step 2: We obtain the weak convergence of $\{z^{n,\nu}(\cdot)\}$. Since $\{z^{n,\nu}(\cdot)\}$ is tight, we can extract convergent subsequences. Select such a subsequence and, for simplicity, still denote it by $z^{n,\nu}(\cdot)$. Choose a sequence of positive real numbers $\{\varsigma_n\}$ such that

$$\varsigma_n\to 0 \quad\text{and}\quad \frac{1}{\varsigma_n}\sum_{j=m(t_n+t+l\varsigma_n)}^{m(t_n+t+(l+1)\varsigma_n)-1}\varepsilon_j\to 1 \quad\text{as}\quad n\to\infty.$$

For $t,s>0$, by virtue of (3.6), as $n\to\infty$,

$$E\left|\sum_{j=m(t_n+t)}^{m(t_n+t+s)-1}\varepsilon_j b_j\pi_\nu(z_j^\nu)\right|\le K\left(\sum_{j=m(t_n+t)}^{m(t_n+t+s)-1}\varepsilon_j\right)O(\delta_{m(t_n+t)})\to 0.$$

Thus the last term in the square brackets of (3.14) has limit 0. Since $\{\rho_n\}$ is a martingale difference sequence,

$$E\left|\sum_{j=m(t_n+t)}^{m(t_n+t+s)-1}\frac{\varepsilon_j}{2\delta_j}\rho_j\pi_\nu(z_j^\nu)\right|^2=\sum_{j=m(t_n+t)}^{m(t_n+t+s)-1}\frac{\varepsilon_j^2}{4\delta_j^2}E|\rho_j\pi_\nu(z_j^\nu)|^2\to 0\quad\text{as } n\to\infty.$$

Therefore, the second term in the square brackets of (3.14) goes to 0 in probability uniform in t as $n\to\infty$. As for the first term on the right-hand side of (3.14),

$$\sum_{j=m(t_n+t)}^{m(t_n+t+s)-1}\varepsilon_j\varphi_z(z_j^\nu)\pi_\nu(z_j^\nu)=\sum_l\varsigma_n\frac{1}{\varsigma_n}\sum_{j=m(t_n+t+l\varsigma_n)}^{m(t_n+t+(l+1)\varsigma_n)-1}\varepsilon_j\varphi_z(z_j^\nu)\pi_j(z_j^\nu).$$

Since $\varphi_z(\cdot)$ and $\pi_\nu(\cdot)$ are smooth functions, we can deduce that the limit of $\frac{1}{\varsigma_n}\sum_{j=m(t_n+t+l\varsigma_n)}^{m(t_n+t+(l+1)\varsigma_n)-1}\varepsilon_j\varphi_z(z_j^\nu)\pi_\nu(z_j^\nu)$ is the same as that of

$$\frac{1}{\varsigma_n}\sum_{j=m(t_n+t+l\varsigma_n)}^{m(t_n+t+(l+1)\varsigma_n)-1}\varepsilon_j\varphi_z(z^\nu_{m(t_n+t+l\varsigma_n)})\pi_\nu(z^\nu_{m(t_n+t+l\varsigma_n)})$$

as $n\to\infty$. Fix $\widetilde{u}$ and let $t_{m(t_n+t+l\varsigma_n)}\to\widetilde{u}$ as $n\to\infty$. We need verify only that

$$\frac{1}{\varsigma_n}\sum_{j=m(t_n+t+l\varsigma_n)}^{m(t_n+t+(l+1)\varsigma_n)-1}\varepsilon_j\varphi_z(z^\nu_{m(t_n+t+l\varsigma_n)})\pi_\nu(z^\nu_{m(t_n+t+l\varsigma_n)})\to\varphi_z(z^\nu(\widetilde{u}))\pi_\nu(z^\nu(\widetilde{u})) \tag{3.15}$$

in probability as $n\to\infty$. For each $\eta>0$, choose a finite number of disjoint sets B_ι^η, $\iota=1,\dots,r$, such that $\cup_{\iota=1}^r B_\iota^\eta$ contains the range of $\{z_n^\nu\}$ and

$$P(z^{n,\nu}(\widetilde{u})\in\partial B_\iota^\eta)=0\quad\text{and}\quad \operatorname{diam}(B_\iota^\eta)\le\eta, \tag{3.16}$$

where ∂B_ι^η denotes the boundary of B_ι^η. Select a point $z_\iota^\eta\in B_\iota^\eta$. Then the term on the left-hand side of (3.15) can be approximated by using a small $\eta>0$ via

$$\sum_{\iota=1}^r I_{\{z^{n,\nu}(\widetilde{u})\in B_\iota^\eta\}}\frac{1}{\varsigma_n}\sum_{j=m(t_n+t+l\varsigma_n)}^{m(t_n+t+(l+1)\varsigma_n)-1}\varepsilon_j\varphi_z(z_\iota^\eta)\pi_\nu(z_\iota^\eta).$$

The interpolation and the choice of ς_n then lead to

$$\sum_l \varsigma_n \frac{1}{\varsigma_n} \sum_{j=m(t_n+t+l\varsigma_n)}^{m(t_n+t+(l+1)\varsigma_n)-1} \varepsilon_j \varphi_z(z_j^\nu)\pi_\nu(z_j^\nu) \to \int_t^{t+s} \varphi_z(z^\nu(\widetilde{u}))\pi_\nu(z^\nu(\widetilde{u}))d\widetilde{u}.$$

Therefore, we obtain that $z^{n,\nu}(\cdot)$ converges weakly to $z^\nu(\cdot)$ as $n \to \infty$ such that

$$z^\nu(t+s) - z^\nu(t) = \int_t^{t+s} \varphi_z(z^\nu(\widetilde{u}))\pi_\nu(z^\nu(\widetilde{u}))d\widetilde{u},$$

and thus the mean ODE $\dot{z}^\nu(t) = \varphi_z(z^\nu(t))\pi_\nu(z^\nu(t))$ is obtained.

Step 3: We obtain the convergence of $z^n(\cdot)$. In fact, we need show only that for each $T > 0$,

$$\limsup_{\nu\to\infty} \limsup_{n\to\infty} P(z^{n,\nu}(t) \neq z^n(t) \text{ for some } t \le T) = 0.$$

This follows from the argument of [12, p. 250]; the details are omitted.

Step 4: We show the convergence of $z^n(s_n + \cdot)$. Choose $T > 0$ and select a convergent subsequence $\{(z^n(s_n + \cdot), z^n(s_n - T + \cdot))\}$ with limit $(z(\cdot), z_T(\cdot))$. It is easily seen that $z(0) = z_T(T)$. The value of $z_T(0)$ may not be known, but the collection of possible $\{z_T(0)\}$ over all T and all convergent subsequences belongs to a set that is tight. The stability of the ODE then implies that for any $\eta > 0$ there is a $0 < T_\eta < \infty$ such that for all $T > T_\eta$, $P(z_T(T) \in N(z_*, \eta)) \ge 1 - \eta$, where $N(z_*, \eta)$ denotes a neighborhood of z_* with radius η. This yields the desired result. □

3.4. Rate of convergence. This section is devoted to the rate of convergence of the algorithm (2.10). The question is studied through a suitably scaled sequence $n^{\kappa_0}(z_n - z_*)$ of the estimation errors, where $\kappa_0 > 0$. Taking $\varepsilon_n = 1/n^{\kappa_1}$ and $\delta_n = \delta/n^{\kappa_2}$ for some $0 < \kappa_2 < \kappa_1 \le 1$ and $\delta > 0$, it is known that the optimal choice is given by $\kappa_0 + \kappa_2 = \kappa_1/2$ and $\kappa_0 = 2\kappa_2$. To be more specific, we take $\varepsilon_n = 1/n$ in what follows. Then $\delta_n = \delta/n^{1/6}$ and $\kappa_0 = 1/3$. To proceed, we define $u_n = n^{1/3}(z_n - z_*)$ and assume that the following condition holds.

(A3) Assume that (A1) and (A2) hold, $z_n \to z_*$ in probability, and $\varphi_{zzz}(\cdot)$ exists and is continuous in a neighborhood of z_*. In addition,
  (a) $\{u_n\}$ is tight;
  (b) all eigenvalues of $\varphi_{zz}(z_*) + (1/3)I$ have negative real parts;
  (c) for each z,

$$\chi_n(z,\xi) = \chi_n(z_*,\xi) + \chi_{n,z}(z_*,\xi)(z - z_*) + \left(\int_0^1 [\chi_{n,z}(z_* + (z_n - z_*)s,\xi) - \chi_{n,z}(z_*,\xi)]ds\right)(z - z_*);$$

  (d) $\{\chi_n(z_*,\xi_n)\}$ is a stationary φ-mixing sequence such that $E|\chi_n(z_*,\xi_n)|^{2+\Delta} < \infty$ for some $\Delta > 0$ and $E\chi_n(z_*,\xi_n) = 0$ and that the mixing measure $\varpi(\cdot)$ is given by

$$\varpi(j) = \sup_{A\in\mathcal{F}^{n+j}} E^{(1+\Delta)/(2+\Delta)}|P(A|\mathcal{F}_n) - P(A)|^{(2+\Delta)/(1+\Delta)},$$

  satisfying $\sum_{j=1}^\infty (\varpi(j))^{\Delta/(1+\Delta)} < \infty$.

Remark 3.4. Similar to what was mentioned in Remark 3.3, (A3)(a) can be obtained by perturbed Liapunov function methods (see [12, section 10.4]). The existence and continuity of $\varphi_{zzz}(\cdot)$ in a neighborhood of z_* allows us to linearize $\varphi(\cdot)$ about z_*. Conditions (A3)(c) and (d) concern the sequence $\chi_n(z,\xi)$. It is important to note that due to the use of the stopping time τ, the $\widehat{\varphi}(z,\xi)$ defined in (2.8) may not be continuous in z. However, we can assume that its expectation is smooth. As with the comments given in section 5, it is easily verified that $\chi_n(z_*,\xi_n)$ is a zero mean sequence and is φ-mixing (in fact, n_0-dependent). Thus (A3)(d) is verified.

Set

$$\rho_n^{*,\iota} = \widehat{\varphi}(z_* + \delta_n e_\iota, \xi_n^+) - \widehat{\varphi}(z_* - \delta_n e_\iota, \xi_n^-) \quad \text{and} \quad \rho_n^* = (\rho_n^{*,1}, \rho_n^{*,2})'.$$

With the stipulation of (2.9), the integrability and the convergence of z_n to z_* imply that

$$E|\rho_n - \rho_n^*|^2 \to 0 \quad \text{as } n \to \infty. \tag{3.17}$$

To proceed, let us state a lemma concerning the asymptotic normality of a scaled sequence of $\{\chi_j(z_*,\xi_j) + \rho_j^*\}$.

LEMMA 3.5. *Under* (A3),

(a) *the following inequalities hold:*

$$|E\chi_j(z_*,\xi_j)\chi_k(z_*,\xi_k)| \le K\left(\varpi(j)\right)^{\Delta/(1+\Delta)},$$
$$E|E(\chi_{n+j}(z_*,\xi_{n+j})|\mathcal{F}_n)| \le K\left(\varpi(j)\right)^{\Delta/(1+\Delta)};$$

(b) *the sequence* $\sum_{j=n}^{m(t_n+t)-1}(\chi_j(z_*,\xi_j)+\rho_j^*)/\sqrt{j}$ *converges weakly to an* $\mathbb{R}^2$*-valued Brownian motion* $\widetilde{w}(\cdot)$ *with covariance* Σt.

Remark 3.6. Note that the proof of part (a) of the lemma follows that of [5, Propositions 7.2.2 and 7.2.4]; part (b) can be proved similarly to [5, Theorem 7.3.1].

Using (2.10), (A1), and $\delta_n = \delta/n^{1/6}$, we obtain

$$\begin{aligned} z_{n+1} - z_* = z_n - z_* &+ \frac{1}{n}\varphi_{zz}(z_*)(z_n - z_*) + \frac{1}{n^{5/6}}\frac{\rho_n}{2\delta} + \frac{1}{n}b_n + \frac{1}{n^{5/6}}\frac{\chi_n(z_n,\xi_n)}{2\delta} \\ &+ \frac{1}{n}\left(\int_0^1 (z_n - z_*)'\varphi_{zzz}(z_* + (z_n - z_*)s)ds\right)(z_n - z_*). \end{aligned} \tag{3.18}$$

Without loss of generality, assume that $\{u_n\}$ is bounded; otherwise, we can use a truncation device as in the proof of the convergence of the algorithm in the previous section. Then we show that the truncated process converges, and finally we conclude that the untruncated process also converges. By virtue of (A3) and using $((n+1)/n)^{1/3} = 1 + (1/(3n)) + O(1/n^2)$, this leads to

$$\begin{aligned} u_{n+1} = u_n &+ \frac{1}{n}\left(\varphi_{zz}(z_*) + \left(\frac{1}{3}\right)I\right)u_n + \frac{1}{n}(n^{1/3}b_n) + \frac{1}{\sqrt{n}}\frac{1}{2\delta}(\chi_n(z_*,\xi_n) + \rho_n^*) \\ &+ \left(\frac{n+1}{n}\right)^{1/3}\frac{1}{n}\left(\int_0^1 u_n'\varphi_{zzz}\left(z_* + \left(\frac{s}{n^{1/3}}\right)u_n\right)ds\right)u_n \\ &+ \left(\frac{n+1}{n}\right)^{1/3}\frac{1}{n^{7/6}}\left(\int_0^1 \left[\chi_{n,z}\left(z_* + \left(\frac{s}{n^{1/3}}\right)u_n, \xi_n\right) - \chi_{n,z}(z_*,\xi_n)\right]ds\right)u_n \\ &+ \left(\frac{n+1}{n}\right)^{1/3}\frac{1}{\sqrt{n}}\frac{1}{2\delta}(\rho_n - \rho_n^*) + \frac{1}{n}o(1 + |u_n|). \end{aligned} \tag{3.19}$$

Take a piecewise constant interpolation

$$u^0(t) = u_n, \qquad t \in [t_n, t_{n+1}), \qquad \text{and} \quad u^n(t) = u^0(t_n + t).$$

It can be demonstrated that, using the definition of interpolation in (3.19), the following three terms,

$$\sum_{j=m(t_n+t)}^{m(t_n+t+s)-1} \left(\frac{j+1}{j}\right)^{1/3} \frac{1}{j} \left(\int_0^1 u_j' \varphi_{zzz}\left(z_* + \left(\frac{s}{j^{1/3}}\right) u_j\right) ds\right) u_j,$$

$$\sum_{j=m(t_n+t)}^{m(t_n+t+s)-1} \left(\frac{j+1}{j}\right)^{1/3} \frac{1}{j^{7/6}} \left(\int_0^1 \left[\chi_{j,z}\left(z_* + \left(\frac{s}{j^{1/3}}\right) u_j, \xi_j\right) - \chi_{j,z}(z_*, \xi_j)\right] ds\right) u_j,$$

$$\sum_{j=m(t_n+t)}^{m(t_n+t+s)-1} \frac{1}{j} o(1 + |u_j|),$$

are asymptotically unimportant and contribute a limit 0 in distribution. Furthermore, noting that $\{\rho_n - \rho_n^*\}$ is a martingale difference sequence,

$$\begin{aligned} &E\left|\sum_{j=m(t_n+t)}^{m(t_n+t+s)-1} \left(\frac{n+1}{n}\right)^{1/3} \frac{1}{\sqrt{j}}(\rho_j - \rho_j^*)\right|^2 \\ &\quad = \sum_{j=m(t_n+t)}^{m(t_n+t+s)-1} \left(\frac{n+1}{n}\right)^{2/3} \frac{1}{j} E|\rho_j - \rho_j^*|^2 \to 0 \quad \text{as } n \to \infty \end{aligned}$$

by (3.17). Then we arrive at

$$\begin{aligned} u^n(t+s) - u^n(t) \;=\; &\sum_{j=m(t_n+t)}^{m(t_n+t+s)-1} \frac{1}{j}\left(\varphi_{zz}(z_*) + \left(\frac{1}{3}\right) I\right) u_j + \sum_{j=m(t_n+t)}^{m(t_n+t+s)-1} \frac{1}{j}(j^{1/3} b_j) \\ &+ \sum_{j=m(t_n+t)}^{m(t_n+t+s)-1} \frac{1}{\sqrt{j}} \frac{1}{2\delta}(\chi_j(z_*, \xi_j) + \rho_j^*) + o(1), \end{aligned} \tag{3.20}$$

where $o(1) \to 0$ in probability uniformly in t.

We proceed to establish the tightness of $u^n(\cdot)$. Under the boundedness of $\{u_n\}$ (being assumed for ease of presentation), for each $\eta > 0$, $t, s > 0$ with $s < \eta$, we obtain

$$\begin{aligned} &E_t^n |u^n(t+s) - u^n(t)|^2 \\ &\le K E_t^n \left|\sum_{j=m(t_n+t)}^{m(t_n+t+s)-1} \frac{1}{j}\left(\varphi_{zz}(z_*) + \left(\frac{1}{3}\right) I\right) u_j\right|^2 + K E_t^n \left|\sum_{j=m(t_n+t)}^{m(t_n+t+s)-1} \frac{1}{j}(j^{1/3} b_j)\right|^2 \\ &\quad + K E_t^n \left|\sum_{j=m(t_n+t)}^{m(t_n+t+s)-1} \frac{1}{\sqrt{j}}(\chi_j(z_*, \xi_j) + \rho_j^*)\right|^2 + o(1) \\ &\le K s^2 + K \sum_{j=m(t_n+t)}^{m(t_n+t+s)-1} \sum_{k>j} \frac{1}{\sqrt{j}} \frac{1}{\sqrt{k}} |E_t^n[\chi_j(z_*, \xi_j) + \rho_j^*][E_{j+1}(\chi_k(z_*, \xi_k) + \rho_k^*)]| + o(1), \end{aligned} \tag{3.21}$$

where $o(1) \to 0$ as $n \to \infty$. An application of part (a) of Lemma 3.5 yields $\lim_{\eta\to 0} \limsup_{n\to\infty} EE_t^n|u^n(t+s) - u^n(t)|^2 = 0$. Hence the tightness follows.

We proceed to characterize the limit process $u(\cdot)$. Consider the bias term. By virtue of (A1) and (A3), a Taylor expansion of b_n^ι leads to

$$\begin{aligned} b_n^\iota &= \frac{\varphi(z_n + \delta_n e_\iota) - \varphi(z_n - \delta_n e_\iota)}{2\delta_n} - \varphi_{z^\iota}(z_n) \\ &= \frac{1}{3!}\varphi_{z^\iota,z^\iota,z^\iota}(z_*) + \frac{1}{3!}\left(\varphi_{z^\iota,z^\iota,z^\iota}(z_n) - \varphi_{z^\iota,z^\iota,z^\iota}(z_*)\right) + o(\delta_n^2), \end{aligned}$$

and the last term above goes to 0 in mean and hence in probability as $n \to \infty$. Noting that $\delta_n^2 = \delta^2/n^{1/3}$, we thus have

$$\begin{aligned} &\sum_{j=m(t_n+t)}^{m(t_n+t+s)-1} \frac{1}{j} j^{1/3} b_j \\ &\quad= \frac{\delta}{3!} \sum_{j=m(t_n+t)}^{m(t_n+t+s)-1} \frac{1}{j} \begin{pmatrix} \varphi_{z^1,z^1,z^1}(z_j) \\ \varphi_{z^2,z^2,z^2}(z_j) \end{pmatrix} + \sum_{j=m(t_n+t)}^{m(t_n+t+s)-1} \frac{1}{j} o(1) \\ &\quad= \frac{\delta}{3!} \sum_l \varsigma_n \frac{1}{\varsigma_n} \sum_{j=m(t_n+t+l\varsigma_n)}^{m(t_n+t+(l+1)\varsigma_n)-1} \frac{1}{j} \begin{pmatrix} \varphi_{z^1,z^1,z^1}(z_*) \\ \varphi_{z^2,z^2,z^2}(z_*) \end{pmatrix} + \sum_{j=m(t_n+t)}^{m(t_n+t+s)-1} \frac{1}{j} o(1) \\ &\qquad+ \frac{\delta}{3!} \sum_l \varsigma_n \frac{1}{\varsigma_n} \sum_{j=m(t_n+t+l\varsigma_n)}^{m(t_n+t+(l+1)\varsigma_n)-1} \frac{1}{j} O(|\varphi_{zzz}(z_n) - \varphi_{zzz}(z_*)|) \\ &\quad= \frac{\delta}{3!} \sum_l \begin{pmatrix} \varphi_{z^1,z^1,z^1}(z_*) \\ \varphi_{z^2,z^2,z^2}(z_*) \end{pmatrix} \varsigma_n \left[\frac{1}{\varsigma_n} \sum_{j=m(t_n+t+l\varsigma_n)}^{m(t_n+t+(l+1)\varsigma_n)-1} \frac{1}{j} \right] + o(1), \end{aligned}$$

where $o(1) \to 0$ in probability as $n \to \infty$ by virtue of $z_n \to z_*$ in probability and the continuity of $\varphi_{zzz}(\cdot)$ (in a neighborhood of z_*). Moreover, since $(1/\varsigma_n) \sum_{j=m(t_n+t+l\varsigma_n)}^{m(t_n+t+(l+1)\varsigma_n)-1} \cdot(1/j) \to 1$ as $n \to \infty$, the limit of the next-to-last term yields

$$\sum_{j=m(t_n+t)}^{m(t_n+t+s)-1} \frac{1}{j} j^{1/3} b_j \to \delta^2 \begin{pmatrix} \varphi_{z^1,z^1,z^1}(z_*) \\ \varphi_{z^2,z^2,z^2}(z_*) \end{pmatrix} s.$$

We summarize what has been proved via the following theorem.

THEOREM 3.7. *Assume that* (A3) *holds. Then* $u^n(\cdot)$ *converges weakly to a diffusion process* $u(\cdot)$ *that is a solution of the stochastic differential equation*

$$du = \left\{ \left(\varphi_{zz}(z_*) + \frac{I}{3} \right) u + \frac{\delta^2}{3!} \begin{pmatrix} \varphi_{z^1,z^1,z^1}(z_*) \\ \varphi_{z^2,z^2,z^2}(z_*) \end{pmatrix} \right\} dt + \frac{1}{2\delta} d\widetilde{w}, \tag{3.22}$$

where $\widetilde{w}(\cdot)$ *is the Brownian motion with covariance* $\Sigma^{1/2}(\Sigma^{1/2})'t = \Sigma t$ *given by Lemma* 3.5.

Remark 3.8. Since (3.22) is linear, it has a unique solution for each initial condition. Note that (3.22) includes a nonzero bias

$$(\delta^2/3!) \begin{pmatrix} \varphi_{z^1,z^1,z^1}(z_*) \\ \varphi_{z^2,z^2,z^2}(z_*) \end{pmatrix}.$$

As a direct consequence of Theorem 3.7, $n^{1/3}(z_n - z_*)$ is asymptotically normally distributed with mean

$$\left(\varphi_{zz}(z_*) + \frac{I}{3}\right)^{-1} \frac{\delta^2}{3!} \begin{pmatrix} \varphi_{z^1,z^1,z^1}(z_*) \\ \varphi_{z^2,z^2,z^2}(z_*) \end{pmatrix}$$

and asymptotic covariance $\widetilde{\Sigma}$, where the covariance is given by

$$\widetilde{\Sigma} = \int_0^\infty \exp\left(\left(\varphi_{zz}(z_*) + \frac{I}{3}\right)t\right) \Sigma \exp\left(\left(\varphi'_{zz}(z_*) + \frac{I}{3}\right)t\right) dt.$$

Note that, due to (A3)(b), the integral above is well defined.

If in lieu of $\varepsilon_n = 1/n$ we use $\varepsilon_n = 1/n^{\kappa_1}$ with $\kappa_1 < 1$, then the limit differential equation becomes

$$du = \left\{\varphi_{zz}(z_*)u + \frac{\delta^2}{3!} \begin{pmatrix} \varphi_{z^1,z^1,z^1}(z_*) \\ \varphi_{z^2,z^2,z^2}(z_*) \end{pmatrix}\right\} dt + \frac{1}{2\delta} d\widetilde{w}.$$

In this case, assuming that $\varphi_{zz}(z_*)$ is stable, we have that $n^{\kappa_0}(z_n - z_*)$ is asymptotically normal with a mean equal to the bias and asymptotic covariance given by

$$\widetilde{\Sigma} = \int_0^\infty \exp(\varphi_{zz}(z_*)t)\Sigma \exp(\varphi'_{zz}(z_*)t)dt.$$

4. Variants of algorithms. This section is devoted to variants of the stochastic optimization algorithms. Using essentially the same approach as in the previous section, we can obtain the convergence and rate of convergence results. In order not to dwell on it, we present the results but omit the details.

4.1. Projection algorithm. To ensure the boundedness of the iterates, one often uses a projection scheme. In addition, the projection schemes may also be used in conjunction with constrained algorithms. For example, one may decide to set a lower bound for the algorithm, e.g., one might choose to sell the stock if there is a 20% loss. Let the boundaries be set so that for all n, $z_n^1 \in [\theta_l^1, \theta_u^1]$ and $z_n^2 \in [\theta_l^2, \theta_u^2]$, where θ_l^ι and θ_u^ι for $\iota = 1, 2$ are some predetermined values. In accordance with Remark 2.1, with the requirement $z_n^i > 0$, we choose the lower boundaries to be strictly positive. That is, $\theta_l^\iota > 0$ for $\iota = 1, 2$. Thus, the constraint region satisfies $[\theta_l^1, \theta_u^1] \times [\theta_l^2, \theta_u^2] \subset (0, \infty) \times (0, \infty)$. The resulting stochastic approximation algorithm becomes one with a projection

$$z_{n+1} = \Pi[z_n + \varepsilon_n D\widehat{\varphi}(z_n, \xi_n)], \tag{4.1}$$

or in component form,

$$z_{n+1}^\iota = \Pi_{[\theta_l^\iota, \theta_u^\iota]}[z_n^\iota + \varepsilon_n (D\widehat{\varphi}(z_n, \xi_n))^\iota], \quad \text{for} \quad \iota = 1, 2,$$

where for each real value x,

$$\Pi_{[\theta_l^\iota, \theta_u^\iota]} x = \begin{cases} \theta_l^\iota & \text{if } x < \theta_l^\iota, \\ \theta_u^\iota & \text{if } x > \theta_u^\iota, \\ x & \text{otherwise.} \end{cases}$$

The idea can be explained as follows. For component ι, after the increment $z_n^\iota + \varepsilon_n D\widehat{\varphi}(z_n^\iota, \xi_n)$ is computed, we compare its value with the bounds θ_l^ι and θ_u^ι. If the

increment is smaller than the lower value θ_l^ι, reset the value to θ_l^ι; if it is larger than the upper value θ_u^ι, reset its value to θ_u^ι; otherwise, keep its value as it was.

Introduce a correction term r_n by defining $\varepsilon_n r_n = z_{n+1} - z_n - \varepsilon_n D\widehat{\varphi}(z_n, \xi_n)$, which is the vector of shortest Euclidean length needed to bring $z_n + \varepsilon_n D\widehat{\varphi}(z_n, \xi_n)$ back to the constraint set $[\theta_l^1, \theta_u^1] \times [\theta_l^2, \theta_u^2]$ if it is outside this set. Thus, (4.1) can be rewritten as

$$z_{n+1} = z_n + \varepsilon_n D\widehat{\varphi}(z_n, \xi_n) + \varepsilon_n r_n. \tag{4.2}$$

We can then carry out the analysis as in the previous section, with the modification of the added correction. In addition to the other interpolations defined in the last section, define $\widetilde{r}^0(t) = \sum_j^{m(t)-1} \varepsilon_j r_j$ and $\widetilde{r}^n(t) = \widetilde{r}^0(t + t_n) - \widetilde{r}^0(t_n)$. Then we can establish the tightness of $(z^n(\cdot), \widetilde{r}^n(\cdot))$ and show that $(z^n(\cdot), \widetilde{r}^n(\cdot))$ converges weakly to $(z(\cdot), \widetilde{r}(\cdot))$ such that $\widetilde{r}(t) = \int_0^t r(s)ds$ and $r(t) = 0$ when $z(t) \in [\theta_l^1, \theta_u^1] \times [\theta_l^2, \theta_u^2]$. The details can be worked out in light of [12, Chapters 5 and 8]. We obtain the following result.

PROPOSITION 4.1. *The following assertions hold.*

(a) *Assume* (A0)–(A2) *and* $z_* \in (\theta_l^1, \theta_u^1) \times (\theta_l^2, \theta_u^2)$. *Then the conclusions of Theorem* 3.2 *continue to hold, with* (3.12) *replaced by the projected ODE*

$$\dot{z}(t) = \varphi_z(z(t)) + r(t).$$

(b) *Assume* (A3) *with* $z_* \in (\theta_l^1, \theta_u^1) \times (\theta_l^2, \theta_u^2)$. *Then the conclusions of Theorem* 3.7 *continue to hold.*

Remark 4.2. Compared with Theorems 3.2 and 3.7, the limit ODE is replaced by the projected ODE. Moreover, we require z_* to be interior to the constraint set. Since the iterates are in the box $[\theta_l^1, \theta_u^1] \times [\theta_l^2, \theta_u^2]$, the tightness assumption of $\{z_n\}$ in Theorem 3.2 is no longer needed.

4.2. Method 2: Finite difference estimates without averages of samples (FDE). This is similar to the gradient estimates presented in section 2.3. Nevertheless, we do not take a sample average as in (2.7). Rather, we use $\widehat{\varphi}(z, \xi_n^\pm) = \widetilde{\varphi}(z, \xi_n^\pm)$. That is, $n_0 = 1$ without using sample averages. The resulting estimate is not expected to be as smooth, and the bias will be larger. However, it does provide us with a reasonable estimate. Moreover, this method is also simpler to use in handling real market data.

Define $Y_n^{\pm,\iota}$ as in (2.9), and use algorithm (2.12). We obtain the convergence and rate of convergence of the algorithm just as in the previous section.

PROPOSITION 4.3. *Under the conditions of Theorems* 3.2 *and* 3.7, *the conclusions of these theorems continue to hold.*

4.3. Method 3: Random directions (RD). This method has been found to be especially efficient for high-dimensional problems; see Spall [17]. The procedure is the following: Generate a sequence of independently and identically distributed (i.i.d.) random vectors $\{d_n\} = \{(d_n^1, d_n^2)\}$ that is independent of all other random processes such that for each $\iota = 1, 2$, d_n^ι is a Bernoulli random variable taking values ± 1 with equal probability 1/2. Replacing e_ι by d_n in the definition of Y_n results in

$$D^d\widehat{\varphi}(z_n, \xi_n) = \widehat{\varphi}(z_n + \delta_n d_n, \xi_n^+) - \widehat{\varphi}(z_n - \delta_n d_n, \xi_n^-). \tag{4.3}$$

The recursive formula then reads

$$z_{n+1} = z_n + \varepsilon_n d_n \frac{D^d\widehat{\varphi}(z_n, \xi_n)}{2\delta_n}. \tag{4.4}$$

Note that, compared with (2.9), the standard unit vectors are replaced by the random direction vectors. The advantage is that, by using the random direction vector d_n, all the components of the numerator of the finite difference are the same; only the denominators differ. That is,

$$d_n \frac{D^d \widehat{\varphi}(z_n, \xi_n)}{2\delta_n} = \begin{pmatrix} \dfrac{\widehat{\varphi}(z_n + \delta_n d_n, \xi_n^+) - \widehat{\varphi}(z_n - \delta_n d_n, \xi_n^-)}{2\delta_n d_n^1} \\ \dfrac{\widehat{\varphi}(z_n + \delta_n d_n, \xi_n^+) - \widehat{\varphi}(z_n - \delta_n d_n, \xi_n^-)}{2\delta_n d_n^2} \end{pmatrix}.$$

PROPOSITION 4.4. *In addition to the conditions of Theorems* 3.2 *and* 3.7, *assume that* $\{d_n\} = \{(d_n^1, d_n^2)\}$ *is a sequence of i.i.d. random variables that is independent of all other random processes of the problem such that for each* $\iota = 1, 2$, d_n^ι *is a Bernoulli random variable taking values* ± 1 *with equal probability* $1/2$. *Then the conclusions of Theorems* 3.2 *and* 3.7 *continue to hold.*

Remark 4.5. We remark that the results in Propositions 4.3 and 4.4 can be extended to the case in which a projection algorithm is used.

5. Numerical results. In this section, we report simulation results and numerical experiments. First we consider a case with $m = 2$ and compare our approach with an analytic solution. Then we use real market data to demonstrate how our method would work if it were used in the market. Using the proposed stochastic optimization procedure, one need not estimate the generator of the underlying "hidden" Markov chain. The calculation is done via only the observed real data, which is advantageous and provides opportunity for on-line recursive estimates.

5.1. Simulation study. This section is devoted to the numerical examples. It is further divided into several parts. By comparing the simulation results with the closed-form solution in the simple case, we demonstrate that the algorithms constructed indeed provide good approximation results.

5.1.1. Comparison with an analytic solution. We take the reward $\Phi(x) = e^x - 1$, $m = 2$, and the generator

$$Q = \begin{pmatrix} -6.04 & 6.04 \\ 8.90 & -8.90 \end{pmatrix}.$$

Choose $\tilde{\varrho} = 4$, $(z^1, z^2)' \in I = [0.01, 0.36] \times [0.01, 2.3]$, $r(1) = 1.50$, $r(2) = -1.61$, $\sigma(1) = 0.44$, $\sigma(2) = 0.63$, and the initial probability of the Markov chain $P(\alpha(0) = 1) = P(\alpha(0) = 2) = 0.5$. Then the analytic solution in [20] is given by $(z_*^1, z_*^2)' = (0.36, 0.277)'$.

In the following simulations, the sequences $\{\varepsilon_n\}$ and $\{\delta_n\}$ are chosen to be $\varepsilon_n = 1/(n + k_0)$ and $\delta_n = 1/(n^{1/6} + k_1)$, respectively, where k_0 and k_1 are some positive integers.

Method 1 (FDEA). In this experiment, we choose $k_0 = 1$, $k_1 = 10$, and $n_0 = 1000$. Recall that n_0 is the number of random samples used in each iteration (see (2.7)). The iterates stop whenever $\varepsilon_n < 0.001$. Several different initial values of $z_0 = (z_0^1, z_0^2)$ are used. Figure 5.1 demonstrates the convergence of the algorithm for two extreme starting points, one from far below and the other from far above the optimal point. In Table 5.1, $\overline{z} = (\overline{z}^1, \overline{z}^2)$ denotes the estimated optimal threshold value. The error

TABLE 5.1
Estimates using the FDEA method.

Initial (z^1, z^2)	(0.05, 0.05)	(0.10, 0.20)	(0.20, 0.40)	(0.10, 0.60)	(0.05, 0.85)
$(\overline{z}^1, \overline{z}^2)$	(0.36, 0.275)	(0.36, 0.269)	(0.36, 0.274)	(0.36, 0.274)	(0.36, 0.271)
$\lvert\overline{z} - z_*\rvert$	0.002	0.008	0.003	0.003	0.006

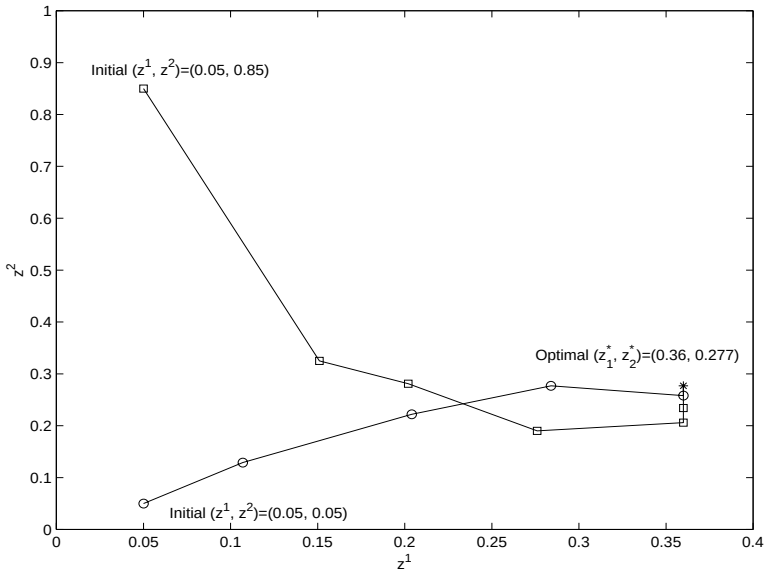


FIG. 5.1. *Convergence: FDEA method (total iteration number = 999).*

$|\overline{z} - z_*|$ is computed by

$$|\overline{z} - z_*| = \sqrt{(\overline{z}^1 - z_*^1)^2 + (\overline{z}^2 - z_*^2)^2}.$$

It can be seen from Table 5.1 that the estimates are insensitive to the initial values of (z^1, z^2); the algorithm leads to accurate estimation of the optimal value.

Method 2 (FDE). We use $k_0 = k_1 = 10$ in this experiment. As mentioned earlier for this method, we take $n_0 = 1$. The results are shown in Table 5.2 and Figure 5.2.

The advantage of this approach is that it does not require taking averages of samples, which is more desirable in practice for handling real data. Although the estimates obtained using the FDE method are not as good as those of the FDEA method, Figure 5.2 does demonstrate that the iterates are getting closer and closer to the optimal threshold value. Since fewer observation or measurement points are needed, this method is often preferred, especially if one is interested in obtaining a rough estimate quickly.

Method 3 (RD). As for Method 2, we use $k_0 = k_1 = 10$, and the gradient estimates are computed based on the random directions method. Table 5.3 and Figure 5.3 show the simulation results. The convergence of using the random direction method is similar to that of using the single sample path algorithm. However, it is slightly better than that of the FDE method (Method 2).

TABLE 5.2
Estimates using the FDE method.

Initial (z^1, z^2)	(0.10, 0.10)	(0.20, 0.20)	(0.30, 0.40)	(0.20, 0.60)	(0.10, 0.80)
$(\overline{z}^1, \overline{z}^2)$	(0.315, 0.276)	(0.314, 0.296)	(0.303, 0.291)	(0.311, 0.307)	(0.308, 0.314)
$\lvert\overline{z} - z_*\rvert$	0.045	0.050	0.059	0.058	0.064

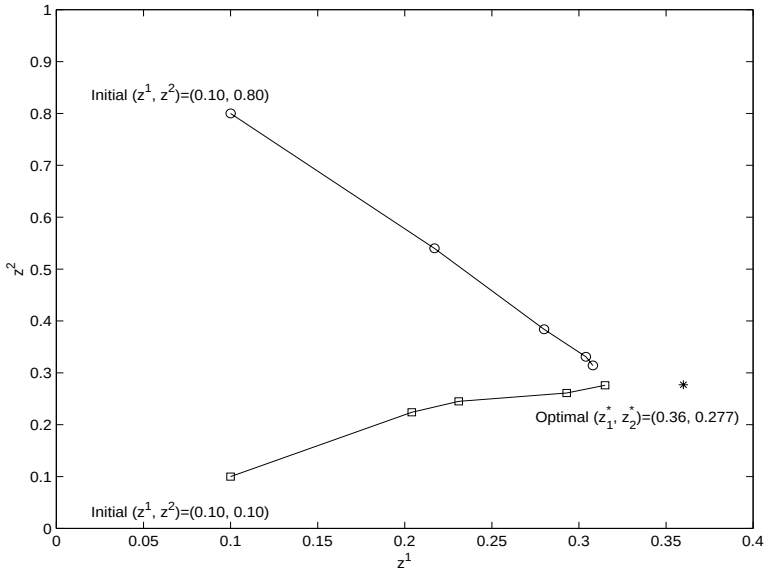


FIG. 5.2. *Convergence: FDE method (total iteration number = 990).*

5.1.2. Using real market data. In this section, we use the FDE method and the RD method to test the algorithms with real market data. We use the closing prices of Cisco (CSCO) (see Figure 5.4) from January 1, 1999, to June, 2001.

First, suppose that we bought the stock on January 3, 2000, at the closing price $S_0 = 54.03$ and wish to use the 1999 prices to find the threshold values $(\overline{z}^1, \overline{z}^2)$ and to determine the target prices $(A, B) = (S_0 \exp(-\overline{z}^1), S_0 \exp(\overline{z}^2))$.

We use our algorithms with constraints $\theta_l^\iota > 0$ for $\iota = 1, 2$. To use the most recent information, let us use the closing prices from July 1, 1999, to December 31, 1999, as observed data in the algorithms. This is consistent with the choice of $\widetilde{\varrho} = 4.0$ in Problem $\mathcal{P}$ defined in (2.4); such a $\widetilde{\varrho}$ corresponds to a half-year average holding duration (see [20]). We take the initial estimate to be $z_0 = (0.2, 0.2)$ and choose the projection region for z to be $[0.01, 0.223] \times [0.01, 2.3]$, corresponding to the stop-loss level of 20%. The computation results are summarized in Table 5.4. In this case, both the FDE and the RD methods give similar results.

Next we consider another set of data. Suppose the stock was bought a year later, on January 2, 2001, at the closing price $S_0 = 33.31$. We would like to use the information of the stock prices for the year 2000 to figure out target prices for the year 2001. We still choose the projection region for z to be $[0.01, 0.223] \times [0.01, 2.3]$, select the initial data $z_0 = (0.2, 0.2)$, and use the data starting from July 3, 2000. The results are given in Table 5.5.

TABLE 5.3
Estimates using the RD method.

Initial (z^1, z^2)	(0.10, 0.10)	(0.20, 0.20)	(0.30, 0.40)	(0.20, 0.60)	(0.10, 0.80)
$(\overline{z}^1, \overline{z}^2)$	(0.313, 0.283)	(0.310, 0.296)	(0.325, 0.301)	(0.319, 0.299)	(0.324, 0.296)
$\lvert\overline{z} - z_*\rvert$	0.047	0.054	0.042	0.047	0.041

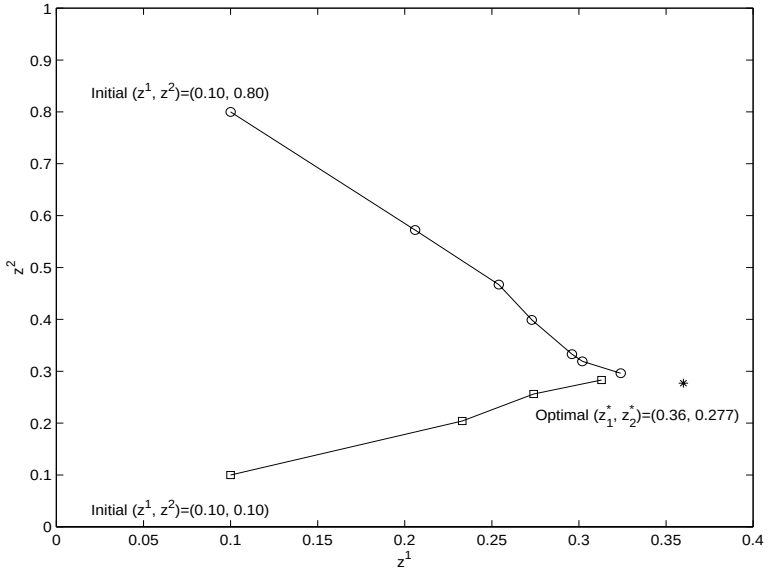


FIG. 5.3. *Convergence: RD method (total iteration number* $= 990$*).*

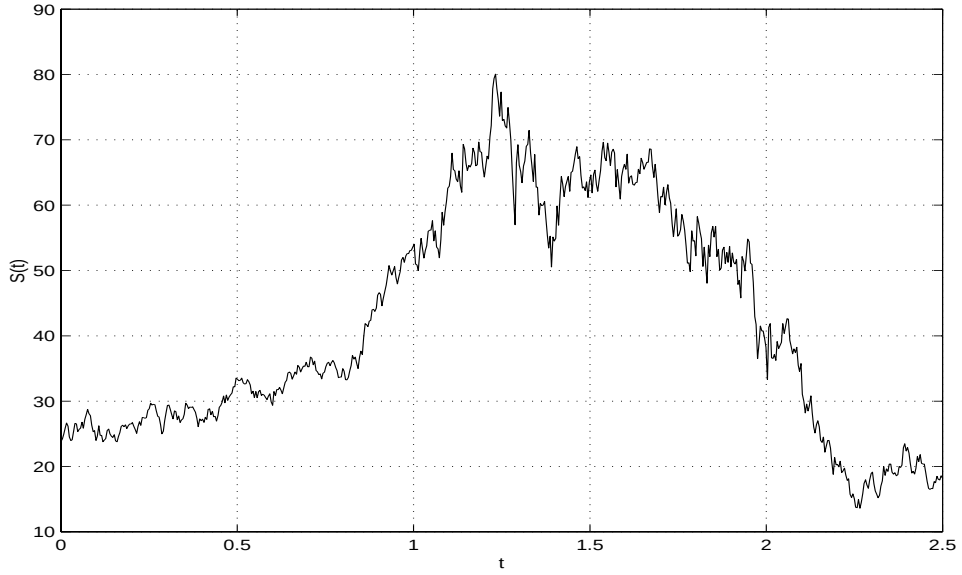


FIG. 5.4. *Closing prices of Cisco stock from January,* 1999*, to June,* 2001.

TABLE 5.4
Target prices of Cisco stock for the year 2000.

	$(\bar{z}^1, \bar{z}^2)$	(A, B)	Sold on	% of return
FDE	(0.2,0.340)	(44.24,75.91)	March 23, 2000	$(75.91 - 54.03)/54.03 = 40.5\%$
RD	(0.187,0.340)	(44.81,75.91)	March 23, 2000	$(75.91 - 54.03)/54.03 = 40.5\%$

TABLE 5.5
Target prices of Cisco stock for the year 2001.

	$(\bar{z}^1, \bar{z}^2)$	(A, B)	Sold on	% of return
FDE	(0.18,0.200)	(27.82,40.68)	Jan 3, 2001	$(40.68 - 33.31)/33.31 = 22.13\%$
RD	(0.18,0.201)	(27.82,40.73)	Jan 3, 2001	$(40.73 - 33.31)/33.31 = 22.28\%$

However, if we bought the stock on January 3, 2001, at the closing price $S_0 = 41.31$, using the FDE method, we obtain $(A, B) = (34.50, 50.46)$. Therefore, the stock should be sold on February 7 at 34.50 with a a loss of $(41.31 - 34.50)/41.31 = 16.48\%$. Similarly, using the RD method, we have $(A, B) = (34.50, 50.51)$, which yields the same result with a loss of 16.48%.

6. Further remarks. A class of stochastic optimization algorithms has been developed for threshold selling rules in stock trading. Although a hybrid GBM model is used, one need not estimate the generator of the Markov chain. As demonstrated by using real data, the algorithm can be applied to on-line estimation so as to provide sound estimates for target prices or stop-loss limits. The approach developed is simple and systematic; it provides a useful guideline for stock liquidation.

Throughout the paper, we have used the hybrid GBM model, with $\alpha(\cdot)$ being a continuous-time Markov chain. It should be pointed out that in the asymptotic studies, other stochastic processes (e.g., certain non-Markovian $\alpha(\cdot)$) can be dealt with. What is needed is that the hidden process can be averaged out.

In this work, we have developed stochastic optimization algorithms using decreasing step sizes $\{\varepsilon_n\}$ and $\{\delta_n\}$. Constant-step-size algorithms with $\varepsilon_n = \varepsilon$ and $\delta_n = \delta_\varepsilon \to 0$ as $\varepsilon \to 0$ and $\varepsilon/\delta_\varepsilon^2 \to 0$ as $\varepsilon \to 0$ may also be considered. Global stochastic optimization algorithms may be developed, but it is known that such an algorithm normally converges much slowly; see, for example, [18]. Further effort may also be devoted to the improvement of the efficiency of the algorithms and to the reduction of variance and bias.

REFERENCES

[1] P. BILLINGSLEY, *Convergence of Probability Measures*, John Wiley, New York, 1968.
[2] H. F. CHEN AND Y. M. ZHU, *Stochastic Approximation*, Shanghai Scientific and Technical Publishers, Shanghai, 1996.
[3] P. CLARK, *A subordinated stochastic process model with finite variance for speculative prices,* Econometrica, 41 (1973), pp. 135–155.
[4] R. J. ELLIOTT AND P. E. KOPP, *Mathematics of Financial Markets*, Springer-Verlag, New York, 1998.
[5] S. N. ETHIER AND T. G. KURTZ, *Markov Processes: Characterization and Convergence*, John Wiley, New York, 1986.
[6] J. P. FOUQUE, G. PAPANICOLAOU, AND K. R. SIRCAR, *Derivatives in Financial Markets with Stochastic Volatility*, Cambridge University Press, London, 2000.

[7] J. D. Hamilton and R. Susmel, *Autoregressive conditional heteroskedasticity and changes in regime,* J. Econometrics, 64 (1994), pp. 307–333.
[8] J. C. Hull, *Options, Futures, and Other Derivatives*, 3rd ed., Prentice-Hall, Upper Saddle River, NJ, 1997.
[9] I. Karatzas, *Lectures on the Mathematics of Finance*, American Mathematical Society, Providence, RI, 1996.
[10] I. Karatzas and S. E. Shreve, *Methods of Mathematical Finance*, Springer-Verlag, New York, 1998.
[11] H. J. Kushner, *Approximation and Weak Convergence Methods for Random Processes, with applications to Stochastic Systems Theory*, MIT Press, Cambridge, MA, 1984.
[12] H. J. Kushner and G. Yin, *Stochastic Approximation Algorithms and Applications*, Springer-Verlag, New York, 1997.
[13] R. C. Merton, *Lifetime portfolio selection under uncertainty: The continuous-time case,* Rev. Econom. Statist., 51 (1969), pp. 247–257.
[14] R. C. Merton, *Option pricing when underlying stock returns are discontinuous,* J. Financial Economics, 3 (1976), pp. 125–144.
[15] M. Musiela and M. Rutkowski, *Martingale Methods in Financial Modelling*, Springer-Verlag, New York, 1997.
[16] P. Praetz, *The distribution of share price changes,* J. Business, 45 (1972), pp. 49–55.
[17] J. C. Spall, *Multivariate stochastic approximation using a simultaneous perturbation gradient approximation,* IEEE Trans. Automat. Control, AC-37 (1992), pp. 332–341.
[18] G. Yin, *Rates of convergence for a class of global stochastic optimization algorithms,* SIAM J. Optim., 10 (1999), pp. 99–120.
[19] G. Yin and Q. Zhang, *Continuous-Time Markov Chains and Applications: A Singular Perturbation Approach*, Springer-Verlag, New York, 1998.
[20] Q. Zhang, *Stock trading: An optimal selling rule,* SIAM J. Control. Optim., 40 (2001), pp. 64–87.

SIAM J. OPTIM.
Vol. 13, No. 1, pp. 264–282

A CLASS OF TRUST-REGION METHODS FOR PARALLEL OPTIMIZATION*

P. D. HOUGH† AND J. C. MEZA†

Abstract. We present a new class of optimization methods that incorporates a parallel direct search (PDS) method within a trust-region Newton framework. This approach combines the inherent parallelism of PDS with the rapid and robust convergence properties of Newton methods. Numerical tests have yielded favorable results for both standard test problems and engineering applications. In addition, the new method appears to be more robust in the presence of noisy functions, which are inherent in many engineering simulations.

Key words. parallel optimization, trust-region methods, direct search methods, nonlinear programming

AMS subject classifications. 65Y05, 68W15, 90C30, 90C53, 90C90

PII. S1052623498343799

1. Introduction.

Optimization of functions derived from the modeling and simulation of some physical process constitutes an important class of problems in many engineering and scientific applications. Often, the computer simulation entails the solution of a system of nonlinear partial differential equations (PDE) in two or three dimensions. Other applications include particle dynamics simulations or problems in chemical kinetics. The main characteristic of these types of problems is that the function evaluation is computationally expensive and dominates the total cost of the optimization problem. Depending on the nature of the application and the solution method employed, there can also be noise associated with the evaluation of the objective function. This noise can usually be reduced, but only at the cost of making the computation time even greater. In many of these applications, derivative information is also not available or must be computed using finite differences, thereby generating noisy gradients. Fortunately, the dimension of the optimization problem in many of these optimal design problems is small (usually on the order of tens of parameters). In this study, we will concentrate on the development of parallel unconstrained optimization algorithms for the solution of these types of problems on small-scale shared memory processors (SMPs), where the number of available processors is comparable to the number of optimization parameters. The rationale for this decision is that, although massively parallel computers are available, the majority of computational power in most industrial or scientific settings consists of small-scale clusters of SMPs or networks of workstations (NOWs) that can be used in a similar capacity.

There have been many attempts at parallelizing nonlinear optimization methods. In the area of Newton methods, one of the earliest attempts at parallelization was the work of Straeter [22], who developed a parallel rank-one updating formula for

*Received by the editors August 19, 1998; accepted for publication (in revised form) December 21, 2001; published electronically August 28, 2002.

http://www.siam.org/journals/siopt/13-1/34379.html

†Computational Sciences and Mathematics Research Department, Sandia National Laboratories, P.O. Box 969, MS 9217, Livermore, CA 94551 (pdhough@ca.sandia.gov, meza@ca.sandia.gov).

the Hessian approximations used in variable metric methods. This formula was later extended by Laarhoven [16] to more general updating formulas. Byrd, Schnabel, and Shultz [2] also proposed parallel quasi-Newton methods based on speculative gradient and Hessian evaluations. Schnabel [21] gave an excellent review of the challenges and limitations in parallel optimization. In that review, Schnabel identified three major levels for introducing parallelism: (i) parallelize the function, gradient, and constraint evaluations; (ii) parallelize the linear algebra; and (iii) parallelize the optimization algorithm at a high level.

In this study, due to the characteristics of the problems mentioned above, we choose to focus on the third option. In particular, the first option is not usually available to us because for many situations we do not have access to the source code for the function or the constraints. In addition, the dimension of the optimization problems of interest is usually small, and therefore parallelizing the linear algebra would not yield any benefits.

The third option, that of parallelizing optimization at a high level, has recently received more attention. Some attempts that fit into this category include methods such as parallel direct search methods [7], genetic algorithms [14], and simulated annealing [13]. These methods are inherently parallel and extremely popular in engineering optimization. Although these search methods can be powerful tools, they suffer from slow convergence and thus may require many function evaluations. In a setting in which each function evaluation may take several CPU hours to compute, this is highly undesirable.

Newton-based methods, on the other hand, have good convergence properties, but there are few options for parallelizing a standard Newton method at a high level. For the purposes of this paper, we will assume that the gradient of the objective function is not available and that finite differences are used to compute any necessary derivative information. This calculation is trivially parallelized, so we focus our attention on finding less apparent opportunities. Another approach to parallelization is the work by Phua and Zeng [19] in which they use a multiple line search, multiple direction algorithm to introduce parallelism into a Newton method. However, it is not clear how robust a line search method would be in a situation where the function and gradient are noisy. Carter [3] has addressed the issue of inexact gradients for another class of algorithms known as trust-region methods and has given conditions under which these algorithms will converge. In a separate paper, Carter presented various numerical results [4] for this class of algorithms.

In this paper, we consider a new class of methods that combines the parallel direct search (PDS) method and the trust-region method to produce a new class of algorithms that takes advantage of the best properties of each approach. In particular, we will show that the rapid convergence rates typical of Newton-type methods are preserved while the advantage of parallelism inherent in the PDS methods is gained. In section 2, we describe the new class of algorithms and in section 3 consider their convergence properties. In section 4, we give numerical results from a set of test problems and an application in optimal design. We conclude in section 5 with a summary and a brief discussion of future research directions.

2. The trust-region PDS algorithm. Before describing the new class of algorithms, we first give a brief overview of the standard trust-region method and the PDS method of Dennis and Torczon [7]. In each iteration of a trust-region method, a quadratic model of the objective function, f, is formed, and a region in which the model is trusted to approximate the actual function accurately is determined. A trial

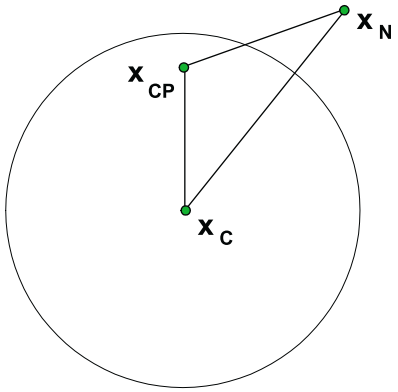


FIG. 1. *Standard trust-region method.*

step is then computed by approximately solving the following subproblem:

$$\min_{\mathbf{s}\in\mathbb{R}^n} \ \psi(\mathbf{s}) = g(\mathbf{x}_c)^T\mathbf{s} + \frac{1}{2}\mathbf{s}^T H_c\mathbf{s}, \tag{1}$$

$$\text{s.t.} \ \ \|\mathbf{s}\|_2 \leq \delta_c,$$

where $\mathbf{x}_c$ is the current point, $\mathbf{s}$ is the step, $g(\mathbf{x}_c)$ is the gradient of f at the current point, $H_c \approx \nabla^2 f(x_c)$ is the Hessian approximation at the current point, and δ_c is the size of the trust region. We will refer to this as the *trust-region subproblem.* It is well known that the step generated at each iteration, k, can be computed using any method, as long as it satisfies a fraction of Cauchy decrease condition according to the quadratic model. In particular, there must exist constants, $\beta > 0$ and $C > 0$, independent of k, for which the step s_k taken at iteration k satisfies

$$\psi_k(\mathbf{s}_k) \leq \beta \left\|g(\mathbf{x}_k)\right\|_2 \min\left(\delta_k, \frac{\|g(\mathbf{x}_k)\|_2}{C}\right). \tag{2}$$

There are several well-known procedures for computing the solution to (1) that satisfy (2). One such example is the dogleg step (see [6, p. 139]), which is a convex combination of the steepest descent direction and the Newton direction. This approach is illustrated in Figure 1.

The PDS algorithm belongs to a class of optimization methods that do not compute derivatives. The PDS algorithm can be briefly described as follows. Starting from an initial simplex, S_0, the function value at each of the vertices in S_0 is computed, and the vertex corresponding to the lowest function value, v_0, is determined. Using an underlying grid structure, S_0 is reflected about v_0, and the function values at the vertices of this rotation simplex, S_r, are compared against the function value at v_0. If one of the vertices in S_r has a function value less than the function value corresponding to v_0, then an expansion step to form a new simplex, S_e, is attempted in which the size of S_r is expanded by some multiple, usually 2. The function values at the vertices of S_e are compared against the lowest function value found in S_r. If a lower function value is encountered, then S_e is accepted as the starting simplex for the next iteration; otherwise, S_r is accepted for the next iteration. If no function value lower than the one corresponding to v_0 is found in S_r, then a contraction simplex is created by reducing the size of S_0 by some multiple, usually 1/2, and is accepted for the next iteration.

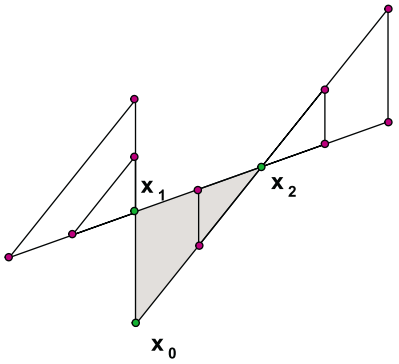


FIG. 2. *PDS method.*

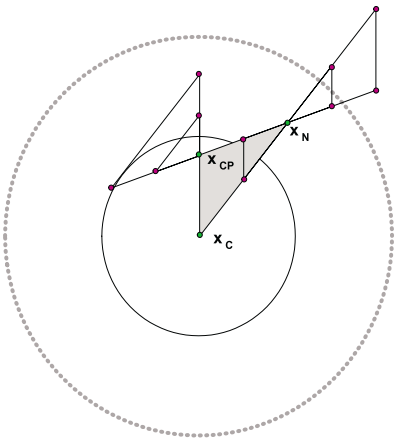


FIG. 3. *TRPDS method.*

Because PDS uses only function comparisons, it is easy to implement and use. Since the rotation, expansion, and contraction steps are all well determined, it is also possible to precompute a set of grid points corresponding to the vertices of the simplices constructed from various combinations of rotations, expansions, and contractions. Given this set of grid points, called a search scheme, the PDS algorithm can compute the function values at all of these vertices in parallel and determine the vertex corresponding to the lowest function value. The number of points used in the search scheme is referred to as the *search scheme size*, and it usually is adjusted to be at least equal to the number of processors available. Figure 2 demonstrates one possible PDS iteration.

Both the trust-region and the PDS methods have advantages and disadvantages, as described in section 1. In order to combine the strengths of these methods, we propose a new class of algorithms which uses the PDS method within a trust-region framework. This type of algorithm, which we will refer to as TRPDS, is illustrated in Figure 3 and is described below. The controlling framework is the same as that for standard trust-region algorithms, but the method of computing the new step is different. Rather than solving the trust-region subproblem, the TRPDS method

approximately solves the following problem:

$$\begin{aligned} &\min_{\mathbf{s}\in\mathbb{R}^n} f(\mathbf{x}_c+\mathbf{s}), \\ &\text{s.t. } \|\mathbf{s}\|_2 \le 2\delta_c, \\ &\qquad \psi(\mathbf{s}) \le \beta \|g(\mathbf{x}_c)\|_2 \min\left(\delta_c, \frac{\|g(\mathbf{x}_c)\|_2}{C}\right), \end{aligned} \tag{3}$$

where $\beta > 0$, $C > 0$, and ψ is defined as in (1). We will refer to this as the *PDS subproblem*. There are several notable differences between this and the standard trust-region approach. The first is that the actual objective function, as opposed to a quadratic model of the objective function, is being minimized. Second, this subproblem is not solved to optimality; only a small amount of decrease is required from the PDS method. Third, the step length is allowed to be twice the size of the trust region to allow for the possibility of taking a step longer than the Newton step, as is sometimes done to accelerate local convergence of singular problems. Further discussions of this acceleration idea can be found in [20] and [15]. Finally, because by design the steps computed by PDS do not satisfy the fraction of Cauchy decrease condition (2), we must include an explicit constraint to enforce this condition.

An overview of the TRPDS algorithm appears below, followed by a discussion of the critical steps.

ALGORITHM 1 (TRPDS).
Given $\mathbf{x}_0$, $\mathbf{g}_0$, H_0, δ_0, and $\eta \in (0,1)$,
for $k = 0, 1, \ldots$ until convergence **do**
  1. Solve $H_k \mathbf{s}_N = -\mathbf{g}_k$.
  for $i = 0, 1, \ldots$ until step accepted **do**
    2. Form an initial simplex using $\mathbf{s}_N$.
    3. Find an approximate solution $\mathbf{s}_i$ to (3) using PDS.
    4. Compute $\rho_i = (f(\mathbf{x}_k + \mathbf{s}_i) - f(\mathbf{x}_k))/\psi_k(\mathbf{s}_i)$.
    if $\rho_i > \eta$, **then**
      5. Accept step and set $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{s}_i$, eval $\mathbf{g}_{k+1}$ and H_{k+1}.
    else
      6. Reject step.
    end if
    7. Update δ.
  end for
end for

Here we use the notation $\mathbf{g}_k = g(\mathbf{x}_k)$ and $H_k = H(\mathbf{x}_k)$, where $H(\cdot)$ is the Hessian approximation. There are several points to consider within this framework. The initial simplex formed in step 2 needs to be chosen carefully. While there is a lot of freedom in the choice of the initial simplex, it will have an impact on the solution to (3) and on the performance of the algorithm. There is also a question as to how accurately we should solve (3). In many applications it may be reasonable to ask for only a small fraction of decrease in the function since each function evaluation is so expensive. This also has a bearing on the decision to accept or reject the new step. Finally, the updating of the trust region must be addressed within the context of this framework. In the following sections, we address each one of these issues.

2.1. Choosing the initial simplex. In step 2 of the TRPDS algorithm we require the formation of an initial simplex. The choice of this simplex is important to the performance of the PDS algorithm and deserves careful consideration. There

are three points that must be included in the simplex: the current point, the Cauchy point, and the Newton point. The Cauchy point is defined to be the minimizer of the quadratic model along the steepest descent direction. Likewise, we define the Newton point, $\mathbf{s}_N$, to be the minimizer of the quadratic model along the Newton direction. The current point must be in the simplex so that PDS can determine whether or not it has found a descent direction. The Cauchy point is required to ensure convergence of the algorithm. Finally, the Newton point is necessary to allow for the possibility of rapid convergence in the limit. In practice, all three of these points are not always included, but instead related points are used. A discussion of these substitutions follows.

Recall that, in the standard trust-region algorithm, the step length is limited by the size of the trust region. When the Cauchy point or the Newton point falls outside of the trust region, it is projected onto the trust region. As we are seeking to preserve as much of the trust-region framework as possible, the construction of the initial simplex includes similar features. There are three different scenarios that must be addressed.

1. *Both the Cauchy point and the Newton point are inside the trust region.* This case is straightforward. The Cauchy point and the Newton point are used in the initial simplex.
2. *The Cauchy point is inside the trust region, and the Newton point is outside the trust region.* The Cauchy point is used in the initial simplex. The dogleg point is computed and replaces the Newton point in the initial simplex.
3. *Both the Cauchy point and the Newton point are outside the trust region.* Both points are projected onto the trust region, and the resulting points are used in the initial simplex.

For a problem of dimension n, PDS requires the initial simplex to have $n+1$ vertices. We have described the selection of only three. The question of how to pick the remaining $n-2$ vertices remains. While there are many logical ways to choose these points, the only real restriction on them is that they be chosen such that the initial simplex is not degenerate. Our current implementation uses $n-2$ vertices from a right angle simplex constructed around the Newton point; i.e., vertices $4, \ldots, n+1$ are defined as follows:

$$\mathbf{v}_{i+3} = \mathbf{x}_N + \delta_c \mathbf{e}_i, \quad i = 1, \ldots, n-2,$$

where $\mathbf{x}_N$ is the Newton point, δ_c is the current trust-region radius, and $\mathbf{e}_i$ is the ith column of the $n \times n$ identity matrix.

When the simplex is constructed in the manner described here, there are two situations in which it may be degenerate. One case can arise in the first iteration of the trust-region method. If the initial Hessian is a multiple of the identity, then the Newton direction and the Cauchy direction will be the same, so the simplex needs to be constructed in a slightly different manner in the first iteration. We use the current point, the Newton point, and $n-1$ of the vertices from the right angle simplex around the Newton point. The other case of degeneracy arises if the edges of the simplex are badly scaled. This is easily corrected by rescaling all of the edges to be the same length as the Newton edge. Note that this allows longer steps than if all edges were rescaled to be the same length as the Cauchy edge.

2.2. Solving the PDS subproblem. One way to think about the TRPDS algorithm is to imagine using an optimization algorithm within an optimization algorithm. As such, the PDS method needs algorithmic parameters in order to solve the

PDS subproblem. In particular, PDS needs information about the search space, and it needs stopping criteria. Recall that the PDS method evaluates the function at a set of predetermined reflection, contraction, and expansion points in order to determine a trial step. This search scheme can be determined ahead of time and need only be generated once; however, during the optimization phase, the PDS method must know how many points in that search scheme to evaluate at each iteration. One possible choice in a parallel setting is to set this number equal to the number of processors that are available. PDS also needs to be aware of the constraints on the step. Clearly, it must know the size of the trust region, as that constrains the step length. As noted earlier, we relax the trust region by a factor of two in order to allow for the possibility of taking a step longer than the Newton step. Finally, PDS must have access to the quadratic model used by the trust-region framework in order to ensure that it generates trial steps that satisfy the fraction of Cauchy decrease constraint.

Since the PDS subproblem is not solved to optimality, we use four criteria to determine when to return a trial step. The first is a simple decrease requirement. If f_c is the function value at the current point, then we return a step when

$$f_t \leq pdstol * f_c, \tag{4}$$

where f_t is the function value at the trial point, and $pdstol < 1$ is the amount of decrease desired. The second is a restriction on how much PDS is allowed to decrease the step length. Ideally, the trust-region framework should maintain control of the step length; however, if stopping criteria for PDS are not chosen appropriately, it is possible for PDS to "hijack" control of the step length. Recall that PDS, in its effort to find decrease, may reduce the size of the simplex. If allowed too many opportunities to shrink the simplex, PDS can return a step that is significantly smaller than the current trust region. As we will see in section 2.4, the trust-region update is based on this step length. As a result, the trust region will become unacceptably small, causing the algorithm to halt prematurely. In order to prevent that from happening, we return a trial step when

$$\|E_t\|_2 \leq etol * \|E_0\|_2 ,$$

where E_t is the longest edge in the simplex producing the trial point, E_0 is the longest edge in the initial simplex, and $etol \leq 1$ is the edge reduction tolerance. If PDS cannot find a trial point that satisfies either of these criteria, then it returns when it has exceeded either the maximum number of function evaluations or the maximum number of iterations allowed.

2.3. Acceptance/rejection of step. Once a step has been computed, it is necessary to determine whether or not it is acceptable. This is handled by the trust-region framework. If the step yields sufficient decrease, then the step will be accepted. Otherwise, the step is rejected. In a standard trust-region method, sufficient decrease is determined by computing ρ_k as given in step 4 of the algorithm and comparing it to some tolerance. If ρ_k is greater than the tolerance, then the decrease is sufficient. There is some flexibility in the choice of this tolerance, and computational expense of the function plays a role in determining the appropriate choice. There is one situation that arises in the TRPDS algorithm that requires a minor modification to this scheme. It is possible that PDS will find no decrease, and thus return a step of zero length. In this case, ρ_k is not computed, and the step is rejected immediately.

2.4. Updating the trust region. The procedure for updating the trust region is based on the strategy proposed in [3]. At each iteration, the trust region is updated as follows:

$$\delta_{k+1} = \begin{cases} \dfrac{\min(\delta_k, \|E\|_2)}{10} & \text{if } \rho_k < \eta_1 \text{ or } \|\mathbf{s}_k\|_2 = 0, \\ \dfrac{\|\mathbf{s}_k\|_2}{2} & \text{if } \eta_1 \le \rho_k \le \eta_2, \\ 2 \cdot \delta_k & \text{if } \eta_3 \le \rho_k \le 2 - \eta_3, \\ \max(2 \cdot \|\mathbf{s}_k\|_2, \delta_k), & \text{otherwise.} \end{cases}$$

Here E is the longest edge of the final PDS simplex, and $\eta_1, \eta_2, \eta_3 \in (0, 1)$. Notice that when the trust-region size is reduced, we incorporate information about the size of the final simplex in order to reduce the possibility of rechecking points that are already known to be unacceptable. In our algorithm, we also impose a maximum on the trust-region size that is allowed at any iteration.

3. Convergence results. It is interesting to note that Algorithm 1 falls into the class of generalized trust-region methods described in [1]. As such, it offers a great deal of flexibility, and the convergence theory is straightforward. In order to apply the work of [1], we first demonstrate how TRPDS fits into the generalized trust-region framework.

The first feature of interest is the approximation model used by the trust-region framework. The model used to approximate the objective function can be any model suitable for the application in question as long as it satisfies the following mild conditions:

$$a_k(\mathbf{x}_k) = f(\mathbf{x}_k), \tag{5}$$

$$\operatorname{grad} a_k(\mathbf{x}_k) = \operatorname{grad} f(\mathbf{x}_k), \tag{6}$$

where a_k is the approximation model at iteration k. In the TRPDS setting, a_k is the quadratic model, so these conditions are clearly satisfied.

The second feature of interest in the generalized trust-region framework is that the step generated at each iteration can be computed using any method as long as it satisfies a fraction of Cauchy decrease condition according to the approximation model being used. In particular, there must exist constants, $\beta > 0$ and $C > 0$, independent of k, for which the step taken at iteration k, $\mathbf{s}_k$, satisfies

$$f(\mathbf{x}_k) - a_k(\mathbf{x}_k + \mathbf{s}_k) \ge \beta \|g(\mathbf{x}_k)\|_2 \min\left(\delta_k, \frac{\|g(\mathbf{x}_k)\|_2}{C}\right), \tag{7}$$

where a_k denotes the model being used to approximate the objective function. To obtain a trial step, we use PDS to solve the PDS subproblem (3), which includes a fraction of Cauchy decrease as an explicit constraint. Therefore, we know that the trial step will satisfy (7). We note that numerical results indicate that ignoring the fraction of Cauchy decrease condition rarely has undesirable effects on the convergence of the algorithm in practice. Given the computational expense of the function, we believe that excluding steps that do not satisfy a fraction of Cauchy decrease should be optional in practice.

With the conditions on the model and the steps satisfied, we can now apply standard trust-region theory.

THEOREM 3.1. *Assume that f is uniformly continuously differentiable, bounded below, and that the Hessian approximations are uniformly bounded. Furthermore, assume that the sequence of iterates generated by Algorithm* 1 *satisfies* (5)–(7). *Then*

$$\lim_{k\to\infty} \inf \|\text{grad}\, f(\mathbf{x}_k)\|_2 = 0.$$

In fact, since Algorithm 1 uses widely accepted criteria for updating the steps within the trust-region framework, it is easy to show that the stronger condition

$$\lim_{k\to\infty} \|\text{grad}\, f(\mathbf{x}_k)\|_2 = 0$$

holds.

This is satisfying, in that the new class of algorithms inherits all of the good theoretical convergence properties of the classical trust-region methods while incorporating all of the practical advantages of PDS for typical engineering optimization problems.

4. Numerical results. In order to evaluate the performance of the TRPDS algorithm, we chose a set of test problems from the literature. One set of 24 problems was obtained from papers by Moré, Garbow, and Hillstrom [18] and Byrd, Schnabel, and Shultz [2]. Another set of 22 problems was obtained from the CUTE test set developed by Conn, Gould, and Toint [5]. A list of the problems used can be found in the appendix. For comparison purposes, we also solved these problems using a standard BFGS trust-region algorithm.

To evaluate the effectiveness of the TRPDS algorithm, we ran a series of tests varying several of the algorithmic parameters. The first set of 24 problems allow a variable dimension, so we ran problems with dimensions of dim $= 4, 8, 16$. The search scheme size (sss) was then chosen so that sss $=$ dim $+ 1, 2 *$ dim, $4 *$ dim, $8 *$ dim. The function tolerance used for the inner PDS iteration was also varied using values of $0.1, 0.5, 0.9$, and 0.99995. The CUTE test problems were run using the default dimension and varying only the sss and the function tolerance for PDS. The combination of all of these experiments resulted in a total of 1504 test cases run, though we present only a representative subset of the results here. In addition, we also present results for the case of noisy functions, as well as the results from an engineering application problem based on a computer model of a chemical vapor deposition furnace.

All tests were run on a 64-processor SGI Origin 2000 with the IRIX 6.5 operating system. The starting points used for these problems were the same as those given in the references, and the gradients were computed using parallel central differences. The BFGS algorithm was run with the number of processors equal to the dimension of the problem, which is the maximal amount of parallelism for our implementation of the parallel central differences. The TRPDS algorithm was run with the number of processors equal to the sss, which we varied in our tests. Algorithmic parameters are listed in Table 1.

The step tolerance, the function tolerance, and the gradient tolerance are used as stopping criteria in termination tests like those found in [8, p. 306].

The next three sections contain the results for all of the test runs: (1) the noise-free case, (2) noisy functions, and (3) a furnace problem.

4.1. Noise-free tests. The results for the noise-free case are contained in Figures 4–7. While reporting the number of iterations may be useful in some settings,

TABLE 1
Algorithmic parameters.

Parameter	**Value**
Initial trust region	$0.1 \cdot \|\mathbf{g}_0\|_2$
Machine epsilon	$2.22045 \cdot 10^{-16}$
Maximum step	4000
Minimum step	$1.49012 \cdot 10^{-8}$
Maximum iter	500
Maximum fcn eval	$1,000,000$
Step tolerance	$1.49012 \cdot 10^{-8}$
Function tolerance	10^{-10}
Gradient tolerance	10^{-6}

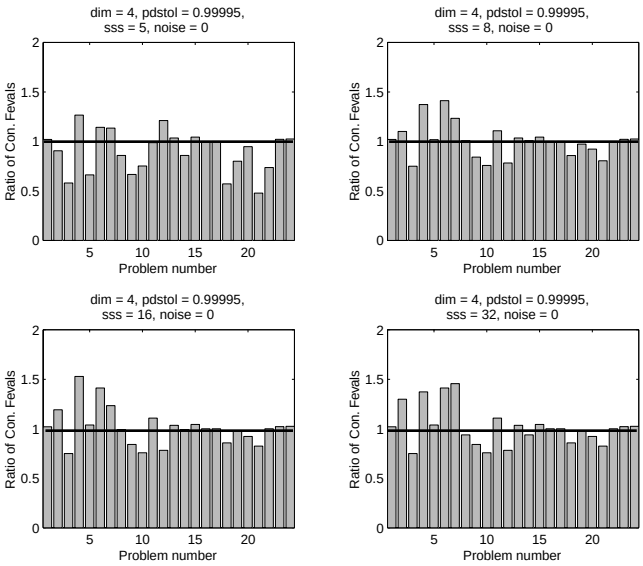


FIG. 4. *Ratios of the number of concurrent function evaluations for BFGS to that for TRPDS for dimension* 4. *Ratios greater than one indicate that the TRPDS algorithm takes fewer concurrent function evaluations and thus less overall time.*

we believe that it is not useful here. In our applications, the most important measure of performance is the total time to solution of the problem. Since the computational cost of the function evaluations dominates the cost of the algorithm, we base our comparison of TRPDS with BFGS on the number of function evaluations. However, since the two algorithms have different degrees of parallelism, comparing the total number of function evaluations required for each is not a fair method of comparison, as this would not reflect the amount of time required to solve the problems. Instead, we compare the number of *concurrent function evaluations*, which are defined as follows. Suppose that p processors are available, where $p \geq 1$. If p independent function evaluations are required, then each processor can be tasked to perform one of them. This means that p function evaluations can be done simultaneously. Thus, we define a *concurrent function evaluation* to be one instance of p function evaluations being

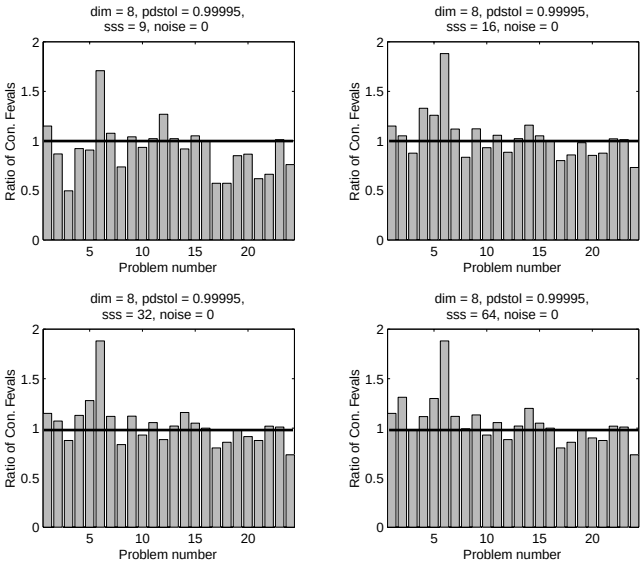


FIG. 5. *Ratios of the number of concurrent function evaluations for BFGS to that for TRPDS for dimension* 8. *Ratios greater than one indicate that the TRPDS algorithm takes fewer concurrent function evaluations and thus less overall time.*

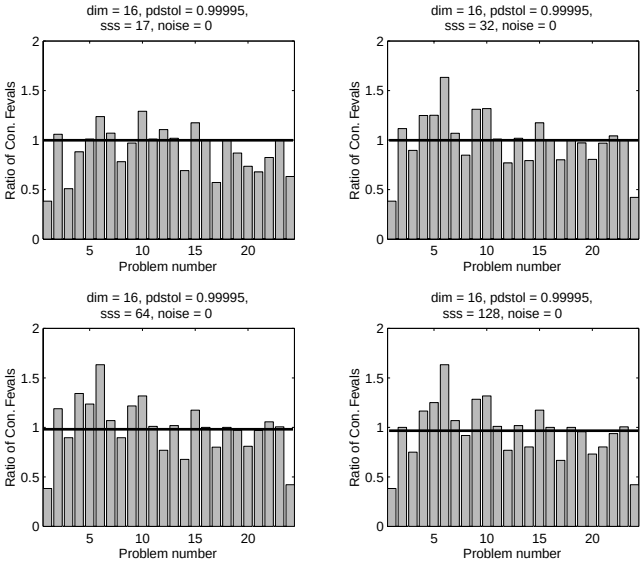


FIG. 6. *Ratios of the number of concurrent function evaluations for BFGS to that for TRPDS for dimension* 16. *Ratios greater than one indicate that the TRPDS algorithm takes fewer concurrent function evaluations and thus less overall time.*

performed simultaneously. Comparing the number of concurrent function evaluations is therefore roughly equivalent to comparing total wall clock time. The results that appear in Figures 4–7 are the ratios of the number of concurrent function evaluations taken by the trust-region method to the number of concurrent function evaluations

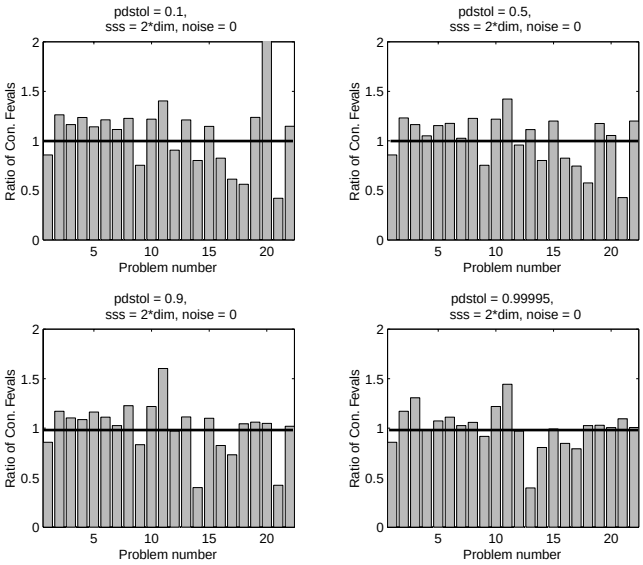


FIG. 7. *Ratios of the number of concurrent function evaluations for BFGS to that for TRPDS for CUTE problems. Ratios greater than one indicate that the TRPDS algorithm takes fewer concurrent function evaluations and thus less overall time.*

taken by the TRPDS algorithm. Ratios greater than one indicate that the TRPDS algorithm will take less overall time.

Overall, we find that TRPDS does at least as well as BFGS in 60% of the test cases. We now take a closer look at the effects of the parameters on the performance of the algorithm. The results shown in Figures 4–6 allow us to make some general observations about the effect of the sss. For a given problem, the size of the search scheme has very little effect on the performance of the algorithm in comparison to the standard trust-region algorithm. There are several factors that may be contributing to this effect. The first is that we have biased the search directions towards the Newton point through our method for constructing the initial simplex. A second factor is the order in which the search scheme points are chosen and evaluated. In the standard PDS algorithm, the reflection points are evaluated first, followed by the contraction points and the expansion points. Since the Newton point is often a good trial point we would not expect nearby points to have a lower function value, and we need to take a large step before finding a better point. Due to the construction of the search pattern, this requires a search scheme larger than those used in our tests. In a situation where we are far away from the solution, or where the function or gradient is noisy, the Newton point might not be a good trial point. In this case, the reflection and contraction points may yield better trial points than the Newton point, and thus the sss represented here might affect the performance of the algorithm. This is a point that will require further investigation. It would also be interesting to try different methods for creating the initial simplex to determine the effect on the algorithm's efficiency.

We have also examined the effects of changing the amount of function decrease requested from the inner PDS iteration by (4). The results are shown for the CUTE problems in Figure 7. We find that changing the amount of function decrease required has very little effect on the performance of TRPDS. The reason is that PDS satisfies

TABLE 2
Convergence tolerances for noisy functions.

Parameter	**Value**
Function tolerance	$\eta/10$
Gradient tolerance	$100\eta^{2/3}$

other stopping criteria before it attains the required amount of function decrease, particularly for larger amounts of decrease.

Schnabel [21] presents an argument for the merits of a line search algorithm with speculative gradient evaluation. This entails using extra processors to compute the components of a finite difference gradient at the current point while the function is being evaluated at that point. He argues that it is difficult to develop a parallel line search algorithm that will perform better than a speculative gradient algorithm, particularly when the dimension of the problem is not much larger than the number of available processors. A similar argument holds for a trust-region approach. Thus, we should also compare TRPDS to a speculative gradient implementation of a trust-region method. We have begun to address this work, and we refer the reader to [11] for an analytical comparison of the methods, and to [12] for preliminary numerical results.

4.2. Noisy functions. Since one of the motivations for proposing this new algorithm was to improve the robustness of Newton methods when applied to noisy functions, we also ran a set of test problems in which random noise was added to the functions. These test problems were generated by computing a function value such that

$$\hat{f}(x) = f(x) + \eta u,$$

where u is a random number from a uniform distribution, $U(0,1)$, and η is the noise level. We ran tests with $\eta = 10^{-9}, 10^{-6}$, and 10^{-3}. One of the more difficult issues in this set of test problems was choosing the stopping criteria. Since we are computing gradients using finite differences, the noise in the function will propagate into the gradients. As such, it is sometimes difficult to detect convergence using the standard stopping criteria. In our case, we used the tolerances given in Table 2.

The results shown in Figures 8–9 allow us to make some general observations about the effect of noise on the algorithms. For the first set of test problems, the TRPDS algorithm either beats or ties the BFGS algorithm in 66% of the test cases. More importantly, TRPDS fails to converge on only 3% of the tests while the BFGS algorithm fails to converge on 17% of the tests. The results from the CUTE test set are not quite as convincing with regard to robustness, although the TRPDS algorithm does win 70% of the time. As in the noise-free case, we suspect that a different choice of the initial simplex and perhaps a larger search scheme size will benefit the TRPDS algorithm in the noisy function case; we are pursuing this line of research.

4.3. Furnace design test problem. As another test case, we chose an optimization problem derived from the design of a vertical, multiwafer furnace. Vertical furnaces can process up to 200 silicon wafers in a single batch and have been used for thin film deposition, oxidation, and other thermal process steps. The evolution of vertical furnaces has been driven by the need for process uniformity (that is, wafer-to-wafer and within-wafer uniformity) and high wafer throughput. A recent variation

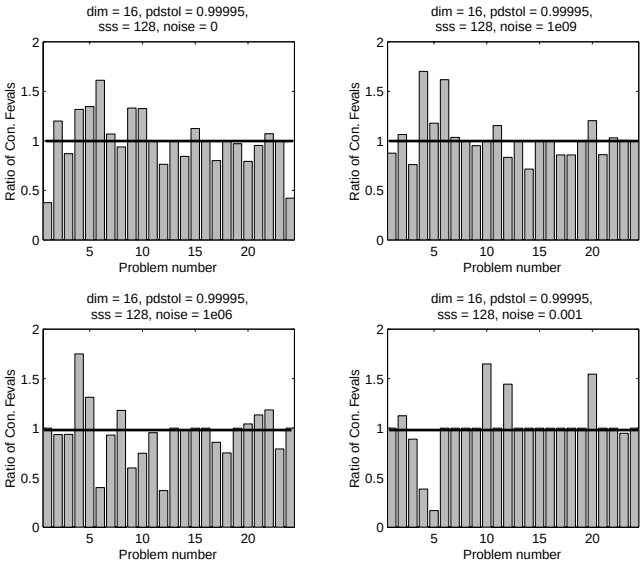


FIG. 8. *Ratios of the number of concurrent function evaluations for BFGS to that for TRPDS for noisy problems. Ratios greater than one indicate that the TRPDS algorithm takes fewer concurrent function evaluations and thus less overall time.*

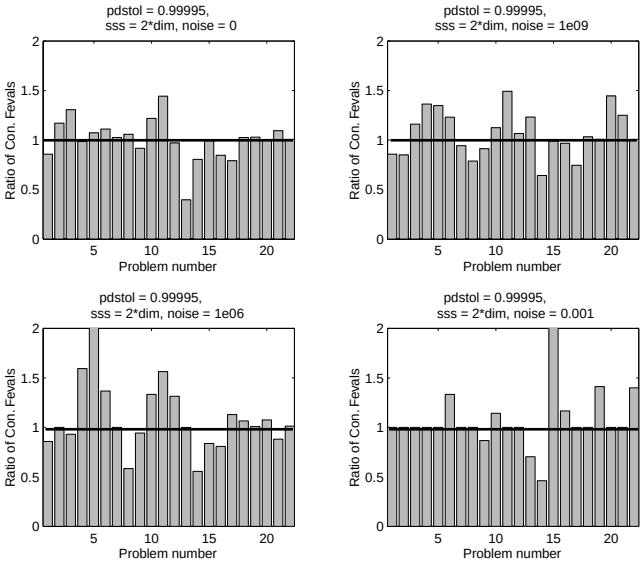


FIG. 9. *Ratios of the number of concurrent function evaluations for BFGS to that for TRPDS for noisy CUTE problems. Ratios greater than one indicate that the TRPDS algorithm takes fewer concurrent function evaluations and thus less overall time.*

of the multiwafer reactor design is the small-batch, fast-ramp (SBFR) furnace. The SBFR is designed to heat up and cool down quickly, thus reducing cycle time and thermal budget. The SBFR consists of a stack of silicon wafers eight inches in diameter enclosed in a vacuum-bearing quartz jar. The stack is radiatively heated by resistive

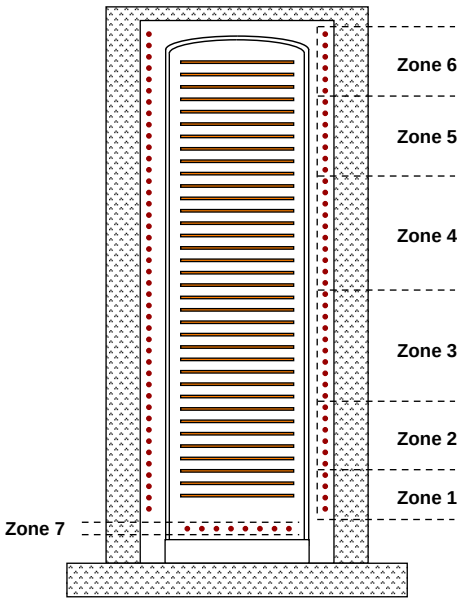


FIG. 10. *Vertical batch furnace with seven control zones.*

coil heaters contained in an insulated canister. The heating coils can be individually controlled or can be ganged together in zones to vary the emitted power along the length of the reactor; a seven-zone configuration is shown in Figure 10. There are six control zones (each containing several heating coils) along the length of the furnace and one heater zone in the base. The zones near the ends of the furnace are usually run hotter than the middle zones to make up for heat loss.

The thermal design optimization problem can be described as follows. Given a set number of fixed heating coils, how can the coils be grouped in the fewest number of control zones such that the temperature deviation about a fixed set-point is minimized? For this example, we concentrate on finding the optimal power settings and related temperature uniformity for a fixed zone configuration. The objective function, F, is defined by a least-squares fit of the N discrete wafer temperatures, $T_{w,i}$, to a prescribed profile, $T_{s,i}$,

$$F(p_j) = \sum_{i=1}^{N} (T_{w,i} - T_{s,i})^2, \tag{8}$$

where p_j are the unknown power parameters.

The engineering heat transfer model used in this example was developed by Houf, Grcar, and Breiland [10] specifically for the analysis of vertical furnaces. (The actual simulation code used in our experiments is called TWAFER.) Given a set of powers, p_j, each call to TWAFER produces a set of temperatures for the entire furnace, from which the wafer temperatures are extracted. The heat transfer formulation is simplified by using mass lumping and one-dimensional approximations. The nonlinear transport equations are solved using the TWOPNT solver [9], which uses a Newton method with a time evolution feature that computes the steady-state solution. By varying the tolerances for the TWOPNT solver, it is possible to increase the accuracy

TABLE 3
Number of nonlinear iterations and wall clock time.

	BFGS		**TRPDS**	
RTOL	**Iter**	**Time**	**Iter**	**Time**
10^{-2}	3*	154.471	100	5217.98
10^{-3}	30*	1603.75	100	6872.32
10^{-4}	64	4470.97	100	8857.03
10^{-5}	31	2729.82	25	2786.17
10^{-6}	39	4145.88	33	4468.22
10^{-7}	34	4516.57	24	3832.82
10^{-8}	31	4612.38	26	4548.66
10^{-9}	31	5152.05	25	4969.36
10^{-10}	31	6044.57	22	5101.78
10^{-11}	31	6576.52	28	6915.97
10^{-12}	31	6600.24	23	5774.78
$*$ indicates the method did not converge				

of the steady-state solution at the cost of increasing the computational time. In particular, we have chosen to vary a parameter that determines the relative convergence tolerance, RTOL, for the steady-state solution of the underlying PDE.

There are many different parameter combinations that have been considered in previous studies of the TWAFER code [17]. For this particular example we used only one configuration, namely, a design problem with seven heater zones: one bottom heater and six equally sized side heaters. Each simulation used a model that contained 100 wafers with ten discretization points per wafer. Our initial guess for the powers was $p_0 = \{100, 200, 300, 2700, 100, 400, 2000\}$.

Table 3 contains the total number of iterations as well as the total wall clock time taken by each method in computing a solution. With respect to the total wall clock time, the new method is competitive with the standard BFGS method in almost all cases. In addition, the new method is more robust for the larger values of RTOL. These values correspond to a very loose convergence tolerance for the PDE solver in the TWAFER modeling code. In these cases, the standard BFGS method did not converge to a solution, while the new method still managed to proceed; the BFGS algorithm terminated with the trust region shrinking below its minimum allowed size. This failure to converge is probably due to large inaccuracies in the gradient evaluations due to the loose tolerances in the PDE solutions.

The resulting power values were then given to the TWAFER simulation using a value of RTOL $= 10^{-12}$. Figures 11–12 show the wafer temperatures that result with the computed powers. The interesting point here is that, even with relatively large values of the parameter RTOL, the resulting temperatures are still quite reasonable. As we have already noted, for these same values of RTOL the BFGS algorithm did not converge.

5. Summary. We have described a new class of algorithms for parallel optimization. The general framework consists of a trust-region model in which a nonstandard subproblem is solved using a PDS method that takes advantage of parallelism to solve the problem more efficiently on multiple processors. This new algorithm can be shown to have the same convergence properties as standard trust-region methods. In addition, the practical properties of PDS methods can be used to solve engineering optimization problems that are noisy or lack analytic derivatives.

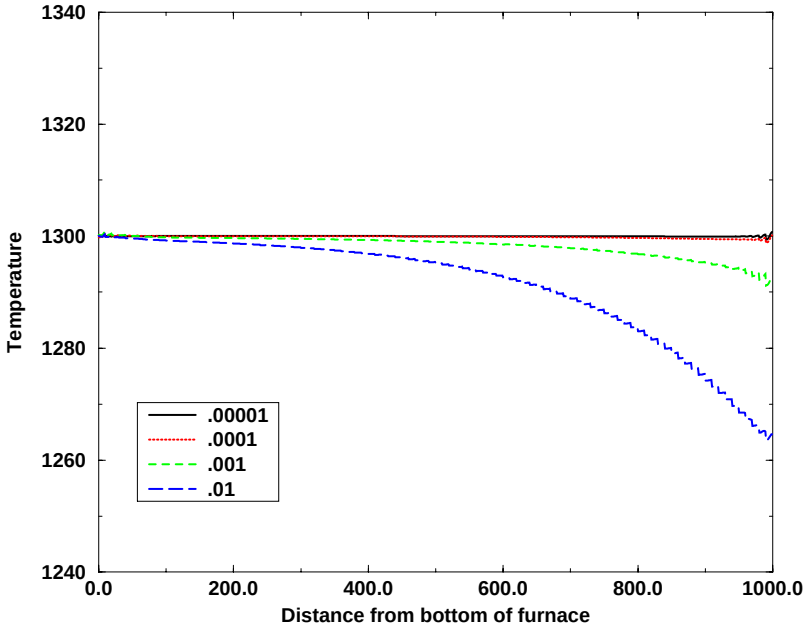


FIG. 11. *Computed temperatures for various values of RTOL using TRPDS.*

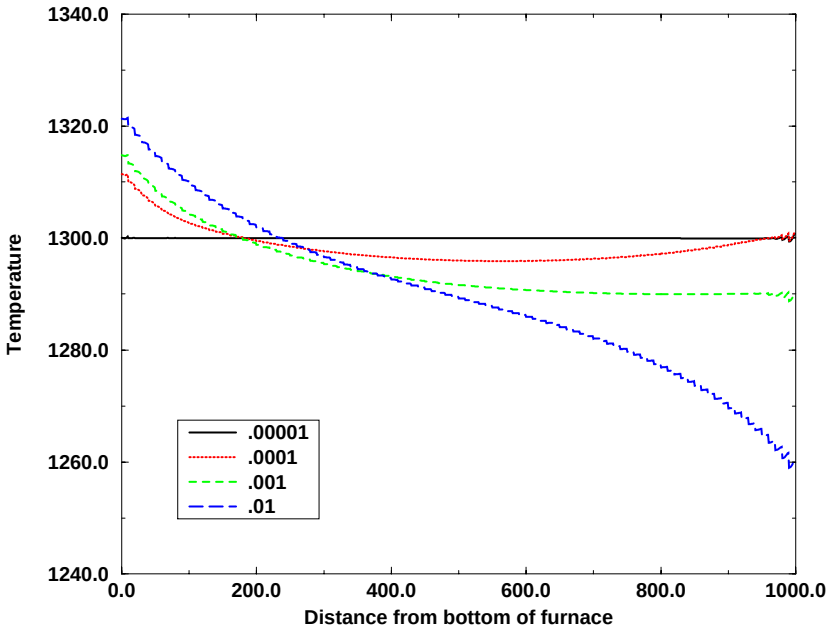


FIG. 12. *Computed temperatures for various values of RTOL using BFGS.*

This new class of algorithms was tested on a standard set of test problems, where it performed favorably against the traditional BFGS method. We also tested this new algorithm on a test case derived from an optimal design problem for a chemical vapor deposition furnace. The results indicate that the new method is competitive with the traditional BFGS method. In addition, the new method is more robust in the presence of noise that is generated by the use of less accurate PDE solvers. This is an important feature since many users would prefer to use less accurate PDE solvers in order to reduce the total computational time.

There are many new options to explore. In particular, it would be useful to develop strategies for bound constrained and general inequality constrained problems. It is also necessary to address issues related to the distribution of function evaluations in order to improve the efficiency of the algorithm. It would also be interesting to explore more general approximation models within this new framework.

Appendix. Tables 4–5 list the problems corresponding to the "Problem number" referred to in the various plots in the paper. Table 4 contains the first set of 24 test problems. Table 5, which also includes the dimension, contains the 22 CUTE test problems.

TABLE 4
Test problems.

Number	Problem
1	almost
2	broyden1a
3	broyden1b
4	broyden2a
5	broyden2b
6	bv
7	chain_singular
8	chain_wood
9	chebyquad
10	cragg_levy
11	epowell
12	erosen
13	gen_brown
14	gen_wood
15	ie
16	lin
17	lin0
18	lin1
19	penalty1
20	penalty2
21	toint_trig
22	tointbroy
23	trig
24	vardim

TABLE 5
CUTE problems.

Number	Problem	Dimension
1	arglinb	10
2	brownal	10
3	cosine	10
4	dixmaana	15
5	dixmaanl	15
6	eigenals	6
7	engval1	2
8	fletcbv2	10
9	freuroth	2
10	mancino	10
11	morebv	10
12	msqrtals	4
13	nondia	10
14	nonmsqrt	9
15	power	10
16	schmvett	3
17	sensors	2
18	sinquad	5
19	sparsine	10
20	sparsqur	10
21	tquartic	5
22	vareigvl	10

Acknowledgments. We wish to thank John Dennis and Matthias Heinkenschloss for many helpful discussions and for pointing out the relationship between our new algorithm and the framework for approximation models. We would also like to thank Vicki Howle and Suzanne Shontz for their work on parallel finite differencing and the comparisons of TRPDS to a speculative gradient trust-region algorithm. Finally, we thank an anonymous referee for many helpful comments and suggestions.

REFERENCES

[1] N. Alexandrov, J. E. Dennis, Jr., R. M. Lewis, and V. Torczon, *A trust region framework for managing the use of approximation models in optimization*, Structural Optim., 15 (1998), pp. 16–12.

[2] R. H. Byrd, R. B. Schnabel, and G. A. Shultz, *Parallel quasi-Newton methods for unconstrained optimization*, Math. Programming, 42 (1988), pp. 273–306.

[3] R. G. Carter, *On the global convergence of trust region algorithms using inexact gradient information*, SIAM J. Numer. Anal., 28 (1991), pp. 251–265.

[4] R. G. Carter, *Numerical experience with a class of algorithms for nonlinear optimization using inexact function and gradient information*, SIAM J. Sci. Comput., 14 (1993), pp. 368–388.

[5] A. R. Conn, N. I. M. Gould, and P. L. Toint, *Testing a class of methods for solving minimization problems with simple bounds on the variables*, Math. Comp., 50 (1988), pp. 399–430.

[6] J. E. Dennis, Jr., and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice–Hall, Englewood Cliffs, NJ, 1983.

[7] J. E. Dennis, Jr., and V. Torczon, *Direct search methods on parallel machines*, SIAM J. Optim., 1 (1991), pp. 448–474.

[8] P. E. Gill, W. Murray, and M. H. Wright, *Practical Optimization*, Academic Press, San Diego, CA, 1981.

[9] J. F. Grcar, *TWOPNT Program for Boundary Value Problems, Version* 3.10, Tech. report SAND91–8230, Sandia National Laboratory, Livermore, CA, 1992.

[10] W. G. Houf, J. F. Grcar, and W. G. Breiland, *A model for low pressure chemical vapor deposition in a hot-wall tubular reactor*, Materials Science Engineering B, Solid State Materials for Advanced Technology, 17 (1993), pp. 163–171.

[11] P. D. Hough and J. C. Meza, *A Class of Trust-Region Methods for Parallel Optimization*, Tech. report SAND98-8245, Sandia National Laboratories, Livermore, CA, 1999.

[12] V. E. Howle, S. M. Shontz, and P. D. Hough, *Some Parallel Extensions to Optimization Methods in OPT++*, Tech. report SAND2000-8877, Sandia National Laboratories, Livermore, CA, 2000.

[13] L. Ingber, *Simulated annealing: Practice versus theory*, Math. Comput. Modelling, 18 (1993), pp. 29–57.

[14] P. Jog, J. Y. Suh, and D. Van Gucht, *Parallel genetic algorithms applied to the traveling salesman problem*, SIAM J. Optim., 1 (1991), pp. 515–529.

[15] C. T. Kelley, *A Shamanskii-like acceleration scheme for nonlinear equations at singular roots*, Math. Comp., 47 (1986), pp. 609–623.

[16] P. J. M. Laarhoven, *Parallel variable metric algorithms for unconstrained optimization*, Math. Programming, 33 (1985), pp. 68–81.

[17] C. D. Moen, P. A. Spence, and J. C. Meza, *Automatic differentiation for gradient-based optimization of radiatively heated microelectronics manufacturing equipment*, in Proceedings of the 6th AIAA/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization, Bellevue, WA, 1996, pp. 1167–1175.

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

[19] P. K. Phua and Y. Zeng, *Parallel Quasi-Newton Algorithms for Large-Scale Optimization*, Tech. report TRB2/95, National University of Singapore, Singapore, 1995.

[20] L. B. Rall, *Convergence of the Newton process to multiple solutions*, Numer. Math., 9 (1966), pp. 23–37.

[21] R. B. Schnabel, *A view of the limitations, opportunities, and challenges in parallel nonlinear optimization*, Parallel Comput., 21 (1995), pp. 875–905.

[22] T. A. Straeter, *A Parallel Variable Metric Optimization Algorithm*, Tech. report NASA TN D-7329, NASA, Langley Research Center, Hampton, VA, 1973.

SIAM J. OPTIM.
Vol. 13, No. 1, pp. 283–297

BLOCK-ITERATIVE ALGORITHMS WITH UNDERRELAXED BREGMAN PROJECTIONS*

YAIR CENSOR† AND GABOR T. HERMAN‡

Abstract. The notion of relaxation is well understood for orthogonal projections onto convex sets. For general Bregman projections it was considered only for hyperplanes, and the question of how to relax Bregman projections onto convex sets that are not linear (i.e., not hyperplanes or half-spaces) has remained open. A definition of the underrelaxation of Bregman projections onto general convex sets is given here, which includes as special cases the underrelaxed orthogonal projections and the underrelaxed Bregman projections onto linear sets as given by De Pierro and Iusem [*J. Optim. Theory Appl.*, 51 (1986), pp. 421–440]. With this new definition, we construct a block-iterative projection algorithmic scheme and prove its convergence to a solution of the convex feasibility problem. The practical importance of relaxation parameters in the application of such projection algorithms to real-world problems is demonstrated on a problem of image reconstruction from projections.

Key words. convex feasibility, projection algorithms, Bregman functions, block-iterative algorithms, underrelaxation

AMS subject classifications. 90C25, 65K05

PII. S1052623401389439

1. Introduction.

The *convex feasibility problem* of finding a point in the nonempty intersection $C := \cap_{i=1}^{m} C_i \neq \emptyset$ of a family of closed convex subsets $C_i \subseteq R^n$, $1 \leq i \leq m$, of the n-dimensional Euclidean space is fundamental in many areas of mathematics and the physical sciences; see, e.g., Stark and Yang [32], Combettes [15], [16], and references therein. It has been used to model significant real-world problems in image reconstruction from projections—see, e.g., the general discussion in Herman [21]; in radiation therapy treatment planning, see Censor, Altschuler, and Powlis [7]; and in crystallography, see Marks, Sinkler, and Landree [28], to name but a few—and has been used under additional names such as *set theoretic estimation* or the *feasible set approach*. A common approach to such problems is to use *projection algorithms* (see, e.g., Bauschke and Borwein [2]), which employ *orthogonal projections* (i.e., nearest point mappings) onto the individual sets C_i.

Flexibility in the actual use of such projection algorithms is often gained by using *relaxation parameters*. If $P_\Omega(z)$ is the orthogonal projection of a point $z \in R^n$ onto a closed convex set $\Omega \subseteq R^n$, i.e.,

$$P_\Omega(z) := \operatorname{argmin}\{\| z - x \|_2 \quad | \quad x \in \Omega\}, \tag{1.1}$$

where $\| \cdot \|_2$ is the Euclidean norm in R^n, and if λ is the so-called *relaxation parameter*, then

$$P_{\Omega,\lambda}(z) := (1 - \lambda)z + \lambda P_\Omega(z) \tag{1.2}$$

*Received by the editors May 16, 2001; accepted for publication (in revised form) March 7, 2002; published electronically August 28, 2002.
http://www.siam.org/journals/siopt/13-1/38943.html

†Department of Mathematics, University of Haifa, Mt. Carmel, Haifa 31905, Israel (yair@math.haifa.ac.il). This author's research was supported by NIH grant HL-28438 and by grants 293/97 and 592/00 of the Israel Science Foundation, founded by the Israel Academy of Sciences and Humanities.

‡Department of Computer Science, The Graduate Center, City University of New York, 365 Fifth Avenue, New York, NY 10016 (gherman@gc.cuny.edu). This author's research was supported by NIH grants HL-28438 and HL-70472.

is the *relaxed projection* of z onto Ω with relaxation λ. In this paper we restrict our attention to the case in which $P_{\Omega,\lambda}(z)$ is a convex combination of z and $P_\Omega(z)$, i.e., when $\lambda \in [0,1]$. This is referred to as *underrelaxation.*

The well-known "projections onto convex sets" (POCS) algorithm for the convex feasibility problem allows such underrelaxation parameters; see Bregman [5], Gubin, Polyak, and Raik [20], Youla [33], and the reviews by Combettes [15], [16]. Starting from an arbitrary initial point $x^0 \in R^n$, the POCS algorithm's iterative step is

$$x^{k+1} = x^k + \lambda_k(P_{C_{i(k)}}(x^k) - x^k), \tag{1.3}$$

where $\{\lambda_k\}_{k\geq 0}$ are relaxation parameters and $\{i(k)\}_{k\geq 0}$ is a *control sequence*, $1 \leq i(k) \leq m$, for all $k \geq 0$, which determines the set $C_{i(k)}$ onto which the current iterate x^k is projected. The effects of relaxation parameters have been studied theoretically; see, e.g., Censor, Eggermont, and Gordon [9]. Their practical effect on early iterates of the POCS algorithm can be dramatic in some real-world situations, as we describe in section 6 below.

Bregman projections onto closed convex sets were introduced and utilized by Censor and Lent [10], based on Bregman's seminal paper [6], and were subsequently used for building sequential and parallel feasibility and optimization algorithms; see, e.g., Censor and Elfving [8], Censor and Reich [11], Censor and Zenios [13], De Pierro and Iusem [17], Kiwiel [24], [25], Bauschke and Borwein [3], and the references therein.

A *Bregman projection* of a point $z \in R^n$ onto a closed convex set $\Omega \subseteq R^n$ with respect to a suitably defined (see, Definition A.1 in the appendix) *Bregman function* f is denoted by $P_\Omega^f(z)$. It is formally defined as

$$P_\Omega^f(z) := \operatorname{argmin}\{D_f(x,z) \mid x \in \Omega \cap \operatorname{cl}S\}, \tag{1.4}$$

where $\operatorname{cl}S$ is the closure of the open convex set S, which is the *zone* of f, and $D_f(x,z)$ is the so-called *Bregman distance*, defined by

$$D_f(x,z) := f(x) - f(z) - \langle \nabla f(z), x - z\rangle \tag{1.5}$$

for all pairs $(x,z) \in \operatorname{cl}S \times S$. If $\Omega \cap \operatorname{cl}S \neq \emptyset$, then (1.4) defines a unique $P_\Omega^f(z) \in \operatorname{cl}S$ for every $z \in S$; see [13, Lemma 2.1.2]. If, in addition, $P_\Omega^f(z) \in S$ for every $z \in S$, then f is called *zone consistent* with respect to Ω.

Orthogonal projections are a special case of Bregman projections, obtained from (1.4) by choosing $f(x) = (1/2)\|x\|^2$ and $S = R^n$ (see, e.g., [13, Example 2.1.1]). However, despite this, relaxation of general (nonorthogonal) Bregman projections has not yet been defined. The lone exception is the special case in which the set Ω is a half-space, which has been described by De Pierro and Iusem [17] in the following manner. Let, for some $a \in R^n$, $a \neq 0$, and $b \in R$,

$$L = \{x \in R^n \mid \langle a, x\rangle \leq b\} \tag{1.6}$$

be a half-space. For a $z \notin L$, De Pierro and Iusem [17] define the *underrelaxed Bregman projection* of z onto L, with respect to a Bregman function f and with relaxation parameter $\rho \in [0,1]$, by

$$P_{L,\rho}^f(z) := P_{\widetilde{L}}^f(z), \tag{1.7}$$

where

$$\widetilde{L} = \{x \in R^n \mid \langle a, x\rangle \leq (1-\rho)\langle a, z\rangle + \rho b\}. \tag{1.8}$$

This means that the relaxed Bregman projection of z onto L is the unrelaxed Bregman projection of z onto a half-space $\widetilde{L}$ whose bounding hyperplane is parallel to that of L and lies between that of L and the point z.

Rewriting (1.6) as

$$L = \{x \in R^n \mid g(x) \leq 0\}, \tag{1.9}$$

with $g(x) := \langle a, x\rangle - b$, we can view (1.7) as the unrelaxed Bregman projection onto the inflated set $\widetilde{L}$ of (1.8), which can be redefined as

$$\widetilde{L} = \{x \in R^n \mid g(x) \leq \varepsilon\}, \tag{1.10}$$

where $\varepsilon = (1-\rho)g(z)$. (Note that it is easy to show that if $\Omega = L$ and $f(x) = (1/2)\|x\|^2$, then $P_{\Omega,\rho}(z)$ as defined by (1.2) is the same as $P^f_{L,\lambda}(z)$ as defined by (1.7) for any $\lambda \in [0,1]$ and any $z \in R^n$.) This approach of projection onto an inflated set would not necessarily work for a set Ω defined by a nonlinear function $g(x)$. This can be seen by taking a planar closed convex set Ω that is defined by an ellipse and considering its inflated set $\widetilde{\Omega}$ to be a confocal ellipse lying between Ω and the point $z \notin \Omega$. Obviously, the orthogonal projection of some $z \notin \Omega$ onto $\widetilde{\Omega}$ in this case is not a relaxed orthogonal projection of z onto Ω, because the two projections do not always lie along the same line.

Thus, we ask the following questions: (i) How should one define a relaxed Bregman projection onto a (not necessarily linear) closed convex set? The new definition should, of course, include as special cases at least both the orthogonal case for general convex sets and the underrelaxed Bregman projections onto half-spaces of De Pierro and Iusem [17] mentioned above. (ii) Can such relaxed Bregman projections be incorporated into a Bregman's projection algorithm for the convex feasibility problem? The Bregman's projection algorithm of [6] (see also [13, Algorithm 5.8.1]) allows only unrelaxed projectionsl; i.e., its iterative step is of the form

$$x^{k+1} = P^f_{C_{i(k)}}(x^k) \quad \text{for all } k \geq 0. \tag{1.11}$$

(iii) Is it possible to construct a block-iterative Bregman projection algorithm that will allow relaxed Bregman projections and variable blocks? Such an algorithm, with dynamically changing blocks, will naturally extend earlier block-iterative projection algorithms, such as the block-iterative ART (Kaczmarz) algorithm of Eggermont, Herman, and Lent [18], and the block-iterative projections (BIP) method of Aharoni and Censor [1] for the convex feasibility problem.

In this paper we constructively answer these three questions. We propose a definition for an underrelaxed Bregman projection onto a closed convex (not necessarily linear) set and prove convergence of a block-iterative projection algorithmic scheme with underrelaxed Bregman projections and dynamically varying blocks. This block-iterative scheme contains, as a new special case, the underrelaxed sequential Bregman projection algorithm for the convex feasibility problem, generalizing the underrelaxed POCS method. The paper is organized as follows. In section 2 we define underrelaxed Bregman projections and analyze some of their properties. In section 3 we present the new block-iterative algorithmic scheme with underrelaxed Bregman projections and prove its convergence in section 4. In section 5 we demonstrate the new block-iterative algorithmic scheme by working out in detail the case with underrelaxed entropy projections. Computational experience with any algorithm that uses underrelaxed nonorthogonal Bregman projections is still missing, but in section 6 we

provide evidence of the advantages of using underrelaxation parameters when working with orthogonal projections in the real-world application of image reconstruction from projections. For the reader's convenience, we attach an appendix that summarizes the definitions and results from the theory of Bregman functions which are used in this paper.

2. Underrelaxation of Bregman projections. We consider that the underrelaxed Bregman projection with Bregman function f and relaxation parameter $\lambda \in [0,1]$ of a point z onto a closed convex set Ω, denoted by $P^f_{\Omega,\lambda}(z)$, should satisfy

$$\nabla f(P^f_{\Omega,\lambda}(z)) = (1-\lambda)\nabla f(z) + \lambda \nabla f(P^f_\Omega(z)). \tag{2.1}$$

The justification for this is that it makes $P^f_{\Omega,\lambda}(z)$ the appropriate (for λ) *convex combination with respect to the Bregman function* f, as defined by Censor and Reich [11, Definiton 4.1], of z and of $P^f_\Omega(z)$. In stating conditions under which (2.1) is a valid definition of $P^f_{\Omega,\lambda}(z)$, we make use of a result which appears in Bauschke and Borwein [3, Fact 2.9] and is based on a finding of Rockafellar [30, Theorem 26.5]; see Theorem A.6 in the appendix of this paper.

PROPOSITION 2.1. *Let $f : R^n \to R$ be a Bregman function with zone $S = \operatorname{int}(\operatorname{dom} f)$, and let $\Omega \subseteq R^n$ be a closed convex set such that $\Omega \cap \operatorname{cl} S \neq \emptyset$. If f is a Legendre function, then for any $z \in S$ and any $\lambda \in [0,1]$ there exists a unique $x \in S$ satisfying*

$$\nabla f(x) = (1-\lambda)\nabla f(z) + \lambda \nabla f(P^f_\Omega(z)). \tag{2.2}$$

Proof. Since f is of the Legendre type, we have, from Theorem A.7 below, that it is zone consistent with respect to Ω. Moreover, [3, Fact 2.9] (see also Theorem A.6 in the appendix below) guarantees that $\nabla f(S)$ is equal to the interior of the domain of the conjugate function f^*. Since $\operatorname{dom} f^*$ is a convex set (see, e.g., Luenberger [26, Proposition 1, p. 196]), its interior is also convex, and thus the right-hand side of (2.2) is in $\nabla f(S)$. Theorem A.6 now ensures the existence and uniqueness of an x in S satisfying (2.2). ∎

In view of this, our definition of $P^f_{\Omega,\lambda}(z)$ is that it is the x whose existence and uniqueness is guaranteed by the proposition. The ability to invert the gradient operator is crucial for the applicability of Proposition 2.1, as well as for the applicability of the algorithmic formula; see (3.4) below, which describes our proposed block-iterative algorithmic scheme. Therefore, using functions which are both Bregman and Legendre (see [3, Remark 5.4]) secures both the zone consistency and gradient invertibility. An anonymous referee made the conjecture that in Proposition 2.1 it may be necessary to assume only that f is of the Bregman/Legendre type, a less restrictive property. Examples of Bregman and Legendre functions are provided in Bauschke and Borwein [3].

REMARK 2.2. *If there exists an $x \in S$ that minimizes*

$$(1-\lambda)D_f(x,z) + \lambda D_f(x, P^f_\Omega(z)) \tag{2.3}$$

over $\operatorname{cl} S$, *then that x satisfies* (2.2), *as follows by substituting for D_f using* (1.5) *and setting the gradient to zero. This provides additional indication of the reasonableness of our definition of underrelaxed Bregman projections.*

For $f(x) = (1/2)\|x\|^2$ with $S = R^n$, our definition of an underrelaxed Bregman projection coincides with the notion of underrelaxation of orthogonal projections. The

next proposition shows that, when Ω is the L of (1.6), our definition of underrelaxation of Bregman projections coincides with the one given by De Pierro and Iusem in [17], provided that their assumptions and the conditions of Proposition 2.1 are met. De Pierro and Iusem made the additional assumption that f is not only zone consistent with respect to the bounding hyperplane of L but also with respect to the bounding hyperplane of any half-space $\widetilde{L}$ as defined in (1.8). (This was termed in [10] the *strong zone consistency of f with respect to the bounding hyperplane of L*; see, e.g., [13, Definition 2.2.1].)

PROPOSITION 2.3. *Let f be a Bregman function with zone S, and Ω be a half-space L as in* (1.6) *satisfying the conditions of Proposition* 2.1. *Assume also that f is strongly zone consistent with respect to the bounding hyperplane of L. Then, for every ρ $(0 \le \rho \le 1)$, there exists a λ $(0 \le \lambda \le 1)$ such that $P^f_{L,\rho}(z)$ of* (1.7) *fulfills* (2.2) *with $\Omega = L$ for every $z \in S$.*

Proof. If $z \in S \cap L$, then there is nothing to prove. Therefore, let $z \in S$ be outside the half-space L. From the well-known characterization of Bregman projections onto hyperplanes (see [13, Lemma 2.2.1]), we know that the projection $P^f_H(z)$ onto the bounding hyperplane H of the half-space L is uniquely determined, along with the (real) *associated projection parameter* θ, by the system

$$\begin{aligned} \nabla f(P^f_H(z)) &= \nabla f(z) + \theta a, \\ \langle P^f_H(z), a\rangle &= b. \end{aligned} \tag{2.4}$$

We claim that $P^f_L(z) = P^f_H(z)$. To see this, first note that the θ of (2.4) is negative (because z is outside L, see [13, Lemma 2.2.2, equation (2.27)]). Now consider any $x \in L$. Multiplying the inequality of (1.6) by the negative of the θ of (2.4), and then using the second line and, subsequently, the first line of (2.4), we get that

$$\langle \nabla f(z) - \nabla f(P^f_H(z)),\ x - P^f_H(z)\rangle \le 0 \quad \text{for all } x \in L \cap \operatorname{cl} S. \tag{2.5}$$

This uniquely characterizes $P^f_H(z)$ as the projection $P^f_L(z)$; see [13, Theorem 2.4.2]. Similarly, by letting $\widetilde{H}$ be the bounding hyperplane of $\widetilde{L}$ and using strong zone consistency, we find that the Bregman projection $P^f_{\widetilde{L}}(z)$ of z onto $\widetilde{L}$ of (1.8) is in fact $P^f_{\widetilde{H}}(z)$ and is uniquely determined, along with the associated projection parameter $\widetilde{\theta}$, by the system

$$\begin{aligned} \nabla f(P^f_{\widetilde{L}}(z)) &= \nabla f(z) + \widetilde{\theta} a, \\ \langle P^f_{\widetilde{L}}(z), a\rangle &= (1-\rho)\langle a, z\rangle + \rho b. \end{aligned} \tag{2.6}$$

Using (1.7) and the first lines of (2.4) and (2.6), we obtain that (recall that $\theta < 0$)

$$\nabla f(P^f_{L,\rho}(z)) = \nabla f(z) + \widetilde{\theta} a = \nabla f(z) + \frac{\widetilde{\theta}}{\theta}(\nabla f(P^f_L(z)) - \nabla f(z)) \tag{2.7}$$

$$= \left(1 - \frac{\widetilde{\theta}}{\theta}\right) \nabla f(z) + \frac{\widetilde{\theta}}{\theta} \nabla f(P^f_L(z)). \tag{2.8}$$

Since θ is negative and by [13, Lemma 2.2.2] $\widetilde{\theta}$ is nonpositive, we have from [13, Lemma 2.2.4] that $\theta \le \widetilde{\theta}$. These facts guarantee that if we define $\lambda = \widetilde{\theta}/\theta$, then $0 \le \lambda \le 1$, which completes the proof. ∎

If $f : R^n \to R$ is a Bregman function with zone S, $\Omega \subseteq R^n$ is a closed convex set such that $\Omega \cap \operatorname{cl} S \neq \emptyset$, and f is zone consistent with respect to Ω, then it follows immediately from (1.4) that P_Ω^f is an *idempotent operator*; i.e.,

$$P_\Omega^f(P_\Omega^f(z)) = P_\Omega^f(z) \tag{2.9}$$

for any $z \in S$. For underrelaxed projections we have the result of the next proposition, which trivially holds for orthogonal projections.

PROPOSITION 2.4. *Let f be a Bregman function with zone S, and Ω be a closed convex set satisfying the conditions of Proposition* 2.1. *Then, for any $z \in S$, we have*

$$P_\Omega^f(P_{\Omega,\lambda}^f(z)) = P_\Omega^f(z) \tag{2.10}$$

for all $\lambda \in [0, 1]$.

Proof. In the case of $\lambda = 1$, (2.10) follows from (2.9). We now assume that $\lambda \in [0, 1)$. The projection $P_\Omega^f(z)$ can be characterized (see Theorem A.4) as the unique element of $\Omega \cap \operatorname{cl} S$ for which

$$\langle \nabla f(z) - \nabla f(P_\Omega^f(z)),\ x - P_\Omega^f(z) \rangle \leq 0 \quad \text{for all } x \in \Omega \cap \operatorname{cl} S. \tag{2.11}$$

Multiplying this by $(1 - \lambda)$ and substituting for $(1 - \lambda)\nabla f(z)$ using (2.1) yields that, for all $x \in \Omega \cap \operatorname{cl} S$,

$$\langle \nabla f(P_{\Omega,\lambda}^f(z)) - \nabla f(P_\Omega^f(z)),\ x - P_\Omega^f(z) \rangle \leq 0. \tag{2.12}$$

Using the characterization of Theorem A.4 in the appendix, we again get (2.10). ∎

3. A block-iterative algorithmic scheme with underrelaxed Bregman projections. In this section we propose a *block-iterative algorithmic scheme* with underrelaxed Bregman projections for the solution of the convex feasibility problem. By *block-iterative* we mean that, at the kth iteration, the next iterate x^{k+1} is generated from the current iterate x^k by using a subset (called a block) of the family of sets $\{C_i\}_{i=1}^m$ of the convex feasibility problem; see, e.g., [13, section 1.1.3]. We use the term *algorithmic scheme* to emphasize that different specific algorithms may be derived by different choices of Bregman functions and by various block structures. For example, if all blocks consist of a single set, then our scheme gives rise to a sequential row-action-type algorithm (cf. [13, Definition 6.2.1] for this term). Taking the other extreme, if we let every block contain all sets, then we obtain a fully simultaneous algorithm. Such a block-iterative scheme for the convex feasibility problem was first proposed by Aharoni and Censor [1], using orthogonal projections onto convex sets. That block-iterative projections (BIP) method generalizes the sequential POCS method of Bregman [5] and Gubin, Polyak, and Raik [20]. (See also Stark and Yang [32] and Censor and Zenios [13] for many more related references.) Our proposed block-iterative scheme extends Aharoni and Censor's BIP method by employing underrelaxed Bregman projections which contain the underrelaxed orthogonal projections as a special case.

Appealing again to the definition of a convex combination with respect to a Bregman function f as defined by Censor and Reich [11, Definiton 4.1], the natural formula for a block-iterative step using underrelaxed Bregman projections is

$$\nabla f(x^{k+1}) = \sum_{i=1}^m v_i^k \nabla f(P_{C_i,\lambda_i^k}^f(x^k)), \tag{3.1}$$

where x^k is the kth iterate, $\lambda_i^k \in [0,1]$ is the relaxation parameter used in the underrelaxed Bregman projection onto the set C_i during the kth iterative step, and the v_i^k are the weights of the convex combination for the kth iterative step (i.e., $v_i^k \geq 0$ for $1 \leq i \leq m$ and $\sum_{i=1}^m v_i^k = 1$). Note that, under the assumptions of Proposition 2.1, if $x^k \in S$, then x^{k+1} is uniquely defined by (3.1) and is also in S.

To simplify notation, from now on we use P_i^f to abbreviate $P_{C_i}^f$. Further, we observe that, according to (2.1),

$$\nabla f(x^{k+1}) = \sum_{i=1}^m v_i^k((1-\lambda_i^k)\nabla f(x^k) + \lambda_i^k \nabla f(P_i^f(x^k))). \tag{3.2}$$

Defining $w_i^k = v_i^k \lambda_i^k$ for $1 \leq i \leq m$, and introducing

$$w_{m+1}^k = 1 - \sum_{i=1}^m w_i^k \quad \text{and} \quad C_{m+1} = R^n, \tag{3.3}$$

we get that

$$\nabla f(x^{k+1}) = \sum_{i=1}^{m+1} w_i^k \nabla f(P_i^f(x^k)), \tag{3.4}$$

with $w_i^k \geq 0$ for $1 \leq i \leq m+1$ and $\sum_{i=1}^{m+1} w_i^k = 1$.

4. A convergence theorem. The following theorem establishes the convergence to a solution of the convex feasibility problem of a sequence generated by any block-iterative algorithm with underrelaxed Bregman projections. The method of proof is closely related to previous proofs of other results in this field; see, e.g., Bauschke and Borwein [3, Theorem 8.1] and Censor and Reich [11, Theorem 3.1]. We will make use of a further condition on the w_i^k of (3.4).

CONDITION 4.1. *Let w_i^k be real numbers for $k \geq 0$ and $1 \leq i \leq m$, and for each k let*

$$I(k) := \{i \mid 1 \leq i \leq m,\ w_i^k > 0\}. \tag{4.1}$$

(i) *There exists an $\varepsilon > 0$ such that $w_i^k \geq \varepsilon$ for all $k \geq 0$ and $i \in I(k)$.*

(ii) *Each i, $1 \leq i \leq m$, is included in infinitely many sets $I(k)$.*

A practitioner might desire to rephrase Condition 4.1 in terms of the weights v_i^k and the relaxation parameters λ_i^k, using (3.3) and the line above it. Condition 4.1(i) states that, for some positive ε, if v_i^k and λ_i^k are both positive, then they are both greater than or equal to ε. It should be noted, however, that Condition 4.1(i) is stronger than the condition used by Aharoni and Censor [1] regarding the weights they used in their BIP method. The (weaker) condition that they use is that "for all $i = 1, 2, \ldots, m$, the series $\sum_{k=0}^\infty v_i^k = +\infty$." The purpose of our condition, as well as that of the condition of [1], is to guarantee that none of the sets C_i is "gradually ignored" by ever-diminishing weights. We do not know whether our convergence result, presented below, can be strengthened by using a condition similar to that of [1]. Notice also that if $\lambda_i^k = 1$ for all $i = 1, 2, \ldots, m$ and all $k \geq 0$, then no underrelaxation takes place and $w_{m+1}^k = 0$ for all $k \geq 0$, leaving only the weights v_i^k to affect the algorithm's progress. Finally, observe that the sequential algorithm is obtained from (3.4) by choosing, for every $k \geq 0$, the weights

$$v_i^k = \begin{cases} 1 & \text{if } i = i(k), \\ 0 & \text{otherwise,} \end{cases} \tag{4.2}$$

where $\{i(k)\}_{k\geq 0}$ is a control sequence such as, e.g., the *cyclic control* defined by $i(k) = k(\mathrm{mod}\, m) + 1$ for all $k \geq 0$.

THEOREM 4.2. *Let $f : R^n \to R$ be a Bregman function, and let $S = \mathrm{int}(\mathrm{dom}\, f)$ be its zone. Let $C_i \subseteq R^n$ be closed convex sets such that $\cap_{i=1}^m C_i \cap \mathrm{cl}\, S \neq \emptyset$. Assume that f is also a Legendre function. For $k \geq 0$, let w_i^k be nonnegative for $1 \leq i \leq m+1$ such that $\sum_{i=1}^{m+1} w_i^k = 1$ and Condition 4.1 is satisfied. Then the sequence $\{x^k\}_{k\geq 0}$ generated by (3.4) from any $x^0 \in S$ converges to a point $x^* \in \cap_{i=1}^m C_i \cap \mathrm{cl}\, S$.*

Proof. The well-definedness of the algorithm described by (3.4) can be shown by a straightforward generalization of the proof of Proposition 2.1 (in which (3.4) is replaced by (2.2)). Legendreness of the function f also ensures, by Theorem A.7 below, the zone consistency of f with respect to each set C_i, a fact which is repeatedly used in this proof. Using (1.5) and (3.4), we have, for every $k \geq 0$ and for any $x \in \mathrm{cl}\, S$,

$$D_f(x, x^{k+1}) = \sum_{i=1}^{m+1} w_i^k (f(x) - f(x^{k+1}) - \langle \nabla f(P_i^f(x^k)),\ x - x^{k+1}\rangle). \tag{4.3}$$

By repeated application of (1.5) to the expression inside the parentheses on the right-hand side of (4.3), we obtain

$$D_f(x, x^{k+1}) = \sum_{i=1}^{m+1} w_i^k (D_f(x, P_i^f(x^k)) - D_f(x^{k+1}, P_i^f(x^k))). \tag{4.4}$$

Therefore,

$$\begin{aligned} D_f(x, x^k) - D_f(x, x^{k+1}) &= \sum_{i=1}^{m+1} w_i^k D_f(x^{k+1}, P_i^f(x^k)) \\ &\quad + \sum_{i=1}^{m+1} w_i^k (D_f(x, x^k) - D_f(x, P_i^f(x^k))). \end{aligned} \tag{4.5}$$

For any point $x \in C_i \cap \mathrm{cl}\, S$, the difference under the sum in the last line fulfills

$$D_f(x, x^k) - D_f(x, P_i^f(x^k)) \geq D_f(P_i^f(x^k), x^k) \geq 0. \tag{4.6}$$

This follows from well-known inequalities in the theory of Bregman distances. The left-hand inequality in (4.6) follows by replacing z, y, and Ω in Theorem A.3 by x, x^k, and C_i, respectively, and the nonnegativity in (4.6) follows from [13, Lemma 2.1.1]. Since all quantities on the right-hand side of (4.5) are nonnegative, we conclude from (4.5) that, for any point $x \in \cap_{i=1}^m C_i \cap \mathrm{cl}\, S$,

$$D_f(x, x^{k+1}) \leq D_f(x, x^k) \quad \text{for all} \quad k \geq 0, \tag{4.7}$$

which means that the sequence $\{x^k\}_{k\geq 0}$ is *D_f-Fejér-monotone* with respect to $\cap_{i=1}^m C_i$ and implies that $\{x^k\}_{k\geq 0}$ is bounded; see [13, p. 108]. Therefore, to conclude the proof, we will show the following: (i) if there exists a cluster point x^* in $C = \cap_{i=1}^m C_i$, then it is the limit of the sequence, and (ii) every cluster point must belong to C.

We first make the observation that, for any $x \in C \cap \mathrm{cl}\, S$, (4.7) and the nonnegativity of $\{D_f(x, x^k)\}_{k\geq 0}$ guarantee the existence of the limit

$$\lim_{k\to\infty} D_f(x, x^k) = \theta. \tag{4.8}$$

To prove (i), let $x^* \in C$ be a cluster point of $\{x^k\}_{k\geq 0}$, and assume that x^{**} is another cluster point, i.e.,

$$\lim_{\substack{k\to\infty \\ k\in K_1}} x^k = x^* \quad \text{and} \quad \lim_{\substack{k\to\infty \\ k\in K_2}} x^k = x^{**}, \tag{4.9}$$

with infinite $K_1 \subseteq N$ and infinite $K_2 \subseteq N$ and $N := \{0, 1, 2, \dots\}$. Since $x^k \in S$ for all $k \geq 0$, $x^* \in \operatorname{cl} S$, and thus (4.8) holds for $x = x^*$ (for some θ). Applying the property of Bregman functions given by Definition A.1(iv) to the subsequence defined by $k \in K_1$, we get that in fact

$$\lim_{k\to\infty} D_f(x^*, x^k) = 0, \tag{4.10}$$

which is true, in particular, for the subsequence defined by $k \in K_2$. Then, using another property of Bregman functions given by Definition A.1(v), $x^* = x^{**}$ follows.

To prove (ii), assume, by way of negation, that

$$\lim_{l\to\infty} x^{k_l} = x^* \quad \text{and} \quad x^* \notin C. \tag{4.11}$$

Define

$$I_{in} := \{i \mid 1 \leq i \leq m, \;\; x^* \in C_i\}, \tag{4.12}$$

$$I_{out} := \{i \mid 1 \leq i \leq m, \;\; x^* \notin C_i\}. \tag{4.13}$$

Because of Condition 4.1(ii), we may assume without loss of generality (passing to a subsequence if necessary) that, for every $l = 1, 2, \dots$,

$$I(k_l) \cup I(k_l + 1) \cup \cdots \cup I(k_{l+1} - 1) = \{i \mid 1 \leq i \leq m\}. \tag{4.14}$$

For every $l = 1, 2, \dots$, let μ_l be the smallest element in the set

$$\{k_l, k_l + 1, k_l + 2, \dots, k_{l+1} - 1\} \tag{4.15}$$

such that

$$I(\mu_l) \cap I_{out} \neq \emptyset. \tag{4.16}$$

Such an element exists by (4.14) and since, by (4.11) and (4.13), $I_{out} \neq \emptyset$.

We want to show now that the sequence $\{x^{\mu_l}\}_{l\geq 0}$ also converges to x^*. By definition, $k_l \leq \mu_l$ for all $l = 1, 2, \dots$. If $\nu \in [k_l, \mu_l)$, then

$$I(\nu) \subseteq I_{in}, \tag{4.17}$$

and so, from (4.5), for any $x \in \operatorname{cl} S$,

$$\begin{aligned} D_f(x, x^\nu) - D_f(x, x^{\nu+1}) = &\sum_{i\in I_{in}} w_i^\nu D_f(x^{\nu+1}, P_i^f(x^\nu)) \\ &+ \sum_{i\in I_{in}} w_i^\nu (D_f(x, x^\nu) - D_f(x, P_i^f(x^\nu))). \end{aligned} \tag{4.18}$$

For any point $x \in \cap_{i \in I_{in}} C_i \cap \operatorname{cl} S$, it follows from (4.6) and (4.18) that, for $\nu \in [k_l, \mu_l)$,

$$D_f(x, x^{\nu+1}) \leq D_f(x, x^{\nu}). \tag{4.19}$$

In other words, with x replaced by x^*, we have, for all $l = 1, 2, \ldots$,

$$\begin{aligned} 0 &\leq D_f(x^*, x^{\mu_l}) \leq D_f(x^*, x^{\mu_l - 1}) \\ &\leq \cdots \leq D_f(x^*, x^{k_l+1}) \leq D_f(x^*, x^{k_l}). \end{aligned} \tag{4.20}$$

Letting $l \to \infty$ in (4.20) yields, by (4.11) and Definition A.1(iv) in the appendix,

$$\lim_{l \to \infty} D_f(x^*, x^{\mu_l}) = 0. \tag{4.21}$$

As a subsequence of the whole sequence $\{x^k\}_{k \geq 0}$, which is bounded, $\{x^{\mu_l}\}_{l \geq 0}$ is bounded and thus has a cluster point. Combining Definition A.1(v) with (4.21) shows that any convergent subsequence of $\{x^{\mu_l}\}_{l \geq 0}$ must converge to x^*; hence

$$\lim_{l \to \infty} x^{\mu_l} = x^*. \tag{4.22}$$

From (4.16) it follows that there exists an index $\hat{\imath} \in I_{out}$ such that $\hat{\imath} \in I(\mu_l)$ for infinitely many indices l. Removing from the sequence $\{\mu_l\}_{l \geq 0}$ all elements μ_l for which $\hat{\imath} \notin I(\mu_l)$, we end up with a new infinite sequence $\{\mu_l\}_{l \geq 0}$ such that $\hat{\imath} \in I(\mu_l) \cap I_{out}$ for $l = 1, 2, \ldots$. Taking an arbitrary $x \in C \cap \operatorname{cl} S$, consider the limits of both sides of (4.5) for the new sequence $\{\mu_l\}_{l \geq 0}$. Due to (4.8), the left-hand side converges to zero, and, therefore, so must the right-hand side. Since all quantities on the right-hand side are nonnegative and $w_{\hat{\imath}}^{\mu_l} \geq \varepsilon > 0$ (for all $l = 1, 2, \ldots$) by Condition 4.1(i), we obtain that

$$\lim_{l \to \infty} (D_f(x, x^{\mu_l}) - D_f(x, P_{\hat{\imath}}^f(x^{\mu_l}))) = 0. \tag{4.23}$$

From (4.6) we obtain

$$\lim_{l \to \infty} D_f(P_{\hat{\imath}}^f(x^{\mu_l}), x^{\mu_l}) = 0. \tag{4.24}$$

If we could show that $\{P_{\hat{\imath}}^f(x^{\mu_l})\}_{l \geq 0}$ is bounded, then (4.22) and (4.24) would imply, by using again Definition A.1(v), that

$$\lim_{l \to \infty} P_{\hat{\imath}}^f(x^{\mu_l}) = x^*, \tag{4.25}$$

which means that $x^* \in C_{\hat{\imath}}$, yielding the sought-after contradiction with the choice of $\hat{\imath}$ made above.

Therefore, we conclude the proof by showing that $\{P_{\hat{\imath}}^f(x^{\mu_l})\}_{l \geq 0}$ is bounded. Indeed, (4.6) for $i = \hat{\imath}$, with $k = \mu_l$ and for $x \in C \cap \operatorname{cl} S$, shows that

$$D_f(x, P_{\hat{\imath}}^f(x^{\mu_l})) \leq D_f(x, x^{\mu_l}) - D_f(P_{\hat{\imath}}^f(x^{\mu_l}), x^{\mu_l}) \quad \text{for every} \quad l \geq 0. \tag{4.26}$$

Applying (4.7) and (4.24) to (4.26) shows that $\{D_f(x, P_{\hat{\imath}}^f(x^{\mu_l}))\}_{l \geq 0}$ is bounded, which, by Definition A.1(iii) in the appendix, implies that $\{P_{\hat{\imath}}^f(x^{\mu_l})\}_{l \geq 0}$ is bounded, and this concludes the proof. □

5. An example: Block-iterative underrelaxed entropy projections. A well-known Bregman function is the negative "$x \log x$" entropy (also called "Shannon's entropy") function; see [13, Example 2.1.2] and the many references given to the literature on this topic in that book, or consult the book by Fang, Rajasekera, and Tsao [19] and its references. The "$x \log x$" entropy has been used in numerous applications in science and engineering, up to and including recent work in the field of computational machine learning; see, e.g., Collins, Shapire, and Singer [14]. It is denoted by $\operatorname{ent} x$ and maps the nonnegative orthant R^n_+ into R according to

$$\operatorname{ent} x := -\sum_{j=1}^{n} x_j \log x_j, \tag{5.1}$$

where "log" denotes the natural logarithmic function and, by definition, $0 \log 0 = 0$. Its negative, $f(x) = \sum_{j=1}^{n} x_j \log x_j$, is a Bregman function with zone $S = \operatorname{int} R^n_+$ (see [13, Lemma 2.1.3]), the jth component of whose gradient is $\partial f / \partial x_j = 1 + \log x_j$.

In order to derive a block-iterative algorithm with underrelaxed Bregman entropy projections for the iterative solution of a linear system of equations $Ax = b$, we consider the sets

$$C_i = \{x \mid \langle a^i, x \rangle = b_i\} \qquad \text{for } i = 1, 2, \ldots, m, \tag{5.2}$$

where $a^i \in R^n$ is the ith column of the transposed matrix A^T and $b_i \in R$ is the ith component of $b \in R^m$. The iterative step (3.4) takes the form

$$\log x_j^{k+1} = \sum_{i=1}^{m+1} w_i^k \log(P_i^f(x^k))_j \qquad \text{for } j = 1, 2, \ldots, n. \tag{5.3}$$

Using the first line of (2.4) (with H, z, a, and θ replaced by C_i, x^k, a^i, and θ_i^k, respectively), substituting into (5.3), and taking exponents, we obtain

$$x_j^{k+1} = x_j^k \prod_{i=1}^{m} \exp(w_i^k \theta_i^k a_j^i) \qquad \text{for } j = 1, 2, \ldots, n, \tag{5.4}$$

where θ_i^k is the Bregman parameter associated with the "entropy projection" of x^k onto the ith hyperplane C_i. If one replaces the θ_i^k's in the iterative step (5.4) with the quantities

$$d_i^k := \log \frac{b_i}{\langle a^i, x^k \rangle} \tag{5.5}$$

for all i and all k, then the resulting formula resembles the iterative step formula of the block-iterative MART algorithm of Censor and Segman [12] (see also [13, Algorithm 6.7.1, equation (6.124)]), the difference being the lack of underrelaxation parameters and of variable block structure and composition in the latter.

6. On the practical usefulness of underrelaxation parameters. In this section we demonstrate the importance of underrelaxation parameters in the field of image reconstruction from projections. Projection algorithms have been used to solve the fully discretized model in this field, and experimental work has shown again and again that there are great advantages in using underrelaxation of the projections. For a recent example in the area of positron emission tomography (PET), see Obi et

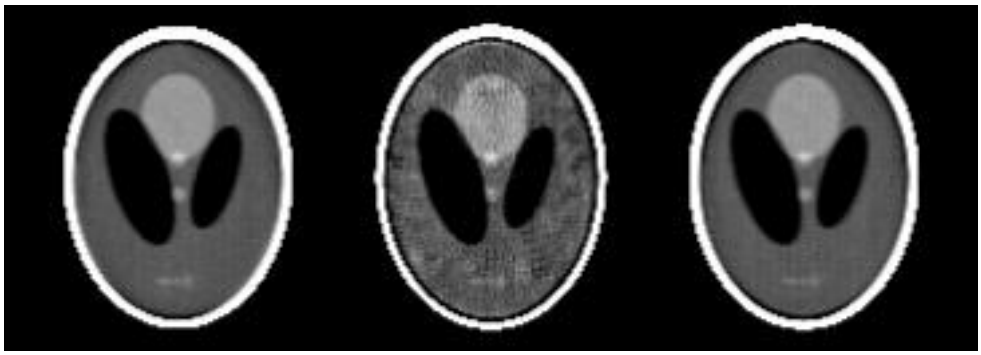


FIG. 6.1. *Slices of the phantom (left), reconstruction with no relaxation (center), and reconstruction with underrelaxation (right). See the text for details.*

al. [29], and in the area of electron microscopy, see Marabini, Herman, and Carazo [27]. Since in such practical applications the data are physically collected, and thus the feasibility condition in Theorem 4.2 cannot be guaranteed, we report here on an experiment that illustrates the usefulness of underrelaxation when all conditions of Theorem 4.2 are satisfied.

The experiment has been performed with the algebraic reconstruction technique (ART) described in Herman, Matej, and Carvalho [23, equation (6)] for the purpose of image reconstruction from x-ray data obtained by a scanner utilizing a helical cone-beam data collection geometry. In terms of our block-iterative step formula (3.1), $f(x) = (1/2)\|x\|^2$, the C_i are hyperplanes, and, for every $k \geq 0$, $v_i^k = 1$ for exactly one $i = i(k)$ and is zero otherwise, and the relaxation parameters $\lambda_i^k = \lambda$ are constant. We will be comparing the values $\lambda = 1$, that is, no relaxation, with $\lambda = 0.01$, which amounts to quite strong underrelaxation. The number of hyperplanes is $m = 4{,}915{,}000$, and the dimensionality of the image vector x is $n = 965{,}887$. To insure feasibility, we used the object reconstructed in [23] from the not necessarily feasible data used in that paper. (In other words, we replaced the system of equations $Ax = b$ that was treated in [23] by the system $Ax = As$, where s is the output of the algorithm reported in [23].) This reconstructed object is to be interpreted as values within a rectangular region of the three-dimensional space; a graphical representation of a single slice through this region is shown on the left of Figure 6.1. In this graphical representation, all values less than or equal to 1.00 are shown as black, all values greater than or equal to 1.04 are shown as white, and the intermediate values are represented by grey levels. For our purposes, this previously reconstructed object is the "phantom" (test image), which is the object (vector) in the intersection of $4,915,000$ hyperplanes whose descriptions are known to our program.

Under the conditions of this special case of (3.1), it is known that the algorithm (provided that it is started with the same vector) should, in the limit, converge to the same vector irrespective of which of the two investigated values of the relaxation parameter is chosen (assuming perfect computer accuracy); this follows, e.g., from Herman, Lent, and Lutz [22, Corollary 1] or from Bauschke et al. [4, Fact 2.2]. However, for such large problems, the algorithm is computationally intensive, and thus it is important that one should get to a reasonable solution in relatively few steps. For those who have not had experience with such projection algorithms, it may come as a surprise that underrelaxation is actually useful for this purpose. We illustrate this in Figure 6.1, in which the central and right images show our new reconstructions,

using no relaxation, i.e., $\lambda = 1$, and underrelaxation, i.e., $\lambda = 0.01$, respectively. The iteration index k at which the algorithm was stopped is the same in both cases, $k = 16m$; i.e., we have cycled through the data 16 times. The same slice through the three-dimensional region is shown in all three images, represented in the same way. The quality of the underrelaxed reconstruction is so good that it is practically indistinguishable from the phantom; this is certainly not the case for the reconstruction with no relaxation. This is also reflected by numerical calculations: considering only those locations in space (not only in the slice shown in Figure 6.1) for which the values of the phantom are in the range $[1.00, 1.04]$, the Euclidean distance between the phantom and reconstruction is 2.2 in the underrelaxed case and it is 6.9 in the no-relaxation case.

Thus this experiment, satisfying the conditions of Theorem 4.2, confirms the previously reported results in applications: a small relaxation parameter allows us to get to a high quality reconstruction faster than is possible with no relaxation.

Appendix. Some definitions and results from Bregman function theory. In this appendix we review some definitions and results from the theory of Bregman functions used in this paper.

DEFINITION A.1. *Let S be a nonempty open convex set in R^n with closure* $\operatorname{cl} S$. *Let $f : \operatorname{cl} S \to R$ be a differentiable function, and define $D_f(x, z) : \operatorname{cl} S \times S \to R$ by*

$$D_f(x, z) = f(x) - f(z) - \langle \nabla f(z), x - z \rangle.$$

We say that f is a Bregman function with zone S, and that D_f is the Bregman distance associated with it, if the following conditions are satisfied:

(i) *f is continuous and strictly convex on* $\operatorname{cl} S$*;*

(ii) *f is continuously differentiable on S;*

(iii) *for any $x \in \operatorname{cl} S$ the level sets $\{y \in S \mid D_f(x, y) \le \alpha\}$ are bounded;*

(iv) *if $y^k \in S$ and $\lim_{k\to\infty} y^k = y^*$, then $\lim_{k\to\infty} D_f(y^*, y^k) = 0$;*

(v) *if $x^k \in \operatorname{cl} S$ and $y^k \in S$, with $\{x^k\}$ bounded, $\lim_{k\to\infty} y^k = y^*$, and $\lim_{k\to\infty} D_f(x^k, y^k) = 0$, then $\lim_{k\to\infty} x^k = y^*$.*

REMARK A.2. (i) *It can be shown that, if the Bregman function f is separable, then the condition that $\{x^k\}$ be bounded in Definition* A.1(v) *is redundant.*

(ii) *As noted by Bauschke and Borwein* [3], *conditions* (i)–(v) *of Definition* A.1 *imply that for any $y \in S$ the level sets $\{x \in \operatorname{cl} S \mid D_f(x, y) \le \alpha\}$ are also bounded.*

(iii) *Solodov and Svaiter* [31] *showed recently that condition* (v) *of Definition* A.1 *is redundant (i.e., it follows from the remaining conditions).*

Let Ω be a closed convex set in R^n and $z \in S$ a given point. The *Bregman projection* of z onto Ω is the point $P_\Omega^f(z) \in \Omega$ which minimizes $D_f(x, z)$ over all $x \in \Omega \cap \operatorname{cl} S$. Bregman projections exist and are unique, provided that the set Ω is closed and convex and that $\Omega \cap \operatorname{cl} S$ is nonempty (see, e.g., [13, Lemma 2.1.2]). Furthermore, we assume that $P_\Omega^f(z) \in S$ whenever $z \in S$. (This is commonly called *zone consistency.*) The useful inequality expressed in the next theorem then holds; see, e.g., [13, Theorem 2.4.1].

THEOREM A.3. *Let f be a Bregman function with zone S, and let $\Omega \subseteq R^n$ be a closed convex set such that $\Omega \cap \operatorname{cl} S \neq \emptyset$. Assume that f is zone consistent with respect to Ω, and let $z \in \Omega \cap \operatorname{cl} S$ be given. Then for any $y \in S$ the inequality*

$$D_f(z, y) - D_f(z, P_\Omega^f(y)) \ge D_f(P_\Omega^f(y), y) \tag{A.1}$$

holds.

The next result is a characterization of Bregman projections onto convex sets, given in [13, Theorem 2.4.2].

THEOREM A.4. *Under the assumptions of Theorem A.3, for any $y \in S$ the point $P_\Omega^f(y)$ is the Bregman projection of y onto Ω with respect to f if and only if*

$$\langle \nabla f(y) - \nabla f(P_\Omega^f(y)),\ x - P_\Omega^f(y) \rangle \le 0 \qquad \text{for all } x \in \Omega \cap \operatorname{cl} S. \tag{A.2}$$

We make use in this paper of Legendre functions and some of their basic properties. Therefore, we give here a definition from Bauschke and Borwein [3, Definition 2.8]; see also Rockafellar [30, section 26].

DEFINITION A.5. *Suppose that f is a closed convex proper function on R^n. Then f is a Legendre function if it is both essentially smooth and essentially strictly convex, i.e., f satisfies the following properties:*

(i) $\operatorname{int}(\operatorname{dom} f) \neq \emptyset$;

(ii) *f is differentiable on* $\operatorname{int}(\operatorname{dom} f)$;

(iii) *for every $x \in bd(\operatorname{dom} f)$ and every $y \in \operatorname{int}(\operatorname{dom} f)$*

$$\lim_{t \to 0^+} \langle \nabla f(x + t(y - x)), y - x \rangle = -\infty; \tag{A.3}$$

(iv) *f is strictly convex on* $\operatorname{int}(\operatorname{dom} f)$.

The next result, characterizing and describing Legendre functions, is quoted from [3, Fact 2.9] and based on [30, Theorem 26.5].

THEOREM A.6. *A convex function f is a Legendre function if and only if its conjugate f^* is. In this case, the gradient mapping*

$$\nabla f : \operatorname{int}(\operatorname{dom} f) \to \operatorname{int}(\operatorname{dom} f^*) \tag{A.4}$$

is a topological isomorphism with inverse mapping $(\nabla f)^{-1} = \nabla f^*$.

Finally, we quote from Bauschke and Borwein [3, Theorem 3.14] the following important fact.

THEOREM A.7. *If f is a Legendre function and $S = \operatorname{int}(\operatorname{dom} f)$, then f is zone consistent with respect to any closed convex set Ω such that $\Omega \cap clS \neq \emptyset$.*

Acknowledgments. Initial work on this paper was done in collaboration with Charles Byrne from the Department of Mathematical Sciences at the University of Massachusetts, Lowell, and we gratefully acknowledge that his contributions greatly helped in formulating the material that appears in this final version. We thank Bruno Motta de Carvalho for his help with creating the images. Part of the work of Y. Censor was done during a visit to the Department of Mathematics of the University of Linköping, Linköping, Sweden; the support and hospitality of Tommy Elfving and of Åke Björck, head of the Numerical Analysis Group there, are gratefully acknowledged. We also thank the referees for their constructive and helpful comments on earlier versions of this paper.

REFERENCES

[1] R. AHARONI AND Y. CENSOR, *Block-iterative projection methods for parallel computation of solutions to convex feasibility problems,* Linear Algebra Appl., 120 (1989), pp. 165–175.

[2] H.H. BAUSCHKE AND J.M. BORWEIN, *On projection algorithms for solving convex feasibility problems,* SIAM Rev., 38 (1996), pp. 367–426.

[3] H.H. BAUSCHKE AND J.M. BORWEIN, *Legendre functions and the method of random Bregman projections,* J. Convex Anal., 4 (1997), pp. 27–67.

[4] H.H. BAUSCHKE, F. DEUTSCH, H. HUNDAL, AND S.-H. PARK, *Fejér monotonicity and weak convergence of an accelerated method of projections,* in Constructive, Experimental, and Nonlinear Analysis, M. Thera, ed., AMS, Providence, RI, 2000, pp. 1–6.

[5] L.M. BREGMAN, *The method of successive projections for finding a common point of convex sets,* Soviet Math. Dokl., 6 (1965), pp. 688–692.

[6] L.M. Bregman, *The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming,* Comput. Math. Math. Phys., 7 (1967), pp. 200–217.
[7] Y. Censor, M.D. Altschuler, and W.D. Powlis, *On the use of Cimmino's simultaneous projections method for computing a solution of the inverse problem in radiation therapy treatment planning,* Inverse Problems, 4 (1988), pp. 607–623.
[8] Y. Censor and T. Elfving, *A multiprojections algorithm using Bregman projections in a product space,* Numer. Algorithms, 8 (1994), pp. 221–239.
[9] Y. Censor, P.P.B. Eggermont, and D. Gordon, *Strong underrelaxation in Kaczmarz's method for inconsistent systems,* Numer. Math., 41 (1983), pp. 83–92.
[10] Y. Censor and A. Lent, *An iterative row-action method for interval convex programming,* J. Optim. Theory Appl., 34 (1981), pp. 321–353.
[11] Y. Censor and S. Reich, *Iterations of paracontractions and firmly nonexpansive operators with applications to feasibility and optimization,* Optimization, 37 (1996), pp. 323–339.
[12] Y. Censor and J. Segman, *On block-iterative entropy maximization,* J. Inform. Optim. Sci., 8 (1987), pp. 275–291.
[13] Y. Censor and S.A. Zenios, *Parallel Optimization: Theory, Algorithms, and Applications*, Oxford University Press, New York, 1997.
[14] M. Collins, R.E. Shapire, and Y. Singer, *Logistic regression, AdaBoost and Bregman distances,* Machine Learning, 48 (2002), pp. 253–285.
[15] P.L. Combettes, *The foundations of set-theoretic estimation,* Proc. IEEE, 81 (1993), pp. 182–208.
[16] P.L. Combettes, *The convex feasibility problem in image recovery,* Adv. in Imaging and Electron Phys., 95 (1996), pp. 155–270.
[17] A.R. De Pierro and A.N. Iusem, *A relaxed version of Bregman's method for convex programming,* J. Optim. Theory Appl., 51 (1986), pp. 421–440.
[18] P.P.B. Eggermont, G.T. Herman, and A. Lent, *Iterative algorithms for large partitioned linear systems, with applications to image reconstruction,* Linear Algebra Appl., 40 (1981), pp. 37–67.
[19] S.-C. Fang, R.J. Rajasekera, and H.-S.J. Tsao, *Entropy Optimization and Mathematical Programming*, Kluwer Academic Publishers, Boston, MA, 1997.
[20] L. Gubin, B. Polyak, and E. Raik, *The method of projections for finding the common point of convex sets,* Comput. Math. Math. Phys., 7 (1967), pp. 1–24.
[21] G.T. Herman, *Image Reconstruction from Projections: The Fundamentals of Computerized Tomography*, Academic Press, New York, 1980.
[22] G.T. Herman, A. Lent, and P.H. Lutz, *Relaxation methods for image reconstruction,* Comm. ACM, 21 (1978), pp. 152–158.
[23] G.T. Herman, S. Matej, and B.M. Carvalho, *Algebraic reconstruction techniques using smooth basis functions for helical cone-beam tomography,* in Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications, D. Butnariu, Y. Censor, and S. Reich, eds., Elsevier Science Publishers, Amsterdam, The Netherlands, 2001, pp. 307–324.
[24] K. Kiwiel, *Generalized Bregman projections in convex feasibility problems,* J. Optim. Theory Appl., 96 (1998), pp. 139–157.
[25] K. Kiwiel, *Free-steering relaxation methods for problems with strictly convex costs and linear constraints,* Math. Oper. Res., 22 (1997), pp. 326–349.
[26] D.G. Luenberger, *Optimization by Vector Space Methods*, John Wiley & Sons, New York, 1969.
[27] R. Marabini, G.T. Herman, and J.M. Carazo, *3D reconstruction in electron microscopy using ART with smooth spherically symmetric volume elements (blobs),* Ultramicroscopy, 72 (1998), pp. 53–65.
[28] L.D. Marks, W. Sinkler, and E. Landree, *A feasible set approach to the crystallographic phase problem,* Acta Cryst. Sect. A, 55 (1999), pp. 601–612.
[29] T. Obi, S. Matej, R.M. Lewitt, and G.T. Herman, *2.5D simultaneous multislice reconstruction by series expansion methods from Fourier rebinned PET data,* IEEE Trans. Medical Imaging, 19 (2000), pp. 474–484.
[30] R.T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[31] M.V. Solodov and B.F. Svaiter, *An inexact hybrid generalized proximal point algorithm and some new results on the theory of Bregman functions,* Math. Oper. Res., 25 (2000), pp. 214–230.
[32] H. Stark and Y. Yang, *Vector Space Projections: A Numerical Approach to Signal and Image Processing, Neural Nets, and Optics,* John Wiley & Sons, New York, 1998.
[33] D.C. Youla, *Mathematical theory of image restoration by the method of convex projections,* in Image Recovery: Theory and Applications, H. Stark, ed., Academic Press, Orlando, FL, 1987, pp. 29–77.

SIAM J. OPTIM.
Vol. 13, No. 1, pp. 298–308

ON THE DISTANCE BETWEEN TWO ELLIPSOIDS*

ANHUA LIN† AND SHIH-PING HAN†

Abstract. We study the fundamental geometric problem of finding the distance between two ellipsoids. An algorithm is proposed for computing the distance and locating the two closest points. The algorithm is based on a local approximation of the two ellipsoids by balls. It is simple, geometric in nature, and has excellent convergence properties.

Key words. distance, ellipsoid, convex, optimization, algorithm

AMS subject classifications. 49M37, 65K05, 90C25, 90C30

PII. S1052623401396510

1. Introduction. In this paper we are concerned with the geometric problem of finding the distance between two ellipsoids. The problem is a fundamental optimization problem which can be expressed in the following form:

$$\begin{aligned} &\min && \|x-y\| \\ &\text{subject to} && x \in E_1, \\ &&& y \in E_2, \end{aligned} \tag{1.1}$$

where $E_1 := \{x : q_1(x) \le 0\}$ and $E_2 : \{y : q_2(y) \le 0\}$ are two given ellipsoids determined by the two quadratic functions

$$q_1(x) := \frac{1}{2}x^T A_1 x + b_1^T x + \alpha_1 \quad \text{and} \quad q_2(y) := \frac{1}{2}y^T A_2 y + b_2^T y + \alpha_2,$$

with positive definite symmetric matrices A_1 and A_2, vectors b_1 and b_2, and scalars α_1 and α_2. The norm $\|\cdot\|$ is the Euclidean norm. Though simple, the problem is nonlinear in nature and cannot be effectively solved by any linear technique such as a linear programming or quadratic programming method. On the other hand, it is highly structured, and any general nonlinear programming method that fails to exploit its structure may also not be very efficient. As an attempt to overcome such difficulties, we present in this paper a special algorithm for solving problem (1.1). The algorithm, which is an extension of our algorithm for finding the projection of a point on an ellipsoid [4], is extremely simple, easy to implement, and has excellent convergence properties.

In section 2 we describe our algorithm. In section 3, to facilitate our analysis of the algorithm, we characterize the optimal solution in terms of angles between some vectors. The convergence analysis of the algorithm is given in section 4, and some comments about the implementation and computational experiment are given in section 5. A proof for a key lemma is given in the appendix.

2. The algorithm. To describe the algorithm, we first introduce some notation. As usual, the angle between two nonzero vectors x and y is defined to be

$$\theta(x,y) := \arccos\left(\frac{x^T y}{\|x\|\|y\|}\right), \qquad 0 \le \theta(x,y) \le \pi,$$

*Received by the editors October 16, 2001; accepted for publication (in revised form) February 27, 2002; published electronically August 28, 2002.
http://www.siam.org/journals/siopt/13-1/39651.html
†Department of Mathematical Sciences, The Johns Hopkins University, Baltimore, MD 21218 (alin@mts.jhu.edu, han@mts.jhu.edu).

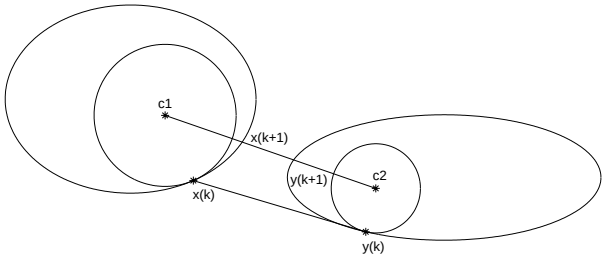


FIG. 1.

and a ball with center y and radius β is defined by

$$B(y;\beta) := \{x : \|y - x\| \le \beta\}.$$

Also, we use $d(E_1, E_2)$ to denote the distance between E_1 and E_2.

In the algorithm we will generate two sequences of points $\{x_k\}$ and $\{y_k\}$ on the boundaries of the two ellipsoids $\Omega(E_1)$ and $\Omega(E_2)$, respectively, and it will be shown that $\lim_{k\to\infty} \|x_k - y_k\| = d(E_1, E_2)$.

The algorithm is an iterative process. To avoid some cumbersome superscripts or subscripts, we use the undaunted symbols such as x and y to denote our current computed vectors, and the barred symbols $\bar{x}$ and $\bar{y}$ to denote the new vectors. More specifically, we use x, y for x^k, y^k and use $\bar{x}$, $\bar{y}$ for x^{k+1}, y^{k+1}, etc.

The algorithm can now be described geometrically as follows. At the kth iteration, having two points $x \in \Omega(E_1)$ and $y \in \Omega(E_2)$, we construct a ball $B(c_1; r_1)$ completely inside the ellipsoid E_1 and tangent to E_1 at x, and a ball $B(c_2; r_2)$ completely inside the ellipsoid E_2 and tangent to E_2 at y (see Figure 1). Then we check whether the line segment $[c_1, c_2]$ between the two centers is entirely contained in $E_1 \cup E_2$. If it is, then the two ellipsoids have a nonempty intersection and the distance $d(E_1, E_2) = 0$; otherwise, we continue and compute the new point $\bar{x}$ as the intersection of the line segment $[c_1, c_2]$ with the boundary $\Omega(E_1)$, and also $\bar{y}$ as the intersection of $[c_1, c_2]$ with the boundary $\Omega(E_2)$.

Two issues need to be addressed to make the algorithm viable. First, can the two balls $B(c_1, r_1)$ and $B(c_2, r_2)$ be easily constructed? Second, how can we check $[c_1, c_2] \subset E_1 \cup E_2$ and compute new estimates $\bar{x}$ and $\bar{y}$? The first issue can be resolved by the following lemma. (A slightly different but equivalent version of the lemma is given in [4]. However, because the proof is short and because we want the paper to be self-contained, we include its proof here.)

LEMMA 2.1. *Let $E := \{w : \frac{1}{2}w^T A w + b^T w + \alpha \le 0\}$ be a nonempty ellipsoid determined by a positive definite symmetric matrix A, a vector b, and a scalar α. Let z be a point on the boundary $\Omega(E)$ of E; then for any $0 < \gamma \le \frac{1}{\rho(A)}$*

$$B(z - \gamma(Az + b); \gamma\|Az + b\|) \subset E,$$

where $\rho(A)$ is the spectral radius of A.

Proof. Let $q(x) := \frac{1}{2}x^T A x + b^T x + \alpha$, and let y be any vector on the boundary of the ball $B(z - \gamma(Az + b); \gamma\|Az + b\|)$. It suffices to show $q(y) \le 0$. We have

$$\|y - z + \gamma(Az + b)\|^2 = \gamma^2 \|Az + b\|^2.$$

By expanding both sides of the above equality, we get

$$\|y-z\|^2 + 2\gamma(y-z)^T(Az+b) = 0.$$

Equivalently, we have

$$\nabla q(z)^T(y-z) = -\frac{1}{2\gamma}\|y-z\|^2.$$

Now, we can show $q(y) \le 0$ when $0 < \gamma \le \frac{1}{\rho(A)}$ by

$$\begin{aligned} q(y) &= q(z) + \nabla q(z)^T(y-z) + \frac{1}{2}(y-z)^T A(y-z) \\ &\le q(z) - \frac{1}{2\gamma}\|y-z\|^2 + \frac{\rho(A)}{2}\|y-z\|^2 \\ &= \frac{1}{2}\left(\rho(A) - \frac{1}{\gamma}\right)\|y-z\|^2 \\ &\le 0. \qquad \square \end{aligned}$$

From the above lemma we can construct the two balls in the ellipsoids by choosing the centers c_1 and c_2 as $x - \gamma_1(A_1x + b_1)$ and $y - \gamma_2(A_2y + b_2)$ and choosing the radii r_1 and r_2 as $\gamma_1\|A_1x + b_1\|$ and $\gamma_2\|A_2y + b_2\|$, respectively, with γ_1 and γ_2 satisfying

$$0 < \tau \le \gamma_1 \le \frac{1}{\rho(A_1)} \quad \text{and} \quad 0 < \tau \le \gamma_2 \le \frac{1}{\rho(A_2)}. \tag{2.1}$$

Here τ is a prescribed small fixed number. Such a lower bound is needed because, for the method to work properly, it is also required that the radii of the balls be bounded away from zero. Obviously, we can choose a matrix norm $\|\cdot\|$ and set γ_1 and γ_2 to be

$$\gamma_1 = \frac{1}{\|A_1\|}, \qquad \gamma_2 = \frac{1}{\|A_2\|}.$$

For such a choice, the γ's are independent of the iteration, and the lower bound τ can be

$$\tau = \min\left\{\frac{1}{\|A_1\|}, \frac{1}{\|A_2\|}\right\}.$$

Of course, the 1-norm and the ∞-norm are particularly useful here because of their ease of computation.

As for the second issue of checking the condition $[c_1, c_2] \subset E_1 \cup E_2$ and the computation of new estimates $\bar{x}$ and $\bar{y}$, we compute two stepsizes t_1 and t_2 by

$$\begin{aligned} t_1 &= \max\{t \in [0,1] : (1-t)c_1 + tc_2 \in E_1\}, \\ t_2 &= \min\{t \in [0,1] : (1-t)c_1 + tc_2 \in E_2\}. \end{aligned} \tag{2.2}$$

Because $[c_1, c_2] \cap E_1$ and $[c_1, c_2] \cap E_2$ are both nonempty closed sets, t_1 and t_2 are well defined and can be easily computed by solving two one-dimensional quadratic equations.

If $t_2 \le t_1$, then E_1 and E_2 have a nonempty intersection. In this case $d(E_1, E_2) = 0$ and we are done. If $t_2 > t_1$, then we let $\bar{x} = c_1 + t_1(c_2 - c_1)$ and $\bar{y} = c_1 + t_2(c_2 - c_1)$.

In this case, we have $\bar{x} \in \Omega(E_1)$, $\bar{y} \in \Omega(E_2)$, and the open line segment $(\bar{x}, \bar{y})$ has no intersection with the union $E_1 \cup E_2$.

Now we summarize the above discussion and give a precise description of the algorithm below.

ALGORITHM 1.

Initiation. Start from an interior point c_1 in E_1 and an interior point c_2 in E_2. The natural choices for these two points are the centers $-A_1^{-1}b_1$ and $-A_2^{-1}b_2$ of the two ellipsoids, respectively.

General steps. At the kth iteration, having an interior point c_1 of E_1 and an interior point c_2 of E_2, we proceed as follows:

1. We solve two one-dimensional quadratic equations to get the stepsizes t_1 and t_2 as given in (2.2).

2. If $t_2 \le t_1$, we are done, and we set $d(E_1, E_2) = 0$. In this case, any point $c_1 + t(c_2 - c_1)$ with $t_2 \le t \le t_1$ is in $E_1 \cap E_2$. Otherwise, we compute new points $\bar{x}$ and $\bar{y}$ by

$$\bar{x} = c_1 + t_1(c_2 - c_1), \qquad \bar{y} = c_1 + t_2(c_2 - c_1).$$

3. We compute θ_1 and θ_2 by

$$\theta_1 = \theta(\bar{y} - \bar{x}, A_1\bar{x} + b_1), \qquad \theta_2 = \theta(\bar{x} - \bar{y}, A_2\bar{y} + b_2).$$

If $\theta_1 = \theta_2 = 0$, then terminate.

4. We compute the new centers $\bar{c}_1$ and $\bar{c}_2$ by

$$\bar{c}_1 = \bar{x} - \gamma_1(A_1\bar{x} + b_1), \qquad \bar{c}_2 = \bar{y} - \gamma_2(A_2\bar{y} + b_2),$$

with γ_1 and γ_2 satisfying (2.1).

We note here that the algorithm will terminate in a finite number of iterations when any of the two situations occurs: (1) $t_2 \le t_1$ or (2) $t_2 > t_1$ but $\theta_1 = \theta_2 = 0$. When $t_2 \le t_1$, as mentioned before, an intersection point of E_1 and E_2 is found and we stop. When $t_2 > t_1$ and $\theta_1 = \theta_2 = 0$, the new points $\bar{x}$ and $\bar{y}$ are distinct, and it will be shown in Corollary 3.6 that, in this case, the two sets E_1 and E_2 are disjoint and the pair $(\bar{x}, \bar{y})$ is the unique optimal solution of problem (1.1). We will also justify the usage of the angle values θ_1 and θ_2 for determining convergence in Theorem 3.5. Of course, in practice we check $\theta_1 \le \epsilon$ and $\theta_2 \le \epsilon$ instead of $\theta_1 = \theta_2 = 0$.

3. Optimality conditions via angles. To study the convergence properties of the algorithm, we need to characterize the optimal solution of problem (1.1). Recall that a pair (x^*, y^*) is a Karush–Kuhn–Tucker point of problem (1.1) if there exist Lagrange multipliers λ and μ such that

$$\begin{cases} x^* - y^* + \lambda(A_1x^* + b_1) & = & 0, \\ y^* - x^* + \mu(A_2y^* + b_2) & = & 0, \\ \lambda(\frac{1}{2}x^{*T}A_1x^* + b_1^Tx^* + \alpha_1) & = & 0, \\ \mu(\frac{1}{2}y^{*T}A_2y^* + b_2^Ty^* + \alpha_2) & = & 0, \\ \frac{1}{2}x^{*T}A_1x^* + b_1^Tx^* + \alpha_1 & \le & 0, \\ \frac{1}{2}y^{*T}A_2y^* + b_2^Ty^* + \alpha_2 & \le & 0, \\ \lambda \ge 0 \quad \text{and} \quad \mu \ge 0. \end{cases}$$

We assume that the matrices A_1 and A_2 are positive definite and that both ellipsoids E_1 and E_2 are nonempty. Under these assumptions, the problem is a well-defined separable convex optimization problem. Therefore, a pair (x^*, y^*) is an optimal solution of problem (1.1) if and only if (x^*, y^*) is also a Karush–Kuhn–Tucker point [1, 5]. When $E_1 \cap E_2 \neq \emptyset$, then for any z in $E_1 \cap E_2$ the pair (z, z) is a Karush–Kuhn–Tucker point with both multipliers $\lambda = 0$ and $\mu = 0$. In this section, we are more concerned with the case in which $E_1 \cap E_2 = \emptyset$. In this case, the optimal value $d(E_1, E_2) > 0$, and the optimal solution (x^*, y^*) is unique.

For our convergence analysis of the algorithm, we need to consider an equivalent optimality condition in terms of angles. Since $E_1 \cap E_2 = \emptyset$, we must have $x^* \neq y^*$. Hence it follows from the first two Karush–Kuhn–Tucker equations that λ and μ must be strictly positive. This implies that $x^* \in \Omega(E_1)$ and $y^* \in \Omega(E_2)$. Moreover, it again follows from the first two equations that the three vectors $y^* - x^*$, $A_1x^* + b_1$, and $-(A_2y^* + b_2)$ are all in the same direction. Therefore, under the assumption that $E_1 \cap E_2 = \emptyset$, the optimal solution (x^*, y^*) also satisfies the following angle conditions:

$$\begin{cases} x^* \in \Omega(E_1) \quad \text{and} \quad y^* \in \Omega(E_2), \\ \theta(y^* - x^*, A_1x^* + b_1) = 0, \\ \theta(x^* - y^*, A_2y^* + b_2) = 0. \end{cases} \tag{3.1}$$

We note here that the above angle condition is also sufficient for optimality. Actually, we need to establish the stronger and more useful result that, if two points $x \in \Omega(E_1)$ and $y \in \Omega(E_2)$ have small angles $\theta(y - x, A_1x + b_1)$ and $\theta(x - y, A_2y + b_2)$, the pair (x, y) is close to the optimal solution (x^*, y^*). To show this we need two simple lemmas.

LEMMA 3.1. *If u and v are two nonzero vectors in R^n with $\|u\| = \|v\|$ and $\theta = \theta(u, v)$, then*

$$\|u - v\| = 2\|u\| \sin \frac{1}{2}\theta.$$

Proof. The result follows immediately from the following equations:

$$\begin{aligned} \|u - v\|^2 &= \|u\|^2 - 2u^T v + \|v\|^2 \\ &= \|u\|^2 - 2\|u\|\|v\| \cos\theta + \|v\|^2 \\ &= (\|u\| - \|v\|)^2 + 2\|u\|\|v\|(1 - \cos\theta) \\ &= 4\|u\|^2 \sin^2 \frac{\theta}{2}. \end{aligned}$$ □

LEMMA 3.2. *Let z, p, and h be vectors in R^n with $p \neq 0$, $h \neq 0$, and let $\theta = \theta(p, h)$ and ϕ be a scalar. If $z^T p \leq \phi$, then*

$$z^T h \leq 2\|z\|\|h\| \sin \frac{\theta}{2} + \frac{\|h\|}{\|p\|}\phi.$$

Proof. It follows from the previous lemma that

$$\begin{aligned} z^T h &= z^T \left(h - \frac{\|h\|}{\|p\|} p \right) + \frac{\|h\|}{\|p\|} z^T p \\ &\leq \|z\| \left\| h - \frac{\|h\|}{\|p\|} p \right\| + \frac{\|h\|}{\|p\|} \phi \\ &= 2\|z\|\|h\| \sin \frac{\theta}{2} + \frac{\|h\|}{\|p\|}\phi. \end{aligned}$$ □

To facilitate our presentation of our next key theorem, we introduce some notation. Let κ_1 and κ_2 be the smallest eigenvalues of A_1 and A_2, respectively. Then, for any $z \in R^n$,

$$\kappa_1||z||^2 \le z^T A_1 z \quad \text{and} \quad \kappa_2||z||^2 \le z^T A_2 z.$$

We also define η_1 and η_2 as the following bounds:

$$\eta_1 := \max_{z\in\Omega(E_1)} \|A_1 z + b_1\| \quad \text{and} \quad \eta_2 := \max_{z\in\Omega(E_2)} \|A_2 z + b_2\|.$$

Notice that η_i is zero if and only if the corresponding ellipsoid is a singleton. We will assume that both ellipsoids contain interior points and thus that both η's are positive. Now we give the key theorem below.

THEOREM 3.3. *Let both E_1 and E_2 have interior points. If $x \in \Omega(E_1)$, $y \in \Omega(E_2)$, and $x \neq y$, then for any $u \in E_1$ and $v \in E_2$*

$$\|y-x\| + \frac{\kappa_1}{2\eta_1}\|u-x\|^2 + \frac{\kappa_2}{2\eta_2}\|v-y\|^2 \le \|u-v\| + 2\|u-x\|\sin\frac{\theta_1}{2} + 2\|v-y\|\sin\frac{\theta_2}{2},$$

where $\theta_1 = \theta(y-x, A_1x+b_1)$ and $\theta_2 = \theta(x-y, A_2y+b_2)$.

Proof. As before, we let $q_1(w) := \frac{1}{2}w^T A_1 w + b_1^T w + \alpha_1$. Because $x \in \Omega(E_1)$ and $u \in E_1$, we have

$$\begin{aligned} 0 &\ge q_1(u) - q_1(x) \\ &= (u-x)^T \nabla q_1(x) + \frac{1}{2}(u-x)^T A_1 (u-x) \\ &\ge (u-x)^T (A_1 x + b_1) + \frac{\kappa_1}{2}\|u-x\|^2. \end{aligned}$$

Therefore, we have

$$(u-x)^T(A_1x+b_1) \le -\frac{\kappa_1}{2}\|u-x\|^2.$$

We apply Lemma 3.2 to the above inequality with $\phi = -\frac{\kappa_1}{2}\|u-x\|^2$, $z = (u-x)$, $p = A_1x + b_1$, and $h = y - x$ to get

$$(u-x)^T(y-x) \le 2\|u-x\|\|y-x\|\sin\frac{\theta_1}{2} - \frac{\kappa_1\|y-x\|}{2\eta_1}\|u-x\|^2.$$

Similarly, for the case of E_2, we interchange the roles of x and y and replace u by v to get

$$(v-y)^T(x-y) \le 2\|v-y\|\|x-y\|\sin\frac{\theta_2}{2} - \frac{\kappa_2\|y-x\|}{2\eta_2}\|v-y\|^2.$$

Adding the above two inequalities and rearranging terms, we have

$$\begin{aligned} &\|y-x\|^2 + \frac{\kappa_1\|y-x\|}{2\eta_1}\|u-x\|^2 + \frac{\kappa_2\|y-x\|}{2\eta_2}\|v-y\|^2 \\ &\quad\le (v-u)^T(y-x) + 2\|u-x\|\|y-x\|\sin\frac{\theta_1}{2} + 2\|v-y\|\|x-y\|\sin\frac{\theta_2}{2} \\ &\quad\le \|u-v\|\|y-x\| + 2\|u-x\|\|y-x\|\sin\frac{\theta_1}{2} + 2\|v-y\|\|x-y\|\sin\frac{\theta_2}{2}. \end{aligned}$$

The desired result follows immediately when we divide the above inequality by $\|y - x\|$. ∎

Some useful results follow from the above theorem. First, we show that the angle condition (3.1) is a necessary and sufficient condition for optimality.

THEOREM 3.4. *Let E_1 and E_2 have nonempty interiors and $E_1 \cap E_2 = \emptyset$. The pair (x^*, y^*) satisfies the angle condition* (3.1) *if and only if (x^*, y^*) is the optimal solution of problem* (1.1).

Proof. The necessary part has already been given in the paragraph prior to the angle condition (3.1). Now we show the sufficient part. We apply the above theorem to (x^*, y^*) and use the assumption that $\theta_1 = \theta_2 = 0$ to get the result that for any $u \in E_1$ and $v \in E_2$, $\|x^* - y^*\| \le \|u - v\|$. Therefore the pair (x^*, y^*) is optimal. ∎

For our convergence analysis the following result is useful.

THEOREM 3.5. *Let E_1 and E_2 have nonempty interiors and $E_1 \cap E_2 = \emptyset$, and let (x^*, y^*) be the optimal solution of problem* (1.1). *Then, for any $x \in \Omega(E_1)$ and $y \in \Omega(E_2)$,*

$$\frac{\kappa_1}{\eta_1}\|x - x^*\|^2 + \frac{\kappa_2}{\eta_2}\|y - y^*\|^2 \le 4\sigma\left(\sin\frac{\theta_1}{2} + \sin\frac{\theta_2}{2}\right),$$

where $\theta_1 = \theta(y - x, A_1x + b_1)$, $\theta_2 = \theta(x - y, A_2y + b_2)$, and

$$\sigma := \max\{\|u - v\| : u \in E_1 \quad \textit{and} \quad v \in E_2\}. \tag{3.2}$$

Proof. This follows from Theorem 3.3 by letting $u = x^*$ and $v = y^*$ and using the fact that $\|x - y\| \ge \|x^* - y^*\|$. ∎

The following corollary gives a justification for using $\theta_1 = \theta_2 = 0$ as the stopping criterion of the algorithm.

COROLLARY 3.6. *Let E_1 and E_2 have nonempty interiors. If $x \in \Omega(E_1)$ and $y \in \Omega(E_2)$ such that $x \ne y$ and $\theta(y - x, A_1x + b_1) = \theta(x - y, A_2y + b_2) = 0$, then $E_1 \cap E_2 = \emptyset$ and the pair (x, y) is the unique optimal solution of problem* (1.1).

Proof. Suppose that $E_1 \cap E_2 \ne \emptyset$; then let $z \in E_1 \cap E_2$. We apply Theorem 3.3 to the pair (x, y) with $u = v = z$. It follows from the inequality of Theorem 3.3 that at least one of θ_1 and θ_2 is nonzero, which contradicts our assumption that $\theta_1 = \theta_2 = 0$. Therefore, we have $E_1 \cap E_2 = \emptyset$, and problem (1.1) has a unique optimal solution. Then the optimality of the pair (x, y) follows immediately from Theorem 3.5. ∎

4. Convergence analysis. The distance between any two balls $B(c_1, r_1)$ and $B(c_2, r_2)$ is easy to find. Indeed, if we assume that $r_1 > 0$, $r_2 > 0$ and the two balls are disjoint, then the closest two points on the boundaries $\Omega(B(c_1, r_1))$ and $\Omega(B(c_2, r_2))$ are given respectively by

$$\hat{x} = c_1 + \frac{r_1}{\|c_1 - c_2\|}(c_2 - c_1) \quad \text{and} \quad \hat{y} = c_2 + \frac{r_2}{\|c_1 - c_2\|}(c_1 - c_2). \tag{4.1}$$

Our algorithm can be viewed as if the two ellipsoids were being iteratively approximated locally by balls. Therefore, for our convergence analysis, we first analyze how the distance between any two points on the boundaries of the two balls, say $x \in \Omega(B(c_1, r_1))$ and $y \in \Omega(B(c_2, r_2))$, is related to the shortest distance $\|\hat{x} - \hat{y}\|$ in terms of the angles $\theta(y - x, x - c_1)$ and $\theta(x - y, y - c_2)$. This result is contained in the following fundamental lemma, which is not only useful for our analysis but also geometrically interesting in its own right. But the proof is long and provides little insight into the algorithm; hence we include the proof in the appendix.

LEMMA 4.1. *Let $B(c_1, r_1)$ and $B(c_2, r_2)$ be two disjoint balls with $r_1 > 0$ and $r_2 > 0$, and let $\hat{x}$ and $\hat{y}$ be defined as in* (4.1). *For any two points $x \in \Omega(B(c_1, r_1))$ and $y \in \Omega(B(c_2, r_2))$,*

$$\|x - y\| - \|\hat{x} - \hat{y}\| \geq \frac{4}{\eta}\left(r_1 d \sin^2 \frac{\theta_1}{2} + r_2 d \sin^2 \frac{\theta_2}{2} + r_1 r_2 \sin^2 \frac{(\theta_1 - \theta_2)}{2}\right),$$

where $d = \|x - y\|$, $\theta_1 = \theta(y - x, x - c_1)$, $\theta_1 = \theta(x - y, y - c_2)$, and $\eta = r_1 + r_2 + d + \|c_1 - c_2\|$.

We now give our convergence theorem. It is noted here that if $E_1 \cap E_2 \neq \emptyset$, then any point in the intersection is considered a solution.

THEOREM 4.2. *If E_1 and E_2 have nonempty interiors, then either any sequence generated by the algorithm terminates at, or any accumulation point of the sequence is a solution of, problem* (1.1). *Furthermore, if $E_1 \cap E_2 = \emptyset$, then the sequence converges to the unique solution of problem* (1.1).

Proof. Suppose the algorithm terminates in a finite number of iterations. In this case, as mentioned in the paragraph following the description of the algorithm, either $t_1 \geq t_1$ or $t_1 < t_2$, but $\theta_1 = \theta_2 = 0$. If $t_1 \geq t_2$, then all the points $c_1 + t(c_2 - c_1)$ with $t \in [t_2, t_1]$ are in $E_1 \cap E_2$. If $t_1 < t_2$ but $\theta_1 = \theta_2 = 0$, then it follows from Corollary 3.6 that $E_1 \cap E_2 = \emptyset$ and the new pair $(\bar{x}, \bar{y})$ is the unique optimal solution of problem (1.1).

We now consider the case in which the generated sequence is infinite. In this case, we have $t_1 < t_2$ in each iteration, and the two balls $B(c_1, r_1)$ and $B(c_2, r_2)$ are mutually disjoint and entirely contained in the ellipsoids E_1 and E_2, respectively. Therefore, we have

$$\|c_1 - c_2\| \geq r_1 + r_2 + \|\bar{x} - \bar{y}\|.$$

On the other hand, by the triangle inequality, we also have

$$\begin{aligned}\|c_1 - c_2\| &= \|c_1 - x + y - y + x - c_2\| \\ &\leq \|c_1 - x\| + \|x - y\| + \|y - c_2\| \\ &\leq r_1 + r_2 + \|x - y\|.\end{aligned}$$

Then, from the two inequalities above, we have the monotonicity property:

$$\|x - y\| \geq \|\bar{x} - \bar{y}\|.$$

Therefore, the sequence of distances $\{\|x^k - y^k\|\}$ is monotone and hence converges, say to d^*. Consider the two cases $d^* = 0$ and $d^* \neq 0$.

For the case $d^* = 0$, let (x^*, y^*) be an accumulation point of $\{(x^k, y^k)\}$. Then there is a subsequence $\{(x^{k_m}, y^{k_m})\}$ converging to (x^*, y^*). Clearly $x^* = y^*$ because

$$\lim_{k_m \to \infty} \|x^{k_m} - y^{k_m}\| = d^* = 0.$$

The fact $x^* \in E_1 \cap E_2$ follows from $\{(x^{k_m})\} \subset E_1$, $\{(y^{k_m})\} \subset E_2$, and that E_1 and E_2 are two closed sets.

We now consider the case $d^* \neq 0$. Because the boundaries $\Omega(E_1)$ and $\Omega(E_2)$ are compact and because the stepsizes γ's are bounded away from zero, the radii r_1 and r_2 remain bounded below from zero throughout the computation. More specifically, there exists a positive number δ such that $r_1 \geq \delta$ and $r_2 \geq \delta$ in each iteration. Let

$\sigma = \max\{\|u - v\| : u \in E_1 \text{ and } v \in E_2\}$ be defined as in (3.2). Clearly, σ is an upper bound for all the generated quantities r_1, r_2, $\|c_1 - c_2\|$, and $\|x - y\|$. In our algorithm, we have $x \in \Omega(E_1)$ and $y \in \Omega(E_2)$. Then it follows from the previous lemma that

$$\begin{aligned}\|x - y\| - \|\bar{x} - \bar{y}\| &\geq \|x - y\| - \|\hat{x} - \hat{y}\| \\ &\geq \frac{1}{\sigma}\left(\delta d^* \sin^2 \frac{\theta_1}{2} + \delta d^* \sin^2 \frac{\theta_2}{2} + \delta^2 \sin^2 \frac{(\theta_1 - \theta_2)}{2}\right).\end{aligned}$$

From the convergence of $\{\|x^k - y^k\|\}$, we have that

$$\lim_{k\to\infty} (\|x^k - y^k\| - \|x^{k+1} - y^{k+1}\|) = 0.$$

This result, combined with the above inequality, implies that

$$\lim_{k\to\infty} \theta(y^k - x^k, A_1 x^k + b_1) = 0 \quad \text{and} \quad \lim_{k\to\infty} \theta(x^k - y^k, A_2 y^k + b_2) = 0.$$

Then it follows immediately from Theorem 3.5 that $\{(x^k, y^k)\}$ converges to the unique optimal solution of problem (1.1). □

5. Discussion. The algorithm proposed here is simple and easy to implement. The major work in each iteration is merely the solution of two one-dimensional quadratic equations and the computation of two angles. We implemented it in MATLAB and did some preliminary testing. We found the algorithm to be very reliable and to work very well generally. But the convergence may become a little slow when the ellipsoids are small, thin, and far apart. This is mainly because the balls generated inside the ellipsoids become too tiny to produce substantial improvement in each iteration. Though the requirement that the balls be completely inside the ellipsoids is sufficient for convergence, it is not necessary. It is desirable to design an acceleration technique that can avoid such restriction and allow more adequate improvements.

The algorithm is well suited for large sparse matrices, because the computation involves only the matrices themselves and does not need factorization. Of course, for this situation, we have to choose vectors other than the centers of the ellipsoids as the initial points.

According to our convergence theorem, the algorithm will work when both ellipsoids have interiors. Therefore, we can check this condition by evaluating the two quadratic functions $q_1(x)$ and $q_2(x)$ at the centers $A_1^{-1}b_1$ and $A_2^{-1}b_2$, respectively. We apply the algorithm only when both values are strictly negative. If one is negative and the other is zero, the problem is reduced to the projection of a point to an ellipsoid. For this case, an algorithm such as the one proposed in [2, 3, 4, 6, 7, 8, 9] should apply. Of course, when both values are nonnegative, this is just a trivial case and no further computation is needed.

Appendix. The proof of Lemma 4.1. Here we give a proof for our main lemma. For doing this, we define two parallel hyperplanes which pass through the centers c_1 and c_2 of the two balls $B(c_1; r_1)$ and $B(c_2; r_2)$, respectively, and have a common normal vector $w = (x - y)/\|x - y\|$:

$$H_1 = \{z : w^T z = w^T c_1\} \quad \text{and} \quad H_2 = \{z : w^T z = w^T c_2\}.$$

The orthogonal projectors onto these two hyperplanes H_1 and H_2 are

$$\begin{aligned} P_1(z) &= (w^T c_1)w + (I - ww^T)z, \\ P_2(z) &= (w^T c_2)w + (I - ww^T)z. \end{aligned}$$

We now give some useful lemmas.

LEMMA A.1. *Let $B(c_1, r_1)$ and $B(c_2, r_2)$ be disjoint and $x \in \Omega(B(c_1, r_1))$, $y \in \Omega(B(c_2, r_2))$, and $d = \|x - y\|$. Let $\theta_1 = \theta(y - x, x - c_1)$ and $\theta_2 = \theta(x - y, y - c_2)$. Then*

(1) $\|P_1(x) - c_1\| = r_1 \sin\theta_1$ *and* $\|P_2(y) - c_2\| = r_2 \sin\theta_2$;

(2) $P_1(x) - P_2(y) = ww^T(c_1 - c_2) = (r_1 \cos\theta_1 + r_2 \cos\theta_2 + d)w$.

Proof. We use the fact that the two vectors $ww^T(x - c_1)$ and $(I - ww^T)(x - c_1)$ are orthogonal to get

$$\begin{aligned} r_1^2 &= \|x - c_1\|^2 \\ &= \|ww^T(x - c_1) + (I - ww^T)(x - c_1)\|^2 \\ &= \|ww^T(x - c_1)\|^2 + \|(I - ww^T)(x - c_1)\|^2 \\ &= (\|x - c_1\| \cos\theta_1)^2 + \|ww^T c_1 + (I - ww^T)x - c_1\|^2 \\ &= (r_1 \cos\theta_1)^2 + \|P_1(x) - c_1\|^2. \end{aligned}$$

Therefore, it follows that $\|P_1(x) - c_1\| = (r_1^2 - r_1^2 \cos^2\theta_1)^{\frac{1}{2}} = r_1 \sin\theta_1$.

The result $\|P_2(y) - c_2\| = r_2 \sin\theta_2$ can be proven similarly. To prove (2), we notice that $(I - ww^T)(x - y) = 0$, and we have

$$\begin{aligned} P_1(x) - P_2(y) &= (w^T c_1)w + (I - ww^T)x - (w^T c_2)w - (I - ww^T)y \\ &= ww^T(c_1 - c_2) + (I - ww^T)(x - y) \\ &= ww^T(c_1 - c_2). \end{aligned}$$

Then the above equality, in turn, implies that

$$\begin{aligned} P_1(x) - P_2(y) &= ww^T((c_1 - x) + (x - y) + (y - c_2)) \\ &= w^T(c_1 - x)w + w^T(x - y)w + w^T(y - c_2)w \\ &= (r_1 \cos\theta_1 + \|x - y\| + r_2 \cos\theta_2)w. \qquad \square \end{aligned}$$

LEMMA A.2. *Let the assumptions of the previous lemma hold; then*

$$\|c_1 - c_2\|^2 \le (r_1 \cos\theta_1 + r_2 \cos\theta_2 + d)^2 + (r_1 \sin\theta_1 + r_2 \sin\theta_2)^2.$$

Proof. A direct calculation shows that the two vectors $P_1(x) - P_2(y)$ and $c_1 - P_1(x) + P_2(y) - c_2$ are orthogonal. Therefore, by the previous lemma and the triangle inequality,

$$\begin{aligned} \|c_1 - c_2\|^2 &= \|P_1(x) - P_2(y) + c_1 - P_1(x) + P_2(y) - c_2\|^2 \\ &= \|P_1(x) - P_2(y)\|^2 + \|c_1 - P_1(x) + P_2(y) - c_2\|^2 \\ &\le (r_1 \cos\theta_1 + r_2 \cos\theta_2 + d)^2 + (\|c_1 - P_1(x)\| + \|P_2(y) - c_2\|)^2 \\ &= (r_1 \cos\theta_1 + r_2 \cos\theta_2 + d)^2 + (r_1 \sin\theta_1 + r_2 \sin\theta_2)^2. \qquad \square \end{aligned}$$

We now give a proof of our main lemma.

Proof of Lemma 4.1. We use the fact that $\|c_1 - c_2\| = r_1 + r_2 + \|\hat{x} - \hat{y}\|$ to get

$$\begin{aligned} \|x - y\| - \|\hat{x} - \hat{y}\| &= (d + r_1 + r_2) - (r_1 + r_2 + \|\hat{x} - \hat{y}\|) \\ &= (d + r_1 + r_2) - \|c_1 - c_2\| \\ &= \frac{1}{\eta}((d + r_1 + r_2)^2 - \|c_1 - c_2\|^2). \end{aligned}$$

Using Lemma A.2 and by expansion and simplification, we get

$$\begin{aligned}(d+r_1+r_2)^2 - \|c_1-c_2\|^2 &\geq (d+r_1+r_2)^2 - (r_1\cos\theta_1 + r_2\cos\theta_2 + d)^2 - (r_1\sin\theta_1 + r_2\sin\theta_2)^2 \\ &= 4\left(r_1 d\sin^2\frac{\theta_1}{2} + r_2 d\sin^2\frac{\theta_2}{2} + r_1 r_2\sin^2\frac{(\theta_1-\theta_2)}{2}\right).\end{aligned}$$

Incorporating the above inequality into the previous equality, we then get the desired result. □

REFERENCES

[1] R. Fletcher, *Practical Methods of Optimization*, 2nd ed., Wiley, New York, 1991.

[2] B. S. He, *Solving trust region problems in large scale optimization*, J. Comput. Math., 18 (2000), pp. 1–12.

[3] L. T. Hoai An, *An efficient algorithm for globally minimizing a quadratic function under convex quadratic constraints*, Math. Programming, 87 (2000), pp. 401–426.

[4] A. Lin and S.-P. Han, *Projection on an Ellipsoid*, submitted, 2001.

[5] O. L. Mangasarian, *Nonlinear Programming*, Classics Appl. Math. 10, SIAM, Philadelphia, 1994.

[6] J. J. Moré, *The Levenberg-Marquardt algorithm: Implementation and theory*, in Numerical Analysis, Lecture Notes in Math. 630, G. A. Watson, ed., Springer-Verlag, Berlin, 1977, pp. 105–116.

[7] J. J. Moré and D. C. Sorenson, *Computing a trust region step*, SIAM J. Sci. Statist. Comput., 4 (1983), pp. 553–572.

[8] T. P. Dinh and L. T. Hoai An, *A D.C. optimization algorithm for solving the trust-region subproblem*, SIAM J. Optim., 8 (1998), pp. 476–505.

[9] D. Sorensen, *Minimization of a large-scale quadratic function subject to a spherical constraint*, SIAM J. Optim., 7 (1997), pp. 141–161.

SIAM J. OPTIM.
Vol. 13, No. 2, pp. 309–320

IMPROVED COMPLEXITY FOR MAXIMUM VOLUME INSCRIBED ELLIPSOIDS*

KURT M. ANSTREICHER†

Abstract. Let $\mathcal{P} = \{x \mid Ax \leq b\}$, where A is an $m \times n$ matrix. We assume that $\mathcal{P}$ contains a ball of radius one centered at the origin and is itself contained in a ball of radius R centered at the origin. We consider the problem of approximating the maximum volume ellipsoid inscribed in $\mathcal{P}$. Such ellipsoids have a number of interesting applications, including the inscribed ellipsoid method for convex optimization. We describe an optimization algorithm that obtains an ellipsoid whose volume is at least a factor $e^{-\epsilon}$ of the maximum possible in $O(m^{3.5}\ln(mR/\epsilon))$ operations. Our result provides an alternative to a saddlepoint-based approach with the same complexity, developed by Nemirovskii. We also show that a further reduction in complexity can be obtained by first computing an approximation of the analytic center of $\mathcal{P}$.

Key words. maximum volume inscribed ellipsoid, inscribed ellipsoid method

AMS subject classifications. 90C25, 90C22

PII. S1052623401390902

1. Introduction.

Let $\mathcal{P} = \{x \mid Ax \leq b\}$, where A is an $m \times n$ matrix. We assume that $\mathcal{P}$ is bounded with a nonempty interior. It is then known [6] that there is a unique ellipsoid $E^* \subset \mathcal{P}$ of maximum volume. We say that an ellipsoid $E \subset \mathcal{P}$ is γ-*maximal* if $\mathrm{Vol}(E) \geq \gamma\,\mathrm{Vol}(E^*)$, where $0 < \gamma < 1$ and $\mathrm{Vol}(\cdot)$ denotes n-dimensional volume. In this paper we consider the complexity of computing a γ-maximal inscribed ellipsoid for $\mathcal{P}$. For convenience in stating complexity results, we often write $\gamma = e^{-\epsilon}$ (as in [12, section 6.5]), where $\epsilon > 0$.

There are a number of interesting applications of γ-maximal ellipsoids. For example, the computation of a γ-maximal ellipsoid, with $\gamma > 0.92$, is required on each iteration of the inscribed ellipsoid algorithm (IEM) for convex programming [14]. The IEM minimizes a convex function over an n-dimensional cube to relative accuracy ν in $O(n\ln(n/\nu))$ iterations, each requiring evaluation of the function and a subgradient. The order of this complexity, also achieved by the volumetric cutting plane algorithm [1, 15], is optimal (see [13]).

Another application of γ-maximal ellipsoids is to provide a "rounding" of $\mathcal{P}$. It is known that, for the maximum volume inscribed ellipsoid (MVIE) E^*,

$$E^* \subset \mathcal{P} \subset nE^*,$$

where, for an ellipsoid E and positive scalar τ, τE denotes the dilation of E about its center by the factor τ. For a γ-maximal ellipsoid E it can be shown [14] that

$$E \subset \mathcal{P} \subset n\left(\frac{1+3\sqrt{1-\gamma}}{\gamma}\right)E.$$

Roundings of this type are required in several contexts, including Lenstra's algorithm for integer programming in fixed dimension [10] and randomized algorithms for volume

*Received by the editors June 18, 2001; accepted for publication (in revised form) February 24, 2002; published electronically September 12, 2002.
http://www.siam.org/journals/siopt/13-2/39090.html
†Department of Management Sciences, University of Iowa, Iowa City, IA 52242 (kurt-anstreicher@uiowa.edu).

computation [7]. Alternative methodologies for obtaining $O(n)$-roundings of $\mathcal{P}$ include the shallow cut ellipsoid algorithm [3, section 4.6] and the volumetric cutting plane algorithm [2].

Assume that $\mathcal{P}$ contains a ball of radius one centered at the origin and is itself contained in a ball of radius R centered at the origin. Using the ellipsoid algorithm, an $e^{-\epsilon}$-maximal inscribed ellipsoid can be computed in $O(n^6(n^2+m)\ln(nR/\epsilon))$ operations [14]. For reasonable m this complexity was substantially improved to

$$O\left(m^{2.5}(n^2+m)\ln\left(\frac{mR}{\epsilon}\right)\right) \tag{1}$$

operations by Nesterov and Nemirovskii [12], using an interior-point algorithm with a specialized "rescaling" technique to lower the work required on each iteration. A further reduction to

$$O\left(m^{3.5}\ln\left(\frac{mR}{\epsilon}\right)\ln\left(\frac{n\ln R}{\epsilon}\right)\right) \tag{2}$$

operations was achieved by Khachiyan and Todd [9], who apply an interior-point algorithm to a sequence of problems, each of which requires less work per iteration than the original problem considered by [12]. Nemirovskii [11] lowers the complexity of obtaining an $e^{-\epsilon}$-maximal inscribed ellipsoid to

$$O\left(m^{3.5}\ln\left(\frac{mR}{\epsilon}\right)\right) \tag{3}$$

operations. The approach taken in [11] uses Lagrangian duality to reformulate the MVIE problem as a saddlepoint problem, and shows that the self-concordance theory developed in [12] can be adapted to analyze an algorithm for computing an approximate saddlepoint. The advantage of [11] is that the theory developed for saddlepoint problems is very general. However, the analysis required to develop this theory is quite extensive.

In this paper we devise an optimization algorithm for which the complexity of obtaining an $e^{-\epsilon}$-maximal inscribed ellipsoid is also (3). Our work can be viewed as a further improvement of the previous optimization-based results of [12] and [9] and provides an alternative to the saddlepoint-based approach of [11]. We also show that, by first computing an approximation of the analytic center of $\mathcal{P}$, we can further reduce the effect of the parameter R, resulting in a total complexity of

$$O\left((mn^2+m^{1.5}n)\ln(R)+m^{3.5}\ln\left(\frac{m}{\epsilon}\right)\right) \tag{4}$$

operations. The difference between (3) and (4) is certainly of interest since, under standard assumptions, bounds on R may be exponential in n [3, Lemma 3.1.25].

Several novel formulations of the MVIE problem are considered in [16]. Primal-dual algorithms based on two of these formulations are analyzed in [17], and their numerical performance is compared to original and modified versions of the algorithm from [9] on instances up to size $n = 500$, $m = 1200$. The performance of one of the primal-dual algorithms is found to be superior to the other three methods.

A problem related to that of computing an $e^{-\epsilon}$-maximal inscribed ellipsoid for $\mathcal{P}$ is that of computing an e^{ϵ}-minimal circumscribing ellipsoid for the convex hull of m, given points in $\Re^n$. Khachiyan [8] shows that the latter problem can be solved in

$$O\left(m^{3.5}\ln\left(\frac{m}{\epsilon}\right)\right)$$

operations; note that this bound is independent of the parameter R.

Notation. If A and B are symmetric matrices, $A \preceq B$ denotes that $B - A$ is positive semidefinite, and $A \prec B$ denotes that $B - A$ is positive definite. The trace of a matrix A is denoted $\mathrm{tr}(A)$, $A \bullet B = \mathrm{tr}(AB^T)$, and $\|A\|$ denotes the Frobenius norm, $\|A\| = \sqrt{A \bullet A}$. We use ldet A to denote $\ln(\det A)$). The Kronecker product of matrices A and B is denoted $A \otimes B$. If A is an $m \times n$ matrix, $\mathbf{vec}(A)$ is the vector in $\Re^{mn}$ formed by "stacking" the columns of A atop one another in the natural order. We use $B(x, r)$ to denote the closed ball of radius r centered at $x \in \Re^n$.

2. Preliminaries. In this section we give definitions and basic results from [12] that will be required in what follows.

DEFINITION 2.1. *Let G be a closed, convex set in $\Re^N$, and let $f(\cdot) : \mathrm{Int}(G) \to \Re$ be a C^3 convex function. Then $f(\cdot)$ is said to be strongly* 1*-self-concordant (hereafter abbreviated strongly self-concordant) on* $\mathrm{Int}(G)$ *if $f(x_k) \to \infty$ for any sequence $\{x_k\}$ converging to a boundary point of G, and*

$$|D^3 f(x)[h,h,h]| \leq 2\left(D^2 f(x)[h,h]\right)^{3/2}$$

for every $x \in \mathrm{Int}(G)$ and $h \in \Re^N$.

Assume that $f(\cdot)$ is strongly self-concordant on $\mathrm{Int}(G)$, and that G is bounded. It can then be shown that $\nabla^2 f(x)$ is nonsingular for every $x \in \mathrm{Int}(G)$. For $x \in \mathrm{Int}(G)$ define the *Newton direction* for $f(\cdot)$ at x to be

$$p(x) \;=\; -[\nabla^2 f(x)]^{-1} \nabla^T f(x),$$

and the *Newton decrement* for $f(\cdot)$ at x to be

$$\lambda(x) = \left(\nabla f(x)[\nabla^2 f(x)]^{-1}\nabla^T f(x)\right)^{1/2}.$$

As shown in [12], for a strongly self-concordant function the Newton decrement provides good information regarding the difference between $f(x)$ and the minimum of $f(\cdot)$ over G. Note that if G is compact, then $f_{\min} = \min\{f(x) \,|\, x \in G\}$ is attained at a unique interior point of G.

LEMMA 2.2. *Let $G \subset \Re^N$ be a compact convex set, and assume that $f(\cdot)$ is strongly self-concordant on* $\mathrm{Int}(G)$*. Let $x \in \mathrm{Int}(G)$, $\lambda = \lambda(x)$, $p = p(x)$.*

1. *If $\lambda \leq 1/3$, then $f(x) - f_{\min} \leq \lambda^2/(1 - 5.0625\lambda^2)$.*
2. *If $x^+ = x + [1/(1+\lambda)]p$, then $x^+ \in \mathrm{Int}(G)$, and $f(x^+) \leq f(x) - [\lambda - \ln(1+\lambda)]$.*

Proof. Part 1 is proved in [4, Lemma 2.22]; a weaker estimate is given in [12, Theorem 2.2.2]. Part 2 is proved in [12, Proposition 2.2.2] and [4, Lemma 2.24]. ∎

DEFINITION 2.3. *Let $G \subset \Re^N$ be a compact convex set, and $F(\cdot) : \mathrm{Int}(G) \to \Re$. Then $F(\cdot)$ is called a ϑ-self-concordant barrier for G if $F(\cdot)$ is strongly self-concordant on* $\mathrm{Int}(G)$ *and $\lambda^2(x) \leq \vartheta$ for every $x \in \mathrm{Int}(G)$.*

It is very well known that the complexity of linear and quadratic optimization over G is characterized by the parameter ϑ. In particular, if $f(\cdot)$ is a convex quadratic function and $F(\cdot)$ is a computable ϑ-self-concordant barrier for G, then given a suitable initial interior point $x^0 \in G$ and lower bound $z^0 \leq f_{\min}$, an interior-point algorithm based on $F(\cdot)$ can be used to obtain an x having $f(x) - f_{\min} \leq \epsilon[f(x^0) - z^0]$ in $O(\sqrt{\vartheta}|\ln\epsilon|)$ iterations, each requiring a Newton step for a linear combination of $f(\cdot)$ and $F(\cdot)$.

To analyze the complexity of optimizing more general convex functions over G, [12] uses the concept of *β-compatibility* between a convex objective $f(\cdot)$ and a barrier

$F(\cdot)$ for G. The details are not important here, but we note that the complexity of approximately minimizing $f(\cdot)$ over G involves the parameters β and ϑ as well as characteristics of the initial point.

3. The MVIE problem. As described in [12], the problem of computing the MVIE for a polyhedral set $\mathcal{P}$ can be cast as the convex programming problem

$$\begin{aligned} \min\ & -\operatorname{ldet} Y \\ \text{s.t.}\ & \|Ya_i\| \le (b_i - a_i^T y), \quad i = 1, \ldots, m, \\ & Y \succeq 0, \end{aligned} \tag{5}$$

where a_i^T denotes the ith row of A. A feasible solution to (5) with objective value within ϵ of optimality provides an $e^{-\epsilon}$-maximal inscribed ellipsoid of the form $\{y + Yz \,|\, \|z\| \le 1\}$. The complexity analysis in [12] uses the $2(m+n)$-self-concordant barrier

$$-2\operatorname{ldet} Y - \sum_{i=1}^{m} \ln((b_i - a_i^T y)^2 - a_i^T Y^2 a_i)$$

and the fact that $f(Y) = -\operatorname{ldet} Y$ is 1-compatible with this barrier. The main difficulty with this approach is that the resulting Newton equations are relatively expensive to form and solve. In [12] a "speed-up" based on rescaling the matrix Y is used to reduce this complexity to $O(m^2(n^2+m))$ operations per iteration, resulting in the overall complexity (1).

Letting $X = Y^2$, (5) is equivalent to the optimization problem

$$\begin{aligned} \min\ & -\operatorname{ldet} X \\ \text{s.t.}\ & a_i^T X a_i \le (b_i - a_i^T x)^2, \quad i = 1, \ldots, m, \\ & a_i^T x \le b_i, \quad i = 1, \ldots, m, \\ & X \succeq 0, \end{aligned} \tag{6}$$

and a solution of (6) with objective within 2ϵ of optimality produces an $e^{-\epsilon}$-maximal inscribed ellipsoid. Unfortunately the constraints of (6), while linear in X, are not all convex in x. Note that although the linear constraints $a_i^T x \le b_i$, $i = 1, \ldots, m$, are needed to correctly formulate (6), these constraints are implicitly enforced if they hold initially and strict inequality is maintained in the remaining constraints. The approach taken in [9] is to approximately solve a sequence of problems of the form

$$\begin{aligned} P(y): \quad \min\ & -\operatorname{ldet} X \\ \text{s.t.}\ & a_i^T X a_i \le (b_i - a_i^T y)(b_i - a_i^T x), \quad i = 1, \ldots, m, \\ & X \succeq 0, \end{aligned} \tag{7}$$

where y is an interior point of $\mathcal{P}$. The process is initialized using $y_0 = 0$, and if (x_k, X_k) is the approximate solution of $P(y_k)$, then $y_{k+1} = (1/2)(x_k + y_k)$. In [9] the convergence of $\{y_k\}$ is shown to be very rapid. Moreover, the barrier

$$-\operatorname{ldet} X - \sum_{i=1}^{m} \ln((b_i - a_i^T y)(b_i - a_i^T x) - a_i^T X a_i) \tag{8}$$

for $P(y)$ is $(m+n)$-self-concordant, $f(X) = -\operatorname{ldet} X$ is $O(1)$-compatible with this barrier, and the Newton direction required on each iteration can be computed in only

$O(m^3)$ operations. This reduces the complexity of finding an $e^{-\epsilon}$-maximal inscribed ellipsoid for $\mathcal{P}$ to (2).

The approach we take here uses the family of barriers (8) as in [9] but avoids solving the sequence of problems $P(y_k)$. In this way we reduce the computation required for each Newton step to $O(m^3)$ operations but avoid the factor $\ln((n \ln R)/\epsilon)$ in (2). We also show that "pre-rounding" $\mathcal{P}$ by first computing an approximation of the analytic center of $\mathcal{P}$ can be used to further reduce the effect of the parameter R, resulting in the complexity (4).

4. Main stage. Let G denote the feasible region of (6), and for $y \in \mathrm{Int}(\mathcal{P})$ let $G(y)$ denote the feasible region of $P(y)$, from (7). For $(x, X) \in \mathrm{Int}(G(y))$ and $t \geq 1$ let

$$F_t(y; x, X) = -t \operatorname{ldet} X - \sum_{i=1}^{m} \ln\big((b_i - a_i^T y)(b_i - a_i^T x)\big) - \sum_{i=1}^{m} \ln\big((b_i - a_i^T y)(b_i - a_i^T x) - a_i^T X a_i\big). \tag{9}$$

It is then straightforward to show that for any $y \in \mathrm{Int}(P)$, $F_t(y; \cdot, \cdot)$ is strongly self-concordant on $\mathrm{Int}(G(y))$. In working with $F_t(y; x, X)$, we consider the components of y to be fixed parameters, while those of (x, X) are variables. Let $[p_t(y; x, X), P_t(y; x, X)]$ denote the Newton direction for $F_t(y; \cdot, \cdot)$ at (x, X), and let $\lambda_t(y; x, X)$ be the corresponding Newton decrement. In this section we describe and analyze the "main stage" of our barrier algorithm for obtaining an $e^{-\epsilon}$-maximal ellipsoid. The main stage is initialized with $t_0 = 1$ and a point (x_0, X_0) such that $x_0 \in \mathrm{Int}(\mathcal{P})$, $(x_0, X_0) \in \mathrm{Int}(G(x_0))$, and $\lambda_1(x_0; x_0, X_0) \leq 0.15$. The problem of obtaining such an initial point is considered in the next section. The main stage algorithm, described in pseudocode below, is a variant of the standard barrier algorithm for convex optimization analyzed in [12]. The novelty of the algorithm here is that the Newton direction used on each inner iteration is obtained from a barrier function $F_t(x; \cdot, \cdot)$ that depends on the current x.

ALGORITHM 1 (Main stage for MVIE).
Given $k = 0$, x_0, X_0, $t_0 = 1$, $t_{\max}$, $\theta > 0$.
Do until $t_k \geq t_{\max}$ (outer iteration)
  $t = t_{k+1} = (1 + \theta)t_k$, $x = x_k$, $X = X_k$.
  Do until $\lambda_t(x; x, X) \leq 0.15$ (inner iteration)
    $p = p_t(x; x, X)$, $P = P_t(x; x, X)$,
    $x = x + (\alpha/2)p$, $X = X + \alpha P$.
  End
  $x_{k+1} = x, X_{k+1} = X$, $k = k + 1$.
End

The steplength α on each inner iteration can be taken to be any value that produces at least the descent in $F_t(\cdot; \cdot, \cdot)$ obtained using $\alpha = 1/(1 + \lambda_t(x; x, X))$; see Lemma 4.3 below. The use of $\alpha/2$ in the step for x may seem strange at first sight but is a simple consequence of the use of the direction $p = p_t(x; \cdot, \cdot)$ based on $F_t(x; \cdot, \cdot)$. In particular, note that, for each i,

$$(b_i - a_i^T x)(b_i - a_i^T (x + \alpha p)) = (b_i - a_i^T x)^2 - \alpha (a_i^T p)(b_i - a_i^T x),$$
$$\left(b_i - a_i^T \left(x + \frac{\alpha}{2} p\right)\right)^2 = (b_i - a_i^T x)^2 - \alpha (a_i^T p)(b_i - a_i^T x) + \alpha^2 \frac{(a_i^T p)^2}{4}.$$

As a result, using $x^+ = x + (\alpha/2)p$, $X^+ = X + \alpha P$ in a step from $(x; x, X)$ to $(x^+; x^+, X^+)$ has the same first-order effect on $F_t(\cdot;\cdot,\cdot)$ as using $x^{++} = x + \alpha p$ and a step to $(x; x^{++}, X^+)$.

Our analysis of the main stage algorithm for MVIE is based on the well-known analysis of the barrier algorithm from [12] (see also [4]). The following result facilitates the use of directions based on the family of barrier functions $F_t(y;\cdot,\cdot)$.

LEMMA 4.1. *Suppose that x and y are interior points of $\mathcal{P}$, and let $f(X) = -\operatorname{ldet} X$, $G(y,x) = \{X \succeq 0 \,|\, a_i^T X a_i \leq (b_i - a_i^T y)(b_i - a_i^T x),\ i = 1,\ldots,m\}$. Define $\phi(y,x) = \min_{X \in G(y,x)} f(X)$ and $\phi_t(y,x) = \min_{X \in G(y,x)} F_t(y;x,X)$, $t \geq 1$. Then*

1. $\phi(y,x) \leq \frac{1}{2}[\phi(y,y) + \phi(x,x)]$,
2. $\phi_t(y,x) \leq \frac{1}{2}[\phi_t(y,y) + \phi_t(x,x)]$.

Proof. Part 1 is proved in [9, p. 144], and we use a similar argument to prove part 2 here. Assume that $X \succ 0$ is the minimizer of $F_t(x;x,\cdot)$ and that $Y \succ 0$ is the minimizer of $F_t(y;y,\cdot)$. By a change of coordinates we may assume without loss of generality that X and Y are diagonal. For each $i = 1,\ldots,m$ the inequalities $a_i^T X a_i \leq (b_i - a_i^T x)^2$ and $a_i^T Y a_i \leq (b_i - a_i^T y)^2$ together imply that

$$\begin{aligned} a_i^T (XY)^{1/2} a_i &\leq \|X^{1/2} a_i\| \, \|Y^{1/2} a_i\| \\ &= [(a_i^T X a_i)(a_i^T Y a_i)]^{1/2} \\ &\leq (b_i - a_i^T x)(b_i - a_i^T y) \end{aligned} \tag{10}$$

and $\operatorname{ldet}(XY)^{1/2} = (1/2)(\operatorname{ldet} X + \operatorname{ldet} Y)$. To prove that $\phi_t(y,x) \leq \frac{1}{2}[\phi_t(y,y) + \phi_t(x,x)]$, it then suffices to show that, for each $i = 1,\ldots,m$,

$$\begin{aligned} &-\ln\big((b_i - a_i^T x)(b_i - a_i^T y) - a_i^T (XY)^{1/2} a_i\big) \\ &\quad \leq -\frac{1}{2}\ln\big((b_i - a_i^T x)^2 - a_i^T X a_i\big) - \frac{1}{2}\ln\big((b_i - a_i^T y)^2 - a_i^T Y a_i\big), \end{aligned}$$

which is equivalent to

$$\begin{aligned} &\big((b_i - a_i^T x)(b_i - a_i^T y) - a_i^T (XY)^{1/2} a_i\big)^2 \\ &\qquad \geq \big((b_i - a_i^T x)^2 - a_i^T X a_i\big)\big((b_i - a_i^T y)^2 - a_i^T Y a_i\big). \end{aligned} \tag{11}$$

Using the first inequality in (10), to prove (11) it suffices to show that

$$\begin{aligned} &\big((b_i - a_i^T x)(b_i - a_i^T y) - [(a_i^T X a_i)(a_i^T Y a_i)]^{1/2}\big)^2 \\ &\qquad \geq \big((b_i - a_i^T x)^2 - a_i^T X a_i\big)\big((b_i - a_i^T y)^2 - a_i^T Y a_i\big), \end{aligned}$$

which reduces to $[(b_i - a_i^T x)(a_i^T Y a_i)^{1/2} - (b_i - a_i^T y)(a_i^T X a_i)^{1/2}]^2 \geq 0$. ∎

We now use Lemma 4.1 and standard results on self-concordant functions to bound the possible reduction in $F_t(\cdot;\cdot,\cdot)$ and $f(\cdot)$ when the Newton decrement is sufficiently small.

LEMMA 4.2. *Suppose that $t > 1$, $x \in \operatorname{Int}(\mathcal{P})$, $(x,X) \in \operatorname{Int}(G(x))$, and $\lambda_t(x;x,X) \leq \lambda \leq 1/3$. Let $\delta = \delta(\lambda) = \lambda^2/(1 - 5.0625\lambda^2)$. Then*

1. *$F_t(y;y,Y) \geq F_t(x;x,X) - 2\delta$ for all $(y,Y) \in G$.*
2. *$f(X) \leq f_{\min} + [12m + 2\delta]/(t-1)$.*

Proof. From part 1 of Lemma 2.2, we have

$$F_t(x;y,Y) \geq F_t(x;x,X) - \delta \tag{12}$$

for any $(y, Y) \in G(x)$. Part 2 of Lemma 4.1 then implies that, for any $y \in \text{Int}(\mathcal{P})$,

$$\begin{aligned}\phi_t(y,y) &\geq 2\phi_t(y,x) - \phi_t(x,x)\\ &\geq 2(F_t(x;x,X) - \delta) - F_t(x;x,X)\\ &= F_t(x;x,X) - 2\delta,\end{aligned}$$

which proves part 1. Next, note that $F_t(x;x,X) = (t-1)f(X) + F_1(x;x,X)$, where $F_1(x;\cdot,\cdot)$ is a $(2m+n)$-self-concordant barrier for $G(x)$. From (12) and a standard argument (see, for example, [12, p. 75]) we conclude that, for all $(y, Y) \in G(x)$,

$$f(Y) \geq f(X) - \frac{2(2m+n)+\delta}{t-1},$$

and therefore $\phi(x,y) \geq f(X) - (6m+\delta)/(t-1)$ for all $y \in \text{Int}(\mathcal{P})$, since $m > n$. Part 1 of Lemma 4.1 then implies

$$\begin{aligned}\phi(y,y) &\geq 2\phi(y,x) - \phi(x,x)\\ &\geq 2\left(f(X) - \frac{6m+\delta}{t-1}\right) - f(X)\\ &= f(X) - \frac{12m+2\delta}{t-1},\end{aligned}$$

which proves part 2. □

The next lemma considers the effect of increasing t on the Newton decrement for $F_t(x;x,X)$, and the reduction in $F_t(\cdot;\cdot,\cdot)$ that can be assured if the Newton decrement is not sufficiently small.

LEMMA 4.3. *For $t \geq 1$ let $\lambda_t(y;x,X)$ be the Newton decrement for $F_t(y;\cdot,\cdot)$ at (x,X).*

1. *Suppose that $\lambda_t(x;x,X) \leq .15$, and let $t^+ = (1+\theta)t$, $\theta \geq 0$. Then $\lambda_{t^+}(x;x,X) \leq .15(1+\theta) + \sqrt{2m}\theta$.*

2. *Suppose that $\lambda_t(x;x,X) = \lambda > .15$. Let $\alpha = 1/(1+\lambda)$, $x^+ = x + (\alpha/2)p_t(x;x,X)$, $X^+ = X + \alpha P_t(x;x,X)$. Then $(x^+, X^+) \in \text{Int}(G)$, and $F_t(x^+;x^+,X^+) \leq F_t(x;x,X) - 0.01$.*

Proof. Part 1 is proved in [4, Lemma 2.25]. Since $F_t(x;\cdot,\cdot)$ is strongly self-concordant for $t \geq 1$, part 2 of Lemma 2.2 implies that if $x^{++} = \alpha p_t(x;x,X)$, then $(x^{++}, X^+) \in \text{Int}(G(x))$ and

$$F_t(x;x^{++},X^+) \leq F_t(x;x,X) - (\lambda - \ln(1+\lambda)) \leq F_t(x;x,X) - 0.01. \tag{13}$$

It is also easy to see that, for any x and y in $\mathcal{P}$, and $i = 1,\ldots,m$,

$$(b_i - a_i^T x)(b_i - a_i^T y) \leq \left(b_i - a_i^T \frac{(x+y)}{2}\right)^2,$$

and therefore if $(y, Y) \in G(x)$, then

$$F_t\left(\frac{x+y}{2};\frac{x+y}{2},Y\right) \leq F_t(x;y,Y). \tag{14}$$

Part 2 is completed by combining (13) and (14), with $(y, Y) = (x^{++}, X^+)$. □

We can now combine Lemmas 4.2 and 4.3 to obtain the final complexity result for the main stage.

THEOREM 4.4. *Let $\theta = 0.07/\sqrt{m}$. Then the main stage algorithm (Algorithm 1) requires $O(1)$ inner iterations per outer iteration. Moreover, for $t_{\max} = O(m/\epsilon)$ the algorithm terminates with an $e^{-\epsilon}$-maximal ellipsoid in $O(m^{.5}\ln(m/\epsilon))$ outer iterations, requiring a total of $O(m^{3.5}\ln(m/\epsilon))$ operations.*

Proof. From part 1 of Lemma 4.3, at the start of each sequence of inner iterations we have

$$\lambda_{t_{k+1}}(x_k; x_k, X_k) \leq .15\left(1 + \frac{.07}{\sqrt{m}}\right) + .1 < 0.26.$$

Part 1 of Lemma 4.2 then implies that, for any $(x, X) \in G$,

$$F_{t_{k+1}}(x; x, X) > F_{t_{k+1}}(x_k; x_k, X_k) - .21,$$

and from Part 2 of Lemma 4.3 there can be at most 20 inner iterations on each outer iteration.

From part 2 of Lemma 4.2, to obtain an $e^{-\epsilon}$-maximal inscribed ellipsoid, it suffices to terminate the algorithm using $t_{\max} = O(m/\epsilon)$, which requires $O(m^{.5}\ln(m/\epsilon))$ outer iterations using $\theta = .07/\sqrt{m}$. Finally, it can be shown using a small modification of the argument used in [9, section 6] that each inner iteration can be executed in $O(m^3)$ operations. □

To close this section we note two details regarding the definition of $F_t(y; x, X)$, from (9). First, it would be more "standard" to replace the term $-t\,\mathrm{ldet}\,X$ with $-(t+1)\,\mathrm{ldet}\,X$, and to work with $t \geq 0$ instead of $t \geq 1$. However, in this case the main stage algorithm must be initialized at a sufficiently large $t_0 > 0$, which would unnecessarily complicate the analysis of the preliminary stage in the next section. Second, the analysis of this section does not require the term

$$-\sum_{i=1}^{m} \ln\big((b_i - a_i^T y)(b_i - a_i^T x)\big) \tag{15}$$

in (9), and in fact this term slightly degrades the theoretical performance of the main stage algorithm. However, (15) is very helpful for the analysis of the preliminary stage.

5. Preliminary stage. In this section we consider the preliminary stage of our barrier algorithm for the MVIE problem. The goal of the preliminary stage is to produce the initial point (x_0, X_0) required by the main stage algorithm of the previous section (Algorithm 1). Our preliminary stage is based on the general preliminary stage described in [12, section 3.2.3], except that we work with directions based on $F_1(y; x, X)$ so as to keep the work per iteration to $O(m^3)$.

The preliminary stage is initialized at $x_0 = 0$ and a suitable X_0 to be described below. Let

$$c = -\nabla_x F_1(0; 0, X_0)^T = -\sum_{i=1}^{m}\left(\frac{b_i}{\Delta_{i,0}} + \frac{1}{b_i}\right)a_i, \tag{16a}$$

$$C = -\nabla_X F_1(0; 0, X_0) = X_0^{-1} - \sum_{i=1}^{m}\frac{1}{\Delta_{i,0}}a_i a_i^T, \tag{16b}$$

where $\Delta_{i,0} = b_i^2 - a_i^T X_0 a_i$. For $t \leq 1$ define the preliminary stage barrier function

$$F_t^0(y; x, X) = t(c^T(x + y) + C \bullet X) + F_1(y; x, X).$$

Let $[p_t^0(y;x,X), P_t^0(y;x,X)]$ denote the Newton direction for $F_t^0(y;\cdot,\cdot)$ at (x,X), and $\lambda_t^0(y;x,X)$ the corresponding Newton decrement. Note that, by construction, $\lambda_1^0(0;0,X_0)=0$. The preliminary stage algorithm, given below, is very similar to the main stage, except that the preliminary stage uses $F_t^0(\cdot;\cdot,\cdot)$, and t is decreased rather than increased on each outer iteration.

ALGORITHM 2 (Preliminary stage for MVIE).
Given $k=0$, $x_0=0$, X_0, $t_0=1$, $t_{\min}$, $\theta>0$.
Do until $t_k \le t_{\min}$ (outer iteration)
  $t=t_{k+1}=(1-\theta)t_k$, $x=x_k$, $X=X_k$.
  Do until $\lambda_t^0(x;x,X)\le 0.15$ (inner iteration)
    $p=p_t^0(x;x,X)$, $P=P_t^0(x;x,X)$,
    $x=x+(\alpha/2)p$, $X=X+\alpha P$.
  End
  $x_{k+1}=x, X_{k+1}=X$, $k=k+1$
End

To analyze the preliminary stage, we require an extension of part 2 of Lemma 4.1 that applies to $F_t^0(y;x,X)$. This turns out to be straightforward under the assumption that $C\succeq 0$.

LEMMA 5.1. *For interior points x and y of $\mathcal{P}$, let $\phi_t^0(y,x)=\min_{X\in G(y,x)} F_t^0(y;x,X)$, $0<t\le 1$. Assume that $C\succeq 0$. Then $\phi_t^0(y,x)\le \frac{1}{2}[\phi_t^0(y,y)+\phi_t^0(x,x)]$.*

Proof. The proof is identical to the proof of part 2 of Lemma 4.1. Note that the change of coordinates that simultaneously diagonalizes X and Y preserves the semidefiniteness of C. After this change of coordinates, we have

$$C\bullet X=\sum_{i=1}^m C_{ii}X_{ii},\quad C\bullet Y=\sum_{i=1}^m C_{ii}Y_{ii},\quad C\bullet (XY)^{1/2}=\sum_{i=1}^m C_{ii}\sqrt{X_{ii}Y_{ii}},$$

where $C_{ii}\ge 0$, $i=1,\dots,m$. It immediately follows that $C\bullet (XY)^{1/2}\le (1/2)(C\bullet X + C\bullet Y)$. ∎

For $C\succeq 0$ and a given value of $t_{\min}$, the analysis of the preliminary stage is very similar to the analysis of the main stage given in the previous section, and it is omitted here. The final complexity result has the following form.

THEOREM 5.2. *Assume that $C\succeq 0$. Then for $\theta=\eta/\sqrt{m}$, where $\eta>0$ is an appropriate positive constant, the preliminary stage requires $O(m^{.5}\ln(1/t_{\min}))$ outer iterations, $O(1)$ inner iterations per outer iteration, and a total of $O(m^{3.5}\ln(1/t_{\min}))$ operations.*

To complete the analysis of the preliminary stage, we must show that X_0 can be chosen so that $C\succeq 0$ and characterize the value $t_{\min}$ so that termination of the preliminary stage produces a suitable initial point for the main stage.

LEMMA 5.3. *Assume that $B(0,1)\subset\mathcal{P}$, and let $X_0=\frac{1}{m+1}I$. Then $C\succeq 0$.*

Proof. By the assumption that $B(0,1)\subset\mathcal{P}$, we must have $b_i\ge\|a_i\|$, $i=1,\dots,m$. Then, for each i,

$$\Delta_{i,0}=b_i^2-a_i^TX_0a_i=b_i^2-\frac{1}{m+1}\|a_i\|^2\ge b_i^2\frac{m}{m+1}, \tag{17}$$

and

$$C\succeq (m+1)I-\frac{m+1}{m}\sum_{i=1}^m\frac{1}{b_i^2}a_ia_i^T\succeq (m+1)\left(I-\frac{1}{m}\sum_{i=1}^m\frac{1}{\|a_i\|^2}a_ia_i^T\right)\succeq 0$$

follows from (16b). ∎

It remains to obtain a lower bound on the required value of $t_{\min}$. To this end, note that for any strongly self-concordant function $F(\cdot)$ defined on the interior of a compact, convex set $G \subset \Re^N$, the Newton decrement $\lambda(x)$ at $x \in \mathrm{Int}(G)$ is equal to the solution value in the optimization problem

$$\begin{aligned} \max \;& DF(x)[h] \\ \text{s.t.} \;& D^2F(x)[h,h] \leq 1. \end{aligned}$$

It follows from this characterization, and from the fact that $F_t^0(y;x,X) = t(c^Tx + c^Ty + C \bullet X) + F_1(y;x,X)$, that, for any $(x,X) \in \mathrm{Int}(G)$,

$$\lambda_1(x;x,X) \leq \lambda_t^0(x;x,X) + tv(x,X), \tag{18}$$

where

$$\begin{aligned} v(x,X) = \max \;& c^Th + C \bullet H \\ \text{s.t.} \;& D^2F_1(x;x,X)[(h,H),(h,H)] \leq 1. \end{aligned} \tag{19}$$

Thus, to obtain a lower bound on the required value $t_{\min}$, we require an upper bound for $v(\cdot,\cdot)$.

LEMMA 5.4. *Assume that $B(0,1) \subset \mathcal{P} \subset B(0,R)$, and let $X_0 = \frac{1}{m+1}I$. Then at any $(x,X) \in \mathrm{Int}(G)$, $v(x,X) \leq (4+\sqrt{n})(m+1)R^2$.*

Proof. It is straightforward to compute (see, for example, [9]) that

$$\begin{aligned} D^2F_1(x;x,X)[(h,H),(h,H)] = \mathbf{vec}(H)^T(X^{-1} \otimes X^{-1})\,\mathbf{vec}(H) + \sum_{i=1}^m \frac{1}{s_i^2}(a_i^Th)^2 \\ + \sum_{i=1}^m \frac{1}{\Delta_i^2}(a_i^THa_i + s_ia_i^Th)^2, \end{aligned} \tag{20}$$

where $s_i = b_i - a_i^Tx$ and $\Delta_i = (b_i - a_i^Tx)^2 - a_i^TXa_i$. From the assumption that $\mathcal{P} \subset B(0,R)$ and from the relationship between (5) and (6), we must have $X \preceq R^2I$, and thus $X^{-1} \succeq (1/R^2)I$ and $X^{-1} \otimes X^{-1} \succeq (1/R^4)I$ (see [5, p. 252]). Therefore

$$\mathbf{vec}(H)^T(X^{-1} \otimes X^{-1})\,\mathbf{vec}(H) \geq (1/R^4)\|H\|^2$$

for any H. From (20), if (h,H) is feasible in (19), then we clearly have $\|H\|^2 \leq R^4$, and thus

$$C \bullet H \leq \|C\|\|H\| \leq (m+1)\sqrt{n}R^2, \tag{21}$$

since $0 \preceq C \preceq X_0^{-1}$. In addition, the fact that $\mathcal{P} \subset B(0,R)$ implies that for any $h \neq 0$ there is an index $i = i(h)$ such that $a_i^Th/\|h\| \geq b_i/R$, implying that $b_i \leq R\|a_i\|$, and $a_i^Th \geq b_i\|h\|/R \geq \|a_i\|\,\|h\|/R$. For this same index it must also be that $s_i = b_i - a_i^Tx \leq b_i + R\|a_i\| \leq 2R\|a_i\|$. Combining these facts, we conclude that, for any h,

$$\sum_{i=1}^m \frac{1}{s_i^2}(a_i^Th)^2 \geq \left(\frac{1}{2R\|a_{i(h)}\|}\right)^2 \left(\frac{\|a_{i(h)}\|\,\|h\|}{R}\right)^2 = \frac{\|h\|^2}{4R^4}.$$

Therefore if (h,H) is feasible in (19), it must be that $\|h\|^2 \leq 4R^4$. Moreover, from (17) we have

$$\frac{b_i}{\Delta_{i,0}} + \frac{1}{b_i} \leq \frac{2m+1}{mb_i} < \frac{2(m+1)}{m\|a_i\|}$$

for each i, implying that $\|c\| \leq 2(m+1)$. We conclude that if (h, H) is feasible in (19), then

$$c^T h \leq \|c\| \, \|h\| \leq 4(m+1)R^2. \tag{22}$$

The proof is completed by combining (21) and (22). ∎

THEOREM 5.5. *Assume that $B(0,1) \subset \mathcal{P} \subset B(0,R)$, and let $X_0 = \frac{1}{m+1} I$. Then for $1/t_{\min} = O(\sqrt{n} m R^2)$ the preliminary stage terminates with an (x, X) having $\lambda_1(x; x, X) \leq .26$, using a total of $O(m^{3.5} \ln(mR))$ operations.*

Proof. This follows immediately from Theorem 5.2, (18), and Lemma 5.4. ∎

Given (x, X) with $\lambda_1(x; x, X) \leq .26$, using at most 20 inner iterations of the main stage algorithm (see the proof of Theorem 4.4), we can obtain (x, X) having $\lambda_1(x; x, X) \leq .15$, as required to initialize the main stage.

Note that the second term on the right-hand side of (20) arises from the presence of (15) in the definition of $F_t(\cdot; \cdot, \cdot)$, and this term is responsible for the bound (22) used in the proof of Theorem 5.2. Without (15) we would be forced to rely on the third term of (20) to bound $c^T h$. Such an analysis may be possible but appears to require that s_i be bounded away from zero for each i. It is interesting to note that a similar issue appears in the analysis of [9]; see the proof of [9, Theorem 3].

Combining Theorems 4.4 and 5.5, we immediately obtain the overall complexity bound (3). Note that the parameter R only appears in the complexity bound for the preliminary stage. We next show that the effect of R can be reduced by first computing an approximation of the ordinary analytic center of $\mathcal{P}$.

The analytic center of $\mathcal{P}$ is the minimizer of the logarithmic barrier function

$$F(x) = -\sum_{i=1}^{m} \ln(b_i - a_i^T x). \tag{23}$$

Let $\lambda(x)$ be the Newton decrement for $F(\cdot)$ at x. It is then well known [12, section 3.2.3] that, under the assumption that $B(0,1) \subset \mathcal{P} \subset B(0,R)$, a point x with $\lambda(x) < .2$ can be computed using $O(m^{.5} \ln(mR))$ damped Newton steps. Moreover, it is straightforward to show (for example, using a small modification of the proof of [2, Lemma 3.1]) that, for such an x,

$$E(x, \nabla^2 F(x), 1) \subset \mathcal{P} \subset E(x, \nabla^2 F(x), 1.25m),$$

where for a positive definite matrix H, $E(x, H, r) = \{y \,|\, (y-x)^T H (y-x) \leq r^2\}$. Using a change of coordinates (that scales volume by a constant factor), we can then move x to the origin and obtain $R = 1.25m$. Finally, by using "partial updating" of the Newton equations required on each iteration, the total complexity of obtaining the approximate analytic center x is

$$O((mn^2 + m^{1.5} n) \ln(mR))$$

operations; this complexity is comprised of a total of $O(m \ln(mR))$ updating steps, each requiring $O(n^2)$ work, and $O(mn)$ other operations per iteration (see [12, Chapter 8] or [4, Chapter 4]). It follows that, by first computing an approximation of the analytic center of $\mathcal{P}$, we reduce to (4) the overall complexity of computing an $e^{-\epsilon}$-maximal ellipsoid. It is worth noting that the technique described here can also be used to reduce the effect of R on the complexities of previous algorithms for the MVIE problem; see, for example, the methods of [11, 12, 9]. Finally, in the application to the inscribed ellipsoid method, $R = O(n)$ holds without loss of generality [14].

Acknowledgments. I am grateful to Mike Todd for several helpful suggestions, and to Yin Zhang for drawing my attention to [11].

REFERENCES

[1] K.M. Anstreicher, *On Vaidya's volumetric cutting plane method for convex programming*, Math. Oper. Res., 22 (1997), pp. 63–89.
[2] K.M. Anstreicher, *Ellipsoidal approximations of convex sets based on the volumetric barrier*, Math. Oper. Res., 24 (1999), pp. 193–203.
[3] M. Grötschel, L. Lovász, and A. Schrijver, *Geometric Algorithms and Combinatorial Optimization*, Springer-Verlag, Berlin, 1988.
[4] D. den Hertog, *Interior Point Approach to Linear, Quadratic, and Convex Programming—Algorithms and Complexity*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1994.
[5] R.A. Horn and C.R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.
[6] F. John, *Extremum problems with inequalities as subsidiary conditions*, in Studies and Essays, Presented to R. Courant on His 60th Birthday, January 8, 1948, Wiley Interscience, New York, pp. 187–204.
[7] R. Kannan, L. Lovász, and M. Simonovits, *Random walks and an $O^*(n^5)$ volume algorithm for convex bodies*, Random Structures Algorithms, 11 (1997), pp. 1–50.
[8] L.G. Khachiyan, *Rounding of polytopes in the real number model of computation*, Math. Oper. Res., 21 (1996), pp. 307–320.
[9] L.G. Khachiyan and M.J. Todd, *On the complexity of approximating the maximal volume inscribed ellipsoid for a polytope*, Math. Programming, 61 (1993), pp. 137–159.
[10] H.W. Lenstra, Jr., *Integer programming with a fixed number of variables*, Math. Oper. Res., 8 (1983), pp. 538–548.
[11] A. Nemirovskii, *On Self-Concordant Concave-Convex Functions*, Faculty of Industrial Engineering and Management, Technion, Haifa, Israel, 1997.
[12] Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM Stud. Appl. Math. 13, SIAM, Philadelphia, 1994.
[13] A.S. Nemirovskii and D.B. Yudin, *Problem Complexity and Method Efficiency in Optimization*, John Wiley, Chichester, UK, 1983.
[14] S.P. Tarasov, L.G. Khachiyan, and I.I. Erlich, *The method of inscribed ellipsoids*, Soviet Math. Dokl., 27 (1988), pp. 226–230.
[15] P.M. Vaidya, *A new algorithm for minimizing convex functions over convex sets*, Math. Programming, 73 (1996), pp. 291–341.
[16] Y. Zhang, *An Interior-Point Algorithm for the Maximum-Volume Ellipsoid Problem*, Technical report TR98-15, Center for Computational and Applied Mathematics, Rice University, Houston, TX, 1998.
[17] Y. Zhang and L. Gao, *On Numerical Solution of the Maximum Volume Ellipsoid Problem*, Technical report TR01-15, Center for Computational and Applied Mathematics, Rice University, Houston, TX, 2001.

SIAM J. OPTIM.
Vol. 13, No. 2, pp. 321–334

PRIMAL-DUAL ACTIVE SET STRATEGY FOR A GENERAL CLASS OF CONSTRAINED OPTIMAL CONTROL PROBLEMS*

K. KUNISCH† AND A. RÖSCH‡

Abstract. An algorithm for efficient solution of optimal control problems with pointwise constraints is introduced. It is based on an active set strategy using primal as well as dual variables. Under certain assumptions a decrease of an appropriately chosen merit function and convergence of the algorithm are proved. Numerical examples for a parabolic optimal control problem demonstrate the efficiency of the proposed method for control-constrained problems.

Key words. constrained optimal control problems, active set strategy, augmented Lagrangians, primal-dual methods

AMS subject classifications. 49J20, 49M29

PII. S1052623499358008

1. Introduction.

In this paper we discuss an active set strategy for the solution of infinite dimensional quadratic optimal control problems with linear equality constraints and pointwise affine inequality constraints. The active set strategy simultaneously uses primal and dual variables and differs significantly from active set strategies that involve primal variables only; see [17], for example. For control-constrained problems the practical behavior of this strategy is characterized by infeasible iterations. Often only the last iteration is feasible. The algorithm stops at a feasible and optimal point.

To specify the problem, let Y and U be Hilbert spaces corresponding to the state- and control space of a controlled dynamical system, and assume that U has the form $U = L^2(\Sigma)$, where Σ is a bounded set in $\mathbb{R}^n$. We investigate the optimal control problem involving control constraints:

$$\min\ F(y,u) = \frac{1}{2}\|y - y_d\|_Y^2 + \frac{\mu}{2}\|u\|_U^2 \tag{1.1}$$

subject to

$$y = Su + f \tag{1.2}$$

and

$$u \in U_{ad} = \{u \in U = L^2(\Sigma) :\ a \le u(x) \le b \text{ a.e. in } \Sigma\}, \tag{1.3}$$

where $f, y_d \in Y$, $\mu > 0$, and $-\infty \le a < b \le \infty$ and at least a or b is finite. Throughout, it is assumed that $S : U \to Y$ is a linear and compact operator.

The theory to be developed can easily be extended to more general quadratic functionals and to box constraints that are functions. In [1] we considered a specific case of problem (1.1)–(1.3), in which the mapping S arises from an elliptic optimal

*Received by the editors June 14, 1999; accepted for publication (in revised form) February 7, 2002; published electronically September 12, 2002.
http://www.siam.org/journals/siopt/13-2/35800.html
†Institute of Mathematics, University of Graz, A-8010 Graz, Austria (karl.kunisch@uni-graz.at).
This author was supported in part by the SFB 03, "Optimierung und Kontrolle."
‡Technical University of Berlin, Faculty II – Mathematics and Natural Sciences, Straße des 17. Juni 136, D-10623 Berlin, Germany (roesch@math.tu-berlin.de).

control problem with distributed control. In the present paper we extend the results from [1] to the situation in which the relationship between controls and outputs is described by (1.2). Moreover, the present analysis goes significantly beyond that of [1] because strong convergence of the iterates is proved. In [1], sufficient conditions for the strict decay of the costs were given, but convergence of the control could be proved only for discretized finite dimensional problems.

While we hope that the results are interesting in their own right, one major application is in the context of sequential quadratic programming (SQP-) techniques. These methods are by now well established for solving general nonlinear optimal control problems; see, for instance, Heinkenschloss and Tröltzsch [8], Kelley and Sachs [10], Kunisch and Sachs [14], Tröltzsch [21], and the references therein. The SQP-algorithm is sequential, and each of its iterations requires the solution of a quadratic minimization problem subject to linearized constraints. If these auxiliary problems contain inequality constraints with infinite dimensional image space, their efficient algorithmic solution is still a significant challenge.

Let us briefly note some alternative approaches to the numerical treatment of constrained optimal control problems. Projected gradient and projected Newton methods are treated in [18], [19] and [11], [12], respectively. In [22], affine scaling methods are analyzed for optimal control problems with L^p-bounds. Trust region methods for projected gradient and projected Newton methods are studied in [13].

Our primary focus, however, is different from that of most of the above-mentioned papers. The algorithm that we propose is rather direct; its computational efficiency and convergence proof do not rely on a globalization strategy. It is motivated by a generalized Moreau–Yosida-type approximation of the constraints. For unilaterally constrained problems the algorithm coincides with the dual active set strategy proposed in [7]. For the latter, local quadratic convergence for semidiscretized optimal control problems governed by ordinary differential equations is proved in [7].

The paper is organized as follows. In section 2 the optimality system for control-constrained problems is given in a form that is suggestive for the proposed algorithm. Section 3 contains the formulation of the algorithm and its basic properties. The global convergence analysis is presented in section 4, and numerical results are given in section 5.

2. Regularization and optimality system. In this section an optimality system for (1.1)–(1.3) that will be suggestive for the proposed algorithm is derived. The optimal control problem (1.1)–(1.3) can be equivalently expressed as

$$\text{(P)} \qquad \begin{cases} \min\ F(y,u) + I_{U_{ad}}(u), \\ u \in L^2(\Sigma), \quad y = Su + f, \end{cases}$$

where $I_{U_{ad}}\colon L^2(\Sigma) \to \mathbb{R} \cup \{\infty\}$ is the indicator function of U_{ad} given by

$$I_{U_{ad}}(u) = \begin{cases} 0 & \text{if } u \in U_{ad}, \\ \infty & \text{otherwise.} \end{cases}$$

In contrast to (1.1)–(1.3), problem (P) does not contain explicit inequality constraints. The cost functional, however, is not differentiable.

We next regularize (P) by smoothing $I_{U_{ad}}$ appropriately. Following [9], we define for a proper, lower semicontinuous function $\varphi : H \mapsto \mathbb{R} \cup \{\infty\}$ the generalized Moreau–

Yosida approximation

$$\varphi_c(u,\lambda) = \inf_{v\in H}\left\{\varphi(u-v) + (\lambda,v)_H + \frac{c}{2}\|v\|_H^2\right\}.$$

Here H denotes a Hilbert space, $(\cdot,\cdot)_H$ stands for the inner product in H, $\lambda \in H$, and $c > 0$. It is known that the mapping $u \mapsto \varphi_c(u,\lambda)$ is differentiable, with Lipschitz continuous derivative given by

$$\varphi_c'(u,\lambda) = A_{1/c}\left(u + \frac{1}{c}\lambda\right),$$

where the prime denotes the differentiation with respect to u. The operator A_γ, $\gamma > 0$, denotes the Yosida approximation to the subdifferential $\partial\varphi = A$ given by $A_\gamma = \frac{1}{\gamma}(I - J_\gamma)$, with $J_\gamma = (I + \gamma A)^{-1}$ (see [5]).

We apply this procedure for $H = L^2(\Sigma)$ and $\varphi = I_{U_{ad}}$. To characterize the derivative of $I_{U_{ad}}$, we introduce the projection $P : \mathbb{R} \mapsto \mathbb{R}$ by

$$P(r) = \begin{cases} r & \text{if } a \le r \le b, \\ a & \text{if } r < a, \\ b & \text{if } r > b. \end{cases}$$

LEMMA 1. *For every $\lambda \in L^2(\Sigma)$ and $c > 0$ the derivative of $I_{U_{ad},c}$ is given by*

$$I'_{U_{ad},c}(u,\lambda) = c\left(u + \frac{1}{c}\lambda - P\left(u + \frac{1}{c}\lambda\right)\right).$$

For a proof we refer to Bertsekas [3] or to [9], for instance.

To derive a first form of the optimality system, we formulate the regularized problem with $\lambda = 0$ and $c > 0$:

$$(\mathrm{P}_c) \qquad \begin{cases} \min\ F(y,u) + I_{U_{ad},c}(u,0), \\ u \in L^2(\Sigma), \quad y = Su + f. \end{cases}$$

Clearly (P_c) has a unique solution $(y_c, u_c) \in Y \times U$ for every $c > 0$. Identifying the dual space of Y with itself, we denote by $S^* : Y \to U$ the adjoint of S. The *adjoint state* $p \in U$ is defined by

$$p = S^*(y_d - y). \tag{2.1}$$

Due to the differentiability of $I_{U_{ad},c}$, the optimality system for (P_c) is found to be

$$\left.\begin{array}{rcl} y_c &=& Su_c + f, \\ p_c &=& S^*(y_d - y_c), \\ p_c &=& \mu u_c + c(u_c - P(u_c)). \end{array}\right\} \tag{2.2}$$

Next we pass to the limit as $c \to \infty$. Since $\mu > 0$, the set $\{\|u_c\|_U\}_{c\ge 1}$ is bounded. Consequently, $\{\|y_c\|_Y\}_{c>0}$ and $\{\|p_c\|_U\}_{c>0}$ are precompact in Y and U. Hence there exist $(y^*, p^*) \in Y \times U$ such that

$$(y_c, p_c) \to (y^*, p^*) \qquad \text{in } Y \times U \tag{2.3}$$

on a subsequence with respect to c, which is again denoted by c.

In the following lemma it is shown that $\{u_c\}_{c>0}$ is a Cauchy sequence in $L^2(\Sigma)$, provided that $\{p_c\}_{c>0}$ is a Cauchy sequence in $L^2(\Sigma)$. This implies the existence of $u^* \in L^2(\Sigma)$ such that $u_c \to u^*$ in $L^2(\Sigma)$. For the statement and proof of the lemma we assume that $-\infty < a < b < \infty$. The cases in which $a = -\infty$ or $b = \infty$ can be treated with minor modifications.

LEMMA 2. *Let $0 < c < d$. Then*

$$\|u_c - u_d\|_{L^2(\Sigma)} \le \frac{1}{\mu}\|p_c - p_d\|_{L^2(\Sigma)} + \frac{1}{c+\mu}(\|p_c\|_{L^2(\Sigma)} + \mu \max(|a|,|b|)).$$

Proof. Note that, as a consequence of (2.2),

$$u_c(x) = \begin{cases} \frac{1}{\mu}p_c(x) & \text{if } a \le u_c(x) \le b \quad (\mu a \le p_c(x) \le \mu b), \\ \frac{p_c(x)+ca}{\mu+c} & \text{if } u_c(x) < a \qquad (p_c(x) < \mu a), \\ \frac{p_c(x)+cb}{\mu+c} & \text{if } u_c(x) > b \qquad (p_c(x) > \mu b), \end{cases} \tag{2.4}$$

pointwise a.e. We define the real valued function

$$h_c(z) = \begin{cases} \frac{z}{\mu} & \text{if } \mu a \le z \le \mu b, \\ \frac{z+ca}{\mu+c} & \text{if } z < \mu a, \\ \frac{z+cb}{\mu+c} & \text{if } z > \mu b \end{cases}$$

and observe that

$$|h_d(z_1) - h_d(z_2)| \le \frac{1}{\mu}|z_1 - z_2|$$

for all z_1, z_2. Furthermore, we find

$$|h_c(z_1) - h_d(z_1)| \le \frac{d-c}{(\mu+c)(\mu+d)} \min(|z_1 - \mu a|, |z_1 - \mu b|).$$

These inequalities imply the estimate

$$|h_c(z_1) - h_d(z_2)| \le \frac{1}{\mu}|z_1 - z_2| + \frac{1}{\mu+c} \min(|z_1 - \mu a|, |z_1 - \mu b|).$$

Using $\min(|z_1 - \mu a|, |z_1 - \mu b|) \le |z_1| + \mu \max(|a|,|b|)$, the assertion follows. ∎

We next pass to the limit in (2.2) and obtain

$$\left.\begin{array}{rcl} y^* &=& Su^* + f, \\ p^* &=& S^*(y_d - y^*), \\ p^* &=& \mu u^* + \lambda^*, \end{array}\right\} \tag{2.5}$$

where λ^* is defined through $c(u_c - P(u_c)) \to \lambda^*$ in $L^2(\Sigma)$ for $c \to \infty$. An easy investigation shows that the Lagrange multiplier λ^* satisfies the following properties:

$$\begin{array}{lll} \lambda^*(t,x) = 0 & \text{if} & \mu a \le p^*(t,x) \le \mu b, \\ \lambda^*(t,x) < 0 & \text{if} & p^*(t,x) < \mu a, \\ \lambda^*(t,x) > 0 & \text{if} & p^*(t,x) > \mu b. \end{array}$$

Similarly one verifies, utilizing (2.4), that

$$\left.\begin{array}{lll} \lambda^*(x) = 0 & \text{if} & a < u^*(x) < b, \\ \lambda^*(x) \le 0 & \text{if} & u^*(x) = a, \\ \lambda^*(x) \ge 0 & \text{if} & u^*(x) = b. \end{array}\right\} \tag{2.6}$$

Note that (2.6) can be expressed as $\lambda^* \in \partial I_{U_{ad},c}$, where ∂ stands for the subdifferential. This differential inclusion is not a convenient starting point for the construction of a numerical algorithm. By a general result of convex analysis (see, e.g., [9]), the differential inclusion can equivalently be expressed as an equation for λ^* given by

$$\lambda^* = I'_{U_{ad},c}(u^*, \lambda^*)$$

for each $c > 0$. In view of Lemma 1, this is equivalent to

$$\lambda^* = c\left(u^* + \frac{1}{c}\lambda^* - P\left(u^* + \frac{1}{c}\lambda^*\right)\right) \tag{2.7}$$

for each $c > 0$. Note that the equivalence between (2.6) and (2.7) can also be verified by a direct computation. The optimality system for (P) can now be specified.

PROPOSITION 1. *A necessary and sufficient condition for the optimality of u^* with associated state y^* is the existence of $(p^*, \lambda^*) \in U \times U$ such that (2.5), (2.7) hold for $(u^*, y^*, p^*, \lambda^*)$.*

Proof. Let us observe that

$$I_{U_{ad},c}(x, 0) = \begin{cases} \frac{c}{2}(x-b)^2 & \text{if } x > b, \\ 0 & \text{if } a \le x \le b, \\ \frac{c}{2}(x-a)^2 & \text{if } x < a. \end{cases} \tag{2.8}$$

Utilizing this fact, we argue that the limit u^* of every convergent subsequence of solutions $\{u_c\}$ to (P_c) is a solution of (P). In fact,

$$F(y_c, u_c) + I_{U_{ad},c}(u_c) \le F(\bar{y}, \bar{u})$$

for every c, where $\bar{u}$ denotes the solution to (P) and $\bar{y} = S\bar{u} + f$. Consequently, $I_{U_{ad},c}(u_c)$ is bounded, and by (2.8), therefore, $u^* \in U_{ad}$. Continuity of F implies that

$$F(y^*, u^*) \le \liminf_{c\to\infty} F(y_c, u_c) + I_{U_{ad},c}(u_c) \le F(\bar{y}, \bar{u}),$$

where $y^* = Su^* + f$. It follows that u^* is a solution to (P), and by uniqueness, $\bar{u} = u^*$ and $\bar{y} = y^*$. The fact that (2.5), (2.7) represents a necessary optimality condition for (P) now follows from the arguments before the statement of Proposition 1. The uniqueness of $(u^*, y^*, p^*, \lambda^*)$ is due to the fact that the solution u^* to (P) is unique and from the optimality system itself. The sufficiency of (2.5), (2.7) follows from a direct computation, which is left to the reader. ∎

3. Presentation of the algorithm. In this section we present the primal-dual active set algorithm and discuss some of its basic properties. For this purpose we introduce the active and inactive sets for the solution to (P) and define

$$A_+^* = \{x \in \Sigma :\ u^*(x) = b\}, \qquad A_-^* = \{x \in \Sigma :\ u^*(x) = a\},$$

$$\text{and} \quad I^* = \{x \in \Sigma :\ a < u^*(x) < b\}.$$

Here and below, the set theoretic definitions are understood in the almost everywhere sense. The proposed strategy is based on the multiplier rule (2.7). Given (u_{n-1}, λ_{n-1}), the active sets for the new iterate are chosen according to

$$A_n^+ = \left\{x \in \Sigma :\ u_{n-1}(x) + \frac{\lambda_{n-1}(x)}{c} > b\right\}, \tag{3.1}$$

$$A_n^- = \left\{x \in \Sigma :\ u_{n-1}(x) + \frac{\lambda_{n-1}(x)}{c} < a\right\}. \tag{3.2}$$

We recall that $\lambda^* \geq 0$ on A_+^*, and $\lambda^* \leq 0$ on A_-^*. The update strategies for A_n^+ and A_n^- are the key ingredients of the proposed algorithm. The complete algorithm is specified next.

ALGORITHM.

1. Initialization: Choose u_0 and λ_0, and set $n = 1$.
2. Determine the active sets according to (3.1), (3.2), and set $I_n = \Sigma \setminus (A_n^+ \cup A_n^-)$.
3. If $n \geq 2$, $A_n^+ = A_{n-1}^+$, $A_n^- = A_{n-1}^-$, and $I_n = I_{n-1}$, then STOP.
4. Else, find $(y_n, p_n) \in Y \times U$ such that

$$y_n = Su_n + f,$$
$$p_n = S^*(y_d - y_n),$$

where

$$u_n(x) = \begin{cases} b & \text{if } x \in A_n^+, \\ a & \text{if } x \in A_n^-, \\ \frac{p_n}{\mu} & \text{if } x \in I_n. \end{cases}$$

5. Set $\lambda_n = p_n - \mu u_n$, update $n := n + 1$, and goto 2.

Numerical experiments show that the number of iterations required by the algorithm before it stops in step 3 is rather insensitive to the choice of u_0. Among other possibilities, we tested $u_0 = 0$ or $u_0 = b$, as well as u_0 the solution to the unconstrained problem. We point out that the iterates of the algorithm satisfy the complementarity condition,

$$(u_n - a)(u_n - b)\lambda_n = 0 \quad \text{a.e. on } \Sigma,$$

for every $n \geq 1$ and, in the case of unilateral constraints, the active set strategy is independent of c as soon as $n \geq 2$.

Note that the system of step 4 constitutes the first order optimality condition for

$$\begin{cases} \min F(y, u), \\ u \in U, \quad y = Su + f, \\ u = b \text{ on } A_n^+, \quad u = a \text{ on } A_n^-, \end{cases}$$

and hence existence of a unique solution to the equations in 4 follows. The stopping rule of step 3 is justified next.

THEOREM 1. *If there exists an index $n \geq 1$ such that $A_n^+ = A_{n+1}^+$, $A_n^- = A_{n+1}^-$, and $I_n = I_{n+1}$, then the algorithm stops and the last iterate satisfies*

$$y_n = Su_n + f, \tag{3.3}$$

$$p_n = S^*(y_d - y_n), \tag{3.4}$$

$$u_n = \begin{cases} b & \text{if } x \in A_n^+, \\ a & \text{if } x \in A_n^-, \\ \frac{p_n}{\mu} & \text{if } x \in I_n, \end{cases} \tag{3.5}$$

$$\lambda_n = p_n - \mu u_n, \quad u_n \in U_{ad}, \tag{3.6}$$

with

$$\lambda_n = 0 \text{ on } I_n, \qquad \lambda_n > 0 \text{ on } A_n^+, \qquad \text{and } \lambda_n < 0 \text{ on } A_n^-. \tag{3.7}$$

Therefore the last iterate $(u_n, y_n, p_n, \lambda_n)$ is the solution of the original optimality system (2.5)–(2.7).

Proof. If the assumptions are fulfilled, then the algorithm stops due to step 3. Moreover, the last iterate satisfies (3.3)–(3.6) by construction except possibly for $u_n \in U_{ad}$. Thus we have to prove $u_n \in U_{ad}$ and (3.7). On I_n we have $\lambda_n = 0$ due to step 5 of the algorithm. Moreover, $a \le u_n + \frac{\lambda_n}{c} = u_n \le b$, since $I_n = I_{n+1}$. On A_n^+ we get $u_n = b$ and $u_n + \frac{\lambda_n}{c} > b$ since $A_n^+ = A_{n+1}^+$. Therefore $\lambda_n > 0$ on A_n^+.

Analogously one argues that $\lambda_n < 0$ on A_n^-. It follows that $u_n \in U_{ad}$ and (3.7) holds. Finally, (3.7) together with the definition of u_n yields (2.7), with (λ^*, u^*) replaced by (λ_n, u_n), and $(u_n, y_n, p_n, \lambda_n)$ satisfies the optimality system (2.5), (2.7). ∎

Before demonstrating the extreme efficiency of the proposed algorithm, we analyze its convergence in the following section. For a general purpose code the algorithm must be completed with a stopping criterion. In all the numerical examples that we tested for parabolic control problems, some of which we present in section 5, the algorithm self-terminated in step 3. Section 4 contains a sufficient condition for an appropriately chosen merit function. If the control space discretized as finite dimensional space, then this can be utilized to argue that the algorithm stops in step 3.

4. Convergence analysis. In this section we show that an appropriately chosen augmented Lagrangian functional acts as a merit function for the proposed algorithm. For technical simplicity we restrict ourselves to the unilateral case and choose

$$U_{ad} = \{u \in L^2(\Sigma) : \ u \le b\}.$$

Consequently there is only one active set $A_n = A_n^+$. For $c > 0$ we define the augmented Lagrange functionals

$$\begin{aligned} L_c(y, u, \lambda) &= F(y, u) + (\lambda, g_c(u, \lambda))_{L^2(\Sigma)} + \tfrac{c}{2}\|g_c(u, \lambda)\|^2_{L^2(\Sigma)}, \\ L_c^+(y, u, \lambda) &= L_c(y, u, \lambda^+), \end{aligned}$$

where $\lambda^+ = \max(0, \lambda)$ and $g_c(u, \lambda) = \max(g(u), -\frac{\lambda}{c})$, with $g(u) = u - b$. For convenience we henceforth replace $\|\cdot\|_{L^2(\Sigma)}$ by $\|\cdot\|_\Sigma$.

The proof of the following technical lemma can be obtained by a straightforward adaptation of the proof of Lemma 3.1 of [9], for example.

LEMMA 3. *For all pairs* (u, y) *fulfilling* (1.2) *we have*

$$F(y_n, u_n) - F(y, u) = -\frac{1}{2}\|y - y_n\|_Y^2 - \frac{\mu}{2}\|u - u_n\|_\Sigma^2 + (\lambda_n, u - u_n)_{A_n}. \tag{4.1}$$

Let us define the sets

$$S_{n-1} = \{x \in A_{n-1} : \ \lambda_{n-1}(x) \le 0\} \quad \text{and} \quad T_{n-1} = \{x \in I_{n-1} : \ u_{n-1}(x) > b\}.$$

S_{n-1} is that part of the currently active set which in the next iteration will be set inactive according to the sign of the Lagrange multiplier. Similarly, T_{n-1} is that part of the inactive set which will be set active in the next step of the iteration. We note that

$$\Sigma = (I_{n-1} \setminus T_{n-1}) \cup T_{n-1} \cup S_{n-1} \cup (A_{n-1} \setminus S_{n-1}) \tag{4.2}$$

defines a decomposition of Σ into mutually disjoint subsets, and, moreover, that

$$I_n = (I_{n-1} \setminus T_{n-1}) \cup S_{n-1}, \qquad A_n = T_{n-1} \cup (A_{n-1} \setminus S_{n-1}). \tag{4.3}$$

Moreover, we define

$$I_n^* = (I_{n-1} \setminus T_{n-1}) \cup \{x \in S_{n-1} : \ u_n(x) \geq b\}.$$

In the following theorem, a sufficient condition for the decay of L_c^+ is established.

THEOREM 2. *If*

$$\mu + \gamma < c < \mu - \frac{\mu^2}{\gamma} + \frac{\mu^2}{\|S^*\|^2} \tag{4.4}$$

for some $\gamma > 0$, then $L_c^+(y_n, u_n, \lambda_n) < L_c^+(y_{n-1}, u_{n-1}, \lambda_{n-1})$ or $(y_n, u_n) = (y_{n-1}, u_{n-1})$ and the algorithm stops.

Proof. The proof of Theorem 2 requires only small modifications of Theorem 3.1 in [1]. □

Remark. For our further investigations we need two inequalities, which we cite from [1]. Analogously to inequality (3.8) in [1], we obtain

$$\|u_n - u_{n-1}\|_{I_n^*} \leq \frac{\|S^*\|}{\mu} \|y_n - y_{n-1}\|_Y. \tag{4.5}$$

Moreover, in the proof of Theorem 3.1 of [1] it is verified that

$$L_c^+(y_n, u_n, \lambda_n) - L_c^+(y_{n-1}, u_{n-1}, \lambda_{n-1}) \leq -c_Y \|y_{n-1} - y_n\|_Y^2 - c_U \|u_{n-1} - u_n\|_{T_{n-1}}^2 \tag{4.6}$$

with certain positive constants c_Y and c_U.

Note that, for the choice $\gamma = \mu$, condition (4.4) is equivalent to

$$2\mu < c < \frac{\mu^2}{\|S^*\|^2}.$$

So far we have given a sufficient condition for L_c^+ to act as a merit function for which the algorithm has a strict descent property. In particular, this eliminates the possibility of chattering of the algorithm: It will not return to the same active set a second time. Our next aim is to prove strong convergence of $\{u_n, y_n, p_n, \lambda_n\}$ to $\{u^*, y^*, p^*, \lambda^*\}$.

COROLLARY 1. *If (4.4) holds, the sequences $\{u_n\}$, $\{p_n\}$, $\{\lambda_n\}$ are bounded in $L^2(\Sigma)$, and $\{y_n\}$ is bounded in Y.*

Proof. Let $n \geq 2$. We first investigate the term $g_c(u_{n-1}, \lambda_{n-1}^+) = \max(u_{n-1} - b, -\lambda_{n-1}^+/c)$. Since $(u_{n-1} - b)\lambda_{n-1} = 0$ pointwise a.e., we have $g_c(u_{n-1}, \lambda_{n-1}^+) \geq 0$.

Clearly, $g_c(u_{n-1}, \lambda_{n-1}^+) > 0$ if and only if $u_{n-1} > b$, which is the case if and only if $x \in T_{n-1}$. It follows that $g_c(u_{n-1}, \lambda_{n-1}^+) = u_{n-1} - b = u_{n-1} - u_n$ and

$$(\lambda_{n-1}^+, g_c(u_{n-1}, \lambda_{n-1}^+))_\Sigma + \frac{c}{2} \|g_c(u_{n-1}, \lambda_{n-1}^+)\|_\Sigma^2 = \frac{c}{2} \|u_{n-1} - u_n\|_{T_{n-1}}^2.$$

As a consequence we obtain

$$L_c^+(y_{n-1}, u_{n-1}, \lambda_{n-1}) = F(y_{n-1}, u_{n-1}) + \frac{c}{2} \|u_{n-1} - u_n\|_{T_{n-1}}^2 \geq 0.$$

This inequality, the structure of F, and the decrease of the merit function imply the boundedness of the sequence $\{u_n\}$. From this fact the assertion follows easily. □

COROLLARY 2. *With* (4.4) *holding, the sequence* $\|u_n - u_{n-1}\|_\Sigma$ *tends to* 0 *for* $n \to \infty$.

Proof. We decompose Σ according to

$$\Sigma = I_n^* \cup T_{n-1} \cup (A_{n-1} \setminus S_{n-1}) \cup S_{n-1}^-, \tag{4.7}$$

where $S_{n-1}^- = \{x \in S_{n-1} : u_n(x) < b\}$. Without loss of generality we assume that $n \geq 3$. On $A_{n-1} \setminus S_{n-1}$ we have $u_n = u_{n-1} = b$, and therefore $\|u_n - u_{n-1}\|_{A_{n-1}\setminus S_{n-1}} = 0$.

To investigate the set S_{n-1}^-, note that $S_{n-1}^- \subset S_{n-1} \subset A_{n-1}$ and hence $u_{n-2} + \frac{\lambda_{n-2}}{c} > b$. If $u_{n-2} = b$, this implies $\lambda_{n-2} = p_{n-2} - \mu u_{n-2} > 0$, and therefore $\frac{p_{n-2}}{\mu} > b$. Otherwise, we have $\lambda_{n-2} = 0$ and $u_{n-2} = \frac{p_{n-2}}{\mu} > b$. For this reason the inequality $\frac{p_{n-2}}{\mu} > b$ holds on S_{n-1}^-. We continue the estimation on the set S_{n-1}^-:

$$\|u_n - u_{n-1}\|_{S_{n-1}^-} = \|b - u_n\|_{S_{n-1}^-} \leq \left\| \frac{p_{n-2}}{\mu} - \frac{p_n}{\mu} \right\|_{S_{n-1}^-} \leq \frac{\|S^*\|}{\mu} \|y_{n-2} - y_n\|_Y.$$

In the following estimate, we use (4.5) and (4.7):

$$\begin{aligned}
\|u_n - u_{n-1}\|_\Sigma^2 &= \|u_n - u_{n-1}\|_{I_n^*}^2 + \|u_n - u_{n-1}\|_{T_{n-1}}^2 \\
&\quad + \|u_n - u_{n-1}\|_{A_{n-1}\setminus S_{n-1}}^2 + \|u_n - u_{n-1}\|_{S_{n-1}^-}^2 \\
&\leq \frac{\|S^*\|^2}{\mu^2} \|y_n - y_{n-1}\|_Y^2 + \|u_n - u_{n-1}\|_{T_{n-1}}^2 + \frac{\|S^*\|^2}{\mu^2} \|y_{n-2} - y_n\|_Y^2.
\end{aligned}$$

From (4.6) we deduce the existence of a constant c_1 such that

$$\frac{\|S^*\|^2}{\mu^2} \|y_n - y_{n-1}\|_Y^2 + \|u_n - u_{n-1}\|_{T_{n-1}}^2 \leq c_1 (L_c^+(y_{n-1}, u_{n-1}, \lambda_{n-1}) - L_c^+(y_n, u_n \lambda_n))$$

for all $n \geq 2$, and consequently

$$\|u_n - u_{n-1}\|_\Sigma^2 \leq c_1 (L_c^+(y_{n-1}, u_{n-1}, \lambda_{n-1}) - L_c^+(y_n, u_n, \lambda_n)) + \frac{\|S^*\|^2}{\mu^2} \|y_{n-2} - y_n\|_Y^2.$$

Moreover, there also exists a positive c_2 such that

$$\begin{aligned}
\frac{\|S^*\|}{\mu} \|y_{n-2} - y_n\|_Y &\leq c_2 \sqrt{L_c^+(y_{n-2}, u_{n-2}, \lambda_{n-2}) - L_c^+(y_{n-1}, u_{n-1}, \lambda_{n-1})} \\
&\quad + c_2 \sqrt{L_c^+(y_{n-1}, u_{n-1}, \lambda_{n-1}) - L_c^+(y_n, u_n, \lambda_n)}.
\end{aligned}$$

Inserting this estimate into the previous one, we obtain

$$\begin{aligned}
\|u_n - u_{n-1}\|_\Sigma^2 &\leq c_1 (L_c^+(y_{n-1}, u_{n-1}, \lambda_{n-1}) - L_c^+(y_n, u_n, \lambda_n)) \\
&\quad + 2c_2^2 (L_c^+(y_{n-2}, u_{n-2}, \lambda_{n-2}) - L_c^+(y_n, u_n, \lambda_n)).
\end{aligned}$$

The right-hand side of this inequality tends to zero as $n \to \infty$ since the merit function decreases and is bounded below. ∎

COROLLARY 3. *If* (4.4) *is satisfied, then the sequence* $\|u_{n-1} - P(u_{n-1} + \frac{1}{c}\lambda_{n-1})\|_\Sigma$ *tends to* 0 *as* $n \to \infty$.

Proof. We define on Σ the function $q := \lambda_{n-1} - c(u_{n-1} + \frac{1}{c}\lambda_{n-1} - P(u_{n-1} + \frac{1}{c}\lambda_{n-1})) = -c(u_{n-1} - P(u_{n-1} + \frac{1}{c}\lambda_{n-1}))$. On $A_{n-1} \setminus S_{n-1}$ we have $u_{n-1} = b$ and

$\lambda_{n-1} \geq 0$, and therefore $q = 0$. On $I_{n-1} \setminus T_{n-1}$ we find $u_{n-1} \leq b$ and $\lambda_{n-1} = 0$, and consequently $q = 0$. On S_{n-1} we have $u_{n-1} = b$ and $\lambda_{n-1} < 0$. We therefore obtain $P(u_{n-1} + \frac{1}{c}\lambda_{n-1}) = u_{n-1} + \frac{1}{c}\lambda_{n-1}$, and therefore $q = \lambda_{n-1} = p_{n-1} - \mu u_{n-1}$. Using $p_n = \mu u_n$, we find

$$\begin{aligned}\|q\|_{S_{n-1}} &= \|p_{n-1} - \mu u_{n-1}\|_{S_{n-1}} \\ &= \|p_{n-1} - p_n - \mu(u_{n-1} - u_n)\|_{S_{n-1}} \\ &\leq \|p_{n-1} - p_n\|_{S_{n-1}} + \mu\|(u_{n-1} - u_n)\|_{S_{n-1}} \\ &\leq \|S^*\|\|y_{n-1} - y_n\|_Y + \mu\|(u_{n-1} - u_n)\|_\Sigma.\end{aligned}$$

On the remaining set T_{n-1} we have $\lambda_{n-1} = 0$ and $u_{n-1} > b$, and thus $q = -c(u_{n-1} - b) = -c(u_{n-1} - u_n)$. Consequently, we obtain

$$\|q\|_{T_{n-1}} = c\|u_{n-1} - u_n\|_{T_{n-1}} \leq c\|(u_{n-1} - u_n)\|_\Sigma.$$

Summarizing these estimates, we find

$$\|q\|_\Sigma^2 \leq (\|S^*\|\|y_{n-1} - y_n\|_Y + \mu\|(u_{n-1} - u_n)\|_\Sigma)^2 + c^2\|(u_{n-1} - u_n)\|_\Sigma^2.$$

By Corollary 2, the right-hand side of this inequality tends to 0, and the assertion follows. □

THEOREM 3. *If* (4.4) *holds, then the sequence* $\{u_n,\ y_n,\ p_n,\ \lambda_n\}$ *converges in* $L^2(\Sigma) \times Y \times L^2(\Sigma) \times L^2(\Sigma)$ *to the optimal solution* $(u^*,\ y^*,\ p^*,\ \lambda^*)$.

Proof. By Corollary 1, the sequence $\{u_n, y_n, p_n, \lambda_n\}$ is bounded in $L^2(\Sigma) \times Y \times L^2(\Sigma) \times L^2(\Sigma)$, and therefore there exists a subsequence, denoted by the same symbol, that converges weakly to some element $(\bar{u}, \bar{y}, \bar{p}, \bar{\lambda})$ in $L^2(\Sigma) \times Y \times L^2(\Sigma) \times L^2(\Sigma)$. From Corollary 3 we know that $\|u_n - P(u_n + \frac{1}{c}\lambda_n)\|_\Sigma \to 0$. We investigate the projection term $P(u_n + \frac{1}{c}\lambda_n)$. If $\lambda_n = 0$, then $u_n = \frac{p_n}{\mu}$, and

$$u_n - P\left(u_n + \frac{1}{c}\lambda_n\right) = u_n - P\left(\frac{p_n}{\mu}\right). \tag{4.8}$$

Otherwise, we have $u_n = b$ and $\lambda_n = p_n - \mu b$. If $\frac{p_n}{\mu} \geq b$, then we get

$$u_n - P\left(u_n + \frac{1}{c}\lambda_n\right) = u_n - P\left(\frac{p_n}{\mu}\right). \tag{4.9}$$

It remains to investigate the case $\frac{p_n}{\mu} < b$. We find $P(u_n + \frac{1}{c}\lambda_n) = P(b + \frac{p_n - \mu b}{c}) = b + \frac{p_n - \mu b}{c}$. Inserting $u_n = b$, it follows that

$$u_n - P\left(u_n + \frac{1}{c}\lambda_n\right) = \frac{\mu b - p_n}{c} = \frac{\mu}{c}\left(u_n - \frac{p_n}{\mu}\right),$$

and consequently

$$u_n - P\left(u_n + \frac{1}{c}\lambda_n\right) = \frac{\mu}{c}\left(u_n - P\left(\frac{p_n}{\mu}\right)\right). \tag{4.10}$$

Combining (4.8)–(4.10), we find

$$\left\|\left(u_n - P\left(\frac{p_n}{\mu}\right)\right)\right\|_\Sigma \leq \varrho\left\|u_n - P\left(u_n + \frac{1}{c}\lambda_n\right)\right\|_\Sigma, \tag{4.11}$$

where $\varrho = \max(1, \frac{c}{\mu})$. By Corollary 3, this implies $\|(u_n - P(\frac{p_n}{\mu}))\|_\Sigma \to 0$. The compactness of S and the weak convergence of $u_n \to \bar{u}$ imply the strong convergence of $p_n \to \bar{p}$. Consequently $u_n \to \bar{u}$ strongly, and in addition

$$\bar{u} = P\left(\frac{\bar{p}}{\mu}\right). \tag{4.12}$$

From the strong convergence of u_n and p_n we easily deduce the strong convergence of λ_n and y_n in $L^2(\Sigma)$ and Y. The iterates $(u_n, y_n, p_n, \lambda_n)$ satisfy (2.5). Passing to the limit, these equations are also true for $(\bar{u}, \bar{y}, \bar{p}, \bar{\lambda})$. Using the strong convergence of $\{u_n\}$ and $\{\lambda_n\}$, and the fact that $\|u_n - P(u_n + \frac{1}{c}\lambda_n)\|_\Sigma \to 0$, we obtain $\bar{u} = P(\bar{u} + \frac{1}{c}\bar{\lambda})$, which is equivalent to

$$\bar{\lambda} = c\left(\bar{u} + \frac{1}{c}\bar{\lambda} - P\left(\bar{u} + \frac{1}{c}\bar{\lambda}\right)\right). \tag{4.13}$$

For that reason, $(\bar{u}, \bar{y}, \bar{p}, \bar{\lambda})$ is the unique solution of the optimality system (2.5), (2.7). ∎

If the control space is discretized, then the descent property of Theorem 2 can be used to argue convergence in a finite number of steps. We assume that the control space is approximated by a system of functions that match to the control constraints, in the sense that the pointwise bounds of U_{ad} can be directly imposed on the coefficients in the expansion of u as is the case, for instance, for piecewise constant or piecewise linear functions. We only discretize with respect to the control u and denote the resulting semidiscretized problems by (P_k).

COROLLARY 4. *Assume that* (4.4) *holds and that the algorithm is discretized as mentioned above. Then the solution of the semidiscretized problem* (P_k) *is obtained in finitely many steps.*

Proof. Because of the discretization of the control u, there are only a finite number of possibilities for the sets A_n and I_n. Theorem 2 implies that the algorithm does not return to the same set except in the case in which it stops in step 3 and the solution is found. Since there are only finitely many configurations of active and inactive sets, this situation must occur after finitely many steps. ∎

5. Applications and numerical experiments. The general formulation (1.1)–(1.3) is applicable to a diverse class of optimal control problems with pointwise constraints on the controls.

The assumption of compactness of the control-to-observation mapping is not restrictive and can be established for many specific cases [15]. The case of elliptic optimal control problems was considered in [1]. Here we argue the applicability of our results and demonstrate the efficiency of the algorithm for constrained parabolic control problems of the type

$$\min \quad \frac{1}{2}\|y(T,\cdot) - y_d\|^2_{L^2(\Omega)} + \frac{\mu}{2}\|u\|^2_{L^2(\Sigma)} \tag{5.1}$$

subject to

$$\left.\begin{aligned} y_t &= \Delta_x y \quad \text{in } Q,\\ \tfrac{\partial y}{\partial n} &= u \quad \text{on } \Sigma,\\ y(0,.) &= 0 \quad \text{in } \Omega,\\ a \le u &\le b, \end{aligned}\right\} \tag{5.2}$$

where $y_d \in L_2(\Omega)$, $Q = (0,T) \times \Omega$, and Ω is a bounded domain in $\mathbb{R}^m$ with sufficient smooth boundary $\partial\Omega$. The general theory can be applied with $\Sigma = (0,T) \times \partial\Omega$, $Y = L^2(\Omega)$, $f = 0$, and $Su = y(T,\cdot)$, where, for given $u \in L^2(\Sigma)$, we denote by y the unique solution to (5.2). The operator S is compact. Indeed, for every $u \in L^2(\Sigma)$, (5.2) admits a unique solution $y \in L^2(0,T;H^1(\Omega)) \cap C(0,T;L^2(\Omega))$ with $y_t \in L^2(0,T,H^1(\Omega)^*)$. Moreover, for every $\varepsilon \in (0,\frac{1}{2})$ there exists a constant K_ε such that

$$\|y(T,\cdot)\|_{H^\varepsilon(\Omega)} \le K_\varepsilon \|u\|_{L_2(\Sigma)} \qquad \text{for all } u \in L^2(\Sigma);$$

see [20], for instance. Since $H^\varepsilon(\Omega)$ embeds compactly into $L^2(\Omega)$ for $\varepsilon > 0$, the compactness of the operator S follows. For the numerical tests, we consider the case $a = -1$, $b = 1$, and $\Omega = (0,1) \subset \mathbb{R}$, with boundary condition of the form

$$\frac{\partial y}{\partial n}(t,0) = 0, \quad \frac{\partial y}{\partial n}(t,1) = u \quad \text{on } (0,T). \tag{5.3}$$

The implementation of the algorithm requires discretization of both the control- and the state space variables. For the sake of studying the behavior of the algorithm, we chose independent, uniform grids for the control- and state space variables. The discretization itself was carried out by finite differences using the standard three-point star approximation for the Laplacian and the implicit Euler method in time. The number of degrees of freedom is denoted by n_x for the spatial direction of y, and by n_u and n_t for the temporal directions of u and y. There are many possibilities for numerically realizing the system arising in step 4 of the algorithm. For testing, we proceeded by precomputing a matrix M_1, which describes the mapping of the control coordinates u_i to the solutions of the primal equation at time T, and a second matrix M_2, which assigns to all coordinates of possible right-hand sides in the adjoint equation their solutions on the control boundary. Once M_1 and M_2 are available, it is simple to solve the discretized form of the system arising in step 4 of the algorithm.

Example 1. Here y_d is chosen to be a discontinuous function

$$y_d = \begin{cases} 1 & \text{on } [0,0.5), \\ 0 & \text{on } [0.5,1]. \end{cases}$$

Note that y_d is not in the range of S. In Table 1 we give the number of iterations (It.) before the algorithm stops in step 3 for various choices of n_u, n_t, n_x, and μ. We observe that consistently only very few iterations are required before the algorithm stops in a mesh-independent number of iterations. Figure 1 contains the plot of the optimal control for $\mu = 0.1$ and $\mu = 0.01$. Let us emphasize that termination of the algorithm in step 3 implies that the exact solution of the discretized problem has been obtained.

Example 2. The data for this example are those of Example 1, except that $y_d = 1$. The results can be found in Table 2 and Figure 2 for $\mu = 0.001$.

Example 3. Here we consider

$$\min \quad \frac{1}{2}\|y(T,\cdot) - y_d\|^2_{L^2(\Omega)} + \frac{\mu}{2}\|u\|^2_{L^2(\Sigma)} \tag{5.4}$$

subject to

TABLE 1

μ	n_u	n_t	n_x	It.
0.1	50	150	100	3
0.1	100	300	100	3
0.1	100	300	200	3
0.01	50	150	100	4
0.01	100	300	100	4
0.01	100	300	200	4
0.001	50	150	100	4
0.001	100	300	100	3
0.001	100	300	200	3

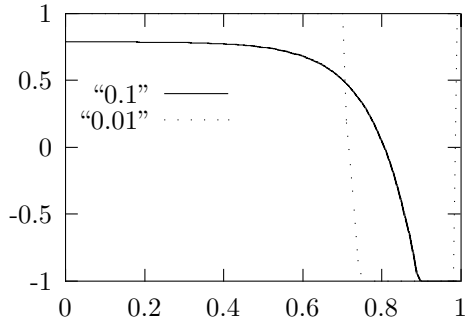


FIG. 1

TABLE 2

μ	n_u	n_t	n_x	It.
0.1	50	150	100	4
0.1	100	300	100	3
0.1	100	300	200	4
0.01	50	150	100	5
0.01	100	300	100	5
0.01	100	300	200	6
0.001	50	150	100	5
0.001	100	300	100	5
0.001	100	300	200	5

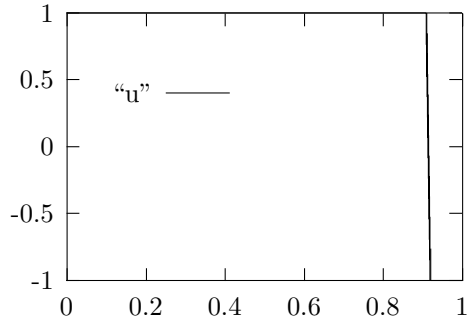


FIG. 2

$$\left.\begin{aligned} y_t &= \Delta_x y && \text{in } Q,\\ \tfrac{\partial y}{\partial n}(t,0) &= 0 && \text{on } [0,T],\\ \tfrac{\partial y}{\partial n}(t,1) &= u - y(t,1) && \text{on } [0,T],\\ y(0,x) &= 0 && \text{in } \Omega,\\ -1 \le u &\le 1, \end{aligned}\right\} \tag{5.5}$$

where $\Omega = (0,1)$ and $y_d = \frac{1}{2}(1-x^2)$. This example was first published in [16] and then also studied in [6]. Table 3 summarizes some of the results that we obtained with our algorithm. Again, convergence is achieved within very few iterations for a wide range of values for μ, n_u, n_t, n_x. Figure 3 contains the optimal control for $\mu = 0.001$.

In [6], several algorithms were tested for their behavior when applied to Example 3. The projection method described by Bertsekas [4] turned out to be the most efficient one with respect to the number of global iterations. Our algorithm never required more iterations than the Bertsekas algorithm. There are many similarities between the two algorithms; for a precise comparison, we refer to the discussion in [2]. Convergence properties of the Bertsekas algorithm applied to infinite dimensional problems are analyzed in [11], [12]. The convergence analysis in those papers is completely different from that in the present one.

Acknowledgment. The authors would like to thank the referees for their many helpful comments and criticism.

TABLE 3

μ	n_u	n_t	n_x	It.
0.01	100	200	200	4
0.001	50	150	100	4
0.001	50	150	200	5
0.001	100	200	200	5
0.001	100	300	200	5
0.0001	50	150	100	5
0.0001	50	150	200	6
0.0001	100	200	200	7
0.0001	100	300	200	8

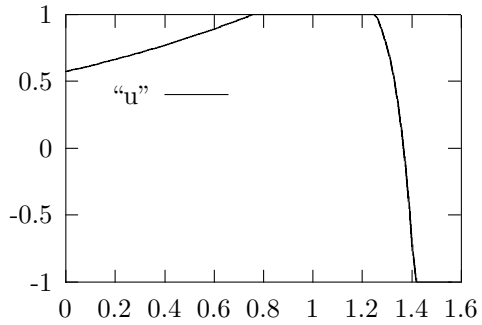


FIG. 3

REFERENCES

[1] M. BERGOUNIOUX, K. ITO, AND K. KUNISCH, *Primal-dual strategy for constrained optimal control problems,* SIAM J. Control Optim., 37 (1999), pp. 1176–1194.

[2] M. BERGOUNIOUX AND K. KUNISCH, *Primal-Dual Strategy for State-Constrained Optimal Control Problems,* Comput. Optim. Appl., 22 (2002), pp. 193–224.

[3] D. P. BERTSEKAS, *Constrained Optimization and Lagrange Multiplier Methods,* Academic Press, Paris, 1982.

[4] D. P. BERTSEKAS, *Projected Newton methods for optimization problems with simple constraints,* SIAM J. Control Optim., 20 (1982), pp. 221–246.

[5] H. BREZIS, *Operateurs Maximaux Monotones et Semi-groupes de Contractions dans les Espaces de Hilberts,* North–Holland, Amsterdam, 1973.

[6] H. GOLDBERG AND F. TRÖLTZSCH, *On a Lagrange-Newton method for a nonlinear parabolic boundary control problem,* Optim. Methods Softw., 8 (1998), pp. 225–247.

[7] W. W. HAGER AND G. D. IANULESCU, *Dual approximations in optimal control,* SIAM J. Control Optim., 22 (1984), pp. 423–465.

[8] M. HEINKENSCHLOSS AND F. TRÖLTZSCH, *Analysis of the Lagrange-SQP-Newton method for the control of a phase field equation,* Control Cybernet., 28 (1999), pp. 178–211.

[9] K. ITO AND K. KUNISCH, *Augmented Lagrangian methods for nonsmooth, convex optimization in Hilbert spaces,* Nonlinear Anal., 41 (2000), pp. 591–616.

[10] C. KELLEY AND E. SACHS, *Approximate quasi-Newton methods,* Math. Programming, 48 (1990), pp. 41–70.

[11] C. KELLEY AND E. SACHS, *Multilevel algorithms for constrained compact fixed point problems,* SIAM J. Sci. Comput., 15 (1994), pp. 645–667.

[12] C. T. KELLEY AND E. W. SACHS, *Solution of optimal control problems by a pointwise projected Newton method,* SIAM J. Control Optim., 33 (1995), pp. 1731–1757.

[13] C. T. KELLEY AND E. W. SACHS, *A trust region method for parabolic boundary control problems,* SIAM J. Optim., 9 (1999), pp. 1064–1081.

[14] K. KUNISCH AND E. W. SACHS, *Reduced SQP methods for parameter identification problems,* SIAM J. Numer. Anal., 29 (1992), pp. 1793–1820.

[15] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations,* Springer-Verlag, Berlin, 1971.

[16] K. SCHITTKOWSKI, *Numerical solution of a time-optimal parabolic boundary-value control problem,* J. Optim. Theory Appl., 27 (1979), pp. 271–290.

[17] K. SCHITTKOWSKI, *On the convergence of a sequential quadratic programming method with an augmented Lagrangian line search function,* Math. Operationsforsch. Statist. Ser. Optim., 14 (1983), pp. 197–216.

[18] T. TIAN AND J. C. DUNN, *On the gradient projection method for optimal control problems with nonnegative L^2 inputs,* SIAM J. Control Optim., 32 (1994), pp. 516–537.

[19] PH. L. TOINT, *Global convergence of a class of trust-region methods for nonconvex minimization in Hilbert space,* IMA J. Numer. Anal., 8 (1988), pp. 231–252.

[20] F. TRÖLTZSCH, *On convergence of semidiscrete Ritz–Galerkin schemes applied to the boundary control of parabolic equations with non-linear boundary conditions,* Z. Angew. Math. Mech., 72 (1992), pp. 291–301.

[21] F. TRÖLTZSCH, *An SQP method for the optimal control of a nonlinear heat equation,* Control Cybernet., 23 (1994), pp. 267–288.

[22] M. ULBRICH, S. ULBRICH, AND M. HEINKENSCHLOSS, *Global convergence of trust-region interior-point algorithms for infinite-dimensional nonconvex minimization subject to pointwise bounds,* SIAM J. Control Optim., 37 (1999), pp. 731–764.

SIAM J. OPTIM.
Vol. 13, No. 2, pp. 335–367

TRANSITIVE PACKING: A UNIFYING CONCEPT IN COMBINATORIAL OPTIMIZATION*

RUDOLF MÜLLER† AND ANDREAS S. SCHULZ‡

Abstract. This paper attempts to provide a better understanding of the facial structure of polyhedra previously investigated separately. It introduces the notion of transitive packing and the transitive packing polytope. Polytopes that turn out to be special cases of the transitive packing polytope include the node packing, acyclic subdigraph, bipartite subgraph, planar subgraph, clique partitioning, partition, transitive acyclic subdigraph, interval order, and relatively transitive subgraph polytopes. We give cutting plane proofs for several rich classes of valid inequalities of the transitive packing polytope, thereby introducing generalized cycle, generalized clique, generalized antihole, generalized antiweb, and odd partition inequalities. On the one hand, these classes subsume several known classes of valid inequalities for several special cases; on the other hand, they yield many new inequalities for several other special cases. For some of the classes we also prove a lower bound on their Gomory–Chvátal rank. Finally, we relate the concept of transitive packing to generalized (set) packing and covering, as well as to balanced and ideal matrices.

Key words. combinatorial optimization, polyhedral combinatorics, 0/1-polytope, Gomory–Chvátal cut, transitive packing, independence system

AMS subject classifications. 90C57, 90C27, 90C10

PII. S1052623499363256

1. Introduction. Various types of packing problems and related polyhedra play a central role in combinatorial optimization. Due to both a large variety of practical applications and their interesting structural properties, they have received considerable attention in the literature; see, e.g., [3, 43] for an overview. One of the classic examples is the *node packing problem* in graphs and the associated *node packing polytope*. (Alternative names are *vertex packing*, *stable set*, *coclique*, *anticlique*, or *independent set* problem and polytope, respectively.) The node packing problem on a finite, undirected, loopless graph G with node weights is the problem of finding a subset of mutually nonadjacent nodes such that the total weight of the selected subset is maximal. If we denote by A the edge-node incidence matrix of the graph G, it can be formulated as

$$\begin{aligned} &\text{maximize} && cx \\ &\text{subject to} && Ax \leqslant \mathbb{1}, \\ &&& x_u \in \{0,1\}, \end{aligned} \tag{1.1}$$

where c is an arbitrary vector of weights and $\mathbb{1}$ denotes (here and henceforth) the all-one vector of compatible dimension. The node packing polytope is defined as the

*Received by the editors October 19, 1999; accepted for publication (in revised form) February 4, 2002; published electronically September 12, 2002. This work is based on [49, Chapter 4]; an extended abstract appeared in [36].

http://www.siam.org/journals/siopt/13-2/36325.html

†Maastricht University, Faculty of Economics and Business Administration, P.O. Box 616, 6200 MD Maastricht, The Netherlands (r.muller@ke.unimaas.nl). This author was previously affiliated with Humboldt Universität Berlin.

‡Massachusetts Institute of Technology, Sloan School of Management, E53-361, 77 Massachusetts Avenue, Cambridge, MA 02139-4307 (schulz@mit.edu). This author was previously affiliated with the Department of Mathematics, Technische Universität Berlin. Most of this work was done while this author was supported by the Deutsche Forschungsgemeinschaft via the Graduate Program "Computational Discrete Mathematics," grant We 1265/2-1.

convex hull of feasible solutions to (1.1) and has been studied in, among other works, [26, 37, 42, 52].

The node packing problem can be extended to hypergraphs, where it reads

$$\begin{array}{ll} \text{maximize} & cx \\ \text{subject to} & Ax \leqslant p_A - \mathbb{1}, \\ & x_u \in \{0,1\}, \end{array} \tag{1.2}$$

and A is now an arbitrary 0/1 matrix (the edge-node incidence matrix of the hypergraph), and the ith component of the vector p_A gives the number of positive entries in row i of the matrix A. If A does not contain a zero row, the undominated rows of A can be interpreted as the incidence vectors of the circuits of an *independence system*. Hence, problem (1.2) can be seen as the problem of finding an independent set of maximal weight. The convex hull of incidence vectors of independent sets (solutions to (1.2)) is known as the *independence system polytope*. Substantial work has been done to find classes of valid inequalities for the independence system polytope, mainly based on the study of special configurations of the family of circuits. Among these are, to name a few, the acyclic subdigraph polytope [25, 29], the bipartite subgraph polytope [4], and the planar subgraph polytope [30]. We refer the reader to [20, 32] and [1, 2, 16, 39, 45] for the study of the facial structure of the independence system polytope in general.

In section 2, we introduce an extension of the node packing problem in hypergraphs, called *transitive packing*, by taking transitive elements into account. The problems we consider can be described as

$$\begin{array}{ll} \text{maximize} & cx \\ \text{subject to} & Ax \leqslant p_A - \mathbb{1}, \\ & x_u \in \{0,1\}, \end{array} \tag{1.3}$$

where A is now an arbitrary $0/\pm 1$ matrix, and the ith component of the vector p_A gives the number of positive entries in row i of the matrix A. Many combinatorial optimization problems can be modeled as *transitive packing problems*. We do not (and cannot) list all problems that fit with this novel framework, but we name a few of them that we are going to revisit later. Indeed, besides those that can be interpreted as finding an independent set of maximal weight, there are the clique partitioning problem [27, 28, 41], the partition problem [10], the transitive acyclic subdigraph problem [34], the interval graph completion problem [35, 49], and the relatively transitive subgraph problem [31, 50, 51].

One of our main purposes is to derive broad classes of valid inequalities for the *transitive packing polytope*, the convex hull of feasible solutions to (1.3). In section 4, we present *generalized cycle*, *generalized clique*, *generalized antihole*, *generalized antiweb*, and *odd partition inequalities*, which are valid for the transitive packing polytope. These classes explain and classify many known inequalities for polytopes that fit with this general framework. Thereby, we emphasize the relations between, and the common structure of (inequalities for), different polyhedra, formerly independently studied, and we provide new insights as well as new inequalities for some of the special polytopes that arise from certain hypergraphs and choices of transitive elements. We show how the knowledge of structural properties of the transitive packing polytope makes it possible to derive results for these special problems.

We derive most of the inequalities for the transitive packing polytope by *integer rounding*. This provides *cutting plane proofs* for many of the known inequalities for special polytopes that have not been observed before. It may also be seen as a guide for using certain patterns of the (initial) constraint matrix A to obtain new inequalities in a systematic way. The latter property might be of some importance for solving general 0/1 integer programs. Moreover, the derivation of the inequalities may be seen as a guideline for generalizing each valid inequality for the node packing polytope whose cutting plane proof is known.

Section 5 is concerned with an interesting subclass of the transitive packing polytopes, formed by those whose corresponding hypergraph is actually a graph. In section 6, we discuss the separation problem associated with the classes of inequalities introduced before. Finally, in section 8 we recall the strong relation between *set covering* and independence system polytopes, point out its extension to *generalized set covering* and transitive packing polytopes, translate our results into this context, and briefly discuss the relation of our work to $0/\pm 1$ matrices that are *balanced* or *ideal*.

Subsequent to the original introduction of transitive packing [49, 36], Borndörfer and Weismantel [7, 8] introduced another scheme that also helps to explain and classify inequalities within the context of a packing polytope and to get cutting plane proofs. We refer to [48] for a discussion of similarities and differences between this scheme and transitive packing.

2. The transitive packing polytope. A *hypergraph* is an ordered pair $(N, \mathcal{H})$, where N is a finite ground set, the set of *nodes*, and $\mathcal{H}$ is a collection of distinct subsets of N, the set of (*hyper*)*edges*. We only deal with hypergraphs without loops, i.e., we always assume that $|H| \geqslant 2$ for all $H \in \mathcal{H}$. We refer to [6] for a thorough introduction to hypergraphs. Here, we are interested in hypergraphs with additional node subsets associated with each edge.

DEFINITION 2.1. *Let $(N, \mathcal{H})$ be a hypergraph, and let $\mathrm{tr} : \mathcal{H} \to 2^N$ be a mapping from the set of edges to the powerset of N, with the property that $\mathrm{tr}(H) \subseteq N \setminus H$. We call the ordered triple $(N, \mathcal{H}, \mathrm{tr})$ an* extended hypergraph*, and $\mathrm{tr}(H)$ the set of* transitive elements *associated with the edge H.*

In the special case that $\mathrm{tr}(H) = \emptyset$ for all $H \in \mathcal{H}$, we often simply write $(N, \mathcal{H})$ instead of $(N, \mathcal{H}, \mathrm{tr})$. We are interested in packing nodes of an extended hypergraph whereby the restrictions imposed by the edges may be compensated by picking transitive elements. This is made precise by the following definition.

DEFINITION 2.2. *Let $(N, \mathcal{H}, \mathrm{tr})$ be an extended hypergraph. A subset S of the nodes is a* transitive packing *(in $(N, \mathcal{H}, \mathrm{tr})$) if, for every $H \in \mathcal{H}$ such that $H \subseteq S$, there exists a node $u \in S \cap \mathrm{tr}(H)$.*

In other words, a transitive packing S is a set of nodes that contains an edge only if S contains at least one node from the set of transitive elements associated with that edge. Given, in addition to $(N, \mathcal{H}, \mathrm{tr})$, a weight function $c : N \to \mathbb{Q}$, the (*maximum weight*) *transitive packing problem* consists of finding a transitive packing $S \subseteq N$ of maximal weight $c(S)$. As indicated in the introduction, the maximum weight transitive packing problem is equivalent to the integer linear programming problem

$$\begin{aligned} &\text{maximize} && cx \\ &\text{subject to} && x(H) - x(\mathrm{tr}(H)) \leqslant |H| - 1 \quad \text{for all } H \in \mathcal{H}, && (2.1) \\ & && x \leqslant \mathbb{1}, && (2.2) \\ & && x \geqslant 0, && (2.3) \\ & && x \in \mathbb{Z}^N. && (2.4) \end{aligned}$$

Note that the constraint matrix of the inequalities (2.1) is the edge-node incidence matrix of the hypergraph $(N, \mathcal{H})$, with additional -1's for the transitive elements of the edge represented by the particular row. We call the inequalities (2.1) *transitivity constraints*.

In the following, we study the transitive packing polytope $P_{\mathrm{TP}}(N, \mathcal{H}, \mathrm{tr})$ of the extended hypergraph $(N, \mathcal{H}, \mathrm{tr})$, which is defined as the convex hull of the incidence vectors of transitive packings in $(N, \mathcal{H}, \mathrm{tr})$, i.e.,

$$P_{\mathrm{TP}}(N, \mathcal{H}, \mathrm{tr}) := \mathrm{conv}\{\chi^S \in \mathbb{R}^N : S \text{ transitive packing in } (N, \mathcal{H}, \mathrm{tr})\}.$$

In other words, $P_{\mathrm{TP}}(N, \mathcal{H}, \mathrm{tr})$ is equal to the integer hull of the feasible solutions to (2.1)–(2.3). At this point, it seems reasonable to introduce a few examples to illustrate the applicability of the results to be presented. Of course, if $\mathrm{tr}(H) = \emptyset$ and $|H| = 2$ for all edges $H \in \mathcal{H}$, a transitive packing reduces to an ordinary node packing in the graph $(N, \mathcal{H})$. However, to motivate hypergraphs and transitive elements, we show now that the acyclic subdigraph polytope as well as the clique partitioning polytope and the partition polytope can be obtained by special choices of the hypergraph and the transitive elements. Other examples will be discussed in section 7.

The acyclic subdigraph polytope. An instance of the *acyclic subdigraph problem* consists of a directed graph $D = (V, A)$ and a weight function $c : A \to \mathbb{Q}$. The objective is to determine a set of arcs $B \subseteq A$ such that the digraph (V, B) is acyclic, i.e., does not contain a directed cycle, and such that $c(B)$ is as large as possible. The *acyclic subdigraph polytope* is the convex hull of incidence vectors of acyclic arc subsets of A. It was studied by Grötschel, Jünger, and Reinelt (see [24, 25, 29]) and Goemans and Hall [23]. If we choose the arc set A of the digraph D as the node set of the hypergraph, if we declare the directed cycles in D as the edges of this hypergraph, and if we let $\mathrm{tr}(H) = \emptyset$ for all $H \in \mathcal{H}$, the acyclic subdigraph polytope appears as a special transitive packing polytope.

The clique partitioning polytope. An instance of the *clique partitioning problem* consists of an undirected graph $G = (V, E)$ and a weight function $c : E \to \mathbb{Q}$. A set $F \subseteq E$ of edges is called a *clique partitioning* of G if there is a partition of V into nonempty, disjoint sets $W_1, W_2, \ldots, W_k$ such that the subgraph induced by each W_i is a clique and such that $F = \bigcup_{i=1}^{k} \{\{u, v\} : u, v \in W_i, u \neq v\}$. Equivalently, a clique partitioning is a subrelation of the symmetric relation represented by G that is an equivalence relation, i.e., in particular transitive. The weight of such a clique partitioning F is $c(F)$. The task is to determine a clique partitioning of minimal weight. (Of course, since we do not restrict the objective function, we could have written that we want to find a clique partitioning of maximal weight as we always do in the context of transitive packing. However, for historical reasons we chose this variant.) The *clique partitioning polytope* is the convex hull of the incidence vectors of all clique partitionings in G. It was introduced and studied by Grötschel and Wakabayashi [27, 28] and has recently been further investigated by Oosten, Rutten, and Spieksma [41]. To show that it is an instance of a transitive packing polytope, it is sufficient to deal with a graph instead of a hypergraph. Indeed, we take as the set N of nodes the edges of G, and two nodes are adjacent (form a hyperedge) if and only if the associated edges are incident in the original graph G. That is, the extended hypergraph we consider is precisely the *line graph* of G, and the transitive element that we attach to a pair of incident edges $\{u, v\}$, $\{v, w\}$ in G is the edge $\{u, w\}$ if it exists.

The partition polytope. An instance of the *graph partitioning problem* consists of an undirected, connected graph $G = (V, E)$, a weight function $c : E \to \mathbb{Q}$, and an integer $r \leqslant |V|$. An r-partition of the node set V is a set of node subsets $N_1, N_2, \ldots, N_r$ such that $N_i \cap N_j = \emptyset$ (for all $i \neq j$) and $\cup_{i=1}^{r} N_i = V$. Some of the subsets N_i may be empty. The weight of an r-partition is the total weight of the edges with end points in two different subsets. The goal is to determine an r-partition of minimal weight. Chopra and Rao [10] have studied polytopes for several variations of this problem. We consider one of them here. This case arises when $r = |V|$. For a complete graph G, this problem is equivalent to the clique partitioning problem. For arbitrary graphs G, Chopra and Rao define the *partition polytope* as the convex hull of the incidence vectors of all sets of edges in G which are not cut by an r-partition. It follows from [10, Lemma 2.2] that the partition polytope arises as a transitive packing polytope by taking the edges of G as the set N and by letting every $(|C|-1)$-cardinality subset of edges of a cycle C in G be the edges of the hypergraph $\mathcal{H}$. The transitive set related to such a hyperedge contains exactly the missing edge from the cycle C.

Before studying the transitive packing polytope, we shall discuss an algorithmic aspect of the concept of transitive packings. How is $(N, \mathcal{H}, \mathrm{tr})$ given? Having in mind problems like the acyclic subdigraph problem, it does not seem to be satisfactory to assume that it is given as a list of hyperedges and their transitive elements. Indeed, the number of directed cycles in a digraph can be exponential in the number of nodes. From the point of view of polyhedral combinatorics, it rather seems to be reasonable to assume that the linear programming problem arising from (2.1)–(2.4) by dropping the integrality constraint (2.4) is solvable in time polynomially bounded in $|N|$ and the input size of c. This means, given a point $x \in \mathbb{Q}^N$ contained in the unit hypercube, we assume that the separation problem formed by x and the class of inequalities (2.1) is solvable in polynomial time. In particular, this guarantees that the decision version of the transitive packing problem belongs to the class NP. Since the node packing problem on graphs is NP-hard, the same holds for the transitive packing problem.

Let us continue with the study of the transitive packing polytope. Since the empty set as well as all singletons of N are transitive packings, we immediately obtain the following result.

PROPOSITION 2.3. *Let $(N, \mathcal{H}, \mathrm{tr})$ be an extended hypergraph.*

(i) *The transitive packing polytope $P_{\mathrm{TP}}(N, \mathcal{H}, \mathrm{tr}) \subseteq \mathbb{R}^N$ is full dimensional, i.e.,* $\dim(P_{\mathrm{TP}}(N, \mathcal{H}, \mathrm{tr})) = |N|$.

(ii) *The nonnegativity constraint $x_u \geqslant 0$ defines a facet of $P_{\mathrm{TP}}(N, \mathcal{H}, \mathrm{tr})$ for each node $u \in N$.*

Because of the transitive elements, it is more difficult to characterize the facet defining inequalities of type $x_u \leqslant 1$ for $u \in N$. Clearly, all these inequalities are facet defining if $|H| \geqslant 3$ for all edges $H \in \mathcal{H}$. But as soon as $\{u, v\} \in \mathcal{H}$ and $\mathrm{tr}(\{u, v\}) = \emptyset$, for instance, the face induced by $x_u \leqslant 1$ is properly contained in the facet defined by $x_v \geqslant 0$. But even if $\mathrm{tr}(\{u, v\}) \neq \emptyset$, it may happen that whenever u is chosen, we cannot choose another element. While it is possible to give a concise characterization in the absence of transitive elements, we are content with a sufficient condition in the general case.

LEMMA 2.4. *Let $P_{\mathrm{TP}}(N, \mathcal{H}, \mathrm{tr})$ be the transitive packing polytope associated with the extended hypergraph $(N, \mathcal{H}, \mathrm{tr})$.*

(i) *If $\mathrm{tr}(H)$ is the empty set for all edges $H \in \mathcal{H}$ such that $|H| = 2$, then an inequality $x_u \leqslant 1$ with $u \in N$ defines a facet of $P_{\mathrm{TP}}(N, \mathcal{H}, \mathrm{tr})$ if and only if $|H| \geqslant 3$ for all edges $H \in \mathcal{H}$ that contain u.*

(ii) *Let $u \in N$. If there exists for all edges $\{u,v\} \in \mathcal{H}$ a node $w \in \mathrm{tr}(\{u,v\})$ such that neither $\{u,w\} \in \mathcal{H}$, $\{v,w\} \in \mathcal{H}$, nor $\{u,v,w\} \in \mathcal{H}$, then the inequality $x_u \leqslant 1$ defines a facet of $P_{\mathrm{TP}}(N,\mathcal{H},\mathrm{tr})$.*

Proof. In case (i), the incidence vectors of the transitive packings $\{u\}$ and $\{u,v\}$ for all $v \in N \setminus \{u\}$ provide the needed set of linearly independent vectors. In case (ii), we proceed as follows. Besides $\{u\}$, we first choose a set $\{u,w\}$ such that $\{u,w\} \notin \mathcal{H}$. (Notice that our assumptions imply the existence of such a node w.) Then, by taking $\{u,v,w\}$, we collect all nodes $v \in N$ such that $\{u,v\} \in \mathcal{H}$, $w \in \mathrm{tr}(\{u,v\})$, $\{v,w\} \notin \mathcal{H}$, and $\{u,v,w\} \notin \mathcal{H}$. Now, we may forget these nodes v and the node w and continue with the remaining nodes in the same manner. Since $\{u,v\} \in \mathcal{H}$ for the nodes v above, they cannot occur in the role of w. Hence, the incidence vectors of the constructed transitive packings are linearly independent. □

We illuminate Lemma 2.4 by applying it to the node packing, the acyclic subdigraph, the clique partitioning, and the partition polytopes. For the node packing polytope of a graph G, (i) says that an inequality $x_u \leqslant 1$ is facet defining for a node u if and only if u is isolated, i.e., if G does not contain an edge incident to u. This is a special case of the well-known fact that a clique inequality defines a facet if and only if the clique is maximal [42]. Given a digraph $D = (V,A)$ and an arc $(u,v) \in A$, Lemma 2.4(i) implies that $x_{uv} \leqslant 1$ defines a facet of the acyclic subdigraph polytope of D if and only if $(v,u) \notin A$. This was shown before by Grötschel, Jünger, and Reinelt [25]. If G is a graph without isolated edges, the assumption of Lemma 2.4(ii) is never met by an edge of the clique partitioning polytope of G. Indeed, Grötschel and Wakabayashi [28] proved that no upper bound constraint defines a facet of this polytope. Finally, Lemma 2.4(ii) also tells us that $x_e \leqslant 1$ defines a facet of the partition polytope if the edge e does not belong to any cycle of length 3.

We conclude this first section on the transitive packing polytope by observing that a transitivity constraint $x(H') - x(\mathrm{tr}(H')) \leqslant |H'| - 1$ is dominated by $x(H) - x(\mathrm{tr}(H)) \leqslant |H| - 1$ if $H \subseteq H'$ and $\mathrm{tr}(H) \subseteq \mathrm{tr}(H')$.

3. The independence system polytope. So far we have mentioned only in the introduction that the transitive packing problem subsumes independent set problems. This section is intended to recall the needed definitions and to explain the relation in detail. An *independence system* is a pair $(N,\mathcal{I})$, with ground set N and a family $\mathcal{I}$ of subsets of N, that contains the empty set and is closed under set inclusion; i.e., for any set $I \in \mathcal{I}$ every subset $I' \subseteq I$ belongs also to $\mathcal{I}$. The elements of $\mathcal{I}$ are called *independent sets*. A subset of N that does not belong to $\mathcal{I}$ is called *dependent*, and the minimal dependent sets (with respect to set inclusion) are the *circuits* of the independence system. The collection of circuits forms a *clutter*, i.e., a family of sets such that no two of them are comparable with respect to set inclusion. Since a subset of N is independent if and only if it does not contain a circuit, an independence system is fully characterized by the family of its circuits. Conversely, every clutter $\mathcal{C} \subseteq 2^N$ determines a unique independence system with ground set N and $\{I \subseteq N : C \not\subseteq I \text{ for all } C \in \mathcal{C}\}$ as the family of its independent sets. The *independence system polytope* is defined as the convex hull of all incidence vectors of independent sets. It coincides with the transitive packing polytope $P_{\mathrm{TP}}(N,\mathcal{H})$, where $\mathrm{tr}(H) = \emptyset$ for all $H \in \mathcal{H}$, and $\mathcal{H}$ is the set of circuits. (To be accurate, this is only true when we make the standard assumption that all singletons are independent. Remember that we have defined the transitive packing polytope only for hypergraphs without loops.) In the following we will sometimes speak of independent sets instead of transitive packings, of circuits instead of edges, and of circuit constraints instead of transitivity (or packing)

constraints when dealing with the special case formed by transitive packing problem instances without transitive elements. As an example of an independence system, we may consider the one defined by the acyclic arc subsets of a digraph. The dicycles are one-to-one with the circuits, and the independence system polytope is the acyclic subdigraph polytope.

Given a hypergraph $(N, \mathcal{H})$, we define its *upper closure* $\mathcal{H}^+$ and its *reduction* $\mathcal{H}^-$ as $\mathcal{H}^+ := \{H' \subseteq N : \text{there exists an } H \in \mathcal{H} \text{ such that } H \subseteq H'\}$ and $\mathcal{H}^- := \{H \in \mathcal{H} : \text{there exists no } H' \in \mathcal{H} \text{ such that } H' \subset H\}$, respectively. Notice that $P_{\mathrm{TP}}(N, \mathcal{H}^+) = P_{\mathrm{TP}}(N, \mathcal{H}) = P_{\mathrm{TP}}(N, \mathcal{H}^-)$. These notions prove useful for characterizing the facet defining packing constraints. Observe that for clutters, for instance the circuits of independence systems, we have $\mathcal{H} = \mathcal{H}^-$.

THEOREM 3.1. *Let $(N, \mathcal{H})$ be a hypergraph. For $H \in \mathcal{H}$, the inequality $x(H) \leqslant |H| - 1$ defines a facet of $P_{\mathrm{TP}}(N, \mathcal{H})$ if and only if $H \in \mathcal{H}^-$ and for all $u \in N \setminus H$ there exists an $H' \subset H$ with $|H'| = |H| - 1$ such that $H' \cup \{u\} \notin \mathcal{H}^+$.*

Proof. Necessity of the stated condition is obvious; otherwise, the face under consideration would be the intersection of some other faces. To show sufficiency we take first the incidence vectors of all $|H|$ subsets of H of size $|H| - 1$. According to the assumption, for each node $u \in N \setminus H$ there exists a subset H' of H of size $|H| - 1$ such that $H' \cup \{u\}$ is independent. Adding the corresponding incidence vectors to our former set completes the proof. □

Theorem 3.1 implies, in particular, that all dicycle inequalities of the acyclic subdigraph polytope are facet defining. A direct proof of this result is given in [25].

Subclasses of the classes of valid inequalities that we introduce in the next section for the transitive packing polytope have been presented earlier for the independence system polytope; generalized cycle, generalized clique, and generalized antihole inequalities by Euler, Jünger, and Reinelt [20], and generalized antiweb inequalities by Laurent [32]. It will turn out that our inequalities are more general, even if we restrict ourselves to the independence system polytope. Nevertheless, in order to keep the terminology simple, we will give the new inequalities the same names and point out the restrictions that lead to the known inequalities, respectively. So far, no cutting plane proofs have been presented for the formerly known inequalities.

4. Valid inequalities. Let $P \subseteq \mathbb{R}^N$ be a rational polyhedron, for instance the initial relaxation of $P_{\mathrm{TP}}(N, \mathcal{H}, \mathrm{tr})$ defined by (2.1)–(2.3). One way to produce a characterization of the integer hull P_{I} of P by means of linear inequalities is integer rounding. For a thorough discussion of this topic, its history, and its applications to integer programming and combinatorial optimization, we refer the reader to the textbooks of Cook, Cunningham, Pulleyblank, and Schrijver [15, Chapter 6.7] and of Nemhauser and Wolsey [38, Chapter II.1] and to Schrijver [47, Chapter 23]. Here, we briefly review the basic definitions that will be needed later on.

If we set

$$P' := \{x \in P : ax \leqslant \beta \text{ for all } a \in \mathbb{Z}^N, \beta \in \mathbb{Z} \text{ with } \max\{ax : x \in P\} < \beta + 1\},$$

then P' can be seen as obtained from P by one step of rounding. In particular, if $P = \{x \in \mathbb{R}^N : Ax \leqslant b\}$ for an integer matrix A and integer right-hand side b, then

$$P' = \{x \in \mathbb{R}^N : \lambda Ax \leqslant \lfloor \lambda b \rfloor \text{ for all vectors } \lambda \geqslant 0 \text{ with } \lambda A \in \mathbb{Z}^N\}.$$

Obviously, the integer hull P_{I} of P, i.e., the convex hull of the integral points in P, is contained in P'. Furthermore $P' = P$ if and only if $P = P_{\mathrm{I}}$. If we define $P^{(0)} := P$

and, recursively, $P^{(t+1)} := (P^{(t)})'$ for all nonnegative integers t, then $P_{\mathrm{I}} \subseteq P^{(t)}$ for all nonnegative integers t. Schrijver [46] showed that P' is again a polyhedron and that there is a nonnegative integer t such that $P^{(t)} = P_{\mathrm{I}}$. The (*Gomory–Chvátal*) *rank* of P is the smallest t such that $P^{(t)} = P_{\mathrm{I}}$. Let $ax \leqslant \beta$ be a valid inequality for P_{I}. Its *depth* relative to P is the smallest d such that $ax \leqslant \beta$ is valid for $P^{(d)}$. Therefore the rank of P equals the maximal depth, relative to P, of an inequality valid for P_{I}.

Let $Ax \leqslant b$ be a system of linear inequalities, and let $cx \leqslant \delta$ be an inequality. Moreover, let $c_1x \leqslant \delta_1, c_2x \leqslant \delta_2, \ldots, c_mx \leqslant \delta_m$ be a sequence of linear inequalities such that each vector c_i, $i = 1, \ldots, m$, is integral, $c_m = c$, $\delta_m = \delta$, and for $i = 1, \ldots, m$ the inequality $c_ix \leqslant \delta_i'$ is a nonnegative linear combination of the inequalities $Ax \leqslant b$, $c_1x \leqslant \delta_1, \ldots, c_{i-1}x \leqslant \delta_{i-1}$ for some δ_i' with $\lfloor \delta_i' \rfloor \leqslant \delta_i$. Such a sequence is called a *cutting plane proof* of $cx \leqslant \delta$ from $Ax \leqslant b$, and m is the *length* of this proof. The depth of the final inequality $cx \leqslant \delta$ is the *depth* of the proof. Every integer solution of $Ax \leqslant b$ satisfies $cx \leqslant \delta$. Let $P = \{x : Ax \leqslant b\}$. Since $P^{(t)} = P_{\mathrm{I}}$ for some t, the converse is true as soon as P_{I} is nonempty. That is, every inequality $cx \leqslant \delta$ with c integral and valid for P_{I} has a cutting plane proof from $Ax \leqslant b$. Clearly, the length of a cutting plane proof of a valid inequality for P_{I} is at least its depth; however, the length can be significantly bigger (see, e.g., [12]).

The idea of deriving cutting planes by rounding based on the exploitation of problem structure can, in particular, be used to obtain valid inequalities for the transitive packing polytope. Thereby, we also show that many inequalities valid for the polytopes which arise from $P_{\mathrm{TP}}(N, \mathcal{H}, \mathrm{tr})$ by certain choices of $(N, \mathcal{H}, \mathrm{tr})$ have short and insightful cutting plane proofs from the initial relaxation (2.1)–(2.3).

4.1. Generalized cycle inequalities. We first use cycles of the hypergraph $(N, \mathcal{H})$ to obtain a class of valid inequalities for the transitive packing polytope, each of which has a cutting plane proof from (2.1)–(2.3) of length 1. Recall that a *cycle* in a hypergraph is a sequence of vertices and of edges of the form $(u_1, H_1, u_2, H_2, \ldots, u_k, H_k, u_{k+1})$ such that the vertices $u_1, \ldots, u_k$ are distinct, $u_{k+1} = u_1$, the edges $H_1, \ldots, H_k$ are distinct, and for $i = 1, \ldots, k$ both u_i and u_{i+1} are contained in H_i. We start, however, with a few more assumptions.

Definition 4.1. *Let $(N, \mathcal{H})$ be a hypergraph, and let q, s, and r be positive integers such that $q \geqslant 2$ and $1 \leqslant r \leqslant q-1$. For convenience, we set $k := sq + r$. Let $N_1, \ldots, N_k$ be a sequence of pairwise disjoint nonempty subsets of N. For $i = 1, \ldots, k$, let $H_i \in \mathcal{H}$ be an edge such that $\bigcup_{j=i}^{i+q-1} N_j \subseteq H_i$. (Indices greater than k are taken modulo $k+1$ and shifted by $+1$.) We denote by C the union of all these edges H_i, $C := \bigcup_{i=1}^{k} H_i$, and by $m(u)$ the multiplicity of a node $u \in C$ in this edge collection, i.e., $m(u) := |\{i \in \{1, \ldots, k\} : u \in H_i\}|$. We assume that $m(u) \leqslant q$ for all nodes $u \in C$. Then we call the hypergraph $(C, \{H_i : i = 1, 2, \ldots, k\})$ a* generalized (k, q)-cycle (*contained in $(N, \mathcal{H})$*).

To illuminate this definition, Figures 4.1 and 4.2 show a generalized $(10, 4)$-cycle and two generalized $(5, 2)$-cycles, respectively. Observe that every generalized cycle is a cycle of the hypergraph, but not vice versa. In fact, the name is a concession to the literature, where already a substructure of the generalized cycles just introduced got this name; see [20]. We now develop an inequality supported by a generalized cycle and its set of transitive elements. So let $(C, \{H_i : i = 1, 2, \ldots, k\})$ be a generalized (k, q)-cycle in $(N, \mathcal{H}, \mathrm{tr})$, and assume that the set $\mathrm{tr}(C) := \bigcup_{i=1}^{k} \mathrm{tr}(H_i)$ of transitive elements does not interact with C itself, i.e., $\mathrm{tr}(H_i) \cap C = \emptyset$ for $i = 1, \ldots, k$. To simplify the notation, we denote by $n(u) := |\{i \in \{1, \ldots, k\} : u \in \mathrm{tr}(H_i)\}|$ the multiplicity of a node $u \in N \setminus C$ with respect to the transitive sets of the edges of the

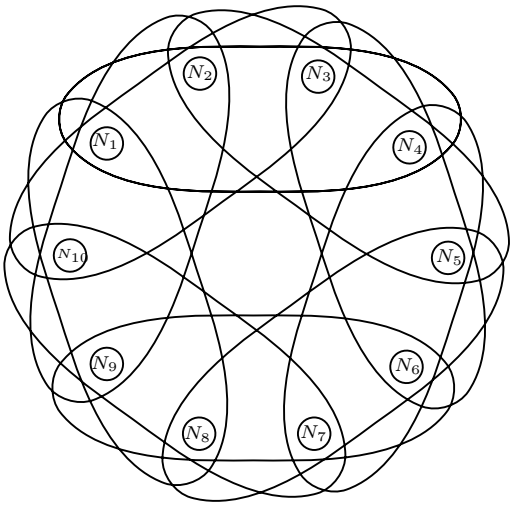


FIG. 4.1. *A generalized* $(10, 4)$*-cycle with* $C = \bigcup_{i=1}^{k} N_i$.

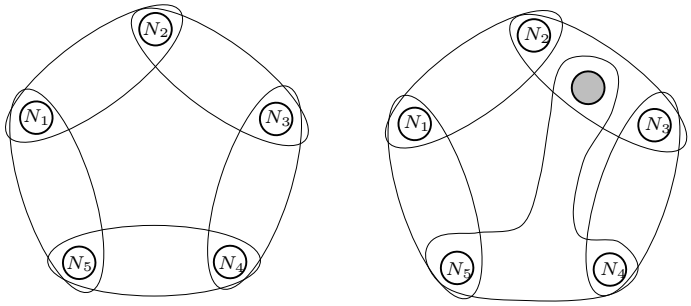


FIG. 4.2. *Two generalized* $(5, 2)$*-cycles. The second one illustrates the case* $C \supset \bigcup_{i=1}^{k} N_i$.

cycle. Furthermore, we let $\lceil \alpha \rceil_q$ be the smallest integer that is bigger than or equal to the scalar α as well as divisible by q.

Adding the transitivity constraints associated with the edges of the generalized (k, q)-cycle,

$$\sum_{u \in H_i} x_u - \sum_{u \in \mathrm{tr}(H_i)} x_u \leqslant |H_i| - 1 \quad \text{for } i = 1, \ldots, k,$$

an appropriate multiple of upper bound constraints,

$$(q - m(u))x_u \leqslant q - m(u) \quad \text{for } u \in C \setminus \bigcup_{i=1}^{k} N_i,$$

as well as an appropriate multiple of nonnegativity constraints,

$$-(\lceil n(u) \rceil_q - n(u))x_u \leqslant 0 \quad \text{for } u \in \mathrm{tr}(C) \text{ with } n(u) \not\equiv 0 \bmod q,$$

and dividing the result by q, we obtain

$$\sum_{u \in C} x_u - \sum_{u \in \mathrm{tr}(C)} \frac{\lceil n(u) \rceil_q}{q} x_u \leqslant \frac{q|C| - k}{q}.$$

Rounding down the right-hand side completes the proof of the following result.

THEOREM 4.2. *Let $(N, \mathcal{H}, \mathrm{tr})$ be an extended hypergraph, and let, for $k > q$, $k \not\equiv 0 \bmod q$, the hypergraph $(C, \{H_i : i = 1, 2, \ldots, k\})$ be a generalized (k, q)-cycle in $(N, \mathcal{H})$ such that $\mathrm{tr}(H_i) \cap C = \emptyset$ for $i = 1, \ldots, k$. Then, the generalized (k, q)-cycle inequality*

$$\sum_{u \in C} x_u - \sum_{u \in \mathrm{tr}(C)} \frac{\lceil n(u) \rceil_q}{q} x_u \leqslant |C| - \left\lceil \frac{k}{q} \right\rceil \tag{4.1}$$

is valid for the transitive packing polytope $P_{\mathrm{TP}}(N, \mathcal{H}, \mathrm{tr})$.

We now relate this first class of inequalities for the transitive packing polytope $P_{\mathrm{TP}}(N, \mathcal{H}, \mathrm{tr})$ to the four selected examples. For the node packing polytope, we obtain exactly the *odd cycle inequalities* introduced by Padberg [42]. This is true because all edges of the (hyper)graph have size 2, and hence all sets N_i have to be singletons. If C is the set of nodes of an odd cycle in a graph G, then the associated odd cycle inequality reads

$$x(C) \leqslant \frac{|C| - 1}{2}.$$

The *Möbius ladder inequalities* form a quite prominent class of facet defining inequalities for the acyclic subdigraph polytope. The support of any of these inequalities is defined as follows.

DEFINITION 4.3 (see [25]). *Let $C_1, C_2, \ldots, C_k$ be a sequence of different dicycles in a digraph $D = (V, A)$ such that the following hold:*

(1) *$k \geqslant 3$ and k odd.*

(2) *C_i and C_{i+1}, $i \in \{1, 2, \ldots, k-1\}$, have a directed path P_i in common; C_1 and C_k have a directed path P_k in common.*

(3) *Given any dicycle C_j, $j \in \{1, 2, \ldots, k\}$, set $I_j := \{1, 2, \ldots, k\} \cap (\{j-2, j-4, j-6, \ldots\} \cup \{j+1, j+3, j+5, \ldots\})$. (Indices greater than k are taken modulo $k+1$ and shifted by $+1$; indices less than 0 are first shifted by -1 and then taken modulo $k+1$.) Then every set $(\bigcup_{i=1}^{k} C_i) \setminus \{a_i : i \in I_j\}$ contains exactly one dicycle (namely, C_j), where a_i, $i \in I_j$, is any arc contained in the dipath P_i.*

(4) *The largest acyclic arc set in $\bigcup_{i=1}^{k} C_i$ has cardinality $|\bigcup_{i=1}^{k} C_i| - \frac{k+1}{2}$.*

*Then the arc set $M := \bigcup_{i=1}^{k} C_i$ is called a (k-)*Möbius ladder.

From Definition 4.3(4) it follows that for any k-Möbius ladder M contained in a digraph D the Möbius ladder inequality

$$x(M) \leqslant |M| - \frac{k+1}{2} \tag{4.2}$$

is valid for the acyclic subdigraph polytope of D. Definition 4.3(3)–(4) seem to be rather unhandy. There exists a large subclass, however, where these conditions are naturally satisfied. Let $C_1, C_2, \ldots, C_k$, $k \geqslant 5$, be a sequence of directed cycles satisfying (1) and (2). If no two different dicycles C_i and C_j, with $j \neq i-1, i+1$, share a node, Grötschel, Jünger, and Reinelt [25] observed that the union of these dicycles forms a Möbius ladder. Such a situation is depicted in Figure 4.3. We now prove that this subclass is contained in the class of generalized cycle inequalities, as has essentially been shown in the context of the independence system polytope by Euler, Jünger, and Reinelt [20].

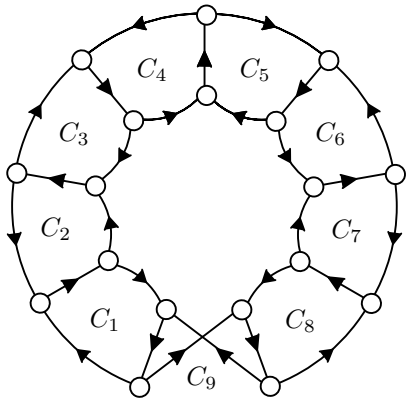


FIG. 4.3. *A 9-Möbius ladder.*

THEOREM 4.4. *Let D be a digraph, and let, for $k \geqslant 5$, $C_1, C_2, \ldots, C_k$ be a sequence of different dicycles in D satisfying Definition* 4.3(1)–(2). *If no two different dicycles C_i and C_j, with $j \neq i-1, i+1$, have a node in common $(i, j = 1, 2, \ldots, k)$, the Möbius ladder inequality* (4.2) *is contained in the class of generalized $(k, 2)$-cycle inequalities for the acyclic subdigraph polytope of D.*

Proof. We make use of the notation introduced in the discussion of the generalized cycle inequalities. We choose $q = 2$ and let k be the number of dicycles. The sets N_i, $i = 1, 2, \ldots, k$, are defined by the arcs forming the dipaths P_i, respectively. For $i = 1, 2, \ldots, k$, the arc sets N_i and N_{i+1} are contained in the hyperedge given by the dicycle C_{i+1}. Observe that no arc in $M = \bigcup_{i=1}^{k} C_i$ occurs in more than two dicycles. The claim now follows from Theorem 4.2. □

Theorem 4.4 throws some light on the Möbius ladder inequalities. The way we derived the generalized cycle inequalities explains, in particular, why the sequence of dicycles should be odd, as was already observed by Grötschel, Jünger, and Reinelt: "For even k, the construction does not give anything interesting" [25, p. 34]. Notice that Theorem 4.4 remains true for those Möbius ladders where each triple of the dicycles $C_1, C_2, \ldots, C_k$ does not have a common arc.

In the case of the clique partitioning polytope, we are obviously restricted to generalized $(k, 2)$-cycles, as the underlying hypergraph is actually a graph, the line graph of the given graph $G = (V, E)$. Nevertheless, this class contains two known classes of valid inequalities. Both are facet defining if G is a complete graph. The first class is formed by the 2-*chorded odd cycle inequalities* introduced by Grötschel and Wakabayashi [28]. Let $C = \{e_1, e_2, \ldots, e_k\}$ be the set of edges of an odd cycle in G, say $e_i = \{u_i, u_{i+1}\}$, and let $\mathrm{tr}(C) = \{\{u_i, u_{i+2}\} \in E : i = 1, 2, \ldots, k\}$ be its set of 2-chords (transitive elements). (As before, indices greater than k are taken modulo $k+1$ and shifted by $+1$.) By observing that $C \cap \mathrm{tr}(C) = \emptyset$, we may apply Theorem 4.2 and obtain the 2-chorded odd cycle inequality

$$\sum_{i=1}^{k} x_{\{u_i, u_{i+1}\}} - \sum_{\substack{i=1 \\ \{u_i,\, u_{i+2}\} \in E}}^{k} x_{\{u_i, u_{i+2}\}} \leqslant \frac{k-1}{2}.$$

However, even structures that are not cycles in G lead to generalized $(k, 2)$-cycle inequalities. For $k \geqslant 3$ odd, assume that G contains the star formed by the sequence

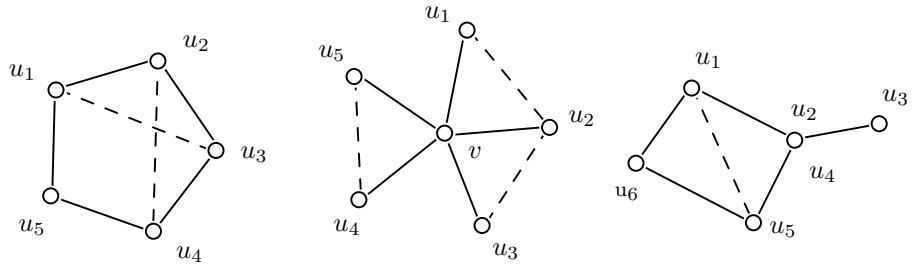


FIG. 4.4. *Generalized* $(5,2)$*-cycles for the clique partitioning polytope. The first is a* 2*-chorded odd cycle, the second is an odd wheel. The third is neither a* 2*-chorded odd cycle nor an odd wheel. The dotted edges indicate existing transitive edges* (*i.e., coefficient* -1 *in the associated inequalities*).

$\{v, u_i\}$, $i = 1, 2, \ldots, k$, of incident edges. Let $\mathrm{tr}(C)$ denote the associated set of 2-chords, i.e., $\mathrm{tr}(C) = \{\{u_i, u_{i+1}\} \in E : i = 1, 2, \ldots, k\}$. Again we have $\mathrm{tr}(C) \cap C = \emptyset$, and Theorem 4.2 implies that the *odd wheel inequality*

$$\sum_{i=1}^{k} x_{\{v,u_i\}} - \sum_{\substack{i=1 \\ \{u_i,\, u_{i+1}\} \in E}}^{k} x_{\{u_i,u_{i+1}\}} \leqslant \frac{k-1}{2}$$

is valid for the clique partitioning polytope. It was introduced and shown to be facet defining if G is complete by Chopra and Rao [10].

There are other structures that may form generalized $(k,2)$-cycles in the line graph of G; see, for instance, Figure 4.4. We can summarize our observations as follows.

THEOREM 4.5. *The class of generalized* $(k,2)$*-cycle inequalities for the clique partitioning polytope properly contains all* 2*-chorded odd cycle inequalities and all odd wheel inequalities.*

The odd wheel inequalities remain valid and facet defining for the partition polytope [10], where they also form a subclass of the generalized $(k,2)$-cycle inequalities. In fact, it is immediate that they can be generalized such that the spokes of the wheel are paths instead of single edges. Moreover, from the class of generalized cycle inequalities we get what we may call *q-chorded cycle inequalities*, a generalization of the 2-chorded odd cycle inequalities of the clique partitioning polytope. Consider a cycle of length k in G, with nodes $1, \ldots, k$. Assume that G also contains the edges $\{i, i+q\}, i = 1, \ldots, k$. Then we define the *q-chorded cycle inequality* as

$$\sum_{i=1}^{k} x_{\{i,i+1\}} - \sum_{i=1}^{k} x_{\{i,i+q\}} \leqslant k - \left\lceil \frac{k}{q} \right\rceil.$$

Again, the edges $\{i, i+1\}$ may be replaced by paths.

We return to the study of the transitive packing polytope in general. Under different types of weak assumptions it is possible to show that the generalized cycle inequality (4.1) has depth 1 relative to (2.1)–(2.3). We present one condition that turns out to be widely applicable. We still use the notation introduced during the definition of a generalized cycle.

LEMMA 4.6. *Let* $(N, \mathcal{H}, \mathrm{tr})$ *be an extended hypergraph; let* $k > q$, $k \not\equiv 0 \bmod q$*; and let* $H_1, \ldots, H_k$ *be the sequence of edges of a generalized* (k,q)*-cycle with node set*

C in $(N,\mathcal{H})$. Assume that $\mathrm{tr}(H_i)\cap C=\emptyset$ for $i=1,2,\dots,k$. If one of the following two conditions is satisfied, then the depth of the generalized (k,q)-cycle inequality (4.1) *relative to* (2.1)–(2.3) *is* 1.

(i) *Every edge $H\in\mathcal{H}\setminus\{H_1,\dots,H_k\}$ with $H\subseteq C$ satisfies $|\mathrm{tr}(H)\cap C|\geqslant 2$.*

(ii) *The generalized cycle satisfies $C=\bigcup_{i=1}^{k}N_i$ and $|N_i|=1$ for $i=1,2,\dots,k$, and every edge $H\in\mathcal{H}\setminus\{H_1,\dots,H_k\}$ with $H\subseteq C$ satisfies $|H|=q$.*

Proof. The same proof works for both cases. For $i=1,\dots,k$ we let u_i be an arbitrary representative of the node subset N_i, i.e., $u_i\in N_i$. We define the point $x\in\mathbb{R}^N$ as follows:

$$x_u:=\begin{cases}(q-1)/q & \text{if } u\in\{u_1,\dots,u_k\},\\ 1 & \text{if } u\in C\setminus\{u_1,\dots,u_k\},\\ 0 & \text{otherwise.}\end{cases}$$

Whereas x belongs to the initial linear relaxation of $P_{\mathrm{TP}}(N,\mathcal{H},\mathrm{tr})$, i.e., satisfies the inequalities (2.1)–(2.3), it violates inequality (4.1). Hence this inequality is not implied by the initial system. □

Notice that Lemma 4.6(ii) is satisfied in the case of the node packing and the clique partitioning polytopes.

Euler, Jünger, and Reinelt [20] introduced generalized cycle inequalities for the independence system polytope and showed that they are facet defining for the independence system induced by the edges of the generalized cycle. The generalized cycles presented here, restricted to independence systems, extend theirs, since they assumed that the nodes of $C\setminus\bigcup_{i=1}^{k}N_i$ are arranged in a certain sequence corresponding to that of the sets N_i.

Finally, we introduce a class of inequalities also supported by generalized cycles, which are in general weaker than the generalized cycle inequalities. This class arises from the class of generalized cycle inequalities when we pay no attention to repetitions of transitive elements. We call this class of valid inequalities *weak* generalized cycle inequalities. For ease of referencing, we state this as a lemma.

LEMMA 4.7. *Let $(N,\mathcal{H},\mathrm{tr})$ be an extended hypergraph, and let, for $k>q$, $k\not\equiv 0 \bmod q$, the hypergraph $(C,\{H_i : i=1,2,\dots,k\})$ be a generalized (k,q)-cycle in $(N,\mathcal{H})$ such that $\mathrm{tr}(H_i)\cap C=\emptyset$ for $i=1,\dots,k$. Then, the weak generalized (k,q)-cycle inequality*

$$\sum_{u\in C}x_u-\sum_{u\in\mathrm{tr}(C)}n(u)x_u\leqslant|C|-\left\lceil\frac{k}{q}\right\rceil$$

is valid for the transitive packing polytope $P_{\mathrm{TP}}(N,\mathcal{H},\mathrm{tr})$.

Clearly, in the case $n(u)\leqslant 1$ for all nodes $u\in N$, a generalized (k,q)-cycle inequality and its weak version coincide.

4.2. Generalized clique inequalities. A second well-known class of valid inequalities for the node packing polytope are *clique inequalities*; see, e.g., [42]. Such an inequality is supported by a clique C in the given graph and is of the form

$$x(C)\leqslant 1.$$

It defines a facet if and only if the clique is maximal (with respect to set inclusion). We now describe how the clique inequalities can be extended to the transitive packing polytope.

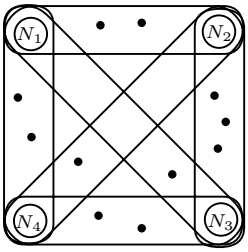


FIG. 4.5. *A* $(4,2)$*-clique. The points indicate other nodes of the clique.*

DEFINITION 4.8. *Let* $(N,\mathcal{H})$ *be a hypergraph, and let* $N_1,\dots,N_k$*, for integers* $k \geqslant q \geqslant 2$*, be a collection of mutually disjoint nonempty subsets of the node set* N*. For each* q*-element subset* $\{i_1,\dots,i_q\} \subseteq \{1,\dots,k\}$ *of indices, we let* $H_{i_1,\dots,i_q} \in \mathcal{H}$ *be an edge such that* $\bigcup_{j=1}^{q} N_{i_j} \subseteq H_{i_1,\dots,i_q}$*. We assume that the edges in any collection of intersecting edges all have one common index. Let* C *be the union of these edges,* $C := \bigcup_{1\leqslant i_1<i_2<\cdots<i_q\leqslant k} H_{i_1,\dots,i_q}$*. Then, we call the hypergraph*

$$(C, \{H_{i_1,\dots,i_q} : 1 \leqslant i_1 < i_2 < \cdots < i_q \leqslant k\})$$

a generalized (k,q)-clique (*contained in* $(N,\mathcal{H})$).

Figure 4.5 depicts a generalized $(4,2)$-clique. Observe that the class of generalized $(3,2)$-cliques coincides with that of generalized $(3,2)$-cycles. Whenever we deal with generalized cliques in the context of extended hypergraphs, we assume that C and its set $\mathrm{tr}(C) := \bigcup_{1\leqslant i_1<i_2<\cdots<i_q\leqslant k} \mathrm{tr}(H_{i_1,\dots,i_q})$ of transitive elements are disjoint, i.e., $\mathrm{tr}(H_{i_1,\dots,i_q}) \cap C = \emptyset$ for all $1 \leqslant i_1 < i_2 < \cdots < i_q \leqslant k$. We denote by $\mathrm{mtr}(C)$ the multiset that arises from the union of the transitive elements $\mathrm{tr}(H_{i_1,\dots,i_q})$. In other words, the multiplicity of a node $u \in \mathrm{mtr}(C)$ is precisely the number of edges $H_{i_1,\dots,i_q}$ of which u is a transitive element.

THEOREM 4.9. *Let* $(N,\mathcal{H},\mathrm{tr})$ *be an extended hypergraph, and let, for* $k \geqslant q \geqslant 2$*, the hypergraph* $(C, \{H_{i_1,\dots,i_q} : 1 \leqslant i_1 < i_2 < \cdots < i_q \leqslant k\})$ *be a generalized* (k,q)*-clique in* $(N,\mathcal{H})$ *such that* $\mathrm{tr}(H_{i_1,\dots,i_q}) \cap C = \emptyset$ *for* $1 \leqslant i_1 < i_2 < \cdots < i_q \leqslant k$*. Then the generalized* (k,q)*-clique inequality*

$$x(C) - x(\mathrm{mtr}(C)) \leqslant |C| - k + q - 1 \tag{4.3}$$

is valid for $P_{\mathrm{TP}}(N,\mathcal{H},\mathrm{tr})$.

Proof. The proof is by induction on the size k of the generalized clique. Observe that for $k = q$ inequality (4.3) coincides with a transitivity constraint. In order to show its validity for $k > q$, we consider all $\binom{k}{\ell}$ generalized (ℓ,q)-cliques that are induced by the ℓ-element subsets of $\{N_1,\dots,N_k\}$ for $\ell := \lfloor k(q-1)/q \rfloor + 1$. If we take the sum of their corresponding generalized (ℓ,q)-clique inequalities, we obtain an inequality whose support coincides with $C \cup \mathrm{tr}(C)$. Due to the assumptions on the relation of edges, a node $u \in N_i$ for some $i \in \{1,\dots,k\}$ has coefficient $\binom{k-1}{\ell-1}$. The coefficient of a node $u \in C \setminus \bigcup_{i=1}^{k} N_i$ is less than or equal to $\binom{k-1}{\ell-1}$. The coefficient of each element in the multiset $\mathrm{mtr}(C)$ is $\binom{k-q}{\ell-q}$. In order to bring these coefficients into a line, we add suitable multiples of the upper bound inequalities $x_u \leqslant 1$ for nodes $u \in C \setminus \bigcup_{i=1}^{k} N_i$, and of the nonnegativity constraints $x_u \geqslant 0$ for $u \in \mathrm{mtr}(C)$. The

resulting inequality then becomes

$$\binom{k-1}{\ell-1}\big(x(C)-x(\mathrm{mtr}(C))\big) \leqslant \binom{k-1}{\ell-1}|C| + \binom{k}{\ell}(q-\ell-1).$$

Dividing this new inequality by $\binom{k-1}{\ell-1}$ results in

$$x(C)-x(\mathrm{mtr}(C)) \leqslant |C| - k + q - 1 + \frac{k-\ell}{\ell}(q-1),$$

and by the choice of ℓ we can truncate the last term of the right-hand side to 0. □

Observe that in the case $q = 2$, the size ℓ of the generalized cliques to be considered in the proof of Theorem 4.9 is $\ell = \lceil \frac{k+1}{2} \rceil$. This implies that the depth of the presented cutting plane proof is at most $\lceil \log(k-1) \rceil$. After drawing some conclusions from Theorem 4.9 for the acyclic subdigraph polytope and the clique partitioning polytope, we show that this bound is almost the best possible.

Again, if we consider the case of independence systems, the definition of generalized cliques given above is slightly more general than that of Euler, Jünger, and Reinelt [20]. They assumed that a node $u \in C \setminus \bigcup_{i=1}^{k} N_i$ cannot be contained in more than $\binom{k-1}{q-1} - 1$ edges (with common subindex) of the generalized (k,q)-clique. They showed that the corresponding generalized clique inequalities are facet inducing for the independence system with ground set C and circuits $H_{i_1,\ldots,i_q}$.

Euler, Jünger, and Reinelt also observed that in the case of the acyclic subdigraph polytope the simple k-fence inequalities are contained in the class of generalized clique inequalities. We now show that even the k-fence inequalities (not necessarily simple) are contained in the class of generalized $(k,2)$-clique inequalities.

A *simple k-fence* ($k \geqslant 3$) is a digraph that is isomorphic to the digraph $F = (U, B_1 \cup B_2)$ on $2k$ nodes $U = \{u_1, u_2, \ldots, u_{2k}\}$, where

$$B_1 = \{(u_i, u_{k+i}) : i = 1, \ldots, k\},$$
$$B_2 = \bigcup_{i=1}^{k} \{(u_{k+i}, v) : v \in \{u_1, \ldots, u_k\} \setminus \{u_i\}\}.$$

Adopting the notation of [25], we call the arcs in B_1 *pales* and the arcs in B_2 *pickets*. A *k-fence* is a digraph that arises from a simple k-fence by repeated subdivision of arcs; i.e., an arc (u,v) may be replaced by (u,w) and (w,v), where w is a new node, and so on. To keep the notation simple, we assume that $F = (U, B_1 \cup B_2)$ is a k-fence and call the arcs on the directed paths from u_i to u_{k+i} pales and those on the directed paths from u_{k+i} to v, $v \neq u_i$, pickets as well. If D is a digraph that contains the k-fence F, the *k-fence inequality*

$$x(B_1 \cup B_2) \leqslant |B_1 \cup B_2| - k + 1 \tag{4.4}$$

defines a facet of the acyclic subdigraph polytope of D; see [25].

THEOREM 4.10. *Let D be a digraph, and let $F = (U, B_1 \cup B_2)$ be a k-fence contained in D. Then the k-fence inequality* (4.4) *is contained in the class of generalized $(k,2)$-clique inequalities for the acyclic subdigraph polytope of D.*

Proof. We continue to use the notation introduced when we defined generalized cliques. We set N_i to be the set of pales on the path from u_i to u_{k+i} for $i = 1, 2, \ldots, k$. Furthermore, for $1 \leqslant i < j \leqslant k$, we define H_{ij} to be the dicycle in F formed by the

set of pales on the paths from u_i to u_{k+i} and from u_j to u_{k+j} as well as the pickets on the paths from u_{k+i} to u_j and u_{k+j} to u_i. Thus the k-fence F defines a generalized $(k,2)$-clique, and its k-fence inequality coincides with the corresponding generalized $(k,2)$-clique inequality. □

Whereas the class of generalized (k,q)-clique inequalities for the acyclic subdigraph polytope is richer than the class of k-fence inequalities, the class of generalized $(k,2)$-clique inequalities turns out to be precisely the class of $(1,k)$-2-*partition inequalities* for the clique partitioning polytope of a graph $G=(V,E)$. (Here, $q>2$ is not possible.) The latter inequalities are due to Grötschel and Wakabayashi [28] and are of the following form. Let $v,u_1,u_2,\dots,u_k \in V$ be a set of $k+1$ vertices such that $\{u_i,v\} \in E$ for $i=1,2,\dots,k$. Then the inequality

$$\sum_{i=1}^{k} x_{\{u_i,v\}} - \sum_{\substack{1 \leqslant i < j \leqslant k \\ \{u_i,\,u_j\} \in E}} x_{\{u_i,u_j\}} \leqslant 1 \tag{4.5}$$

is valid for the clique partitioning polytope. It is facet defining if G is complete; see [28].

THEOREM 4.11. *The class of generalized $(k,2)$-clique inequalities for the clique partitioning polytope of a graph G coincides with the class of $(1,k)$-2-partition inequalities.*

Proof. Let us first consider a $(1,k)$-2-partition inequality (4.5). Since the edges $\{u_i,v\}$ and $\{u_j,v\}$ for $i,j=1,2,\dots,k$, $i \neq j$, form a hyperedge and since the transitive edges of these hyperedges are distinct from the edges $\{u_i,v\} \in E$ for $i=1,2,\dots,k$, this inequality is a generalized $(k,2)$-clique inequality. On the other hand, a generalized $(k,2)$-clique of the line graph of G always leads to the support of a $(1,k)$-2-partition inequality: since all participating edges in G have to be pairwise incident, either they share one common node or we have $k=3$. In the former case they form the support of a $(1,k)$-2-partition inequality. The latter case contradicts the assumption that the generalized clique and its transitive elements do not intersect. □

For the partition polytope, there can exist generalized (k,q)-clique inequalities for any q.

We are now about to show that the depth of the generalized $(k,2)$-clique inequalities tends to infinity with k.

THEOREM 4.12. *Let $(C,\{H_{ij} : 1 \leqslant i < j \leqslant k\})$ be a generalized $(k,2)$-clique of the extended hypergraph $(N,\mathcal{H},\mathrm{tr})$. Assume that $N_i = \big(\bigcap_{i<j\leqslant k} H_{ij}\big) \bigcap \big(\bigcap_{1\leqslant j<i} H_{ji}\big)$, for $i=1,2,\dots,k$, and that each edge $H \in \mathcal{H}$ such that $H \subseteq C$ satisfies $N_i \cup N_j \subseteq H$ for some $i,j \in \{1,2,\dots,k\}$, $i \neq j$. Then the depth of the generalized $(k,2)$-clique inequality (4.3) relative to (2.1)–(2.3) is at least $\log k - 1$.*

In order to prove this theorem we make use of the following lemma of Chvátal, Cook, and Hartmann [12].

LEMMA 4.13 (see [12]). *Let P be a rational polyhedron in $\mathbb{R}^N$. Let y and z be points in $\mathbb{R}^N$, and let $\mu_1,\mu_2,\dots,\mu_d$ be positive numbers. Furthermore, for $t=0,1,\dots,d$ set*

$$x^{(t)} := y - \sum_{i=1}^{t} \frac{1}{\mu_i} z.$$

If $y \in P$ and if, for all $t=1,\dots,d$, every inequality $ax \leqslant \beta$ valid for $P \cap \mathbb{Z}^N$ with $a \in \mathbb{Z}^N$ and $az < \mu_t$ satisfies $ax^{(t)} \leqslant \beta$, then $x^{(t)} \in P^{(t)}$ for all $t=0,1,\dots,d$.

Proof of Theorem 4.12. For $i = 1, \dots, k$ let u_i be an arbitrary representative of the node subset N_i, i.e., $u_i \in N_i$. Let C_1 be the union of these nodes u_i, $C_1 := \bigcup_{i=1}^{k} \{u_i\}$. Moreover, denote by C_2 the rest of the generalized $(k, 2)$-clique C, that is, $C_2 := C \setminus C_1$. For a nonnegative integer t we define

$$x^{(t)} := \chi^{C_2} + 2^{-(t+1)} \chi^{C_1}.$$

If $t < \log k - 1$, then

$$x^{(t)}(C) - x^{(t)}(\mathrm{mtr}(C)) = \chi^{C_2}\chi^{C_2} + 2^{-(t+1)}\chi^{C_1}\chi^{C_1} = |C| - k + 2^{-(t+1)}k > |C| - k + 1,$$

and so $x^{(t)}$ fails to satisfy the generalized $(k, 2)$-clique inequality (4.3). It remains to show that $x^{(t)} \in P^{(t)}$ for all t. For this we use Lemma 4.13 with $y := \chi^{C_2} + \frac{1}{2}\chi^{C_1}$, $z := \chi^{C_1}$, and $\mu_t := 2^{t+1}$. Observe that y is a solution to (2.1)–(2.3). Now consider an arbitrary inequality $ax \leqslant \beta$, valid for $P_{\mathrm{TP}}(N, \mathcal{H}, \mathrm{tr})$ and such that $a \in \mathbb{Z}^N$ and $a\chi^{C_1} < \mu_t$. We need to verify that $ax^{(t)} \leqslant \beta$. Whereas this is obvious if $a\chi^{C_1} \leqslant 0$, in the case $a\chi^{C_1} > 0$ we have

$$ax^{(t)} = a\chi^{C_2} + \frac{1}{\mu_t} a\chi^{C_1} < a\chi^{C_2} + 1 \leqslant a(\chi^{C_2} + \chi^{\{u_i\}}) \leqslant \beta$$

for a representative u_i such that $a_{u_i} \geqslant 1$. The last inequality follows from $\chi^{C_2} + \chi^{\{u_i\}} \in P_{\mathrm{TP}}(N, \mathcal{H}, \mathrm{tr})$. □

Theorem 4.12 was proved before for the special instances formed by the clique inequalities of the node packing polytope [11] and by the simple k-fence inequalities of the acyclic subdigraph polytope [12]. Notice that the assumption of Theorem 4.12 is also satisfied by the k-fence inequalities since each dicycle contained in a fence uses pales between at least two different pairs of nodes. Moreover, Theorem 4.12 also applies to the $(1, k)$-2-partition inequalities of the clique partitioning polytope.

4.3. Generalized antihole inequalities. Another class of valid inequalities for the node packing polytope is supported by odd antiholes. An *odd antihole* in a graph is the complement of an odd cycle of length at least five without a chord. Let O denote the set of vertices of an odd antihole. Then the *odd antihole inequality* associated with O is

$$x(O) \leqslant 2.$$

Again, it turns out that these inequalities form a special case of a more general principle.

DEFINITION 4.14. *Let $(N, \mathcal{H})$ be a hypergraph, and let q and s be integers such that $s \geqslant q \geqslant 2$. For convenience, we set $k := qs + 1$. Let $N_1, N_2, \dots, N_k$ be a sequence of mutually disjoint nonempty subsets of the node set N. Moreover, for each $\ell \in \{1, 2, \dots, k\}$ and for every q-element set of indices $\{i_1, i_2, \dots, i_q\} \subseteq \{\ell, \ell + 1, \dots, \ell + s - 1\}$ (where indices greater than k are taken modulo $k + 1$ and shifted by $+1$) we let the set $H^{\ell}_{i_1, i_2, \dots, i_q}$ be an edge such that $\bigcup_{j=1}^{q} N_{i_j} \subseteq H^{\ell}_{i_1, i_2, \dots, i_q}$. In addition, we assume, for each $\ell \in \{1, 2, \dots, k\}$, that the edges in any collection of intersecting edges of type $H^{\ell}_{i_1, i_2, \dots, i_q}$ all have one common (sub)index. We denote by O^{ℓ} the union of these edges, $O^{\ell} := \bigcup_{\ell \leqslant i_1 < i_2 < \cdots < i_q \leqslant \ell + s - 1} H^{\ell}_{i_1, i_2, \dots, i_q}$, and by O the union of all these edges, $O := \bigcup_{\ell=1}^{k} O^{\ell}$. Moreover, let $\widetilde{m}(u) := |\{\ell \in \{1, 2, \dots, k\} : u \in O^{\ell}\}|$ for a node $u \in O$. We assume that $\widetilde{m}(u) \leqslant s$ for all nodes $u \in O$. Then the hypergraph*

$$(O, \{H^{\ell}_{i_1, i_2, \dots, i_q} : \ell \leqslant i_1 < i_2 < \cdots < i_q \leqslant \ell + s - 1 \text{ for some } \ell \in \{1, 2, \dots, k\}\})$$

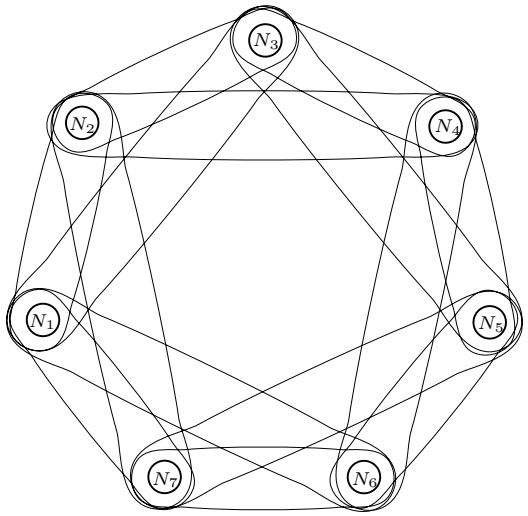


FIG. 4.6. *A generalized* $(3,2)$*-antihole (with* $O = \bigcup_{i=1}^{7} N_i$*).*

is called a generalized (s,q)-antihole (*contained in* $(N,\mathcal{H})$).

Figure 4.6 depicts a generalized $(3,2)$-antihole. Notice that it may happen that the same edge wears different names. For instance, if $O = \bigcup_{i=1}^{k} N_i$ and $q < s$, then $H^{\ell}_{\ell+s-q,\ldots,\ell+s-1} = H^{\ell+1}_{\ell+s-q,\ldots,\ell+s-1}$. Given a generalized antihole that is contained in a given extended hypergraph, we define $\widetilde{n}(u)$ to be the multiplicity of a node u contained in the transitive sets associated with that generalized antihole, i.e., $\widetilde{n}(u) := |\{H^{\ell}_{i_1,i_2,\ldots,i_q} : u \in \mathrm{tr}(H^{\ell}_{i_1,i_2,\ldots,i_q}) \text{ for some } \ell \in \{1,2,\ldots,k\}, \ell \leqslant i_1 < i_2 < \cdots < i_q \leqslant \ell+s-1\}|$. Thus, if the same edge occurs more often under different names, we count the number of names. We set $\mathrm{tr}(O) := \bigcup_{\ell=1}^{k} (\bigcup_{\ell \leqslant i_1 < i_2 < \cdots < i_q \leqslant \ell+s-1} \mathrm{tr}(H^{\ell}_{i_1,i_2,\ldots,i_q}))$.

THEOREM 4.15. *Let* $(N,\mathcal{H},\mathrm{tr})$ *be an extended hypergraph, and let the hypergraph* $(O, \{H^{\ell}_{i_1,i_2,\ldots,i_q} : \ell \leqslant i_1 < i_2 < \cdots < i_q \leqslant \ell+s-1 \text{ for some } \ell \in \{1,2,\ldots,k\}\})$ *be a generalized* (s,q)*-antihole in* $(N,\mathcal{H})$ *such that* $\mathrm{tr}(O) \cap O = \emptyset$*. Then, the generalized* (s,q)*-antihole inequality*

$$\sum_{u \in O} x_u - \sum_{u \in \mathrm{tr}(O)} \frac{\lceil \widetilde{n}(u) \rceil_s}{s} x_u \leqslant |O| - q(s-q+1) - 1 \tag{4.6}$$

is valid for $P_{\mathrm{TP}}(N,\mathcal{H},\mathrm{tr})$*. It has a cutting plane proof from* (2.1)–(2.3) *of depth at most* $\lceil \log(s-1) \rceil + 1$*.*

Proof. Let $N_1, N_2, \ldots, N_k$ be the sequence of nodes underlying the generalized (s,q)-antihole. Notice that for every $\ell \in \{1,2,\ldots,k\}$ the edges $\{H^{\ell}_{i_1,i_2,\ldots,i_q} : \ell \leqslant i_1 < i_2 < \cdots < i_q \leqslant \ell+s-1\}$ induce a generalized (s,q)-clique. Each set N_i is contained in precisely s of these k cliques. By adding up the k associated (s,q)-clique inequalities, the appropriate number of upper bound constraints $x_u \leqslant 1$ for $u \in O \setminus \bigcup_{i=1}^{k} N_i$ (namely, $s - \widetilde{m}(u)$ many), as well as the appropriate number of nonnegativity constraints for each element $u \in \mathrm{tr}(O)$ (namely, $\lceil \widetilde{n}(u) \rceil_s - \widetilde{n}(u)$), we obtain that

$$s \sum_{u \in O} x_u - \sum_{u \in \mathrm{tr}(O)} \lceil \widetilde{n}(u) \rceil_s x_u \leqslant s|O| - k(s-q+1)$$

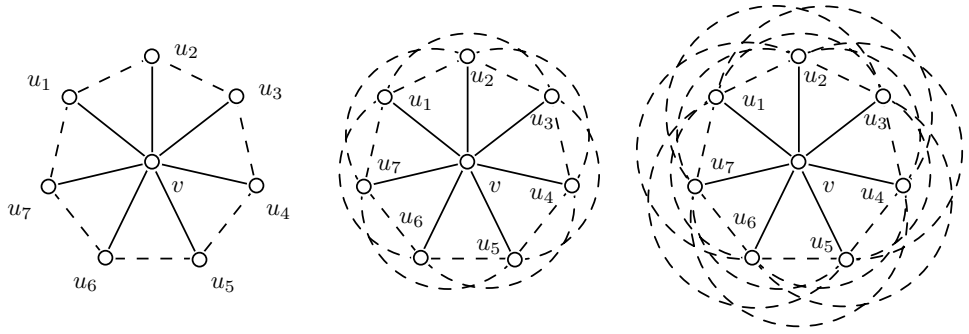


FIG. 4.7. *From left to right: a generalized cycle, a generalized antihole, and a generalized clique with associated transitive elements in the case of the clique partitioning problem.*

is valid for $P_{\mathrm{TP}}(N, \mathcal{H}, \mathrm{tr})$. Division by s and taking the floor of the right-hand side gives the desired inequality. The bound on the depth of its cutting plane proof follows immediately from that for the generalized clique inequalities. □

To see that we indeed derive from Theorem 4.15 the usual odd antihole inequalities for the node packing polytope of a graph G, we proceed as follows. Let O be the node set of an odd antihole in G, $O = \{u_1, u_2, \ldots, u_k\}$, and assume that u_ℓ and $u_{\ell+s}$ as well as u_ℓ and $u_{\ell+s+1}$ are not adjacent, for $\ell = 1, 2, \ldots, k$. We now relate this to a generalized antihole. Clearly, $q = 2$, and hence $|O| = k = 2s + 1$. It remains to identify the edges. For $\ell \in \{1, 2, \ldots, k\}$ we take as edges H^ℓ_{ij} the edges of the clique induced by the nodes $u_\ell, u_{\ell+1}, \ldots, u_{\ell+s-1}$. Notice that several edges in G are taken more than once but under different names. Finally, observe that the right-hand side of (4.6) simplifies to 2.

Since line graphs do not contain odd antiholes (with more than five nodes), there do not exist generalized antihole inequalities for the clique partitioning polytope when we assume that $s \geqslant 3$ and that u_ℓ and $u_{\ell+s}$ as well as u_ℓ and $u_{\ell+s+1}$ are not linked by a hyperedge, for each $\ell = 1, 2, \ldots, k$. Others may well exist; see, for instance, Figure 4.7. We record this as a lemma.

LEMMA 4.16. *Let $G = (V, E)$ be a graph, and let $v, u_1, u_2, \ldots, u_k \in V$, be distinct nodes such that $\{v, u_i\} \in E$, for $i = 1, 2, \ldots, k$, for $k = 2s + 1$, and $s \geqslant 2$. Define the set $T := \{\{u_i, u_j\} \in E : \ell \leqslant i < j \leqslant \ell + s - 1$ for some $\ell \in \{1, 2, \ldots, k\}\}$. The inequality*

$$\sum_{i=1}^{k} x_{\{v,u_i\}} - \sum_{\{u_i,u_j\}\in T} x_{\{u_i,u_j\}} \leqslant 2$$

is valid for the clique partitioning polytope of G. (Again, indices greater than k are taken modulo $k + 1$ and shifted by $+1$.)

We note that generalized $(2, 2)$-antihole inequalities of the transitive packing polytope coincide with generalized $(5, 2)$-cycle inequalities. So far, antihole inequalities have not been exploited for the acyclic subdigraph polytope or the partition polytope.

4.4. Generalized antiweb inequalities. The main idea in the derivation of the generalized antihole inequalities was to combine generalized clique inequalities in a manner oriented on the cutting plane proof of generalized cycle inequalities. This can be generalized and leads for the node packing polytope to the antiweb inequalities [52].

For integers $1 \leqslant s \leqslant k$, a (k,s)-*antiweb* is a graph with node set $W = \{u_1, u_2, \ldots, u_k\}$ such that each node u_i is adjacent to all other nodes but not to the $\max\{0, k-2s+1\}$ nodes $u_{i+s}, u_{i+s+1}, \ldots, u_{i+k-s}$. (Again, indices greater than k are taken modulo $k+1$ and shifted by $+1$.) The associated *antiweb inequality* is

$$x(W) \leqslant \left\lfloor \frac{k}{s} \right\rfloor.$$

We proceed by introducing special hypergraphs that we call generalized antiwebs.

DEFINITION 4.17. *Let $(N,\mathcal{H})$ be a hypergraph, and let k, s, and q be integers such that $k \geqslant s \geqslant q \geqslant 2$. Let $N_1, N_2, \ldots, N_k$ be a sequence of mutually disjoint nonempty subsets of the node set N. For each $\ell \in \{1,2,\ldots,k\}$ and each q-element set of indices $\{i_1, i_2, \ldots, i_q\} \subseteq \{\ell, \ell+1, \ldots, \ell+s-1\}$ (where indices are taken modulo $k+1$ and shifted by $+1$), we let $H^\ell_{i_1,i_2,\ldots,i_q} \in \mathcal{H}$ be an edge such that $\bigcup_{j=1}^q N_{i_j} \subseteq H^\ell_{i_1,i_2,\ldots,i_q}$. In addition, we assume, for each $\ell \in \{1,2,\ldots,k\}$, that the edges in any collection of intersecting edges of type $H^\ell_{i_1,i_2,\ldots,i_q}$ all have one common (sub)index. For each ℓ, we denote by W^ℓ the union of the associated edges, $W^\ell := \bigcup_{\ell \leqslant i_1 < i_2 < \cdots < i_q \leqslant \ell+s-1} H^\ell_{i_1,i_2,\ldots,i_q}$. Moreover, we let W denote the union of all these edges, $W := \bigcup_{\ell=1}^k W^\ell$. Again, for $u \in W$ we let $\widetilde{m}(u)$ be the multiplicity of u with respect to its occurrence in W^ℓ, $\ell = 1,2,\ldots,k$, i.e., $\widetilde{m}(u) := |\{\ell \in \{1,2,\ldots,k\} : u \in W^\ell\}|$. If $\widetilde{m}(u) \leqslant s$ for all $u \in W$, then we call the hypergraph*

$$(W, \{H^\ell_{i_1,i_2,\ldots,i_q} : \ell \leqslant i_1 < i_2 < \cdots < i_q \leqslant \ell+s-1 \text{ for some } \ell \in \{1,2,\ldots,k\}\})$$

a generalized (k,s,q)-antiweb (*contained in* $(N,\mathcal{H})$).

THEOREM 4.18. *Let $(N,\mathcal{H},\mathrm{tr})$ be an extended hypergraph, and let the hypergraph $(W, \{H^\ell_{i_1,i_2,\ldots,i_q} : \ell \leqslant i_1 < i_2 < \cdots < i_q \leqslant \ell+s-1 \text{ for some } \ell \in \{1,2,\ldots,k\}\})$ be a generalized (k,s,q)-antiweb in $(N,\mathcal{H})$ such that $\mathrm{tr}(W) \cap W = \emptyset$. Then, the generalized (k,s,q)-antiweb inequality*

$$\sum_{u \in W} x_u - \sum_{u \in \mathrm{tr}(W)} \frac{\lceil \widetilde{n}(u) \rceil_s}{s} x_u \leqslant \left\lfloor \frac{s|W| - k(s-q+1)}{s} \right\rfloor \tag{4.7}$$

is valid for $P_{\mathrm{TP}}(N,\mathcal{H},\mathrm{tr})$. It has a cutting plane proof from (2.1)–(2.3) of depth at most $\lceil \log(s-1) \rceil + 1$. Here, $\mathrm{tr}(W) := \bigcup_{\ell=1}^k (\bigcup_{\ell \leqslant i_1 < i_2 < \cdots < i_q \leqslant \ell+s-1} \mathrm{tr}(H^\ell_{i_1,i_2,\ldots,i_q}))$ and $\widetilde{n}(u) := |\{H^\ell_{i_1,i_2,\ldots,i_q} : u \in \mathrm{tr}(H^\ell_{i_1,i_2,\ldots,i_q})$ for some $\ell \in \{1,2,\ldots,k\}$, $\ell \leqslant i_1 < i_2 < \cdots < i_q \leqslant \ell+s-1\}|$.

Proof. The cutting plane proof goes along the line of the proof of the validity of generalized antihole inequalities (Theorem 4.15) and is therefore omitted. □

It follows from their construction that generalized (k,s,q)-antiweb inequalities subsume all the former classes of inequalities for the transitive packing polytope $P_{\mathrm{TP}}(N,\mathcal{H},\mathrm{tr})$. In fact,

- if $q = s$ and if s does not divide k, we obtain the class of generalized (k,q)-cycle inequalities;
- if $s = k$, the class of generalized antiweb inequalities contains the class of generalized (k,q)-clique inequalities;
- if $k = qs+1$, we have the class of generalized (s,q)-antihole inequalities.

Laurent [32] previously extended antiwebs to the independence system polytope; however, the inequalities (4.7) restricted to this setting are more general. Laurent

used one-element sets N_i and edges that are precisely the union of q of these. She showed that such an inequality is facet defining for the polytope associated with the independence system defined by the circuits of her antiweb.

4.5. Odd partition inequalities. In this section, we introduce another new class of inequalities for the transitive packing polytope. It is an extension of a class of inequalities recently proposed by Caprara and Fischetti [9] for the acyclic subdigraph polytope.

Assume that we are given an extended hypergraph $(N, \mathcal{H}, \mathrm{tr})$. Let $H_1, \ldots, H_k$ be a collection of distinct edges of $\mathcal{H}$, and let $m(u)$ and $n(u)$ denote the multiplicity of a node $u \in N$ in this collection and the associated set of transitive elements, respectively. That is, $m(u) := |\{i \in \{1, \ldots, k\} : u \in H_i\}|$ and $n(u) := |\{i \in \{1, \ldots, k\} : u \in \mathrm{tr}(H_i)\}|$. We denote the difference of these two numbers by $d(u)$, $d(u) := m(u) - n(u)$. Let W be the union of all the nodes involved, $W := \bigcup_{i=1}^{k}(H_i \cup \mathrm{tr}(H_i))$, and let W^{odd} be the set of those nodes that occur either in an odd number of edges H_i or in an odd number of transitive sets $\mathrm{tr}(H_i)$ but not both, $W^{\mathrm{odd}} := \{u \in W : d(u) \text{ odd}\}$. Furthermore, let $(W_1^{\mathrm{odd}}, W_2^{\mathrm{odd}})$ be a partition of W^{odd} such that $\sum_{i=1}^{k}|H_i| + |W_1^{\mathrm{odd}}| - k$ is odd. ($W_1^{\mathrm{odd}} = \emptyset$ or $W_2^{\mathrm{odd}} = \emptyset$ is possible.)

Taking the sum of the constraints

$$\begin{aligned} \sum_{u \in H_i} x_u - \sum_{u \in \mathrm{tr}(H_i)} x_u &\leqslant |H_i| - 1 \quad \text{for } i = 1, \ldots, k, \\ x_u &\leqslant 1 \qquad \text{for } u \in W_1^{\mathrm{odd}}, \\ -x_u &\leqslant 0 \qquad \text{for } u \in W_2^{\mathrm{odd}}, \end{aligned}$$

and dividing the result by 2, we obtain

$$\begin{aligned} \sum_{u \in W \setminus W^{\mathrm{odd}}} \frac{d(u)}{2} x_u + \sum_{u \in W_1^{\mathrm{odd}}} \frac{d(u)+1}{2} x_u \\ + \sum_{u \in W_2^{\mathrm{odd}}} \frac{d(u)-1}{2} x_u \leqslant \frac{\sum_{i=1}^{k}|H_i| + |W_1^{\mathrm{odd}}| - k}{2}. \end{aligned} \tag{4.8}$$

Rounding down the right-hand side gives the following inequality that is valid for the transitive packing polytope $P_{\mathrm{TP}}(N, \mathcal{H}, \mathrm{tr})$,

$$\begin{aligned} \sum_{u \in W \setminus W^{\mathrm{odd}}} \frac{d(u)}{2} x_u + \sum_{u \in W_1^{\mathrm{odd}}} \frac{d(u)+1}{2} x_u \\ + \sum_{u \in W_2^{\mathrm{odd}}} \frac{d(u)-1}{2} x_u \leqslant \frac{\sum_{i=1}^{k}|H_i| + |W_1^{\mathrm{odd}}| - k - 1}{2}. \end{aligned} \tag{4.9}$$

We call inequalities of type (4.9) *odd partition inequalities*. We continue by pointing out some special cases in which inequality (4.9) is dominated by other inequalities, as well as some other cases in which it has depth 1 relative to (2.1)–(2.3) and is therefore interesting.

LEMMA 4.19. *Let $(N, \mathcal{H}, \mathrm{tr})$ be a hypergraph with associated transitive elements, and let $H_1, \ldots, H_k$ be a collection of distinct edges of $\mathcal{H}$. If $(H_k \cup \mathrm{tr}(H_k)) \cap \bigcup_{i=1}^{k-1}(H_i \cup \mathrm{tr}(H_i)) = \emptyset$, then the odd partition inequality* (4.9) *for $H_1, \ldots, H_k$ is implied by the initial inequalities* (2.1)–(2.3) *and inequality* (4.9) *for $H_1, \ldots, H_{k-1}$.*

Proof. Observe first that $H_k \cup \mathrm{tr}(H_k) \subseteq W^{\mathrm{odd}}$. Thus the left-hand side of inequality (4.9) can be expressed as follows:

$$\sum_{u \in W \setminus W^{\mathrm{odd}}} \frac{d(u)}{2} x_u + \sum_{u \in W_1^{\mathrm{odd}} \setminus (H_k \cup \mathrm{tr}(H_k))} \frac{d(u)+1}{2} x_u + \sum_{u \in W_2^{\mathrm{odd}} \setminus (H_k \cup \mathrm{tr}(H_k))} \frac{d(u)-1}{2} x_u + \sum_{u \in H_k \cap W_1^{\mathrm{odd}}} x_u - \sum_{u \in \mathrm{tr}(H_k) \cap W_2^{\mathrm{odd}}} x_u.$$

Notice that the first three terms precisely form the left-hand side of inequality (4.8) for $H_1, \ldots, H_{k-1}$ (where we use the natural restriction of W_1^{odd} and W_2^{odd}). We continue by distinguishing three cases, namely,

(i) $|(H_k \cap W_2^{\mathrm{odd}}) \cup (\mathrm{tr}(H_k) \cap W_1^{\mathrm{odd}})| \geqslant 2$,
(ii) $|(H_k \cap W_2^{\mathrm{odd}}) \cup (\mathrm{tr}(H_k) \cap W_1^{\mathrm{odd}})| = 1$, and finally,
(iii) $|(H_k \cap W_2^{\mathrm{odd}}) \cup (\mathrm{tr}(H_k) \cap W_1^{\mathrm{odd}})| = 0$.

In case (i), we add to inequality (4.8) for $H_1, \ldots, H_{k-1}$ the inequalities

$$x_u \leqslant 1 \text{ for } u \in H_k \cap W_1^{\mathrm{odd}} \quad \text{and} \quad -x_u \leqslant 0 \text{ for } u \in \mathrm{tr}(H_k) \cap W_2^{\mathrm{odd}}.$$

Then the left-hand side of the resulting inequality coincides with that of inequality (4.9). The numerator of the right-hand side is

$$\sum_{i=1}^{k-1} |H_i| + |W_1^{\mathrm{odd}} \setminus (H_k \cup \mathrm{tr}(H_k))| - k + 1 + 2|H_k \cap W_1^{\mathrm{odd}}| = \sum_{i=1}^{k} |H_i| + |W_1^{\mathrm{odd}}| - k + 1 - (|H_k \cap W_2^{\mathrm{odd}}| + |\mathrm{tr}(H_k) \cap W_1^{\mathrm{odd}}|),$$

which is, because of assumption (i), less than or equal to

$$\sum_{i=1}^{k} |H_i| + |W_1^{\mathrm{odd}}| - k - 1,$$

which is the numerator of the right-hand side of inequality (4.9) for $H_1, \ldots, H_k$. Hence in this case inequality (4.9) has depth 0 relative to (2.1)–(2.3).

Since we assumed $\sum_{i=1}^{k} |H_i| + |W_1^{\mathrm{odd}}| - k$ to be odd in order to derive inequality (4.9), the assumption in case (ii) guarantees that the numerator of the right-hand side of inequality (4.8) for $H_1, \ldots, H_{k-1}$ will be odd, too. Hence, the following inequality is valid for $P_{\mathrm{TP}}(N, \mathcal{H}, \mathrm{tr})$, which is inequality (4.9) for $H_1, \ldots, H_{k-1}$:

$$\sum_{u \in W \setminus W^{\mathrm{odd}}} \frac{d(u)}{2} x_u + \sum_{u \in W_1^{\mathrm{odd}} \setminus (H_k \cup \mathrm{tr}(H_k))} \frac{d(u)+1}{2} x_u + \sum_{u \in W_2^{\mathrm{odd}} \setminus (H_k \cup \mathrm{tr}(H_k))} \frac{d(u)-1}{2} x_u \leqslant \frac{\sum_{i=1}^{k-1} |H_i| + |W_1^{\mathrm{odd}} \setminus (H_k \cup \mathrm{tr}(H_k))| - k}{2}.$$

By adding to this inequality the inequalities

$$x_u \leqslant 1 \text{ for } u \in H_k \cap W_1^{\mathrm{odd}} \quad \text{and} \quad -x_u \leqslant 0 \text{ for } u \in \mathrm{tr}(H_k) \cap W_2^{\mathrm{odd}},$$

we obtain inequality (4.9), which is therefore implied by (4.9) for $H_1, \ldots, H_{k-1}$ and the bound constraints (2.2) and (2.3).

In case (iii), we simply add the transitivity constraint (2.1) for H_k to inequality (4.8) for $H_1, \ldots, H_{k-1}$. It follows that inequality (4.9) again has depth 0 relative to system (2.1)–(2.3). □

Lemma 4.19 reflects, in particular, the trivial fact that we cannot hope to obtain a stronger inequality by adding inequalities with mutually disjoint support. We now present a condition that is sufficient to ensure that inequality (4.9) has depth 1, which leads us back to cycles in the hypergraph $(N, \mathcal{H})$.

LEMMA 4.20. *Let $(N, \mathcal{H}, \mathrm{tr})$ be an extended hypergraph, and let $H_1, \ldots, H_k$ be a collection of distinct edges in $\mathcal{H}$, $k \geqslant 2$. Let the sets W^{odd}, W_1^{odd}, and W_2^{odd} be defined as before. Assume that $\mathrm{tr}(H_j) \cap \bigcup_{i=1}^k H_i = \emptyset$ for $j = 1, \ldots, k$. If*

- *there exist k distinct nodes $u_1, \ldots, u_k \in N$ such that $u_i \in H_i \cap H_{i+1}$ but $u_i \notin H_j$ for $j \neq i, i+1$,*
- *the transitive set $\mathrm{tr}(H)$ of an edge $H \neq H_i$ $(i = 1, \ldots, k)$ that satisfies $H \subseteq \bigcup_{i=1}^k H_i$ intersects $\bigcup_{i=1}^k H_i$ either in at least one node different from $u_1, \ldots, u_k$ or in at least two nodes from $u_1, \ldots, u_k$, and*
- *$W_1^{\mathrm{odd}} \subseteq \left(\bigcup_{i=1}^k H_i\right) \setminus \{u_1, u_2, \ldots, u_k\}$,*

then the depth of the odd partition inequality (4.9) *relative to* (2.1)–(2.3) *is* 1.

Proof. Define the point $x \in \mathbb{R}^N$ as follows:

$$x_u := \begin{cases} 1/2 & \text{if } u \in \{u_1, \ldots, u_k\}, \\ 1 & \text{if } u \in \left(\bigcup_{i=1}^k H_i\right) \setminus \{u_1, \ldots, u_k\}, \\ 0 & \text{otherwise.} \end{cases}$$

Whereas x belongs to the initial linear relaxation of $P_{\mathrm{TP}}(N, \mathcal{H}, \mathrm{tr})$, i.e., satisfies inequalities (2.1)–(2.3), it violates inequality (4.9). Hence this inequality is not implied by the initial system. □

As mentioned before, Caprara and Fischetti [9] introduced the odd partition inequalities for the acyclic subdigraph polytope in order to show that a subclass of the Möbius ladder inequalities can be derived from the initial relaxation by a cutting plane proof of length 1, where all coefficients used are either 0 or $\frac{1}{2}$. Indeed, if $(C, \{H_i : i = 1, 2, \ldots, k\})$ is a generalized $(k, 2)$-cycle, we obtain the associated generalized $(k, 2)$-cycle inequality as an odd partition inequality by setting $W_1^{\mathrm{odd}} := \{u \in C : m(u) \text{ odd}\}$ and $W_2^{\mathrm{odd}} := \{u \in \mathrm{tr}(C) : n(u) \text{ odd}\}$. In section 4.1, we showed that the subclass of Möbius ladder inequalities where each triple of participating dicycles has an empty intersection is contained in the class of generalized $(k, 2)$-cycle inequalities for the acyclic subdigraph polytope. This implies Caprara and Fischetti's result.

5. Transitive packing in graphs. An important subproblem of the transitive packing problem is formed by the instances where the given hypergraph is actually a graph. This section is devoted to discussing the polytopes associated with these instances in more detail. To avoid confusion, we still use the notation $(N, \mathcal{H}, \mathrm{tr})$ but assume throughout this section that $|H| = 2$ for all $H \in \mathcal{H}$. We call the triple $(N, \mathcal{H}, \mathrm{tr})$ an *extended graph*. The transitive packing polytope is then given as

$$P_{\mathrm{TP}}(N, \mathcal{H}, \mathrm{tr}) = \mathrm{conv}\left\{ x \in \{0,1\}^N : x_u + x_v - \sum_{w \in \mathrm{tr}(\{u,v\})} x_w \leqslant 1 \text{ for } \{u, v\} \in \mathcal{H} \right\}.$$

Recall that both the node packing polytope and the clique partitioning polytope are of this flavor. For the node packing polytope, it is known that all facet defining inequalities with right-hand side 1 are clique inequalities; see [42]. This remains true for the transitive packing polytope of the following extended graphs.

THEOREM 5.1. *Let $(N, \mathcal{H}, \mathrm{tr})$ be an extended graph such that for every clique C in $(N, \mathcal{H})$ the following condition is satisfied:*

Each node $u \in \mathrm{tr}(C)$ belongs to $\mathrm{tr}(\{v, w\})$ for a unique edge $\{v, w\}$ induced by C and satisfies either

- *$\{u, v\}, \{u, w\} \notin \mathcal{H}$, or*
- *$\{u, v\} \notin \mathcal{H}$, $\{u, w\} \in \mathcal{H}$, and $v \in \mathrm{tr}(\{u, w\})$, or*
- *$\{u, w\} \notin \mathcal{H}$, $\{u, v\} \in \mathcal{H}$, and $w \in \mathrm{tr}(\{u, v\})$, or*
- *$\{u, v\}, \{u, w\} \in \mathcal{H}$, and $v \in \mathrm{tr}(\{u, w\})$ and $w \in \mathrm{tr}(\{u, v\})$.*

Then, any facet defining inequality $cx \leqslant 1$ (with c integral) of the transitive packing polytope $P_{\mathrm{TP}}(N, \mathcal{H}, \mathrm{tr})$ either is of the form $x_u \leqslant 1$ or is a generalized $(k, 2)$-clique inequality.

Proof. Since every singleton is a transitive packing, the coefficients of the vector c have value at most 1. If c has exactly one coefficient with value 1, indexed by, say, $u \in N$, then $c = \chi^{\{u\}}$. Otherwise, $cx \leqslant 1$ would be dominated by $x_u \leqslant 1$. So we may assume from now on that the number of coefficients of c with value 1 is at least two. Let C be the set of nodes u such that $c_u = 1$. Since $cx \leqslant 1$ is valid, the nodes in C have to be pairwise adjacent, i.e., they induce a clique in $(N, \mathcal{H})$. From this validity it also follows that $\mathrm{tr}(C) \cap C = \emptyset$. It remains to be observed that the coefficient c_u of a transitive element $u \in \mathrm{tr}(C)$ is not zero. This follows from the assumptions with respect to transitive elements and the validity of $cx \leqslant 1$ for $P_{\mathrm{TP}}(N, \mathcal{H}, \mathrm{tr})$. We just need to observe that the node set formed by u and the pair of nodes $v, w \in C$ such that $u \in \mathrm{tr}(\{v, w\})$ is a transitive packing in $(N, \mathcal{H}, \mathrm{tr})$. □

The assumptions made in Theorem 5.1 are satisfied, for instance, by the extended graphs corresponding to instances of the clique partitioning problem. Hence, if a graph G has no isolated edges, $(1, k)$-2-partition inequalities are the only facet defining inequalities with right-hand side 1 of the clique partitioning polytope of G. The latter observation was independently made in [41].

Notice that the assumptions of Theorem 4.12 are always satisfied for transitive packing problems in graphs. Consequently, the generalized $(k, 2)$-clique inequalities have depth at least $\log k - 1$, relative to (2.1)–(2.3).

If the transitive elements of a clique C do not interact with C itself, the clique and its transitive elements form the support of valid inequalities, where the nodes of the cliques have coefficients greater than one.

THEOREM 5.2. *Let $(N, \mathcal{H}, \mathrm{tr})$ be an extended graph, and let C be the node set of a generalized $(k, 2)$-clique in $(N, \mathcal{H})$ such that $\mathrm{tr}(C) \cap C = \emptyset$. Moreover, let $t \geqslant 1$ be an integer. Then, the t-reinforced generalized $(k, 2)$-clique inequality*

$$tx(C) - x(\mathrm{mtr}(C)) \leqslant \frac{t(t+1)}{2} \tag{5.1}$$

is valid for the transitive packing polytope $P_{\mathrm{TP}}(N, \mathcal{H}, \mathrm{tr})$.

Proof. Let x be the incidence vector of a transitive packing in $(N, \mathcal{H}, \mathrm{tr})$, and assume that $x(C) = \mu$. Consequently, $x(\mathrm{mtr}(C)) \geqslant \mu(\mu - 1)/2$. Thus the left-hand side of inequality (5.1) is less than or equal to $t\mu - \mu(\mu - 1)/2$. Since

$$t\mu = \frac{\mu(\mu-1)}{2} + \frac{t(t+1)}{2} - \frac{(t-\mu)(t-\mu+1)}{2}$$

and the last term is nonnegative, x satisfies inequality (5.1). □

The proof of Theorem 5.2 implies immediately that the faces of two nonempty face defining t-reinforced generalized $(k,2)$-clique inequalities with the same support but different values of t in general contain different sets of incidence vectors of transitive packings. The proof also implies a range on t in order to ensure that the intersection of the transitive packing polytope and the hyperplane defined by a t-reinforced generalized $(k,2)$-clique inequality is nonempty.

COROLLARY 5.3. *Let $(N,\mathcal{H},\mathrm{tr})$ be an extended graph, and let C be the node set of a generalized $(k,2)$-clique in $(N,\mathcal{H})$ such that $\mathrm{tr}(C)\cap C=\emptyset$. Let $t\geqslant 1$ be an integer. If the t-reinforced generalized $(k,2)$-clique inequality (5.1) defines a nonempty face of the transitive packing polytope $P_{\mathrm{TP}}(N,\mathcal{H},\mathrm{tr})$, then $t\leqslant |C|$.*

The bound on t can be strengthened if we assume that the t-reinforced generalized $(k,2)$-clique inequality is facet defining.

LEMMA 5.4. *Let $(N,\mathcal{H},\mathrm{tr})$ be an extended graph, and let C be the node set of a generalized $(k,2)$-clique in $(N,\mathcal{H})$ such that $\mathrm{tr}(C)\cap C=\emptyset$. Let $t\geqslant 1$ be an integer. If the t-reinforced generalized $(k,2)$-clique inequality (5.1) induces a facet of the transitive packing polytope $P_{\mathrm{TP}}(N,\mathcal{H},\mathrm{tr})$, then $t\leqslant |C|-2$.*

Proof. The proof is by contradiction. Because of Corollary 5.3, we are left with the cases $t=|C|$ and $t=|C|-1$. In the former case each point x contained in the facet under consideration would satisfy $x(C)=|C|$. Hence this facet would be contained in all faces induced by the upper bound constraints $x_u\leqslant 1$ for $u\in C$, a contradiction. In the latter case the $(|C|-1)$-reinforced generalized $(k,2)$-clique inequality (5.1) turns out to be the sum of all the transitivity constraints induced by pairs of nodes of the clique C, again a contradiction. □

One might ask whether there exist transitive packing polytopes of extended graphs such that the t-reinforced generalized $(k,2)$-clique inequalities are facet defining. This is indeed the case. Oosten, Rutten, and Spieksma [41] showed that the t-reinforced generalized $(k,2)$-clique inequalities define facets of the clique partitioning polytope of a complete graph, for $t\leqslant k-2$ of course.

One appealing aspect of our suggestion to treat suitable problems in the transitive packing context is the opportunity to use knowledge that is available, not only for the transitive packing polytope itself but also for some of its special cases. We ellucidate this by considering a simple example. Let us assume that the underlying graph G of a clique partitioning problem is bipartite. This implies for the associated extended graph $(N,\mathcal{H},\mathrm{tr})$ that $\mathrm{tr}(H)=\emptyset$ for all edges $H\in\mathcal{H}$. In other words, the transitive packing (clique partitioning) polytope of G coincides with the node packing polytope of its line graph $(N,\mathcal{H})$. Since node packings in line graphs correspond one-to-one with matchings in the original graphs, we obtain the following result.

LEMMA 5.5. *Let $G=(V,E)$ be a bipartite graph. The clique partitioning polytope of G is completely characterized by the following linear inequalities:*

$$\begin{aligned} x_e &\geqslant 0 \quad \textit{for all edges } e\in E,\\ x(C) &\leqslant 1 \quad \textit{for all sets } C\subseteq E \textit{ of pairwise incident edges.}\end{aligned}$$

It also follows that the clique partitioning problem on bipartite graphs reduces to a matching problem and can hence be solved in polynomial time. This example is, as already indicated, an instance of a more general point of view. Whenever we can interpret a given problem as a transitive packing problem, and whenever the extended graph (or even hypergraph) of an instance of this problem does not have transitive

elements but does have a structure such that the corresponding node packing (independence system) polytope can explicitly be described by linear inequalities, the same holds for the polytope associated with the original problem.

6. Separation. After introducing several classes of valid inequalities for the transitive packing polytope, one question that arises is whether we can use these inequalities efficiently in cutting plane algorithms for attacking the transitive packing problem. This topic is discussed in this section. We concentrate on generalized cycle and odd partition inequalities.

Given an integer polyhedron $P_I = \text{conv}\{x \in \mathbb{Z}^n : Ax \leqslant b\}$, where $A \in \mathbb{Z}^{m\times n}$ and $b \in \mathbb{Z}^m$, a $\{0, \frac{1}{2}\}$-*Gomory–Chvátal cut* is a valid inequality for P_I of the form $\lambda Ax \leqslant \lfloor \lambda b \rfloor$, with $\lambda \in \{0, \frac{1}{2}\}^m$ and $\lambda A \in \mathbb{Z}^n$. In other words, a $\{0, \frac{1}{2}\}$-Gomory–Chvátal cut has a cutting plane proof of length 1 from $Ax \leqslant b$, and the coefficients in the corresponding linear combination belong to $\{0, \frac{1}{2}\}$ only. Caprara and Fischetti [9] showed that the separation problem for any point $y \in \mathbb{Q}^n$ and the class of $\{0, \frac{1}{2}\}$-Gomory–Chvátal cuts is solvable in time polynomially bounded in the input size of A, b, and y, assuming that A has, at most, two odd coefficients in each row. For 0/1 polytopes P_I this remains true for a relaxation $\{x \in \mathbb{R}^n : A'x \leqslant b'\}$ of $\{x \in \mathbb{R}^n : Ax \leqslant b\}$, where $A'x \leqslant b'$ is obtained from $Ax \leqslant b$ by adding systematically lower bound constraints $x_u \geqslant 0$ and upper bound constraints $x_u \leqslant 1$ such that A' has, at most, two odd coefficients in each row. More precisely, we may replace each inequality $\sum_u a_{iu}x_u \leqslant b_i$ with more than three odd coefficients by

$$a_{iv}x_v + a_{iw}x_w + \sum_{u:a_{iu}\text{ even}} a_{iu}x_u + \sum_{u\in L_i}(a_{iu}-1)x_u + \sum_{u\in U_i}(a_{iu}+1)x_u \leqslant b_i + |U_i|$$

for all elements v, w with odd coefficients and for all (including trivial) partitions (L_i, U_i) of $\{u \in \{1, 2, \ldots, n\}\setminus\{v, w\} : a_{iu} \text{ odd}\}$ for $i = 1, 2, \ldots, m$. Although this leads in general to an exponential number of rows, the separation problem associated with the $\{0, \frac{1}{2}\}$-Gomory–Chvátal cuts of this relaxation can still be solved in polynomial time; see [9]. Observe that a weak generalized $(k, 2)$-cycle inequality can be derived as a $\{0, \frac{1}{2}\}$-Gomory–Chvátal cut of such a relaxation when $|H_i| = 2$ for all edges H_i of the supporting cycle $(C, \{H_i : i = 1, 2, \ldots, k\})$. (Indeed, we do not need the upper bound constraints here.)

THEOREM 6.1. *There exists a polynomial time algorithm that, for any extended hypergraph $(N, \mathcal{H}, \text{tr})$ and for any point $y \in \mathbb{Q}^N$, either asserts that y satisfies all weak generalized $(k, 2)$-cycle inequalities supported by cycles $(C, \{H_i : i = 1, 2, \ldots, k\})$ such that $|H_i| = 2$, $i = 1, 2, \ldots, k$, or finds an inequality violated by y from a class of valid inequalities for $P_{\text{TP}}(N, \mathcal{H}, \text{tr})$ that contains all weak generalized $(k, 2)$-cycle inequalities supported by cycles $(C, \{H_i : i = 1, 2, \ldots, k\})$ such that $|H_i| = 2$, $i = 1, 2, \ldots, k$.*

Notice that this captures, in particular, all transitive packing problems in graphs. It covers, for instance, the 2-chorded odd cycle inequalities and the odd wheel inequalities for the clique partitioning and the partition polytope. The separation problem for the former class has previously been solved in [9, 33], the latter one in [17].

For the odd partition inequalities, we make use of both lower and upper bound constraints. Let us assume that $H_1, H_2, \ldots, H_k$ is the underlying collection of edges and that $d(u)$ odd implies that either $m(u) = 1$ and $n(u) = 0$ or $m(u) = 0$ and $n(u) = 1$ for all nodes $u \in N$. For a given partition $(W_1^{\text{odd}}, W_2^{\text{odd}})$ of W^{odd} the corresponding odd partition inequality can be obtained as a $\{0, \frac{1}{2}\}$-Gomory–Chvátal

cut from the relaxed system

$$\sum_{u\in H_i} x_u - \sum_{u\in \mathrm{tr}(H_i)} x_u + \sum_{u\in (H_i\cup \mathrm{tr}(H_i))\cap W_1^{\mathrm{odd}}} x_u - \sum_{u\in (H_i\cup \mathrm{tr}(H_i))\cap W_2^{\mathrm{odd}}} x_u$$
$$\leqslant |H_i| + |(H_i \cup \mathrm{tr}(H_i)) \cap W_1^{\mathrm{odd}}| - 1$$

for $i = 1, 2, \ldots, k$. For fixed κ, we denote by $\mathcal{C}_\kappa$ the class of odd partition inequalities such that $|H_i| \leqslant \kappa$, such that $d(u)$ odd implies that either $m(u) = 1$ and $n(u) = 0$ or $m(u) = 0$ and $n(u) = 1$ for all nodes $u \in N$, and such that $|(H_i \cup \mathrm{tr}(H_i)) \setminus W^{\mathrm{odd}}| \leqslant 2$ for $i = 1, 2, \ldots, k$. The next observation follows again from Caprara and Fischetti's result.

THEOREM 6.2. *There exists a polynomial time algorithm that, for any extended hypergraph $(N, \mathcal{H}, \mathrm{tr})$, for any fixed constant κ, and for any point $y \in \mathbb{Q}^N$, either asserts that y satisfies all odd partition inequalities in $\mathcal{C}_\kappa$ or finds an inequality violated by y from a class of valid inequalities for $P_{\mathrm{TP}}(N, \mathcal{H}, \mathrm{tr})$ that contains the class $\mathcal{C}_\kappa$ of certain odd partition inequalities.*

7. Special polytopes. In this section, we discuss two more polytopes that arise from the transitive packing polytope by special choices of hypergraphs and transitive elements. The detailed discussion of a third one, the interval order polytope, which inspired the introduction and the study of the transitive packing polytope, is the subject of another paper; see [49, Chapter 5]. The insights obtained for the acyclic subdigraph polytope as well as for the clique partitioning and the partition polytope have been stated during the treatment above. We will not repeat them here. We also do not review special independence system polytopes since this model has been known for years. Instead we concentrate on two recently introduced polytopes that deal with transitive elements.

7.1. The transitive acyclic subdigraph polytope. An instance of the *transitive acyclic subdigraph problem* (or *poset problem*) consists of a directed graph $D = (V, A)$ and a weight function $c : A \to \mathbb{Q}$. The goal is to determine a set of arcs $B \subseteq A$ such that the digraph (V, B) is acyclic and transitively closed, i.e., such that it represents a partially ordered set and such that $c(B)$ is as large as possible. The *transitive acyclic subdigraph polytope* (or *partial order polytope*) of D is the convex hull of 0/1 incidence vectors of all transitive and acyclic arc sets of D. Equivalently, it is the integer hull of the polytope defined by

$$x_{uv} \geqslant 0 \quad \text{for all arcs } (u, v) \in A, \tag{7.1}$$
$$x_{uv} \leqslant 1 \quad \text{for all arcs } (u, v) \in A, \tag{7.2}$$
$$x_{uv} + x_{vu} \leqslant 1 \quad \text{for all pairs } (u, v), (v, u) \in A, \tag{7.3}$$
$$x_{uv} + x_{vw} \leqslant 1 \quad \text{for all } (u, v), (v, w) \in A \text{ such that } (u, w) \notin A, \tag{7.4}$$
$$x_{uv} + x_{vw} - x_{uw} \leqslant 1 \quad \text{for } (u, v), (v, w), (u, w) \in A. \tag{7.5}$$

The transitive acyclic subdigraph polytope was introduced by Müller [33]. It arises as a transitive packing polytope of an extended graph $(N, \mathcal{H}, \mathrm{tr})$ defined as follows: the arc set A of the digraph D forms the node set N, and two nodes $(u_1, v_1), (u_2, v_2) \in A$ are said to be *adjacent* if $v_1 = u_2$ or $u_1 = v_2$ (or both). The transitive element that we associate with a pair of adjacent arcs $(u, v), (v, w) \in A$ is the arc (u, w), if it exists.

It has already been shown in [33] that the transitive acyclic subdigraph polytope is full dimensional, that the nonnegativity constraints (7.1) are facet defining, and

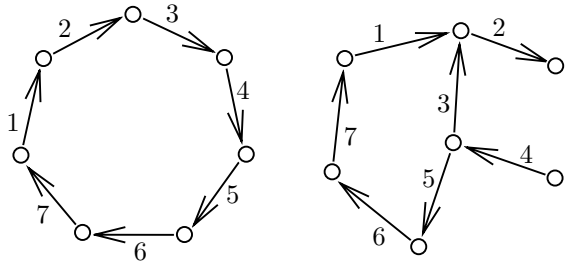


FIG. 7.1. *Two digraphs which are generalized* $(7,2)$*-cycles in the extended graphs corresponding to the transitive acyclic subdigraph problem. The numbers indicate the chosen sequence, respectively.*

that an upper bound constraint $x_{uv} \leqslant 1$ defines a facet if and only if for all $w \in V$ with $(w,u) \in A$ (or $(v,w) \in A$) also $(w,v) \in A$ (respectively, $(u,w) \in A$). The latter condition is precisely the translation of the assumption made in Lemma 2.4(ii). The only known nontrivial class of facet defining inequalities is associated with odd dicycles in D [33]. If $(u_1,u_2),(u_2,u_3),\ldots,(u_{k-1},u_k),(u_k,u_1)$ forms an odd dicycle in D, its *cycle inequality* is

$$\sum_{i=1}^{k} x_{u_i u_{i+1}} - \sum_{\substack{i=1 \\ (u_i,\, u_{i+2}) \in A}}^{k} x_{u_i u_{i+2}} \leqslant \frac{k-1}{2}.$$

These cycle inequalities obviously belong to the class of generalized $(k,2)$-cycle inequalities. However, there is no reason to restrict ourselves to cycles in the digraph D. Figure 7.1 shows an arc configuration that defines a generalized cycle in the extended graph defined above but is no dicycle in D. Hence, we can present a much larger class of valid inequalities for the transitive acyclic subdigraph polytope.

LEMMA 7.1. *Let* $D=(V,A)$ *be a digraph. For* $k \geqslant 3$ *odd, let* $a_1,a_2,\ldots,a_k$ *be a sequence of arcs in* A *such that* a_i,a_{i+1} *are adjacent,* $i=1,2,\ldots,k$*. The inequalities*

$$\sum_{i=1}^{k} x_{a_i} - \sum_{\substack{a \in \mathrm{tr}(\{a_i,\, a_{i+1}\}) \\ \text{for some } i}} \frac{\lceil n(a) \rceil_2}{2} x_a \leqslant \frac{k-1}{2} \quad \text{and} \quad \sum_{i=1}^{k} x_{a_i} - \sum_{i=1}^{k} x_{\mathrm{tr}(\{a_i,a_{i+1}\})} \leqslant \frac{k-1}{2}$$

are valid for the transitive acyclic subdigraph polytope of D*. Here,* $n(a) = |\{i \in \{1,2,\ldots,k\} : a \in \mathrm{tr}(\{a_i,a_{i+1}\})|$*. The latter class of inequalities is contained in a class of valid inequalities for the transitive acyclic subdigraph polytope of* D *for which the corresponding separation problem is solvable in polynomial time.*

We note that there do not exist generalized $(k,2)$-cliques in the case of the transitive acyclic subdigraph polytope for $k \geqslant 4$. We close this section on the transitive acyclic subdigraph polytope with the observation that the transitive acyclic subdigraph polytope of a digraph D whose underlying graph is bipartite is completely described by (7.1)–(7.4). We may argue as follows. First observe that there do not exist transitive arcs. Let black and white be the two color classes of the underlying bipartite graph. The extended graph induced by D is also bipartite. Its color classes are the arcs directed from black to white and the arcs from white to black, respectively. Since it is known that the node packing polytope of a bipartite graph is completely described by the nonnegativity, the upper bound, and the edge constraints, our claim follows.

7.2. The relatively transitive subdigraph polytope. A digraph $D = (V, A)$ is said to be transitively closed, or just transitive, whenever the presence of two arcs (u, v), $(v, w) \in A$ implies the presence of the arc (u, w) in A. A subdigraph (V, B) of a digraph $D = (V, A)$ is called *relatively transitive* if for every dipath from u to v in (V, B) either $(u, v) \in B$ or (u, v) is not in A. We define the *relatively transitive subdigraph polytope* of D as the convex hull of the incidence vectors of all relatively transitive subdigraphs of D or, equivalently, as the integer hull of the polytope defined by

$$x_{uv} \geqslant 0 \qquad \text{for all arcs } (u, v) \in A, \tag{7.6}$$

$$x_{uv} \leqslant 1 \qquad \text{for all arcs } (u, v) \in A, \tag{7.7}$$

$$\sum_{a \in p} x_a - x_{uv} \leqslant |p| - 1 \qquad \text{for all } (u, v) \in A \text{ and for all dipaths } p \in \mathcal{P}^D_{uv}, \tag{7.8}$$

where $\mathcal{P}^D_{uv}$ is the set of dipaths from u to v in D. The size $|p|$ of such a dipath p is the number of its arcs. Shallcross and Bland [51] (see also [50]) studied the convex hull of 0/1 points x whose complements $\overline{x} = \mathbb{1} - x$ satisfy (7.6)–(7.8). If D is transitively closed, these points represent the independent sets of the transitivity antimatroid of D. Shallcross and Bland were motivated by a question raised by Korte and Lovász [31] of whether the convex hull of these incidence vectors has a (computationally) nice description. Shallcross and Bland present some conditions on D such that their polytope, and therefore the relatively transitive subdigraph polytope, is completely described by (7.6)–(7.8). They also point out that maximizing a linear function over the relatively transitive subdigraph polytope is NP-hard in general, thereby answering Korte and Lovász's question to the negative.

The way we introduced the relatively transitive subdigraph polytope makes it likely to be a certain transitive packing polytope. To be precise, let the arc set A of the given digraph $D = (V, A)$ be the node set N of the extended hypergraph to be defined. The hyperedges are formed by the arcs of dipaths from node u to node v for all $u, v \in V$ such that $(u, v) \in A$. Finally, the transitive element associated with such a hyperedge is clearly the arc (u, v). Now, we may translate all the inequalities presented for the transitive packing polytope into this context, thus answering a question of Shallcross and Bland for other valid inequalities for the (complement of the) relatively transitive subdigraph polytope.

8. Concluding remarks. Notice that the inequalities presented above remain valid when we allow for hypergraphs with loops. Then, we cover, for instance, the *cut polytope* (see, e.g., [5, 18]) and the Boolean quadric polytope (e.g., [44]) as well.

It is well known (see [19]) that every *set packing problem*

$$\begin{aligned} &\text{maximize} && cx \\ &\text{subject to} && Ax \leqslant \mathbb{1}, \\ & && x_u \in \{0, 1\}, \end{aligned} \tag{8.1}$$

where A is a matrix of zeros and ones, can be transformed into an equivalent node packing problem on the *intersection graph* of A. Every column becomes a node, and two nodes u and v are joined by an edge if and only if the matrix A contains a row with entry 1 in columns u and v. In other words, the convex hull of feasible solutions to (8.1) (the *set packing polytope* of A) is identical to the node packing polytope of the intersection graph of A. Hence transitive packing covers set packing as well since

it subsumes node packing. However, generalized set packing polytopes [13] do not immediately occur as special instances of transitive packing polytopes. In fact, given a $0/\pm 1$ matrix A and the vector n_A whose components count the number of negative entries in the corresponding rows of A, Conforti and Cornuéjols defined (the integer hull of) $\{x : Ax \leqslant \mathbb{1} - n_A, 0 \leqslant x \leqslant \mathbb{1}\}$ as a *generalized set packing polytope*.

On the other hand, as already pointed out, the transitive packing polytope of an extended hypergraph with no transitive elements reduces to an independence system polytope. There is a close relation between independence system polytopes and set covering polytopes (see, e.g., [32, 39]). A *set covering polytope* is of the form $\operatorname{conv}\{y \in \{0,1\}^n : Ay \geqslant \mathbb{1}\}$, where A is a $0/1$ matrix. The points y in the set covering polytope and the points x in the independence system polytope of the circuit system defined by the undominated rows of A are related by the affine transformation $x = \mathbb{1} - y$. Explicitly, $x \in \operatorname{conv}\{x \in \{0,1\}^n : Ax \leqslant p_A - \mathbb{1}\}$ if and only if $\mathbb{1} - x \in \operatorname{conv}\{y \in \{0,1\}^n : Ay \geqslant \mathbb{1}\}$. Consequently, set covering polytopes and independence system polytopes are equivalent, modulo the above transformation. An implication of this is that any result stated for the independence system polytope can be translated to the set covering polytope and vice versa. Thus the work of Balas and Ng [1, 2], Cornuéjols and Sassano [16], Euler and Mahjoub [21], Nobili and Sassano [39], and Sassano [45] as well as others on the set covering polytope can be seen as contributions to the knowledge concerning the independence system polytope. For instance, the inequalities for the set covering polytope associated with *complete* (q,s)*-roses of order* k [45] turn out to be equivalent to the generalized (k,s,q)-antiweb inequalities of Laurent [32]. This implies especially that our extension of the class of antiweb inequalities for the independence system polytope extends the known rose inequalities for the set covering polytope, too.

If we apply the complementing of variables to the transitive packing polytope $P_{\mathrm{TP}}(N,\mathcal{H},\mathrm{tr}) = \operatorname{conv}\{x \in \{0,1\}^N : Ax \leqslant p_A - \mathbb{1}\}$, where the $0/\pm 1$ matrix A is the extended edge-node incidence matrix of the extended hypergraph $(N,\mathcal{H},\mathrm{tr})$, it turns out to be equivalent (modulo this affine transformation) to the polytope $Q(A) := \operatorname{conv}\{x \in \{0,1\}^N : Ax \geqslant \mathbb{1} - n_A\}$. The natural linear relaxation of the polytope $Q(A)$ has been introduced by Conforti and Cornuéjols [13] in the context of balanced $0/\pm 1$ matrices as the (*fractional*) *generalized set covering polytope*. Conforti and Cornuéjols [13] as well as Nobili and Sassano [40] characterize when the fractional generalized set covering polytope is integral, i.e., when it coincides with the generalized set covering polytope. Our work can be seen as a contribution to the study of the generalized set covering polytope when it is properly contained in the corresponding fractional one. Recall that a $0/\pm 1$ matrix is *balanced* if, in every submatrix with exactly two nonzero entries per row and per column, the sum of the entries is a multiple of four [53]. We refer to Conforti, Cornuéjols, Kapoor, Vuskŏvić, and Rao [14] for a survey of balanced matrices and related concepts. Conforti and Cornuéjols [13] showed that a $0/\pm 1$ matrix A is balanced if and only if the fractional generalized set covering (or packing) polytope is integral for each submatrix of A. An extension of the concept of balanced $0/\pm 1$ matrices is ideal matrices. A $0/\pm 1$ matrix A is *ideal* if its fractional generalized set covering polytope is integral or, equivalently, if its fractional transitive packing polytope is integral. It would be very interesting, for problems that can be interpreted as transitive packing problems, to characterize when the extended edge-node incidence matrices of their associated extended hypergraphs are ideal. Little is known so far about ideal $0/\pm 1$ matrices; see [14, 40].

The way we introduced the transitive packing model and the name we gave to it reflect how we discovered it [49, Chapter 4] but may hide its full generality. To

highlight and to slightly extend the generality of our model, we finally provide another presentation. A *directed hypergraph* is a pair $(N, \mathcal{H})$ consisting of a finite set N of nodes and of a set of directed hyperedges (hyperarcs). A *hyperarc* $(H^+, H^-) \in \mathcal{H}$ consists of two (possibly empty) disjoint subsets of N. For a survey of directed hypergraphs the reader is referred to [22]. Now, consider for $x \in \{0,1\}^N$ the following "directed hypergraph covering" constraints:

$$\overline{x}(H^+) + x(H^-) \geqslant 1 \quad \text{for all hyperarcs } (H^+, H^-) \in \mathcal{H},$$

where $\overline{x} = \mathbb{1} - x$ is the complement of the 0/1 vector x. Observe that this is equivalent to the transitivity constraints (2.1), with $H^+ = H$ and $H^- = \mathrm{tr}(H)$. In particular, this form emphasizes the symmetry of the role of hyperedges and their associated transitive sets. For example, reversing the direction of the hyperarcs simply amounts to exchanging x and $\overline{x}$.

Acknowledgment. The authors are grateful to Maurice Queyranne for suggesting the interpretation of transitive packing in terms of directed hypergraphs.

REFERENCES

[1] E. Balas and S. M. Ng, *On the set covering polytope* I: *All the facets with coefficients in* $\{0,1,2\}$, Math. Programming, 43 (1989), pp. 57–69.
[2] E. Balas and S. M. Ng, *On the set covering polytope* II: *Lifting the facets with coefficients in* $\{0,1,2\}$, Math. Programming, 45 (1989), pp. 1–20.
[3] E. Balas and M. W. Padberg, *Set partitioning: A survey*, SIAM Rev., 18 (1976), pp. 710–760.
[4] F. Barahona, M. Grötschel, and A. R. Mahjoub, *Facets of the bipartite subgraph polytope*, Math. Oper. Res., 10 (1985), pp. 340–358.
[5] F. Barahona and A. R. Mahjoub, *On the cut polytope*, Math. Programming, 36 (1986), pp. 157–173.
[6] C. Berge, *Graphs and Hypergraphs*, North–Holland, Amsterdam, 1973.
[7] R. Borndörfer, *Aspects of Set Packing, Partitioning, and Covering*, Ph.D. thesis, Technische Universität Berlin, Berlin, Germany, 1998.
[8] R. Borndörfer and R. Weismantel, *Set packing relaxations of some integer programs*, Math. Programming, 88 (2000), pp. 425–450.
[9] A. Caprara and M. Fischetti, $\{0, \frac{1}{2}\}$*-Chvátal–Gomory cuts*, Math. Programming, 74 (1996), pp. 221–235.
[10] S. Chopra and M. R. Rao, *The partition problem*, Math. Programming, 59 (1993), pp. 87–115.
[11] V. Chvátal, *Edmonds polytopes and a hierarchy of combinatorial problems*, Discrete Math., 4 (1973), pp. 305–337.
[12] V. Chvátal, W. Cook, and M. Hartmann, *On cutting-plane proofs in combinatorial optimization*, Linear Algebra Appl., 114/115 (1989), pp. 455–499.
[13] M. Conforti and G. Cornuéjols, *Balanced* $0, \pm 1$*-matrices, bicoloring and total dual integrality*, Math. Programming, 71 (1995), pp. 249–258.
[14] M. Conforti, G. Cornuéjols, A. Kapoor, K. Vuskŏvić, and M. R. Rao, *Balanced matrices*, in Mathematical Programming: State of the Art 1994, J. R. Birge and K. G. Murty, eds., University of Michigan, Ann Arbor, MI, 1994, pp. 1–33.
[15] W. J. Cook, W. H. Cunningham, W. R. Pulleyblank, and A. Schrijver, *Combinatorial Optimization*, John Wiley, New York, 1998.
[16] G. Cornuéjols and A. Sassano, *On the* $0,1$ *facets of the set covering polytope*, Math. Programming, 43 (1989), pp. 45–55.
[17] M. Deza, M. Grötschel, and M. Laurent, *Clique-web facets for multicut polytopes*, Math. Oper. Res., 17 (1992), pp. 981–1000.
[18] M. M. Deza and M. Laurent, *Geometry of Cuts and Metrics*, Algorithms Combin. 15, Springer, Berlin, 1997.
[19] J. Edmonds, *Covers and packings in a family of sets*, Bull. Amer. Math. Soc. (N.S.), 68 (1962), pp. 494–499.
[20] R. Euler, M. Jünger, and G. Reinelt, *Generalizations of cliques, odd cycles and anticycles and their relation to independence system polyhedra*, Math. Oper. Res., 12 (1987), pp. 451–462.

[21] R. Euler and A. R. Mahjoub, *Balanced matrices and the set covering polytope*, Arab. J. Sci. Eng. Sect. B Eng., 16 (1991), pp. 269–282.
[22] G. Gallo, G. Longo, S. Pallottino, and S. Nguyen, *Directed hypergraphs and applications*, Discrete Appl. Math., 42 (1993), pp. 177–201.
[23] M. X. Goemans and L. A. Hall, *The strongest facets of the acyclic subgraph polytope are unknown*, in Integer Programming and Combinatorial Optimization, Proceedings of the 5th International IPCO Conference, Vancouver, BC, Canada, W. H. Cunningham, S. T. McCormick, and M. Queyranne, eds., Lecture Notes in Comput. Sci. 1084, Springer, Berlin, 1996, pp. 415–429.
[24] M. Grötschel, M. Jünger, and G. Reinelt, *Acyclic subdigraphs and linear orderings: Polytopes, facets, and cutting plane algorithms*, in Graphs and Order, I. Rival, ed., D. Reidel, Dordrecht, The Netherlands, 1985, pp. 217–266.
[25] M. Grötschel, M. Jünger, and G. Reinelt, *On the acyclic subgraph polytope*, Math. Programming, 33 (1985), pp. 28–42.
[26] M. Grötschel, L. Lovász, and A. Schrijver, *Geometric Algorithms and Combinatorial Optimization*, Algorithms Combin. 2, Springer, Berlin, 1988.
[27] M. Grötschel and Y. Wakabayashi, *A cutting plane algorithm for a clustering problem*, Math. Programming, 45 (1989), pp. 59–96.
[28] M. Grötschel and Y. Wakabayashi, *Facets of the clique partitioning polytope*, Math. Programming, 47 (1990), pp. 367–388.
[29] M. Jünger, *Polyhedral Combinatorics and the Acyclic Subdigraph Problem*, Res. Exp. Math. 7, Heldermann-Verlag, Berlin, 1985.
[30] M. Jünger and P. Mutzel, *Solving the maximum weight planar subgraph problem by branch and cut*, in Integer Programming and Combinatorial Optimization, Proceedings of the 3rd International IPCO Conference, Erice, Italy, G. Rinaldi and L. A. Wolsey, eds., 1993, pp. 479–492.
[31] B. Korte and L. Lovász, *Polyhedral results for antimatroids*, in Combinatorial Mathematics, Proceedings of the 3rd International Conference, G. S. Bloom, R. L. Graham, and J. Malkevitch, eds., Ann. New York Acad. Sci. 555, 1989, pp. 283–295.
[32] M. Laurent, *A generalization of antiwebs to independence systems and their canonical facets*, Math. Programming, 45 (1989), pp. 97–108.
[33] R. Müller, *On the transitive acyclic subdigraph polytope*, in Integer Programming and Combinatorial Optimization, Proceedings of the 3rd International IPCO Conference, Erice, Italy, G. Rinaldi and L. A. Wolsey, eds., 1993, pp. 463–477.
[34] R. Müller, *On the partial order polytope of a digraph*, Math. Programming, 73 (1996), pp. 31–49.
[35] R. Müller and A. S. Schulz, *The interval order polytope of a digraph*, in Integer Programming and Combinatorial Optimization, Proceedings of the 4th International IPCO Conference, Copenhagen, Denmark, E. Balas and J. Clausen, eds., Lecture Notes in Comput. Sci. 920, Springer, Berlin, 1995, pp. 50–64.
[36] R. Müller and A. S. Schulz, *Transitive packing*, in Integer Programming and Combinatorial Optimization, Proceedings of the 5th International IPCO Conference, W. H. Cunningham, S. T. McCormick, and M. Queyranne, eds., Lecture Notes in Comput. Sci. 1084, Springer, Berlin, 1996, pp. 430–444.
[37] G. L. Nemhauser and L. E. Trotter, Jr., *Properties of vertex packing and independence system polyhedra*, Math. Programming, 6 (1974), pp. 48–61.
[38] G. L. Nemhauser and L. A. Wolsey, *Integer and Combinatorial Optimization*, John Wiley, New York, 1988.
[39] P. Nobili and A. Sassano, *Facets and lifting procedures for the set covering polytope*, Math. Programming, 45 (1989), pp. 111–137.
[40] P. Nobili and A. Sassano, $(0, \pm 1)$ *ideal matrices*, in Integer Programming and Combinatorial Optimization, Proceedings of the 4th International IPCO Conference, E. Balas and J. Clausen, eds., Lecture Notes in Comput. Sci. 920, Springer, Berlin, 1995, pp. 344–359.
[41] M. Oosten, J. H. G. C. Rutten, and F. C. R. Spieksma, *The Clique Partitioning Polytope: Facets*, Tech. report, Department of Mathematics, University of Limburg, Maastricht, The Netherlands, 1995.
[42] M. W. Padberg, *On the facial structure of set packing polyhedra*, Math. Programming, 5 (1973), pp. 199–215.
[43] M. W. Padberg, *Covering, packing and knapsack problems*, Ann. Discrete Math., 4 (1979), pp. 265–287.
[44] M. W. Padberg, *The Boolean quadric polytope: Some characteristics, facets and relatives*, Math. Programming, 45 (1989), pp. 139–172.

[45] A. Sassano, *On the facial structure of the set covering polytope*, Math. Programming, 44 (1989), pp. 181–202.
[46] A. Schrijver, *On cutting planes*, in Combinatorics '79, Part II, M. Deza and I. G. Rosenberg, eds., Ann. Discrete Math. 9, North–Holland, Amsterdam, 1980, pp. 291–296.
[47] A. Schrijver, *Theory of Linear and Integer Programming*, John Wiley, Chichester, U.K., 1986.
[48] A. S. Schulz, *Transitive packing vs. set packing relaxations: An explanation*, manuscript.
[49] A. S. Schulz, *Polytopes and Scheduling*, Ph.D. thesis, Technische Universität Berlin, Berlin, Germany, 1996.
[50] D. Shallcross, *Two Investigations in Combinatorial Optimization: Approximate Solution of Large Traveling Salesman Problems Arising from X-Ray Diffraction Experiments; Relatively Transitive Subgraphs*, Ph.D. thesis, Cornell University, Ithaca, NY, 1989.
[51] D. F. Shallcross and R. G. Bland, *On the Polyhedral Structure of Relatively Transitive Subgraphs*, Technical Report 1049, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY, 1993.
[52] L. E. Trotter, Jr., *A class of facet producing graphs for vertex packing polyhedra*, Discrete Math., 12 (1975), pp. 373–388.
[53] K. Truemper, *Alpha-balanced graphs and matrices and GF(3)-representability of matroids*, J. Combin. Theory Ser. B, 32 (1982), pp. 112–139.

SIAM J. OPTIM.
Vol. 13, No. 2, pp. 368–385

COMPLEMENTARITY CONSTRAINT QUALIFICATION VIA THE THEORY OF 2-REGULARITY*

A. F. IZMAILOV† AND M. V. SOLODOV‡

Abstract. We exhibit certain second-order regularity properties of parametric complementarity constraints, which are notorious for being irregular in the classical sense. Our approach leads to a constraint qualification in terms of 2-regularity of the mapping corresponding to the subset of constraints which must be satisfied as equalities around the given feasible point, while no qualification is required for the rest of the constraints. Under this 2-regularity assumption, we derive constructive sufficient conditions for tangent directions to feasible sets defined by complementarity constraints. A special form of primal-dual optimality conditions is also obtained. We further show that our 2-regularity condition always holds under the piecewise Mangasarian–Fromovitz constraint qualification, but not vice versa. Relations with other constraint qualifications and optimality conditions are also discussed. It is shown that our approach can be useful when alternative ones are not applicable.

Key words. equilibrium constraints, tangent cone, constraint qualification, 2-regularity, optimality conditions

AMS subject classifications. 90C30, 90C33

PII. S1052623499365292

1. Introduction.

This paper is devoted to the analysis of local structure of a set defined by parametric complementarity constraints, such as

$$D := \{(x, y) \in \Re^n \times \Re^m \mid g(x, y) \geq 0, y \geq 0, \langle g(x, y), y\rangle = 0\}, \tag{1.1}$$

where $g : \Re^n \times \Re^m \to \Re^m$, and $\langle \cdot, \cdot \rangle$ denotes the usual inner product in an appropriate space. The geometry of sets having this (or similar) structure is important in connection with mathematical programs with complementarity constraints, where D appears as the feasible region:

$$\begin{array}{ll} \text{minimize} & f(x, y) \\ \text{subject to} & (x, y) \in D. \end{array} \tag{1.2}$$

Mathematical programs with equilibrium constraints (MPEC), of which (1.2) is a special case, are a relatively new active field of research. We refer the reader to the monograph [10] for further references. It is known [4], and not difficult to see, that for D given by (1.1) the classical Mangasarian–Fromovitz constraint qualification (MFCQ) [13] does not hold at any feasible point $(x, y) \in D$. Arguably, this is the single most important reason that makes treatment of this problem considerably more difficult when compared to regular (i.e., satisfying some standard constraint qualifications) nonlinear programs. Specifically, the nonsatisfaction of classical constraint qualifications makes a simple constructive description of the tangent cone to the set

*Received by the editors December 13, 1999; accepted for publication (in revised form) February 11, 2002; published electronically September 24, 2002.
http://www.siam.org/journals/siopt/13-2/36529.html

†Computing Center of the Russian Academy of Sciences, Vavilova Street 40, Moscow, GSP-1, Russia (izmaf@ccas.ru). This author's research was supported by the Russian Foundation for Basic Research grants 99-01-00472 and 01-01-00810. The author also thanks IMPA, where he was a visiting professor during the completion of this work.

‡Instituto de Matemática Pura e Aplicada, Estrada Dona Castorina 110, Jardim Botânico, Rio de Janeiro, RJ 22460-320, Brazil (solodov@impa.br). This author's research was supported in part by CNPq grant 300734/95-6, by PRONEX–Optimization, and by FAPERJ.

D (and its dual) difficult. This rules out standard stationarity/optimality conditions. The latter, in turn, makes it problematic to come up with reliable computational algorithms for solving MPEC.

To deal with these issues, a number of approaches have been proposed, among which are [11, 10, 18, 15, 14, 20]. In this paper, we study complementarity constraints from the point of view of the theory of 2-regularity of twice differentiable mappings, which has not been used in this context previously. Our analysis clarifies the role that 2-regularity can play for MPEC, as well as provides an alternative view of some issues related to MPEC constraint qualifications and optimality conditions. We also exhibit some situations when other approaches do not apply but our results appear useful to characterize optimality.

Given any $(x^*, y^*) \in D$, define the three index sets

$$\begin{aligned} I_0 &:= \{i \mid y_i^* = 0 = g_i(x^*, y^*)\}, \\ I_g &:= \{i \mid y_i^* > 0 = g_i(x^*, y^*)\}, \\ I_y &:= \{i \mid y_i^* = 0 < g_i(x^*, y^*)\}, \end{aligned}$$

with I_0 being the *degenerate* set. Note that these index sets depend on the point under consideration. Since this point will be fixed throughout our analysis, we shall omit this dependency in our notation. Note also that locally, the constraints $g_i(x, y) \geq 0, i \in I_y$, are never active, and we can further deal with constraints $y_i = 0, i \in I_y$, explicitly by eliminating the variables. Thus, to simplify the notation, we can assume that

$$I_y = \emptyset.$$

It is further easy to see that there exists a neighborhood $\mathcal{V}$ of $(x^*, y^*) \in D$ such that $D \cap \mathcal{V} = D^* \cap \mathcal{V}$, where

$$D^* := \left\{ (x, y) \;\middle|\; \begin{array}{l} y_i \geq 0, g_i(x, y) \geq 0, y_i g_i(x, y) = 0, \; i \in I_0 \\ g_i(x, y) = 0, \; i \in I_g \end{array} \right\}. \tag{1.3}$$

The sets D and D^* are therefore locally equivalent, and we can work with the representation given by (1.3). Let us define the mapping $F : \Re^n \times \Re^m \to \Re^{|I_0|+|I_g|}$ associated with equality constraints in D^*:

$$F_i(x, y) := \begin{cases} y_i g_i(x, y), & i \in I_0, \\ g_i(x, y), & i \in I_g. \end{cases}$$

In what follows, we show that certain (second-order) regularity properties of this mapping F are sufficient to completely characterize tangent directions to the set D and to derive corresponding optimality conditions for MPEC.

The rest of the paper is organized as follows. In section 2 we review the theory of 2-regularity (for smooth equality constraints) and show, by means of a simple example, that it is relevant in the context of MPEC. Section 3 contains our necessary and sufficient conditions for tangency and a demonstration that this description is always valid under the piecewise MFCQ and can also be useful when the latter condition does not hold. A special form of primal-dual necessary optimality conditions is derived in section 4. This section also contains a comparison with some alternative approaches to MPEC constraint qualifications and optimality conditions.

We next briefly describe our notation. All vectors will be column-vectors. When writing a pair (x, y), we mean a column-vector composed of two vectors x and y.

For a vector v of arbitrary (finite) dimension, v_i will denote its ith component, and for a matrix M, M_i will denote its ith row. By M^T we shall denote the transposed matrix of M. For a differentiable scalar function $\varphi : \Re^n \times \Re^m \to \Re$, $\nabla\varphi(z)$ stands for the row-vector of its partial derivatives at $z = (x, y) \in \Re^n \times \Re^m$ with respect to all the variables, while $\nabla_x\varphi(z)$ will denote the vector of partial derivatives with respect to $x \in \Re^n$ (similarly for y). If φ is twice differentiable, $\nabla^2\varphi(z)$ denotes its Hessian matrix, and $\nabla^2_x\varphi(z)$ its Hessian matrix with respect to x. For a differentiable vector-function $F : \Re^l \to \Re^k$, $F'(z)$ will denote its Jacobian, i.e., the $k \times l$ matrix whose ith row is $\nabla F_i(z)$. If F is twice differentiable, then $F''(z)[h]$ denotes the $k \times l$ matrix whose ith row is $\nabla^2 F_i(z)h, i = 1, \ldots, k$. Furthermore, $F''(z)[h, p] := (F''(z)[h])p$, and $F''(z)[h]^2 := F''(z)[h, h]$. For a linear operator $A : S_1 \to S_2$, $\operatorname{Ker} A = \{w \in S_1 \mid Aw = 0\}$ is its null space, and $\operatorname{Im} A = \{v \in S_2 \mid v = Aw \text{ for some } w \in S_1\}$ is its image space. Given a subspace S of an arbitrary space, we shall denote its orthogonal complement in this space by $S^\perp$. If K is a cone, then $K^* = \{p \mid \langle p, h\rangle \geq 0 \ \forall h \in K\}$ is the (positive) dual cone of K. For a set K, $\operatorname{cl} K$ will stand for its closure. Finally, for a (finite) index set I its cardinality will be denoted by $|I|$.

2. Elements of the theory of 2-regularity. In this section, we describe some results from the theory of 2-regularity for smooth nonlinear mappings [12, 19, 1, 9, 5, 6, 7]. Our emphasis will be on only those constructions that will be used later in the paper. We therefore shall limit our presentation to equality constraints only, to the finite-dimensional setting, and to the case of mappings differentiable at least twice. After introducing the necessary objects, we provide an example which illustrates certain 2-regularity properties of MPEC.

Let $F : \Re^l \to \Re^k$ be differentiable at a point $z^* \in C$, where

$$C := \{z \in \Re^l \mid F(z) = 0\}.$$

We are interested in the tangent cone to the set C at the point $z^* \in C$, denoted $T_C(z^*)$, which is the set of all vectors $h \in \Re^l$ for which there exists a mapping $r : \Re_+ \to \Re^l$ such that

$$z^* + th + r(t) \in C \quad \forall\, t \in \Re_+, \quad \|r(t)\| = o(t).$$

We note that in the context of constructive constraint qualifications there is no distinction between this definition of the tangent cone and the more general Bouligand tangent cone. In the following, we shall use the given definition, as it is more convenient for our development (and leads to the same results as when using the Bouligand cone anyway). It is well known that

$$T_C(z^*) \subset H_1 := \operatorname{Ker} F'(z^*),$$

which is the (first-order) necessary condition of tangency. A sufficient condition is given by the classical Lyusternik theorem: the equality

$$T_C(z^*) = H_1$$

holds if F is (first-order) *regular*, i.e.,

$$\operatorname{Im} F'(z^*) = \Re^k.$$

In the *irregular* case when $\operatorname{Im} F'(z^*) \neq \Re^k$, cone $T_C(z^*)$ can be smaller than H_1. In that case, a more accurate representation is needed.

Suppose now that F is twice differentiable at z^*. Then it can be verified that

$$T_C(z^*) \subset H_2 := H_1 \cap \{h \mid F''(z^*)[h]^2 \in \operatorname{Im} F'(z^*)\},$$

which is the second-order necessary condition of tangency. Let P be the orthogonal projector onto $(\operatorname{Im} F'(z^*))^\perp$ in $\Re^k$. With this notation, we can equivalently write

$$H_2 = \operatorname{Ker} F'(z^*) \cap \{h \mid PF''(z^*)[h]^2 = 0\}. \tag{2.1}$$

DEFINITION 2.1. *Under the assumptions above, the mapping F is called 2-regular at the point z^* with respect to an element $h \in \Re^l$ if*

$$\operatorname{Im}(F'(z^*) + PF''(z^*)[h]) = \Re^k.$$

Obviously, if F is (first-order) regular in the classical sense, i.e., $\operatorname{Im} F'(z^*) = \Re^k$, then it is 2-regular with respect to *every* h. The following generalization of the classical Lyusternik theorem can be found in [12, 19, 1, 9, 5], and more general results under weaker smoothness assumptions, with applications, in [6, 7].

THEOREM 2.2. *Let $F : \Re^l \to \Re^k$ be twice differentiable at a point $z^* \in \Re^l$ such that $F(z^*) = 0$. Assume further that F is 2-regular at z^* with respect to an element $h \in H_2$. Then there exists a mapping $r : \Re_+ \to \Re^l$ such that*

$$F(z^* + th + r(t)) = 0 \quad \forall\, t \in \Re_+, \quad \text{where } \|r(t)\| = o(t).$$

In other words, 2-regularity of F on some h satisfying the second-order necessary conditions of tangency ($h \in H_2$) guarantees that this h is indeed a tangent direction. In particular, if F is 2-regular with respect to every $h \in H_2 \setminus \{0\}$, then

$$T_C(z^*) = H_2.$$

In fact, due to the closedness of the tangent cone, for the above equality to hold it is enough to require that F be 2-regular with respect to every h in some dense subset of H_2 (i.e., $h \in K$ such that $\operatorname{cl} K = H_2$).

Note that this representation of the tangent cone subsumes the classical regular case. Indeed, in that case $\operatorname{Im} F'(z^*) = \Re^k$; hence $(\operatorname{Im} F'(z^*))^\perp = \{0\}$, $P = 0$, and thus $\{h \mid PF''(z^*)[h]^2 = 0\} = \Re^l$, and so we have that $T_C(z^*) = \operatorname{Ker} F'(z^*) = H_1 = H_2$. Of course, the description of tangent directions via constructions of 2-regularity is useful precisely in the absence of classical regularity. It is therefore natural to see what, if anything, the approach of 2-regularity has to offer in the context of MPEC. To quickly convince the reader that 2-regularity is relevant here, we next consider a simple one-dimensional problem and show that sufficient conditions for tangency can be interpreted as 2-regularity of a certain subset of constraints.

Example 2.1. Let $n = m = 1$ and consider the set D given by

$$y \geq 0, \quad g(x, y) \geq 0, \quad 0 = F(x, y) := yg(x, y). \tag{2.2}$$

Obviously, the cause of difficulties in MPEC is the degenerate set I_0, so the case of interest is when $I_0 \neq \emptyset$. Here, this means that $I_0 = \{1\}$ while the other index sets introduced in section 1 are empty. Therefore we shall consider a point $(x^*, y^*) \in \Re^2$ such that

$$y^* = 0, \quad g(x^*, y^*) = 0.$$

As is easy to see, in this case

$$F'(x^*, y^*) = (0, 0).$$

Hence, the last constraint in (2.2) is not regular (even by itself, separately from the other constraints defining D!). Note further that

$$F''(x^*, y^*) = \begin{pmatrix} 0 & \nabla_x g(x^*, y^*) \\ \nabla_x g(x^*, y^*) & 2\nabla_y g(x^*, y^*) \end{pmatrix}.$$

As is easy to see, the necessary conditions for some $(u, v) \in \Re \times \Re$ to be a tangent direction to D at (x^*, y^*) are the following:

$$\begin{gathered} v \geq 0, \quad \nabla_x g(x^*, y^*)u + \nabla_y g(x^*, y^*)v \geq 0, \\ v(\nabla_x g(x^*, y^*)u + \nabla_y g(x^*, y^*)v) = 0. \end{gathered} \tag{2.3}$$

When are these conditions sufficient? Consider the two possible cases.

Case 1: $v = 0$. It is clear that the point $(x^* + tu, y^* + tv) = (x^* + tu, 0)$ satisfies the first and third constraints in (2.2) for any $t > 0$. Suppose that $\nabla_x g(x^*, y^*) \neq 0$. Then, by (2.3), if $u \neq 0$, it holds that $\nabla_x g(x^*, y^*)u > 0$. By differentiability of g, we then obtain for all $t > 0$ small enough that

$$g(x^* + tu, 0) = t\nabla_x g(x^*, y^*)u + o(t) \geq 0.$$

Hence, $(x^* + tu, 0) \in D \; \forall\, t > 0$ small, which means that $h = (u, 0) \in T_D(z^*)$. Observe now that

$$F'(x^*, y^*) + PF''(x^*, y^*)[h] = (0, \nabla_x g(x^*, y^*)u),$$

and condition $\nabla_x g(x^*, y^*) \neq 0$ means precisely 2-regularity of F with respect to $h = (u, 0)$, where $u \neq 0$.

Case 2: $v > 0$. By (2.3), $\langle \nabla g(x^*, y^*), (u, v) \rangle = 0$. Suppose that $\nabla g(x^*, y^*) \neq 0$. Then g is regular in the classical sense, and the standard Lyusternik theorem implies that $h = (u, v)$ is a tangent direction to the set $\{(x, y) \mid g(x, y) = 0\}$: there exists some mapping $r(\cdot) = (r^1(\cdot), r^2(\cdot))$ such that

$$g(z^* + th + r(t)) = 0, \quad \|r(t)\| = o(t).$$

Furthermore, for any $t > 0$ sufficiently small

$$y^* + tv + r^1(t) = tv + o(t) \geq 0.$$

Hence, the point $z^* + th + r(t)$ satisfies all constraints in (2.2). Therefore, $h \in T_D(z^*)$. Again observe that

$$\begin{aligned} F'(x^*, y^*) + PF''(x^*, y^*)[h] &= (\nabla_x g(x^*, y^*)v, \nabla_x g(x^*, y^*)u + 2\nabla_y g(x^*, y^*)v) \\ &= (\nabla_x g(x^*, y^*)v, \nabla_y g(x^*, y^*)v). \end{aligned}$$

As is now easy to see, condition $\nabla g(x^*, y^*) \neq 0$ is precisely 2-regularity of F with respect to $h = (u, v)$, where $v \neq 0$. (Note that if $\nabla_x g(x^*, y^*) = 0$, then necessarily $\nabla_y g(x^*, y^*) = 0$ (by (2.3), because $v > 0$). Hence, under our assumption, case $v > 0$ can occur only if $\nabla_x g(x^*, y^*) \neq 0$.)

We have demonstrated that in both cases the assumption which makes necessary conditions of tangency for some direction h to be sufficient can be interpreted as 2-regularity of the mapping F with respect to this h.

In the next section we show that the intuition derived from this simple one-dimensional case is essentially valid in general. Specifically, 2-regularity of constraints that have to be satisfied as equalities (in a neighborhood of the point being considered) yields a precise description of the tangent cone to complementarity constraints at this point.

3. 2-regularity for MPEC. We start with recalling some useful objects from the MPEC theory [11, 10, 18, 15]. The standard *necessary* condition of tangency in MPEC is

$$T_D(z^*) \subset L, \tag{3.1}$$

where L is the "linearized cone":

$$L := \left\{ (u,v) \in \Re^n \times \Re^m \;\middle|\; \begin{array}{ll} \langle \nabla g_i(z^*), (u,v) \rangle \geq 0, v_i \geq 0, & i \in I_0, \\ v_i \langle \nabla g_i(z^*), (u,v) \rangle = 0, & i \in I_0, \\ \langle \nabla g_i(z^*), (u,v) \rangle = 0, & i \in I_g \end{array} \right\}. \tag{3.2}$$

Linearization here is understood differently from the usual notion in nonlinear programming. Indeed, L is not a polyhedral cone, except when the degenerate set I_0 is empty. Instead, L consists of a finite union of polyhedral cones. It is known [11] that (3.1) holds as equality when $g(\cdot,\cdot)$ is an affine function or when $z^* \in D$ is (strongly) regular in the sense of [17].

It is also clear that $T_D(z^*) = L$ under some *piecewise* constraint qualification. Let (R,Q) be the family of partitions of the degenerate set I_0, i.e., $R \cup Q = I_0$, $R \cap Q = \emptyset$. Associated with each partition (R,Q), define the *branch* of the feasible set

$$D_{RQ} := \left\{ (x,y) \;\middle|\; \begin{array}{ll} g_i(x,y) \geq 0, y_i = 0, & i \in R, \\ g_i(x,y) = 0, y_i \geq 0, & i \in Q \cup I_g \end{array} \right\}. \tag{3.3}$$

It is easy to see that

$$\bigcup_{(R,Q)} (D_{RQ} \cap \mathcal{V}) = D \cap \mathcal{V}, \quad T_D(z^*) = \bigcup_{(R,Q)} T_{D_{RQ}}(z^*), \tag{3.4}$$

where $\mathcal{V}$ is a neighborhood of z^*. If one assumes that, say, the MFCQ is satisfied at z^* for constraints defining each of the sets D_{RQ}, then the corresponding branch of L given by

$$L_{RQ} := \left\{ (u,v) \;\middle|\; \begin{array}{ll} v_i = 0, \langle \nabla g_i(z^*), (u,v) \rangle \geq 0, & i \in R, \\ \langle \nabla g_i(z^*), (u,v) \rangle = 0, & i \in Q \cup I_g, \\ v_i \geq 0, & i \in Q, \end{array} \right\} \tag{3.5}$$

will represent necessary *and* sufficient conditions of tangency for this D_{RQ}. Putting together all the pieces, one then obtains that $T_D(z^*) = L$.

In general, however, (3.1) does not hold as equality, and so further analysis is needed to obtain a precise description of the tangent directions. We next show how this can be done with the help of 2-regularity outlined in section 2.

Recall the mapping $F : \Re^n \times \Re^m \to \Re^{|I_0|+|I_g|}$ associated with equality constraints in D^* (which is locally equivalent to D):

$$F_i(x,y) := \begin{cases} y_i g_i(x,y), & i \in I_0, \\ g_i(x,y), & i \in I_g. \end{cases} \tag{3.6}$$

With this notation,

$$D^* := \left\{ (x,y) \;\middle|\; \begin{array}{ll} y_i \geq 0, g_i(x,y) \geq 0, & i \in I_0, \\ F(x,y) = 0 & \end{array} \right\}.$$

Note that the ith row of $F'(x^*,y^*)$ is

$$F'(x^*,y^*)_i = \nabla F_i(x^*,y^*) = \begin{cases} 0, & i \in I_0, \\ \nabla g_i(x^*,y^*), & i \in I_g. \end{cases}$$

Observe that F cannot be regular unless the degenerate set I_0 is empty. On the other hand, it is quite natural to study F from the point of view of 2-regularity. Note, however, that F represents only part of the constraints. Indeed, application of 2-regularity (and arguably, of any general regularity concept) to MPEC cannot be straightforward. This can already be seen from the one-dimensional example considered in section 2. The current general theory for irregular problems does not permit mixed equality and inequality constraints, except in some very special cases. Within the general theory available, if one is to insist on mixed constraints, then all irregularity has to be induced by equalities with inequalities being regular [2, 9], or vice versa [8]. Yet it can be verified that for MPEC even these two extreme cases cannot be applied. Fortunately, and thanks to the special complementarity structure of D, it appears possible to develop a special approach for MPEC, different from the general theory. We next show that whenever tangent directions can be characterized via 2-regularity for the subset of constraints in D^* corresponding to $F(x,y) = 0$, then these directions are automatically tangent for the complete set of constraints, without any further assumptions involving these constraints.

We start by computing some objects described in section 2. Let A_g be the $|I_g| \times (n+m)$ matrix with rows $\nabla g_i(z^*), i \in I_g$. With this notation,

$$\operatorname{Ker} F'(z^*) = \operatorname{Ker} A_g.$$

Furthermore,

$$\operatorname{Im} F'(z^*) = \{(0,w) \in \Re^{|I_0|} \times \Re^{|I_g|} \mid w \in \operatorname{Im} A_g\},$$

$$(\operatorname{Im} F'(z^*))^\perp = \{(p,q) \in \Re^{|I_0|} \times \Re^{|I_g|} \mid q \in (\operatorname{Im} A_g)^\perp\}.$$

Now, let P be the orthogonal projector onto $(\operatorname{Im} F'(z^*))^\perp$. Then

$$P(x,y) = (x, P_g y),$$

where P_g is the orthogonal projector onto $(\operatorname{Im} A_g)^\perp$. Differentiating F twice, it can be verified that for any $(u,v) \in \Re^n \times \Re^m$,

$$F''(z^*)[(u,v)] = \begin{pmatrix} B_0[(u,v)] \\ B_g[(u,v)] \end{pmatrix},$$

where $B_0[(u,v)]$ is the $|I_0| \times (n+m)$ matrix with rows

$$B_0[(u,v)]_i = v_i \nabla g_i(z^*) + \langle \nabla g_i(z^*), (u,v) \rangle e^i, \quad i \in I_0, \tag{3.7}$$

with $e^i \in \Re^{n+m}$ being the vector of zeros except for the $(n+i)$th component, which is equal to one; and $B_g[(u,v)]$ is the $|I_g| \times (n+m)$ matrix with rows

$$B_g[(u,v)]_i = \nabla^2 g_i(z^*)(u,v), \quad i \in I_g. \tag{3.8}$$

Observe further that

$$(B_0[(u,v)]^2)_i = 2v_i \langle \nabla g_i(z^*), (u,v) \rangle. \tag{3.9}$$

Taking into account this information, we have that

$$PF''(z^*)[(u,v)]^2 = 0 \quad \Leftrightarrow \quad B_0[(u,v)]^2 = 0, \ P_g B_g[(u,v)]^2 = 0,$$

which in turn is equivalent to

$$v_i \langle \nabla g_i(z^*), (u,v) \rangle = 0, \quad i \in I_0, \quad P_g B_g[(u,v)]^2 = 0.$$

We can now state necessary conditions for tangent directions to equality constraints in D^*, in terms of cone H_2 defined by (2.1):

$$\begin{array}{l} \langle \nabla g_i(z^*), (u,v) \rangle = 0, \quad i \in I_g, \\ v_i \langle \nabla g_i(z^*), (u,v) \rangle = 0, \quad i \in I_0, \\ P_g B_g[(u,v)]^2 = 0. \end{array}$$

Putting these together with standard necessary conditions for inequality constraints in D^*, we obtain

$$T_D(z^*) \subset H,$$

where

$$H := \left\{ (u,v) \;\middle|\; \begin{array}{l} \langle \nabla g_i(z^*), (u,v) \rangle \geq 0, v_i \geq 0, \quad i \in I_0, \\ v_i \langle \nabla g_i(z^*), (u,v) \rangle = 0, \quad i \in I_0, \\ \langle \nabla g_i(z^*), (u,v) \rangle = 0, \quad i \in I_g, \\ P_g B_g[(u,v)]^2 = 0 \end{array} \right\}. \tag{3.10}$$

Note that the above is not a standard set of necessary conditions: we have a mix of second-order conditions for equality constraints with first-order conditions for inequality constraints. In general, H is not piecewise polyhedral, although it can be such under some natural assumptions discussed below. But perhaps more interesting is the difficult case when H is not piecewise polyhedral. Even in that case, as we shall see, (3.10) gives a precise (i.e., $T_D(z^*) = H$) constructive description of tangency whenever the mapping F is 2-regular. This can further be used for deriving a special form of primal-dual optimality conditions in the situation where other techniques are not applicable.

We are now ready to state our description of the tangent cone to complementarity constraints.

THEOREM 3.1. *Let g be twice differentiable at $z^* \in D$, where D is given by* (1.1). *Then the following statements hold.*

1. *If* $h \in T_D(z^*)$, *then* $h \in H$.
2. *If* F *defined by* (3.6) *is* 2*-regular at* z^* *with respect to* $h \in H$, *then* $h \in T_D(z^*)$.

Proof. In view of the preceding discussion, the first assertion of the theorem requires no further justification.

Suppose now that F is 2-regular with respect to some $h = (u, v) \in H$. By Theorem 2.2, it then follows that h is tangent to the set $\{z \mid F(z) = 0\}$:

$$F(z^* + th + r(t)) = 0, \quad \|r(t)\| = o(t),$$

where $r(\cdot) = (r^1(\cdot), r^2(\cdot))$ is the mapping from Theorem 2.2. By the definition of F, this is equivalent to

$$\begin{aligned}(3.11) \qquad (y_i^* + tv_i + r_i^2(t))g_i(x^* + tu + r^1(t), y^* + tv + r^2(t)) = 0, \quad i \in I_0,\\ g_i(x^* + tv + r^1(t), y^* + tv + r^2(t)) = 0, \quad i \in I_g.\end{aligned}$$

To prove that $h \in T_D(z^*)$, it remains to show that $z^* + th + r(t)$ satisfies the remaining inequality constraints (for $t > 0$ sufficiently small). According to prior calculations,

$$F'(z^*) + PF''(z^*)[(u, v)] = \begin{pmatrix} B_0[(u, v)] \\ A_g + P_g B_g[(u, v)] \end{pmatrix},$$

and the assumption of 2-regularity means that the matrix on the right-hand side has full row rank. In particular, $B_0[(u, v)]$ must have full row rank. Hence (see (3.7)),

$$(3.12) \qquad v_i \nabla g_i(z^*) + \langle \nabla g_i(z^*), (u, v)\rangle e^i, \ i \in I_0, \text{ are linearly independent.}$$

Define

$$I_0^0 := \{i \in I_0 \mid v_i = 0\}, \quad I_0^1 := I_0 \setminus I_0^0 = \{i \in I_0 \mid v_i > 0\}.$$

It immediately follows from (3.10) that

$$\langle \nabla g_i(z^*), (u, v)\rangle = 0, \quad i \in I_0^1.$$

Note that the linear independence of vectors in (3.12) corresponding to $i \in I_0^0$ means that $\langle \nabla g_i(z^*), (u, v)\rangle \neq 0$, $i \in I_0^0$. Further, taking into account (3.10), this is equivalent to

$$(3.13) \qquad \langle \nabla g_i(z^*), (u, v)\rangle > 0, \quad i \in I_0^0.$$

Using (3.13), differentiability of g implies that for all $t > 0$ sufficiently small

$$g_i(x^* + tv + r^1(t), y^* + tv + r^2(t)) = t\langle \nabla g_i(z^*), (u, v)\rangle + o(t) > 0, \quad i \in I_0^0.$$

The latter relation, together with (3.11), yields

$$y_i^* + tv_i + r_i^2(t) = r_i^2(t) = 0, \quad i \in I_0^0.$$

Similarly, for all $t > 0$ small enough,

$$y_i^* + tv_i + r_i^2(t) = tv_i + r_i^2(t) > 0, \quad i \in I_0^1,$$

and so (3.11) implies that

$$g_i(x^* + tv + r^1(t), y^* + tv + r^2(t)) = 0, \quad i \in I_0^1.$$

In particular, we have established that for all $t > 0$ sufficiently small, $(x, y) = z^* + th + r(t)$ satisfies the inequality constraints $y_i \geq 0, g_i(x, y) \geq 0, i \in I_0 = I_0^0 \cup I_0^1$, which completes the proof that $h \in T_D(z^*)$. ∎

COROLLARY 3.2. *The following statements hold.*

1. *If there exists some set $K \subset H$ such that F is 2-regular with respect to every $h \in K$ and $\operatorname{cl} K = H$, then $T_D(z^*) = H$.*
2. *If*

$$\nabla g_i(z^*), \quad i \in I_g, \quad \text{are linearly independent,} \tag{3.14}$$

then $H = L$. In particular, under the assumption of the previous item, $T_D(z^*) = L$.

Proof. The first assertion follows from Theorem 3.1 and closedness of the tangent cone. For the second assertion, just notice that (3.14) implies that $(\operatorname{Im} A_g)^\perp = \{0\}$, and so $P_g = 0$. Hence, $H = L$. □

Therefore, whenever there exists a dense subset of H with respect to which F is 2-regular, H gives a precise description of the tangent cone. If (3.14) also holds, necessary conditions for tangency take a simpler piecewise polyhedral form.

We next show that the piecewise MFCQ is a sufficient condition for 2-regularity of F on a certain dense subset of H, as well as for (3.14). A branch D_{RQ}, defined by (3.3), satisfies MFCQ if

$$\begin{array}{l} \nabla g_i(z^*), \quad i \in Q \cup I_g, \quad e^i, i \in R, \quad \text{are linearly independent, and} \\ \exists\, (u,v) \in \Re^n \times \Re^m \text{ such that} \\ v_i = 0, \quad \langle \nabla g_i(z^*), (u,v) \rangle > 0, \quad i \in R, \\ \langle \nabla g_i(z^*), (u,v) \rangle = 0, \quad i \in Q \cup I_g, \\ v_i > 0, \quad i \in Q, \end{array} \tag{3.15}$$

where $e^i = (0, 0, \ldots, 1, \ldots, 0) \in \Re^{n+m}$ with one in the $(n+i)$th position. Under the piecewise MFCQ, for each branch D_{RQ} the (classical) linearized cone L_{RQ}, defined by (3.5), represents all the tangent directions. Note that, as discussed above, the fact that a piecewise constraint qualification implies the equality $T_D(z^*) = L$ is fairly obvious. The significance of the following result lies not in establishing this equality but in providing a new interpretation for piecewise MFCQ in terms of 2-regularity of a subset of constraints of the original nondecomposed problem. Since piecewise MFCQ is arguably the most natural condition guaranteeing that $T_D(z^*) = L$, this goes to show that the introduced 2-regularity condition is also natural for MPEC. Furthermore, piecewise MFCQ is merely a sufficient condition for 2-regularity, as would be demonstrated by an example.

THEOREM 3.3. *Suppose that piecewise MFCQ holds at $z^* \in D$. Then F is 2-regular with respect to every $h = (u,v) \in \Re^n \times \Re^m$ satisfying (3.15) for some pair (R,Q), and it holds that $T_D(z^*) = H = L$.*

Proof. Consider an arbitrary branch D_{RQ} of the feasible set D, and an arbitrary $h = (u,v)$ satisfying (3.15). We have to show that the rows of matrix

$$F'(z^*) + PF''(z^*)[(u,v)] = \begin{pmatrix} B_0[(u,v)] \\ A_g + P_g B_g[(u,v)] \end{pmatrix} = \begin{pmatrix} B_0[(u,v)] \\ A_g \end{pmatrix}$$

are linearly independent. By (3.7) and (3.15), we have that

$$B_0[(u,v)]_i = \begin{cases} \langle \nabla g_i(z^*), (u,v) \rangle e^i, & i \in R, \\ v_i \nabla g_i(z^*), & i \in Q. \end{cases}$$

By (3.15), it further holds that

$$\begin{array}{l} \langle \nabla g_i(z^*), (u,v) \rangle > 0, \quad i \in R, \\ v_i > 0, \quad i \in Q. \end{array}$$

The linear independence condition in (3.15) now implies the desired linear independence of the rows of $F'(z^*) + PF''(z^*)[(u, v)]$.

Taking into account that MFCQ subsumes (3.14), the last assertion follows immediately from Corollary 3.2. □

The following example demonstrates that 2-regularity of F is a weaker condition than piecewise MFCQ, i.e., the former may be satisfied when the latter is not. Furthermore, it shows that our results can yield a relatively simple description of the tangent cone even in the difficult situation when piecewise constraint qualifications fail.

Example 3.1. Let $n = 1$, $m = 2$, and consider the set D given by (1.1), where

$$g_1(x, y) = x + y_1, \quad g_2(x, y) = x^2 - (y_2 - a)^2,$$

with $a > 0$ being any given number. Consider the point $z^* = (x^*, y^*)$ with $x^* = 0$, $y_1^* = 0$, $y_2^* = a$. We have that $z^* \in D$ and $I_0 = \{1\}$, $I_g = \{2\}$,

$$H = \left\{ (u, v) \;\middle|\; \begin{array}{l} v_1 \geq 0,\ u + v_1 \geq 0, \\ v_1(u + v_1) = 0, \\ u^2 - v_2^2 = 0 \end{array} \right\}.$$

Since $\nabla g_2(z^*) = 0$, it is clear from (3.15) that the piecewise MFCQ does not hold at z^*. By direct computations, we have that

$$F'(z^*) + PF''(z^*)[(u, v)] = \begin{pmatrix} v_1 & u + 2v_1 & 0 \\ 2u & 0 & 2v_2 \end{pmatrix}.$$

It can be easily seen that F is 2-regular with respect to any element in $H \setminus \{0\}$, and hence, $T_D(z^*) = H$.

It is further interesting to note that even though $H \neq L$, the structure of H is no more complex than that of L. Indeed, observe that $u^2 - v_2^2 = 0 \Leftrightarrow u = v_2$ or $u = -v_2$. Hence, cone H is piecewise polyhedral, just like L. However, $H = T_D(z^*) \neq L$. This shows that 2-regularity can give a constructive description of the tangent cone when $T_D(z^*) \neq L$, even without going beyond the situation where the tangent cone is piecewise polyhedral.

We conclude this example by noting that it remains valid if we perturb g by any function δ such that $\delta(z) = o(\|z - z^*\|^2)$. In particular, this would not affect 2-regularity properties of F or constraints defining cone H (this is because the first and second derivatives at z^* would not change). After such a perturbation, it would still hold that $H = T_D(z^*) \neq L$.

4. Optimality conditions. In this section we show that 2-regularity can be used to derive optimality conditions, including a special form of primal-dual conditions, in situations when alternative approaches may not be applicable. Our results are meant as a complement to (rather than a substitute for) other types of optimality conditions for MPEC, some of which we shall discuss below.

We start with the primal form of optimality conditions, which can be derived in a standard way using the relations

$$\operatorname{cl} K \subset T_D(x^*) \subset H,$$

where

$$K := \{h \in H \mid F \text{ is 2-regular with respect to } h\}.$$

THEOREM 4.1. *Let f be once differentiable and g be twice differentiable at a point $z^* \in D$.*

If z^* is a local solution of (1.2)*, then*

$$\langle \nabla f(z^*), h\rangle \geq 0 \quad \forall\, h \in \operatorname{cl} K. \tag{4.1}$$

If it holds that

$$\langle \nabla f(z^*), h\rangle > 0 \quad \forall\, h \in H \setminus \{0\}, \tag{4.2}$$

then z^* is an isolated local solution of (1.2).

Note that when $\operatorname{cl} K = H = L$ (e.g., under the assumptions of Corollary 3.2), we recover B-stationarity as defined in [18], which is generally considered one of the natural primal stationarity concepts for MPEC (see also Example 4.1).

To obtain a primal-dual form of optimality conditions, one way is to compute the dual cone of K. In general, this is a difficult problem, as this cone need not be even piecewise polyhedral. However, we are able to derive special primal-dual optimality conditions using a different technique. This is possible under the assumption that there exists some $\bar{h} \in K$ for which inequality in (4.1) holds as equality (a "critical" direction). This assumption is justifiable in this context, and it is quite common in the literature, e.g., [3, 16]. If $\langle \nabla f(z^*), \bar{h}\rangle > 0$ for a given $\bar{h} \in K$, then it is clear that the same property holds for all h in some neighborhood $\mathcal{U}$ of $\bar{h}$. Hence, on $\mathcal{U}$ there is no contradiction with necessary optimality conditions, and z^* is still a candidate for being a solution. There is no more local information to be extracted using this $\bar{h}$. On the other hand, if $\langle \nabla f(z^*), \bar{h}\rangle = 0$, a further investigation is needed, and indeed more local information relevant to optimality of z^* can be obtained. Note that in the following result, 2-regularity of F is assumed with respect to this $\bar{h}$ only, and in particular, the knowledge of the entire tangent cone is not necessary. We first state our optimality conditions using the same objects as in 2-regularity constructions in section 3, which allows a more compact form. After the proof, we shall rewrite them in terms of the original problem data and give comparisons with some other results in the literature.

THEOREM 4.2. *Let f be once and g be twice differentiable at a point $z^* \in D$, which is a local minimizer for problem* (1.2)*. Assume further that*

$$\exists\, \bar{h} \in K \text{ such that } \langle \nabla f(z^*), \bar{h}\rangle = 0. \tag{4.3}$$

Then there exist unique $\bar{\lambda} \in \Re^{|I_0|}$ and $\bar{\mu}^1$, $\bar{\mu}^2 \in \Re^{|I_g|}$ (all depending on $\bar{h}$) such that

$$\nabla f(z^*) + (B_0[\bar{h}])^T \bar{\lambda} + A_g^T \bar{\mu}^1 + (B_g[\bar{h}])^T \bar{\mu}^2 = 0, \tag{4.4}$$

$$\bar{\mu}^1 \in \operatorname{Im} A_g, \quad \bar{\mu}^2 \in (\operatorname{Im} A_g)^{\perp}, \tag{4.5}$$

where the index sets I_0 and I_g are as defined earlier.

Moreover, the multiplier $\bar{\mu}^2$ satisfies the following additional condition: for every h such that

$$h \in \operatorname{Ker} A_g, \quad B_0[\bar{h}, h] = 0, \quad B_g[\bar{h}, h] \in \operatorname{Im} A_g, \tag{4.6}$$

it holds that

$$\langle \bar{\mu}^2, B_g[h]^2\rangle \geq 0. \tag{4.7}$$

Proof. It can be easily seen from Definition 2.1 that if 2-regularity of F holds for some $\bar{h} = (\bar{u}, \bar{v}) \in H$, then it holds for all h sufficiently close to $\bar{h}$. In other words, there exists a neighborhood $\mathcal{U}$ of $\bar{h}$ in $\Re^n \times \Re^m$ such that

$$H \cap \mathcal{U} \subset K.$$

Hence, by Theorem 4.1, we have that

$$\langle \nabla f(z^*),\, h\rangle \geq 0 \quad \forall\, h \in H \cap \mathcal{U}.$$

Observe that the latter relation and (4.3) imply that $\bar{h}$ is a local solution of the optimization problem

$$\min_{h \in H} \langle \nabla f(z^*),\, h\rangle. \tag{4.8}$$

The feasible set H of problem (4.8), given by (3.10), can be equivalently written in the following form (recalling definitions of B_0 and A_g):

$$H = \left\{ h = (u,\, v) \;\middle|\; \begin{array}{l} v_i \geq 0,\, \langle \nabla g_i(z^*),\, h\rangle \geq 0, \quad i \in I_0, \\ \frac{1}{2} B_0[h]^2 = 0, \\ A_g h + \frac{1}{2} P_g B_g[h]^2 = 0 \end{array} \right\}. \tag{4.9}$$

Here we have also taken into account the following obvious observation: for any h the elements $A_g h$ and $\frac{1}{2} P_g B_g[h]^2$ belong to orthogonal subspaces in $\Re^{|I_g|}$, and hence, they both are equal to zero if and only if their sum is equal to zero.

Next we show that in a neighborhood of $\bar{h}$ the set H is completely defined by equality constraints only, so in our local considerations we can omit the inequality constraints in (4.9). Indeed, recall that since F is 2-regular with respect to $\bar{h}$, we have by (3.13) that

$$\langle \nabla g_i(z^*),\, \bar{h}\rangle > 0, \quad i \in I_0^0, \tag{4.10}$$

where

$$I_0^0 = \{i \in I_0 \mid \bar{v}_i = 0\}. \tag{4.11}$$

Let h satisfy the equality $B_0[h]^2 = 0$. By (3.9), we then obtain that

$$v_i \langle \nabla g_i(z^*),\, h\rangle = 0, \quad i \in I_0.$$

This relation and (4.10) imply that if $h = (u, v)$ is sufficiently close to $\bar{h}$, then it must hold that

$$\begin{array}{l} v_i = 0,\, \langle \nabla g_i(z^*),\, h\rangle > 0, \quad i \in I_0^0, \\ v_i > 0,\, \langle \nabla g_i(z^*),\, h\rangle = 0, \quad i \in I_0^1, \end{array} \tag{4.12}$$

where

$$I_0^1 = I_0 \setminus I_0^0 = \{i \in I_0 \mid \bar{v}_i > 0\}. \tag{4.13}$$

We have therefore established that there exists a neighborhood $\mathcal{V}$ of $\bar{h}$ such that whenever $h \in \mathcal{V}$ satisfies the equality constraints in (4.9), it also satisfies the inequality constraints (by (4.12)). In other words, H is completely defined by its equality constraints (locally, in $\mathcal{V}$). Writing these equality constraints as $\Phi(h) = 0$, where

$$\Phi(h) = \begin{pmatrix} \frac{1}{2} B_0[h]^2 \\ A_g h + \frac{1}{2} P_g B_g[h]^2 \end{pmatrix},$$

we conclude that problem (4.8) is locally equivalent to

$$\min_{\Phi(h)=0} \langle \nabla f(z^*), h\rangle. \tag{4.14}$$

Furthermore,

$$\Phi'(\bar{h}) = \begin{pmatrix} B_0[\bar{h}] \\ A_g + P_g B_g[\bar{h}] \end{pmatrix}, \tag{4.15}$$

and, as we have verified in the proof of Theorem 3.1, the assumption of 2-regularity of F with respect to $\bar{h}$ means that the matrix on the right-hand side has full row rank. Thus Φ is regular at $\bar{h}$ in the classical sense. This means that classical first- and second-order optimality conditions for regular equality-constrained problems are applicable for problem (4.14) at its local solution $\bar{h}$. To this end, define the standard Lagrangian for (4.14):

$$\begin{aligned} L(h,\, \lambda,\, \mu) &= \langle \nabla f(z^*),\, h\rangle + \langle (\lambda,\, \mu),\, \Phi(h)\rangle \\ &= \langle \nabla f(z^*),\, h\rangle + \langle \lambda,\, \tfrac{1}{2} B_0[h]^2\rangle + \langle \mu,\, A_g h + \tfrac{1}{2} P_g B_g[h]^2\rangle. \end{aligned}$$

Representing μ as $\mu = \mu^1 + \mu^2$, where $\mu^1 \in \operatorname{Im} A_g$ and $\mu^2 \in (\operatorname{Im} A_g)^\perp$, the Lagrangian can be further rewritten as

$$L(h,\, \lambda,\, \mu) = \langle \nabla f(z^*),\, h\rangle + \frac{1}{2}\langle \lambda,\, B_0[h]^2\rangle + \langle \mu^1,\, A_g h\rangle + \frac{1}{2}\langle \mu^2,\, B_g[h]^2\rangle. \tag{4.16}$$

Now, by the classical first-order necessary optimality conditions, there exists a (unique, by necessity) pair $(\bar{\lambda},\, \bar{\mu}) \in \Re^{|I_0|} \times \Re^{|I_g|}$ such that

$$\nabla_h L(\bar{h},\, \bar{\lambda},\, \bar{\mu}) = 0,$$

and, by the second-order necessary optimality conditions, it further holds that

$$\langle \nabla_h^2 L(\bar{h},\, \bar{\lambda},\, \bar{\mu})h,\, h\rangle \geq 0 \quad \forall\, h \in \operatorname{Ker} \Phi'(\bar{h}).$$

Computing all the derivatives involved (by taking into account (4.15) and (4.16)), we obtain the asserted results. □

Using the definition of A_g, (3.7), (3.8), and (4.12) (for $h = \bar{h}$), we can rewrite our optimality conditions explicitly in terms of the problem data. Specifically, for (4.4) and the second relation in (4.5), we obtain

$$\nabla f(z^*) + \sum_{i \in I_0^0} \tilde{\lambda}_i e^i + \sum_{i \in I_0^1} \tilde{\lambda}_i \nabla g_i(z^*) + \sum_{i \in I_g} (\bar{\mu}_i^1 \nabla g_i(z^*) + \bar{\mu}_i^2 \nabla^2 g_i(z^*)\bar{h}) = 0 \tag{4.17}$$

and

$$\sum_{i \in I_g} \bar{\mu}_i^2 \nabla g_i(z^*) = 0,$$

where e^i is as defined earlier, the index sets I_0^0 and I_0^1 are defined in (4.11) and (4.13), respectively, and

$$\tilde{\lambda}_i = \begin{cases} \bar{\lambda}_i \langle \nabla g_i(z^*),\, \bar{h}\rangle, & i \in I_0^0, \\ \bar{\lambda}_i \bar{v}_i, & i \in I_0^1. \end{cases}$$

Similarly, we can rewrite the first two conditions on $h = (u, v)$ in (4.6) as follows:

$$\langle \nabla g_i(z^*),\, h\rangle = 0, \quad i \in I_g,$$

and

$$v_i = 0, \quad i \in I_0^0, \quad \langle \nabla g_i(z^*),\, h\rangle = 0, \quad i \in I_0^1,$$

respectively. Finally, taking into account (3.9), condition (4.7) takes the form

$$\sum_{i \in I_g} \bar{\mu}_i^2 \langle \nabla^2 g_i(z^*)h,\, h\rangle \geq 0.$$

Note that if the linear independence condition (3.14) is satisfied, then necessarily $\bar{\mu}^2 = 0$, and our optimality conditions take a simpler form. We next comment on some other approaches to constraint qualifications and optimality conditions for MPEC.

First, note that Theorem 4.2 subsumes optimality conditions for the *nondegenerate* case of $I_0 = \emptyset$, but in a way different from conditions based on disjunctive programming [11, 10]. Observe that if $I_0 = \emptyset$, then F is composed of components $g_i, i \in I_g$. Further, if $\nabla g_i(z^*), i \in I_g$, are linearly independent (condition (3.14)), then F is regular at z^* in the classical sense and, hence, 2-regular with respect to every h. We can then choose $\bar{h} = 0$ in (4.3). With this choice, the primal-dual characterization of optimality given by Theorem 4.2 reduces to $\nabla f(z^*) = \sum_{i \in I_g} \bar{\mu}_i^1 \nabla g_i(z^*)$, which is precisely the Karush–Kuhn–Tucker conditions in the situation where $I_0 = \emptyset$ and $I_y = \emptyset$. (Recall that we set $I_y = \emptyset$ merely to simplify notation; if $I_y \neq \emptyset$, then the gradients of $y_i = 0, i \in I_y$, enter into play in the obvious way.)

In [15], some conditions are obtained which ensure that the tangent cone and/or its dual can be computed as corresponding cones of certain standard nonlinear programs associated with MPEC. In particular, these cones are therefore polyhedral. Since this need not be the case in our development, it is clear that the two approaches are principally different. The motivation of [15] is to identify situations where MPEC optimality can be characterized without an excessive combinatorial burden of disjunctive programming. Note that our constraint qualification does not require decomposition of the feasible region for its verification. In this sense, it also contributes to the same goal.

We next comment on some results in [18] and, in particular, show that our approach can be useful to verify optimality when [18] may not be applicable, and vice versa. We start with some general comments and then give a number of examples.

For the feasible set D given by (1.1), all constraint qualifications considered in [18] are equivalent to the following linear independence constraint qualification (LICQ):

$$\nabla g_i(z^*),\, i \in I_0 \cup I_g, \quad e^i,\, i \in I_0, \quad \text{are linearly independent.} \tag{4.18}$$

Under this assumption, the following primal-dual necessary conditions hold:

$$\nabla f(z^*) = \sum_{i \in I_0 \cup I_g} \bar{\lambda}_i \nabla g_i(z^*) + \sum_{i \in I_0} \bar{\mu}_i e^i, \tag{4.19}$$
$$\bar{\lambda}_i \geq 0, \quad \bar{\mu}_i \geq 0, \quad i \in I_0.$$

To compare this result with ours, first note that LICQ (4.18) implies piecewise MFCQ, which is stronger than 2-regularity (recall Example 3.1). On the other hand, for our

primal-dual conditions we need a tangent direction $\bar{h}$ which is critical (condition (4.3)). It is therefore clear that the two constraint qualifications are not directly comparable. In other words, one can be satisfied when the other is not, and vice versa. Secondly, note that primal-dual conditions (4.17) and (4.19) are also essentially different. If (3.14) holds (which is certainly implied by LICQ), as discussed above, we have that $\bar{\mu}^2 = 0$. Observe now that, since $I_0^0 \cup I_0^1 = I_0$, the right-hand side of (4.17) involves twice fewer gradients of constraints from the index set I_0 than (4.19). That is, in (4.17) the multipliers corresponding to half of the constraints in I_0 are claimed to be zero. In this respect, (4.17) is sharper than (4.19). On the other hand, (4.19) contains additional conditions on the signs of the multipliers in I_0. Hence, the two optimality conditions are essentially different, neither one of them being stronger than the other.

We next give some examples. Note that in the first example below, LICQ (4.18) holds, which is a much stronger condition than is needed to apply our results. Thus this example is certainly not the most suitable for showing the utility of our approach. We include it mainly because it is extensively used in the literature, e.g., [18]. Examples 4.2 and 4.3 better illustrate the differences and our contribution.

Example 4.1. Let $n = m = 1$ and

$$f(x, y) = \alpha x + \beta y, \quad g(x, y) = x,$$

where α, $\beta \in \Re$ are parameters. In the terminology of [18], the point $z^* = (x^*, y^*) = 0$ is a B-stationary point (equivalent here to strongly stationary) if and only if $\alpha \geq 0$, $\beta \geq 0$. This is a primal stationarity condition. Concerning primal-dual stationarity, z^* is a weak stationary point for any α and β; it is a C-stationary point if and only if either $\alpha \geq 0$, $\beta \geq 0$ or $\alpha \leq 0$, $\beta \leq 0$; and it is an M-stationary point if and only if $\alpha \geq 0$, $\beta \geq 0$ or $\alpha\beta = 0$.

As is easy to see, $I_0 = \{1\}$, $F'(z^*) = 0$, $H = \{(u, v) \mid u \geq 0, v \geq 0, uv = 0\}$, and F is 2-regular at z^* with respect to all $(u, v) \in H \backslash \{0\}$ because $F'(z^*)+PF''(z^*)[(u, v)] = (v, u)$.

We start by illustrating our primal optimality conditions. If $\alpha > 0$, $\beta > 0$, then (4.2) holds, and the second assertion of Theorem 4.1 guarantees that z^* is an isolated local minimizer. The first assertion of Theorem 4.1 shows that z^* cannot be a minimizer unless $\alpha \geq 0$, $\beta \geq 0$. The latter is consistent with B-stationarity because in this example (3.14) holds, and $T_D(z^*) = \operatorname{cl} K = H = L$ (recall the comment after Theorem 4.1). Example 4.2 below shows that Theorem 4.1 is in fact useful beyond B-stationarity in the sense of [18].

Consider now primal-dual conditions.

Let $\alpha = 0$ and $\beta > 0$, in which case z^* is a minimizer. Note that $\bar{h} = (\bar{u}, \bar{v}) = (1, 0)$ satisfies (4.3). We have that $I_0^0 = \{1\}$, $I_0^1 = \emptyset$, $\tilde{\lambda}_1 = \bar{\lambda}_1 \langle \nabla g(z^*), \bar{h} \rangle = \bar{\lambda}_1$, and (4.17) holds with $\tilde{\lambda}_1 = -\beta$. All the stationarity conditions from [18] are also satisfied here.

Let $\alpha = 1$ and $\beta = -1$, in which case z^* is not a minimizer. In (4.3) we can take $\bar{h} = (\bar{u}, \bar{v}) = (1, 1)$, so that $I_0^1 = \{1\}$, $I_0^0 = \emptyset$, and $\tilde{\lambda}_1 = \bar{\lambda}_1 \bar{v}_1 = \bar{\lambda}_1$. There exists no $\tilde{\lambda}_1$ such that $(1, -1) + \tilde{\lambda}_1(1, 0) = 0$, and so (4.17) does not hold. Therefore this nonoptimal z^* is not stationary in our sense. Note that it is weakly stationary but not C- or M-stationary.

In the case $\alpha = 0$ and $\beta < 0$, the point z^* is not a minimizer. But all the stationarity concepts of [18] hold, and it can be seen that Theorem 4.2 also applies. In this case, none of those stationarity concepts is able to detect nonoptimality of z^*.

The following examples are intended to show that the presented optimality conditions can be useful in situations where other approaches (e.g., [15, 18, 6]) do not apply.

Example 4.2. Let g, x^*, and y^* be as defined in Example 3.1 and the objective function be given by $f(x,y) = \alpha x + \beta_1 y_1 + \beta_2 y_2 + o(\|(x,\, y)\|)$, where α, β_1, β_2 are parameters. Recall that not only (4.18) but even the much weaker condition (3.14) does not hold in this example. In particular, the constraint qualifications used in [15, 18] do not hold, and the corresponding results do not apply.

According to Example 3.1, the cone H consists of four rays spanned by $(1,\, 0,\, 1)$, $(1,\, 0,\, -1)$, $(-1,\, 1,\, 1)$, and $(-1,\, 1,\, -1)$, and the mapping F is 2-regular at z^* with respect to every element in $H \setminus \{0\}$. From the primal necessary optimality conditions stated in Theorem 4.1, it can be derived that z^* cannot be a local minimizer unless

$$0 \le \alpha \le \beta_1, \quad |\beta_2| \le \min\{a,\, |\beta_1 - a|\}.$$

Sufficient condition (4.2) is satisfied if and only if both of these inequalities hold strictly.

Next, for $\bar{h} = (1,\, 0,\, \pm 1)$, equality in (4.3) takes place if and only if $\alpha \pm \beta_2 = 0$, and (4.4), (4.5) hold with $\bar{\lambda} = -\beta_1$, $\bar{\mu}^1 = 0$, $\bar{\mu}^2 = \pm\beta_2/2$. For $\bar{h} = (-1,\, 1,\, \pm 1)$, equality in (4.3) takes place if and only if $-\alpha + \beta_1 \pm \beta_2 = 0$, and (4.4), (4.5) hold with the same multipliers.

"Second-order" condition (4.7) does not provide any additional information in this example, as for every $\bar{h} \in H \setminus \{0\}$, the subspace comprised by elements h satisfying (4.6) coincides with $\mathrm{span}\{\bar{h}\}$, and hence $B_g[h]^2 = 0$ for such h.

We next demonstrate the utility of the "second-order" condition (4.7). Note that this example also highlights the differences of the present paper as compared to optimality conditions in [6], where a general theory of 2-regularity is developed for $C^{1,1}$ mappings, and optimality conditions for MPEC are obtained after reformulating the constraints as $C^{1,1}$ equations.

Example 4.3. Let $n = m = 2$,

$$f(x,\, y) = x_1 + y_1 - y_2, \quad g_1(x,\, y) = x_1 + y_1, \quad g_2(x,\, y) = x_1^2 + x_2^2 - (y_2 - a)^2,$$

where $a > 0$ is a parameter. Consider the point $z^* = (x^*,\, y^*) \in D$ with $x^* = 0$, $y_1^* = 0$, $y_2^* = a$. It can be verified that this z^* is not a minimizer, but the stationarity conditions given in [6, section 6] hold. Results from [15, 18] are not applicable (because $\nabla g_2(z^*) = 0$ and thus constraint qualifications do not hold). We next show that Theorem 4.2 correctly classifies z^* as nonoptimal.

We have that $I_0 = \{1\}$, $I_g = \{2\}$,

$$H = \left\{ (u,\, v) \;\middle|\; \begin{array}{l} v_1 \ge 0,\, u_1 + v_1 \ge 0, \\ v_1(u_1 + v_1) = 0, \\ u_1^2 + u_2^2 - v_2^2 = 0 \end{array} \right\}.$$

It is easy to see that F is 2-regular with respect to all elements $h = (u,\, v) \in H$ satisfying $u_1 \ne 0$.

Take, e.g., $\bar{h} = (1,\, 0,\, 0,\, 1) \in H$. Then equality in (4.3) is satisfied, and (4.4), (4.5) hold with $\bar{\lambda} = -1$, $\bar{\mu}^1 = 0$, $\bar{\mu}^2 = -1$. Hence, z^* is still a candidate for being optimal. A different analysis based on [6] leads to the same conclusion up to this point, and no further analysis is possible based on that reference.

However, using Theorem 4.2, we see that linear equalities (4.6) hold for elements $h = (u,\, v)$ such that $v_1 = 0$, $u_1 - v_2 = 0$. For such h, it holds that

$$\bar{\mu}^2 B_g[h]^2 = -u_2^2 < 0$$

for any $u_2 \neq 0$. Hence, condition (4.7) is violated here, and z^* cannot be a local minimizer.

REFERENCES

[1] E.R. Avakov, *Extremum conditions for smooth problems with equality-type constraints*, Comput. Math. Math. Phys., 25 (1985), pp. 24–32.

[2] E.R. Avakov, *Necessary extremum conditions for smooth abnormal problems with equality and inequality constraints*, Math. Notes, 45 (1989), pp. 431–437.

[3] J.F. Ben-Tal and J. Zowe, *A unified theory of first- and second-order optimality conditions for extremum problems in topological vector spaces*, Math. Programming Stud., 19 (1982), pp. 39–76.

[4] Y. Chen and M. Florian, *The nonlinear bilevel programming problem: Formulations, regularity and optimality conditions*, Optimization, 32 (1995), pp. 193–209.

[5] A.F. Izmailov, *On certain generalizations of Morse's lemma*, Proc. Steklov Inst. Math., 220 (1998), pp. 138–153.

[6] 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., (2002), to appear.

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

[8] A.F. Izmailov and M.V. Solodov, *Optimality conditions for irregular inequality-constrained problems*, SIAM J. Control Optim., 40 (2001), pp. 1280–1295.

[9] U. Ledzewicz and H. Schättler, *High-order approximations and generalized necessary conditions for optimality*, SIAM J. Control Optim., 37 (1998), pp. 33–53.

[10] Z.-Q. Luo, J.-S. Pang, and D. Ralph, *Mathematical Programs with Equilibrium Constraints*, Cambridge University Press, Cambridge, UK, 1996.

[11] Z.-Q. Luo, J.-S. Pang, D. Ralph, and S.-Q. Wu, *Exact penalization and stationarity conditions of mathematical programs with equilibrium constraints*, Math. Programming, 75 (1996), pp. 19–76.

[12] R.J. Magnus, *On the local structure of the zero-set of a Banach space valued mapping*, J. Funct. Anal., 22 (1976), pp. 58–72.

[13] O.L. Mangasarian and S. Fromovitz, *The Fritz John necessary optimality conditions in the presence of equality and inequality constraints*, J. Math. Anal. Appl., 7 (1967), pp. 37–47.

[14] J.V. Outrata, *Optimality conditions for a class of mathematical programs with equilibrium constraints*, Math. Oper. Res., 24 (1999), pp. 627–644.

[15] J.-S. Pang and M. Fukushima, *Complementarity constraint qualifications and simplified B-stationarity conditions for mathematical programs with equilibrium constraints*, Comput. Optim. Appl., 13 (1999), pp. 111–136.

[16] J.-P. Penot, *Second-order conditions for optimization problems with constraints*, SIAM J. Control Optim., 37 (1999), pp. 303–318.

[17] S.M. Robinson, *Strongly regular generalized equations*, Math. Oper. Res., 5 (1980), pp. 43–62.

[18] H. Scheel and S. Scholtes, *Mathematical programs with complementarity constraints: Stationarity, optimality and sensitivity*, Math. Oper. Res., 25 (2000), pp. 1–22.

[19] A.A. Tretyakov, *Necessary and sufficient conditions for optimality of pth order*, Comput. Math. Math. Phys., 24 (1984), pp. 123–127.

[20] J.J. Ye, *Optimality conditions for optimization problems with complementarity constraints*, SIAM J. Optim., 9 (1999), pp. 374–387.

SIAM J. OPTIM.
Vol. 13, No. 2, pp. 386–405

SUPERLINEARLY CONVERGENT ALGORITHMS FOR SOLVING SINGULAR EQUATIONS AND SMOOTH REFORMULATIONS OF COMPLEMENTARITY PROBLEMS*

A. F. IZMAILOV† AND M. V. SOLODOV‡

Abstract. We propose a new algorithm for solving smooth nonlinear equations in the case where their solutions can be singular. Compared to other techniques for computing singular solutions, a distinctive feature of our approach is that we do not employ second derivatives of the equation mapping in the algorithm and we do not assume their existence in the convergence analysis. Important examples of once but not twice differentiable equations whose solutions are inherently singular are smooth equation-based reformulations of the nonlinear complementarity problems. Reformulations of complementarity problems serve both as illustration of and motivation for our approach, and one of them we consider in detail. We show that the proposed method possesses local superlinear/quadratic convergence under reasonable assumptions. We further demonstrate that these assumptions are in general not weaker and not stronger than regularity conditions employed in the context of other superlinearly convergent Newton-type algorithms for solving complementarity problems, which are typically based on nonsmooth reformulations. Therefore our approach appears to be an interesting complement to the existing ones.

Key words. nonlinear equations, singularity, regularity, complementarity, reformulation, superlinear convergence

AMS subject classifications. 90C30, 46T20, 47J07, 90C33

PII. S1052623401372946

1. Introduction.

In this paper, we are interested in solving nonlinear equations in the case where their solutions can be singular and smoothness requirements are weaker than those usually assumed in this context. Our development is partially motivated by the nonlinear complementarity problem, which we consider in detail, and for which our method takes a particularly simple and readily implementable form.

Let $F : V \to \mathbf{R}^n$ be a given mapping, where V is a neighborhood of a point $\bar{x}$ in $\mathbf{R}^n$, with $\bar{x}$ being a solution of the system of equations

$$F(x) = 0. \tag{1.1}$$

In the following, F is assumed to be once (but not necessarily twice) differentiable on V. In this setting, $\bar{x}$ is referred to as *singular solution* if the linear operator $F'(\bar{x})$ is singular, i.e.,

$$\det F'(\bar{x}) = 0,$$

or, equivalently,

$$\operatorname{corank} F'(\bar{x}) = \dim \ker F'(\bar{x}) > 0.$$

*Received by the editors June 15, 2001; accepted for publication (in revised form) March 28, 2002; published electronically September 24, 2002.

http://www.siam.org/journals/siopt/13-2/37294.html

†Computing Center of the Russian Academy of Sciences, Vavilova Str. 40, Moscow, GSP-1, Russia (izmaf@ccas.ru). The research of this author was supported by the Russian Foundation for Basic Research grants 99-01-00472 and 01-01-00810. This author also thanks IMPA, where he was a visiting professor during the completion of this work.

‡Instituto de Matemática Pura e Aplicada, Estrada Dona Castorina 110, Jardim Botânico, Rio de Janeiro, RJ 22460-320, Brazil (solodov@impa.br). The research of this author was supported in part by CNPq grant 300734/95-6, by PRONEX–Optimization, and by FAPERJ.

In other cases, $\bar{x}$ is referred to as a *regular solution*.

Singularity gives rise to numerous difficulties. It is well known that for Newton-type methods, at best one can guarantee linear convergence rate to a singular solution [6, 7, 9]. Moreover, it is not sufficient to choose a starting point only close enough to a solution (usually the set of appropriate starting points does not contain a full neighborhood of the solution, although this set is normally rather "dense" [18]). We refer the reader to the survey [19] and references therein. Another difficulty typical in this context is related to possible instability of a singular solution with respect to perturbations of F [27]. Certain special approaches to overcome those difficulties have been developed in the last two decades, but they employ second derivatives of F. Concerning methods for computing singular solutions, we cite [8, 20, 19, 43, 14, 1] and the more recent proposals in [26, 22, 21, 2, 27, 4] (of course, this list does not mention all contributions in this field).

One of the motivations for our new approach to solving singular equations lies in applications to the classical nonlinear complementarity problem (NCP) [37, 12, 13], which is to find an $x \in \mathbf{R}^n$ such that

$$g(x) \geq 0, \quad x \geq 0, \quad \langle g(x), x\rangle = 0, \tag{1.2}$$

where $g : \mathbf{R}^n \to \mathbf{R}^n$ is smooth. One of the most useful approaches to numerical and theoretical treatment of the NCP consists of reformulating it as a system of smooth or nonsmooth equations [35, 29, 46]. One possible choice of a smooth reformulation is given by the following function (for other choices, see section 5.1):

$$F : \mathbf{R}^n \to \mathbf{R}^n, \quad F_i(x) = 2g_i(x)x_i - (\min\{0,\, g_i(x) + x_i\})^2, \qquad i = 1, \ldots, n. \tag{1.3}$$

It is easy to check that for this mapping the solution set of the system of equations (1.1) coincides with the solution set of the NCP (1.2) [29, 47]. If $\bar{x}$ is a solution of the NCP, by direct computations (see section 3), we obtain that

$$F_i'(\bar{x}) = 2\begin{cases} 0 & \text{if } i \in I_0, \\ \bar{x}_i g_i'(\bar{x}) & \text{if } i \in I_1, \\ g_i(\bar{x})e^i & \text{if } i \in I_2, \end{cases} \tag{1.4}$$

where $e^1, \ldots, e^n$ denotes the standard basis in $\mathbf{R}^n$ and the index sets I_0, I_1, and I_2 are defined by

$$\begin{aligned} I_0 &:= \{i = 1, \ldots, n \mid g_i(\bar{x}) = 0,\, \bar{x}_i = 0\}, \\ I_1 &:= \{i = 1, \ldots, n \mid g_i(\bar{x}) = 0,\, \bar{x}_i > 0\}, \\ I_2 &:= \{i = 1, \ldots, n \mid g_i(\bar{x}) > 0,\, \bar{x}_i = 0\}. \end{aligned}$$

It is immediately clear that $F'(\bar{x})$ cannot be nonsingular, unless the index set I_0 is empty. The latter *strict complementarity* assumption is regarded as rather restrictive. Therefore, smooth NCP reformulation provided by (1.3) gives rise to *inherently* singular solutions of the corresponding system of equations. In fact, it is known that *any* other smooth NCP reformulation has the same singularity properties [31] (see also section 5.1). Furthermore, it is clear that F is once differentiable with Lipschitz-continuous derivative (if g is twice continuously differentiable), but F is not twice differentiable when $I_0 \neq \emptyset$. This is also a common property shared by all useful smooth reformulations; e.g., see the collection [16]. Thus NCP reformulations provide an interesting example of once differentiable nonlinear equations whose solutions

are inherently singular. As discussed above, application of standard numerical techniques (e.g., Newton methods) in this context is prone to difficulties (and even failure) because of singularity. On the other hand, known special approaches to computing singular solutions are inapplicable, since these require second derivatives of F. This is the apparent reason why superlinearly convergent Newton-type algorithms for solving the NCP are typically based on nonsmooth equation reformulations and nonsmooth Newton methods (see [13] for a discussion and some references). In this paper, we show that it is, in fact, possible to devise superlinearly convergent algorithms based on the smooth NCP reformulations. Specifically, we propose an alternative approach based on computing singular solutions of the smooth reformulation stated above, and show that conditions needed for convergence of our method are principally different from those required for convergence of known nonsmooth algorithms. Thus the two can be considered as a complement to each other.

We complete this section with some notation, which is fairly standard. We denote by $\mathcal{L}^n$ the space of linear operators from $\mathbf{R}^n$ to $\mathbf{R}^n$. For $A \in \mathcal{L}^n$, let $\ker A = \{x \in \mathbf{R}^n \mid Ax = 0\}$ stand for its kernel (null space), and $\operatorname{im} A = \{Ax \mid x \in \mathbf{R}^n\}$ stand for its image (range space). For a bilinear mapping $B : \mathbf{R}^n \times \mathbf{R}^n \to \mathbf{R}^n$ and an element $p \in \mathbf{R}^n$, we define the linear operator $B[p] \in \mathcal{L}^n$ by $B[p]\xi = B[p, \xi]$. Recall that *symmetric* bilinear mappings and linear operators of the form $p \to B[p] : \mathbf{R}^n \to \mathcal{L}^n$ are in isometrically isomorphic correspondence to each other, i.e., the correspondence is one-to-one, linear, and it preserves the norm. Therefore, in what follows we shall not be making a formal distinction between those objects. Given a set S in a vector space, we denote by $\operatorname{conv} S$ its convex hull and by $\operatorname{span} S$ its linear hull. Finally, by E we denote the identity operator in $\mathbf{R}^n$.

2. A general approach to solving singular equations. We start with describing an approach to computing singular solutions of twice differentiable nonlinear equations, which was developed in [26, 22, 27]. We then extend it to the setting of once differentiable mappings, and in the next section show how it applies to solving complementarity problems.

A solution $\bar{x}$ of (1.1) being regular is equivalent to saying that $\operatorname{im} F'(\bar{x}) = \mathbf{R}^n$, while singularity means that $\operatorname{im} F'(\bar{x}) \neq \mathbf{R}^n$. In this situation, one possibility for "regularizing" a singular solution $\bar{x}$ is to add to the left-hand side of (1.1) another term, which vanishes at $\bar{x}$ (so that $\bar{x}$ remains a solution), and such that its Jacobian at $\bar{x}$ "compensates" for the singularity of $F'(\bar{x})$ (so as to complement $\operatorname{im} F'(\bar{x})$ in $\mathbf{R}^n$). It is natural to base this extra term on the information about the first derivative of F.

To this end, define the mappings $P : V \to \mathcal{L}^n$, $h : V \to \mathbf{R}^n$, and

$$\Phi : V \to \mathbf{R}^n, \quad \Phi(x) = F(x) + P(x)F'(x)h(x), \tag{2.1}$$

and consider the equation

$$\Phi(x) = 0. \tag{2.2}$$

Suppose that $P(\cdot)$ is defined in such a way that for $\bar{P} = P(\bar{x})$ it holds that

$$\operatorname{im} F'(\bar{x}) \subset \ker \bar{P}. \tag{2.3}$$

Then, by the structure of Φ, solution $\bar{x}$ of (1.1) is also a solution for (2.2). Furthermore, if F is sufficiently smooth (at least twice differentiable at $\bar{x}$), then under appropriate assumptions on the first two derivatives of F at $\bar{x}$, and on $P(\cdot)$ and $h(\cdot)$,

it is possible to ensure that Φ is differentiable at $\bar{x}$, and $\bar{x}$ is a regular solution of (2.2). As these assumptions will not be used in this paper, we omit the details, referring the reader to [26, 27]. The regular solution $\bar{x}$ of (2.2) can be computed by means of effective special methods [26, 22, 27], or by conventional numerical techniques (the latter would typically require stronger assumptions, in order to ensure differentiability of Φ not only at $\bar{x}$ but also in its neighborhood). There exist certain general techniques to define $P(\cdot)$ and $h(\cdot)$ with necessary properties (see [26, 27]). However, when one has additional information about the structure of singularity of F at $\bar{x}$ (e.g., recall (1.4) for the NCP reformulation), it can often be used to choose $P(\cdot)$ and $h(\cdot)$ in a particularly simple and constructive way. One such application is precisely the NCP, where the subspace $\operatorname{im} F'(\bar{x})$ can be identified (locally, but without knowing $\bar{x}$), and so the two mappings can be chosen constant (see section 3).

In this paper, we shall focus exclusively on the case where it is possible to choose $P(\cdot) \equiv \bar{P}$ on V, with some $\bar{P} \in \mathcal{L}^n$ satisfying (2.3). We emphasize that, of course, $\bar{P}$ should be determined without knowing the exact solution $\bar{x}$. The simplest case when this is possible is when we know that $\operatorname{corank} F'(\bar{x}) = n$ (i.e., $F'(\bar{x}) = 0$), or when we are interested in determining a solution specifically with this particular type of singularity. In that case, it is natural to take $\bar{P} = E$. In section 3, we show how an appropriate $\bar{P}$ for the NCP reformulation can be determined using information available at any point close enough to a solution (but without knowing the solution itself). In general, if $P(\cdot)$ is defined as a constant $\bar{P}$ satisfying (2.3), one can also usually take $h(\cdot) \equiv p$, with $p \in \mathbf{R}^n \setminus \{0\}$ being an arbitrary element. Indeed, with those choices the function defined by (2.1) takes the form

$$\Phi(x) = \Phi_p(x) := F(x) + \bar{P}F'(x)p, \quad x \in V, \tag{2.4}$$

and $\bar{x}$ is still a solution of (2.2), due to (2.3). If F is twice differentiable at $\bar{x}$, then it is clear that Φ is differentiable at this point, and

$$\Phi'(\bar{x}) = F'(\bar{x}) + \bar{P}F''(\bar{x})[p]. \tag{2.5}$$

Therefore, $\bar{x}$ is a regular solution of (2.2) if the linear operator in the right-hand side of (2.5) is nonsingular. This is possible under appropriate assumptions. Since the case of twice differentiable F is not the subject of this paper, we shall not discuss technical details. We only note that nonsingularity of (2.5) subsumes the condition

$$\operatorname{im} F'(\bar{x}) + \operatorname{im} \bar{P} = \mathbf{R}^n.$$

Observe that the latter relation implies that (2.3) must hold as equality. Summarizing, we obtain the following assumptions on the choice of $\bar{P}$:

$$\ker \bar{P} = \operatorname{im} F'(\bar{x}), \quad \mathbf{R}^n = \operatorname{im} F'(\bar{x}) \oplus \operatorname{im} \bar{P}. \tag{2.6}$$

These assumptions clearly hold if, for example, $\bar{P}$ is the projector onto some complement of $\operatorname{im} F'(\bar{x})$ in $\mathbf{R}^n$ parallel to $\operatorname{im} F'(\bar{x})$. With this choice, nonsingularity of (2.5) formally coincides with the notion of 2-regularity of Φ at $\bar{x}$ with respect to $p \in \mathbf{R}^n$, in the sense of [23, 3, 27]. We note, however, that this connection does not seem conceptually important, and in fact, appears to be in some sense a coincidence. Indeed, in the case of once differentiable mappings considered below, the nonsingularity condition that would be required no longer has any direct relation to 2-regularity for mappings with Lipschitzian derivatives, as defined in [24, 25].

As a final note, we remark that it can be shown (by a simple argument; see [26, 27]) that if there exists at least one element $p \in \mathbf{R}^n$ such that the operator (2.5) is nonsingular, then it will be so for almost every $p \in \mathbf{R}^n$.

We conclude the discussion of the twice differentiable case by the following example, which is very simple but serves to illustrate the basic idea.

Example 2.1. Let $n = 1$ and $F : V \to \mathbf{R}$ be twice continuously differentiable on V, where V is a neighborhood of $\bar{x} \in \mathbf{R}$ which is a singular solution of (1.1). The latter means here that $F(\bar{x}) = F'(\bar{x}) = 0$. Taking $\bar{P} = E$ and any $p \in \mathbf{R} \setminus \{0\}$, we obtain the following regularized equation: $\Phi(x) = F(x) + F'(x)p = 0$. Obviously, $\Phi(\bar{x}) = 0$ and $\Phi'(\bar{x}) = F''(\bar{x})p$, which is distinct from zero for any $p \in \mathbf{R} \setminus \{0\}$, provided $F''(\bar{x}) \neq 0$. This shows that in this example, if $F''(\bar{x}) \neq 0$, singularity can be easily dealt with by using the second-order information.

In the approach outlined above, F is assumed to be twice differentiable. Suppose now that F is once (but not twice) differentiable, and its first derivative is Lipschitz-continuous on V. Then Φ defined by (2.4) is also Lipschitz-continuous on V, and it is natural to try to apply to the corresponding equation (2.2) the *generalized (nonsmooth) Newton method* [32, 33, 41, 42, 40, 28]. We emphasize that we shall use the nonsmooth Newton method to solve a (nonsmooth) regularization of a smooth equation. In the context of NCP, this should be compared to the more traditional approach of solving an inherently nonsmooth reformulation by the nonsmooth Newton method. As we shall show in section 4, the two different approaches lead to two different regularity conditions, neither of which is weaker or stronger than the other.

Let $\partial\Phi(x)$ denote the Clarke's *generalized Jacobian* [5] of Φ at $x \in V$. That is,

$$\partial\Phi(x) = \operatorname{conv} \partial_B\Phi(x),$$

where $\partial_B\Phi(x)$ stands for the *B-subdifferential* [45] of Φ at x, which is the set

$$\partial_B\Phi(x) = \{H \in \mathcal{L}^n \mid \exists \{x^k\} \subset D_\Phi \ : \ x^k \to x \text{ and } \Phi'(x^k) \to H\},$$

with $D_\Phi \subset V$ being the set of points at which Φ is differentiable. With this notation, the nonsmooth Newton method is the following iterative procedure:

$$x^{k+1} = x^k - (H(x^k))^{-1}\Phi(x^k), \quad H(x^k) \in \partial\Phi(x^k), \quad k = 0, 1, \ldots . \tag{2.7}$$

It is well known [42, 40, 28] that if

(i) Φ is semismooth [36] at $\bar{x}$, and

(ii) all the linear operators comprising $\partial\Phi(\bar{x})$ are nonsingular,

then the process (2.7) is locally well defined and superlinearly convergent to $\bar{x}$. Moreover, if Φ is strongly semismooth [36], then the rate of convergence is quadratic. The regularity condition (ii) can be relaxed if a more specific rule of determining $H(x^k) \in \partial\Phi(x^k)$ is employed. For example, if one chooses $H(x^k) \in \partial_B\Phi(x^k)$, then it is enough to assume *BD-regularity*, i.e., that all elements in $\partial_B\Phi(\bar{x})$ are nonsingular [41].

In applications, Φ usually has some special (tractable) structure, and at each iterate x^k we are interested in obtaining just one, preferably easily computable, $H(x^k) \in \partial\Phi(x^k)$. This would be precisely the case here. The choice of an element in $\partial\Phi(x)$ that we suggest to use in the nonsmooth Newton method for solving (2.2), with Φ given by (2.4), is the following:

$$H(x) = H_p(x) := F'(x) + (\bar{P}F')'(x; p), \quad x \in V, \tag{2.8}$$

where $(\bar{P}F')'(x;\, p)$ denotes the usual directional derivative of the mapping $\bar{P}F'(\cdot)$ at $x \in V$ with respect to a direction $p \in \mathbf{R}^n$. In section 3, we show that this $H(x)$ is explicitly and easily computable for the NCP reformulations. The validity of the choice suggested in (2.8) for an element of $\partial\Phi(x)$ is actually not so obvious. The possibility of choosing the directional derivative $(\bar{P}F')'(x;\, p)$ as an element in the generalized Jacobian of $\bar{P}F'(x)p$ is based on the following fact. At a point $x \in V$ where $\bar{P}F'(\cdot)$ is differentiable, its derivative is in fact the second derivative of $\bar{P}F(\cdot)$. Due to this, $(\bar{P}F')'(x)$ can be considered as a symmetric bilinear mapping. This symmetry will be essential in the proof of Lemma 2.1 below. For a mapping $x \to Q(x)p$, where $p \in \mathbf{R}^n$ and $x \to Q(x) : \mathbf{R}^n \to \mathcal{L}^n$ is an arbitrary Lipschitzian mapping, the inclusion $Q'(x;\, p) \in \partial(Q(x)p)$ can be in general invalid.

LEMMA 2.1. *Suppose that $F : V \to \mathbf{R}^n$ has a Lipschitzian derivative on V, where V is an open set in $\mathbf{R}^n$. Assume that for some $\bar{P} \in \mathcal{L}^n$ the mapping $\bar{P}F' : V \to \mathcal{L}^n$ is directionally differentiable at a point $x \in V$ with respect to a direction $p \in \mathbf{R}^n$.*

Then $H(x) \in \partial\Phi(x)$, where H and Φ are defined in (2.8) *and* (2.4)*, respectively.*

Proof. Since $\bar{P}F'(\cdot)$ is clearly Lipschitz-continuous, using further the assumption that $\bar{P}F'$ is directionally differentiable at x with respect to p, it follows that there exists a linear operator $B \in \partial(\bar{P}F')(x)$ $(B : \mathbf{R}^n \to \mathcal{L}^n)$ such that

$$(\bar{P}F')'(x;\, p) = Bp. \tag{2.9}$$

The above conclusion can be deduced from [42, Lemma 2.2(ii)] after identifying the space $\mathcal{L}^n$ with the equivalent space $\mathbf{R}^m$, $m = n^2$, and using the equivalence of the norms in finite-dimensional spaces.

By the definition of the generalized Jacobian, $B \in \partial(\bar{P}F')(x)$ means that there exist an integer m, sequences $\{x^{i,k}\} \subset V$, and numbers λ_i, $i = 1, \ldots, m$, with the following properties: $\lambda_i \geq 0\,, \sum_{i=1}^m \lambda_i = 1$, $\bar{P}F'(\cdot)$ is differentiable at each $x^{i,k}$, and

$$x = \lim_{k\to\infty} x^{i,k}, \quad B = \sum_{i=1}^m \lambda_i \lim_{k\to\infty} (\bar{P}F')'(x^{i,k}), \tag{2.10}$$

where the limits in the right-hand side of the second equality exist for each $i = 1, \ldots, m$.

Note that differentiability of $\bar{P}F'(\cdot)$ at each $x^{i,k}$ means that the mapping $\bar{P}F : V \to \mathbf{R}^n$ is twice differentiable at these points. Taking into account the symmetry of the bilinear mapping representing the second derivative, we conclude that

$$\Phi'(x^{i,k}) = F'(x^{i,k}) + (\bar{P}F')'(x^{i,k})p.$$

Therefore,

$$\begin{aligned}
\sum_{i=1}^m \lambda_i \lim_{k\to\infty} \Phi'(x^{i,k}) &= F'(x) + \sum_{i=1}^m \lambda_i \lim_{k\to\infty} (\bar{P}F')'(x^{i,k})p \\
&= F'(x) + Bp \\
&= F'(x) + (\bar{P}F')'(x;\, p) \\
&= H(x),
\end{aligned}$$

where the second equality follows from (2.10), and the third from (2.9). Using the definition of the generalized Jacobian, we conclude that $H(x) \in \partial\Phi(x)$. ∎

Remark 2.1. There exists another way to construct the regularized equation $\Phi(x) = 0$, which can have advantages in certain situations over the one described above. Specifically, the mapping Φ defined by (2.4) can be modified as follows:

$$\Phi(x) := (E - \bar{P})F(x) + \bar{P}F'(x)p, \quad x \in V. \tag{2.11}$$

It is clear that, with this definition, $\bar{x}$ is still a solution of $\Phi(x) = 0$. Modifying H accordingly, we have

$$H(x) := (E - \bar{P})F'(x) + (\bar{P}F')'(x; p), \quad x \in V. \tag{2.12}$$

Furthermore, it is clear that Lemma 2.1 is still valid with Φ and H defined by (2.11) and (2.12). Finally, it is easy to see that since $\bar{P}F'(\bar{x}) = 0$, the possible limits of $H(x)$ as $x \to \bar{x}$ are the same, whether H is defined by (2.8) or (2.12). Hence, the regularity condition at $\bar{x}$ that would be needed for the superlinear convergence of our method is again the same, whether the method is applied to one regularized equation or the other.

The possible advantage of the modified equation is the following. If the singularity of $F'(\bar{x})$ has a certain structure, then not all the components of F may need to be computed in (2.11). Furthermore, (2.12) can also take a simpler form in that case. For example, suppose that $F'(\bar{x})$ is such that $\bar{P}$ satisfying (2.3) can be chosen as the orthogonal projector onto the subspace span $\{e^i\,,\, i \in I\}$, where $e^1, \ldots, e^n$ is the standard basis in $\mathbf{R}^n$ and $I \subset \{1, \ldots, n\}$. Then $E - \bar{P}$ is the orthogonal projector onto span $\{e^i\,,\, i \in \{1, \ldots, n\} \setminus I\}$. It is easy to see that, in this case, (2.11) would not require computing the function values $F_i(x)$, $i \in I$. Furthermore, the derivatives of F_i, $i \in I$, would not appear in (2.12), and so this part would also be simplified. This feature would be further illustrated in the context of NCP in section 3.

Next, we shall also consider the following modification of the Newton algorithm (2.7), which will be useful for solving the NCP reformulation in section 3:

(2.13)

$$x^{k+1} = x^k - (\tilde{H}(x^k))^{-1}\Phi(x^k), \quad \|\tilde{H}(x^k) - H(x^k)\| = O(\|x^k - \bar{x}\|), \quad H(x^k) \in \partial\Phi(x^k), \quad k = 0, 1, \ldots .$$

This modification is essentially motivated by the idea of "truncating" elements of the (generalized) Jacobian by omitting the terms which vanish at the solution $\bar{x}$. These terms typically involve some higher-order derivatives of the problem data (in the context of NCP (1.2), the second derivatives of g), and so it can be advantageous not to compute them, if possible.

Note that the regularity condition which is typically employed in nonsmooth Newton methods consists of saying that every element in the generalized Jacobian $\partial\Phi(\bar{x})$ (or the B-subdifferential $\partial_B\Phi(\bar{x})$) is nonsingular (recall condition (ii) stated above). This seems to be unnecessarily restrictive, because in most implementable algorithms some specific rule to choose $H(x^k) \in \partial\Phi(x^k)$ is used. We shall therefore replace the traditional condition by a weaker one. Specifically, we shall assume that all the possible limits of $H(x^k)$ as $x^k \to \bar{x}$ are nonsingular, where $H(x^k)$ is precisely the element given by (2.8) (or by (2.12)). To this end, we shall define the set

$$\Delta\Phi_p(\bar{x}) := \{\bar{H} \in \mathcal{L}^n \mid \exists\{x^k\} \subset V \ : \ x^k \to \bar{x}, \ \ H_p(x^k) \to \bar{H}\}.$$

Our regularity assumption would be that elements in $\Delta\Phi_p(\bar{x})$ are nonsingular. We remind the reader that this set is the same for both choices of H, i.e., (2.8) and

(2.12). We point out that, unlike in the twice differentiable case, this regularity condition cannot be related to the notion of 2-regularity [24, 25] of Φ at $\bar{x}$.

With Lemma 2.1 in hand, convergence of algorithms (2.7) and (2.13), with the data defined in (2.4) and (2.8) or (2.11) and (2.12), can be established similarly to [41], but taking into account the modified nonsingularity assumption.

THEOREM 2.2. *Suppose $F : V \to \mathbf{R}^n$ has a Lipschitz-continuous derivative on V, where V is a neighborhood of a solution $\bar{x}$ of (1.1). Let $\bar{P} \in \mathcal{L}^n$ satisfy (2.3). Assume further that the mapping $\bar{P}F' : V \to \mathcal{L}^n$ is directionally differentiable with respect to a direction $p \in \mathbf{R}^n$ at any point in V, and the mapping $\bar{P}F'(\cdot)p : V \to \mathbf{R}^n$ is semismooth at $\bar{x}$. Let Φ and H be defined by (2.4) and (2.8), or (2.11) and (2.12). Assume further that all linear operators comprising $\Delta\Phi_p(\bar{x})$ are nonsingular.*

Then the iterates given by (2.7) or (2.13) are locally well defined and converge to $\bar{x}$ superlinearly. If, in addition, the mapping $\bar{P}F'(\cdot)p$ is strongly semismooth at $\bar{x}$, then the rate of convergence is quadratic.

Proof. It is easy to see that under our regularity assumption, $(H(\cdot))^{-1}$ is locally uniformly bounded. Indeed, assume the contrary, i.e., that there exists a sequence $\{x^k\} \subset V$ such that $x^k \to \bar{x}$, and the sequence $\{(H(x^k))^{-1}\}$ is unbounded (this subsumes the possibility that some elements of the latter sequence are not even well defined). Recall that the generalized Jacobian is locally bounded [5]. Since, by Lemma 2.1, $H(x^k) \in \partial\Phi(x^k)$ for every k, it follows that the sequence $\{H(x^k)\}$ is bounded. Hence, we can assume that $\{H(x^k)\}$ converges to some $\bar{H} \in \Delta\Phi_p(\bar{x})$, where the inclusion is by the very definition of the set $\Delta\Phi_p(\bar{x})$. But then $\bar{H}$ is nonsingular, which is in contradiction with the earlier assumption that $\{(H(x^k))^{-1}\}$ is unbounded.

Consider first algorithm (2.7), and suppose that the iterates are well defined up to some index $k \geq 0$. We have that

$$\begin{aligned}\|x^{k+1} - \bar{x}\| &= \|(H(x^k))^{-1}(\Phi(x^k) - \Phi(\bar{x}) - H(x^k)(x^k - \bar{x}))\| \\ &\leq M\|\Phi(x^k) - \Phi(\bar{x}) - H(x^k)(x^k - \bar{x})\|,\end{aligned}$$

where $M > 0$. Note that when $\bar{P}F'(\cdot)p$ is (strongly) semismooth, so is $\Phi(\cdot)$. It is known [40, Proposition 1] that semismoothness of Φ at $\bar{x}$ implies that

$$\sup_{H \in \partial\Phi(\bar{x}+\xi)} \|\Phi(\bar{x} + \xi) - \Phi(\bar{x}) - H\xi\| = o(\|\xi\|)$$

(the latter property was introduced in the context of the nonsmooth Newton methods in [33]). Using Lemma 2.1 and combining the last two relations, well-definedness of the whole sequence $\{x^k\}$ and its superlinear convergence to $\bar{x}$ follow by a standard argument.

In the strongly semismooth case, one has that

$$\sup_{H \in \partial\Phi(\bar{x}+\xi)} \|\Phi(\bar{x} + \xi) - \Phi(\bar{x}) - H\xi\| = O(\|\xi\|^2),$$

and so convergence is quadratic.

Consider now the iterates $\{x^k\}$ generated by (2.13). By our regularity assumption and the classical results of linear analysis, the condition

$$\|\tilde{H}(x^k) - H(x^k)\| = O(\|x^k - \bar{x}\|)$$

implies that

$$\|(\tilde{H}(x^k))^{-1} - (H(x^k))^{-1}\| = O(\|x^k - \bar{x}\|).$$

Hence,

$$\|(\tilde{H}(x^k))^{-1}\Phi(x^k) - (H(x^k))^{-1}\Phi(x^k)\| \le \|(\tilde{H}(x^k))^{-1} - (H(x^k))^{-1}\|\|\Phi(x^k) - \Phi(\bar{x})\| = O(\|x^k - \bar{x}\|^2),$$

where the Lipschitz-continuity of Φ was also used. It follows that the difference between the original and modified steps is of the second order. By the obvious argument, it can now be easily seen that the modified algorithm has the same convergence rate as the original one. □

Note that, in principle, our regularity condition depends not only on the structure of the singularity of F at $\bar{x}$, but also on the choice of p. Implementation of this approach presumes that there exists at least one $p \in \mathbf{R}^n$ for which this condition is satisfied. Furthermore, a way to choose such p should be available. Fortunately, a typical situation is the following. The *existence* of one suitable p can usually be established under some reasonable regularity assumption. Then, given the existence of one such p, it can further be proven that the set of appropriate elements is, in fact, open and dense in the whole space. Hence, p can be chosen arbitrary, with the understanding that almost any is suitable. We shall come back to this issue in section 4, where regularity conditions for NCP are discussed. In the computational experience of [22, 26], where conceptually related methods for smooth operator equations are considered, a random choice of p does the job. Even though this choice certainly affects the rate and range of convergence, the differences between different choices are usually not dramatic.

Finally, we remark that the development presented above can be extended to the case when $P(\cdot)$ is not necessarily constant, but it is a Lipschitzian mapping satisfying (2.6) with $\bar{P} = P(\bar{x})$. In that case, we would have to provide a technique to define such $P(\cdot)$ in the general setting. Such techniques are possible, but they go beyond the scope of the present paper. Here we are mainly concerned with a specific application of our approach to the NCP, which we consider next.

3. Algorithm for the NCP. Consider the NCP (1.2), and its reformulation as a system of smooth equations (1.1), given by (1.3). For convenience, we restate the associated function F, which is

$$F : \mathbf{R}^n \to \mathbf{R}^n, \quad F_i(x) = 2g_i(x)x_i - (\min\{0,\, g_i(x) + x_i\})^2, \qquad i = 1, \ldots, n.$$

We choose a specific reformulation for the clarity of presentation. In section 5.1, we show that our analysis is intrinsic and extends to other smooth reformulations.

Let $\bar{x} \in \mathbf{R}^n$ be a solution of NCP. Suppose that g is twice continuously differentiable in some neighborhood V of $\bar{x}$ in $\mathbf{R}^n$. Then it is easy to see that F has a Lipschitz-continuous derivative on V, which is given by

$$F_i'(x) = 2(x_i g_i'(x) + g_i(x)e^i - \min\{0,\, g_i(x) + x_i\}(g_i'(x) + e^i)), \quad i = 1, \ldots, n, \quad x \in V, \tag{3.1}$$

where $e^1, \ldots e^n$ is the standard basis in $\mathbf{R}^n$. Recalling the three index sets

$$I_0 := \{i = 1, \ldots, n \mid g_i(\bar{x}) = 0,\, \bar{x}_i = 0\},$$
$$I_1 := \{i = 1, \ldots, n \mid g_i(\bar{x}) = 0,\, \bar{x}_i > 0\},$$
$$I_2 := \{i = 1, \ldots, n \mid g_i(\bar{x}) > 0,\, \bar{x}_i = 0\},$$

from (3.1) we immediately obtain that

$$F_i'(\bar{x}) = 2\begin{cases} 0 & \text{if } i \in I_0, \\ \bar{x}_i g_i'(\bar{x}) & \text{if } i \in I_1, \\ g_i(\bar{x})e^i & \text{if } i \in I_2. \end{cases} \tag{3.2}$$

As already discussed in section 1, the Jacobian $F'(\bar{x})$ is necessarily singular whenever $I_0 \neq \emptyset$, the latter being the usual situation for complementarity problems of interest. Furthermore, F is not twice differentiable. Hence, smooth NCP reformulations fall precisely within the framework of section 2. Such equations cannot be effectively solved by previously available methods, and so our approach comes into play. We next show that in the setting of NCP, the general algorithm introduced in section 2 takes a simple implementable form.

Given the structure of $F'(\bar{x})$, we have that

$$\operatorname{im} F'(\bar{x}) \subset \operatorname{span}\{e^i,\ i \in I_1 \cup I_2\}.$$

Then the natural choice of $\bar{P}$ satisfying $\operatorname{im} F'(\bar{x}) \subset \ker \bar{P}$ (recall condition (2.3)) is the operator with the matrix representation consisting of rows

$$\bar{P}_i = \begin{cases} e^i & \text{if } i \in I_0, \\ 0 & \text{if } i \in I_1 \cup I_2. \end{cases} \tag{3.3}$$

At the end of this section, we shall show how to define $\bar{P}$ without knowing the solution $\bar{x}$ (clearly, this task reduces to identifying the index set I_0). This is possible by means of *error bound* analysis. A sufficient condition for our error bound is weaker than b-regularity [39], which is currently the weakest assumption under which Newton methods for nonsmooth NCP reformulations are known to be (superlinearly) convergent [30, 34].

Once $\bar{P}$ is defined according to (3.3), we fix $p \in \mathbf{R}^n \setminus \{0\}$ arbitrarily. Then the function Φ defined by (2.4) takes the form

$$\Phi_i(x) = \begin{cases} F_i(x) + \langle F_i'(x),\, p\rangle & \text{if } i \in I_0, \\ F_i(x) & \text{if } i \in I_1 \cup I_2, \end{cases} \qquad x \in V. \tag{3.4}$$

According to section 2, $\bar{x}$ is a solution of $\Phi(x) = 0$, which is our "regularized" equation. We proceed to derive explicit forms for iterations of algorithms (2.7) and (2.13), and the regularity condition needed for their convergence.

First, by (2.8) and (3.3), the matrix representation of $H(x)$, which is the element of $\partial\Phi(x)$ employed in algorithm (2.7), consists of rows

$$H_i(x) = \begin{cases} F_i'(x) + (F_i')'(x;\, p) & \text{if } i \in I_0, \\ F_i'(x) & \text{if } i \in I_1 \cup I_2, \end{cases} \qquad x \in V. \tag{3.5}$$

Furthermore, the directional derivatives employed in (3.5) exist and can be obtained explicitly from (3.1):

$$\begin{aligned} (F_i')'(x;\, p) = {} & 2(x_i g_i''(x)p + p_i g_i'(x) + \langle g_i'(x),\, p\rangle e^i \\ & - \min\{0,\, g_i(x) + x_i\} g_i''(x)p - \gamma_i(x,\, p)(g_i'(x) + e^i)), \\ & i = 1, \ldots, n,\ x \in V, \end{aligned} \tag{3.6}$$

where

$$\gamma_i(x,\,p) = \begin{cases} \langle g_i'(x),\,p\rangle + p_i & \text{if } g_i(x) + x_i < 0, \\ \min\{0,\,\langle g_i'(x),\,p\rangle + p_i\} & \text{if } g_i(x) + x_i = 0, \\ 0 & \text{if } g_i(x) + x_i > 0, \end{cases} \tag{3.7}$$
$$i = 1,\,\ldots,\,n,\; x \in V.$$

Note that, according to (3.5), one has to compute $(F_i')'(\cdot;\,p)$ only for $i \in I_0$. Another useful observation which would suggest truncation of the Jacobian to be discussed later is that for $i \in I_0$ all the terms in (3.6) involving the second derivatives of g vanish at $\bar{x}$.

Furthermore, taking into account (3.5), (3.2), (3.6), and (3.7), we conclude that the matrix representation of an arbitrary limit point $\bar{H}$ of $H(x)$ as $x \to \bar{x}$ consists of rows

$$\bar{H}_i = 2 \begin{cases} -\langle g_i'(\bar{x}),\,p\rangle g_i'(\bar{x}) - p_i e^i \;\text{ or }\; p_i g_i'(\bar{x}) + \langle g_i'(\bar{x}),\,p\rangle e^i & \text{if } i \in I_0, \\ \bar{x}_i g_i'(\bar{x}) & \text{if } i \in I_1, \\ g_i(\bar{x}) e^i & \text{if } i \in I_2. \end{cases}$$

Hence, we can state the following *sufficient* condition for nonsingularity of every linear operator in $\Delta\Phi_p(\bar{x})$. Denote by $\mathcal{J}$ the collection of pairs of index sets $(J_1,\,J_2)$ such that $J_1 \cup J_2 = I_0$, $J_1 \cap J_2 = \emptyset$. Our regularity condition consists of saying that, for every pair of index sets $(J_1,\,J_2) \in \mathcal{J}$, it holds that

$$\left.\begin{array}{l} \langle g_i'(\bar{x}),\,p\rangle g_i'(\bar{x}) + p_i e^i,\; i \in J_1 \\ p_i g_i'(\bar{x}) + \langle g_i'(\bar{x}),\,p\rangle e^i,\; i \in J_2 \\ g_i'(\bar{x}),\; i \in I_1 \\ e^i,\; i \in I_2 \end{array}\right\} \quad \text{are linearly independent in } \mathbf{R}^n. \tag{3.8}$$

In section 4, we shall discuss the relation between this condition and other regularity conditions for the NCP, as well as compare convergence results of our algorithm with convergence results of other locally superlinearly convergent equation-based methods for solving NCP.

Under our assumptions, semismoothness of $\bar{P}F'(\cdot)p$ follows readily from (3.1) and standard calculus of semismooth mappings [36, Theorem 5]. Moreover, under the additional assumption of Lipschitz-continuity of $g''(\cdot)$ on V, $\bar{P}F'(\cdot)p$ is strongly semismooth, which follows from results on the superposition of strongly semismooth mappings [15, Theorem 19]. Hence, $\Phi(\cdot)$ is (strongly) semismooth.

Note that all the elements involved in the iteration scheme (2.7) are computed in this section by explicit formulas. In principle, computing H via (3.5)–(3.7) involves second derivatives of g. However, as already noted above, the terms containing second derivatives of g tend to zero as $x \to \bar{x}$. This suggests the idea of modifying the process by omitting these terms, which leads to the method represented by (2.13). We shall also take into account the structure of $\bar{P}$ and make use of Remark 2.1.

Note that for $\bar{P}$ given by (3.3) we have that $(E - \bar{P})$ is the orthogonal projector onto span $\{e^i\,,\; i \in I_1 \cup I_2\}$. According to (2.11), we can therefore redefine

$$\Phi_i(x) = \begin{cases} \langle F_i'(x),\,p\rangle & \text{if } i \in I_0, \\ F_i(x) & \text{if } i \in I_1 \cup I_2, \end{cases} \qquad x \in V. \tag{3.9}$$

Taking into account (2.12) and omitting further the terms that vanish at $\bar{x}$, we can take

(3.10)
$$\tilde{H}_i(x) = \begin{cases} 2(p_i g_i'(x) + \langle g_i'(x), p\rangle e^i - \gamma_i(x, p)(g_i'(x) + e^i)) & \text{if } i \in I_0, \\ F_i'(x) & \text{if } i \in I_1 \cup I_2, \end{cases} \qquad x \in V.$$

Comparing expressions (3.9) and (3.10) with (3.4) and (3.5), one can easily observe that the former are simpler and require fewer computations.

Furthermore, under our smoothness assumptions, it is easy to see that

$$\|\tilde{H}(x) - H(x)\| = O(\|x - \bar{x}\|),$$

and so the modified Newton method given by (2.13) is applicable.

We next give a formal statement of the convergence result for our methods applied to NCP, which is a corollary of Theorem 2.2.

THEOREM 3.1. *Let $g : V \to \mathbf{R}^n$ be a twice continuously differentiable mapping on V, V being a neighborhood of a solution $\bar{x}$ of the NCP* (1.2). *Assume that for some $p \in \mathbf{R}^n$ condition* (3.8) *is satisfied for every pair of index sets $(J_1, J_2) \in \mathcal{J}$.*

Then the iterates given by (2.7) *or* (2.13) *(with all the objects as defined in this section) converge to $\bar{x}$ locally superlinearly. If, in addition, the second derivative of g is Lipschitz-continuous on V, then the rate of convergence is quadratic.*

We next show how to construct $\bar{P}$ without knowing the solution $\bar{x}$. Given the structure of $\bar{P}$ (see (3.3)), it is clear that this task reduces to correct identification of the degenerate set I_0. This can be done with the help of *error bounds*, as described next (our approach is in the spirit of the technique developed in [10] for identification of active constraints in nonlinear programming). To our knowledge, the weakest condition under which a local error bound for NCP is currently available is the 2-regularity of F given by (1.3) at the NCP solution $\bar{x}$ [25]. Specifically, if F is 2-regular at $\bar{x}$, then there exist a neighborhood U of $\bar{x}$ in $\mathbf{R}^n$ and a constant $M_1 > 0$ such that

$$\|x - \bar{x}\| \le M_1 \|F(x)\|^{1/2} \quad \forall\, x \in U. \tag{3.11}$$

We shall not introduce the notion of 2-regularity formally here, as this would require an extensive discussion. We emphasize only that the bound (3.11) may hold when the so-called natural residual $\min\{x, g(x)\}$ does not provide an error bound, and always holds when it does (see [25], and in particular [25, Example 1]). Hence, the 2-regularity of F is a weaker assumption than the R_0-type property or semistability, which in the case of NCP are both equivalent to an error bound in terms of the natural residual [38]. And it is further weaker than b-regularity; see [25].

We note that in Lemma 3.2 below we could also use other error bounds for identifying I_0. However, they would require either stronger local assumptions or global assumptions.

LEMMA 3.2. *Suppose that $\bar{x}$ is a solution of NCP, g is Lipschitz-continuous on V, where V is a neighborhood of $\bar{x}$. Suppose finally that the local error bound* (3.11) *holds. Then for any $\alpha \in (0, 1)$ there exists a neighborhood U of $\bar{x}$ such that*

$$\{i \in \{1, \ldots, n\} \mid |\max\{g_i(x), x_i\}| \le \|F(x)\|^{\alpha/2}\} = I_0 \quad \forall x \in U. \tag{3.12}$$

Proof. It is easy to observe that there exist some $M_2 > 0$ and some neighborhood U of $\bar{x}$ such that

$$\text{for } i \in I_0,\ |\max\{g_i(x), x_i\}| = |\max\{g_i(x), x_i\} - \max\{g_i(\bar{x}), \bar{x}_i\}| \le M_2 \|x - \bar{x}\| \quad \forall x \in U,$$

where the inequality follows from the Lipschitz-continuity of the functions involved.

Therefore, by (3.11) (possibly adjusting the neighborhood U), for an arbitrary fixed $\alpha \in (0, 1)$ we have that

$$\text{for } i \in I_0, \ |\max\{g_i(x), x_i\}| \le M_1 M_2 \|F(x)\|^{1/2} \le \|F(x)\|^{\alpha/2} \quad \forall\, x \in U.$$

In particular, the quantity in the left-hand side of the inequality above tends to zero as x tends to $\bar{x}$. On the other hand, it is clear that there exists $\varepsilon > 0$ such that

$$\text{for } i \in I_1 \cup I_2, \ |\max\{g_i(x), x_i\}| \ge \varepsilon > 0 \quad \forall\, x \in U$$

(U should be adjusted again, if necessary). Combining those facts, we obtain (3.12) for U sufficiently small. □

By Lemma 3.2, the index set I_0, and hence the mapping $\bar{P}$, are correctly identified by (3.12), provided one has a point close enough to the solution. We note that this requirement of closeness to solution is completely consistent with the setting of the paper, since the subjects under consideration are superlinearly convergent Newton-like methods, which are local by nature.

Finally, we mention other considerations that can also be useful for identifying I_0. Sometimes the cardinality r of I_0 may be known from a priori analysis of the problem, or one can be interested in finding an NCP solution with a given cardinality of I_0. Then for any $x \in \mathbf{R}^n$ sufficiently close to $\bar{x}$, the set I_0 coincides with the set of indices corresponding to the r smallest values of $|\max\{g_i(x), x_i\}|$. In this case, no error bound is needed to identify I_0. We note that, in the present setting, cardinality of I_0 is closely related to corank of singularity. In the literature on numerical methods for solving singular equations, the assumption that corank of singularity is known is absolutely standard [20, 19, 43, 14, 1, 2]. In the complementarity literature, on the other hand, assumptions about cardinality of I_0 are not common, except possibly for $I_0 = \emptyset$.

4. Regularity conditions. In this section we compare our approach with other Newton-type methods that solve one linear system at each iteration. The weakest condition under which there exists a locally superlinearly convergent Newton-type algorithm for solving a (nonsmooth) equation reformulation of the NCP is the *b*-regularity assumption, which can be stated as follows: for every pair of index sets $(J_1, J_2) \in \mathcal{J}$, it holds that

$$\left.\begin{array}{ll} g_i'(\bar{x}), & i \in J_1 \cup I_1 \\ e^i, & i \in J_2 \cup I_2 \end{array}\right\} \quad \text{are linearly independent in } \mathbf{R}^n.$$

Under this assumption, the natural residual mapping $x \to \min\{x, g(x)\} : \mathbf{R}^n \to \mathbf{R}^n$, is *BD*-regular at $\bar{x}$. Furthermore, it is also (strongly) semismooth under standard assumptions on g. Hence, the nonsmooth Newton method (2.7) based on it converges locally superlinearly [30, 34]. Note that Newton methods applied to another popular reformulation based on the Fischer–Burmeister function [17, 11] require for convergence the stronger *R-regularity* [44] assumption; see [34].

In what follows, we compare our regularity condition (3.8) with *b*-regularity and show that they are essentially different. In general, neither is weaker or stronger than the other. This implies that our approach based on the smooth NCP reformulation is a complement to nonsmooth reformulations, and vice versa, as each approach can be successful in situations when the other is not.

The next result is important to obtaining an insight into the nature of our regularity condition (3.8). We start with the following definition.

DEFINITION 4.1. *A solution $\bar{x}$ of the NCP* (1.2) *is referred to as* quasi-regular *if for every pair of index sets $(J_1, J_2) \in \mathcal{J}$ there exists an element $p = p(J_1, J_2) \in \mathbf{R}^n$ such that* (3.8) *is satisfied.*

PROPOSITION 4.2. *Suppose that the solution $\bar{x}$ of the NCP* (1.2) *is quasi-regular. Then there exists a universal $p \in \mathbf{R}^n$ which satisfies* (3.8) *for every pair $(J_1, J_2) \in \mathcal{J}$. Moreover, the set of such p is open and dense in $\mathbf{R}^n$.*

Proof. Fix a pair $(J_1, J_2) \in \mathcal{J}$, and consider the determinant of the system of vectors in (3.8) as a function of p. This function is a polynomial on $\mathbf{R}^n$, and this polynomial is not everywhere zero, since it is not zero at $p(J_1, J_2)$ (see Definition 4.1). But then the set where the polynomial is not zero is obviously open and dense in $\mathbf{R}^n$. Moreover, the intersection of such sets corresponding to pairs $(J_1, J_2) \in \mathcal{J}$ is also open and dense, since it is a finite intersection of open and dense sets. □

It follows that if $\bar{x}$ is a quasi-regular solution of NCP in the sense of Definition 4.1, then even picking a random $p \in \mathbf{R}^n$, one is extremely unlikely to pick a "wrong" p (as the set of wrong elements is "thin"). Hence, under the assumption of quasiregularity, for the implementation of the algorithm described in section 3 we can choose $p \in \mathbf{R}^n \setminus \{0\}$ arbitrarily, with the understanding that almost every $p \in \mathbf{R}^n$ is appropriate. In particular, for all practical purposes, we can think of quasiregularity as the regularity condition needed for superlinear convergence of our algorithm. We next investigate the relationship between quasiregularity and b-regularity.

First, we show that if the cardinality of I_0 is equal to one, then quasiregularity is in fact weaker than b-regularity.

PROPOSITION 4.3. *Suppose that $\bar{x}$ is a b-regular solution of the NCP, and the cardinality of I_0 is equal to one. Then $\bar{x}$ is quasi-regular.*

Proof. Let $I_0 = \{i_0\}$ and denote $L = \operatorname{span}\{g_i'(\bar{x}),\ i \in I_1,\ e^i,\ i \in I_2\}$. In this setting, b-regularity clearly means that

$$g_{i_0}'(\bar{x}) \notin L, \quad e^{i_0} \notin L,$$

corresponding to the two possible choices of $(J_1, J_2) \in \mathcal{J}$. It follows that

$$\forall\, q \in L^\perp \setminus \{0\}, \quad \langle g_{i_0}'(\bar{x}),\, q\rangle \neq 0, \quad q_{i_0} \neq 0. \tag{4.1}$$

Assume for a contradiction that $\bar{x}$ is not quasi-regular. Then by Definition 4.1, there exists a pair $(J_1, J_2) \in \mathcal{J}$ such that for every $p \in \mathbf{R}^n$ condition (3.8) is violated. This means that either

$$\langle g_{i_0}'(\bar{x}),\, p\rangle g_{i_0}'(\bar{x}) + p_{i_0} e^{i_0} \in L$$

or

$$p_{i_0} g_{i_0}'(\bar{x}) + \langle g_{i_0}'(\bar{x}),\, p\rangle e^{i_0} \in L.$$

Taking any $q \in L^\perp \setminus \{0\}$, we deduce that for every $p \in \mathbf{R}^n$ either

$$\langle g_{i_0}'(\bar{x}),\, q\rangle \langle g_{i_0}'(\bar{x}),\, p\rangle + q_{i_0} p_{i_0} = 0$$

or

$$\langle g_{i_0}'(\bar{x}),\, q\rangle p_{i_0} + q_{i_0} \langle g_{i_0}'(\bar{x}),\, p\rangle = 0.$$

Setting $p = q$, we then obtain that either

$$\langle g_{i_0}'(\bar{x}),\, q\rangle^2 + q_{i_0}^2 = 0$$

or

$$\langle g'_{i_0}(\bar{x}),\, q\rangle q_{i_0} = 0,$$

which contradicts b-regularity, because of (4.1). □

It is easy to see that in the setting of Proposition 4.3, the quasiregularity condition can be satisfied without b-regularity. For example, let $g'_{i_0}(\bar{x}) \in L$, but $e^{i_0} \notin L$. Then b-regularity is violated. On the other hand, quasiregularity here is equivalent to saying that there exist elements $p^1,\, p^2 \in \mathbf{R}^n$ (corresponding to the two possible choices of $(J_1,\, J_2) \in \mathcal{J}$) such that

$$p^1_{i_0} \neq 0, \quad \langle g'_{i_0}(\bar{x}),\, p^2\rangle \neq 0,$$

which is satisfied for almost any p^1 and p^2, provided $g'_{i_0}(\bar{x}) \neq 0$. It is also quite clear that just choosing p^1 and p^2 randomly should do the job.

In general, i.e., in the cases of higher cardinality of I_0, b-regularity and quasiregularity become different, not directly related conditions. In particular, neither is stronger or weaker than the other, as illustrated by the following examples.

Example 4.1. Let $n = 2$, $I_0 = \{1,\, 2\}$, and

$$g'_1(\bar{x}) = (1,\, \sqrt{2}), \quad g'_2(\bar{x}) = (\sqrt{2},\, 1).$$

Then b-regularity is obvious, but

$$\langle g'_i(\bar{x}),\, p\rangle g'_i(\bar{x}) + p_i e^i = (2p_1 + \sqrt{2}p_2,\, \sqrt{2}p_1 + 2p_2) \quad \forall\, i = 1,\, 2, \quad \forall\, p \in \mathbf{R}^2.$$

This means that for $J_1 = I_0$, $J_2 = \emptyset$, (3.8) does not hold for any p, and so the quasiregularity condition is not satisfied.

Example 4.2. Let $n = 2$, $I_0 = \{1,\, 2\}$, and

$$g'_1(\bar{x}) = e^2, \quad g'_2(\bar{x}) \notin \operatorname{span}\{e^1\}.$$

Then b-regularity does not hold (the linear independence condition is violated for $J_1 = \{1\}$, $J_2 = \{2\}$), but quasiregularity is satisfied, which can be shown by straightforward computations. We omit the details, as they do not provide any further insight.

We complete our discussion with a sufficient condition for quasiregularity of $\bar{x}$, which is meaningful when the cardinality of I_0 is not greater than $n/2$, half dimensionality of the space. Specifically, suppose that

$$g'_i(\bar{x}),\, e^i,\, i \in I_0, \text{ are linearly independent in } \mathbf{R}^n \tag{4.2}$$

and

$$\exists\, (\bar{J}_1,\, \bar{J}_2) \in \mathcal{J} \text{ s.t. } \left.\begin{array}{ll} g'_i(\bar{x}), & i \in \bar{J}_1 \cup I_1 \\ e^i, & i \in \bar{J}_2 \cup I_2 \end{array}\right\} \text{ are linearly independent in } \mathbf{R}^n. \tag{4.3}$$

It is clear that (4.3) is subsumed by b-regularity (where it must hold for *all* partitions of I_0). It is also not difficult to see that (4.3) is necessary for quasiregularity of $\bar{x}$. Hence, this assumption does not introduce any additional restrictions with respect to the regularity conditions under consideration. Furthermore, for nonpathological problems the cardinality of I_0 should not be too large compared to the dimensionality of the space, and so condition (4.2) should not be difficult to satisfy. Therefore, (4.2) and (4.3) appear to be not restrictive.

PROPOSITION 4.4. *Suppose that* (4.2) *and* (4.3) *hold. Then* $\bar{x}$ *is a quasi-regular solution of NCP.*

Proof. Take any pair of index sets $(J_1, J_2) \in \mathcal{J}$, and consider the system of (twice the cardinality of I_0) linear equations

$$\begin{cases} \langle g_i'(\bar{x}), p\rangle = 1, & p_i = 0, \quad i \in J_1 \cap \bar{J}_1, \\ \langle g_i'(\bar{x}), p\rangle = 0, & p_i = 1, \quad i \in J_1 \cap \bar{J}_2, \\ \langle g_i'(\bar{x}), p\rangle = 0, & p_i = 1, \quad i \in J_2 \cap \bar{J}_1, \\ \langle g_i'(\bar{x}), p\rangle = 1, & p_i = 0, \quad i \in J_2 \cap \bar{J}_2 \end{cases}$$

in the variable $p \in \mathbf{R}^n$. Under the assumption (4.2), this system has a solution $p = p(J_1, J_2)$. Observe further that substituting this p into (3.8) reduces the system of vectors appearing in (3.8) precisely to the system of vectors appearing in (4.3), which is linearly independent by the hypothesis. □

Again, it is easy to see that the latter sufficient condition for quasiregularity of $\bar{x}$ can hold without b-regularity. On the other hand, in general it is not implied by b-regularity. In particular, b-regularity need not imply (4.2).

Summarizing the preceding discussion, we conclude that the regularity assumption required for the algorithm proposed in section 3 for solving the NCP is different from b-regularity, which is the typical assumption in the context of nonsmooth Newton-type methods for solving nonsmooth NCP reformulations. In fact, the two assumptions are of a rather distinct nature. This is not surprising, considering that they result from approaches which are also quite different.

5. Some further applications. The general approach presented in section 2 can also be useful in other problems where complementarity is present. Below we outline applications to another class of smooth reformulations of NCP (different from (1.3)) and to the mixed complementarity problems. We limit this discussion to exhibiting the structure of singularity associated with the smooth equation reformulations of those problems. Deriving the resulting regularity conditions and comparing them to known ones requires too much space. Without going into detail, we claim that regularity assumptions needed for our approach would again be different from assumptions of Newton methods for nonsmooth equations.

5.1. Other NCP reformulations. The analysis presented in sections 3 and 4 for NCP is intrinsic in the sense that it is also applicable to smooth reformulations other than the one given by (1.3). Indeed, following [35], consider the family of functions

$$F : \mathbf{R}^n \to \mathbf{R}^n, \quad F_i(x) = \theta(g_i(x)) + \theta(x_i) - \theta(|g_i(x) - x_i|), \qquad i = 1, \ldots, n,$$

where $\theta : \mathbf{R} \to \mathbf{R}$ is any strictly increasing function such that $\theta(0) = 0$. It can be checked that the NCP solution set coincides with zeros of F. As an aside, note that reformulation (1.3) cannot be written in the form stated above, so the two are really different.

Suppose further that θ is differentiable on $\mathbf{R}$ with $\theta'(0) = 0$ and $\theta'(t) > 0$ for $t > 0$. For example, we could take

$$\theta(t) = t|t|.$$

Let $\bar{x}$ be some solution of NCP, and V be its neighborhood. If g is twice continuously differentiable on V and θ' is Lipschitz-continuous, then the derivative of F is Lipschitz-

continuous near $\bar{x}$, and it is given by

$$F_i'(x) = \theta'(g_i(x))g_i'(x) + \theta'(x_i)e^i - \operatorname{sign}(g_i(x) - x_i)\theta'(|g_i(x) - x_i|)(g_i'(x) - e^i),$$
$$i = 1, \ldots, n, \quad x \in V.$$

As is easy to see,

$$F_i'(\bar{x}) = \begin{cases} 0 & \text{if } i \in I_0, \\ \theta'(\bar{x}_i)g_i'(\bar{x}) & \text{if } i \in I_1, \\ \theta'(g_i(\bar{x}))e^i & \text{if } i \in I_2. \end{cases}$$

Since $\theta'(t) > 0$ for any $t > 0$, we conclude that the structure of singularity here is absolutely identical to that for F given by (1.3) (recall (1.4)). In particular,

$$\operatorname{im} F'(\bar{x}) \subset \operatorname{span}\{e^i,\ i \in I_1 \cup I_2\},$$

and all the objects and the analysis in sections 3 and 4 can be derived in a similar fashion.

5.2. Mixed complementarity problems. The mixed complementarity problem (MCP) is a variational inequality on a (generalized) box $B = \{x \in \mathbf{R}^n \mid l \le x \le u\}$, where $l_i \in [-\infty, +\infty)$ and $u_i \in (-\infty, +\infty]$ are such that $l_i < u_i$, $i = 1, \ldots, n$. Specifically, the problem is to find

$$x \in B \text{ such that } \langle g(x), y - x\rangle \ge 0 \qquad \forall y \in B,$$

where $g : \mathbf{R}^n \to \mathbf{R}^n$. It can be seen that this is equivalent to $x \in \mathbf{R}^n$ satisfying the following conditions: for every $i = 1, \ldots, n$,

$$\begin{array}{l} \text{if } g_i(x) > 0, \text{ then } x_i = l_i; \\ \text{if } g_i(x) < 0, \text{ then } x_i = u_i; \\ \text{if } g_i(x) = 0, \text{ then } l_i \le x_i \le u_i. \end{array}$$

NCP is a special case of MCP corresponding to $l_i = 0$, $u_i = +\infty$, $i = 1, \ldots, n$. Define

$$\psi : \mathbf{R}^2 \to \mathbf{R}, \quad \psi(a, b) = 2ab - (\min\{0, a + b\})^2.$$

We claim that solutions of MCP coincide with zeros of the function $F : \mathbf{R}^n \to \mathbf{R}^n$ whose components are given by

$$F_i(x) = \begin{cases} \psi(g_i(x), x_i - l_i), & i \in I_l := \{i \mid l_i > -\infty, u_i = +\infty\}, \\ \psi(-g_i(x), u_i - x_i), & i \in I_u := \{i \mid l_i = -\infty, u_i < +\infty\}, \\ g_i(x), & i \in I_g := \{i \mid l_i = -\infty, u_i = +\infty\}, \\ \psi(-\psi(-g_i(x), u_i - x_i), x_i - l_i), & i \in I_{lu} := \{i \mid l_i > -\infty, u_i < +\infty\}. \end{cases}$$

We omit the proof, which can be carried out by direct verification. Let $\bar{x}$ be some solution of MCP, and V be its neighborhood. If g is twice continuously differentiable on V, then the derivative of F is Lipschitz-continuous near $\bar{x}$. Defining

$$I_0 := \{i = 1, \ldots, n \mid g_i(\bar{x}) = 0\} \cap \{i = 1, \ldots, n \mid \bar{x}_i = l_i \text{ or } \bar{x}_i = u_i\},$$
$$I_1 := \{i = 1, \ldots, n \mid g_i(\bar{x}) = 0\} \setminus I_0,$$
$$I_2 := \{i = 1, \ldots, n \mid g_i(\bar{x}) \ne 0\},$$

it can be verified that

$$F_i'(\bar{x}) = \begin{cases} 0 & \text{if } i \in I_0, \\ \rho_i g_i'(\bar{x}) & \text{if } i \in I_1, \\ \nu_i e^i & \text{if } i \in I_2, \end{cases}$$

where

$$\rho_i = \begin{cases} 2(\bar{x}_i - l_i), & i \in \{i \in I_l \mid g_i(\bar{x}) = 0,\ \bar{x}_i > l_i\}, \\ -2(u_i - \bar{x}_i), & i \in \{i \in I_u \mid g_i(\bar{x}) = 0,\ \bar{x}_i < u_i\}, \\ 1, & i \in I_g, \\ 4(\bar{x}_i - l_i)(u_i - \bar{x}_i), & i \in \{i \in I_{lu} \mid g_i(\bar{x}) = 0,\ l_i < \bar{x}_i < u_i\}, \end{cases}$$

$$\nu_i = \begin{cases} g_i(\bar{x}), & i \in \{i \in I_l \cup I_u \mid g_i(\bar{x}) \neq 0\}, \\ -4g_i(\bar{x})(u_i - l_i), & i \in \{i \in I_{lu} \mid g_i(\bar{x}) < 0,\ \bar{x}_i = u_i\}, \\ 4g_i(\bar{x})(u_i - l_i) + 2(\min\{0,\ u_i - l_i - g_i(\bar{x})\})^2, & i \in \{i \in I_{lu} \mid g_i(\bar{x}) > 0,\ \bar{x}_i = l_i\}. \end{cases}$$

In particular, $\rho_i \neq 0$, $i \in I_1$, and $\nu_i \neq 0$, $i \in I_2$. Observing the structure of $F'(\bar{x})$, further analysis can now follow the ideas of sections 3 and 4.

6. Concluding remarks. We have presented a new approach to solving smooth singular equations. Unlike previously available algorithms, our method is applicable when the equation mapping is not necessarily twice differentiable. Important examples of once differentiable singular equations are reformulations of the NCPs, which we have studied in detail. In particular, we have demonstrated that in the case of NCP our method takes a readily implementable simple form. Furthermore, the structure of singularity can be completely identified by means of local error bound analysis, without knowing the solution itself. It was further shown that the regularity condition required for the superlinear convergence of the presented algorithm is different from conditions needed for the nonsmooth Newton methods applied to nonsmooth NCP reformulations. Thus the two approaches should be regarded as complementing each other. Finally, it was demonstrated that the main ideas of this paper should also be applicable to other problems where complementarity structures are present.

REFERENCES

[1] E.L. Allgower and K. Böhmer, *Resolving singular nonlinear equations,* Rocky Mountain J. Math., 18 (1988), pp. 225–268.

[2] E.L. Allgower, K. Böhmer, A. Hoy, and V. Janovský, *Direct methods for solving singular nonlinear equations,* ZAMM Z. Angew. Math. Mech., 79 (1999), pp. 219–231.

[3] A.V. Arutyunov, *Optimality Conditions: Abnormal and Degenerate Problems*, Kluwer Academic Publishers, Dordrecht, the Netherlands, 2000.

[4] O.A. Brezhneva, A.F. Izmailov, A.A. Tretyakov, and A. Khmura, *An approach to finding singular solutions to a general system of nonlinear equations,* Comput. Math. Math. Phys., 40 (2000), pp. 347–358.

[5] F.H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley and Sons, New York, 1983.

[6] D.W. Decker and C.T. Kelley, *Newton's method at singular points.* I, SIAM J. Numer. Anal., 17 (1980), pp. 66–70.

[7] D.W. Decker and C.T. Kelley, *Newton's method at singular points.* II, SIAM J. Numer. Anal., 17 (1980), pp. 465–471.

[8] D.W. Decker and C.T. Kelley, *Convergence acceleration for Newton's method at singular points,* SIAM J. Numer. Anal., 19 (1982), pp. 219–229.

[9] D.W. Decker, H.B. Keller, and C.T. Kelley, *Convergence rates for Newton's method at singular points,* SIAM J. Numer. Anal., 20 (1983), pp. 296–314.

[10] F. Facchinei, A. Fischer, and C. Kanzow, *On the accurate identification of active constraints,* SIAM J. Optim., 9 (1999), pp. 14–32.
[11] F. Facchinei and J. Soares, *A new merit function for nonlinear complementarity problems and a related algorithm,* SIAM J. Optim., 7 (1997), pp. 225–247.
[12] M.C. Ferris and J.-S. Pang, eds., *Complementarity and Variational Problems: State of the Art*, SIAM, Philadelphia, 1997.
[13] M.C. Ferris and C. Kanzow, *Complementarity and related problems*, in Handbook of Applied Optimization, P.M. Pardalos and M.G.C. Resende, eds., Oxford University Press, New York, 2002, pp. 514–530.
[14] J.P. Fink and W.C. Rheinboldt, *A geometric framework for the numerical study of singular points,* SIAM J. Numer. Anal., 24 (1987), pp. 618–633.
[15] A. Fischer, *Solution of monotone complementarity problems with locally Lipschitzian functions,* Math. Program., 76 (1997), pp. 513–532.
[16] M. Fukushima and L. Qi, eds., *Reformulation—Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods,* Kluwer Academic Publishers, Dordrecht, the Netherlands, 1999.
[17] C. Geiger and C. Kanzow, *On the resolution of monotone complementarity problems,* Comput. Optim. Appl., 5 (1996), pp. 155–173.
[18] A.O. Griewank, *Starlike domains of convergence for Newton's method at singularities,* Numer. Math., 35 (1980), pp. 95–111.
[19] A.O. Griewank, *On solving nonlinear equations with simple singularities or nearly singular solutions,* SIAM Rev., 27 (1985), pp. 537–563.
[20] A. Griewank and G.W. Reddien, *Characterization and computation of generalized turning points,* SIAM J. Numer. Anal., 21 (1984), pp. 176–185.
[21] M. Hermann, P. Kunkel, and W. Middelman, *Augmented systems for computation of singular points in Banach space problems,* ZAMM Z. Angew. Math. Mech., 78 (1998), pp. 39–50.
[22] A.F. Izmailov, *Stable methods for finding* 2*-regular solutions of nonlinear operator equations,* Comput. Math. Math. Phys., 36 (1996), pp. 1183–1192.
[23] A.F. Izmailov, *On certain generalizations of Morse's lemma,* Proc. Steklov Inst. Math., 220 (1998), pp. 138–153.
[24] 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., to appear.
[25] A.F. Izmailov and M.V. Solodov, *Error bounds for* 2*-regular mappings with Lipschitzian derivatives and their applications,* Math. Program., 89 (2001), pp. 413–435.
[26] A.F. Izmailov and A.A. Tretyakov, *Local regularization of certain classes of nonlinear operator equations,* Comput. Math. Math. Phys., 36 (1996), pp. 835–846.
[27] A.F. Izmailov and A.A. Tretyakov, 2*-Regular Solutions of Nonlinear Problems. Theory and Numerical Methods*, Fizmatlit, Moscow, 1999, (in Russian).
[28] H. Jiang, L. Qi, X. Chen, and D. Sun, *Semismoothness and superlinear convergence in nonsmooth optimization and nonsmooth equations*, in Nonlinear Optimization and Applications, G. Di Pillo and F. Giannessi, eds., Plenum Press, New York, 1996, pp. 197–212.
[29] C. Kanzow, *Some equation-based methods for the nonlinear complementarity problem,* Optim. Methods Soft., 3 (1994), pp. 327–340.
[30] C. Kanzow and M. Fukushima, *Solving box constrained variational inequality problems by using the natural residual with D-gap function globalization,* Oper. Res. Lett., 23 (1998), pp. 45–51.
[31] C. Kanzow and H. Kleinmichel, *A new class of semismooth Newton-type methods for nonlinear complementarity problems,* Comput. Optim. Appl., 11 (1998), pp. 227–251.
[32] B. Kummer, *Newton's method for nondifferentiable functions,* in Advances in Mathematical Optimization, J. Guddat, B. Bank, H. Hollatz, P. Kall, D. Klatte, B. Kummer, K. Lommatzsch, K. Tammer, M. Vlach, and K. Zimmermann, eds., Math. Res. 45, Akademie-Verlag, Berlin, 1988, pp. 114–125.
[33] B. Kummer, *Newton's method based on generalized derivatives for nonsmooth functions,* in Advances in Optimization, W. Oettli and D. Pallaschke, eds., Springer-Verlag, Berlin, 1992, pp. 171–194.
[34] T. De Luca, F. Facchinei, and C. Kanzow, *A theoretical and numerical comparison of some semismooth algorithms for complementarity problems,* Comput. Optim. Appl., 16 (2000), pp. 173–205.
[35] O.L. Mangasarian, *Equivalence of the complementarity problem to a system of nonlinear equations,* SIAM J. Appl. Math., 31 (1976), pp. 89–92.
[36] R. Mifflin, *Semismooth and semiconvex functions in constrained optimization,* SIAM J. Control Optim., 15 (1977), pp. 959–972.

[37] J.-S. Pang, *Complementarity problems*, in Handbook of Global Optimization, R. Horst and P. Pardalos, eds., Kluwer Academic Publishers, Boston, MA, 1995, pp. 271–338.
[38] J.-S. Pang, *private communication*, January 2001.
[39] J.-S. Pang and S.A. Gabriel, *NE/SQP: A robust algorithm for the nonlinear complementarity problem,* Math. Program., 60 (1993), pp. 295–337.
[40] J.-S. Pang and L. Qi, *Nonsmooth equations: Motivation and algorithms, SIAM J. Optim.*, 3 (1993), pp. 443–465.
[41] L. Qi, *Convergence analysis of some algorithms for solving nonsmooth equations,* Math. Oper. Res., 18 (1993), pp. 227–244.
[42] L. Qi and J. Sun, *A nonsmooth version of Newton's method,* Math. Program., 58 (1993), pp. 353–367.
[43] P.J. Rabier and G.W. Reddien, *Characterization and computation of singular points with maximum rank deficiency,* SIAM J. Numer. Anal., 23 (1986), pp. 1040–1051.
[44] S.M. Robinson, *Strongly regular generalized equations,* Math. Oper. Res., 5 (1980), pp. 43–62.
[45] S.M. Robinson, *Local structure of feasible sets in nonlinear programming. Part* III: *Stability and sensitivity,* Math. Programming Stud., 30 (1987), pp. 45–66.
[46] D. Sun and L. Qi, *On NCP-functions,* Comput. Optim. Appl., 13 (1999), pp. 201–220.
[47] P. Tseng, *Growth behavior of a class of merit functions for the nonlinear complementarity problem,* J. Optim. Theory Appl., 89 (1996), pp. 17–37.

SIAM J. OPTIM.
Vol. 13, No. 2, pp. 406–431

SECOND-ORDER NECESSARY AND SUFFICIENT OPTIMALITY CONDITIONS FOR OPTIMIZATION PROBLEMS AND APPLICATIONS TO CONTROL THEORY*

EDUARDO CASAS† AND FREDI TRÖLTZSCH‡

Abstract. This paper deals with a class of nonlinear optimization problems in a function space, where the solution is restricted by pointwise upper and lower bounds and by finitely many equality and inequality constraints of functional type. Second-order necessary and sufficient optimality conditions are established, where the cone of critical directions is arbitrarily close to the form which is expected from the optimization in finite dimensional spaces. The results are applied to some optimal control problems for ordinary and partial differential equations.

Key words. necessary and sufficient optimality conditions, control of differential equations, state constraints

AMS subject classifications. 49K20, 35J25, 90C45, 90C48

PII. S1052623400367698

1. Introduction. Let $(X, \mathcal{S}, \mu)$ be a measure space with $\mu(X) < +\infty$. In this paper we will study the following optimization problem:

$$\text{(P)} \qquad \begin{cases} \text{minimize } J(u), \\ u_a(x) \le u(x) \le u_b(x) & \text{a.e. } x \in X, \\ G_j(u) = 0, & 1 \le j \le m_1, \\ G_j(u) \le 0, & m_1 + 1 \le j \le m, \end{cases}$$

where $u_a, u_b \in L^\infty(X)$ and $J, G_j : L^\infty(X) \longrightarrow \mathbb{R}$ are given functions with differentiability properties to be fixed later. We will state necessary and sufficient optimality conditions for a local minimum of (P). Our main goal is to reduce the classical gap between the necessary and sufficient conditions for optimization problems in Banach spaces. We shall prove some optimality conditions very close to the ones for finite dimensional optimization problems. In the case of finite dimensions, strongly active inequality constraints (i.e., with strictly positive Lagrange multipliers) are considered in the critical cone by associated linearized equality constraints. Roughly speaking, this is what we are able to extend to infinite dimensions. Due to the lack of compactness, the classical proof of the sufficiency theorem known for finite dimensions cannot be transferred to the case of general Banach spaces. Our direct method of proof is able to overcome this difficulty. To our best knowledge, this result has not yet been presented in the literature. Of course, the bound constraints $u_a(x) \le u(x) \le u_b(x)$ introduce some additional difficulties in the study because they constitute an infinite number of constraints. In section 2 we introduce a slightly stronger regularity assumption than that considered in the Kuhn–Tucker theorem, which allows us to deal with the bound constraints.

*Received by the editors February 8, 2000; accepted for publication (in revised form) February 7, 2002; published electronically September 24, 2002.

http://www.siam.org/journals/siopt/13-2/36769.html

†Departamento de Matemática Aplicada y Ciencias de la Computación, E.T.S.I. Industriales y de Telecomunicación, Universidad de Cantabria, 39005 Santander, Spain (eduardo.casas@unican.es). This author was partially supported by Dirección General de Enseñanza Superior e Investigación Científica (Spain).

‡Fakultät für Mathematik, Technische Universität Chemnitz, D-09107 Chemnitz, Germany (f.troeltzsch@mathematik.tu-chemnitz.de).

In section 4 we discuss the application of our general results to different types of optimal control problems. We consider the control of ODEs as well as that of partial differential equations of elliptic and parabolic type.

2. Necessary optimality conditions. In this section we will assume that $\bar{u}$ is a local solution of (P), which means that there exists a real number $r > 0$ such that for every feasible point of (P), with $\|u - \bar{u}\|_{L^\infty(X)} < r$, we have that $J(\bar{u}) \le J(u)$.

For every $\varepsilon > 0$, we denote the set of points at which the bound constraints are *ε-inactive* by

$$X_\varepsilon = \{x \in X : u_a(x) + \varepsilon \le \bar{u}(x) \le u_b(x) - \varepsilon\}.$$

We make the following regularity assumption:

$$\begin{cases} \exists \varepsilon_{\bar{u}} > 0 \text{ and } \{h_j\}_{j\in I_0} \subset L^\infty(X), \quad \text{with supp } h_j \subset X_{\varepsilon_{\bar{u}}}, \\ \text{such that } G_i'(\bar{u})h_j = \delta_{ij}, \quad i,j \in I_0, \end{cases} \tag{2.1}$$

where

$$I_0 = \{j \le m | G_j(\bar{u}) = 0\}.$$

I_0 is the set of indices corresponding to active constraints. We also denote the set of nonactive constraints by I_-

$$I_- = \{j \le m | G_j(\bar{u}) < 0\}.$$

Obviously (2.1) is equivalent to the independence of the derivatives $\{G_j'(\bar{u})\}_{j\in I_0}$ in $L^\infty(X_{\varepsilon_{\bar{u}}})$. Under this assumption we can derive the first-order necessary conditions for optimality satisfied by $\bar{u}$. For the proof, the reader is referred to Bonnans and Casas [3] or Clarke [10].

THEOREM 2.1. *Let us assume that* (2.1) *holds and that J and $\{G_j\}_{j=1}^m$ are of class C^1 in a neighborhood of $\bar{u}$. Then there exist real numbers $\{\bar{\lambda}_j\}_{j=1}^m \subset \mathbb{R}$ such that*

$$\bar{\lambda}_j \ge 0, \quad m_1 + 1 \le j \le m, \qquad \bar{\lambda}_j = 0 \quad \text{if } j \in I_-, \tag{2.2}$$

$$\left\langle J'(\bar{u}) + \sum_{j=1}^m \bar{\lambda}_j G_j'(\bar{u}), u - \bar{u} \right\rangle \ge 0 \qquad \forall u_a \le u \le u_b. \tag{2.3}$$

Since we want to establish some optimality conditions useful for the study of control problems, we need to take into account the two-norm discrepancy; for this question, see, for instance, Ioffe [17] and Maurer [19]. Then we have to impose some additional assumptions on the functions J and G_j.

(A1) There exist functions $f, g_j \in L^2(X)$, $1 \le j \le m$, such that for every $h \in L^\infty(X)$

$$J'(\bar{u})h = \int_X f(x)h(x)d\mu(x) \quad \text{and} \quad G_j'(\bar{u})h = \int_X g_j(x)h(x)d\mu(x), \quad 1 \le j \le m. \tag{2.4}$$

(A2) If $\{h_k\}_{k=1}^\infty \subset L^\infty(X)$ is bounded, $h \in L^\infty(X)$, and $h_k(x) \to h(x)$ a.e. in X, then

$$\left[J''(\bar{u}) + \sum_{j=1}^m \bar{\lambda}_j G_j''(\bar{u})\right] h_k^2 \to \left[J''(\bar{u}) + \sum_{j=1}^m \bar{\lambda}_j G_j''(\bar{u})\right] h^2. \tag{2.5}$$

If we define

$$L(u,\lambda) = J(u) + \sum_{j=1}^{m} \lambda_j G_j(u) \quad \text{and} \quad d(x) = f(x) + \sum_{j=1}^{m} \bar{\lambda}_j g_j(x), \tag{2.6}$$

then

$$\frac{\partial L}{\partial u}(\bar{u},\bar{\lambda})h = \left[J'(\bar{u}) + \sum_{j=1}^{m} \bar{\lambda}_j G_j'(\bar{u}) \right] h = \int_X d(x)h(x)d\mu(x) \quad \forall h \in L^\infty(X). \tag{2.7}$$

From (2.3) we deduce that

$$d(x) = \begin{cases} 0 & \text{for almost every } x \in X, \text{ where } u_a(x) < \bar{u}(x) < u_b(x), \\ \geq 0 & \text{for almost every } x \in X, \text{ where } \bar{u}(x) = u_a(x), \\ \leq 0 & \text{for almost every } x \in X, \text{ where } \bar{u}(x) = u_b(x). \end{cases} \tag{2.8}$$

Associated with d, we set

$$X^0 = \{x \in X : |d(x)| > 0\}. \tag{2.9}$$

Given $\{\bar{\lambda}_j\}_{j=1}^m$ by Theorem 2.1, we define the *cone of critical directions*

$$C_{\bar{u}}^0 = \{h \in L^\infty(X) \text{ satisfying (2.11) and } h(x) = 0 \text{ for almost every } x \in X^0\}, \tag{2.10}$$

with

$$\begin{cases} G_j'(\bar{u})h = 0 & \text{if } (j \leq m_1) \text{ or } (j > m_1,\ G_j(\bar{u}) = 0, \text{ and } \bar{\lambda}_j > 0), \\ G_j'(\bar{u})h \leq 0 & \text{if } j > m_1,\ G_j(\bar{u}) = 0, \text{ and } \bar{\lambda}_j = 0, \\ h(x) = \begin{cases} \geq 0 & \text{if } \bar{u}(x) = u_a(x), \\ \leq 0 & \text{if } \bar{u}(x) = u_b(x). \end{cases} \end{cases} \tag{2.11}$$

In the following theorem we state the necessary second-order optimality conditions.

THEOREM 2.2. *Assume that* (2.1), (A1), *and* (A2) *hold;* $\{\bar{\lambda}_j\}_{j=1}^m$ *are the Lagrange multipliers satisfying* (2.2) *and* (2.3); *and* J *and* $\{G_j\}_{j=1}^m$ *are of class* C^2 *in a neighborhood of* $\bar{u}$. *Then the following inequality is satisfied:*

$$\frac{\partial^2 L}{\partial u^2}(\bar{u},\bar{\lambda})h^2 \geq 0 \quad \forall h \in C_{\bar{u}}^0. \tag{2.12}$$

To prove this theorem we will make use of the following lemma.

LEMMA 2.3. *Let us assume that* (2.1) *holds and that* J *and* $\{G_j\}_{j=1}^m$ *are of class* C^2 *in a neighborhood of* $\bar{u}$. *Let* $h \in L^\infty(X)$ *satisfy* $G_j'(\bar{u})h = 0$ *for every* $j \in I$, *where* I *is an arbitrary subset of* I_0. *Then there exist a number* $\varepsilon_h > 0$ *and* C^2*-functions* $\gamma_j : (-\varepsilon_h, +\varepsilon_h) \longrightarrow \mathbb{R}$, $j \in I$, *such that*

$$\begin{cases} G_j(u_t) = 0,\ j \in I, \text{ and } G_j(u_t) < 0,\ j \notin I_0, \quad \forall |t| \leq \varepsilon_h; \\ \gamma_j(0) = \gamma_j'(0) = 0, \quad j \in I, \end{cases} \tag{2.13}$$

with $u_t = \bar{u} + th + \sum_{j\in I} \gamma_j(t)h_j$, $\{h_j\}_{j\in I}$ *given by* (2.1).

Proof. Let k be the cardinal number of I and let us define $\omega : \mathbb{R} \times \mathbb{R}^k \longrightarrow \mathbb{R}^k$ by

$$\omega(t,\rho) = \left(G_j \left(\bar{u} + th + \sum_{i\in I} \rho_i h_i \right) \right)_{j\in I}.$$

Then ω is of class C^2 in a neighborhood of $(0,0)$,

$$\frac{\partial \omega}{\partial t}(0,0) = (G_j'(\bar{u})h)_{j\in I} = 0 \quad \text{and} \quad \frac{\partial \omega}{\partial \rho}(0,0) = (G_j'(\bar{u})h_i)_{i,j\in I} = \text{identity}.$$

Therefore we can apply the implicit function theorem and deduce the existence of $\varepsilon > 0$ and functions $\gamma_j : (-\varepsilon, +\varepsilon) \longrightarrow \mathbb{R}$ of class C^2, $j \in I$, such that

$$\omega(t, \gamma(t)) = \omega(0,0) = 0 \qquad \forall t \in (-\varepsilon, +\varepsilon) \ \text{ and } \ \gamma(0) = 0,$$

where $\gamma(t) = (\gamma_j(t))_{j\in I}$. Furthermore, by differentiation in the previous identity we get

$$\frac{\partial \omega}{\partial t}(0,0) + \frac{\partial \omega}{\partial \rho}(0,0)\gamma'(0) = 0 \Longrightarrow \gamma'(0) = 0.$$

Taking into account the continuity of γ and G_j and that $\gamma(0) = 0$, we deduce the existence of $\varepsilon_h \le \varepsilon$ such that (2.13) holds for every $t \in (-\varepsilon_h, +\varepsilon_h)$. ∎

Proof of Theorem 2.2. Let us take $h \in C^0_{\bar{u}}$ satisfying

$$h(x) = 0 \quad \text{if } u_a(x) < \bar{u}(x) < u_a(x) + \varepsilon \ \text{ or } \ u_b(x) - \varepsilon < \bar{u}(x) < u_b(x) \tag{2.14}$$

for some $\varepsilon \in (0, \varepsilon_{\bar{u}}]$. We introduce

$$I = \{1, \dots, m_1\} \cup \{j : m_1 + 1 \le j \le m,\ G_j(\bar{u}) = 0,\ \text{and } G_j'(\bar{u})h = 0\}. \tag{2.15}$$

I includes all equality constraints, all strongly active inequality constraints (i.e., $\bar{\lambda}_j > 0$), and, depending on h, possibly some of the weakly active inequality constraints (i.e., $\bar{\lambda}_j = 0$). Then we are under the assumptions of Lemma 2.3. Let us set

$$u_t = \bar{u} + th + \sum_{j\in I} \gamma_j(t) h_j, \quad t \in (-\varepsilon_h, \varepsilon_h).$$

From Lemma 2.3 we know that $G_j(u_t) = 0$ if $j \in I$, and $G_j(u_t) < 0$ if $j \notin I_0$, provided that $t \in (-\varepsilon_h, +\varepsilon_h)$. From (2.11) we deduce that $G_j(\bar{u}) = 0$ and $G_j'(\bar{u})h < 0$ for $j \in I_0 \setminus I$. Therefore we have that $G_j(u_t) < 0$ for every $j \notin I$ and $t \in (0, \varepsilon_0)$, for some $\varepsilon_0 > 0$ small. On the other hand, the assumptions on h, along with the additional condition (2.14) and the fact that $\operatorname{supp} h_j \subset X_{\varepsilon_{\bar{u}}}$, imply that $u_a(x) \le u_t(x) \le u_b(x)$ for $t \ge 0$ small enough. Consequently, by taking $\varepsilon_0 > 0$ sufficiently small, we get that u_t is a feasible control for (P) for every $t \in [0, \varepsilon_0)$. Now we know $G_j(u_t) = 0$ for $j \in I$ and $\bar{\lambda}_j = 0$ for $j \notin I_0$ (cf. (2.2)). According to (2.11) we require $G_j'(\bar{u})h = 0$ for active inequalities with $\bar{\lambda}_j > 0$; hence if i belongs to $I_0 \setminus I$, then $\bar{\lambda}_j = 0$ must hold. This leads to

$$\sum_{j=1}^{m} \bar{\lambda}_j G_j(u_t) = 0 \quad \forall t \in [0, \varepsilon_0).$$

Therefore the function $\phi : [0, +\epsilon_0) \longrightarrow \mathbb{R}$ given by

$$\phi(t) = J(u_t) + \sum_{j=1}^{m} \bar{\lambda}_j G_j(u_t)$$

has a local minimum at 0 and, taking into account that $\gamma_j'(0) = 0$,

$$\begin{aligned}\phi'(0) &= \left(J'(\bar{u}) + \sum_{j=1}^{m} \bar{\lambda}_j G_j'(\bar{u}) \right) \left(h + \sum_{j \in I} \gamma_j'(0) h_j \right) \\ &= \left[J'(\bar{u}) + \sum_{j=1}^{m} \bar{\lambda}_j G_j'(\bar{u}) \right] h = \int_X d(x) h(x) d\mu(x) = 0.\end{aligned}$$

The last identity follows from the fact that h vanishes on X^0. Since the first derivative of ϕ is zero, the following second-order necessary optimality condition must hold:

$$\begin{aligned}0 \le \phi''(0) &= \left[J''(\bar{u}) + \sum_{j=1}^{m} \bar{\lambda}_j G_j''(\bar{u}) \right] h^2 + \left[J'(\bar{u}) + \sum_{j=1}^{m} \bar{\lambda}_j G_j'(\bar{u}) \right] \left(\sum_{i \in I} \gamma_i''(0) h_i \right) \\ &= \left[J''(\bar{u}) + \sum_{j=1}^{m} \bar{\lambda}_j G_j''(\bar{u}) \right] h^2 + \sum_{i \in I} \gamma_i''(0) \int_X d(x) h_i(x) d\mu(x) \\ &= \left[J''(\bar{u}) + \sum_{j=1}^{m} \bar{\lambda}_j G_j''(\bar{u}) \right] h^2 = \frac{\partial^2 L}{\partial u^2}(\bar{u}, \bar{\lambda}) h^2.\end{aligned}$$

Here we have used (A1). Now let us consider $h \in L^\infty(X)$ satisfying (2.11) but not (2.14), i.e., h is any critical direction. The main idea in this case is to approach h by functions h_ε, which belong to the critical cone $C_{\bar{u}}^0$ and satisfy (2.14) as well. Then for every $\varepsilon > 0$, we define $A_\varepsilon = X_\varepsilon \cup \{x \in X : \bar{u}(x) = u_a(x) \text{ or } \bar{u}(x) = u_b(x)\}$. This is the complement of the set of points x satisfying (2.14). Set

$$h_\varepsilon = h\chi_{A_\varepsilon} + \sum_{i \in I} \left[\int_{X \setminus A_\varepsilon} g_i(x) h(x) d\mu(x) \right] h_i = h\chi_{A_\varepsilon} + \hat{h},$$

where χ_{A_ε} is the characteristic function of A_ε and I is given by (2.15). We verify that h_ε belongs to $C_{\bar{u}}^0$, while $h\chi_{A_\varepsilon}$ is possibly not contained in this cone.

Thus for every $j \in I$, using (2.1) and taking $0 < \varepsilon < \varepsilon_{\bar{u}}$, we have

$$\begin{aligned}G_j'(\bar{u}) h_\varepsilon &= \int_X g_j(x)(h\chi_{A_\varepsilon})(x) d\mu(x) + \int_X g_j(x) \hat{h}(x) d\mu(x) \\ &= \int_{A_\varepsilon} g_j(x) h(x) d\mu(x) \\ &\quad + \sum_{i \in I} \left[\int_{X \setminus A_\varepsilon} g_i(x) h(x) d\mu(x) \right] \int_X g_j(x) h_i(x) d\mu(x) \\ &= \int_{A_\varepsilon} g_j(x) h(x) d\mu(x) + \sum_{i \in I} \left[\int_{X \setminus A_\varepsilon} g_i(x) h(x) d\mu(x) \right] \delta_{ji} \\ &= \int_X g_j(x) h(x) d\mu(x) = G_j'(\bar{u}) h = 0.\end{aligned}$$

In the case of $j \in I_0 \setminus I$, we have $G_j'(\bar{u})h < 0$. Then it is enough to take ε sufficiently small to get $G_j'(\bar{u})h_\varepsilon < 0$.

Thus, recalling that $\operatorname{supp} h_j \subset X_{\varepsilon_{\bar{u}}}$, we infer that h_ε satisfies the conditions (2.11) and (2.14); therefore (2.12) holds for each h_ε, $\varepsilon > 0$ small enough.

Finally, it is clear that $h_\varepsilon(x) \to h(x)$ a.e. in X as $\varepsilon \to 0$. Therefore, assumption (A2) allows us to pass to the limit in the second-order optimality conditions satisfied for every h_ε and to conclude (2.12). ∎

3. Sufficient optimality conditions. Whenever nonlinear optimal control problems are solved, second-order sufficient conditions play an essential role in the numerical analysis. For instance, they ensure local convergence of Lagrange–Newton–SQP methods; see Alt and Malanowski [2], Dontchev et al. [11], Ito and Kunisch [18], or Schulz [23], and the references cited therein. Such conditions are important for error estimates as well. We refer, for instance, to Arada, Casas, and Tröltzsch [1] and Hager [15]. Finally, we mention that second-order conditions should be checked numerically to verify local optimality of computed solutions; see Mittelmann [21].

In this section, $\bar{u}$ is a given feasible element for the problem (P). Motivated again by the considerations on the two-norm discrepancy, we have to make some assumptions involving the $L^\infty(X)$ and $L^2(X)$ norms, as follows.

(A3) There exists a positive number $r > 0$ such that J and $\{G_j\}_{j=1}^m$ are of class C^2 in the $L^\infty(X)$-ball $B_r(\bar{u})$, and for every $\eta > 0$ there exists $\varepsilon \in (0, r)$ such that for each $u \in B_r(\bar{u})$, $\|v - \bar{u}\|_{L^\infty(X)} < \varepsilon$, $h, h_1, h_2 \in L^\infty(X)$, and $1 \le j \le m$ we have

$$
\begin{cases}
\left| \left[\dfrac{\partial^2 L}{\partial u^2}(v, \bar{\lambda}) - \dfrac{\partial^2 L}{\partial u^2}(\bar{u}, \bar{\lambda}) \right] h^2 \right| \le \eta \|h\|^2_{L^2(X)}, \\
|J'(u)h| \le M_{0,1}\|h\|_{L^2(X)}, \quad |G_j'(u)h| \le M_{j,1}\|h\|_{L^2(X)}, \\
|J''(u)h_1h_2| \le M_{0,2}\|h_1\|_{L^2(X)}\|h_2\|_{L^2(X)}, \\
|G_j''(u)h_1h_2| \le M_{j,2}\|h_1\|_{L^2(X)}\|h_2\|_{L^2(X)}.
\end{cases} \tag{3.1}
$$

Analogously to (2.9) and (2.10), we define for every $\tau > 0$

$$
X^\tau = \{x \in X : |d(x)| > \tau\}, \tag{3.2}
$$

$$
C_{\bar{u}}^\tau = \{h \in L^\infty(X) \ \text{ satisfying } \ (2.11) \ \text{ and } \ h(x) = 0 \ \text{ a.e. } \ x \in X^\tau\}. \tag{3.3}
$$

The next theorem provides the second-order sufficient optimality conditions of (P). Although they seem to be different from the classical ones, we will prove later that they are equivalent; see Theorem 3.2 and Corollary 3.3.

THEOREM 3.1. *Let $\bar{u}$ be a feasible point for problem* (P) *verifying the first-order necessary conditions* (2.2) *and* (2.3)*, and let us suppose that assumptions* (2.1), (A1), *and* (A3) *hold. Let us also assume that for every $h \in L^\infty(X)$ satisfying* (2.11)

$$
\frac{\partial^2 L}{\partial u^2}(\bar{u}, \bar{\lambda})h^2 \ge \delta_1 \|h\|^2_{L^2(X \setminus X^\tau)} - \delta_2 \|h\|^2_{L^2(X^\tau)} \tag{3.4}
$$

holds for some $\delta_1 > 0$, $\delta_2 \ge 0$, and $\tau > 0$. Then there exist $\varepsilon > 0$ and $\delta > 0$ such that $J(\bar{u}) + \delta\|u - \bar{u}\|^2_{L^2(X)} \le J(u)$ for every feasible point u for (P)*, with $\|u - \bar{u}\|_{L^\infty(X)} < \varepsilon$.*

Proof. (i) *Condition* (3.4) *is stable w.r.t. perturbations of $\bar{u}$.* Without loss of generality, we will assume that $\delta_2 > 0$. From (A3) we deduce the existence of $r_0 \in$

$(0,r)$ such that for all $h \in L^\infty(X)$ and $\|v-\bar u\|_{L^\infty(X)} < r_0$

$$\left|\left[\frac{\partial^2 L}{\partial u^2}(v,\bar\lambda) - \frac{\partial^2 L}{\partial u^2}(\bar u,\bar\lambda)\right] h^2\right| \le \min\left\{\frac{\delta_1}{2},\delta_2\right\}\|h\|^2_{L^2(X)}.$$

From this inequality and (3.4) it follows easily that

$$\frac{\partial^2 L}{\partial u^2}(v,\bar\lambda)h^2 \ge \frac{\delta_1}{2}\|h\|^2_{L^2(X\setminus X^\tau)} - 2\delta_2\|h\|^2_{L^2(X^\tau)} \tag{3.5}$$

for every h satisfying (2.11) and $\|v-\bar u\|_{L^\infty(X)} < r_0$.

(ii) *Some technical definitions.* Let us set

$$M = M_{0,2} + \sum_{j=1}^m |\bar\lambda_j| M_{j,2} \quad\text{and}\quad \rho = \min\left\{1, \frac{\delta_1}{16M}\right\}, \tag{3.6}$$

$$C_1 = \max\left\{\frac{\delta_1}{4}, 2\delta_2\right\} + \frac{3M}{2} + \frac{4M^2}{\delta_1}, \qquad C_2 = \frac{C_1}{2}\max_{j\in I_0}\|h_j\|^2_{L^2(X)}\left[\sum_{j=1}^m M_{j,2}\right]^2, \tag{3.7}$$

$$C_3 = 2C_1 m\mu(X)^{1/2}\max_{j\in I_0}\|h_j\|^2_{L^2(X)}\max_{1\le j\le m} M_{j,1}. \tag{3.8}$$

Finally, we take

$$\varepsilon = \min\left\{r_0, \sqrt{\frac{\delta_1}{64C_2\mu(X)}}, \frac{8\tau}{\delta_1+16\delta_2}, \frac{\rho}{C_3}\min_{j\in I_+, j>m_1}\bar\lambda_j\right\}, \tag{3.9}$$

where

$$I_+ = \{1,\dots,m_1\}\cup\{j>m_1 : G_j(\bar u) = 0 \text{ and } \bar\lambda_j > 0\}.$$

(iii) *Approximation of $u-\bar u$ by elements of the critical cone.* Let u be a feasible point for problem (P), with $\|u-\bar u\|_{L^\infty(X)} < \epsilon$. Then $u-\bar u$ will not, in general, belong to the critical cone. Therefore, we use the representation $u-\bar u = h + h_0$, where h is in the critical cone and h_0 is some small correction.

Let us introduce the set of indices

$$I_u = \{j\in I_0 : G_j'(\bar u)(u-\bar u) > 0 \text{ or } [G_j'(\bar u)(u-\bar u) < 0 \text{ and } j\in I_+]\}.$$

This is the set of indices for which we need to correct $G_j'(\bar u)(u-\bar u)$, since the conditions of the critical cone are not met. We need to carry out this correction for equality constraints if $G_j'(\bar u)(u-\bar u)\ne 0$. We also need to apply this correction for an active inequality constraint satisfying $G_j'(\bar u)(u-\bar u) > 0$ or for a strongly active inequality constraint if $G_j'(\bar u)(u-\bar u) < 0$ holds. We define for all $j\in I_u$

$$\alpha_j = G_j'(\bar u)(u-\bar u), \qquad h_0 = \sum_{j\in I_u}\alpha_j h_j, \qquad\text{and}\quad h = u-\bar u - h_0, \tag{3.10}$$

where the elements h_j are introduced in assumption (2.1). Then h satisfies (2.11). This is seen as follows:

$$G_j'(\bar u)h_0 = \sum_{i\in I_u}\alpha_i G_j'(\bar u)h_i = \sum_{i\in I_u}\alpha_i\delta_{ji}.$$

If $j \notin I_u$, then $\delta_{ji} = 0$ for all $i \in I_u$; hence

$$G_j'(\bar{u})h = G_j'(\bar{u})(u-\bar{u}) - G_j'(\bar{u})h_0 = G_j'(\bar{u})(u-\bar{u}) \begin{cases} = 0 & \text{if } j \le m_1, \\ \le 0 & \text{if } j > m_1 \end{cases}$$

(the last inequality follows from $j \notin I_u$). Thus $G_j'(\bar{u})h$ fulfills the conditions of the critical cone. If $j \in I_u$, then

$$G_j'(\bar{u})h = G_j'(\bar{u})(u-\bar{u}) - \alpha_j \delta_{jj} = \alpha_j - \alpha_j = 0,$$

and $G_j'(\bar{u})h$ also fulfills the conditions of the critical cone.

Let us now estimate h_0 in $L^2(X)$. For every $j \in I_u$ there exists $v_j = \bar{u} + \theta_j(u-\bar{u})$, with $0 < \theta_j < 1$, such that

$$0 \ge G_j(u) = G_j(\bar{u}) + G_j'(\bar{u})(u-\bar{u}) + \frac{1}{2}G_j''(v_j)(u-\bar{u})^2 = \alpha_j + \frac{1}{2}G_j''(v_j)(u-\bar{u})^2. \tag{3.11}$$

If $\alpha_j \ge 0$, we deduce from (3.11) and (3.1) that

$$|\alpha_j| = \alpha_j \le \frac{1}{2}|G_j''(v_j)(u-\bar{u})^2| \le \frac{1}{2}M_{j,2}\|u-\bar{u}\|^2_{L^2(X)}. \tag{3.12}$$

If $\alpha_j < 0$ and $G_j(u) = 0$, we get

$$|\alpha_j| = -\alpha_j = \frac{1}{2}G_j''(v_j)(u-\bar{u})^2 \le \frac{1}{2}M_{j,2}\|u-\bar{u}\|^2_{L^2(X)}. \tag{3.13}$$

Let us define

$$I_u^- = \{j \in I_u : G_j(u) < 0 \text{ and } \alpha_j < 0\}.$$

This is the set of all indices, where we do not obtain an estimate of α_j having the order $\|u-\bar{u}\|^2_{L^2(x)}$. We should notice at this point that $\bar{\lambda}_j > 0$ holds for all $j \in I_u^-$. (Since u must be feasible, j stands for an inequality constraint. Therefore, $0 > \alpha_j = G_j'(\bar{u})(u-\bar{u})$, and $j \in I_u$ implies $j \in I_+$.) Then we have

$$\|h_0\|_{L^2(X)} \le \max_{j\in I_0}\|h_j\|_{L^2(X)} \left[\frac{1}{2}\left(\sum_{j=1}^m M_{j,2}\right)\|u-\bar{u}\|^2_{L^2(X)} + \sum_{j\in I_u^-}|\alpha_j|\right]. \tag{3.14}$$

(iv) *Estimation of* $J(u) - J(\bar{u})$. Using (2.6), (2.7), (3.6), (3.10), and (3.11), for some $v = \bar{u} + \theta(u-\bar{u})$, $0 < \theta < 1$,

$$\begin{aligned} J(u) &= J(u) + \sum_{j=1}^{m_1}\bar{\lambda}_j G_j(u) + \sum_{j=m_1+1}^{m}\bar{\lambda}_j G_j(u) - \sum_{j=m_1+1}^{m}\bar{\lambda}_j G_j(u) \\ &= L(u,\bar{\lambda}) - \sum_{j=m_1+1}^{m}\bar{\lambda}_j G_j(u) \\ &\ge L(u,\bar{\lambda}) - \sum_{j\in I_u^-}\bar{\lambda}_j G_j(u) \ge L(u,\bar{\lambda}) - \rho\sum_{j\in I_u^-}\bar{\lambda}_j G_j(u) \end{aligned}$$

holds, since $\rho < 1$. Therefore,

$$\begin{aligned}
J(u) &\geq L(u,\bar\lambda) - \rho \sum_{j\in I_u^-} \bar\lambda_j G_j(u) = L(\bar u,\bar\lambda) + \frac{\partial L}{\partial u}(\bar u,\bar\lambda)(u-\bar u) + \frac12 \frac{\partial^2 L}{\partial u^2}(v,\bar\lambda)(u-\bar u)^2 \\
&\quad - \rho \sum_{j\in I_u^-} \bar\lambda_j \alpha_j - \frac{\rho}{2} \sum_{j\in I_u^-} \bar\lambda_j G_j''(v_j)(u-\bar u)^2 \\
&= J(\bar u) + \int_X d(x)(u(x)-\bar u(x))d\mu(x) + \frac12 \frac{\partial^2 L}{\partial u^2}(v,\bar\lambda)h^2 \\
&\quad + \frac{\partial^2 L}{\partial u^2}(v,\bar\lambda)hh_0 + \frac12 \frac{\partial^2 L}{\partial u^2}(v,\bar\lambda)h_0^2 + \rho \sum_{j\in I_u^-} \bar\lambda_j|\alpha_j| - \frac{\rho}{2} \sum_{j\in I_u^-} \bar\lambda_j G_j''(v_j)(u-\bar u)^2.
\end{aligned}$$

Now from (2.8), (2.11), (3.1), (3.5), and (3.6) it follows that

$$\begin{aligned}
J(u) &\geq J(\bar u) + \tau \int_{X^\tau} |u(x)-\bar u(x)|d\mu(x) + \frac{\delta_1}{4}\|h\|^2_{L^2(X\setminus X^\tau)} - \delta_2\|h\|^2_{L^2(X^\tau)} \\
&\quad - M\|h_0\|_{L^2(X)}\|h\|_{L^2(X)} - \frac{M}{2}\|h_0\|^2_{L^2(X)} + \rho \sum_{j\in I_u^-} \bar\lambda_j|\alpha_j| \\
&\quad - \frac{\rho}{2}\left[\sum_{j\in I_u^-} \bar\lambda_j M_{j,2}\right] \|u-\bar u\|^2_{L^2(X)} \\
&\geq J(\bar u) + \frac{\tau}{\varepsilon}\|u-\bar u\|^2_{L^2(X^\tau)} + \frac{\delta_1}{8}\|u-\bar u\|^2_{L^2(X\setminus X^\tau)} - \frac{\delta_1}{4}\|h_0\|^2_{L^2(X\setminus X^\tau)} \\
&\quad - 2\delta_2\|u-\bar u\|^2_{L^2(X^\tau)} - 2\delta_2\|h_0\|^2_{L^2(X^\tau)} \\
&\quad - M\|h_0\|_{L^2(X)}\big(\|u-\bar u\|_{L^2(X)} + \|h_0\|_{L^2(X)}\big) - \frac{M}{2}\|h_0\|^2_{L^2(X)} \\
&\quad + \rho \sum_{j\in I_u^-} \bar\lambda_j|\alpha_j| - \frac{\rho}{2}M\|u-\bar u\|^2_{L^2(X)}. \qquad (3.15)
\end{aligned}$$

Using the definition of ε from (3.9), we have

$$\frac{\tau}{\varepsilon} - 2\delta_2 \geq \frac{\delta_1}{8}. \qquad (3.16)$$

On the other hand,

$$\begin{aligned}
M\|h_0\|_{L^2(X)}\|u-\bar u\|_{L^2(X)} &= 2\left[\frac{\sqrt{\delta_1}}{4}\|u-\bar u\|_{L^2(X)}\right]\left[\frac{2M}{\sqrt{\delta_1}}\|h_0\|_{L^2(X)}\right] \\
&\leq \frac{\delta_1}{16}\|u-\bar u\|^2_{L^2(X)} + \frac{4M^2}{\delta_1}\|h_0\|^2_{L^2(X)}. \qquad (3.17)
\end{aligned}$$

From the definitions of C_1 and ρ given in (3.7) and (3.6) along with (3.15), (3.16), and (3.17), we get

$$\begin{aligned}
J(u) &\geq J(\bar u) + \frac{\delta_1}{8}\|u-\bar u\|^2_{L^2(X)} - C_1\|h_0\|^2_{L^2(X)} \\
&\quad - \frac{\delta_1}{16}\|u-\bar u\|^2_{L^2(X)} + \rho \sum_{j\in I_u^-} \bar\lambda_j|\alpha_j| - \frac{\delta_1}{32}\|u-\bar u\|^2_{L^2(X)} \\
&= J(\bar u) + \frac{\delta_1}{32}\|u-\bar u\|^2_{L^2(X)} - C_1\|h_0\|^2_{L^2(X)} + \rho \min_{j\in I_+, j>m_1} \bar\lambda_j \sum_{j\in I_u^-} |\alpha_j|. \qquad (3.18)
\end{aligned}$$

(v) *Two auxiliary estimates and final result*. From (3.7), (3.9), and (3.14) we get, on using $(a+b)^2 \le 2\,(a^2+b^2)$,

$$C_1\|h_0\|^2_{L^2(X)} \le C_1 \max_{j\in I_0}\|h_j\|^2_{L^2(X)}\left[\frac{1}{2}\left(\sum_{j=1}^m M_{j,2}\right)^2\|u-\bar u\|^4_{L^2(X)} + 2\left(\sum_{j\in I_u^-}|\alpha_j|\right)^2\right]$$

$$= C_2\|u-\bar u\|^4_{L^2(X)} + 2C_1\max_{j\in I_0}\|h_j\|^2_{L^2(X)}\left(\sum_{j\in I_u^-}|\alpha_j|\right)^2$$

$$\le C_2\varepsilon^2\mu(X)\|u-\bar u\|^2_{L^2(X)} + 2C_1\max_{j\in I_0}\|h_j\|^2_{L^2(X)}\left(\sum_{j\in I_u^-}|\alpha_j|\right)^2$$

$$\le \frac{\delta_1}{64}\|u-\bar u\|^2_{L^2(X)} + 2C_1\max_{j\in I_0}\|h_j\|^2_{L^2(X)}\left(\sum_{j\in I_u^-}|\alpha_j|\right)^2. \tag{3.19}$$

The definition of α_j given by (3.10) along with assumption (3.1) imply

$$|\alpha_j| \le M_{j,1}\|u-\bar u\|_{L^2(X)} \le M_{j,1}\varepsilon\sqrt{\mu(X)}. \tag{3.20}$$

From (3.8) and the above inequality, we deduce

$$2C_1\max_{j\in I_0}\|h_j\|^2_{L^2(X)}\left(\sum_{j\in I_u^-}|\alpha_j|\right) \le C_3\varepsilon. \tag{3.21}$$

Definition (3.9) and (3.21) lead to

$$\rho\min_{j\in I_+, j>m_1}\bar\lambda_j - 2C_1\max_{j\in I_0}\|h_j\|^2_{L^2(X)}\left(\sum_{j\in I_u^-}|\alpha_j|\right) \ge 0. \tag{3.22}$$

Finally, combining (3.18), (3.19), and (3.22), we conclude the desired result:

$$J(u) \ge J(\bar u) + \frac{\delta_1}{64}\|u-\bar u\|^2_{L^2(X)}. \qquad \square$$

Now we prove the equivalence between the sufficient optimality conditions stated in Theorem 3.1 and the classical ones.

THEOREM 3.2. *Let $\bar u$ be a feasible point of* (P) *satisfying* (2.2) *and* (2.3). *Let $C_{\bar u}$ be the set of elements $h \in L^\infty(X)$ satisfying* (2.11), *and $C^\tau_{\bar u}$ be given by* (3.3). *Let us suppose that assumptions* (2.1), (A1), *and* (A3) *hold. Let $\tau > 0$ be given. Then the following statements are equivalent:*

$$\exists \delta > 0 : \frac{\partial^2 L}{\partial u^2}(\bar u, \bar\lambda)h^2 \ge \delta\|h\|^2_{L^2(X)} \quad \forall h \in C^\tau_{\bar u}, \tag{3.23}$$

$$\exists \delta_1 > 0, \delta_2 \ge 0 : \frac{\partial^2 L}{\partial u^2}(\bar u, \bar\lambda)h^2 \ge \delta_1\|h\|^2_{L^2(X\setminus X^\tau)} - \delta_2\|h\|^2_{L^2(X^\tau)} \quad \forall h \in C_{\bar u}. \tag{3.24}$$

Proof. It is obvious that (3.24) implies (3.23), since $h = 0$ in X^τ if $h \in C^\tau_{\bar u}$. Therefore, it is enough to take $\delta = \delta_1$. Let us prove the opposite implication. Let $h \in C_{\bar u}$. We set $h_\tau = h\chi_{X^\tau}$, where χ_{X^τ} is the characteristic function of X^τ and

$$I_h = \{j \in I_0 : G_j'(\bar u)(h-h_\tau) > 0 \ \text{ or } \ [G_j'(\bar u)(h-h_\tau) < 0 \ \text{ and } \ G_j'(\bar u)h = 0]\}.$$

We define

$$\alpha_j = G_j'(\bar u)(h - h_\tau) \quad \forall j \in I_h, \qquad \hat h = \sum_{j\in I_h} \alpha_j h_j, \qquad \text{and} \quad h_0 = h - h_\tau - \hat h,$$

where the functions h_j are given by (2.1).

Let us see that $h_0 \in C_{\bar u}^\tau$. Since $\operatorname{supp} h_j \subset X_{\varepsilon_{\bar u}}$ and $h - h_\tau = h(1 - \chi_{X^\tau})$, we have that $h_0(x) = 0$ for $x \in X^\tau$. Now we distinguish between the cases $j \in I_h$ and $j \in I_0 \setminus I_h$.

If $j \in I_h$, then

$$G_j'(\bar u)h_0 = G_j'(\bar u)(h - h_\tau) - \sum_{i\in I_h} \alpha_i G_j'(\bar u)h_i = G_j'(\bar u)(h - h_\tau) - \alpha_j = 0.$$

If $j \in I_0 \setminus I_h$, then from the definition of I_h we obtain that $G_j'(\bar u)h_0 = G_j'(\bar u)(h - h_\tau) \le 0$.

If this inequality reduces to an equality $G_j'(\bar u)(h-h_\tau) = 0$, then h_0 verifies that the condition is in $C_{\bar u}^\tau$. In the remaining case in which $j \in I_0 \setminus I_h$ but $G_j'(\bar u)(h - h_\tau) < 0$, using again the definition of I_h, we deduce that $G_j'(\bar u)h < 0$. ($G_j'(\bar u)h = 0$ and $G_j'(\bar u)(h - h_\tau) < 0$ would give $j \in I_h$.) Consequently, since $h \in C_{\bar u}$, we have that $j > m_1$ and $\bar\lambda_j = 0$ (otherwise, $h \in C_{\bar u}^\tau$ and $\bar\lambda_j > 0$ would imply $G_j'(\bar u)h = 0$). Then the inequality $G_j'(\bar u)h_0 < 0$ also means that h_0 shows the condition to be in $C_{\bar u}^\tau$.

We now prove that

$$\|\hat h\|_{L^2(X)} \le C_0 \|h_\tau\|_{L^2(X)}, \tag{3.25}$$

where

$$C_0 = \sum_{j\in I_0} \|g_j\|_{L^2(X)} \|h_j\|_{L^2(X)},$$

g_j being given in (2.4). Indeed, if $\alpha_j > 0$, then

$$|\alpha_j| = \alpha_j = G_j'(\bar u)(h - h_\tau) = G_j'(\bar u)h - G_j'(\bar u)h_\tau \le -G_j'(\bar u)h_\tau \le \|g_j\|_{L^2(X)} \|h_\tau\|_{L^2(X)}.$$

If $\alpha_j < 0$, then from the definition of I_h we have that $G_j'(\bar u)h = 0$; therefore

$$|\alpha_j| = -\alpha_j = -G_j'(\bar u)(h - h_\tau) = G_j'(\bar u)h_\tau \le \|g_j\|_{L^2(X)} \|h_\tau\|_{L^2(X)}.$$

Combining the previous two inequalities and the definition of $\hat h$, we get (3.25).

Finally, taking M as in (3.6), we obtain from (3.23) and (3.25)

$$\begin{aligned}
\frac{\partial^2 L}{\partial u^2}(\bar u, \bar\lambda)h^2 &= \frac{\partial^2 L}{\partial u^2}(\bar u, \bar\lambda)h_0^2 + \frac{\partial^2 L}{\partial u^2}(\bar u, \bar\lambda)(h_\tau + \hat h)^2 + 2\frac{\partial^2 L}{\partial u^2}(\bar u, \bar\lambda)h_0(h_\tau + \hat h) \\
&\ge \delta \|h_0\|_{L^2(X)}^2 - M\|h_\tau + \hat h\|_{L^2(X)}^2 - 2M\|h_0\|_{L^2(X)}\|h_\tau + \hat h\|_{L^2(X)} \\
&\ge \frac{\delta}{2}\|h - h_\tau\|_{L^2(X)}^2 - \delta\|\hat h\|_{L^2(X)}^2 - 2M(\|h_\tau\|_{L^2(X)}^2 + \|\hat h\|_{L^2(X)}^2) \\
&\quad - 2M(\|h - h_\tau\|_{L^2(X)} + \|\hat h\|_{L^2(X)})(\|h_\tau\|_{L^2(X)} + \|\hat h\|_{L^2(X)}) \\
&\ge \frac{\delta}{2}\|h - h_\tau\|_{L^2(X)}^2 - C_0^2\delta\|h_\tau\|_{L^2(X)}^2 - 2M(C_0^2 + 1)\|h_\tau\|_{L^2(X)}^2 \\
&\quad - 2M(C_0 + 1)(\|h - h_\tau\|_{L^2(X)} + C_0\|h_\tau\|_{L^2(X)})\|h_\tau\|_{L^2(X)}
\end{aligned}$$

$$\geq \frac{\delta}{4}\|h-h_\tau\|^2_{L^2(X)} - \left\{C_0^2\delta + 2M(C_0^2+1) + \frac{4M^2(C_0+1)^2}{\delta} + 2M(C_0+1)C_0\right\}\|h_\tau\|^2_{L^2(X)}$$
$$= \delta_1\|h\|^2_{L^2(X\setminus X^\tau)} - \delta_2\|h\|^2_{L^2(X^\tau)},$$

where obviously $\delta_1 > 0$ and $\delta_2 \geq 0$ are independent of $h \in C_{\bar u}$. ∎

The following corollary is an immediate consequence of Theorems 3.1 and 3.2.

COROLLARY 3.3. *Let $\bar u$ be a feasible point for problem* (P) *satisfying* (2.2) *and* (2.3), *and suppose that assumptions* (2.1), (A1), *and* (A3) *hold. Assume also that*

$$\frac{\partial^2 L}{\partial u^2}(\bar u,\bar\lambda)h^2 \geq \delta\|h\|^2_{L^2(X)} \quad \forall h \in C^\tau_{\bar u} \tag{3.26}$$

for some $\delta > 0$ and $\tau > 0$ given. Then there exist $\varepsilon > 0$ and $\alpha > 0$ such that $J(\bar u)+\alpha\|u-\bar u\|^2_{L^2(X)} \leq J(u)$ for every feasible point u for (P), *with $\|u-\bar u\|_{L^\infty(X)} < \varepsilon$.*

Remark 3.4. Comparing the sufficient optimality condition (3.4) with the necessary condition (2.12), we notice the existence of a gap between the two, arising from two facts. First, the constant δ_1 is strictly positive in (3.4), and it can be zero in (2.12), which is the classical situation even in finite dimensions. Second, we cannot substitute, in general, $C^\tau_{\bar u}$, with $\tau > 0$, for $C^0_{\bar u}$ in (3.26), as is done in (2.12), because of the presence of an infinite number of constraints. Quite similar strategies are employed by Maurer and Zowe [20], Maurer [19], Donchev et al. [11], and Dunn [12]. The following example, due to Dunn [13], demonstrates the impossibility of taking $\tau = 0$ in (3.26). Let us consider $X = [0,1]$, $\mathcal{S}$ the σ-algebra of Lebesgue-measurable sets of $[0,1]$, μ the Lebesgue measure in $[0,1]$, and $a(x) = 1 - 2x$. The optimization problem is

$$\begin{cases} \text{minimize } J(u) = \displaystyle\int_0^1 [2a(x)u(x) - \operatorname{sign}(a(x))u(x)^2]dx, \\ u \in L^\infty([0,1]),\ u(x) \geq 0 \text{ a.e. } x \in [0,1]. \end{cases}$$

Let us set $\bar u(x) = \max\{0, -a(x)\}$. Then we have that

$$J'(\bar u)h = \int_0^1 2[a(x) - \operatorname{sign}(a(x))\bar u(x)]h(x)dx = \int_0^{1/2} 2a(x)h(x)dx \geq 0$$

holds for all $h \in L^2([0,1])$, with $h(x) \geq 0$. If we assume that $h(x) = 0$ for $x \in X^0$,

$$J''(\bar u)h^2 = -\int_0^1 2\operatorname{sign}(a(x))h^2(x)dx = 2\int_{1/2}^1 h^2(x)dx - 2\int_0^{1/2} h^2(x)dx = 2\|h\|^2_{L^2(X)}$$

holds, where, following the notation introduced in (2.9),

$$X^0 = \{x \in [0,1] : |d(x)| > 0\} = \left[0, \frac{1}{2}\right).$$

Thus (3.26) holds with $\delta = 2$ and $\tau = 0$. However, $\bar u$ is not a local minimum in $L^\infty([0,1])$. Indeed, let us take for $0 < \varepsilon < \frac{1}{2}$

$$u_\varepsilon(x) = \begin{cases} \bar u(x) + 3\varepsilon & \text{if } x \in \left[\frac{1}{2} - \varepsilon, \frac{1}{2}\right], \\ \bar u(x) & \text{otherwise.} \end{cases}$$

Then we have $J(u_\varepsilon) - J(\bar u) = -3\varepsilon^3 < 0$. The reader can easily check that the only points u satisfying the first-order optimality conditions are given by the formula

$$u(x) = \begin{cases} 0 & \text{if } x \in Z, \\ \operatorname{sign}(a(x))a(x) & \text{otherwise,} \end{cases}$$

where Z is any measurable subset of $[0,1]$ satisfying that $a(x) \geq 0$ for every $x \in Z$. None of these points is a local minimum of the optimization problem. Moreover, if we define $u_k(x) = k \cdot \max\{0, a(x)\}$, then $J(u_k) = k(2-k)/6 \to -\infty$ when $k \to +\infty$.

4. Application to some optimal control problems.

4.1. An abstract control problem. Let, in addition to the measure space (X, S, μ), Y and Z be real Banach spaces; let $A : Y \to Z$ be a linear continuous operator; and let $B : Y \times L^\infty(X) \to Z$ be an operator of class C^2. Moreover, $F, F_j : Y \times L^\infty(X) \to \mathbb{R}$ are functionals of class C^2, $j = 1, \dots, m$. Consider the optimal control problem

$$\text{(OC)} \qquad \begin{cases} \text{minimize } F(y,u), \\ Ay + B(y,u) = 0, \\ u_a(x) \leq u(x) \leq u_b(x) & \text{a.e. } x \in X, \\ F_j(y,u) = 0, & 1 \leq j \leq m_1, \\ F_j(y,u) \leq 0, & m_1 + 1 \leq j \leq m, \end{cases}$$

where the control u is taken from $L^\infty(X)$. We assume that for all $u \in L^\infty(X)$ the equation $Ay + B(y,u) = 0$ admits a unique solution $y \in Y$, so that a control-state mapping $G : u \mapsto y$ is defined. Moreover, the inverse operator $(A + \frac{\partial B}{\partial y}(y,u))^{-1} : Z \to Y$ is assumed to exist for all $(y,u) \in Y \times L^\infty(X)$ as a linear continuous operator. Then the implicit function theorem yields that G is of class C^2 from $L^\infty(X)$ to Y. The first- and second-order derivatives $G'(u)$ and $G''(u)$ are given as follows: Define $y = G(u)$, $z_h = G'(u)h$, and $z_{h_1h_2} := G''(u)[h_1, h_2] := (G''(u)h_1)h_2$. Then z_h is the unique solution of

$$Az + \frac{\partial B}{\partial y}(y,u)z + \frac{\partial B}{\partial u}(y,u)h = 0, \tag{4.1}$$

while $z_{h_1h_2}$ is uniquely determined by

$$\begin{cases} Az + \dfrac{\partial B}{\partial y}(y,u)z = -\left\{ \dfrac{\partial^2 B}{\partial y^2}(y,u)[z_{h_1}, z_{h_2}] + \dfrac{\partial^2 B}{\partial y \partial u}(y,u)[z_{h_1}, h_2] \right. \\ \qquad\qquad \left. + \dfrac{\partial^2 B}{\partial u \partial y}(y,u)[h_1, z_{h_2}] + \dfrac{\partial^2 B}{\partial u^2}(y,u)[h_1, h_2] \right\}. \end{cases} \tag{4.2}$$

We omit the proof, which can easily be transferred from that of Theorem 2.3 in [7]. The abstract control problem (OC) fits in the optimization problem (P) by

$$J(u) := F(G(u), u), \qquad G_j(u) := F_j(G(u), u).$$

In this way, we obtain necessary and/or sufficient conditions for local solutions $(\bar y, \bar u)$ of (OC) by application of Theorems 2.1, 2.2, and 3.1 and Corollary 3.3, provided that the corresponding assumptions (2.1) and (A1)–(A3) are satisfied. We tacitly assume

this in what follows and formulate these results in a way that is convenient for optimal control problems. A *Lagrange function* $\mathcal{L} = \mathcal{L}(y, u, \varphi, \lambda)$ is associated with (OC) by

$$\mathcal{L}(y, u, \varphi, \lambda) = F(y, u) - \langle \varphi, Ay + B(y, u) \rangle + \sum_{j=1}^{m} \lambda_j F_j(y, u), \tag{4.3}$$

where $\varphi \in Z^*$, and $\langle \cdot, \cdot \rangle$ denotes the duality between Z and Z^*. Notice that we must distinguish between L for (P) and $\mathcal{L}$ for (OC). We have

$$J'(\bar{u})h = \frac{\partial F}{\partial y}(\bar{y}, \bar{u})G'(\bar{u})h + \frac{\partial F}{\partial u}(\bar{y}, \bar{u})h$$

and obtain similar expressions for $G_j(\bar{u})h$. Therefore, (2.6) yields

$$\begin{cases} \dfrac{\partial L}{\partial u}(\bar{u}, \bar{\lambda})h = \left[\dfrac{\partial F}{\partial y}(\bar{y}, \bar{u}) + \displaystyle\sum_{j=1}^{m} \bar{\lambda}_j \frac{\partial F_j}{\partial y}(\bar{y}, \bar{u}) \right] G'(\bar{u})h \\ \qquad\qquad + \left[\dfrac{\partial F}{\partial u}(\bar{y}, \bar{u}) + \displaystyle\sum_{j=1}^{m} \bar{\lambda}_j \frac{\partial F_j}{\partial y}(\bar{y}, \bar{u}) \right] h. \end{cases} \tag{4.4}$$

Define an *adjoint state* $\varphi \in Z^*$ by

$$\left[\frac{\partial F}{\partial y}(\bar{y}, \bar{u}) + \sum_{j=1}^{m} \bar{\lambda}_j \frac{\partial F_j}{\partial y}(\bar{y}, \bar{u}) \right] y = \left\langle \bar{\varphi}, Ay + \frac{\partial B}{\partial y}(\bar{y}, \bar{u})y \right\rangle \qquad \forall y \in Y. \tag{4.5}$$

We assume that $\bar{\varphi}$ is well defined by (4.5), which is true in our applications. Notice that (4.5) is equivalent to $\partial \mathcal{L}/\partial y(\bar{y}, \bar{u}, \bar{\varphi}, \bar{\lambda})y = 0$ for all $y \in Y$; that is, $\partial \mathcal{L}/\partial y(\bar{y}, \bar{u}, \bar{\varphi}, \bar{\lambda}) = 0$ in the sense of Y^*. Insert $y = z_h = G'(\bar{u})h$ into (4.5); then y solves (4.1), and the right-hand side of (4.5) is equal to $-\langle \bar{\varphi}, \partial B/\partial u(\bar{y}, \bar{u})h \rangle$. Substituting this for the first item in (4.4), we find that

$$\frac{\partial L}{\partial u}(\bar{u}, \bar{\lambda})h = \frac{\partial \mathcal{L}}{\partial u}(\bar{y}, \bar{u}, \bar{\varphi}, \bar{\lambda})h \tag{4.6}$$

for all $h \in L^\infty(X)$. If (A1) is satisfied, then we deduce from (2.7) that $d(x)$ expresses the derivative $\partial \mathcal{L}/\partial u$, i.e.,

$$\frac{\partial \mathcal{L}}{\partial u}(\bar{y}, \bar{u}, \bar{\varphi}, \bar{\lambda})h = \int_X d(x)h(x)d\mu(x). \tag{4.7}$$

Corollary 4.1. *Define J and G_j, $j = 1, \dots, m$, as above, and let $\bar{u}$ with associated state $\bar{y}$ be a local solution of* (OC). *If the regularity assumption* (2.1) *is fulfilled, then there are Lagrange multipliers $\bar{\lambda}_j$, $j = 1, \dots, m$, such that* (2.2), (2.3) *are satisfied. Assume further that $\bar{\varphi} \in Z^*$ is uniquely determined by* (4.5). *Then* (2.3) *is equivalent, with*

$$\frac{\partial \mathcal{L}}{\partial u}(\bar{y}, \bar{u}, \bar{\varphi}, \bar{\lambda})(u - \bar{u}) \geq 0 \qquad \forall u_a \leq u \leq u_b. \tag{4.8}$$

If additionally (A1) *is satisfied, then $\frac{\partial \mathcal{L}}{\partial u}(\bar{y}, \bar{u}, \bar{\varphi}, \bar{\lambda})$ can be identified with a real function $d = d(x)$, and* (4.8) *admits the form*

$$\int_X d(x)(u(x) - \bar{u}(x)) \geq 0 \qquad \forall u_a \leq u \leq u_b. \tag{4.9}$$

Proof. The statement follows from Theorem 2.1: The variational inequality (4.8) is obtained from (2.3) by (2.6) and (4.6). If (A1) is satisfied, then (4.8) and (4.7) imply (4.9). ∎

Let us now apply the second-order conditions to the control system. We have to express $\partial^2 L/\partial u^2$ in terms of $\mathcal{L}$. From

$$L(u,\lambda) = F(G(u),u) + \sum_{j=1}^{m} \lambda_j F_j(G(u),u)$$

we get, after some straightforward computations,

$$\left\{\begin{aligned} \frac{\partial^2 L}{\partial u^2}(\bar u,\bar\lambda)[h_1,h_2] = &\left[F''(\bar y,\bar u) + \sum_{j=1}^{m} \bar\lambda_j F_j''(\bar y,\bar u)\right][(y_1,h_1),(y_2,h_2)] \\ &+ \left[\frac{\partial F}{\partial y}(\bar y,\bar u) + \sum_{j=1}^{m} \bar\lambda_j \frac{\partial F_j}{\partial y}(\bar y,\bar u)\right] G''(\bar u)[h_1,h_2], \end{aligned}\right. \tag{4.10}$$

where $y_i = G'(\bar u)h_i = z_{h_i}$, $i = 1,2$. We know that $G''(\bar u)[h_1,h_2] = z_{h_1h_2}$, where $z = z_{h_1h_2}$ is the solution of (4.2); hence this term can be reduced to z_{h_1} and z_{h_2}. By definition of $\bar\varphi$, (4.2), and (4.5),

$$\left\{\begin{aligned} \left[\frac{\partial F}{\partial y} + \sum_{j=1}^{m} \bar\lambda_j \frac{\partial F_j}{\partial y}\right] z_{h_1h_2} &= \left\langle \bar\varphi, Az_{h_1h_2} + \frac{\partial B}{\partial y} z_{h_1h_2} \right\rangle \\ &= -\langle \bar\varphi, B''(\bar y,\bar u)[(z_{h_1},h_1),(z_{h_2},h_2)]\rangle \end{aligned}\right.$$

is obtained. Insert this into (4.10); then $y_i = z_{h_i}$ and $z_{h_1h_2} = G''(\bar u)[h_1,h_2]$ give

$$\left\{\begin{aligned} \frac{\partial^2 L}{\partial u^2}(\bar u,\bar\lambda)[h_1,h_2] = &\left[F''(\bar y,\bar u) + \sum_{j=1}^{m} \bar\lambda_j F_j''(\bar y,\bar u)\right][(y_1,h_1),(y_2,h_2)] \\ &- \langle \bar\varphi, B''(\bar y,\bar u)[(y_1,h_1),(y_2,h_2)]\rangle \\ &= \mathcal{L}''_{(y,u)}(\bar y,\bar u,\bar\varphi,\bar\lambda)[(y_1,h_1),(y_2,h_2)]. \end{aligned}\right. \tag{4.11}$$

Notice that in (4.11) the increments (y_i,h_i) cannot be chosen independently, since y_i and h_i are coupled through $y_i = G'(\bar u)h_i = z_{h_i}$. Hence the definition of z_{h_i} shows that the pairs $(y,h) = (y_i,h_i)$ have to solve the *linearized equation*

$$Ay + \frac{\partial B}{\partial y}(\bar y,\bar u)y + \frac{\partial B}{\partial u}(\bar y,\bar u)h = 0. \tag{4.12}$$

Corollary 4.2. *Assume that* (2.1), (A1), *and* (A2) *are satisfied and that* $\bar\varphi \in Z^*$ *is uniquely defined by* (4.5). *Then*

$$\mathcal{L}''_{(y,u)}(\bar y,\bar u,\bar\varphi,\bar\lambda)(y,h)^2 \ge 0 \tag{4.13}$$

holds for all $(y,h) \in Y \times L^\infty(X)$ *that satisfy the linearized equation* (4.12) *and the relations*

$$\left\{\begin{aligned} &\frac{\partial F_j}{\partial y}(\bar y,\bar u)y + \frac{\partial F_j}{\partial u}(\bar y,\bar u)h = 0 \quad \text{if } (j \le m_1) \\ &\qquad\qquad\qquad\qquad\qquad\qquad \text{or } (j > m_1,\ F_j(\bar y,\bar u) = 0,\ \text{and } \bar\lambda_j > 0), \\ &\frac{\partial F_j}{\partial y}(\bar y,\bar u)y + \frac{\partial F_j}{\partial u}(\bar y,\bar u)h \le 0 \quad \text{if } j > m_1,\ F_j(\bar y,\bar u) = 0,\ \text{and } \bar\lambda_j = 0, \end{aligned}\right. \tag{4.14}$$

$$h(x) = \begin{cases} \geq 0 & \text{if } \bar{u}(x) = u_a(x), \\ \leq 0 & \text{if } \bar{u}(x) = u_b(x), \end{cases} \tag{4.15}$$

$$h(x) = 0 \quad \text{if } x \in X^0. \tag{4.16}$$

The second-order sufficient optimality conditions are given by the following.

COROLLARY 4.3. *Let* $(\bar{y}, \bar{u})$ *fulfill all constraints of* (OC) *and, together with* $\bar{\varphi}$ *and* $\bar{\lambda}_j$, $j = 1, \dots, m$, *the first-order optimality conditions stated in Corollary* 4.1. *Assume that* (2.1), (A1), *and* (A3) *hold true. If there exist* $\tau > 0$, $\delta_1 > 0$, *and* $\delta_2 > 0$ *such that*

$$\mathcal{L}''_{(y,u)}(\bar{y}, \bar{u}, \bar{\varphi}, \bar{\lambda})(y, h)^2 \geq \delta_1 \|h\|^2_{L^2(X \setminus X^\tau)} - \delta_2 \|h\|^2_{L^2(X^\tau)} \tag{4.17}$$

holds for all $(y, h) \in Y \times L^\infty(X)$ *that satisfy the linearized equation* (4.12) *and the relations* (4.14), (4.15), *then the conclusions of Theorem* 3.1 *hold true; hence* $\bar{u}$ *is a local solution of* (OC). *Here, the set* X^τ *is defined by* (3.2). *The same conclusion is true if the condition*

$$\mathcal{L}''_{(y,u)}(\bar{y}, \bar{u}, \bar{\varphi}, \bar{\lambda})(y, h)^2 \geq \delta \|h\|^2_{L^2(X)} \tag{4.18}$$

holds instead of (4.17) *with some* $\delta > 0$, *where* $h(x) = 0$ *for all* $x \in X^\tau$ *for some* $\tau > 0$, *and* (y, h) *are subject to* (4.12), (4.14), *and* (4.15).

4.2. Optimal control of ODEs. In this section we discuss an optimal control problem governed by an ODE. We concentrate on a very simplified setting to give the reader an easy insight into the application of the theory. For further problems, we refer to the book by Hestenes [16]. Define

$$F(y, u) = \psi(y(T)) + \int_0^T f_0(t, y(t), u(t))dt,$$
$$F_j(y, u) = \int_0^T f_j(t, y(t), u(t))dt,$$

$j = 1, \dots, m$, and consider the optimal control problem

$$\text{(ODE)} \qquad \begin{cases} \text{minimize } F(y, u), \\ y'(t) + b(t, y(t), u(t)) = 0 & \text{a.e. } t \in (0, T), \\ y(0) = 0, \\ u_a(t) \leq u(t) \leq u_b(t) & \text{a.e. } t \in (0, T), \\ F_j(y, u) = 0, & 1 \leq j \leq m_1, \\ F_j(y, u) \leq 0, & m_1 + 1 \leq j \leq m. \end{cases}$$

Here, T is a fixed time. To reduce the number of technicalities, let us discuss only real-valued functions y and u. The vector-valued case can be handled analogously. For the same reason, we assume that the functions ψ, f_j, and b are of class C^2 on $\mathbb{R}$ and $[0, T] \times \mathbb{R} \times [\min u_a, \max u_b]$, respectively, although weaker Carathéodory-type conditions would suffice. We introduce the state space $Y = \{y \in W^{1,\infty}(0, T) | y(0) = 0\}$ and set

$$(Ay)(t) = y'(t), \qquad (B(y, u))(t) = b(t, y(t), u(t)).$$

A is continuous from Y to $Z = L^\infty(0, T)$, and B is of class C^2 from $Y \times L^\infty(0, T)$ to Z. In this way, (ODE) is related to (OC) as a particular case, where $X = [0, T]$, and μ

is the Lebesgue measure, $d\mu = dt$. For convenience, the variable $t \in X$ is substituted for the variable x, which was used in the former sections.

Let $(\bar{y}, \bar{u}) \in Y \times L^\infty(0,T)$ be our *reference solution*, a given candidate for optimality. For (ODE), the Lagrange function

$$\mathcal{L}(y,u,\varphi,\lambda) = F(y,u) - \int_0^T \varphi(y' + b(t,y,u))dt + \sum_{j=1}^m \lambda_j F_j(y,u) \tag{4.19}$$

is introduced, where $\varphi \in W^{1,\infty}(0,T)$ will be defined by the adjoint equation below. In an obvious way this φ generates a linear functional belonging to Z^*, but it has more regularity than arbitrary functionals of this space.

Remark 4.4. Given the inhomogeneous initial condition $y(0) = y_0$, we have to work with the space $Y = W^{1,\infty}(0,T)$ and must include the initial condition in the definition of A. Then the additional term $\varphi_0(y(0) - y_0)$ would appear in (4.19). This requires some more notational effort. However, the optimality conditions are not changed. Therefore, without loss of generality we confine ourselves to a homogeneous initial condition.

Having in mind the particular form of φ, we see that here (4.5) is nothing more than the definition of the *adjoint equation*

$$\begin{cases} -\varphi' + \dfrac{\partial b}{\partial y}(t,\bar{y},\bar{u})\varphi = \dfrac{\partial f_0}{\partial y}(t,\bar{y},\bar{u}) + \displaystyle\sum_{j=1}^m \bar{\lambda}_j \frac{\partial f_j}{\partial y}(t,\bar{y},\bar{u}), \\ \varphi(T) = \psi'(y(T)). \end{cases} \tag{4.20}$$

It is obvious that (4.20) admits a unique solution $\bar{\varphi} \in W^{1,\infty}(0,T)$. In section 5 we show that (A1) is satisfied for (ODE). We obtain the following derivatives of the Lagrange function:

$$\frac{\partial \mathcal{L}}{\partial u}(\bar{y},\bar{u},\bar{\varphi},\bar{\lambda})h = \int_0^T \left(\frac{\partial f_0}{\partial u} - \bar{\varphi}\frac{\partial b}{\partial u} + \sum_{j=1}^m \bar{\lambda}_j \frac{\partial f_j}{\partial u} h \right) dt \tag{4.21}$$

(all derivatives taken at $(\bar{y},\bar{u})$); hence $\partial\mathcal{L}/\partial u$ can be identified with $d \in L^\infty(0,T)$,

$$d(t) = \left(\frac{\partial f_0}{\partial u} - \bar{\varphi}\frac{\partial b}{\partial u} + \sum_{j=1}^m \bar{\lambda}_j \frac{\partial f_j}{\partial u} \right)(t). \tag{4.22}$$

The second derivative of $\mathcal{L}$ is

$$\begin{cases} \mathcal{L}''_{(y,u)}(\bar{y},\bar{u},\bar{\varphi},\bar{\lambda})[(y_1,h_1),(y_2,h_2)] = \psi''(\bar{y}(T))y_1(T)y_2(T) \\ \quad + \displaystyle\int_0^T \left\{ (y_1,h_1) \left(f_0''(\bar{y},\bar{u}) - \bar{\varphi}b''(\bar{y},\bar{u}) + \sum_{j=1}^m \bar{\lambda}_j f_j''(\bar{y},\bar{u}) \right) (y_2,h_2)^\top \right\} dt, \end{cases} \tag{4.23}$$

where f_0'', b'', f_j'' stand for 2×2 Hessian matrices taken at $(t, \bar{y}(t), \bar{u}(t))$. It is easy to verify that (A2) is satisfied.

The *first-order necessary optimality conditions* are stated in Corollary 4.1. In particular, the following variational inequality has to be satisfied:

$$\int_X d(t)(u(t) - \bar{u}(t))dt \geq 0 \tag{4.24}$$

for all $u_a \le u(t) \le u_b$; hence $\bar{u}(t) = u_a$, where $d(t) > 0$, and $\bar{u}(t) = u_b$, where $d(t) < 0$. (These points form the set X^0.) No information is obtained where d is zero. Roughly speaking, this is the set for which higher-order conditions are needed.

The *second-order necessary conditions* are formulated in Corollary 4.2. We have to specify the linearized equation (4.12) and the form of the derivatives in the relations (4.14). The *linearized equation* is

$$\begin{cases} y' + \dfrac{\partial b}{\partial y}(t,\bar{y},\bar{u})y + \dfrac{\partial b}{\partial u}(t,\bar{y},\bar{u})h = 0, \\ y(0) = 0, \end{cases} \tag{4.25}$$

while

$$\frac{\partial F_j}{\partial y}(\bar{y},\bar{u})y + \frac{\partial F_j}{\partial u}(\bar{y},\bar{u})h = \int_X \left\{ \frac{\partial f_j}{\partial y}(t,\bar{y},\bar{u})y + \frac{\partial f_j}{\partial u}(t,\bar{y},\bar{u})h \right\} dt. \tag{4.26}$$

4.3. Optimal boundary control of an elliptic equation. As a further application, we consider an elliptic control problem. For convenience, we discuss a simplified version and refer for further reading to [9].

Let $\Omega \subset \mathbb{R}^N$ be a bounded domain with boundary Γ of class $C^{0,1}$. Let ν denote the outward unit normal vector at Γ, and ∂_ν be the associated normal derivative. Define

$$F(y,u) = \int_\Omega \gamma_0(x,y(x))dx + \int_\Omega \psi_0(x,y(x))d\mu_0(x) + \int_\Gamma f_0(x,y(x),u(x))dS(x),$$
$$F_j(y,u) = \int_\Omega \gamma_j(x,y(x))dx + \int_\Omega \psi_j(x,y(x))d\mu_j(x) + \int_\Gamma f_j(x,y(x),u(x))dS(x),$$

$j = 1,\dots,m$. We assume that the functions $\gamma_j = \gamma_j(x,y)$, $\psi_j = \psi_j(x,y)$, and $f_j = f_j(x,y,u)$ are of class C^2 on $\bar{\Omega}\times\mathbb{R}$ and $\bar{\Omega}\times\mathbb{R}^2$, respectively. Moreover, real Borel measures μ_j are given on Ω. Here, μ is the Lebesgue surface measure induced on Γ, $d\mu = dS$. The appearance of the measures μ_j in the functionals will heavily influence the verification of assumptions (A1)–(A3). Therefore, the easier case $\psi_j = 0$, $j = 1,\dots,m$, is of interest as well.

Consider the optimal control problem

$$\begin{cases} \text{minimize } F(y,u), & \\ -\Delta y + y = 0 & \text{in } \Omega, \\ \partial_\nu y + b(x,y,u) = 0 & \text{on } \Gamma, \\ u_a(x) \le u(x) \le u_b(x) & \text{a.e. on } \Gamma, \\ F_j(y,u) = 0, & 1 \le j \le m_1, \\ F_j(y,u) \le 0, & m_1+1 \le j \le m. \end{cases} \tag{ELL}$$

In this setting, the *boundary control* u is looked upon in the space $L^\infty(\Gamma)$, hence $X = \Gamma$, while the *state* y belongs to $Y = \{y \in H^1(\Omega) | -\Delta y + y \in L^q(\Omega), \partial_\nu y \in L^p(\Gamma)\}$. (Here $q > N/2$ and $p > N-1$ are given fixed.) Endowing Y with the graph norm, it is known that $Y \subset C(\bar{\Omega})$, the embedding being continuous. Assume that $b = b(x,y,u)$ satisfies the same conditions as the f_j. Additionally, we require that $(\partial b/\partial y)(x,y,u) \ge 0$ on $\Gamma\times\mathbb{R}\times[\min u_a, \max u_b]$. Define

$$A : Y \to L^q(\Omega)\times L^p(\Gamma) \quad \text{and} \quad B : Y\times L^\infty(\Gamma) \to L^q(\Omega)\times L^p(\Gamma)$$

by

$$(Ay) = \begin{pmatrix} -\Delta y + y \\ \partial_\nu y \end{pmatrix} \quad \text{and} \quad B(y,u)(x) = \begin{pmatrix} 0 \\ b(x, y(x), u(x)) \end{pmatrix}.$$

The equation $Ay + B(y,u) = 0$, which is equivalent to our elliptic boundary value problem, admits for each $u \in L^\infty(\Gamma)$ exactly one solution $y \in Y$. The mapping $u \mapsto y$ is of class C^2 from $L^\infty(\Gamma)$ to Y. Now we proceed in the same way as in the preceding section. The Lagrange function is

$$\begin{aligned}\mathcal{L}(y,u,\varphi,\lambda) = F(y,u) &- \int_\Omega (-\Delta y + y)\varphi dx \\ &- \int_\Gamma (\partial_\nu y + b(x,y,u))\varphi dS + \sum_{j=1}^m \lambda_j F_j(y,u),\end{aligned}$$

where $\varphi \in W^{1,s}(\Omega)$ for all $s < \frac{N}{N-1}$ is the *adjoint state*. The adjoint state φ together with its trace $\varphi_{|\Gamma}$ forms a Lagrange multiplier of $Z^* = L^{q'}(\Omega) \times L^{p'}(\Gamma)$ having higher regularity. Here (4.5) reduces to the adjoint equation

$$\begin{cases} -\Delta\varphi + \varphi = \dfrac{\partial \gamma_0}{\partial y} + \dfrac{\partial \psi_0}{\partial y}\mu_0|_\Omega + \displaystyle\sum_{j=1}^m \bar\lambda_j \left(\frac{\partial \gamma_j}{\partial y} + \frac{\partial \psi_j}{\partial y}\mu_j|_\Omega \right), \\ \partial_\nu\varphi + \dfrac{\partial b}{\partial y}\varphi = \dfrac{\partial f_0}{\partial y} + \displaystyle\sum_{j=1}^m \bar\lambda_j \frac{\partial f_j}{\partial y} + \frac{\partial \psi_0}{\partial y}\mu_0|_\Gamma + \sum_{j=1}^m \bar\lambda_j \frac{\partial \psi_j}{\partial y}\mu_j|_\Gamma \end{cases}$$

(all partial derivatives taken at $(x, \bar y(x), \bar u(x))$). This equation has a unique solution $\bar\varphi \in W^{1,s}(\Omega)$ associated with $(\bar y, \bar u, \bar\lambda)$. Notice that for $N = 2$ the Sobolev imbedding theorem yields $\varphi \in L^\sigma(\Omega)$ for all $\sigma < \infty$, but not in general $\varphi \in L^\infty(\Omega)$. For $N \geq 3$ the regularity of φ is even lower. This indicates that we have to discuss assumptions (A1)–(A3) with more care. We shall do this in the last section.

The situation is easier in the case $\psi_j = 0$, $j = 0, \dots, m$. Then all data given in the adjoint equation are bounded and measurable, and the regularity theory of elliptic equations yields $\bar\varphi \in C(\bar\Omega)$ (see [5]).

Let us establish the first- and second-order derivatives of $\mathcal{L}$. We get

$$\frac{\partial \mathcal{L}}{\partial u}(\bar y, \bar u, \bar\varphi, \bar\lambda)h = \int_\Gamma \left(\frac{\partial f_0}{\partial u}(x, \bar y, \bar u) + \sum_{j=1}^m \bar\lambda_j \frac{\partial f_j}{\partial u}(x, \bar y, \bar u) - \bar\varphi \frac{\partial b}{\partial u}(x, \bar y, \bar u) \right) h dS$$

and

$$\begin{aligned}&\mathcal{L}''_{(y,u)}(\bar y, \bar u, \bar\varphi, \bar\lambda)[(y_1,h_1),(y_2,h_2)] \\ &\quad = \int_\Gamma (y_1, h_1) \left(f_0''(x, \bar y, \bar u) + \sum_{j=1}^m \bar\lambda_j f_j''(x, \bar y, \bar u) - \bar\varphi b''(x, \bar y, \bar u) \right) (y_2, h_2)^\top dS \\ &\qquad + \int_\Omega \left(\frac{\partial^2 \gamma_0}{\partial y^2}(x, \bar y) + \sum_{j=1}^m \bar\lambda_j \frac{\partial^2 \gamma_j}{\partial y^2}(x, \bar y) \right) y_1 y_2 dx \\ &\qquad + \int_\Omega \frac{\partial^2 \psi_0}{\partial y^2}(x, \bar y) y_1 y_2 d\mu_0 + \sum_{j=1}^m \bar\lambda_j \frac{\partial^2 \psi_j}{\partial y^2}(x, \bar y) y_1 y_2 d\mu_j.\end{aligned}$$

We observe that, due to our notation, there is almost no difference in the expressions derived for the case of (ODE) in (4.21), (4.23). The first- and second-order conditions for our elliptic problem (ELL) admit the following form: Set

$$d(x) = \frac{\partial f_0}{\partial u}(x, \bar{y}(x), \bar{u}(x)) + \sum_{j=1}^{m} \bar{\lambda}_j \frac{\partial f_j}{\partial u}(x, \bar{y}(x), \bar{u}(x)) - \bar{\varphi}\frac{\partial b}{\partial u}(x, \bar{y}(x), \bar{u}(x)).$$

Then d has the same form as in (4.22). The first- and second-order optimality conditions are given by Corollaries 4.1–4.3. There we set $X = \Gamma$ to obtain all first- and second-order conditions for (ELL). Now the directions (y, h) are coupled through the *linearized boundary value problem*

$$\begin{cases} -\Delta y + y = 0, \\ \partial_\nu y + \dfrac{\partial b}{\partial y}(x, \bar{y}, \bar{u})y + \dfrac{\partial b}{\partial u}(x, \bar{y}, \bar{u})h = 0. \end{cases} \tag{4.27}$$

The derivatives in (4.14), (4.15) admit the form

$$\begin{cases} \dfrac{\partial F_j}{\partial y}(\bar{y}, \bar{u})y + \dfrac{\partial F_j}{\partial u}(\bar{y}, \bar{u})h = \displaystyle\int_\Omega \frac{\partial \gamma_j}{\partial y}(t, \bar{y})y dx + \int_\Omega \frac{\partial \psi_j}{\partial y}(t, \bar{y})y d\mu_j \\ \qquad\qquad + \displaystyle\int_\Gamma \left\{ \frac{\partial f_j}{\partial y}(t, \bar{y}, \bar{u})y + \frac{\partial f_j}{\partial u}(t, \bar{y}, \bar{u})h \right\} dS. \end{cases} \tag{4.28}$$

In this way, we have obtained the second-order sufficient condition for a simplified elliptic control problem. For the discussion of more general problems, we refer to [7], [9]. We should underline again that so far we have stated the optimality condition in a formal way. It remains to verify (A1)–(A3) to make our theory work. Low regularity of the adjoint state φ can be an essential obstacle for this. We refer to section 5.

4.4. Optimal distributed control of a parabolic equation. We confine ourselves to a distributed parabolic control problem. A more general class, including boundary control and boundary observation, is considered in a separate paper by Raymond and Tröltzsch [22]. Let Ω be defined as in the last section, and set $Q = \Omega \times (0, T)$, $\Sigma = \Gamma \times (0, T)$. Define

$$\begin{aligned} F(y, u) &= \int_\Omega \gamma_0(x, y(x, T))dx + \int_\Omega \psi_0(x, y(x, T))d\mu_0(x) \\ &\quad + \int_Q f_0(x, t, y(x, t), u(x, t))dxdt, \\ F_j(y, u) &= \int_Q \psi_j(x, t, y(x, t))d\mu_j(x, t) + \int_Q f_j(x, t, y(x, t), u(x, t))dxdt, \end{aligned}$$

$j = 1, \ldots, m$. We assume again that the functions ψ_j, f_j, and γ_j are of class C^2 on $\bar{Q} \times \mathbb{R}$ and $\bar{Q} \times \mathbb{R}^2$, respectively. Moreover, real Borel measures $\mu_j, j = 0, \ldots, m$, are given on Ω and Q, respectively. Now μ is the Lebesgue measure on Q, $d\mu = dxdt$.

Consider the optimal control problem

$$\text{(PAR)} \qquad \begin{cases} \text{minimize } F(y,u), & \\ \dfrac{\partial y}{\partial t} - \Delta y + b(x,t,y,u) = 0 & \text{in } Q, \\ \partial_\nu y = 0 & \text{on } \Sigma, \\ y(x,0) = 0 & \text{in } \Omega, \\ u_a(x,t) \le u(x,t) \le u_b(x,t) & \text{a.e. on } Q, \\ F_j(y,u) = 0, & 1 \le j \le m_1, \\ F_j(y,u) \le 0, & m_1 + 1 \le j \le m. \end{cases}$$

In this setting, the *distributed control* u is looked upon in the space $L^\infty(Q)$; hence we set $X = Q$. The *state* y belongs to $Y = \{y \in W(0,T) | y(0) = 0, y_t - \Delta y \in L^q(Q), \partial_\nu y \in L^p(\Sigma)\}$, where $q > N/2 + 1$ and $p > N + 1$ are given fixed. It is known that $Y \subset C(\bar{Q})$, the embedding being continuous for the graph norm. Assume that $b = b(x,t,y,u)$ satisfies the same conditions as the f_j. Additionally, we require that $\partial b/\partial y(x,t,y,u) \ge 0$ on $Q \times \mathbb{R} \times [\min u_a, \max u_b]$. Define

$$A : Y \to L^q(Q) \times L^p(\Sigma) \quad \text{and} \quad B : Y \times L^\infty(Q) \to L^q(Q) \times L^p(\Sigma)$$

by

$$Ay = \begin{pmatrix} \dfrac{\partial y}{\partial t} - \Delta y \\ \partial_\nu y \end{pmatrix} \quad \text{and} \quad B(y,u)(x,t) = \begin{pmatrix} b(x,t,y(x,t),u(x,t)) \\ 0 \end{pmatrix}.$$

The equation $Ay + B(y,u) = 0$, which is equivalent to our parabolic initial-boundary value problem, admits for each $u \in L^\infty(Q)$ exactly one solution $y \in Y$. We refer to [5]. The mapping $u \mapsto y$ is of class C^2 from $L^\infty(Q)$ to Y. Here, the Lagrange function is

$$\begin{aligned} \mathcal{L}(y,u,\varphi,\lambda) = F(y,u) &- \int_Q (y_t - \Delta y - b(x,t,y,u))\varphi dx dt \\ &- \int_\Sigma \partial_\nu y \varphi dS dt + \sum_{j=1}^m \lambda_j F_j(y,u), \end{aligned}$$

where φ is the *adjoint state* and dS again denotes the Lebesgue surface measure induced on Γ. Equation (4.5) turns out to be the adjoint equation

$$\begin{cases} -\dfrac{\partial \varphi}{\partial t} - \Delta \varphi + \dfrac{\partial b}{\partial y}\varphi = \dfrac{\partial f_0}{\partial y} + \displaystyle\sum_{j=1}^m \bar{\lambda}_j \left(\frac{\partial f_j}{\partial y} + \frac{\partial \psi_j}{\partial y}\mu_j \right) & \text{in } Q, \\ \partial_\nu \varphi = 0 & \text{in } \Sigma, \\ \varphi(x,T) = \dfrac{\partial \gamma_0}{\partial y}(x, \bar{y}(x,T)) + \dfrac{\partial \psi_0}{\partial y}(x, \bar{y}(x,T))\mu_0 & \text{in } \Omega \end{cases}$$

(all partial derivatives taken at $(x, \bar{y}, \bar{u})$). This equation has a unique solution $\bar{\varphi} \in W^{1,s}(\Omega)$ associated with $(\bar{y}, \bar{u}, \bar{\varphi}, \bar{\lambda})$. If, however, $\psi_j = 0$, $j = 1, \dots, m$, then $\bar{\varphi}$ is more regular, $\bar{\varphi} \in W(0,T) \cap C(\bar{Q})$.

The relevant derivatives of $\mathcal{L}$ are

$$\begin{aligned}
&\frac{\partial \mathcal{L}}{\partial u}(\bar y,\bar u,\bar\varphi,\bar\lambda)h\\
&\quad=\int_Q\left[\frac{\partial f_0}{\partial u}(x,\bar y,\bar u)+\sum_{j=1}^m\bar\lambda_j\frac{\partial f_j}{\partial u}(x,\bar y,\bar u)-\bar\varphi\frac{\partial b}{\partial u}(x,\bar y,\bar u)\right]h\,dxdt\\
&\quad=\int_Q d(x,t)h(x,t)\,dxdt,\\
&\mathcal{L}''_{(y,u)}(\bar y,\bar u,\bar\varphi,\bar\lambda)[(y_1,h_1),(y_2,h_2)]\\
&\quad=\int_Q(y_1,h_1)\left[f_0''(x,\bar y,\bar u)+\sum_{j=1}^m\bar\lambda_j f_j''(x,\bar y,\bar u)-\bar\varphi b''(x,\bar y,\bar u)\right](y_2,h_2)^\top dxdt\\
&\qquad+\int_\Omega\frac{\partial^2\psi_0}{\partial y^2}(x,\bar y(T))y_1(T)y_2(T)d\mu_0+\int_Q\sum_{j=1}^m\bar\lambda_j\frac{\partial^2\psi_j}{\partial y^2}(x,\bar y)y_1y_2d\mu_j\\
&\qquad+\int_\Omega\frac{\partial^2\gamma_0}{\partial y^2}(x,\bar y(T))y_1(T)y_2(T)dx.
\end{aligned}$$

The first- and second-order conditions for the parabolic case are covered by Corollaries 4.1–4.3. We have to substitute Q for X there and replace the variable x by (x,t). Moreover, in the second-order conditions, y and h are coupled through the *linearized initial-boundary value problem*

$$\begin{cases} y_t-\Delta y+\dfrac{\partial b}{\partial y}(x,t,\bar y,\bar u)y+\dfrac{\partial b}{\partial u}(x,t,\bar y,\bar u)h=0,\\ \partial_\nu y=0,\\ y(x,0)=0. \end{cases}\tag{4.29}$$

We leave the calculations of the derivatives in (4.14) to the reader; they are obtained by an obvious modification of (4.28). We should mention again that these optimality conditions are meaningful only if the assumptions (A1)–(A3) are satisfied.

5. Verification of the assumptions. Our theory relies on the general assumptions (A1)–(A3). We shall see that (A1)–(A3) are naturally satisfied for the problem (ODE), while the situation is more complicated in the case of the elliptic or parabolic PDE.

(i) Problem (ODE). (A1). It is obviously sufficient to look at one of the functionals $G_j(u)=F_j(G(u),u)$ to assess the situation. We have

$$G_j'(\bar u)h=\int_0^T\frac{\partial f_j}{\partial y}(t,\bar y,\bar u)y\,dt+\int_0^T\frac{\partial f_j}{\partial u}(t,\bar y,\bar u)h\,dt,\tag{5.1}$$

where $y=G'(\bar u)h$. Here, $\partial f_j/\partial y$, $\partial f_j/\partial u$ are bounded and measurable functions. Moreover, the estimate

$$\|y\|_{C[0,T]}=\|G'(\bar u)h\|_{C[0,T]}\le c\|h\|_{L^2(0,T)}\tag{5.2}$$

holds, since $\|y\|_{C[0,T]}\le c\|y\|_{H^1(0,T)}\le c\|h\|_{L^2(0,T)}$. Thus the mapping $h\mapsto G_j'(\bar u)h$ defines a linear and continuous functional on $L^2(0,T)$. By the Riesz representation

theorem,

$$G_j'(\bar u)h = \int_0^T g_j(t)h(t)dt \tag{5.3}$$

must hold with some $g_j \in L^2(0,T)$; hence (A1) is fulfilled.

(A2). Here, the derivative

$$G_j''(\bar u)[h_1,h_2] = \int_0^T (y_1,h_1)f_j''(t,\bar y,\bar u)(y_2,h_2)^\top dt$$

is characteristic for the discussion. All entries of f_j'' are bounded and measurable. If $h_i^k \to h_i$ in $L^2(0,T)$, $k\to\infty, i=1,2$, then $y_i^k \to y_i$ in $C[0,T]$; hence $G_j''(\bar u)[h_1^k,h_2^k] \to G_j''(\bar u)[h_1,h_2]$. This shows (A2).

(A3). First, we must estimate differences of the type $G_j''(\tilde u) - G_j''(\bar u)$ for $\tilde u$ in a L^∞-neighborhood of $\bar u$. We get

$$|(G_j''(\tilde u) - G_j''(\bar u))h^2| \le \int_0^T |f_j''(t,\tilde y,\tilde u) - f_j''(t,\bar y,\bar u)||(y,h)|^2 dt,$$

where $\tilde y = G(\tilde u), \bar y = G(\bar u), y = G'(\bar u)h$. Due to our assumptions, we find that

$$|[G_j''(\tilde u) - G_j''(\bar u)]h^2| \le \delta(\|y\|^2_{C[0,T]} + \|h\|^2_{L^2(0,T)}) \le c\delta\|h\|^2_{L^2(0,T)}, \tag{5.4}$$

where $\delta \to 0$ as $\|\tilde u - \bar u\|_{L^\infty} \to 0$. Another characteristic part in $\partial^2 L/\partial u^2$ is the coupling of the nonlinearity b with $\bar\varphi$. It is the essential advantage of our simplified case (ODE) that $\bar\varphi \in L^\infty(0,T)$. Therefore, we are justified to estimate

$$\begin{aligned}\left|\int_0^T (y,h)b''(t,\bar y,\bar u)(y,h)^\top \bar\varphi dt\right| &\le c\|\bar\varphi\|_{L^\infty(0,T)}(\|y\|^2_{C[0,T]} + \|h\|^2_{L^2(0,T)}) \\ &\le c\|h\|^2_{L^2(0,T)}.\end{aligned} \tag{5.5}$$

Discussing all second-order terms in this way, we easily verify that (A3) is also satisfied.

(ii) Elliptic problem (ELL). We repeat the discussion of (A1)–(A3) along the lines of (i) but concentrating on the essential differences with the case of (ODE). Here, it holds that

$$\begin{aligned}G_j'(\bar u)h = &\int_\Omega \frac{\partial\gamma_j}{\partial y}(x,\bar y)y dx + \int_\Omega \frac{\partial\psi_j}{\partial y}(x,\bar y)y d\mu_j \\ &+ \int_\Gamma \frac{\partial f_j}{\partial y}(x,\bar y,\bar u)y dS + \int_\Gamma \frac{\partial f_j}{\partial u}(x,\bar y,\bar u)h dS,\end{aligned}$$

where $y = G'(\bar u)h$. In contrast to (5.2), now the mapping $G'(\bar u)$ is not in general continuous from $L^2(\Gamma)$ to $C(\bar\Omega)$. This property only holds for $N = \dim\Omega = 2$ (see [9]). For $N>2$ we assume that Ω_j, the support of μ_j, satisfies $\bar\Omega_j \subset \Omega$. Then the mapping $h \mapsto G'(\bar u)h$ is continuous from $L^2(\Gamma)$ to $C(\bar\Omega_j)$; hence $h \mapsto G_j'(\bar u)h$ is a linear and continuous functional on $L^2(\Gamma)$. The Riesz theorem yields a representation analogous to (5.3). Hence (A1) is shown under additional assumptions on the subdomains Ω_j. (A2) then holds true in the same way. Notice that the restriction to Ω_j is not needed if all ψ_j vanish.

To verify (A3) we need even more restrictions on the data. The situation is easy if $\psi_j = 0$, $j = 1, \dots, m$. Then all given data in the adjoint equation are bounded and measurable, and the regularity theory of elliptic equations yields $\bar\varphi \in C(\bar\Omega)$. In this case, (A3) is obviously satisfied.

Let us now assume that at least one of the ψ_j is not zero. Then the best regularity of the trace $\bar\varphi_{|\Gamma}$ is $\bar\varphi_{|\Gamma} \in L^r(\Gamma)$ for all $r < (N-1)/(N-2)$. For instance, $\varphi \in L^r(\Gamma)$ for all $r < \infty$ is obtained in the case $N = 2$. We therefore cannot assume that $\bar\varphi \in L^\infty(\Omega)$. Regard the elliptic counterpart to (5.5),

$$\begin{aligned}\left|\int_\Gamma (y,h)b''(x,\bar y,\bar u)(y,h)^\top \bar\varphi dS\right| &= \left|\int_\Gamma \bar\varphi\left(\frac{\partial^2 b}{\partial y^2}y^2 + 2\frac{\partial^2 b}{\partial y\partial u}yh + \frac{\partial^2 b}{\partial u^2}h^2\right)dS\right| \\ &\le c\int_\Gamma (|\bar\varphi|y^2 + |\bar\varphi|yh + |\bar\varphi|h^2)dS.\end{aligned} \tag{5.6}$$

This expression has to be estimated for $h \in L^2(\Gamma)$. If $\bar\varphi_{|\Gamma} \notin L^\infty(\Gamma)$, which is the normal case, then we must exclude the third term from (5.6). This means that $\partial^2 b/\partial u^2$ has to disappear—u must appear linearly. Next we consider the second term, where $\|\bar\varphi_{|\Gamma} y\|_{L^2(\Gamma)}$ is estimated against $\|h\|_{L^2(\Gamma)}$. The mapping $h \mapsto y$ is continuous from $L^2(\Gamma)$ to $C(\Gamma)$ ($N = 2$), to $L^r(\Gamma)$ for all $r < \infty$ ($N = 3$), and to $L^r(\Gamma)$ for all $r < 2(N-1)/(N-3)$ ($N > 3$). Therefore, the second term can be estimated iff $N = 2$, while it must be cancelled for $N > 2$. The latter means $\partial^2 b/\partial u\partial y = 0$—here $b = b_1(x,y) + b_2(x)u$ must hold. In the same way we arrive at the surprising fact that for $N > 3$ the first term in (5.6) must vanish, too. In other words, in the case of elliptic boundary control with *pointwise* functionals F_j, we cannot admit nonlinear equations for $N > 3$.

Remark 5.1. We should underline again that these restrictions are not needed if the functionals F_j are sufficiently regular ($\psi_j = 0$, $j = 1, \dots, m$). Moreover, the case of distributed controls permits us to slightly relax the restrictions on the dimension N.

(iii) Parabolic problem (PAR). Once again, (A1)–(A3) are satisfied if $\psi_j = 0$, $j = 1, \dots, m$. This is due to the high regularity $\bar\varphi \in W(0,T) \cap C(\bar Q)$ in this case.

In the opposite case, the problem of regularity is even more delicate than in the elliptic problem. We cannot discuss the general case in detail and refer to the recent paper [22]. Instead of this, let us explain the point for a very particular constraint: Suppose that only one (pointwise) state constraint of the form

$$g_1(y,u) = \int_0^T y(x_1,t)dt = 0$$

is given, where $x_1 \in \Omega$ is a fixed position of observation. To make the theory work, we need some strong restrictions: We assume $N = \dim \Omega = 1$, i.e., $\Omega = (a,b)$, and require that $\partial^2 b/\partial u^2 = 0$ (the control appears linearly). Then the mapping $h \mapsto y = G'(\bar u)h$ is continuous from $L^2(Q)$ to $C(\bar Q)$, and the functional $h \mapsto g_1(y,h)$ is continuous on $L^2(Q)$. We know that $\bar\varphi \in L^s(Q)$ for all $s < 3$. (This follows from Theorem 4.3 in [22] for $N = 1$ and $\alpha = \tilde\alpha$.) Hence $\bar\varphi \notin L^\infty(Q)$, and that is the reason why we cannot admit a control appearing nonlinearly. The estimate of the parabolic counterpart of (5.6) is

$$\begin{aligned}&\left|\int_Q \left(\frac{\partial^2 b}{\partial y^2}\bar\varphi y^2 + 2\frac{\partial^2 b}{\partial y\partial u}\bar\varphi yh\right)dxdt\right| \\ &\quad\le c\|\bar\varphi\|_{L^1(Q)}\|y\|^2_{L^\infty(Q)} + c\|\bar\varphi\|_{L^2(Q)}\|y\|_{L^\infty(Q)}\|h\|_{L^2(Q)} \le c\|h\|^2_{L^2(Q)}.\end{aligned}$$

Discussions of this type reveal that (A1)–(A3) are satisfied. However, we needed very strong assumptions, in particular $N = 1$. The case $N = 2$ can be handled under additional restrictions concerning the appearance of control and observations ("control and observations have disjoint supports"; see [22]).

If there are no pointwise state constraints, the situation is easier, as the reader can check.

Remark 5.2. The second-order conditions established in the previous sections allow us to study L^∞-local solutions. This causes specific difficulties if the optimal control exhibits jumps. Therefore, L^p-optimality conditions can be more interesting. An associated extension to L^p is possible, provided that the control-state mapping $u \mapsto y$ and the objective functional are differentiable from L^p to L^∞. Under associated restrictions (for instance, that the control appear linearly in the state equation and the cost functional be quadratic with respect to the control), this extension to L^p is possible for sufficiently large $p < \infty$. For some associated results we refer the reader to Casas, Tröltzsch, and Unger [8] and Dunn [14].

Remark 5.3. For some optimal control problems, the second-order condition

$$\frac{\partial^2 L}{\partial u^2}(\bar{u}, \bar{\lambda})h^2 > 0 \quad \forall h \in C_{\bar{u}}^0 \setminus \{0\},$$

along with a certain positivity of the second derivative with respect to the control of the Hamiltonian, provide sufficient optimality conditions. The reader is referred to Casas and Mateos [6], where these conditions are proved to be sufficient and equivalent to (3.26); see also Bonnans and Zidani [4]. In particular, if the control appears linearly in the state equation and the cost functional is quadratic and positive with respect to the control, then the above condition is sufficient for optimality.

REFERENCES

[1] N. Arada, E. Casas, and F. Tröltzsch, *Error estimates for a semilinear elliptic control problem*, Comput. Optim. Appl., to appear.

[2] W. Alt and K. Malanowski, *The Lagrange-Newton method for nonlinear optimal control problems*, Comput. Optim. Appl., 2 (1993), pp. 77–100.

[3] J. Bonnans and E. Casas, *Contrôle de systèmes elliptiques semilinéaires comportant des contraintes sur l'état*, in Nonlinear Partial Differential Equations and Their Applications. Collège de France Seminar, Vol. 8, H. Brezis and J. Lions, eds., Longman Scientific and Technical, New York, 1988, pp. 69–86.

[4] J.F. Bonnans and H. Zidani, *Optimal control problems with partially polyhedric constraints*, SIAM J. Control Optim., 37 (1999), pp. 1726–1741.

[5] E. Casas, *Pontryagin's principle for state-constrained boundary control problems of semilinear parabolic equations*, J. Anal. Appl., 15 (1996), pp. 687–707.

[6] E. Casas and M. Mateos, *Second order optimality conditions for semilinear elliptic control problems with finitely many state constraints*, SIAM J. Control Optim., 40 (2002), pp. 1431–1454.

[7] E. Casas and F. Tröltzsch, *Second order necessary optimality conditions for some state-constrained control problems of semilinear elliptic equations*, Appl. Math. Optim., 39 (1999), pp. 211–227.

[8] E. Casas, F. Tröltzsch, and A. Unger, *Second order sufficient optimality conditions for a nonlinear elliptic control problem*, J. Anal. Appl., 15 (1996), pp. 687–707.

[9] E. Casas, F. Tröltzsch, and A. Unger, *Second order sufficient optimality conditions for some state-constrained control problems of semilinear elliptic equations*, SIAM J. Control Optim., 38 (2000), pp. 1369–1391.

[10] F. Clarke, *A new approach to Lagrange multipliers*, Math. Oper. Res., 1 (1976), pp. 165–174.

[11] A.L. Dontchev, W.W. Hager, A.B. Poore, and B. Yang, *Optimality, stability, and convergence in nonlinear control*, Appl. Math. Optim., 31 (1995), pp. 297–326.

[12] J.C. Dunn, *Second-order optimality conditions in sets of L^∞ functions with range in a polyhedron*, SIAM J. Control Optim., 33 (1995), pp. 1603–1635.
[13] J. Dunn, *On second-order sufficient optimality conditions for structured nonlinear programs in infinite-dimensional function spaces*, in Mathematical Programming with Data Perturbations, A. Fiacco, ed., Marcel Dekker, New York, 1998, pp. 83–107.
[14] J.C. Dunn, $\mathbb{L}^2$ *sufficient conditions for end-constrained optimal control problems with inputs in a polyhedron*, SIAM J. Control Optim., 36 (1998), pp. 1833–1851.
[15] W.W. Hager, *Error bounds for Euler approximation of a state and control constrained optimal control problem*, Numer. Funct. Anal. Optim. 21 (2000), pp. 653–682.
[16] M.R. Hestenes, *Calculus of Variations and Optimal Control Theory*, John Wiley, New York, 1966.
[17] 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.
[18] K. Ito and K. Kunisch, *Augmented Lagrangian–SQP methods for nonlinear optimal control problems of tracking type*, SIAM J. Control Optim., 34 (1996), pp. 874–891.
[19] H. Maurer, *First and second-order sufficient optimality conditions in mathematical programming and optimal control*, Math. Programming Study, 14 (1981), pp. 163–177.
[20] H. Maurer and J. Zowe, *First- and second-order conditions in infinite-dimensional programming problems*, Math. Programming, 16 (1979), pp. 98–110.
[21] H.D. Mittelmann, *Verification of second-order sufficient optimality conditions for semilinear elliptic and parabolic control problems*, Comput. Optim. Appl., 20 (2001), pp. 93–110.
[22] J.-P. Raymond and F. Tröltzsch, *Second order sufficient optimality conditions for nonlinear parabolic control problems with state-constraints*, Discrete Contin. Dynam. Systems, 6 (2000), pp. 431–450.
[23] V. Schulz, ed., *SQP Based Direct Discretization Methods for Practical Optimal Control Problems*, J. Comput. Appl. Math., special issue, 120 (2000).

SIAM J. OPTIM.
Vol. 13, No. 2, pp. 432–469

CONVERGENT INFEASIBLE INTERIOR-POINT TRUST-REGION METHODS FOR CONSTRAINED MINIMIZATION*

PAUL TSENG†

Abstract. We study an infeasible primal-dual interior-point trust-region method for constrained minimization. This method uses a log-barrier function for the slack variables and updates the slack variables using second-order correction. We show that if a certain set containing the initial iterate is bounded and the origin is not in the convex hull of the nearly active constraint gradients everywhere on this set, then the iterates remain in this set, and any cluster point of the iterates is a first-order stationary point. Moreover, any subsequence of iterates converging to the cluster point has an asymptotic second-order stationarity property, which, when the constraint functions are affine or when the active constraint gradients are linearly independent, implies that the cluster point is a second-order stationary point. Preliminary numerical experience with the method is reported. A primal method and its extension to semidefinite nonlinear programming is also discussed.

Key words. nonlinear program, log-barrier function, interior-point method, trust-region strategy, first- and second-order stationary points, semidefinite programming

AMS subject classifications. 49M30, 49M37, 65K10, 90C22, 90C26, 90C30, 90C51

PII. S1052623499357945

1. Introduction.

We consider the nonlinear program with inequality constraints:

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & g(x) = [g_1(x) \ \cdots \ g_m(x)]^T \le 0, \end{array} \tag{1.1}$$

where f and $g_1, \ldots, g_m$ ($m \ge 0$) are real-valued and twice continuously differentiable functions defined on $\Re^n$. We define $\nabla g = [\nabla g_1 \ \cdots \ \nabla g_m]$, where ∇g_i denotes the gradient of g_i. For simplicity, we consider only inequality constraints for now. Extensions to incorporate equality constraints and to treat semidefinite nonlinear programs are discussed in sections 8 and 9, respectively.

Define the Lagrangian function

$$l(x, \lambda) := f(x) + g(x)^T \lambda.$$

We say that an $x \in \Re^n$ is a *first-order stationary point* of (1.1) if it satisfies, together with some $\lambda \in \Re^m$ (Lagrange multipliers), the first-order necessary optimality condition for (1.1)

$$g(x) \le 0, \quad \lambda \ge 0, \quad g(x)^T \lambda = 0, \quad \nabla_x l(x, \lambda) = 0, \tag{1.2}$$

and x is a *second-order stationary point* if, in addition, it satisfies, together with λ, the *weak* second-order necessary optimality condition (see, e.g., [3, 18, 26])

$$d^T \nabla_{xx} l(x, \lambda) d \ge 0 \quad \forall d, \qquad \text{with } \nabla g_i(x)^T d = 0 \quad \forall i \in I(x), \tag{1.3}$$

where $I(x) := \{i \in \{1, \ldots, m\} : g_i(x) = 0\}$.

*Received by the editors June 15, 1999; accepted for publication (in revised form) March 4, 2002; published electronically September 24, 2002. This research was supported by National Science Foundation grant CCR-9731273.

http://www.siam.org/journals/siopt/13-2/35794.html

†Department of Mathematics, University of Washington, Seattle, WA 98195 (tseng@math.washington.edu).

A well-known approach to solving (1.1) entails introducing the slack variables $s := -g(x)$ and a logarithmic barrier for the nonnegativity constraints on s to obtain the log-barrier problem:

$$\begin{aligned} &\text{minimize} && f^{\mu}(x,s) := f(x) - \mu \sum_{i=1}^{m} \ln(s_i) \\ &\text{subject to} && g(x) + s = 0, \quad s > 0, \end{aligned} \tag{1.4}$$

where $\mu > 0$ is the barrier parameter. By eliminating s, this may also be written in terms of x only. Then, for a given $\mu > 0$, we solve problem (1.4) inexactly, and then μ is decreased, and we repeat, etc. The asymptotic property of the *exact* local optimal solution of (1.4) as $\mu \to 0$ has been studied by Fiacco and McCormick in their well-known book [19]. Auslender [3] showed that, for penalty approaches in general, a second-order stationary point of the penalized problem approaches in the limit a second-order stationary point of the original problem. Since the work of Karmarkar, interest in the log-barrier approach has been renewed, and extensive studies are made in the cases of linear/quadratic/convex programs and monotone complementarity problems (see, e.g., [34, 41] and references therein).

Recently, there has been considerable interest in extending the above barrier/interior-point approach to the nonconvex case. One such extension, giving rise to the (infeasible) primal-dual methods, entails taking one or two damped Newton steps on a reformulation of the first-order optimality condition for (1.4), and then decreasing μ, and so on. These methods and their global/local convergence were studied by El-Bakry et al. [17], Yamashita [42], Yamashita and Yabe [43], and Akrotirianakis and Rustem [1] (see also [37] for a feasible method). Forsgren and Gill [22] and Gay, Overton, and Wright [23] also studied implementation issues for these methods, including properties of the Newton direction, modified Newton directions (based on adding a suitable positive semidefinite matrix to $\nabla_{xx} l(x,\lambda)$), techniques for calculating the directions, and merit functions for stepsize selection. Vanderbei and Shanno [39] considered a modified Newton direction based on adding a nonnegative multiple of the identity matrix to $\nabla_{xx} l(x,\lambda)$ (see also (2.23)). Promising numerical results with these methods were reported in [23, 39, 42]. Numerical comparison of a primal-dual method, a primal-dual trust-region method, and a primal method on sparse problems was given by Lasdon, Plummer, and Yu [29]. For some methods, local superlinear convergence can also be shown under suitable assumptions. However, as Newton directions may not be defined everywhere, global convergence of methods using Newton directions is difficult to obtain and requires fairly strong assumptions such as positive definiteness or, at least, nonsingularity of $\nabla_{xx} l(x,\lambda)$ and linear independence of $\nabla g_i(x)$, $i \in I(x)$, globally [1, 17, 37]. In addition, only convergence to a first-order stationary point of (1.1) was shown in these references. In [42], only convergence to a first-order stationary point of (1.4) was shown (see also [44] for related results using an ℓ_2 log-barrier function). Conn, Gould, and Toint [13] studied an infeasible primal-dual method for the case of linear constraints. Their method uses modified Newton directions, and global convergence to first-order stationary point was shown under reasonable assumptions. Numerical results on quadratic programs from the CUTE test set were also reported. In a series of papers [6, 7, 8], Byrd, Nocedal, and coworkers proposed methods that combine interior-point approaches, trust-region strategies, and sequential quadratic programming (SQP) techniques. Global convergence to first-order stationary points and local superlinear convergence were studied in, respectively, [6] and [8]. Yamashita, Yabe, and Tanabe [45] proposed a primal-dual interior-point trust-region method and,

under certain assumptions, showed global convergence and local superlinear convergence to a first-order stationary point of (1.1). However, their assumptions seem to be difficult to verify. Implementation issues were studied in [7, 45], with promising numerical results reported. Jarre [28] considered a trust-region strategy for computing a first-order stationary point of (1.4), which involves a line search on the trust-region multiplier at each iteration and requires a strictly feasible starting point. Partial results on convergence to a first-order stationary point of (1.1) as $\mu \to 0$ are also obtained. Additional references on related work are given in [7].

Motivated by the aforementioned work, in this paper we study an infeasible interior-point method, based on the log-barrier approach, for solving (1.1). (By "solving," we mean computing a second-order stationary point.) In our method, the barrier problem (1.4) is solved inexactly using a (new) trust-region strategy. The trust-region strategy maintains and iteratively updates an $(x, s, \lambda) \in \Re^n \times \Re^{2m}_{++}$ and a trust-region radius $\delta > 0$. It proceeds in three phases, with the aim of achieving, respectively, approximate feasibility, centrality, and first- and second-order stationarity for (1.4). At each iteration, a quadratic subproblem with an ℓ_2-ball trust region is solved (inexactly), and second-order correction is used to construct a trial point $(x_{\rm TR}, s_{\rm TR}, \lambda_{\rm TR})$—see (2.1), (2.3), (2.6). A novel feature of the subproblem is the scaling of the infeasibility term by δ^β, with β set to either below 1 or above 1, according to whether we are aiming to achieve feasibility or second-order stationarity for (1.4). We use a merit function to test the trial point for acceptance and adjust the trust-region radius δ accordingly. In phase 3, the merit function is $f^\mu(x, s)$. In phase 1, $f^\mu(x, s)$ is augmented by an ℓ_p-penalty on the infeasibility term $g(x) + s$, and thus the merit function is

$$f^{\mu,\tau}(x, s) := f^\mu(x, s) + \tau \|g(x) + s\|_p, \tag{1.5}$$

with $\tau > 0$ and $1 \le p \le \infty$. In phase 2, $f^\mu(x, s)$ is augmented by an ℓ_∞-penalty on the centrality term $S\lambda - \mu e$, and so the merit function is

$$\bar{f}^{\mu,\bar{\tau}}(x, s, \lambda) := f^\mu(x, s) + \bar{\tau} \|S\lambda - \mu e\|_\infty, \tag{1.6}$$

with $\bar{\tau} > 0$. (Here, S and Λ denote the $m \times m$ diagonal matrices with diagonal entries $s_1, \ldots, s_m$ and $\lambda_1, \ldots, \lambda_m$, respectively; e denotes the vector of 1's.) The function $f^{\mu,\tau}$ has been used with $p = 1$ by Yamashita [42] and Yamashita, Yabe, and Tanabe [45] in their interior-point/trust-region methods, with $p = 2$ by Byrd and Omojokun in their SQP methods (see [6, p. 153]), and with $p = 2$ by Byrd and coworkers [6, 7] in their interior-point/trust-region/SQP methods.

Our trust-region strategy is a primal-dual strategy. We can also consider a primal strategy, in which we set $\lambda_{\rm TR} := \mu(S_{\rm TR})^{-1}e$ in the trial point. Our global convergence results still hold for this primal method—see Note 1. Also, as we shall see in section 9, a primal method can be readily extended to semidefinite nonlinear programming. The method analyzed in [6] is a primal method, though it is remarked that much of its analysis can be extended to a primal-dual method. The result in [29] suggests that the two types of methods may be comparable in practice, unless high solution accuracy is desired. Despite this, primal-dual methods are generally considered to be superior, due partly to their superlinear convergence properties. On the other hand, a primal method can be efficient for certain classes of semidefinite programs, such as semidefinite linear programs in dual standard form arising from combinatorial optimization problems [4].

Under reasonable assumptions on the problem, our method is well defined and generates second-order stationary points. More specifically, define for each $\zeta \geq 0$ the set

$$\mathcal{X}_\zeta := \{x \in \Re^n : \|g(x) + s\|_p \leq \zeta \text{ for some } s \in \Re^m_+\},$$

and $I_\zeta(x) := \{i \in \{1, \ldots, m\} : |g_i(x)| \leq \zeta\}$. Thus, $\mathcal{X}_0$ is the feasible set for (1.1) and $I_0(x) = I(x)$. We show that if $\mathcal{X}_\zeta$ is bounded and a *relaxed* Mangasarian–Fromovitz constraint qualification (MFCQ) holds at each $x \in \mathcal{X}_\zeta$ (see (4.24)), where ζ is a constant depending on the initial infeasibility, then every sequence $\{x^t\}$ generated by our method remains in $\mathcal{X}_\zeta$, and every cluster point $\bar{x}$ and its corresponding Lagrange multiplier vector $\bar{\lambda} \in \Re^m$ satisfy the first-order optimality condition (1.2). If, in addition, f is thrice differentiable at $\bar{x}$, then for any subsequence of the generated iterates $\{(x^t, \lambda^t)\}_{t \in T'}$ converging to some $(\bar{x}, \bar{\lambda})$ and any subsequence of unit directions $\{d^t\}_{t \in T'}$ satisfying $[\nabla g_i(x^t)^T d^t]_{i \in I(\bar{x})} = 0$ for all $t \in T'$, we have

$$\liminf_{\substack{t \to \infty, \\ t \in T'}} (d^t)^T \nabla_{xx} l(x^t, \lambda^t) d^t \;\geq\; 0. \tag{1.7}$$

It follows that if g is affine (i.e., each g_i is affine) or if $\nabla g_i(\bar{x})$, $i \in I(\bar{x})$, are linearly independent, then $(\bar{x}, \bar{\lambda})$ also satisfies (1.3)—see Corollary 6.2.

Thus, in the case of affine g and thrice differentiable f, the boundedness of $\mathcal{X}_\zeta$ and a relaxed MFCQ are the only assumptions needed to ensure that a cluster point $\bar{x}$ exists and is a second-order stationary point of (1.1). If the initial (x, s) is also feasible, i.e., satisfies $g(x) + s = 0$, then our method maintains feasibility at all iterations, and ζ can be taken to be zero in the above assumptions, i.e., it suffices that the feasible set $\mathcal{X}_0$ be bounded and that MFCQ hold at each $x \in \mathcal{X}_0$ (see the discussion after Corollary 6.2). The latter is weaker than the common regularity assumption of linear independence of $\nabla g_i(x)$, $i \in I(x)$. The feasible active-set Newton method of Forsgren and Murray for linear inequality-constrained problems [21] also generates second-order stationary points but assumes, instead of MFCQ, that the problem does not have "primal degenerate" second-order stationary points. In the case of bound constraints (for which regularity holds everywhere), Coleman and Li [11] proposed a feasible affine-scaling trust-region method. They showed that if every cluster point of the generated iterates is "nondegenerate," then at least one cluster point is a second-order stationary point [11, Thm. 3.10(ii)]. A related method was studied by Monteiro and Wang [30] for the case of linear constraints and for f either convex or concave. Under certain nondegeneracy and constant Hessian-range assumptions, their method generates a second-order stationary point [30, Thm. 4.13]. An extension of the Coleman–Li method to a case of implicit bound constraints was considered by Dennis, Heinkenschloss, and Vicente [15]. They showed that if the search directions have certain properties and the iterates are bounded, then at least one cluster point is a second-order stationary point. Other trust-region methods, not using an interior-point approach, have been developed to generate, under reasonable assumptions, second-order stationary points for unconstrained and bound/equality-constrained problems (see [11, 14, 26] and references therein; see also the introduction of [18]).

In the case of nonaffine g, the only other (infeasible) methods we are aware of that can generate a second-order stationary point of (1.1) are the line search methods proposed by Mukai and Polak [32] and by Facchinei and Lucidi [18]. The method in [32] reformulates the inequalities as equalities (by expressing the slacks as the square of artificial variables), which in turn are handled by an exact penalty function. It is

shown that if the generated iterates x are bounded and remain in a "regular set," then every cluster point is a second-order stationary point. The method in [18] uses an exact differentiable penalty function (for inequalities) of Glad and Polak [24], as further studied by Di Pillo and Grippo [16]. It is shown that if (i) the generated iterates x lie in a bounded set, (ii) every point in $\Re^n$ is regular (it is remarked that this can be relaxed to every feasible solution's being regular), and (iii) every first-order stationary point of $x \mapsto \|\max\{g(x), 0\}\|^2$ is a feasible solution, then every cluster point is a second-order stationary point of (1.1). This set of assumptions is quite different from the relaxed MFCQ that we make. Gould and Toint [26] give an example of quadratic f and $g(x) = -x$ for which the local minimizer of $f^\mu(x, -g(x))$ converges to the origin, which does not satisfy the *strong* second-order necessary optimality condition for (1.1).[1] In fact, for quadratic f with rational coefficients and $g(x) = -x$, it is NP-complete to decide whether the origin satisfies the strong second-order necessary condition [33]. This suggests that a weak second-order necessary condition may be the best we can hope to achieve.

After the original version of this paper was written, two independent works appeared that influenced our subsequent revisions. The first is by Conn et al. [12] (also see [14, sect. 13.6]), in which a feasible primal-dual interior-point trust-region method is proposed for nonlinear programming with linear equality constraints and nonlinear inequality constraints. This method allows approximate Hessian computation and approximate solution of the trust-region subproblem. It is shown that, under reasonable assumptions, the outer iterates have an asymptotic second-order stationarity property [12, eq. (87)]. These results, as well as comments from the referees and the editor, motivated us to revise our work to also consider primal-dual methods,[2] approximate Hessian computation, approximate solution of the trust-region subproblem, and a simplification of our second-order stationarity result. The convergence analysis in [12] does not depend on the boundedness of the iterates, though it assumes that the inner iterates satisfy (in our notation) $\|\lambda^k - \mu(S^k)^{-1}e\| \to 0$ with $s^k = -g(x^k)$. This assumption is in some sense approximately enforced in our trust-region strategy via its phase 2. Local superlinear convergence of the method in [12] was recently studied in [25]. A second work by Wächter and Biegler [40] furnishes examples of (1.1), with $n = 1$ variable and $m = 2$ linear and quadratic constraints, for which the infeasible line search based methods in [10, 17, 23, 39, 42, 45] fail to achieve feasibility when started at reasonable infeasible solutions. The same reference notes that the trust-region method of [6] solves the examples. We will show in section 7 that our method also solves, at least numerically, such an example using the troublesome starting points. We note that our trust-region subproblem (2.1) differs from the one used in [6, eq. (2.1)] in several ways: the use of δ^β with $\beta \neq 1$, the absence of Δs from the trust-region constraint, and the absence of bound constraints on Δs. The last obviates the need to solve approximately an indefinite quadratic program (QP). Also, the analysis in [6] establishes convergence to first-order stationary points under a linear independence constraint qualification (LICQ). Our analysis establishes a similar result under a relaxed MFCQ and establishes convergence to second-order stationary points under an LICQ—see Corollary 6.2.

Throughout this paper, $\Re^n$ denotes the space of n-dimensional real column vectors, $\Re^m_{++}$ denotes the positive orthant in $\Re^m$, $\Re^{n\times m}$ denotes the space of $n \times m$ real

[1] This condition replaces the equality $\nabla g_i(x)^T d = 0$ in (1.3) with the inequality $\nabla g_i(x)^T d \leq 0$ for those $i \in I(x)$ with $\lambda_i = 0$.

[2] The original version of this paper considered the simpler primal method.

matrices, and ^{T} denotes transpose. For any $y \in \Re^m$, we denote by y_i the ith component of y, and by $\|y\|_p$, $\|y\|$ the p-, 2-norm of y ($1 \le p \le \infty$). For any $A \in \Re^{n\times m}$, let $\|A\| := \max_{\|y\|=1} \|Ay\|$. (":=" means "define.") For any $I \subseteq \{0, 1, 2, \ldots\}$, we denote by $|I|$ the cardinality of I. We also define the second-order Taylor remainders:

$$\begin{aligned} R_f(x, \Delta x) &:= f(x + \Delta x) - f(x) - \nabla f(x)^T \Delta x - \frac{1}{2}\Delta x^T \nabla^2 f(x) \Delta x, \\ L_g(x, \Delta x) &:= g(x + \Delta x) - g(x) - \nabla g(x)^T \Delta x, \\ R_g(x, \Delta x) &:= L_g(x, \Delta x) - \frac{1}{2}\Delta x^T \nabla^2 g(x) \Delta x, \end{aligned} \tag{1.8}$$

where $\Delta x^T \nabla^2 g(x) \Delta x$ denotes the vector in $\Re^m$ with components $\Delta x^T \nabla^2 g_i(x) \Delta x$, $i = 1, \ldots, m$. We denote the ℓ_2-norm ball by $\mathbb{B} := \{y \in \Re^n : \|y\| \le 1\}$, and denote by Π the class of continuous functions $\pi : \Re_{++} \mapsto \Re_{++}$ satisfying $\lim_{\mu\to 0} \pi(\mu) = 0$. For any $\varpi \ge 1$ and $\gamma \in \Re$, we define $\varpi \otimes \gamma := \varpi\gamma$ if $\gamma \ge 0$, and otherwise $\varpi \otimes \gamma := \gamma/\varpi$. Then $\varpi \otimes \gamma \ge \gamma$.

2. A trust-region strategy for (1.4).

2.1. Motivation. We sketch and motivate the key steps in our trust-region strategy for solving (1.4) inexactly. Given $(x, s, \lambda) \in \Re^n \times \Re^{2m}_{++}$ and trust-region radius $\delta > 0$, we choose a $\beta \ge 0$ and solve the following trust-region subproblem associated with $(x, s, \lambda, \delta, \beta)$:

$$\begin{aligned} \text{minimize} \quad & \nabla f(x)^T \Delta x - \lambda_{\text{P}}^T \Delta s + \frac{1}{2}\Delta x^T M \Delta x + \frac{1}{2}\Delta s^T \Lambda S^{-1} \Delta s \\ \text{subject to} \quad & \|\Delta x\| \le \delta, \quad A^T \Delta x + \Delta s = -\delta^\beta r, \end{aligned} \tag{2.1}$$

where we define

$$\lambda_{\text{P}} := \mu S^{-1} e, \quad M := \nabla_{xx} l(x, \lambda), \quad r := g(x) + s, \quad A := \nabla g(x). \tag{2.2}$$

This subproblem acts as a linear/quadratic approximation of (1.4) at (x, s), with λ approximating the Lagrange multipliers associated with the equality constraints. It can be solved accurately by various techniques [14, 31]. From an approximate solution $(\Delta x, \Delta s)$ of (2.1) and Lagrange multipliers $\lambda_{\text{TR}} \in \Re^m$ associated with the equality constraints, a trial point is generated according to

$$x_{\text{TR}} := x + \Delta x, \qquad s_{\text{TR}} := \max\{s + \Delta s - \Delta s_{\text{CS}}, -g(x_{\text{TR}})\}. \tag{2.3}$$

Here Δs_{CS} is a second-order correction term chosen to account for the absence of curvature information in the linearized constraints and for the discrepancy between λ and λ_{P}. This term, whose exact form will be given later in (2.6), ensures that the quadratic model is accurate up to the second order; see (2.7) and Lemma 3.1(c)–(e). Also, taking the maximum with $-g(x_{\text{TR}})$ ensures that $g(x_{\text{TR}}) + s_{\text{TR}} \ge 0$, which is needed to drive towards feasibility. This mechanism has been employed in the trust-region method of [6] for a similar purpose.

For any feasible solution $(\Delta x, \Delta s)$ of (2.1), let υ denote its objective value. By using the equality constraint in (2.1) to eliminate Δs, we obtain

$$\upsilon = \delta^\beta \lambda_{\text{P}}^T r + c^T \Delta x + \frac{1}{2}\delta^{2\beta} \|\Lambda^{1/2} S^{-1/2} r\|^2 + \delta^\beta r^T \Lambda S^{-1} A^T \Delta x + \frac{1}{2}\Delta x^T Q \Delta x, \tag{2.4}$$

where we define

$$b := S\lambda - \mu e, \quad c := \nabla_x l(x, \lambda_{\rm P}), \quad Q := M + A\Lambda S^{-1}A^T. \tag{2.5}$$

Then (2.1) is equivalent to minimizing (2.4) subject to $\|\Delta x\| \le \delta$. Our trust-region strategy proceeds in three phases. In phase 1, the aim is to maintain $(s, \lambda) > 0$ while driving $\|r\|_p$ below a desired threshold ϵ_1. This is accomplished by accepting the trial point if it achieves $\upsilon - \tau\delta^\beta\|r\|_p + \mathrm{o}(\delta^2)$ descent in the merit function $f^{\mu,\tau}(x, s)$, where υ is the objective value of the approximate solution of (2.1).[3] As this descent amount is driven towards 0, by choosing β judiciously so that r is the dominant term in (2.4), r would be driven towards 0. In phase 2, the aim is to maintain $(s, \lambda) > 0$ and $\|r\|_p \le \epsilon_1$ while driving $\|b\|_\infty$ below a desired threshold ϵ_2. This is accomplished analogously to phase 1, with $f^{\mu,\bar\tau}$ used as the merit function. In phase 3, the aim is to maintain $(s, \lambda) > 0$, $\|r\|_p \le \epsilon_1$, and $\|b\|_\infty \le \epsilon_2$, while driving towards second-order stationarity for (1.4). This is accomplished by accepting the trial point if it achieves $\upsilon + \mathrm{o}(\delta^2)$ descent in the merit function $f^\mu(x, s)$. As this descent amount is driven towards 0, by choosing β and δ judiciously so that $\Delta x^T Q \Delta x$ is the dominant term in (2.4), second-order stationarity would be attained asymptotically.

How can we ensure that the trial point has the desired descent property for the merit function? Specifically, how should we choose $\Delta s_{\rm CS}$ in (2.3), with $(\Delta x, \Delta s)$ being an (approximate) solution of (2.1)? Choosing $\Delta s_{\rm CS} = 0$ does not work since it captures neither the constraint curvature information nor the discrepancy between λ and $\lambda_{\rm P}$, and the resulting $\mathcal{O}(\delta^2)$ error between υ and the descent in the (augmented) barrier function is too large to attain second-order stationarity. Rather, we will choose

$$\Delta s_{\rm CS} := \frac{1}{2\mu} S\Lambda \Delta x^T \nabla^2 g(x) \Delta x + \frac{1}{2\mu} S^{-1}(\Delta S)^2 b, \tag{2.6}$$

with the first term accounting for the constraint curvature information embodied in M, and the second term accounting for the discrepancy between λ and $\lambda_{\rm P}$. (Here, ΔS denotes the $m \times m$ diagonal matrix with diagonal entries $\Delta s_1, \ldots, \Delta s_m$.) This choice arises naturally from our analysis of the trust-region subproblem, and, in particular, it ensures that

$$f^\mu(x_{\rm TR}, s_{\rm TR}) - f^\mu(x, s) \le \upsilon + \mathrm{o}(\delta^2) + \mathcal{O}(\max\{\delta, \delta^\beta\}^3) \tag{2.7}$$

(see Lemmas 3.1(c) and 3.2(c)).

How should we choose β? The most intuitive choice of $\beta = 1$ does not work. This can be seen from (2.4), where choosing $\beta = 1$ makes the feasibility term $\delta^\beta \lambda_{\rm P}^T r$ and the first-order stationarity term $c^T \Delta x$ equally dominant, in which case $\upsilon \to 0$ does not ensure $r \to 0$ or $c \to 0$. Instead, we make at most one of these two terms dominant by choosing β to be either less than 1 or greater than 1, depending on the relative size of $\lambda_{\rm P}^T r$ and $\|c\|$. (Actually, the choice of β matters only as δ approaches zero. For δ above a given threshold, β can be chosen arbitrarily. For example, choosing $\beta = 0$ would yield a primal-dual direction that is useful for fast local convergence.) In particular, if

$$\sigma_2\|c\| < \lambda_{\rm P}^T r$$

for some positive constant σ_2, then choosing β to be less than 1 would make $\delta^\beta \lambda_{\rm P}^T r$ dominant (regardless of the choice of Δx), i.e.,

$$\upsilon = \delta^\beta \lambda_{\rm P}^T r + \mathrm{o}(\delta^\beta),$$

[3] The notations $\mathcal{O}(\delta)$ and $\mathrm{o}(\delta)$ mean, respectively, $\mathcal{O}(\delta)/\delta$ is bounded and $\mathrm{o}(\delta)/\delta \to 0$ as $\delta \to 0$.

so that $\upsilon/\delta^\beta \to 0$ would force $\lambda_{\text{P}}^T r \to 0$. (Also, since $\Delta s = \mathcal{O}(\delta^\beta)$, we need $\beta > 2/3$ so that the error term in (2.7) is $\text{o}(\delta^2)$.) Otherwise, $\sigma_2\|c\| \geq \lambda_{\text{P}}^T r$, and let υ_c be the objective value (2.4) of the Cauchy solution $\Delta x_c := -\delta c/\|c\|$. Choosing β to be greater than 1 would make $-\delta\|c\|$ dominant in υ_c, so by choosing Δx to be no worse than Δx_c, we would have

$$\upsilon \leq \upsilon_c = -\delta\|c\| + \text{o}(\delta).$$

Then $\upsilon/\delta \to 0$ would force $\|c\| \to 0$ and hence $\lambda_{\text{P}}^T r \to 0$. In either case, we obtain $\lambda_{\text{P}}^T r \to 0$. By maintaining $r \geq 0$ and using that λ_{P} is bounded away from zero componentwise, we obtain $r \to 0$. To drive towards first- and second-order stationarity for (1.4), we set β to be strictly between 1 and 2, which makes $\delta^\beta \lambda_{\text{P}}^T r - \delta\|c\|$ dominant in υ_c, implying

$$\upsilon \leq \upsilon_c = \delta^\beta \lambda_{\text{P}}^T r - \delta\|c\| + \text{o}(\delta^\beta) = -\delta^\beta \left(\frac{\|c\|}{\delta^{\beta-1}} - \lambda_{\text{P}}^T r \right) + \text{o}(\delta^\beta).$$

Then $\upsilon/\delta^\beta \to 0$ would force $\|c\|/\delta^{\beta-1} - \lambda_{\text{P}}^T r$ to be asymptotically nonpositive, and, by maintaining r small, this would make c small. In particular, when $\|r\|_p$ and $\|c\|/\delta^{\beta-1} - \lambda_{\text{P}}^T r$ are below $\text{o}(\delta^{2-\beta})$, we would have $\|c\| = \text{o}(\delta)$, and the first four terms on the right-hand side of (2.4) would be $\text{o}(\delta^2)$. Then, by choosing Δx to be no worse than δz, where z is a unit eigenvector associated with the minimum eigenvalue γ_* of Q, we would have

$$\upsilon \leq \text{o}(\delta^2) + \frac{1}{2}\delta^2\gamma_* \leq \text{o}(\delta^2) + \frac{1}{2}\delta^2 d^T Q d \quad \forall d \in \mathbb{B}.$$

Then $A^T d = 0$ would imply $\upsilon \leq \text{o}(\delta^2) + \frac{1}{2}\delta^2 d^T M d$. As $\upsilon/\delta^2 \to 0$, this would yield asymptotic second-order stationarity, provided that $\lambda - \lambda_{\text{P}} \to 0$ (since c depends on λ_{P} while M depends on λ). The latter will be enforced by driving $b = S(\lambda - \lambda_{\text{P}}) \to 0$ in phase 2. In summary, we will choose β to be either β_1 or β_2, with $2/3 < \beta_1 < 1 < \beta_2 < 2$, when δ is below a given threshold; otherwise, we will choose β to be any number in $[0, \beta_2]$.

While the above ideas are intuitively reasonable, there are various details involving approximation and error estimation that need to be addressed. Also, it matters how fast r and b are driven to 0 relative to c and $\Delta x^T Q \Delta x$. They should not go to 0 too slowly and thus overwhelm the latter. They should not go to 0 too fast or the method will resemble a feasible primal method. Moreover, to ensure convergence of (x, λ) to second-order stationary points of (1.1) as $\mu \to 0$, we need the infeasibility term $\lambda_{\text{P}}^T r$ in (2.4) to go to zero faster than $\delta^{2-\beta}$. This in turn requires a careful estimation of the size of δ in terms of μ, $\|b\|_\infty$, $\|c\|$, and υ—see (5.1) and the assumptions A1 and A3 in Proposition 6.1(c). It also necessitates the introduction of a mechanism to ensure that δ does not go to zero too fast relative to μ^2. Specifically, when μ^2 is below a certain threshold, we maintain δ to be at least μ^2 upon entering phase 3 or after each successful (i.e., nonnull) step in phase 3.

Thus, much of the complication in our method and its analysis arise from its infeasible nature, necessitating a careful control and estimation of the infeasibility term r relative to δ, μ, $\|b\|_\infty$, $\|c\|$, and υ. This contrasts with a feasible method, e.g., [12], which does not face such complications. The primal-dual nature of the method is another source of complication. A primal method has a simpler structure and, in particular, does not need phase 2—see Note 1. Alternatively, if we assume

analogously to [12, Assumpt. 10] that $\lambda - \lambda_{\rm P} \to 0$ when $\|c\| \to 0$, then phase 2 would not be needed. While infeasible interior-point methods are preferred over feasible interior-point methods for linear programming, there is no established preference for nonlinear programming. A recent study of this issue is described in [9]. A hybrid method, in which a subset of inequality constraints is maintained as satisfied, seems promising since it offers the greatest flexibility.

2.2. Inexact solution of the trust-region subproblem. It is well known that an exact solution $(\Delta x, \Delta s)$ of (2.1) satisfies, together with Lagrange multipliers $\lambda_{\rm TR} \in \Re^m$, the following system of linear equations [14, 31]:

$$(\gamma I + M)\Delta x + A\lambda_{\rm TR} = -\nabla f(x), \tag{2.8}$$

$$A^T\Delta x + \Delta s = -\delta^\beta r, \tag{2.9}$$

$$S^{-1}\Lambda\Delta s + \lambda_{\rm TR} = \lambda_{\rm P}, \tag{2.10}$$

where $\gamma \geq 0$ is such that $\gamma I + M + A\Lambda S^{-1}A^T$ is positive semidefinite and $\|\Delta x\| \leq \delta$, with equality holding whenever $\gamma > 0$. Here, $\lambda_{\rm P}, M, r, A$ are given by (2.2). By letting $\Delta\lambda := \lambda_{\rm TR} - \lambda$ and eliminating Δs, this can be written as

$$\begin{bmatrix} \gamma I + M & A \\ -A^T & \Lambda^{-1}S \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta\lambda \end{bmatrix} = \begin{bmatrix} -c \\ \delta^\beta r - \Lambda^{-1}b \end{bmatrix},$$

with b, c given by (2.5). For $\beta = 0$, this equation coincides with that used in many infeasible primal-dual interior-point methods [10, 17, 23, 39, 42, 45].

On large problems, it is often desirable to solve (2.1) inexactly, as is discussed in [14, Chap. 7] and references therein. We consider below a notion of an inexact solution that allows for attainment of second-order stationarity. Let $q(\Delta x)$ denote the objective value of (2.1) as a function of Δx, i.e., $q(\Delta x)$ equals the right-hand side of (2.4). Then, (2.1) reduces to

$$\upsilon_* \;:=\; \min_{x \in \delta\mathbb{B}} q(\Delta x).$$

Here we use the notation from (2.2), (2.5). Define the Cauchy solution:

$$\Delta x_{\rm C} := -\delta\bar{c} \qquad \text{with} \quad \bar{c} := \begin{cases} \dfrac{c}{\|c\|} & \text{if } c \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Then $\|\Delta x_{\rm C}\| \leq \delta$ and hence $\upsilon_* \leq q(\Delta x_{\rm C})$. Fix any $\varpi \geq 1$. Let γ_* denote the minimum eigenvalue of Q and let $z \in \Re^n$ be any unit vector satisfying $z^TQz \leq \varpi \otimes \gamma_* + \varpi\mu$. (Since $\gamma_* \leq \varpi \otimes \gamma_*$, we can take z to be any unit eigenvector of Q associated with γ_*.) Let

$$\Delta x_{\rm Q} := \delta z.$$

Then $\|\Delta x_{\rm Q}\| = \delta$ and hence $\upsilon_* \leq q(\Delta x_{\rm Q}) \leq \upsilon_{\rm Q}$, where we define

$$\upsilon_{\rm Q} := \delta^\beta\lambda_{\rm P}^T r + \delta\|c\| + \frac{\delta^{2\beta}}{2}\|\Lambda^{1/2}S^{-1/2}r\|^2 + \delta^{\beta+1}\|A\Lambda S^{-1}r\| + \frac{\delta^2}{2}(\varpi \otimes \gamma_* + \varpi\mu).$$

We prefer to work with $\upsilon_{\rm Q}$ rather than $q(\Delta x_{\rm Q})$, since it involves only eigenvalue, not eigenvector.

For any fixed $\varpi \geq 1$, we say that $(\Delta x, \Delta s, \lambda_{\text{TR}})$ is a *ϖ-approximate solution* of (2.1) if

$$(\Delta x, \Delta s) \text{ is feasible for (2.1)}, \quad \lambda_{\text{TR}} \text{ satisfies (2.10)}, \quad \text{and} \quad q(\Delta x) \leq \varpi \otimes \min\{q(\Delta x_{\text{c}}), \upsilon_{\text{Q}}\}. \tag{2.11}$$

From the preceding discussion, we see that an exact primal-dual solution of (2.1) is a ϖ-approximate solution of (2.1). Also, if Δx_{Q} is available, then setting $\Delta x := \Delta x_{\text{Q}}$ if $q(\Delta x_{\text{Q}}) \leq q(\Delta x_{\text{c}})$, and otherwise setting $\Delta x := \Delta x_{\text{c}}$, yields a ϖ-approximate solution. This solution can be further improved by minimizing $q(\Delta x)$, with $\Delta x \in \delta\mathbb{B}$ restricted to a linear combination of Δx_{c} and Δx_{Q}. This minimization is a two-dimensional trust-region problem. More generally, we can use (2.11) as a termination criterion for any algorithm applied to solve (2.1). In the case in which Q is positive semidefinite and $r = 0$, it is known that an $\Delta x \in \delta\mathbb{B}$ satisfying $q(\Delta x) \leq \upsilon_*/2$ (called the "Steihaug–Toint point") can be computed using a truncated conjugated gradient method; see [14, sect. 7.5.2]. Thus, in this case, we can take $\varpi = 2$. If Q is indefinite, the generalized Lanczos trust-region method of Gould, Lucidi, Roma, and Toint (see [14, sec. 7.5.4]) can be used. The criterion (2.11) differs from that given in [12, eq. (19)], though both involve the minimum eigenvalue of Q.

2.3. Trust-region strategy description. Below we formally describe our three-phase trust-region strategy for solving (1.4) inexactly. Throughout this section, for any $(x^k, s^k, \lambda^k) \in \Re^n \times \Re^{2m}_{++}$, we define analogously to (2.2) and (2.5)

$$\lambda^k_{\text{P}} := \mu(S^k)^{-1}e, \quad M^k := \nabla_{xx} l(x^k, \lambda^k), \quad r^k := g(x^k) + s^k, \quad A^k := \nabla g(x^k), \tag{2.12}$$

$$b^k := S^k\lambda^k - \mu e, \quad c^k := \nabla_x l(x^k, \lambda^k_{\text{P}}), \quad Q^k := M^k + A^k\Lambda^k(S^k)^{-1}(A^k)^T. \tag{2.13}$$

In addition, for any $(\Delta x^k, \Delta s^k, \lambda^k_{\text{TR}}) \in \Re^n \times \Re^{2m}$, we define analogously to (2.6) and (2.3) the following:

$$\begin{aligned} \Delta s^k_{\text{CS}} &:= \frac{1}{2\mu} S^k\Lambda^k(\Delta x^k)^T\nabla^2 g(x^k)\Delta x^k + \frac{1}{2\mu}(S^k)^{-1}(\Delta S^k)^2 b^k, \\ x^k_{\text{TR}} &:= x^k + \Delta x^k, \qquad s^k_{\text{TR}} := \max\{s^k + \Delta s^k - \Delta s^k_{\text{CS}}, -g(x^k_{\text{TR}})\}, \\ r^k_{\text{TR}} &:= g(x^k_{\text{TR}}) + s^k_{\text{TR}}, \qquad b^k_{\text{TR}} := S^k_{\text{TR}}\lambda^k_{\text{TR}} - \mu e. \end{aligned} \tag{2.14}$$

Algorithm 1.

Input. *$\mu > 0$, $1 \leq p \leq \infty$, $1 < \beta_2 < 2$, $0 < \omega_1 < 1$, $\varpi \geq 1$, $0 < \delta_{\text{th}} \leq \delta_{\max}, 0 < \delta_0 \leq \delta_{\max}$, $(x^0, s^0, \lambda^0) \in \Re^n \times \Re^{2m}_{++}$ with $s^0 \geq -g(x^0)$, and termination tolerances $\epsilon_1 > 0, \epsilon_2 > 0, \epsilon_3 > 0, \epsilon_4 > 0$.*

Initialization. *Choose $2/3 < \beta_1 < 1$, $0 < \sigma_1 < 1$, $\sigma_2 > 0$, $\tau > 0$, $\bar{\tau} > 0$, $0 < \eta < 1$. Initialize $k := 0$. Go to phase 1.*

Phase 1 *(towards feasibility).*

1. *If $\|r^k\|_p \leq \epsilon_1$, go to phase 2. Otherwise, go to step 2.*

2. *If $\delta_k > \delta_{\text{th}}$, set β^k to any value in $[0, \beta_2]$; else if*

$$\sigma_2\|c^k\| < (\lambda^k_{\text{P}})^T r^k, \tag{2.15}$$

set $\beta^k := \beta_1$; else set $\beta^k := \beta_2$. Let $(\Delta x^k, \Delta s^k, \lambda^k_{\text{TR}})$ be a ϖ-approximate solution of (2.1) associated with $(x^k, s^k, \lambda^k, \delta_k, \beta^k)$. Let υ^k denote its objective value. Let

$$\rho_k := \frac{f^{\mu,\tau}(x^k_{\text{TR}}, s^k_{\text{TR}}) - f^{\mu,\tau}(x^k, s^k)}{-\tau\sigma_1(\delta_k)^{\beta^k}\|r^k\|_p}.$$

(Here $\rho_k = -\infty$ if $s^k_{\rm TR} \not> 0$.) If $\rho_k \geq \eta$ and $\lambda^k_{\rm TR} > 0$ and

$$\|r^k_{\rm TR}\|_p \leq \max\big\{(1-\sigma_1(\delta_k)^{\beta^k})\|r^k\|_p, \epsilon_1\big\}, \tag{2.16}$$

$$\|b^k_{\rm TR}\|_\infty \leq \max\big\{(1-\sigma_1(\delta_k)^{\beta_1})\|b^k\|_\infty, \epsilon_2\big\}, \tag{2.17}$$

set $(x^{k+1}, s^{k+1}, \lambda^{k+1}) := (x^k_{\rm TR}, s^k_{\rm TR}, \lambda^k_{\rm TR})$; else set $(x^{k+1}, s^{k+1}, \lambda^{k+1}) := (x^k, s^k, \lambda^k)$ ("null step"). Let

$$\delta_{k+1} := \begin{cases} \omega_1\delta_k & \text{if null step,} \\ \text{any } \delta \in [\delta_k, \delta_{\max}] & \text{otherwise.} \end{cases} \tag{2.18}$$

Increment k by 1 and return to step 1.

Phase 2 *(towards centrality).*

1. *If $\|b^k\|_\infty \leq \epsilon_2$, go to phase 3, and if in addition $\delta_k < \mu^2 < \delta_{\rm th}$, reset δ_k to any $\delta \in [\mu^2, \delta_{\max}]$. Otherwise, go to step 2.*

2. *If $\delta_k > \delta_{\rm th}$, set β^k to any value in $[0, \beta_2]$; else set $\beta^k := \beta_2$. Let $(\Delta x^k, \Delta s^k, \lambda^k_{\rm TR})$ be a ϖ-approximate solution of (2.1) associated with $(x^k, s^k, \lambda^k, \delta_k, \beta^k)$. Let υ^k denote its objective value. Let*

$$\rho_k := \frac{\bar f^{\mu,\bar\tau}(x^k_{\rm TR}, s^k_{\rm TR}, \lambda^k_{\rm TR}) - \bar f^{\mu,\bar\tau}(x^k, s^k, \lambda^k)}{-\bar\tau\sigma_1(\delta_k)^{\beta_1}\|b^k\|_\infty}.$$

(Here $\rho_k = -\infty$ if $s^k_{\rm TR} \not> 0$.) If

$$\rho_k \geq \eta \quad \text{and} \quad \lambda^k_{\rm TR} > 0 \quad \text{and} \quad \|r^k_{\rm TR}\|_p \leq \epsilon_1 \quad \text{and} \quad (2.17) \text{ holds,} \tag{2.19}$$

set $(x^{k+1}, s^{k+1}, \lambda^{k+1}) := (x^k_{\rm TR}, s^k_{\rm TR}, \lambda^k_{\rm TR})$; else set $(x^{k+1}, s^{k+1}, \lambda^{k+1}) := (x^k, s^k, \lambda^k)$ ("null step"). Let δ_{k+1} be given by (2.18). Increment k by 1 and return to step 1.

Phase 3 *(towards first- and second-order stationarity).*

1. *If $\delta_k > \delta_{\rm th}$, set β^k to any value in $[0, \beta_2]$; else set $\beta^k := \beta_2$. Let $(\Delta x^k, \Delta s^k, \lambda^k_{\rm TR})$ be a ϖ-approximate solution of (2.1) associated with $(x^k, s^k, \lambda^k, \delta_k, \beta^k)$. Let υ^k denote its objective value. If*

$$\frac{\|c^k\|}{(\delta_k)^{\beta_2-1}} - (\lambda^k_{\rm P})^T r^k \leq \epsilon_3 \quad \text{and} \quad \frac{\upsilon^k}{(\delta_k)^2} \geq -\epsilon_4, \tag{2.20}$$

then terminate phase 3 and output $(x^k, s^k, \lambda^k, \delta_k, \beta^k, \upsilon^k)$. Otherwise, go to step 2.

2. *Let*

$$\rho_k := \begin{cases} \dfrac{f^\mu(x^k_{\rm TR}, s^k_{\rm TR}) - f^\mu(x^k, s^k)}{\upsilon^k} & \text{if } \upsilon^k < 0, \\ -\infty & \text{otherwise.} \end{cases} \tag{2.21}$$

(Here $\rho_k = -\infty$ if $s^k_{\rm TR} \not> 0$.) If

$$\rho_k \geq \eta \quad \text{and} \quad \lambda^k_{\rm TR} > 0 \quad \text{and} \quad \|r^k_{\rm TR}\|_p \leq \epsilon_1 \quad \text{and} \quad \|b^k_{\rm TR}\|_\infty \leq \epsilon_2 \tag{2.22}$$

and the second condition in (2.20) fails, set $(x^{k+1}, s^{k+1}, \lambda^{k+1}) := (x^k_{\rm TR}, s^k_{\rm TR}, \lambda^k_{\rm TR})$; else set $(x^{k+1}, s^{k+1}, \lambda^{k+1}) := (x^k, s^k, \lambda^k)$ ("null step"). If $\delta_k < \mu^2 < \delta_{\rm th}$ and the step is not null, let δ_{k+1} be any $\delta \in [\mu^2, \delta_{\max}]$; else let δ_{k+1} be given by (2.18). Increment k by 1 and return to step 1.

Note 1. Algorithm 1 is a primal-dual strategy in the sense that Δs^k and $\Delta\lambda^k := \lambda^k_{\rm TR} - \lambda^k$ satisfy the primal-dual Newton equation (see (2.10))

$$\Lambda^k \Delta s + S^k \Delta\lambda = \mu e - S^k\lambda^k.$$

We can also consider a *primal* strategy, whereby we instead initialize $\lambda^0 := \mu(S^0)^{-1}e$ and set

$$\lambda^k_{\rm TR} := \mu(S^k_{\rm TR})^{-1}e \qquad \forall k.$$

The resulting method is simpler than Algorithm 1, and, in particular, it maintains $b^k = b^k_{\rm TR} = 0$ and $\lambda^k_{\rm P} = \lambda^k$ for all k, thus allowing phase 2 to be bypassed altogether. It can be seen that our convergence result, namely, Proposition 4.1, also extends to this primal strategy.

Note 2. To improve numerical stability, we can make the substitution $\widetilde{\Delta s} = S^{-1}\Delta s$ in (2.1) and solve for $\widetilde{\Delta s}^k := (S^k)^{-1}\Delta s^k$ instead of Δs^k. Then, we have

$$(s^k_{\rm TR})_i = s^k_i(1+\alpha^k_i), \qquad (\lambda^{k+1}_{\rm P})_i = \begin{cases} (\lambda^k_{\rm P})_i & \text{if } s^{k+1}_i = s^k_i, \\ \dfrac{(\lambda^k_{\rm P})_i}{(1+\alpha^k_i)} & \text{otherwise} \end{cases}$$

for all $i = 1, \ldots, m$, where

$$\alpha^k_i := \max\left\{ \widetilde{\Delta s}^k_i - \frac{(\Delta s^k_{\rm CS})_i}{s^k_i}, \frac{-g_i(x^k_{\rm TR})}{s^k_i} - 1 \right\}.$$

The above equations give a numerically more stable way to compute $s^k_{\rm TR}$ and $\lambda^{k+1}_{\rm P}$.

Note 3. In step 1 of phase 2 and step 2 of phase 3, instead of μ^2 we can more generally use any constant multiple of μ^2. In (2.15), the term $\|c^k\|$ may be replaced more generally by $\|c^k\|/\min\{1, (\delta_k)^{\beta_3 - 1}\}$, with $1 \le \beta_3 < \beta_2$. Proposition 4.1 still applies to this case, though it is unclear what practical advantage it offers. Also, following [12], we can allow $\nabla^2 f(x^k)$ and $\nabla^2 g_i(x^k)$ to be replaced by $n \times n$ symmetric real matrices H^k and H^k_i, $i = 1, \ldots, m$, in the definition of M^k and $\Delta s^k_{\rm CS}$. Proposition 4.1 still holds, provided that

$$\|\nabla^2 f(x^k) - H^k\| \to 0 \quad \text{and} \quad \|H^k_i\| \text{ is bounded}, \qquad i = 1, \ldots, m, \quad \text{as } k \to \infty.$$

Note 4. The equations (2.8)–(2.10) suggest a variant of Algorithm 1 whereby we update γ explicitly and solve (2.8)–(2.10) instead of (2.1). More precisely, given $\gamma_k \ge 0$, we check whether

$$\gamma I + M + A\Lambda S^{-1}A^T \tag{2.23}$$

is positive definite, with $(x, s, \lambda, \gamma, \delta, \beta) = (x^k, s^k, \lambda^k, \gamma_k, \delta_{k-1}, \beta^k)$ and A, M given by (2.2). If yes, we solve (2.8)–(2.10) to obtain $(\Delta x^k, \Delta s^k, \lambda^k_{\rm TR})$ and set $\delta_k := \|\Delta x^k\|$ if $\gamma_k > 0$ and otherwise set $\delta_k := \max\{\|\Delta x^k\|, \delta_{k-1}\}$. (We can check whether (2.23) is positive definite using a single Cholesky factorization. If (2.23) is positive definite, the factorization can be used to solve (2.8)–(2.10).) Otherwise, we choose a $\gamma_{k+1} > \gamma_k$ and set $(x^{k+1}, s^{k+1}, \lambda^{k+1}) := (x^k, s^k, \lambda^k)$ (see [39, sect. 3] for some rules for updating γ). In step 2, an analogous formula for γ_{k+1} would replace (2.18). Proposition 4.1 extends to this variant as long as $\{\delta_{k+1}/\delta_k\}$ is bounded away from 0 and exceeds 1 whenever a successful step is taken at iteration k.

3. Properties of the trust-region subproblem. In this section we study properties of the approximate solutions of (2.1) and the corresponding trial point $(x_{\rm TR}, s_{\rm TR}, \lambda_{\rm TR})$ given by (2.3), (2.6). These properties will be used in the next section to analyze Algorithm 1. We begin with the following lemma estimating the change in $\|r\|_p$, $\|b\|_\infty$, $f^\mu(x,s)$, $f^{\mu,\tau}(x,s)$, and $\bar f^{\mu,\bar\tau}(x,s,\lambda)$ when moving to a trial point. In what follows, we define

$$R(s,\Delta s,\Delta s_{\rm CS}) := \left(\frac{1}{2}\Delta s_{\rm CS} - \Delta s\right)^T S^{-2}\Delta s_{\rm CS} + \sum_{i=1}^m \frac{|\Delta s_i - (\Delta s_{\rm CS})_i|^3}{(s_i)^3}.$$

As we will see in (3.4) below, this remainder estimates the contribution by slack variables to the error between the predicted descent υ (i.e., the objective value of (2.1)) and the actual descent in the log-barrier value $f^\mu(x_{\rm TR}, s_{\rm TR}) - f^\mu(x,s)$.

LEMMA 3.1. *Fix $\mu > 0$, $1 \le p, q \le \infty$ with $1/p + 1/q = 1$, $\beta_1 > 0$, $0 < \sigma_1 < 1$, $\tau > 0$, $\bar\tau > 0$, $0 < \eta < 1$, $\epsilon_1 > 0$, $\epsilon_2 > 0$, and $(x,s,\lambda) \in \Re^n \times \Re^{2m}_{++}$. Let r, b be given by (2.2), (2.5). For any $0 < \delta \le 1$ and $\beta > 0$, let $(\Delta x, \Delta s)$ denote a feasible solution of (2.1) and let υ denote its objective value. Then, the following results hold with $\Delta s_{\rm CS}$, $(x_{\rm TR}, s_{\rm TR})$, $\lambda_{\rm TR}$ given by (2.6), (2.3), (2.10), respectively, and with $r_{\rm TR} := g(x_{\rm TR}) + s_{\rm TR}$, $b_{\rm TR} := S_{\rm TR}\lambda_{\rm TR} - \mu e$, $\widehat{\Delta s} := \Delta s - \Delta s_{\rm CS}$.*

(a) *$r_{\rm TR} \ge 0$. If*

$$\frac{\|L_g(x,\Delta x) - \Delta s_{\rm CS}\|_p}{\delta^\beta} \le (1-\sigma_1)\|r\|_p, \tag{3.1}$$

then $\|r_{\rm TR}\|_p \le (1 - \sigma_1\delta^\beta)\|r\|_p$. If

$$\|r\|_p \le \epsilon_1 \quad \text{and} \quad \frac{\|L_g(x,\Delta x) - \Delta s_{\rm CS}\|_p}{\delta^\beta} \le \epsilon_1, \tag{3.2}$$

then $\|r_{\rm TR}\|_p \le \epsilon_1$.

(b) *$b_{\rm TR} = \Delta S(\lambda_{\rm TR} - \lambda)$. If $\|\Delta s\| < \mu/\|\lambda\|_\infty$, then $\lambda_{\rm TR} > 0$. For any $\theta > 0$, if $r \ge 0$ and*

$$\begin{aligned} \|\Delta s_{\rm CS}\|_\infty + \|L_g(x,\Delta x) - \Delta s_{\rm CS}\|_\infty &\le \frac{\theta}{2\|\lambda_{\rm TR}\|_\infty}, \\ \|S^{-1}\Delta s\|_\infty &\le \frac{\theta}{\sqrt{2\theta(\|b\|_\infty + \mu)} + 2\|b\|_\infty}, \end{aligned} \tag{3.3}$$

then $\|b_{\rm TR}\|_\infty \le \theta$.

(c) *Assume $\|S^{-1}\widehat{\Delta s}\|_\infty \le 2/3$. Then*

$$f^\mu(x_{\rm TR}, s_{\rm TR}) - f^\mu(x,s) \le \upsilon + R_f(x,\Delta x) + \mu R(s,\Delta s,\Delta s_{\rm CS}). \tag{3.4}$$

If $\upsilon < 0$, then the right-hand side of (3.4) is below $\eta\upsilon$ whenever $R_f(x,\Delta x) + \mu R(s,\Delta s,\Delta s_{\rm CS}) \le (\eta - 1)\upsilon$.

(d) *Assume that $\|S^{-1}\widehat{\Delta s}\|_\infty \le 2/3$ and that (3.1) holds. Then*

$$\begin{aligned} &f^{\mu,\tau}(x_{\rm TR}, s_{\rm TR}) - f^{\mu,\tau}(x,s) \\ &\quad \le \upsilon + R_f(x,\Delta x) + \mu R(s,\Delta s,\Delta s_{\rm CS}) - \tau\sigma_1\delta^\beta\|r\|_p. \end{aligned} \tag{3.5}$$

If $\upsilon < 0$, then the right-hand side of (3.5) is below $-\eta\tau\sigma_1\delta^\beta\|r\|_p$ whenever $R_f(x,\Delta x) + \mu R(s,\Delta s,\Delta s_{\rm CS}) \le -\upsilon$. If for some $\chi > 0$,

$$\upsilon \le (1+\chi)\delta^\beta\lambda_{\rm P}^T r \quad \text{and} \quad (1+2\chi)\|\lambda_{\rm P}\|_q \le (1-\eta)\tau\sigma_1, \tag{3.6}$$

then the right-hand side of (3.5) is below $-\eta\tau\sigma_1\delta^\beta\|r\|_p$ whenever $R_f(x,\Delta x) + \mu R(s,\Delta s,\Delta s_{\rm CS}) \le \chi\delta^\beta\lambda_{\rm P}^T r$.

(e) *Assume* $\|S^{-1}\widehat{\Delta s}\|_\infty \le 2/3$, $r \ge 0$, *and that* (3.3) *holds with* $\theta := (1 - \sigma_1\delta^{\beta_1})\|b\|_\infty$. *Then*

$$\bar{f}^{\mu,\bar{\tau}}(x_{\rm TR}, s_{\rm TR}, \lambda_{\rm TR}) - \bar{f}^{\mu,\bar{\tau}}(x, s, \lambda) \le -\eta\bar{\tau}\sigma_1\delta^{\beta_1}\|b\|_\infty$$

whenever $v + R_f(x, \Delta x) + \mu R(s, \Delta s, \Delta s_{\rm CS}) \le (1 - \eta)\bar{\tau}\sigma_1\delta^{\beta_1}\|b\|_\infty$.

Proof. (a) By (2.3), $s_{\rm TR} \ge -g(x_{\rm TR})$ and so $r_{\rm TR} \ge 0$. Also, for each $i \in \{1, \ldots, m\}$, we have $(r_{\rm TR})_i = g_i(x_{\rm TR}) + s_i + \widehat{\Delta s}_i$ if $g_i(x_{\rm TR}) + s_i + \widehat{\Delta s}_i > 0$ and otherwise $(r_{\rm TR})_i = 0$. Hence $|(r_{\rm TR})_i| \le |g_i(x_{\rm TR}) + s_i + \widehat{\Delta s}_i|$, implying

$$\begin{aligned}\|r_{\rm TR}\|_p &\le \|g(x_{\rm TR}) + s + \widehat{\Delta s}\|_p \\ &= \|L_g(x, \Delta x) - \Delta s_{\rm CS} + (1 - \delta^\beta)r\|_p \\ &\le \|L_g(x, \Delta x) - \Delta s_{\rm CS}\|_p + (1 - \delta^\beta)\|r\|_p,\end{aligned}$$

where the equality follows from the equation in (2.1) and the definition of L_g (see (1.8)). Thus, the right-hand side is below $(1 - \sigma_1\delta^\beta)\|r\|_p$ whenever (3.1) holds. Similarly, the right-hand side is below ϵ_1 whenever (3.2) holds.

(b) If $\|\Delta s\| < \mu/\|\lambda\|_\infty$, then (2.10) and $s > 0, \lambda > 0$ imply that $\lambda_{\rm TR} = S^{-1}\Lambda(\mu\Lambda^{-1}e - \Delta s) > 0$. Next, we have from (2.3) and letting $h := g(x_{\rm TR}) + s + \Delta s - \Delta s_{\rm CS}$ that

$$g(x_{\rm TR}) + s_{\rm TR} \;=\; \max\{h, 0\}.$$

Thus, for each $i \in \{1, \ldots, m\}$, either $g_i(x_{\rm TR}) + (s_{\rm TR})_i = h_i$, $h_i \ge 0$, or $g_i(x_{\rm TR}) + (s_{\rm TR})_i = 0$, $h_i < 0$. In the latter case, we have from $h = L_g(x, \Delta x) - \Delta s_{\rm CS} + (1 - \delta^\beta)r$ (see the proof of (a)) and $r \ge 0$ that $L_{g_i}(x, \Delta x) - (\Delta s_{\rm CS})_i \le h_i < 0$. Thus, we have

$$\|g(x_{\rm TR}) + s_{\rm TR} - h\|_\infty \le \|L_g(x, \Delta x) - \Delta s_{\rm CS}\|_\infty.$$

Then

$$\begin{aligned}b_{\rm TR} &= \Lambda_{\rm TR}s_{\rm TR} - \mu e \\ &= \Lambda_{\rm TR}(s + \Delta s) - \mu e - \Lambda_{\rm TR}\Delta s_{\rm CS} + \Lambda_{\rm TR}(g(x_{\rm TR}) + s_{\rm TR} - h),\end{aligned}$$

so that

$$\|b_{\rm TR}\|_\infty \le \|\Lambda_{\rm TR}(s + \Delta s) - \mu e\|_\infty + \|\lambda_{\rm TR}\|_\infty \left(\|\Delta s_{\rm CS}\|_\infty + \|L_g(x, \Delta x) - \Delta s_{\rm CS}\|_\infty\right).$$

By (2.10) with $\Delta\lambda := \lambda_{\rm TR} - \lambda$, we have

$$\begin{aligned}\Lambda_{\rm TR}(s + \Delta s) - \mu e &= (S + \Delta S)(\lambda + \Delta\lambda) - \mu e \\ &= \Delta S\Delta\lambda \\ &= -\Delta SS^{-1}(b + \Lambda\Delta s) \\ &= -(S^{-1}\Delta S)b - (S^{-1}\Delta S)^2 S\lambda,\end{aligned}$$

so that

$$\begin{aligned}&\|b_{\rm TR}\|_\infty - \|\lambda_{\rm TR}\|_\infty(\|\Delta s_{\rm CS}\|_\infty + \|L_g(x, \Delta x) - \Delta s_{\rm CS}\|_\infty) \\ &\qquad \le \|(S^{-1}\Delta S)b\|_\infty + \|(S^{-1}\Delta S)^2 S\lambda\|_\infty \\ &\qquad \le \alpha c_1 + \alpha^2 c_2, \qquad (3.7)\end{aligned}$$

where for simplicity we define $\alpha := \|S^{-1}\Delta s\|_\infty$, $c_1 := \|b\|_\infty$, $c_2 := \|S\lambda\|_\infty$. Then, for any $\theta > 0$, $\|b_{\rm TR}\|_\infty$ is below θ if

$$\|\lambda_{\rm TR}\|_\infty(\|\Delta s_{\rm CS}\|_\infty + \|L_g(x,\Delta x) - \Delta s_{\rm CS}\|_\infty) \le \frac{\theta}{2}$$

and the right-hand side of (3.7) is below $\theta/2$, i.e.,

$$\alpha c_1 + \alpha^2 c_2 \le \frac{\theta}{2}.$$

This is a quadratic inequality in α, which is satisfied if and only if $\alpha \le \bar\alpha$, where

$$\bar\alpha := \frac{\sqrt{c_1^2 + 2\theta c_2} - c_1}{2c_2} = \frac{\theta}{\sqrt{c_1^2 + 2\theta c_2} + c_1} \ge \frac{\theta}{\sqrt{2\theta c_2} + 2c_1}.$$

Using $c_2 = \|b + \mu e\|_\infty \le \|b\|_\infty + \mu$ completes the proof.

(c) We have from properties of the logarithm that, for any $\xi > 0$ and $-1 < a < 1$,

$$\begin{aligned}
-\ln(\xi(1+a)) + \ln(\xi) &= -a + \frac{a^2}{2} - \frac{a^3}{3} + \frac{a^4}{4} - \frac{a^5}{5} + \frac{a^6}{6} + \cdots \\
&\le -a + \frac{a^2}{2} + \frac{|a|^3}{3} + \frac{a^4}{3} + \frac{|a|^5}{3} + \frac{a^6}{3} + \cdots \\
&= -a + \frac{a^2}{2} + \frac{|a|^3}{3(1-|a|)}.
\end{aligned}$$

Since $|\widehat{\Delta s}_i|/s_i \le \frac{2}{3}$ for each i, this yields

$$\begin{aligned}
-\ln((s_{\rm TR})_i) &\le -\ln(s_i + \widehat{\Delta s}_i) \\
&= -\ln\left(s_i\left(1 + \frac{\widehat{\Delta s}_i}{s_i}\right)\right) \\
&\le -\ln(s_i) - \frac{\widehat{\Delta s}_i}{s_i} + \frac{(\widehat{\Delta s}_i)^2}{2(s_i)^2} + \frac{|\widehat{\Delta s}_i|^3}{3(1-|\widehat{\Delta s}_i|/s_i)(s_i)^3} \\
&\le -\ln(s_i) - \frac{\widehat{\Delta s}_i}{s_i} + \frac{(\widehat{\Delta s}_i)^2}{2(s_i)^2} + \frac{|\widehat{\Delta s}_i|^3}{(s_i)^3},
\end{aligned}$$

where the first inequality uses $(s_{\rm TR})_i \ge s_i + \widehat{\Delta s}_i$ and the increasing property of $\ln(\cdot)$. Thus, summing the above inequality over $i = 1, \ldots, m$ and using $\widehat{\Delta s}_i = \Delta s_i - (\Delta s_{\rm CS})_i$ and (2.6) yields

$$\begin{aligned}
-\sum_{i=1}^m \ln((s_{\rm TR})_i) &\le -\sum_{i=1}^m \ln(s_i) - \frac{\widehat{\Delta s}_i}{s_i} + \frac{(\widehat{\Delta s}_i)^2}{2(s_i)^2} + \frac{|\widehat{\Delta s}_i|^3}{(s_i)^3} \\
&= -\sum_{i=1}^m \ln(s_i) - \frac{\Delta s_i}{s_i} + \frac{\lambda_i \Delta x^T \nabla^2 g_i(x)\Delta x}{2\mu} + \frac{\lambda_i(\Delta s_i)^2}{2\mu\ s_i} \\
&\quad + \frac{((\Delta s_{\rm CS})_i/2 - \Delta s_i)\,(\Delta s_{\rm CS})_i}{(s_i)^2} + \frac{|\Delta s_i - (\Delta s_{\rm CS})_i|^3}{(s_i)^3}.
\end{aligned}$$

Also, the definition of $R_f(x,\Delta x)$ (see (1.8)) implies

$$f(x_{\rm TR}) = f(x) + \nabla f(x)^T \Delta x + \frac{1}{2}\Delta x^T \nabla^2 f(x)\Delta x + R_f(x,\Delta x).$$

The above two inequalities, together with the definitions of f^μ and υ, yield (3.4). The subsequent claim readily follows.

(d) Since (3.1) holds, (a) implies $\|r_{\text{TR}}\|_p \le (1-\sigma_1\delta^\beta)\|r\|_p$. Since $\|S^{-1}\widehat{\Delta s}\|_\infty \le 2/3$, (c) implies that (3.4) holds. These two inequalities and the definition of $f^{\mu,\tau}$ (see (1.5)) yield (3.5). If $\upsilon < 0$, then the right-hand side of (3.5) is below $-\tau\sigma_1\delta^\beta\|r\|_p$ whenever $R_f(x,\Delta x)+\mu R(s,\Delta s,\Delta s_{\text{CS}}) \le -\upsilon$. Since $0<\eta<1$, the first claim readily follows. If for some $\chi>0$, (3.6) holds, then, whenever $R_f(x,\Delta x)+\mu R(s,\Delta s,\Delta s_{\text{CS}}) \le \chi\delta^\beta\lambda_{\text{P}}^T r$, we have from (3.5) that

$$\begin{aligned} f^{\mu,\tau}(x_{\text{TR}},s_{\text{TR}}) - f^{\mu,\tau}(x,s) &\le (1+2\chi)\delta^\beta\lambda_{\text{P}}^T r - \tau\sigma_1\delta^\beta\|r\|_p \\ &\le (1+2\chi)\delta^\beta\|\lambda_{\text{P}}\|_q\|r\|_p - \tau\sigma_1\delta^\beta\|r\|_p \\ &\le -\eta\tau\sigma_1\delta^\beta\|r\|_p. \end{aligned}$$

(e) Since $r\ge 0$ and (3.3) holds with $\theta := (1-\sigma_1\delta^{\beta_1})\|b\|_\infty$, (b) implies $\|b_{\text{TR}}\|_\infty \le (1-\sigma_1\delta^{\beta_1})\|b\|_\infty$. Since $\|S^{-1}\widehat{\Delta s}\|_\infty \le 2/3$, (c) implies that (3.4) holds. These two inequalities and the definition of $\widehat{f^{\mu,\tau}}$ (see (1.6)) yield the desired conclusion. ∎

We next have the following technical lemma that estimates the (scaled) stepsize and the remainder terms in terms of the trust-region radius δ.

LEMMA 3.2. *Fix $\mu>0$, $1\le p\le\infty$, $\beta_2>0$, $\delta_{\max}>0$, and $(x,s,\lambda)\in\Re^n\times\Re^{2m}_{++}$. For any $\delta\in(0,\delta_{\max}]$ and $\beta\in[0,\beta_2]$, let $(\Delta x,\Delta s)$ denote a feasible solution of (2.1) with λ_{P}, M, r, A given by (2.2). Then, the following results hold with b, Δs_{CS} given by (2.5), (2.6), and $\widehat{\Delta s} := \Delta s - \Delta s_{\text{CS}}$, $\hat\delta := \max\{\delta,\delta^\beta\}$, $\hat\delta_{\max} := \max\{\delta_{\max},(\delta_{\max})^{\beta_2},1\}$, $C_1 := \max\left\{\sum_{i=1}^m\|\nabla^2 g_i(x)\|, \|A^T\|+\sqrt{m}\|r\|_p\right\}$, $C_2 := 2\max\{C_1,C_1^2\}\max\{1,\|\lambda\|_\infty, \|\lambda_{\text{P}}\|_\infty^\ell\}_{\ell=1}^2$, $C_3 := \max\{C_2^2\hat\delta_{\max}, C_2^3\hat\delta_{\max}^3\}$.*

(a)

$$\|\Delta s_{\text{CS}}\|_\infty \le \frac{\hat\delta^2}{2\mu^2}(C_1\|b\|_\infty\mu + C_1\mu^2 + C_2\|b\|_\infty),$$

$$\|L_g(x,\Delta x) - \Delta s_{\text{CS}}\|_p \le \|R_g(x,\Delta x)\|_p + C_2\hat\delta^2\|b\|_\infty\left(\frac{1}{\mu}+\frac{1}{\mu^2}\right).$$

(b)

$$\|\Delta s\| \le C_1\hat\delta, \quad \|S^{-1}\Delta s\| \le \frac{C_2}{2}\frac{\hat\delta}{\mu}, \quad \|S^{-1}\widehat{\Delta s}\| \le C_2\hat\delta_{\max}\frac{\hat\delta}{\mu}\left(1+\frac{\|b\|_\infty}{\mu^2}\right).$$

(c)

$$R(s,\Delta s,\Delta s_{\text{CS}}) \le C_3\frac{\hat\delta^3}{\mu^3}\left(\mu\left(1+\frac{\|b\|_\infty}{\mu^2}\right)^2 + \left(1+\frac{\|b\|_\infty}{\mu^2}\right)^3\right).$$

Proof. Using $\|r\|\le\sqrt{m}\|r\|_p$ and the feasibility condition for (2.1), we have

$$\begin{aligned} \|\Delta s\| &\le \|A^T\|\delta + \|r\|\delta^\beta \\ &\le C_1\hat\delta, \\ \|S^{-1}\Delta s\| &\le \frac{\|\Delta s\|}{\min_i s_i} \\ &\le \frac{C_1\hat\delta}{\min_i s_i} \\ &= \frac{C_1\|\lambda_{\text{P}}\|_\infty\hat\delta}{\mu}. \end{aligned}$$

Also, we have

$$\begin{aligned}\|S^{-1}\Delta s_{\rm CS}\| &= \frac{1}{2\mu}\|\Lambda\Delta x^T\nabla^2 g(x)\Delta x + (S^{-1}\Delta S)^2 b\|\\ &\le \frac{1}{2\mu}\|\lambda\|_\infty\|\Delta x^T\nabla^2 g(x)\Delta x\|_1 + \frac{1}{2\mu}\|(S^{-1}\Delta S)S^{-1}\Delta s\|_1\|b\|_\infty\\ &\le \frac{1}{2\mu}\|\lambda\|_\infty C_1\delta^2 + \frac{1}{2\mu}\|S^{-1}\Delta s\|^2\|b\|_\infty\\ &\le \frac{1}{2\mu}\|\lambda\|_\infty C_1\delta^2 + \frac{1}{2\mu}\left(C_1\|\lambda_{\rm P}\|_\infty\frac{\hat\delta}{\mu}\right)^2\|b\|_\infty\\ &\le \frac{1}{2}C_2\frac{\hat\delta^2}{\mu} + C_2\frac{\hat\delta^2}{\mu^3}\|b\|_\infty\\ &= C_2\frac{\hat\delta^2}{\mu}\left(\frac{1}{2}+\frac{\|b\|_\infty}{\mu^2}\right),\end{aligned}$$

where the second inequality uses $\|\Delta x^T\nabla^2 g(x)\Delta x\|_1 = \sum_{i=1}^m |(\Delta x)^T\nabla^2 g_i(x)\Delta x| \le C_1\delta^2$.

(a) We have from the above estimates that

$$\begin{aligned}\|\Delta s_{\rm CS}\|_\infty &= \frac{1}{2\mu}\|S\Lambda\Delta x^T\nabla^2 g(x)\Delta x + (S^{-1}\Delta S)\Delta S b\|_\infty\\ &\le \frac{1}{2\mu}\|S\lambda\|_\infty\|\Delta x^T\nabla^2 g(x)\Delta x\|_1 + \frac{1}{2\mu}\|(S^{-1}\Delta S)\Delta s\|_\infty\|b\|_\infty\\ &\le \frac{1}{2\mu}(\|b\|_\infty+\mu)C_1\delta^2 + \frac{1}{2\mu}\|S^{-1}\Delta s\|_\infty\|\Delta s\|_\infty\|b\|_\infty\\ &\le \frac{1}{2\mu}(\|b\|_\infty+\mu)C_1\delta^2 + \frac{1}{2\mu}\left(C_1\|\lambda_{\rm P}\|_\infty\frac{\hat\delta}{\mu}\right)\left(C_1\hat\delta\right)\|b\|_\infty\\ &\le \frac{1}{2\mu}(\|b\|_\infty+\mu)C_1\delta^2 + \frac{1}{2}C_2\left(\frac{\hat\delta}{\mu^2}\right)^2\|b\|_\infty.\end{aligned}$$

Further,

$$\begin{aligned}&\|L_g(x,\Delta x) - \Delta s_{\rm CS}\|_p\\ &= \left\|R_g(x,\Delta x) - \frac{1}{2\mu}(S\Lambda-\mu I)\Delta x^T\nabla^2 g(x)\Delta x - \frac{1}{2\mu}S^{-1}(\Delta S)^2 b\right\|_p\\ &\le \|R_g(x,\Delta x)\|_p + \frac{1}{2\mu}\|S\lambda-\mu e\|_\infty\|\Delta x^T\nabla^2 g(x)\Delta x\|_1 + \frac{1}{2\mu}\|(S^{-1}\Delta S)\Delta s\|_1\|b\|_\infty\\ &\le \|R_g(x,\Delta x)\|_p + \frac{1}{2\mu}\|b\|_\infty C_1\delta^2 + \frac{1}{2\mu}\|S^{-1}\Delta s\|\|\Delta s\|\|b\|_\infty\\ &\le \|R_g(x,\Delta x)\|_p + \frac{1}{2\mu}\|b\|_\infty C_1\delta^2 + \frac{1}{2\mu}\left(C_1\|\lambda_{\rm P}\|_\infty\frac{\hat\delta}{\mu}\right)\left(C_1\hat\delta\right)\|b\|_\infty\\ &\le \|R_g(x,\Delta x)\|_p + C_2\|b\|_\infty\left(\frac{\hat\delta^2}{\mu}+\frac{\hat\delta^2}{\mu^2}\right),\end{aligned}$$

where the first inequality uses $\|\cdot\|_p \le \|\cdot\|_1$, the second inequality uses $b = S\lambda - \mu e$, and $\|(S^{-1}\Delta S)\Delta s\|_1 = (S^{-1}\Delta s)^T(\Delta s) \le \|S^{-1}\Delta s\|\|\Delta s\|$.

(b) We have from the above estimates and $1 \le \hat{\delta}_{\max}$, $\hat{\delta} \le \hat{\delta}_{\max}$ that

$$\begin{aligned}\|S^{-1}\widehat{\Delta s}\| &\le \|S^{-1}\Delta s\| + \|S^{-1}\Delta s_{\mathrm{CS}}\| \\ &\le C_1\|\lambda_{\mathrm{P}}\|_\infty \frac{\hat{\delta}}{\mu} + C_2\frac{\hat{\delta}^2}{\mu}\left(\frac{1}{2} + \frac{\|b\|_\infty}{\mu^2}\right) \\ &\le \frac{1}{2}C_2\frac{\hat{\delta}}{\mu} + C_2\frac{\hat{\delta}^2}{\mu}\left(\frac{1}{2} + \frac{\|b\|_\infty}{\mu^2}\right) \\ &\le C_2\hat{\delta}_{\max}\frac{\hat{\delta}}{\mu}\left(1 + \frac{\|b\|_\infty}{\mu^2}\right).\end{aligned}$$

(c) We have

$$\begin{aligned}R(s, \Delta s, \Delta s_{\mathrm{CS}}) &= \left(\frac{1}{2}\Delta s_{\mathrm{CS}} - \Delta s\right)^T S^{-2}\Delta s_{\mathrm{CS}} + \sum_{i=1}^m \left|\frac{\widehat{\Delta s}_i}{s_i}\right|^3 \\ &\le \left(\|S^{-1}\Delta s_{\mathrm{CS}}\| + \|S^{-1}\Delta s\|\right)\|S^{-1}\Delta s_{\mathrm{CS}}\| + \|S^{-1}\widehat{\Delta s}\|^3 \\ &\le C_2^2\hat{\delta}_{\max}\frac{\hat{\delta}^3}{\mu^2}\left(1 + \frac{\|b\|_\infty}{\mu^2}\right)\left(\frac{1}{2} + \frac{\|b\|_\infty}{\mu^2}\right) + C_2^3\hat{\delta}_{\max}^3\frac{\hat{\delta}^3}{\mu^3}\left(1 + \frac{\|b\|_\infty}{\mu^2}\right)^3 \\ &\le C_3\frac{\hat{\delta}^3}{\mu^3}\left(\mu\left(1 + \frac{\|b\|_\infty}{\mu^2}\right)^2 + \left(1 + \frac{\|b\|_\infty}{\mu^2}\right)^3\right),\end{aligned}$$

where the first inequality uses $\|\cdot\|_3 \le \|\cdot\|$, the second inequality uses the proof of (b), and the last inequality uses the definition of C_3. ∎

4. Convergence of trust-region strategy. Using Lemmas 3.1 and 3.2, we show below that Algorithm 1 terminates finitely under suitable assumptions. Moreover, δ_k is not too small at termination. The latter will be used to establish convergence to a second-order stationary point of (1.1) as $\mu \to 0$; see Corollary 6.2.

PROPOSITION 4.1. *Fix any $\mu > 0$, and let $1 \le p \le \infty$ and $\{(x^k, s^k, \lambda^k, \delta_k, \beta^k, \Delta x^k, \Delta s^k, \Delta s^k_{\mathrm{CS}}, x^k_{\mathrm{TR}}, s^k_{\mathrm{TR}}, \lambda^k_{\mathrm{TR}})\}_{k=0,1,\ldots}$ be generated by Algorithm 1, with $\lambda^k_{\mathrm{P}}, r^k, A^k$ given by (2.12). Suppose that $\{x^k\}$ is bounded. Then the following results hold:*

(a) *If there exists $\chi > 0$ such that, during phase 1,*

$$(1 + 2\chi)\|\lambda^k_{\mathrm{P}}\|_q \le (1 - \eta)\tau\sigma_1 \quad \text{whenever (2.15) holds,} \tag{4.1}$$

where $1/p + 1/q = 1$, then phase 1 terminates finitely, i.e., there exists an index k such that $\|r^k\|_p \le \epsilon_1$.

(b) *Phase 2 terminates finitely, i.e., there exists an index k such that $\|r^k\|_p \le \epsilon_1$ and $\|b^k\|_\infty \le \epsilon_2$.*

(c) *Phase 3 terminates finitely, i.e., there exists an index k such that $\|r^k\|_p \le \epsilon_1$, $\|b^k\|_\infty \le \epsilon_2$, and (2.20) holds.*

(d) *If $\mu^2 < \delta_{\mathrm{th}}$, then, for the index k of (c), either $\hat{\delta} := \delta_k/\omega_1 \ge \min\{1, \mu^2\}$ or there exists a $\Delta x \in \hat{\delta}\,\mathbb{B}$ satisfying one of the following:*

$$C_1^k\hat{\delta} \ge \frac{\mu}{\|\lambda^k\|_\infty}, \tag{4.2}$$

$$\frac{\|R_g(x^k, \Delta x)\|_p}{\hat{\delta}^{\beta_2}} + C_2^k\hat{\delta}^{2-\beta_2}\left(\frac{1}{\mu} + \frac{1}{\mu^2}\right)\epsilon_2 > \epsilon_1, \tag{4.3}$$

$$\frac{C_2^k\hat\delta^2}{2\mu^2}(\mu^2+3\mu\epsilon_2+3\epsilon_2)+\|R_g(x^k,\Delta x)\|_p>\frac{\epsilon_2}{2\nu}, \tag{4.4}$$

$$C_2^k\hat\delta>\frac{2\mu}{\sqrt{2\mu/\epsilon_2}+4}, \tag{4.5}$$

$$C_2^k\hat\delta_{\max}\frac{\hat\delta}{\mu}\left(1+\frac{\epsilon_2}{\mu^2}\right)\ge\frac{2}{3},\ \textit{or} \tag{4.6}$$

$$\frac{R_f(x^k,\Delta x)}{\hat\delta^2}+C_3^k\frac{\hat\delta}{\mu^2}\left(\mu\left(1+\frac{\epsilon_2}{\mu^2}\right)^2+\left(1+\frac{\epsilon_2}{\mu^2}\right)^3\right)\ge(1-\eta)\epsilon_4, \tag{4.7}$$

where $\hat\delta_{\max}$ *is defined as in Lemma* 3.2, $C_1^k:=\max\{\sum_{i=1}^m\|\nabla^2 g_i(x^k)\|,\ \|(A^k)^T\|+\sqrt{m}\|r^k\|_p\}$, $C_2^k:=2\max\{C_1^k,(C_1^k)^2\}\max\{1,\|\lambda^k\|_\infty,\|\lambda_{\rm P}^k\|_\infty^\ell\}_{\ell=1}^2$, $C_3^k:=\max\{(C_2^k)^2\hat\delta_{\max},(C_2^k)^3\hat\delta_{\max}^3\}$, $\nu:=\|\lambda_{\rm P}^k\|_\infty+\|\lambda^k\|_\infty C_2^k\hat\delta/\mu$.

Proof. Let b^k,c^k,Q^k and $r_{\rm TR}^k,b_{\rm TR}^k$ be given by (2.13) and (2.14). Also, let $\hat\delta_k:=\max\{\delta_k,(\delta_k)^{\beta^k}\}$ and $\widehat{\Delta s}^k:=\Delta s^k-\Delta s_{\rm CS}^k$. Since $s^0\ge-g(x^0)$ and $s_{\rm TR}^k\ge-g(x_{\rm TR}^k)$ for all k, then $s^k\ge-g(x^k)$ or, equivalently, $r^k\ge0$ for all k.

For each $k\in\{0,1,\ldots\}$, $(\Delta x^k,\Delta s^k,\lambda_{\rm TR}^k)$ is a ϖ-approximate solution of (2.1) associated with $(x^k,s^k,\lambda^k,\delta_k,\beta^k)$, and υ^k is its objective value. Letting

$$\begin{aligned}q_k(\Delta x):=\ &(\delta_k)^{\beta^k}(\lambda_{\rm P}^k)^Tr^k+(c^k)^T\Delta x+\frac{1}{2}(\delta_k)^{2\beta^k}\|(\Lambda^k)^{1/2}(S^k)^{-1/2}r^k\|^2\\&+(\delta_k)^{\beta^k}(r^k)^T\Lambda^k(S^k)^{-1}(A^k)^T\Delta x+\frac{1}{2}\Delta x^TQ^k\Delta x\quad\forall\Delta x\in\delta_k\mathbb{B}\end{aligned} \tag{4.8}$$

(cf. (2.4)), this implies that

$$\upsilon^k=q_k(\Delta x^k)\ \le\ \varpi\otimes q_k(\Delta x_{\rm c}^k)\ =\ \varpi_k q_k(\Delta x_{\rm c}^k)$$

for some $\varpi_k\in\{\varpi,1/\varpi\}$, where $\Delta x_{\rm c}^k:=-\delta_k\bar c^k$ and $\bar c^k:=c^k/\|c^k\|$ if $c^k\ne0$, and otherwise $\bar c^k:=0$. In the case of $\beta^k=\beta_2$, dividing both sides by $(\delta_k)^{\beta_2}$ and rearranging terms yields

$$\begin{aligned}\frac{\|c^k\|}{(\delta_k)^{\beta_2-1}}-(\lambda_{\rm P}^k)^Tr^k\le\ &\frac{1}{2}(\delta_k)^{\beta_2}\|(\Lambda^k)^{1/2}(S^k)^{-1/2}r^k\|^2-\frac{\upsilon^k/\varpi_k}{(\delta_k)^{\beta_2}}\\&-\delta_k(r^k)^T\Lambda^k(S^k)^{-1}(A^k)^T\bar c^k+\frac{1}{2}(\delta_k)^{2-\beta_2}(\bar c^k)^TQ^k\bar c^k.\end{aligned} \tag{4.9}$$

(a) Suppose that phase 1 does not terminate finitely. Then $\|r^k\|_p>\epsilon_1$ for all $k=0,1,\ldots$. Let $K_1:=\{k\in\{0,1,\ldots\}:\ \rho_k\ge\eta$ and $\lambda_{\rm TR}^k>0$ and (2.16), (2.17) hold$\}$. Then, by the definition of ρ_k and the updating rule,

$$\begin{aligned}&f^{\mu,\tau}(x^{k+1},s^{k+1})\le f^{\mu,\tau}(x^k,s^k)-\eta\tau\sigma_1(\delta_k)^{\beta^k}\|r^k\|_p&&\forall k\in K_1,\\&(x^{k+1},s^{k+1})=(x^k,s^k)&&\forall k\notin K_1.\end{aligned} \tag{4.10}$$

Also, (2.16) implies that $\|r^k\|_p$ is monotonically decreasing, and (2.17) implies that $\|b^k\|_\infty$ is monotonically decreasing whenever it is above ϵ_2. Thus $\{r^k\}$ and $\{b^k\}$ are bounded. Then $\{C_1^k\}$ is bounded.

By (4.10), $f^{\mu,\tau}(x^k,s^k)$ is monotonically decreasing with k. Since $\{x^k\}$ is bounded, the definitions of $f^{\mu,\tau}$ (see (1.5)) imply that $-\sum_{i=1}^m\ln(s_i^k)$ must be bounded above. Since $\{r^k\}$ is bounded so that $\{s^k\}$ is bounded, this implies that s^k is bounded away from zero componentwise, and thus $\{(S^k)^{-1}\}$ and $\{\lambda_{\rm P}^k\}$ are bounded. Then

$\{f^{\mu,\tau}(x^k, s^k)\}$ is bounded below and thus must converge. Also, since $\{b^k\}$ is bounded, the boundedness of $\{(S^k)^{-1}\}$ implies that $\{\lambda^k\}$ is bounded. Then $\{C_2^k\}$, $\{C_3^k\}$, and $\{\lambda_{\text{TR}}^k\}$ are bounded.

If $|K_1| < \infty$, then the updating rule (2.18) implies $\delta_{k+1} = \omega_1 \delta_k$ for all k sufficiently large, so that $\delta_k \to 0$. If $|K_1| = \infty$, then since $\{f^{\mu,\tau}(x^k, s^k)\}$ converges, by (4.10) and $\|r^k\|_p > \epsilon_1$ for all k, we have $\{\delta_k\}_{k \in K_1} \to 0$. Since $\delta_{k+1} \geq \delta_k$ for all $k \in K_1$, there must exist a subsequence K of $\{0, 1, \ldots\} \backslash K_1$ such that

$$\{\delta_k\}_{k \in K} \to 0. \tag{4.11}$$

By taking $k \in K$ sufficiently large, if necessary, we can assume that $\delta_k \leq \min\{1, \delta_{\text{th}}\}$ for all $k \in K$, so that β^k equals either β_1 or β_2 for $k \in K$. For each $k \in K$, either $\rho_k < \eta$ or $\lambda_{\text{TR}}^k \not> 0$ or (2.16) fails or (2.17) fails. If $\lambda_{\text{TR}}^k \not> 0$, then Lemmas 3.1(b) and 3.2(b) imply $C_1^k \hat{\delta}_k \geq \|\Delta s^k\| \geq \mu / \|\lambda^k\|_\infty$. Since $\{C_1^k\}$ and $\{\lambda^k\}$ are bounded and, by (4.11), $\{\hat{\delta}_k\}_{k \in K} \to 0$, this cannot hold for an infinite number of $k \in K$. If (2.16) fails, then Lemma 3.1(a) implies

$$\frac{\|L_g(x^k, \Delta x^k) - \Delta s_{\text{CS}}^k\|_p}{(\delta_k)^{\beta^k}} > (1 - \sigma_1)\|r^k\|_p > (1 - \sigma_1)\epsilon_1. \tag{4.12}$$

Since $\delta_k \leq \min\{1, \delta_{\text{th}}\}$, either $\beta^k = \beta_1$ with $\hat{\delta}_k = (\delta_k)^{\beta_1}$, or $\beta^k = \beta_2$ with $\hat{\delta}_k = \delta_k$. Since $\|\Delta x^k\| \leq \delta_k \to 0$ as $k \in K \to \infty$ (short for "$k \in K$, $k \to \infty$"), Lemma 3.2(a) implies that

$$\frac{\|L_g(x^k, \Delta x^k) - \Delta s_{\text{CS}}^k\|_p}{(\delta_k)^{\beta^k}} \leq \frac{\|R_g(x^k, \Delta x^k)\|_p}{(\delta_k)^{\beta^k}} + C_2^k \frac{(\hat{\delta}_k)^2}{(\delta_k)^{\beta^k}} \|b^k\|_\infty \left(\frac{1}{\mu} + \frac{1}{\mu^2}\right) \to 0 \quad \text{as } k \in K \to \infty, \tag{4.13}$$

and hence (4.12) cannot hold for an infinite number of $k \in K$. Also, Lemma 3.2(a)–(b), the boundedness of $\{C_1^k\}$, $\{C_2^k\}$, $\{b^k\}$, and $\{\hat{\delta}_k\}_{k \in K} \to 0$ imply

$$\begin{aligned} \|\Delta s_{\text{CS}}^k\|_\infty &\leq \frac{(\hat{\delta}_k)^2}{2\mu^2}(C_1^k \|b^k\|_\infty \mu + C_1^k \mu^2 + C_2^k \|b^k\|_\infty) \to 0 \qquad \text{as } k \in K \to \infty, \\ \|(S^k)^{-1} \Delta s^k\| &\leq \frac{C_2^k}{2} \frac{\hat{\delta}_k}{\mu} \to 0 \qquad \text{as } k \in K \to \infty. \end{aligned} \tag{4.14}$$

Also, (4.13) implies $\{\|L_g(x^k, \Delta x^k) - \Delta s_{\text{CS}}^k\|_\infty\}_{k \in K} \to 0$. Then, by Lemma 3.1(b) and the boundedness of $\{\lambda_{\text{TR}}^k\}$ and $\{b^k\}$, (2.17) holds for all $k \in K$ sufficiently large. Thus, it must be that $\rho_k < \eta$ for all $k \in K$ sufficiently large. By passing to a subsequence if necessary, we can assume that either (i) $\rho_k < \eta$ and $\beta^k = \beta_1$ for all $k \in K$ or (ii) $\rho_k < \eta$ and $\beta^k = \beta_2$ for all $k \in K$. Since x^k, b^k, C_3^k are bounded for all k and $\hat{\delta}_k = \delta_k$ or $\hat{\delta}_k = (\delta_k)^{\beta_1}$ for all $k \in K$, then $\beta_1 > 2/3$, $\|\Delta x^k\| \leq \delta_k \to 0$ as $k \in K \to \infty$, and Lemma 3.2(c) imply that

$$\begin{aligned} &\frac{R_f(x^k, \Delta x^k)}{(\delta_k)^2} \to 0 \quad \text{as } k \in K \to \infty, \\ &\frac{R(s^k, \Delta s^k, \Delta s_{\text{CS}}^k)}{(\delta_k)^2} \leq C_3^k \frac{(\hat{\delta}_k)^3}{(\delta_k)^2} \left(\frac{(C_4^k)^2}{\mu^2} + \frac{(C_4^k)^3}{\mu^3}\right) \to 0 \quad \text{as } k \in K \to \infty, \end{aligned} \tag{4.15}$$

where we denote for simplicity $C_4^k := 1 + \|b^k\|_\infty/\mu^2$. Similarly, we have from Lemma 3.2(b) and the boundedness of $\{C_2^k\}$ and $\{C_4^k\}$ that

$$\|(S^k)^{-1}\widehat{\Delta s}^k\| \le C_2^k \hat{\delta}_{\max}\hat{\delta}_k \frac{C_4^k}{\mu} \to 0 \quad \text{as } k \in K \to \infty. \tag{4.16}$$

Thus, we can assume by taking k sufficiently large that $\|(S^k)^{-1}\widehat{\Delta s}^k\|_\infty \le 2/3$ for all $k \in K$.

In case (i), for each $k \in K$, we have from $\delta_k \le \delta_{\text{th}}$ and $\beta^k = \beta_1$ that (2.15) holds, so, by (4.1), $(1+2\chi)\|\lambda_{\text{P}}^k\|_q \le (1-\eta)\tau\sigma_1$. Since $\rho_k < \eta$, Lemma 3.1(d) implies that either (ia) (4.12) holds, or (ib) $\upsilon^k > (1+\chi)(\delta_k)^{\beta_1}(\lambda_{\text{P}}^k)^T r^k$, or (ic)

$$R_f(x^k, \Delta x^k) + \mu R(s^k, \Delta s^k, \Delta s_{\text{CS}}^k) > \chi(\delta_k)^{\beta_1}(\lambda_{\text{P}}^k)^T r^k. \tag{4.17}$$

As argued earlier, subcase (ia) cannot occur for an infinite number of $k \in K$. Then, by further passing to a subsequence if necessary, we can assume that either (ib) holds for all $k \in K$ or (ic) holds for all $k \in K$. In subcase (ib), since $\upsilon^k = q_k(\Delta x^k)$, dividing both sides by $(\delta_k)^{\beta_1}$ and using (4.8) and $\|\Delta x^k\| \le \delta_k$ gives

$$\begin{aligned}(\delta_k)^{1-\beta_1}\|c^k\| + \frac{(\delta_k)^{\beta_1}}{2}\|(\Lambda^k)^{1/2}(S^k)^{-1/2}r^k\|^2 + \delta_k\|A^k\Lambda^k(S^k)^{-1}r^k\| + \frac{(\delta_k)^{2-\beta_1}}{2}\|Q^k\| \\ > \chi(\lambda_{\text{P}}^k)^T r^k \ \ge\ 0,\end{aligned}$$

where the second inequality uses $r^k \ge 0$ and $\lambda_{\text{P}}^k > 0$. This together with $\beta_1 < 1$, (4.11), and the boundedness of $\lambda^k,\ r^k, c^k, A^k, Q^k,\ (S^k)^{-1}$ implies that $\{(\lambda_{\text{P}}^k)^T r^k\}_{k\in K} \to 0$. In subcase (ic), dividing both sides of (4.17) by $(\delta_k)^{\beta_1}$ and using (4.15) yields $\{(\lambda_{\text{P}}^k)^T r^k\}_{k\in K} \to 0$.

In case (ii), if $\upsilon^k < 0$, then $\rho_k < \eta$ and Lemma 3.1(d) imply that either (4.12) holds or

$$R_f(x^k, \Delta x^k) + \mu R(s^k, \Delta s^k, \Delta s_{\text{CS}}^k) \ >\ (\eta - 1)\upsilon^k. \tag{4.18}$$

As argued earlier, (4.12) cannot occur for an infinite number of $k \in K$, so either (4.18) holds or $\upsilon^k \ge 0$ for all $k \in K$ sufficiently large. In the former case, dividing both sides of (4.18) by $(\delta_k)^2$ and using (4.15) yields

$$\liminf_{k\in K\to\infty} \frac{\upsilon^k}{(\delta_k)^2} \ge 0.$$

In the latter case, the above clearly also holds. Since (4.9) holds for all $k \in K$, this, together with $\beta_2 < 2$, (4.11), and the boundedness of $\lambda^k,\ \lambda_{\text{P}}^k,\ r^k, \bar{c}^k, A^k, Q^k,\ (S^k)^{-1}$ for all k, yields

$$\limsup_{k\in K\to\infty} \frac{\|c^k\|}{(\delta_k)^{\beta_2-1}} - (\lambda_{\text{P}}^k)^T r^k \le 0. \tag{4.19}$$

For each $k \in K$, since $\delta_k \le \delta_{\text{th}}$ and $\beta^k = \beta_2$ so that (2.15) fails, we must have $(\lambda_{\text{P}}^k)^T r^k \le \sigma_2\|c^k\|$. Then (4.19) implies that $\{\|c^k\|/(\delta_k)^{\beta_2-1}\}_{k\in K} \to 0$ and hence $\{(\lambda_{\text{P}}^k)^T r^k\}_{k\in K} \to 0$.

Thus, in either case (i) or (ii), we have $\{(\lambda_{\text{P}}^k)^T r^k\}_{k\in K} \to 0$. Since $\lambda_{\text{P}}^k = \mu(S^k)^{-1}e > 0$ and $r^k \ge 0$ so that

$$(\lambda_{\text{P}}^k)^T r^k \ge \left(\min_i(\lambda_{\text{P}}^k)_i\right)\|r^k\|_p = \frac{\mu\|r^k\|_p}{\left(\max_i s_i^k\right)} \quad \forall k,$$

this, together with the boundedness of $\{s^k\}$, implies $\{\|r^k\|_p\}_{k\in K} \to 0$. Thus $\|r^k\|_p < \epsilon_1$ for all $k \in K$ sufficiently large, contradicting $\|r^k\|_p > \epsilon_1$ for all k.

(b) Suppose that phase 2 does not terminate finitely. By part (a), phase 1 terminates finitely, so there exists a $k_2 \in \{0, 1, \ldots\}$ such that $k \geq k_2$ if and only if $\|r^k\|_p \leq \epsilon_1$. Let

$$K_2 := \{k \geq k_2 : (2.19) \text{ holds}\}.$$

Then, by the definition of ρ_k and the updating rule,

$$\begin{array}{ll} \bar{f}^{\mu,\bar{\tau}}(x^{k+1}, s^{k+1}, \lambda^{k+1}) \leq \bar{f}^{\mu,\bar{\tau}}(x^k, s^k, \lambda^k) - \eta\bar{\tau}\sigma_1(\delta_k)^{\beta_1}\|b^k\|_\infty & \forall k \in K_2, \\ (x^{k+1}, s^{k+1}, \lambda^{k+1}) = (x^k, s^k, \lambda^k) & \forall k \in \{k_2, k_2+1, \ldots\}\backslash K_2. \end{array} \tag{4.20}$$

Since $\|r^k\|_p \leq \epsilon_1$ and, by (2.17), $\|b^k\|_\infty$ is monotonically decreasing for $k \geq k_2$, $\{r^k\}$ and $\{b^k\}$ are bounded. Since $\{x^k\}$ is bounded and, by (4.20), $\bar{f}^{\mu,\bar{\tau}}(x^k, s^k, \lambda^k)$ is monotonically decreasing for $k \geq k_2$, an argument similar to that for part (a) shows that s^k, $(S^k)^{-1}$, $\lambda^k_{\rm P}$, λ^k are bounded for $k \geq k_2$ and that $\{\bar{f}^{\mu,\bar{\tau}}(x^k, s^k, \lambda^k)\}_{k\geq k_2}$ converges. Then C^k_1, C^k_2, C^k_3, $\lambda^k_{\rm TR}$ are bounded for $k \geq k_2$.

If $|K_2| < \infty$, then the updating rule (2.18) implies $\delta_{k+1} = \omega_1\delta_k$ for all $k \geq k_2$ sufficiently large, so that $\delta_k \to 0$. If $|K_2| = \infty$, then since $\{\bar{f}^{\mu,\bar{\tau}}(x^k, s^k, \lambda^k)\}_{k\geq k_2}$ converges and $\|b^k\|_\infty > \epsilon_2$ for all $k \geq k_2$, by (4.20), $\{\delta_k\}_{k\in K_2} \to 0$. Since $\delta_{k+1} \geq \delta_k$ for all $k \in K_2$, this shows there must exist a subsequence K of $\{k_2, k_2 + 1, \ldots\}\backslash K_2$ such that (cf. (4.11))

$$\{\delta_k\}_{k\in K} \to 0.$$

By taking k sufficiently large, we can further assume that $\delta_k \leq \min\{1, \delta_{\rm th}\}$ for all $k \in K$, so that $\beta^k = \beta_2$ and $\hat{\delta}_k = \delta_k$ for all $k \in K$. For each $k \in K$, we have from $k \notin K_2$ that (2.19) fails. Since $\{C^k_1\}$ and $\{\lambda^k\}$ are bounded, by using Lemmas 3.1(b) and 3.2(b), we obtain as in the proof of part (a) that $\lambda^k_{\rm TR} > 0$ for all $k \in K$ sufficiently large. If $\|r^k_{\rm TR}\|_p > \epsilon_1$, then Lemma 3.1(a) would imply

$$\frac{\|L_g(x^k, \Delta x^k) - \Delta s^k_{\rm CS}\|_p}{(\delta_k)^{\beta_2}} > \epsilon_1.$$

Since $\{C^k_2\}$, $\{b^k\}$ are bounded and $\|\Delta x^k\| \leq \delta_k \to 0$ as $k \in K \to \infty$, we obtain from Lemma 3.2(a) that (4.13) holds; thus the above inequality cannot occur for an infinite number of $k \in K$. Also, Lemma 3.2(a)–(b) and the boundedness of $\{C^k_1\}$, $\{C^k_2\}$, $\{b^k\}$ imply that (4.14) holds, and (4.13) implies $\{\|L_g(x^k, \Delta x^k) - \Delta s^k_{\rm CS}\|_\infty\}_{k\in K} \to 0$. Then, by Lemma 3.1(b) and the boundedness of $\{\lambda^k_{\rm TR}\}$ and $\{b^k\}$, (2.17) holds for all $k \in K$ sufficiently large. Thus, it must be that $\rho_k < \eta$ for all $k \in K$ sufficiently large. By a similar reasoning, (4.16) holds, so we can assume, by taking $k \in K$ sufficiently large, that $\rho_k < \eta$ and $\|(S^k)^{-1}\widehat{\Delta s}^k\|_\infty \leq 2/3$ for all $k \in K$. Then, for each $k \in K$, Lemma 3.1(e) implies that

$$\upsilon^k + R_f(x^k, \Delta x^k) + \mu R(s^k, \Delta s^k, \Delta s^k_{\rm CS}) > (1 - \eta)\bar{\tau}\sigma_1(\delta_k)^{\beta_1}\|b^k\|_\infty.$$

Also, by Lemma 3.2(c), (4.15) holds, so dividing both sides by $(\delta_k)^{\beta_1}$ and using (4.15) yields

$$\liminf_{k\in K\to\infty} \frac{\upsilon^k}{(\delta_k)^{\beta_1}} \geq (1 - \eta)\bar{\tau}\sigma_1\|b^k\|_\infty > (1 - \eta)\bar{\tau}\sigma_1\epsilon_2.$$

On the other hand, since $v^k = q_k(\Delta x^k)$ and $\|\Delta x^k\| \le \delta_k$, (4.8), together with $\beta^k = \beta_2 > 1 > \beta_1$ and the boundedness of $\lambda^k, \lambda^k_{\rm P}$, r^k, c^k, A^k, Q^k, $(S^k)^{-1}$ for all k, implies $\lim_{k\in K\to\infty} v^k/(\delta_k)^{\beta_1} = 0$, a contradiction.

(c) Suppose that phase 3 does not terminate finitely. By part (b), phase 2 terminates finitely, so there exists a $k_3 \in \{0, 1, \ldots\}$ such that $k \ge k_3$ if and only if $\|r^k\|_p \le \epsilon_1$ and $\|b^k\|_\infty \le \epsilon_2$. Let

$$K_3 := \left\{ k \ge k_3 : (2.22) \text{ holds and } \frac{v^k}{(\delta_k)^2} < -\epsilon_4 \right\}.$$

Then, by the definition of ρ_k and the updating rule,

$$\begin{aligned} &f^\mu(x^{k+1}, s^{k+1}) \le f^\mu(x^k, s^k) + \eta v^k \ \text{ and } \ v^k < 0 \quad \forall k \in K_3, \\ &(x^{k+1}, s^{k+1}) = (x^k, s^k) \qquad\qquad\qquad\qquad\quad \forall k \in \{k_3, k_3+1, \ldots\}\backslash K_3. \end{aligned} \tag{4.21}$$

Since $\|r^k\|_p \le \epsilon_1$ and $\|b^k\|_\infty \le \epsilon_2$ for $k \ge k_3$, $\{r^k\}$ and $\{b^k\}$ are bounded. Since $\{x^k\}$ is bounded and, by (4.21), $f^\mu(x^k, s^k)$ is monotonically decreasing for $k \ge k_3$, an argument similar to that for part (a) shows that s^k, $(S^k)^{-1}$, $\lambda^k_{\rm P}$, λ^k are bounded for $k \ge k_3$, and that $\{f^\mu(x^k, s^k)\}_{k\ge k_3}$ converges. Then C^k_1, C^k_2, C^k_3, $\lambda^k_{\rm TR}$ are bounded for $k \ge k_3$.

We claim that there exists a $\bar\delta > 0$ such that

$$\frac{v^k}{(\delta_k)^2} < -\epsilon_4 \quad \forall k \ge k_3 \text{ with } \delta_k \le \bar\delta. \tag{4.22}$$

(If not, there would exist a subsequence K of $\{k_3, k_3+1, \ldots\}$ such that $v^k/(\delta_k)^2 \ge -\epsilon_4$ for all $k \in K$ and $\{\delta_k\}_{k\in K} \to 0$. Then, for all $k \in K$ sufficiently large so that $\delta_k \le \delta_{\rm th}$ and hence $\beta^k = \beta_2 < 2$, (4.9) together with the boundedness of λ^k, r^k, $\bar c^k, A^k, Q^k$, $(S^k)^{-1}$ would imply (4.19). Thus (2.20) would hold for all $k \in K$ sufficiently large, contradicting the nontermination of phase 3.) If $|K_3| < \infty$, then the updating rule (2.18) implies $\delta_{k+1} = \omega_1\delta_k$ for all $k \ge k_3$ sufficiently large, so that $\delta_k \to 0$. If $|K_3| = \infty$, then since $\{f^\mu(x^k, s^k)\}_{k\ge k_3}$ converges, by (4.21), $\{v^k\}_{k\in K_3} \to 0$. This and the definition of K_3 imply $\{\delta_k\}_{k\in K_3} \to 0$. Since $\delta_{k+1} \ge \delta_k$ for all $k \in K_3$, this shows there must exist a subsequence K of $\{k_3, k_3+1, \ldots\}\backslash K_3$ such that (cf. (4.11))

$$\{\delta_k\}_{k\in K} \to 0.$$

By (4.22), we can assume that $v^k/(\delta_k)^2 < -\epsilon_4$ for all $k \in K$, so that (2.22) fails for all $k \in K$. By taking k sufficiently large, we can further assume that $\delta_k \le \min\{1, \delta_{\rm th}\}$ for all $k \in K$, so that $\beta^k = \beta_2$ and $\hat\delta_k = \delta_k$ for all $k \in K$. Since $\{C^k_1\}$ and $\{\lambda^k\}$ are bounded, by using Lemmas 3.1(b) and 3.2(b), we obtain as in the proof of part (a) that $\lambda^k_{\rm TR} > 0$ for all $k \in K$ sufficiently large. Since $\{C^k_2\}$ and $\{b^k\}$ are bounded, by using Lemmas 3.1(a) and 3.2(a), we obtain as in the proof of part (b) that (4.13) holds and $\|r^k_{\rm TR}\|_p \le \epsilon_1$ for all $k \in K$ sufficiently large. Also, Lemma 3.2(a)–(b) and the boundedness of $\{C^k_1\}$, $\{C^k_2\}$, $\{b^k\}$ imply that (4.14) holds, and (4.13) implies $\{\|L_g(x^k, \Delta x^k) - \Delta s^k_{\rm CS}\|_\infty\}_{k\in K} \to 0$. Then, by Lemma 3.1(b) (with $\theta := \epsilon_2$) and the boundedness of $\{\lambda^k_{\rm TR}\}$ and $\{b^k\}$, we obtain similarly to the proof of part (b) that $\|b^k_{\rm TR}\|_\infty \le \epsilon_2$ for all $k \in K$ sufficiently large. Thus, it must be that $\rho_k < \eta$ for all $k \in K$ sufficiently large. By similar reasoning, (4.16) holds, so we can assume, by taking $k \in K$ sufficiently large, that $\rho_k < \eta$ and $\|(S^k)^{-1}\widehat{\Delta s}^k\|_\infty \le 2/3$ for all $k \in K$.

Then, for each $k \in K$, $v^k < 0$ and Lemma 3.1(c) imply that (4.18) holds. Also, by Lemma 3.2(c), (4.15) holds. Dividing both sides of (4.18) by $(\delta_k)^2$ and using (4.15) yields

$$\liminf_{k \in K \to \infty} \frac{v^k}{(\delta_k)^2} \geq 0,$$

which contradicts $v^k/(\delta_k)^2 < -\epsilon_4$ for all $k \in K$.

(d) Assume that $\mu^2 < \delta_{\text{th}}$. Let k index the final iteration of phase 3 of Algorithm 1; i.e., (2.20) holds. If $\delta_k \geq \omega_1 \min\{1, \mu^2\}$, then the desired conclusion follows. Suppose instead $\delta_k < \omega_1 \min\{1, \mu^2\}$. If a successful (i.e., nonnull) step is taken at iteration $k-1$, then $\mu^2 > \delta_k \geq \delta_{k-1}$ and the updating rule would imply $\delta_k \geq \mu^2$, a contradiction. Thus it must be that a null step is taken at iteration $k-1$, i.e., $(x^k, s^k, \lambda^k) = (x^{k-1}, s^{k-1}, \lambda^{k-1})$. Then $\delta_k = \omega_1 \delta_{k-1}$, and we establish below the desired conclusion with $\hat{\delta} := \delta_{k-1}$ and $\Delta x := \Delta x^{k-1}$.

Using the first inequality of (2.20), as well as $\delta_k < \delta_{k-1}$ and $(\lambda_{\text{P}}^k, r^k, b^k, c^k) = (\lambda_{\text{P}}^{k-1}, r^{k-1}, b^{k-1}, c^{k-1})$ and the criterion for entering phase 3, we have

$$\|r^{k-1}\|_p \leq \epsilon_1, \quad \|b^{k-1}\| \leq \epsilon_2, \quad \text{and} \quad \frac{\|c^{k-1}\|}{(\delta_{k-1})^{\beta_2 - 1}} - (\lambda_{\text{P}}^{k-1})^T r^{k-1} \leq \epsilon_3.$$

Since the method did not terminate at iteration $k-1$, this implies that we must have

$$\frac{v^{k-1}}{(\delta_{k-1})^2} < -\epsilon_4.$$

Since a null step is taken at iteration $k-1$, this in turn implies that one of the following holds:

$$\text{(i) } \lambda_{\text{TR}}^{k-1} \not> 0, \quad \text{(ii) } \|r_{\text{TR}}^{k-1}\|_p > \epsilon_1, \quad \text{(iii) } \|b_{\text{TR}}^{k-1}\|_\infty > \epsilon_2, \quad \text{(iv) } \rho_{k-1} < \eta.$$

Since $\delta_{k-1} = \delta_k/\omega_1 < \min\{1, \mu^2\} < \delta_{\text{th}}$, then $\beta^{k-1} = \beta_2$ and $\hat{\delta}_{k-1} = \delta_{k-1}$. In case (i), Lemmas 3.1(b), 3.2(b) and $\lambda^k = \lambda^{k-1}$, $C_1^k = C_1^{k-1}$ imply

$$C_1^k \delta_{k-1} \geq \|\Delta s^{k-1}\| \geq \frac{\mu}{\|\lambda^k\|_\infty},$$

and so (4.2) holds. In case (ii), Lemma 3.1(a) and $x^k = x^{k-1}$ imply

$$\frac{\|L_g(x^k, \Delta x^{k-1}) - \Delta s_{\text{CS}}^{k-1}\|_p}{(\delta_{k-1})^{\beta_2}} > \epsilon_1,$$

and so (4.3) follows from Lemma 3.2(a). In case (iii), Lemma 3.1(b), $s^k = s^{k-1}$, and $\|b^{k-1}\|_\infty \leq \epsilon_2$ imply that

$$\begin{aligned} &\text{either } \|\Delta s_{\text{CS}}^{k-1}\|_\infty + \|L_g(x^k, \Delta x^{k-1}) - \Delta s_{\text{CS}}^{k-1}\|_\infty > \frac{\epsilon_2}{2\|\lambda_{\text{TR}}^{k-1}\|_\infty} \\ &\text{or} \quad \|(S^k)^{-1} \Delta s^{k-1}\|_\infty > \frac{1}{\sqrt{2 + 2\mu/\epsilon_2} + 2} \geq \frac{1}{\sqrt{2\mu/\epsilon_2} + 4}. \end{aligned}$$

Also, $\|\lambda_{\text{TR}}^{k-1}\|_\infty = \|\lambda_{\text{P}}^k - (S^k)^{-1} \Lambda^k \Delta s^{k-1}\|_\infty \leq \|\lambda_{\text{P}}^k\|_\infty + \|\lambda^k\|_\infty \|(S^k)^{-1} \Delta s^{k-1}\|_\infty$. Then (4.4)–(4.5) follow from Lemma 3.2(a)–(b), and $C_1^k = C_1^{k-1} \leq C_2^k = C_2^{k-1}$,

$\hat{\delta}_{k-1} = \delta_{k-1}$, $\|b^{k-1}\|_\infty \le \epsilon_2$. In case (iv), since $v^{k-1} < 0$, Lemma 3.1(c) and $x^k = x^{k-1}$, $s^k = s^{k-1}$ imply that

$$\text{either } \|(S^k)^{-1}\widehat{\Delta s}^{k-1}\|_\infty > \frac{2}{3}$$
$$\text{or} \quad R_f(x^k, \Delta x^{k-1}) + \mu R(s^k, \Delta s^{k-1}, \Delta s^{k-1}_{\rm CS}) > (\eta-1)v^{k-1} > (1-\eta)\epsilon_4(\delta_{k-1})^2.$$

Then (4.6)–(4.7) follow from Lemma 3.2(b)–(c) and $C_3^k = C_3^{k-1}$, $\hat{\delta}_{k-1} = \delta_{k-1}$, $\|b^{k-1}\|_\infty \le \epsilon_2$. □

Proposition 4.1 assumes that $\{x^k\}$ is bounded and $\|\lambda_{\rm P}^k\|_q$ is strictly below τ whenever (2.15) holds during phase 1. While the first assumption is reasonable (see Corollary 6.2 and the subsequent discussion), it is less obvious whether the second assumption is reasonable, since $\{\lambda_{\rm P}^k\}$ depends on τ. The following lemma gives a sufficient condition for $\|\lambda_{\rm P}^k\|_q$ to be bounded whenever (2.15) holds, *independent* of τ. Then, it suffices to choose τ large enough, namely,

$$\tau > \sup_{(2.15)\ \text{holds}} \frac{\|\lambda_{\rm P}^k\|_q}{((1-\eta)\sigma_1)},$$

and (4.1) would hold for any sufficiently small $\chi > 0$. More generally, if g_i is affine and $r_i^0 = 0$ for some i, then we have $r_i^k = 0$ for all k, and g_i can be removed from consideration when applying this lemma. In particular, if g is affine and $r^0 = 0$, i.e., the starting point is an interior feasible solution, then $r^k = 0$ for all k, and thus phase 1 is never entered and the second assumption is automatically satisfied.

LEMMA 4.2. *Suppose there exist $\zeta \ge 0$ and $\xi > 0$ such that $\mathcal{X}_\zeta$ is bounded and*

$$\xi\mathbb{B} \cap \{\textit{convex hull of } \nabla g_i(x),\ i \in I_\zeta(x)\} = \emptyset \quad \forall x \in \mathcal{X}_\zeta. \tag{4.23}$$

Then, for any $\mu > 0$, there exists $\kappa_{\zeta,\xi,\mu} > 0$ such that $\|\lambda_{\rm P}^k\|_q \le \kappa_{\zeta,\xi,\mu}$ for all $\sigma_2 \ge \zeta/\xi$, all $1 \le p, q \le \infty$ with $1/p + 1/q = 1$, and all sequences $(x^k, s^k, \lambda_{\rm P}^k) \in \Re^n \times \Re^{2m}_{++}$, $k = 0, 1, \ldots$, satisfying (2.15), $\|r^k\|_p \le \zeta$, and $S^k\lambda_{\rm P}^k \le \mu e$, where r^k and c^k are given by (2.12), (2.13).

Proof. Suppose the claim is false so that there exist $\sigma_2 \ge \zeta/\xi$, $1 \le p, q \le \infty$ with $1/p + 1/q = 1$, and $(x^k, s^k, \lambda_{\rm P}^k) \in \Re^n \times \Re^{2m}_{++}$, $k = 0, 1, \ldots$, satisfying (2.15), $\|r^k\|_p \le \zeta$, and $S^k\lambda_{\rm P}^k \le \mu e$, but $\|\lambda_{\rm P}^k\|_q \to \infty$. Then, $x^k \in \mathcal{X}_\zeta$ for all k, and so $\{x^k\}$ is bounded. Since $\|r^k\|_p \le \zeta$ for all k, $\{s^k\}$ is bounded. By passing to a subsequence, if necessary, we can assume that $(x^k, s^k, \lambda_{\rm P}^k/\|\lambda_{\rm P}^k\|_1)$ converges to some $(\bar{x}, \bar{s}, \bar{\lambda})$. Then $\|\bar{\lambda}\|_1 = 1$ and, since $\lambda_{\rm P}^k > 0$ for all k, $\bar{\lambda} \ge 0$. Let $J := \{i \in \{1, \ldots, m\} : \bar{\lambda}_i > 0\}$. Then, for each $i \in J$, $(\lambda_{\rm P}^k)_i \to \infty$, so $s_i^k(\lambda_{\rm P}^k)_i \le \mu$ implies $s_i^k \to 0$. This, together with

$$|g_i(x^k) + s_i^k| = |r_i^k| \le \|r^k\|_p \le \zeta$$

for all k, yields in the limit $|g_i(\bar{x})| \le \zeta$. Thus, $J \subseteq I_\zeta(\bar{x})$. Also, (2.15) implies

$$\|\nabla f(x^k) + \nabla g(x^k)\lambda_{\rm P}^k\| = \|c^k\| < \frac{(\lambda_{\rm P}^k)^T r^k}{\sigma_2} \le \frac{\|\lambda_{\rm P}^k\|_q \zeta}{\sigma_2}$$

for all k, and thus dividing both sides by $\|\lambda_{\rm P}^k\|_1$ and using $\|\cdot\|_q \le \|\cdot\|_1$ yields in the limit $\|\nabla g(\bar{x})\bar{\lambda}\| \le \zeta/\sigma_2 \le \xi$. This contradicts (4.23). □

A compactness argument shows that the assumption of Lemma 4.2 is *equivalent* to $\mathcal{X}_\zeta$ being bounded and

$$0 \notin \{\text{convex hull of } \nabla g_i(x),\ i \in I_\zeta(x)\} \tag{4.24}$$

for all $x \in \mathcal{X}_\zeta$.[4] By the Farkas lemma, (4.24) is equivalent to the existence of a $d \in \Re^n$ satisfying $\nabla g_i(x)^T d < 0$ for all $i \in I_\zeta(x)$. Thus, when $\zeta = 0$, (4.24) reduces to the well-known MFCQ at x; see, e.g., [5]. When $\zeta > 0$, (4.24) may be viewed as a relaxed MFCQ.

To illustrate (4.23), consider the following example.

Example 1. Assume

$$n = 2, \quad m = 3, \quad g_1(x) = -x_2, \quad g_2(x) = x_2 - 1, \quad g_3(x) = x_1^2 - x_2.$$

Then $\mathcal{X}_\zeta$ (with $p = \infty$) is bounded for any $\zeta \geq 0$. For $0 \leq \zeta < 1/2$, it can be seen that $I_\zeta(x) \in \{\{1\}, \{2\}, \{3\}, \{1,3\}, \{2,3\}\}$ for all $x \in \mathcal{X}_\zeta$, from which a bit of calculus shows that (4.23) holds for any $0 < \xi < (1 + 1/(1-2\zeta))^{-1}$. Notice that $\nabla g_i(x)$, $i \in I(x)$, are *not* linearly independent at the feasible solution $x = [0\ 0]^T$.

5. An interior-point trust-region method for (1.1). Below we describe formally our method for solving (1.1), which uses Algorithm 1 to solve (1.4) inexactly in the inner iterations. The barrier parameter μ is adjusted in the outer iterations.

ALGORITHM 2.

0. *Choose* $\mu_1 > 0$, $1 \leq p \leq \infty$, $1 < \beta_2 < 2$, $0 < \omega_1 < 1$, $\varpi \geq 1$, $0 < \delta_{\rm th} \leq \delta_0 \leq \delta_{\max}$, $(x^0, s^0, \lambda^0) \in \Re^n \times \Re^{2m}_{++}$ *with* $s^0 \geq -g(x^0)$, $0 < \omega_2 < 1$, *and functions* $\pi_1, \pi_2, \pi_3, \pi_4 \in \Pi$. *Initialize* $t := 1$. *Go to step* 1.

1. *Apply Algorithm* 1 *with input* μ_t, p, β_2, ω_1, ϖ, $\delta_{\rm th}$, δ_{t-1}, $\delta_{\max}$, $(x^{t-1}, s^{t-1}, \lambda^{t-1})$ *and termination tolerances* $\pi_j(\mu_t)$, $j = 1,2,3,4$. *The method generates an* $(x^t, s^t, \lambda^t) \in \Re^n \times \Re^{2m}_{++}$ *and* $\delta_t \in (0, \delta_{\max}]$, $\beta^t \in [0, \beta_2]$, *and* $\upsilon^t \in \Re$ *satisfying*

$$\begin{aligned} &\|r^t\|_p \leq \pi_1(\mu_t), \qquad \|b^t\|_\infty \leq \pi_2(\mu_t), \\ &\frac{\|c^t\|}{(\delta_t)^{\beta_2 - 1}} \leq (\lambda_{\rm P}^t)^T r^t + \pi_3(\mu_t), \qquad \frac{\upsilon^t}{(\delta_t)^2} \geq -\pi_4(\mu_t), \end{aligned} \tag{5.1}$$

where $r^t := g(x^t) + s^t$, $\lambda_{\rm P}^t := \mu_t (S^t)^{-1} e$, $b^t := S^t \lambda^t - \mu_t e$, $c^t := \nabla_x l(x^t, \lambda_{\rm P}^t)$, *and* υ^t *is the objective value of some* ϖ*-approximate solution of* (2.1) *associated with* $(x^t, s^t, \lambda^t, \delta_t, \beta^t)$. *Moreover,* $\beta^t = \beta_2$ *whenever* $\delta_t \leq \delta_{\rm th}$. *Go to step* 2.

2. *If* μ_t *is below a desired threshold, terminate the method. Otherwise, choose any* $0 < \mu_{t+1} \leq \omega_2 \mu_t$. *Increment* t *by* 1 *and return to step* 1.

To obtain the desired convergence results, we need $\pi_j(\mu) \to 0$ as $\mu \to 0$, $j = 1,2,3,4$, sufficiently fast but not too fast relative to each other. In particular, we make the following (relative) local growth assumptions on $\pi_1, \ldots, \pi_4$:

A1.

(a) $\lim_{\mu \to 0} \pi_1(\mu)/\mu = 0$ and $\limsup_{\mu \to 0} \pi_3(\mu)/\pi_1(\mu) < \infty$.

(b) $\lim_{\mu \to 0} \pi_2(\mu)/\mu^2 = 0$.

(c) $\lim_{\mu \to 0} \pi_1(\mu)/(\mu \pi_2(\mu))^{(2-\beta_2)/2} = 0$.

(d) $\lim_{\mu \to 0} \pi_1(\mu)/(\mu^2 \pi_4(\mu))^{2-\beta_2} = 0$.

Notice that A1 is satisfied if we choose any $\pi_4 \in \Pi$ and any $\pi_0 \in \Pi$ and then set

$$\pi_2(\mu) := \pi_0(\mu)\mu^2, \quad \pi_1(\mu) := \pi_3(\mu) := \pi_0(\mu) \cdot \min\left\{\mu, (\mu \pi_2(\mu))^{\frac{(2-\beta_2)}{2}}, (\mu^2 \pi_4(\mu))^{2-\beta_2}\right\}.$$

[4] Assume that $\mathcal{X}_\zeta$ is bounded but (4.23) is false, so that there exist $I \subseteq \{1, \ldots, m\}$, $x^k \in \mathcal{X}_\zeta$, and $(\lambda_{\rm P}^k)_i \geq 0$ for $i \in I$ such that $I_\zeta(x^k) = I$, $\sum_{i \in I} (\lambda_{\rm P}^k)_i = 1$ for all k, and $\sum_{i \in I} \nabla g_i(x^k)(\lambda_{\rm P}^k)_i \to 0$. Then, any cluster point $(\bar{x}, \bar{\lambda}_i)_{i \in I}$ satisfies $\bar{x} \in \mathcal{X}_\zeta$, $\bar{\lambda}_i \geq 0$ for $i \in I$, and $I_\zeta(\bar{x}) = I$, $\sum_{i \in I} \bar{\lambda}_i = 1$, $\sum_{i \in I} \nabla g_i(\bar{x}) \bar{\lambda}_i = 0$, and thus (4.24) does not hold for all $x \in \mathcal{X}_\zeta$. The converse is obvious.

6. Convergence of the interior-point trust-region method. We first have the following proposition, which is algorithm-independent. The global convergence of Algorithm 2 follows as its corollary.

PROPOSITION 6.1. *Fix $1 \le p, q \le \infty$ with $1/p + 1/q = 1$, $1 < \beta_2 < 2$, $\varpi \ge 1$, $0 < \delta_{\rm th} \le \delta_{\max}$, and $\pi_1, \pi_2, \pi_3, \pi_4 \in \Pi$. Suppose for each $t = 1, 2, \ldots$ that we have $\mu_t > 0$, $(x^t, s^t, \lambda^t) \in \Re^n \times \Re^{2m}_{++}$, $\delta_t \in (0, \delta_{\max}]$, $\beta^t \in [0, \beta_2]$, and $v^t \in \Re$ satisfying (5.1) and that v^t is the objective value of some ϖ-approximate solution of (2.1) associated with $(x^t, s^t, \lambda^t, \delta_t, \beta^t)$, with $r^t, \lambda^t_{\rm P}, b^t, c^t$ defined as in Algorithm 2. Suppose further that $\mu_t \to 0$ and there exists a subsequence T of $\{1, 2, \ldots\}$ such that $\{x^t\}_{t\in T}$ converges to some $\bar{x}$. Then the following results hold:*

(a) *$\bar{x}$ is a feasible solution of (1.1).*

(b) *If $\bar{x}$ satisfies (4.24) with $\zeta = 0$, then $\{\lambda^t_{\rm P}\}_{t\in T}$ is bounded and every cluster point $\bar{\lambda}$ satisfies (1.2).*

(c) *If $\pi_1, \ldots, \pi_4$ satisfy* A1 *and, in addition to the assumption of* (b)*, we have $\beta^t = \beta_2$ whenever $\delta_t \le \delta_{\rm th}$, and the following assumptions* A2 *and* A3 *are satisfied, then $\{\lambda^t\}_{t\in T'} \to \bar{\lambda}$ and (1.7) holds for any $T' \subseteq T$ such that $\{(x^t, \lambda^t_{\rm P})\}_{t\in T'} \to (\bar{x}, \bar{\lambda})$ and any $d^t \in \Re^n$ satisfying both $\|d^t\| = 1$ and $[\nabla g_i(x^t)^T d^t]_{i\in I(\bar{x})} = 0$ for all $t \in T'$.*

A2. $\limsup_{\substack{x\to\bar{x}\\ \delta\to 0}} \sup_{\|\Delta x\|\le\delta} R_f(x, \Delta x)/\delta^3 < \infty$.

A3. *For each t with $(\mu_t)^2 < \delta_{\rm th}$, either $\hat{\delta}_t := \delta_t/\omega_1 \ge \min\{1, (\mu_t)^2\}$ or there exists a $\Delta x^t \in \hat{\delta}_t\, \mathbb{B}$ satisfying one of the following:*

$$C_1^t \hat{\delta}_t \ge \frac{\mu_t}{\|\lambda^t\|_\infty}, \tag{6.1}$$

$$\frac{\|R_g(x^t, \Delta x^t)\|_p}{(\hat{\delta}_t)^{\beta_2}} + C_2^t (\hat{\delta}_t)^{2-\beta_2} \left(\frac{1}{\mu_t} + \frac{1}{(\mu_t)^2}\right) \pi_2(\mu_t) > \pi_1(\mu_t), \tag{6.2}$$

$$C_2^t \frac{(\hat{\delta}_t)^2}{2(\mu_t)^2} \pi_5(\mu^t) + \|R_g(x^t, \Delta x^t)\|_p > \frac{\pi_2(\mu_t)}{2\nu_t} m, \tag{6.3}$$

$$C_2^t \hat{\delta}_t > \frac{2\mu_t}{\sqrt{2\mu_t/\pi_2(\mu_t)} + 4}, \tag{6.4}$$

$$C_2^t \hat{\delta}_{\max} \frac{\hat{\delta}_t}{\mu_t} C_4^t \ge \frac{2}{3}, \tag{6.5}$$

$$\frac{R_f(x^t, \Delta x^t)}{(\hat{\delta}_t)^2} + C_3^t \frac{\hat{\delta}_t}{(\mu_t)^2} \left(\mu_t \left(C_4^t\right)^2 + \left(C_4^t\right)^3\right) \ge (1 - \eta)\pi_4(\mu_t), \tag{6.6}$$

where $\pi_5(\mu) := \mu^2 + 3(\mu + 1)\pi_2(\mu)$, $\hat{\delta}_{\max}$ is defined as in Lemma 3.2, $A^t := \nabla g(x^t)$, $C_1^t := \max\left\{\sum_{i=1}^m \|\nabla^2 g_i(x^t)\|, \|(A^t)^T\| + \sqrt{m}\pi_1(\mu_t)\right\}$, $C_2^t := 2\max\{C_1^t, (C_1^t)^2\}\max\{1, \|\lambda^t\|_\infty, \|\lambda^t_{\rm P}\|^\ell_\infty\}^2_{\ell=1}$, $C_3^t := \max\{(C_2^t)^2\hat{\delta}_{\max}, (C_2^t)^3\hat{\delta}^3_{\max}\}$, $C_4^t := 1 + \pi_2(\mu_t)/(\mu_t)^2$, and $\nu_t := \|\lambda^t_{\rm P}\|_\infty + \|\lambda^t\|_\infty C_2^t \hat{\delta}^t/\mu_t$.

Proof. By passing to a subsequence if necessary, we can without loss of generality assume that $x^t \to \bar{x}$.

(a) Since $s^t > 0$ and $\|r^t\|_p \le \pi_1(\mu_t) \to 0$, it follows from $x^t \to \bar{x}$ that $g(\bar{x}) = -\lim_{t\to\infty} s^t \le 0$.

(b) If $\{\lambda^t_{\rm P}\}$ is not bounded, so that there exists a subsequence T of $\{1, 2, \ldots\}$ such that $\{\|\lambda^t_{\rm P}\|_1\}_{t\in T} \to \infty$, then the first and third inequalities in (5.1) and $\|\cdot\|_q \le \|\cdot\|_1$

would yield

$$\frac{\|\nabla f(x^t)+\nabla g(x^t)\lambda_{\rm P}^t\|}{\|\lambda_{\rm P}^t\|_1}=\frac{\|c^t\|}{\|\lambda_{\rm P}^t\|_1}\le\frac{\|\lambda_{\rm P}^t\|_q\pi_1(\mu_t)+\pi_3(\mu_t)}{\|\lambda_{\rm P}^t\|_1}(\delta_t)^{\beta_2-1}\to 0 \text{ as } t\in T\to\infty.$$

Passing to the limit yields $\nabla g(\bar{x})\bar{\lambda}=0$, where $\bar{\lambda}\ge 0$ is any cluster point of $\{\lambda_{\rm P}^t/\|\lambda_{\rm P}^t\|_1\}_{t\in T}$. Since $\|\bar{\lambda}\|_1=1$ and, for each i with $\bar{\lambda}_i>0$, we have $\{(\lambda_{\rm P}^t)_i\}_{t\in T}\to\infty$ so that $g_i(\bar{x})=-\lim_{t\to\infty}s_i^t=-\lim_{t\to\infty}\mu_t/(\lambda_{\rm P}^t)_i=0$, this would contradict $\bar{x}$ satisfying (4.24) with $\zeta=0$. Hence $\{\lambda_{\rm P}^t\}$ is bounded.

Since $\{\lambda_{\rm P}^t\}$ is bounded, then the boundedness of $\{\delta_t\}$ and the first and third inequalities in (5.1) yield

$$\|c^t\|\le(\|\lambda_{\rm P}^t\|\pi_1(\mu_t)+\pi_3(\mu_t))(\delta_t)^{\beta_2-1}\to 0. \tag{6.7}$$

Since $S^t\lambda_{\rm P}^t=\mu_t e\to 0$ and $s^t\to -g(\bar{x})$, this implies that any cluster point $\bar{\lambda}$ of $\{\lambda_{\rm P}^t\}$ satisfies $\bar{\lambda}\ge 0$, $g(\bar{x})^T\bar{\lambda}=0$, and $\nabla_x l(\bar{x},\bar{\lambda})=0$.

(c) Consider any $T'\subseteq T$ such that $\{(x^t,\lambda_{\rm P}^t)\}_{t\in T'}\to(\bar{x},\bar{\lambda})$ and any $d^t\in\Re^n$ satisfying $\|d^t\|=1$ and $[\nabla g_i(x^t)^T d^t]_{i\in I(\bar{x})}=0$ for all $t\in T'$. Since

$$\frac{\pi_2(\mu_t)}{\mu_t}\ge\frac{\|b^t\|_\infty}{\mu_t}=\frac{\|S^t(\lambda^t-\lambda_{\rm P}^t)\|_\infty}{\mu_t}\ge\frac{\min_i s_i^t\|\lambda^t-\lambda_{\rm P}^t\|_\infty}{\mu_t}=\frac{\|\lambda^t-\lambda_{\rm P}^t\|_\infty}{\|\lambda_{\rm P}^t\|_\infty},$$

we have $\|\lambda^t-\lambda_{\rm P}^t\|_\infty\le\|\lambda_{\rm P}^t\|_\infty\pi_2(\mu_t)/\mu_t$. Since $\{\lambda_{\rm P}^t\}_{t\in T'}\to\bar{\lambda}$ and $\lim_{\mu\to 0}\pi_2(\mu)/\mu=0$ by A1(b), we obtain $\{\lambda^t\}_{t\in T'}\to\bar{\lambda}$.

For each $t\in T'$, we have from $[\nabla g_i(x^t)^T d^t]_{i\in I(\bar{x})}=0$ that

$$(d^t)^T\left(\frac{\lambda_i^t}{s_i^t}\nabla g_i(x^t)\nabla g_i(x^t)^T\right)d^t=0\quad\forall i\in I(\bar{x}). \tag{6.8}$$

For each $i\notin I(\bar{x})$, we have $s_i^t\to -g_i(\bar{x})>0$ and $\lambda_i^t\to\bar{\lambda}_i=0$ as $t\in T'\to\infty$, so $\lambda_i^t/s_i^t\to 0$, and hence the left-hand side of (6.8) tends to zero. Summing this term over all $i=1,\ldots,m$ yields

$$(d^t)^T A^t\Lambda^t(S^t)^{-1}(A^t)^T d^t\to 0\quad\text{ as } t\in T'\to\infty. \tag{6.9}$$

For each $t\in T'$, we have from the definition of υ^t (see (2.11)) that $\upsilon^t\le\varpi\otimes\upsilon_{\rm Q}^t=\varpi_t\upsilon_{\rm Q}^t$ for some $\varpi_t\in\{\varpi,1/\varpi\}$, where

$$\begin{aligned}\upsilon_{\rm Q}^t:=&(\delta_t)^{\beta^t}(\lambda_{\rm P}^t)^T r^t+\delta_t\|c^t\|+\frac{1}{2}(\delta_t)^{2\beta^t}\|(\Lambda^t)^{1/2}(S^t)^{-1/2}r^t\|^2\\&+(\delta_t)^{\beta^t+1}\|A^t\Lambda^t(S^t)^{-1}r^t\|+\frac{1}{2}(\delta_t)^2(\varpi\otimes\gamma_*^t+\varpi\mu_t),\end{aligned}$$

$Q^t:=\nabla_{xx}l(x^t,\lambda^t)+A^t\Lambda^t(S^t)^{-1}(A^t)^T$, and γ_*^t denotes the minimum eigenvalue of Q^t. Also, $\varpi\otimes\gamma_*^t=\varphi_t\gamma_*^t$ for some $\varphi_t\in\{\varpi,1/\varpi\}$. Using (5.1) and $\gamma_*^t\le(d^t)^TQ^td^t$ yields

$$\begin{aligned}-\frac{\pi_4(\mu_t)}{\varpi_t}&\le\frac{\upsilon^t}{\varpi_t(\delta_t)^2}\\&\le\frac{\upsilon_{\rm Q}^t}{(\delta_t)^2}\\&\le\frac{(\lambda_{\rm P}^t)^Tr^t}{(\delta_t)^{2-\beta^t}}+\frac{\|c^t\|}{\delta_t}+\frac{1}{2}(\delta_t)^{2\beta^t-2}\|(\Lambda^t)^{1/2}(S^t)^{-1/2}r^t\|^2\\&\quad+(\delta_t)^{\beta^t-1}\|A^t\Lambda^t(S^t)^{-1}r^t\|+\frac{1}{2}(\varphi_t(d^t)^TQ^td^t+\varpi\mu_t).\end{aligned} \tag{6.10}$$

Moreover, (5.1) and $\lambda_{\rm P}^T r \le \|\lambda_{\rm P}\|_q \|r\|_p$ give

$$\begin{aligned}\frac{(\lambda_{\rm P}^t)^T r^t}{(\delta_t)^{2-\beta^t}} + \frac{\|c^t\|}{\delta_t} &\le \frac{(\lambda_{\rm P}^t)^T r^t}{(\delta_t)^{2-\beta^t}} + \frac{(\lambda_{\rm P}^t)^T r^t + \pi_3(\mu_t)}{(\delta_t)^{2-\beta_2}} \\ &\le \frac{\|\lambda_{\rm P}^t\|_q \pi_1(\mu_t)}{(\delta_t)^{2-\beta^t}} + \frac{\|\lambda_{\rm P}^t\|_q \pi_1(\mu_t) + \pi_3(\mu_t)}{(\delta_t)^{2-\beta_2}}. \end{aligned} \tag{6.11}$$

Since $\mu_t \to 0$, by taking t sufficiently large if necessary, we can assume that $(\mu_t)^2 < \delta_{\rm th}$ for all $t \in T'$. For each $t \in T'$, either (i) $\hat\delta_t \ge \min\{1, \delta_{\rm th}\}$, (ii) $(\mu_t)^2 \le \hat\delta_t < \min\{1, \delta_{\rm th}\}$, or (iii) $\hat\delta_t < \min\{1, (\mu_t)^2\}$. In cases (ii) and (iii), we have $\delta_t = \omega_1 \hat\delta_t < \delta_{\rm th}$ and hence $\beta^t = \beta_2$. In case (iii), since $(\mu_t)^2 < \delta_{\rm th}$, A3 implies that there exists $\Delta x^t \in \hat\delta_t\, \mathbb{B}$ satisfying either (iiia) (6.1), (iiib) (6.2), (iiic) (6.3), (iiid) (6.4), (iiie) (6.5), or (iiif) (6.6), with π_5, A^t, $C_1^t, C_2^t, C_3^t, C_4^t$, ν_t as defined therein. By further passing to a subsequence if necessary, we can assume that we are either in case (i) for all $t \in T'$, in case (ii) for all $t \in T'$, or in subcase (iiia) for all $t \in T'$ or ..., etc. We show below that the right-hand side of (6.11) tends to zero as $t \in T' \to \infty$ in each of these cases and subcases. Since $x^t, \lambda_{\rm P}^t, \lambda^t$ are bounded for all t, then C_1^t, C_2^t, C_3^t are bounded for all $t \in T'$.

In case (i), since $\min\{1, \delta_{\rm th}\} \le \hat\delta_t = \delta_t/\omega_1$ for all $t \in T'$, the boundedness of $\{\lambda_{\rm P}^t\}$ and $\pi_1(\mu_t) \to 0$, $\pi_3(\mu_t) \to 0$ imply that the right-hand side of (6.11) tends to zero as $t \in T' \to \infty$. In cases (ii) and (iii), since $\{\lambda_{\rm P}^t\}$ is bounded, $\beta^t = \beta_2$, and $\limsup_{\mu\to 0} \pi_3(\mu)/\pi_1(\mu) < \infty$ by A1(a), it suffices to show that $\pi_1(\mu_t)/(\delta_t)^{2-\beta_2} \to 0$ or, equivalently, $\pi_1(\mu_t)/(\hat\delta_t)^{2-\beta_2} \to 0$ as $t \in T' \to \infty$. We do this below.

In case (ii), we have from A1(d) that

$$\frac{\pi_1(\mu_t)}{(\hat\delta_t)^{2-\beta_2}} \le \frac{\pi_1(\mu_t)}{((\mu_t)^2)^{2-\beta_2}} \;\to\; 0 \quad \text{as } t \in T' \to \infty.$$

In subcase (iiia), we have from (6.1), A1(d), and the boundedness of $\{C_1^t\}$, $\{\lambda^t\}$ that

$$\frac{\pi_1(\mu_t)}{(\hat\delta_t)^{2-\beta_2}} \le \frac{\pi_1(\mu_t)}{(\mu_t)^{2-\beta_2}} \left(C_1^t \|\lambda^t\|_\infty\right)^{2-\beta_2} \to 0 \quad \text{as } t \in T' \to \infty.$$

In subcase (iiib), since $\|\Delta x^t\| \le \hat\delta_t < (\mu_t)^2 \to 0$ and $\{C_2^t\}$ is bounded, we obtain from (6.2) and A1(b) that

$$\frac{\pi_1(\mu_t)}{(\hat\delta_t)^{2-\beta_2}} < \frac{\|R_g(x^t, \Delta x^t)\|_p}{(\hat\delta_t)^2} + C_2^t \left(\frac{1}{\mu_t} + \frac{1}{(\mu_t)^2}\right) \pi_2(\mu_t) \to 0 \quad \text{as } t \in T' \to \infty.$$

In subcase (iiic), we see from A1(b) that $\lim_{\mu\to 0} \pi_5(\mu)/\mu^2 = 1$. Since $\{\|R_g(x^t, \Delta x^t)\|_p / (\hat\delta_t)^2\}$ and $\{C_2^t\}$ are bounded, then (6.3) implies that $\{\pi_2(\mu_t)/((\hat\delta_t)^2 \nu_t)\}_{t\in T'}$ is bounded. Since $\nu_t = \|\lambda_{\rm P}^t\|_\infty + \|\lambda^t\|_\infty C_2^t \hat\delta^t/\mu_t$ and $\{\lambda_{\rm P}^t\}$, $\{\lambda^t\}$, $\{C_2^t\}$ are bounded, this implies, passing to a subsequence if necessary, that either $\{\pi_2(\mu_t)/(\hat\delta_t)^2\}_{t\in T'}$ or $\{\mu_t \pi_2(\mu_t)/(\hat\delta_t)^3\}_{t\in T'}$ is bounded. Then we obtain from A1(c) that

$$\text{either} \quad \frac{\pi_1(\mu_t)}{(\hat\delta_t)^{2-\beta_2}} = \frac{\pi_1(\mu_t)}{(\pi_2(\mu_t))^{(2-\beta_2)/2}} \cdot \left(\frac{\pi_2(\mu_t)}{(\hat\delta_t)^2}\right)^{(2-\beta_2)/2} \to 0$$

$$\text{or} \quad \frac{\pi_1(\mu_t)}{(\hat\delta_t)^{2-\beta_2}} = \frac{\pi_1(\mu_t)}{(\mu_t \pi_2(\mu_t))^{(2-\beta_2)/3}} \cdot \left(\frac{\mu_t \pi_2(\mu_t)}{(\hat\delta_t)^3}\right)^{(2-\beta_2)/3} \to 0 \quad \text{as } t \in T' \to \infty.$$

In subcase (iiid), we have from (6.4) and the boundedness of $\{C_2^t\}$ and $\{\pi_2(\mu_t)/\mu_t\}$ $\to 0$ that $\{\sqrt{\mu_t \pi_2(\mu_t)}/\hat{\delta}_t\}_{t\in T'}$ is bounded, and thus A1(c) yields

$$\frac{\pi_1(\mu_t)}{(\hat{\delta}_t)^{2-\beta_2}} = \frac{\pi_1(\mu_t)}{(\mu_t \pi_2(\mu_t))^{(2-\beta_2)/2}} \cdot \left(\frac{\sqrt{\mu_t \pi_2(\mu_t)}}{\hat{\delta}_t} \right)^{2-\beta_2} \to 0 \quad \text{as } t \in T' \to \infty.$$

In subcase (iiie), we have from A1(b) that $\{C_4^t\} \to 1$. Then (6.5) and the boundedness of $\{C_2^t\}$ imply that $\{\mu_t/\hat{\delta}_t\}_{t\in T'}$ is bounded. This, together with A1(a) and $\beta_2 > 1$, yields

$$\frac{\pi_1(\mu_t)}{(\hat{\delta}_t)^{2-\beta_2}} = \frac{\pi_1(\mu_t)}{(\mu_t)^{2-\beta_2}} \cdot \left(\frac{\mu_t}{\hat{\delta}_t} \right)^{2-\beta_2} \to 0 \quad \text{as } t \in T' \to \infty.$$

In subcase (iiif), we have, upon multiplying both sides of (6.6) by $(\mu_t)^2/\hat{\delta}_t$, that

$$(\mu_t)^2 \frac{R_f(x^t, \Delta x^t)}{(\hat{\delta}_t)^3} + C_3^t \left(\mu_t \left(C_4^t \right)^2 + \left(C_4^t \right)^3 \right) \geq (1-\eta) \frac{(\mu_t)^2 \pi_4(\mu_t)}{\hat{\delta}_t}.$$

Since $\{C_3^t\}$ is bounded and, by A2, $\{R_f(x^t, \Delta x^t)/(\hat{\delta}_t)^3\}$ is bounded and, by A1(b), $\{C_4^t\} \to 1$, we have that $\{(\mu_t)^2 \pi_4(\mu_t)/\hat{\delta}_t\}_{t\in T'}$ is bounded. This in turn, together with A1(d), implies that

$$\frac{\pi_1(\mu_t)}{(\hat{\delta}_t)^{2-\beta_2}} = \frac{\pi_1(\mu_t)}{((\mu_t)^2 \pi_4(\mu_t))^{2-\beta_2}} \cdot \left(\frac{(\mu_t)^2 \pi_4(\mu_t)}{\hat{\delta}_t} \right)^{2-\beta_2} \to 0 \quad \text{as } t \in T' \to \infty.$$

Thus, in all cases and subcases, we obtain that the right-hand side of (6.11) tends to zero as $t \in T' \to \infty$. Also, the first and third inequalities of (5.1) and assumption A1(a) imply

$$r^t \to 0, \quad c^t \to 0, \quad \text{and} \quad \|(S^t)^{-1} r^t\| = \frac{\|\tilde{\Lambda}^t r^t\|}{\mu_t} \leq \frac{\|\lambda_{\mathrm{P}}^t\|_\infty \|r^t\|}{\mu_t} \to 0.$$

These together with $\pi_4(\mu_t) \to 0$, $\{(x^t, \lambda^t, \lambda_{\mathrm{P}}^t)\}_{t\in T'} \to (\bar{x}, \bar{\lambda}, \bar{\lambda})$, and $\beta^t = \beta_2 > 1$ in cases (ii) and (iii) yield from (6.10) that $\liminf_{t\to\infty, t\in T'} (d^t)^T Q^t d^t \geq 0$. Using (6.9), this in turn yields (1.7). ∎

Combining Propositions 4.1 and 6.1 and Lemma 4.2, we obtain the following global convergence result for Algorithm 2.

COROLLARY 6.2. *Suppose that $\mathcal{X}_\zeta$ is bounded and (4.24) holds for all $x \in \mathcal{X}_\zeta$, where $\zeta := \max\{\|g(x^0)+s^0\|_p, \max_{\mu\leq\mu_1} \pi_1(\mu)\}$ and $(\mu_1, x^0, s^0, \lambda^0)$ denotes the starting iterate for Algorithm 2. Then $\{(\mu_t, x^t, s^t, \lambda^t, \delta_t)\}_{t=1,2,\ldots}$, generated by Algorithm 2, with σ_2 and τ chosen sufficiently large within Algorithm 1, is defined. Moreover, $x^t \in \mathcal{X}_\zeta$ for all t, $\{\lambda^t\}$ is bounded, and the following results hold for every cluster point $(\bar{x}, \bar{\lambda})$ of $\{(x^t, \lambda^t)\}$:*

(a) *$(\bar{x}, \bar{\lambda})$ satisfies (1.2).*

(b) *If $\pi_1, \ldots, \pi_4$ satisfy A1 and f satisfies A2 in Proposition 6.1(c), then (1.7) holds for any subsequence T' of $\{1, 2, \ldots\}$ such that $\{(x^t, \lambda^t)\}_{t\in T'}$ converges to $(\bar{x}, \bar{\lambda})$ and for any subsequence of unit directions $\{d^t\}_{t\in T'}$ satisfying $[\nabla g_i(x^t)^T d^t]_{i\in I(\bar{x})} = 0$ for all $t \in T'$.*

Proof. The assumptions on $\mathcal{X}_\zeta$ imply that (4.23) holds for some $\xi > 0$. Suppose at the beginning of an iteration t of Algorithm 2 that we have $\|g(x^{t-1}) + s^{t-1}\|_p \le \zeta$ and $s^{t-1} \ge -g(x^{t-1})$. Then when we apply Algorithm 1 at step 1 with $\sigma_2 \ge \zeta/\xi$, Lemma 4.2 and the fact that Algorithm 1 maintains $\|g(x) + s\|_p$ to be monotonically decreasing whenever $\|g(x) + s\|_p > \pi_1(\mu_t)$ imply that each inner iterate $(x^k, \lambda_{\rm P}^k)$ generated by Algorithm 1 satisfies $x^k \in \mathcal{X}_\zeta$ and $\|\lambda_{\rm P}^k\|_q$ is bounded above (independent of τ) whenever (2.15) holds. Then, by choosing τ in Algorithm 1 sufficiently large, (4.1) holds for all such k. An induction argument then shows that the inner iterates generated by Algorithm 1 are defined and satisfy (4.1) at all inner iterations k, and thus, by Proposition 4.1(a), Algorithm 1 terminates finitely with an (x^t, s^t, λ^t) and δ_t, υ^t satisfying $\|g(x^t) + s^t\|_p \le \zeta$ (so $x^t \in \mathcal{X}_\zeta$), as well as (5.1) and $s^t \ge -g(x^t)$.

Since x^t lies in the closed and bounded set $\mathcal{X}_\zeta$, $\bar{x}$ also lies in $\mathcal{X}_\zeta$. Since (5.1) is satisfied for all t, Proposition 6.1(a) yields that $\bar{x}$ is a feasible solution of (1.1). Since $\bar{x}$ satisfies (4.24) with $\zeta = 0$, Proposition 6.1(b) yields that $\{\lambda_{\rm P}^t\}$ is bounded and every cluster point $\bar{\lambda}$, together with $\bar{x}$, satisfies (1.2). Proposition 4.1(d) implies that assumption A3 holds for all t. Since $\pi_1, \ldots, \pi_4$ satisfy A1 and f satisfies A2, Proposition 6.1(c) yields that $\{\lambda^t\}$ is bounded and, in fact, $\lambda^t - \lambda_{\rm P}^t \to 0$. This proves (a). Then (b) follows from Proposition 6.1(c). □

Notice that A2 is satisfied if f is thrice differentiable at $\bar{x}$. If g is affine so that $\nabla g(x^t) = \nabla g(\bar{x})$ for all t, Corollary 6.2(b) yields that $(\bar{x}, \bar{\lambda})$ satisfies the second-order stationarity condition (1.3). If $\nabla g_i(\bar{x})$, $i \in I(\bar{x})$, are linearly independent, then for any $\bar{d}$ in the null space of $[\nabla g_i(\bar{x})^T]_{i \in I(\bar{x})}$ there exists a d^t in the null space of $[\nabla g_i(x^t)^T]_{i \in I(\bar{x})}$ that converges to $\bar{d}$ as $x^t \to \bar{x}$. (For example, we can take d^t to be the orthogonal projection of $\bar{d}$ onto the latter null space.) Using this d^t in Corollary 6.2(b) yields that $(\bar{x}, \bar{\lambda})$ satisfies the second-order stationarity condition (1.3). Thus, in the above two cases, we obtain that $\bar{x}$ is a second-order stationary point of (1.1).

To summarize, if (i) f is thrice-differentiable, (ii) either g is affine or $\nabla g_i(x)$, $i \in I(x)$, are linearly independent for all $x \in \mathcal{X}_0$, and (iii) $\mathcal{X}_\zeta$ is bounded and (4.24) holds for all $x \in \mathcal{X}_\zeta$, with ζ defined as in Corollary 6.2 (so that ζ depends on the initial iterate for Algorithm 2), then $\{x^t\}$ generated by Algorithm 2, with $\pi_1, \pi_2, \pi_3, \pi_4$ chosen to satisfy assumption A1 (and with σ_2, τ chosen sufficiently large within Algorithm 1), is defined and bounded, and all cluster points are second-order stationary points of (1.1). If g is affine and the initial iterate is feasible, i.e., $g(x^0) + s^0 = 0$, then the method maintains feasibility at all iterations, and the condition (iii) can be refined to the feasible set $\mathcal{X}_0$ being bounded and the MFCQ holding at all $x \in \mathcal{X}_0$. This contrasts with the sufficient conditions for the methods of [11, 21, 30]. If g is given by Example 1 and f is any thrice-differentiable function defined on $\Re^2$, then

$$\max\left\{\|g(x^0) + s^0\|_\infty, \max_{\mu \le \mu_1} \pi_1(\mu)\right\} < \frac{1}{2}$$

is sufficient for $\{x^t\}$ generated by Algorithm 2 (with the same provision as above) to be defined and bounded, and for all cluster points to be second-order stationary points of (1.1). In practice, we can initialize σ_2, τ to any positive value and then increase both by a constant factor whenever at some iteration k of Algorithm 1 we find that $\tau \le \|\lambda_{\rm P}^k\|_q/((1-\eta)\sigma_1)$. It can be seen that this does not affect the convergence results of Corollary 6.2.

Lastly, following [12] and analogous to the discussions in Note 3, we can allow $\nabla^2 f(x^t)$ and $\nabla^2 g_i(x^t)$ to be replaced by $n \times n$ symmetric real matrices H^t and H_i^t, $i = 1, \ldots, m$, in Algorithm 2. Our convergence result and, in particular, Proposition

6.1 and Corollary 6.2, still hold, provided that

$$\|\nabla^2 f(x^t) - H^t\| \to 0 \quad \text{and} \quad \|\nabla^2 g_i(x^t) - H_i^t\| \to 0,\ i = 1, \ldots, m, \qquad \text{as } t \to \infty.$$

7. Preliminary numerical experience. We implemented a version of Algorithm 2 in Matlab and ran it on five Hock and Schittkowski (HS) problems [27] as well as one of the Wächter–Biegler (WB) examples [40]. Our tests use only five HS problems due to the time required to set up the gradients and Hessians of f and $g_1, \ldots, g_m$ for each problem. As such, our reported numerical experience should be considered as very preliminary. However, the intention of our tests is mainly to validate the convergence theory and to gain some understanding of the numerical behavior of the method. Notice that the last two problems, HS108 and HS116, are nontrivial, and thus they provide a reasonably good test of our implementation.

Parameter choices. In our Matlab implementation of Algorithm 2, we choose

$$\mu_1 = 10, \quad p = 2, \quad \omega_1 = 0.3, \quad \omega_2 = \frac{1}{4}, \quad \delta_{\text{thr}} = 1, \quad \delta_0 = \delta_{\max} = 100,$$

as well as $\pi_4(\mu) = \max\{\mu^{0.2}, 10\mu\}$, $\pi_2(\mu) = \mu^2\pi_4(\mu)$,[5] $\pi_1(\mu) = \pi_3(\mu) = \pi_4(\mu) \cdot \min\{\mu, (\mu\pi_2(\mu))^{(2-\beta_2)/2}, 1000(\mu^2\pi_4(\mu))^{2-\beta_2}\}$. For a given starting x^0, we set

$$s^0 = \max\{-g(x^0), s^0_{\min} e\} \tag{7.1}$$

with $s^0_{\min} > 0$ a user-chosen parameter. We update $\mu_{t+1} = \omega_2\mu_t$ for all t. In Algorithm 1, we further set

$$\beta_1 = 0.9, \quad \beta_2 = \begin{cases} 1.1 & \text{in phase 1,} \\ 1.9 & \text{in phase 2 or 3,} \end{cases} \qquad \sigma_1 = \eta = 0.1, \quad \bar{\tau} = 0.1,$$

and we initialize both σ_2 and τ to 10 and increase them by a factor of $1/\omega_1$ whenever $\tau \le \|\lambda^k_{\text{P}}\|/((1-\eta)\sigma_1)$.[6] We set $\beta^k = 0$ whenever $\delta_k > \delta_{\text{th}}$. If a successful step is taken at iteration k, we set δ_{k+1} to be either $\min\{\delta_k/\omega_1, \delta_{\max}\}$ or δ_k, depending on whether $\rho_k \ge 0.95$. We remark that the above parameter choices were made without too much fine-tuning and can conceivably be improved.

Solving the trust-region subproblem. In one implementation, we solve (2.1) inexactly by using binary search to find a $\gamma \ge 0$ such that $\gamma I + M + A\Lambda S^{-1}A^T$ is positive semidefinite and the solution of (2.8)–(2.10) satisfies $|\|\Delta x\| - \delta|/\delta < 10^{-9}$ whenever $\gamma > 10^{-15}$. In a second implementation, we solve (2.1) inexactly using a version of the Moré–Sorensen method [31]. The performance of the two implementations are generally comparable in terms of number of iterations and solution accuracy. An exception occurs on HS108 and HS116 with $s^0_{\min} = 1$, where the second implementation requires roughly twice the number of iterations. We do not yet have an explanation for this. The numerical results reported below are for the first implementation. In the future, we hope to try more efficient solution strategies, such as those described in [14], in conjunction with the inexact solution idea of subsection 2.2.

[5] Because this choice of $\pi_2(\mu)$ becomes small rapidly with μ, we replace it by $\pi_2(\mu) = \mu\pi_4(\mu)$, provided $\|(S^k)^{-1}\Delta s^{k-1}\|/\delta_k$ is below a fixed threshold, where k indexes the final iteration of Algorithm 1.

[6] We set β_2 near 1 in phase 1 in order to prevent $1 - \sigma_1(\delta_k)^{\beta^k}$ in (2.16) from being treated as 1 by Matlab when δ_k is small. Using different β_2 in different phases does not affect the convergence properties of Algorithm 1, as given in Proposition 4.1.

Initialization. For the HS problems, the starting x^0 is as given in [27]. The WB example is

$$n = 1, \quad m = 2, \quad f(x) = x, \quad g_1(x) = -x^2, \quad g_2(x) = -x + 1, \quad x^0 = -1,$$

which corresponds to setting $a = 0$, $b = 1$ in [40, eq. (8)]. It can be verified that x^0, together with s^0 given by (7.1), satisfies the assumptions of Theorem 1 in [40], so that an interior-point method of the type described therein would not converge to a feasible solution when started at (x^0, s^0).

Acceleration step. Our initial implementation found the correct solution on all problems except HS116. On HS116, $\{x^k\}$ remains bounded while $\|\lambda_{\rm P}^k\| \to \infty$ during phase 1 of Algorithm 1. This suggests that (4.24) fails for the ζ given in Corollary 6.2, which is plausible since ζ is quite large on this problem. To improve convergence, we added line search based acceleration steps to Algorithm 1, which we describe below. The addition of these steps does not affect the convergence properties of Algorithm 1, as given in Proposition 4.1.

The acceleration step is activated at iteration k of Algorithm 1 whenever a null step is about to be taken. Then, instead of taking the null step, we compute a second trial point $(x_{\rm TR2}^k, s_{\rm TR2}^k, \lambda_{\rm TR2}^k)$ and test it for acceptance. This second trial point is computed as follows: Let $(\overline{\Delta x}^k, \overline{\Delta s}^k, \bar\lambda^k)$ be a solution of (2.8)–(2.10) with $(x, s, \lambda) = (x^k, s^k, \lambda^k)$, $\beta = 0$, and

$$\gamma = \begin{cases} 0 & \text{in phase 1,} \\ \gamma_k & \text{in phase 2 or 3,} \end{cases}$$

where γ_k denotes the trust-region Lagrange multiplier that is obtained in computing $(\Delta x^k, \Delta s^k, \lambda^k)$.[7] Let

$$\overline{\Delta\lambda}^k := \bar\lambda^k - \lambda^k, \qquad t_k := 0.95 \min\left\{ \min_{i:\overline{\Delta s}_i^k < 0} \left(\frac{-s_i^k}{\overline{\Delta s}_i^k} \right) \min_{i:\overline{\Delta\lambda}_i^k < 0} \left(\frac{-\lambda_i^k}{\overline{\Delta\lambda}_i^k} \right) \right\}.$$

Then, starting at $t = t_k$, we test

$$x_{\rm TR2}^k := x^k + t\Delta x^k, \qquad s_{\rm TR2}^k := \max\{s^k + t\overline{\Delta s}^k, -g(\overline{x}^k)\}, \qquad \lambda_{\rm TR2}^k := \lambda^k + t\overline{\Delta\lambda}^k$$

for acceptance and, if not accepted, we replace t by $t/2$ until t is below $\max\{0.01, 0.1/\|\overline{\Delta x}^k\|\}$. If none of the trial points is accepted, a null step is taken. The addition of this step accelerates convergence significantly on HS116 and, to a lesser extent, on other problems.

Numerical results. The performance of the final Matlab implementation (with the acceleration steps) on the five HS problems, numbered as in [27] and the WB example, is summarized in Table 7.1. (Here b denotes the number of bound constraints, which are treated as inequalities.) As can be seen from the table, our method seems robust in the sense that reasonable solution accuracy is achieved in a reasonable number of iterations. Furthermore, the generated solution is in each case a second-order stationary point, though possibly not a global optimal solution (which is unavoidable for a local method). The final values of σ_2 and τ vary. On HS116, they are quite

[7] Thus, if we are in phase 1 and $\gamma_k = 0$, or if we are in phase 2 or 3, then both $(\Delta x^k, \Delta s^k, \lambda_{\rm TR}^k)$ and $(\overline{\Delta x}^k, \overline{\Delta s}^k, \bar\lambda^k)$ are solutions of linear equations with the same left-hand matrix, and thus the same matrix factorization can be used to compute both.

TABLE 7.1

Performance of Algorithm 2 on five HS problems and a WB example, as indicated by the number of gradient and Hessian evaluations of $f, g_1, \ldots, g_m$ (neval), the number of trust-region subproblems solved (nsolv), the number of additional matrix factorizations in the acceleration steps (nacc), the infeasibility $\|\max\{0, g(x)\}\|$ (infeas), and the objective value $f(x)$ upon termination (obj). The number in parentheses is the final μ value. (†Convergence to a second-order stationary point $(-0.5, 0.7071)$. ‡Convergence to a second-order stationary point $(0.5, 0.4, \ldots, 1)$.)

Problem	n, m, b	$s^0_{\min}$	**neval/nsolv/nacc/infeas/obj**
HS1	2,1,1	0.01	$29/29/0/0/3{\cdot}10^{-9}$ (10^{-4})
		1	$29/29/0/0/3{\cdot}10^{-9}$ (10^{-4})
HS16	2,5,3	0.01	$28/41/0/2{\cdot}10^{-7}/23.149^\dagger$ (10^{-3})
		1	$47/71/0/2{\cdot}10^{-7}/23.149^\dagger$ (10^{-3})
HS43	4,3,0	0.01	$14/17/0/7{\cdot}10^{-6}/-43.980$ $(5{\cdot}10^{-3})$
		1	$14/17/0/7{\cdot}10^{-6}/-43.980$ $(5{\cdot}10^{-3})$
HS108	9,14,1	0.01	$51/69/12/2{\cdot}10^{-6}/-0.6477^\ddagger$ $(5{\cdot}10^{-4})$
		1	$49/69/12/2{\cdot}10^{-6}/-0.6477^\ddagger$ $(5{\cdot}10^{-4})$
HS116	13,41,26	0.01	$398/443/47/2{\cdot}10^{-4}/97.620$ (10^{-3})
		1	$124/168/31/2{\cdot}10^{-4}/97.620$ (10^{-3})
WB	1,1,1	0.01	$32/47/12/4{\cdot}10^{-10}/1.000$ $(5{\cdot}10^{-4})$
		1	$24/37/2/4{\cdot}10^{-10}/1.000$ $(5{\cdot}10^{-4})$

high, near 45000. Most of the iterations are spent in phase 3, followed by phase 1 and then phase 2. For example, on HS16 with $s^0_{\min} = 1$, the number of iterations spent in phases 1, 2, and 3 are, respectively, 12, 5, and 54.

The performance of our method on the HS problems, while respectable, is not nearly as good as those reported for some other interior-point methods [23, 45]. Nonetheless, we feel encouraged by the robustness of the results and their consistency with the theory. For a comparison, the method in [23] reportedly has difficulty solving HS116 and, according to [40], the methods in [10, 17, 23, 39, 42, 45] fail to solve the WB example. A nice feature of our method is that it readily allows for incorporation of acceleration steps, while maintaining its theoretical convergence properties. This suggests that it can serve as a safeguard in combination with other primal-dual methods in order to enhance the robustness of the latter. Also, similar to [23, sect. 7], instead of introducing slack variables for all inequality constraints, we can do this only for those inequality constraints such that $g_i(x^0) \geq 0$. For the remaining inequality constraints, we can use the barrier penalty $-\ln(-g_i(x))$ directly. Such a hybrid method may be more efficient. Lastly, Dominique Orban [35] has suggested that preconditioners can be taken into account by considering symmetric positive definite matrices which define norms that are uniformly equivalent to the ℓ_2-norm.

8. Incorporating equality constraints. Algorithm 2 may be extended to solve a nonlinear program with both equality and inequality constraints:

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & g(x) \leq 0, \quad h(x) = 0, \end{array} \tag{8.1}$$

where $f : \Re^n \mapsto \Re$, $g : \Re^n \mapsto \Re^m$, and $h : \Re^n \mapsto \Re^{m'}$ ($m \geq 0$, $m' \geq 0$) are twice continuously differentiable functions. We sketch the extensions below.

If h is affine and the starting x^0 satisfies $h(x^0) = 0$, then Algorithm 2 may be extended to solve (8.1) by adding the constraints $\nabla h(x)^T \Delta x = 0$ to (2.1), thus maintaining that the iterates lie in the linear manifold $\{x \in \Re^n : h(x) = 0\}$. Convergence

results analogous to Corollary 6.2 may be obtained by replacing (4.24) with

$$0 \notin \{\nabla g(x)\lambda + \nabla h(x)\psi : \|\lambda\|_1 + \|\psi\|_1 = 1,\ \lambda \geq 0,\ \lambda_i = 0\ \ \forall i \notin I_\zeta(x)\} \tag{8.2}$$

and further restricting d^t in (1.7) to satisfy $\nabla h(x^t)^T d^t = 0$, etc.

In general, Algorithm 2 may be extended to solve (8.1) using a penalty approach. In particular, when we apply Algorithm 1 at iteration t of Algorithm 2, we would replace f in the objective by, for example, the ℓ_2-penalty function (see, e.g., [19, p. 63]):

$$f^t(x) := f(x) + \frac{1}{2\mu_t}\|h(x)\|^2.$$

Then, with (5.1) accordingly modified, it can be seen that part (a) of Proposition 6.1 still holds, while part (b) holds provided that (4.24) is replaced by (8.2) and that (1.2) is replaced by the corresponding first-order necessary optimality condition for (8.1). By using the formula for $\nabla^2 f^t(x)$ and defining

$$l(x, \lambda, \psi) := f(x) + g(x)^T\lambda + h(x)^T\psi,$$

part (c) can be analogously extended by further assuming that h is thrice differentiable and further restricting d^t to satisfy $\nabla h(x^t)^T d^t = 0$ for all $t \in T'$. Corollary 6.2 may be extended accordingly. Also, for the primal method (see Note 1), it appears that these results can be extended to the barrier function $1/(\cdot)^\kappa$ ($\kappa > 0$) in place of $-\ln(\cdot)$.

9. Extension of the primal method to a semidefinite nonlinear program. The primal version of our method, as described in Note 1, and its convergence properties may be extended to the following semidefinite nonlinear program:

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & -g(x) \succeq 0, \end{array}$$

where $f : \Re^n \mapsto \Re$ and $g = [g_{ij}]_{i,j=1}^m : \Re^n \mapsto \mathcal{S}$ are twice continuously differentiable functions and, for $A \in \mathcal{S}$, "$A \succeq 0$" (respectively, "$A \succ 0$") means that A is positive semidefinite (respectively, positive definite) [2, 20, 28, 34, 38]. Here $\mathcal{W}$ denotes the space of $m \times m$ block-diagonal real matrices with k blocks of sizes $m_1, \ldots, m_k$, respectively (the blocks are fixed), and $\mathcal{S}$ denotes the subspace comprising those $A \in \mathcal{W}$ that are symmetric, i.e., $A^T = A$. We sketch the extension below. We speculate that the primal-dual method, i.e., Algorithm 1, can be extended in an analogous manner. However, the analysis is more involved since there are many choices for the primal-dual search direction in the case of semidefinite programs.

We endow $\mathcal{W}$ with the inner product and norm

$$\langle A, B\rangle := \mathrm{tr}[A^T B], \quad \|A\| := \sqrt{\langle A, A\rangle} \qquad \forall A, B \in \mathcal{W},$$

where $\mathrm{tr}[\cdot]$ denotes the matrix trace. For $A \in \mathcal{S}$, we also define $\|A\|_p := \|[\alpha_i]_{i=1}^m\|_p$ ($1 \leq p \leq \infty$), where $\alpha_1, \ldots, \alpha_m$ are the eigenvalues of A. Thus, $\|A\| = \|A\|_2 = (\sum_{i=1}^m \|A_i\|^2)^{1/2}$, where A_i denotes the ith column of A. Also, define $\mathcal{O} := \{A \in \mathcal{W} : A^T A = I\}$ and $\mathcal{S}_{++} := \{A \in \mathcal{S} : A \succ 0\}$.

Accordingly, we define the Lagrangian $l(x, \Lambda) := f(x) + \langle g(x), \Lambda\rangle$, with $\Lambda \in \mathcal{S}$. The first-order stationarity condition is

$$-g(x) \succeq 0, \quad \Lambda \succeq 0, \quad \langle g(x), \Lambda\rangle = 0, \quad \nabla_x l(x, \Lambda) = 0.$$

We work with the barrier problem:

$$\begin{array}{ll}\text{minimize} & f^\mu(x,s) := f(x) - \mu \ln(\det[S]) \\ \text{subject to} & g(x) + S = 0, \quad S \succ 0.\end{array}$$

The trust-region subproblem is accordingly

$$\begin{array}{ll}\text{minimize} & \nabla f(x)^T \Delta x - \mu\langle S^{-1}, \Delta S\rangle + \frac{1}{2}\Delta x^T \nabla_{xx} l(x,\Lambda)\Delta x + \frac{1}{2}\|\Lambda^{1/2}\Delta S S^{-1/2}\|^2 \\ \text{subject to} & \|\Delta x\| \le \delta,\ \nabla g(x)^T \Delta x + \Delta S = -\delta^\beta (g(x) + S),\end{array}$$

where $\Delta x \mapsto \nabla g(x)^T \Delta x$ is a linear mapping from $\Re^n$ to $\mathcal{S}$. From an approximate solution $(\Delta x, \Delta S)$, we generate a trial point,

$$x_{\text{TR}} := x + \Delta x, \quad S_{\text{TR}} := S + \widehat{\Delta S} + [-g(x_{\text{TR}}) - (S + \widehat{\Delta S})]_+, \quad \Lambda_{\text{TR}} := \mu (S_{\text{TR}})^{-1},$$

with $\widehat{\Delta S} := \Delta S - \Delta x^T \nabla^2 g(x) \Delta x/2$, and we test whether to accept $(x_{\text{TR}}, S_{\text{TR}}, \Lambda_{\text{TR}})$ as the new iterate or to decrease δ and repeat, etc. Here, $\Delta x^T \nabla^2 g(x) \Delta x := [\Delta x^T \nabla^2 g_{ij}(x) \Delta x]_{i,j=1}^m$, and $[A]_+$ denotes the matrix obtained by replacing, in the spectral decomposition of A, the negative eigenvalues of A by zero; see [36, sect. 2] for properties of $[\cdot]_+$. Analogously, we use the ℓ_2 augmented barrier function:

$$f^{\mu,\tau}(x,S) := f^\mu(x,S) + \tau \|g(x) + S\|.$$

Then, Algorithms 1 and 2 extend in a straightforward manner. Moreover, Lemma 3.1(a) extends by noting that $g(x_{\text{TR}}) + S_{\text{TR}} = [g(x_{\text{TR}}) + S + \widehat{\Delta S}]_+$, and hence $\|g(x_{\text{TR}}) + S_{\text{TR}}\| \le \|g(x_{\text{TR}}) + S + \widehat{\Delta S}\|$. Lemma 3.1(c) extends by using the Taylor series expansion (e.g., [38, eq. (45)]), giving that

$$\begin{aligned} &-\ln(\det[S + \widehat{\Delta S}]) \\ &\quad \le -\ln(\det[S]) - \langle S^{-1}, \widehat{\Delta S}\rangle + \frac{1}{2}\|S^{-1/2}\widehat{\Delta S} S^{-1/2}\|^2 + \sum_{j=3}^{\infty} \frac{(-1)^j}{j} \text{tr}[(S^{-1}\widehat{\Delta S})^j] \end{aligned}$$

and assuming $\|S^{-1/2}\widehat{\Delta S} S^{-1/2}\| \le 2/3$. Lemma 3.1(d) extends accordingly, as does Proposition 4.1, with $\epsilon_2 = 0$ and omitting (4.4), (4.5).

To extend assumption (4.24) and Lemma 4.2, we define, for each $\zeta \ge 0$, $x \in \Re^n$, and $P \in \mathcal{O}$, $I_\zeta(x,P) := \{i \in \{1,\ldots,m\} : \|[P^T g(x) P]_i\| \le \zeta\}$. Then, we replace (4.24) by

$$0 \notin \left\{\nabla_x \langle g(x), \Lambda\rangle : \Lambda \succeq 0, \|\Lambda\|_1 = 1, [P^T \Lambda P]_i = 0\ \forall i \notin I_\zeta(x,P)\right\}\ \forall P \in \mathcal{O},$$

and Lemma 4.2 extends analogously. In particular, for any $\zeta \ge 0, \xi > 0, \mu > 0$, suppose that there exist $\sigma_2 \ge \zeta/\xi$ and a sequence $(x^k, S^k, \mu^k) \in \Re^n \times \mathcal{S}_{++} \times \Re_{++}$, $k = 0, 1, \ldots$, such that $\mu^k \le \mu$ and

$$\|g(x^k) + S^k\| \le \zeta, \quad \sigma_2 \|\nabla_x l(x^k, \Lambda^k)\| < \langle \Lambda^k, g(x^k) + S^k\rangle \quad \forall k,$$

with $\Lambda^k := \mu^k (S^k)^{-1}$, but $\|\Lambda^k\| \to \infty$. Then we would obtain as in the proof of Lemma 4.2 that any cluster point $(\bar{x}, \bar{S}, \bar{\Lambda})$ of $(x^k, S^k, \Lambda^k/\|\Lambda^k\|_1)$ satisfies $\|\bar{\Lambda}\|_1 = 1$, $\bar{S} \succeq 0, \bar{\Lambda} \succeq 0$, and $\bar{S}\bar{\Lambda} = 0$. Then, for any $P \in \mathcal{O}$ such that $P^T \bar{S} P$ is diagonal, and letting $I := \{i \in \{1,\ldots,m\} : [P^T \bar{S} P]_i = 0\}$, it can be seen that $[P^T \bar{\Lambda} P]_i = 0$ for $i \notin I$. Also, for each $i \in I$, we have

$$\zeta \ge \|g(\bar{x}) + \bar{S}\| = \|P^T (g(\bar{x}) + \bar{S}) P\| \ge \|[P^T (g(\bar{x}) + \bar{S}) P]_i\| = \|[P^T g(\bar{x}) P]_i\|.$$

Thus $I \subseteq I_\zeta(x,P)$. Also, we have from $\|\nabla_x l(x^k, \Lambda^k)\| < \langle \Lambda^k, g(x^k) + S^k\rangle/\sigma_2 \le \|\Lambda^k\|\zeta/\sigma_2$ that, in the limit, $\|\nabla_x \langle g(x), \bar{\Lambda}\rangle\|_{x=\bar{x}} \le \zeta/\sigma_2 \le \xi$.

Parts of (a) and (b) of Proposition 6.1 extend readily, thus yielding analogous results on convergence to first-order stationary points. Part (c) can be similarly extended to show that, under a modification to the stated assumptions, (1.7) holds for any subsequence $\{(x^t, \Lambda^t)\}_{t\in T'} \to (\bar{x}, \bar{\lambda})$, any subsequence $P^t \in \mathcal{O}$ with $(P^t)^T S^t P^t$ diagonal for all $t \in T'$ and $\{P^t\}_{t\in T'} \to$ some $\bar{P}$, and any subsequence $d^t \in \Re^n$ with $\|d^t\| = 1$ and $[(P^t)^T(\nabla g(x^t)^T d^t)P^t]_{i\in I_0(\bar{x},\bar{P})} = 0$ for all $t \in T'$. The modification entails omitting A1(b), A1(c), (6.3), and (6.4) and setting $\pi_2 \equiv 0$. To our knowledge, such an asymptotic second-order stationarity result has not been obtained previously.

Acknowledgments. The author thanks the referees and, in particular, Dominique Orban and the Associate Editor, Nick Gould, for their helpful comments and suggestions which led to significant improvements in the paper.

REFERENCES

[1] I. Akrotirianakis and B. Rustem, *A Globally Convergent Interior Point Algorithm for General Non-linear Programming Problems*, Technical report 97/14, Department of Computing, Imperial College of Science, Technology and Medicine, London, United Kingdom, 1997.

[2] F. Alizadeh, *Interior point methods in semidefinite programming with applications to combinatorial optimization*, SIAM J. Optim., 5 (1995), pp. 13–51.

[3] A. Auslender, *Penalty methods for computing points that satisfy second order necessary points*, Math. Programming, 17 (1979), pp. 229–238.

[4] S. J. Benson, Y. Ye, and X. Zhang, *Solving large-scale sparse semidefinite programs for combinatorial optimization*, SIAM J. Optim., 10 (2000), pp. 443–461.

[5] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed., Athena Scientific, Belmont, MA, 1999.

[6] R. H. Byrd, J. C. Gilbert, and J. Nocedal, *A trust region method based on interior point techniques for nonlinear programming*, Math. Program., 89 (2000), pp. 149–185.

[7] R. H. Byrd, M. E. Hribar, and J. Nocedal, *An interior point algorithm for large-scale nonlinear programming*, SIAM J. Optim., 9 (1999), pp. 877–900.

[8] R. H. Byrd, G. Liu, and J. Nocedal, *On the local behavior of an interior point method for nonlinear programming*, in Numerical Analysis 1997, D. F. Griffiths and D. J. Higham, eds., Addison Wesley Longman, Reading, MA, 1998, pp. 37–56.

[9] R. H. Byrd, J. Nocedal, and R. A. Waltz, *Feasible interior methods using slacks for nonlinear optimization*, Technical report, Optimization Technology Center, Northwestern University, Evanston, IL, 2000.

[10] A. M. Cervantes, A. Wächter, R. Tütüncü, and L. T. Biegler, *A reduced space interior point strategy for optimization of differential algebraic systems*, Comput. Chem. Eng., 24 (2000), pp. 39–51.

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

[12] A. R. Conn, N. I. M. Gould, D. Orban, and P. L. Toint, *A primal-dual trust-region algorithm for non-convex nonlinear programming*, Math. Program., 87 (2000), pp. 215–249.

[13] A. R. Conn, N. I. M. Gould, and P. L. Toint, *A primal-dual algorithm for minimizing a nonconvex function subject to bound and linear equality constraints*, in Nonlinear Optimization and Related Topics, G. Di Pillo and F. Giannessi, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000, pp. 15–49.

[14] A. R. Conn, N. I. M. Gould, and Ph. L. Toint, *Trust-Region Methods*, MPS/SIAM Ser. Optim. MPOI, SIAM, Philadelphia, 2000.

[15] J. E. Dennis, M. Heinkenschloss, and L. N. Vicente, *Trust-region interior-point SQP algorithms for a class of nonlinear programming problems*, SIAM J. Control Optim., 36 (1998), pp. 1750–1794.

[16] G. Di Pillo and L. Grippo, *A continuously differentiable exact penalty function for nonlinear programming problems with inequality constraints*, SIAM J. Control Optim., 23 (1985), pp. 72–84.

[17] A. S. El-Bakry, R. A. Tapia, T. Tsuchiya, and Y. Zhang, *On the formulation and theory of the Newton interior-point method for nonlinear programming*, J. Optim. Theory Appl., 89 (1996), pp. 507–541.

[18] F. Facchinei and S. Lucidi, *Convergence to second order stationary points in inequality constrained optimization*, Math. Oper. Res., 23 (1998), pp. 746–766.

[19] A. V. Fiacco and G. P. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley and Sons, New York, 1968; republished as Classics Appl. Math. 4, SIAM, Philadelphia, 1990.
[20] R. Fletcher, *Semi-definite matrix constraints in optimization*, SIAM J. Control Optim., 23 (1985), pp. 493–513.
[21] A. Forsgren and W. Murray, *Newton methods for large-scale linear inequality-constrained minimization*, SIAM J. Optim., 7 (1997), pp. 162–176.
[22] A. Forsgren and P. E. Gill, *Primal-dual interior methods for nonconvex nonlinear programming*, SIAM J. Optim., 8 (1998), pp. 1132–1152.
[23] D. M. Gay, M. L. Overton, and M. H. Wright, *A primal-dual interior method for nonconvex nonlinear programming*, in Advances in Nonlinear Programming, Appl. Optim. 14, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 31–56.
[24] T. Glad and E. Polak, *A multiplier method with automatic limitation of penalty growth*, Math Programming, 17 (1979), pp. 140–155.
[25] 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.
[26] N. I. M. Gould and Ph. L. Toint, *A note on the second-order convergence of optimization algorithms using barrier functions*, Math. Program., 85 (1999), pp. 433–438.
[27] W. Hock and K. Schittkowski, *Test Examples for Nonlinear Programming Codes*, Springer-Verlag, Berlin, Germany, 1981.
[28] F. Jarre, *A QQP-minimization method for semidefinite and smooth nonconvex programs*, Technical report, Institut für Angewandte Mathematik und Statistik, Universität Würzburg, Würzburg, Germany, 1998.
[29] L. S. Lasdon, J. Plummer, and G. Yu, *Primal-dual and primal interior point algorithms for general nonlinear programs*, ORSA J. Comput., 7 (1995), pp. 321–332.
[30] R. D. C. Monteiro and Y. Wang, *Trust region affine scaling algorithms for linearly constrained convex and concave programs*, Math. Programming, 80 (1998), pp. 283–313.
[31] J. J. Moré and D. C. Sorensen, *Computing a trust region step*, SIAM J. Sci. Statist. Comput., 4 (1983), pp. 553–572.
[32] H. Mukai and E. Polak, *A second-order method for the general nonlinear programming problem*, J. Optim. Theory Appl., 26 (1978), pp. 515–532.
[33] K. G. Murty and S. N. Kabadi, *Some NP-complete problems in quadratic and nonlinear programming*, Math. Programming, 39 (1987), pp. 117–129.
[34] Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM Stud. Appl. Math. 13, SIAM, Philadelphia, 1994.
[35] D. Orban, *private communication*, CERFACS, Toulouse, France, 2000.
[36] P. Tseng, *Merit functions for semi-definite complementarity problems*, Math. Programming, 83 (1998), pp. 159–185.
[37] T. Urban, A. L. Tits, and C. T. Lawrence, *A primal-dual interior-point method for nonconvex optimization with multiple logarithmic barrier parameters and with strong convergence properties*, Report, Electrical Engineering Department, University of Maryland, College Park, MD, 1998.
[38] L. Vandenberghe and S. Boyd, *Semidefinite programming*, SIAM Rev., 38 (1996), pp. 49–95.
[39] R. V. Vanderbei and D. F. Shanno, *An interior-point algorithm for nonconvex nonlinear programming*, Comput. Optim. Appl., 13 (1999), pp. 231–252.
[40] A. Wächter and L. T. Biegler, *Failure of global convergence for a class of interior point methods for nonlinear programming*, Math. Programming, 88 (2000), pp. 565–574.
[41] S. J. Wright, *Primal-Dual Interior-Point Methods*, SIAM, Philadelphia, 1997.
[42] H. Yamashita, *A globally convergent primal-dual interior point method for constrained optimization*, Technical report, Mathematical Systems Institute Inc., Shinjuku-ku, Tokyo, Japan, 1995.
[43] H. Yamashita and H. Yabe, *Superlinear and quadratic convergence of some primal-dual interior point methods for constrained optimization*, Math. Programming, 75 (1996), pp. 377–397.
[44] H. Yamashita and H. Yabe, *An interior point method with a primal-dual ℓ_2 barrier penalty function for nonlinear optimization*, Technical report, Mathematical Systems Institute Inc., Shinjuku-ku, Tokyo, Japan, 1999.
[45] H. Yamashita, H. Yabe, and T. Tanabe, *A globally and superlinearly convergent primal-dual interior point trust region method for large scale constrained optimization*, Technical report, Mathematical Systems Institute Inc., Shinjuku-ku, Tokyo, Japan, 1997 (revised 1998).

SIAM J. OPTIM.
Vol. 13, No. 2, pp. 470–497

MODIFYING SQP FOR DEGENERATE PROBLEMS*

STEPHEN J. WRIGHT†

Abstract. Most local convergence analyses of the sequential quadratic programming (SQP) algorithm for nonlinear programming make strong assumptions about the solution, namely, that the active constraint gradients are linearly independent and that there are no weakly active constraints. In this paper, we establish a framework for variants of SQP that retain the characteristic superlinear convergence rate even when these assumptions are relaxed, proving general convergence results and placing some recently proposed SQP variants in this framework. We discuss the reasons for which implementations of SQP often continue to exhibit good local convergence behavior even when the assumptions commonly made in the analysis are violated. Finally, we describe a new algorithm that formalizes and extends standard SQP implementation techniques, and we prove convergence results for this method also.

Key words. nonlinear programming, sequential quadratic programming, degeneracy

AMS subject classifications. 90C33, 90C30, 49M45

PII. S1052623498333731

1. Introduction.

We investigate local convergence properties of variants of the sequential quadratic programming (SQP) algorithm applied to the nonlinear programming problem

$$\text{NLP:} \qquad \min_z \phi(z) \qquad \text{subject to } g(z) \le 0, \tag{1.1}$$

where $\phi : \mathsf{R}^n \to \mathsf{R}$ and $g : \mathsf{R}^n \to \mathsf{R}^m$ are twice Lipschitz continuously differentiable functions. We are interested in *degenerate* problems: those for which the active constraint gradients at the solution are linearly dependent and/or the strict complementarity condition fails to hold.

We showed in [18] that even when strict complementarity (SC), second-order sufficient conditions, and a constraint qualification hold, nonuniqueness of the optimal multiplier can produce nonsuperlinear behavior of SQP. Motivated by this observation and by the fact that primal-dual interior-point algorithms for related problems converge superlinearly under the conditions just described [20, 16], we proposed a stabilized SQP (sSQP) method [18] and proved a local superlinear convergence result, later enhanced by Hager [11]. Independently, Fischer [8] proposed an algorithm into which a special procedure for choosing the Lagrange multiplier estimate is inserted between iterations of SQP. He proved superlinear convergence under slightly different assumptions from ours.

Our purposes in this paper are twofold. First, we introduce a common framework, which we call iSQP (for inexact SQP), that allows for a unified analysis of the stabilization procedures of the preceding paragraph. We prove general convergence results for methods in the iSQP framework, highlighting the effect on the convergence rate of changes between successive Lagrange multiplier estimates.

*Received by the editors February 4, 1998; accepted for publication (in revised form) February 25, 2002; published electronically October 1, 2002. This work was supported by the Mathematical, Information, and Computational Sciences Division subprogram of the Office of Computational and Technology Research, U.S. Department of Energy, under contract W-31-109-Eng-38, and by the National Science Foundation under grants CDA-9726385 and ACI-0082065.

http://www.siam.org/journals/siopt/13-2/33373.html

†Computer Sciences Department, 1210 W. Dayton Street, University of Wisconsin, Madison, WI 53706 (swright@cs.wisc.edu).

Our second goal requires a little more explanation. Implementations of SQP (for example, SNOPT [10]) often continue to exhibit good local convergence behavior even on degenerate problems, even though such problems fail to satisfy the standard assumptions made in local convergence analyses and even though theoretical examples of poor convergence behavior are easy to construct (see [18]). The iSQP framework proves to be useful in providing some theoretical support for this good practical performance. We find that the strategy of using the active (or working) set from the quadratic prgramming (QP) subproblem at the previous iteration as the initial active set for the current iteration is important in explaining the good behavior, as is the fact that the solver of the QP subproblem is allowed to return a slightly infeasible answer. Further, we propose and analyze an algorithm called SQPsws, in which the techniques used in existing implementations are formalized to produce an algorithm whose local convergence is superlinear under certain assumptions.

The main point of difference between the basic SQP algorithm as presented here and the versions that are implemented in standard software is that the implementations usually make use of quasi-Newton Hessian approximations, whereas we assume here that exact Hessians are used. Still, we believe that our observations below are relevant to the quasi-Newton case and, in particular, that quasi-Newton versions of the various algorithms discussed here would exhibit fast local convergence. Extension of the analysis to this case would, however, not be trivial, since it would have to take into account such factors as the effects of degeneracy on the quasi-Newton updates, so we leave this issue for possible future work. Another feature of the description presented here is that we focus on the local properties of the SQP approach and ignore the various algorithmic devices used to ensure global convergence.

The remainder of the paper is structured as follows. In section 2, we outline first-order optimality conditions and define various terms and assumptions that are used in the remainder of the paper. Section 3 defines the various second-order sufficient conditions that are required by the algorithms described in later sections. The iSQP framework is defined in section 4, where we also prove a useful result about the active set identified by the iSQP subproblem. Section 5 contains the main results about convergence of algorithms in the iSQP framework. Brief discussions of the stabilized SQP algorithm and Fischer's approach are given in sections 6 and 7, respectively, where we outline how both methods fit into the iSQP framework. Finally, the new algorithm SQPsws is described and some superlinear convergence results for it are proved in section 8.

2. Assumptions, notation, and basic results. We now review the optimality conditions for (1.1) and discuss various assumptions that are used in subsequent sections. These include second-order sufficient conditions of various types, along with complementarity conditions and the Mangasarian–Fromovitz constraint qualification (MFCQ). Finally, we quote a result that plays a key role in the analysis of the remainder of the paper—that MFCQ is equivalent to boundedness of the set of optimal Lagrange multipliers.

The Lagrangian for (1.1) is

$$\mathcal{L}(z, \lambda) = \phi(z) + \lambda^T g(z), \tag{2.1}$$

where $\lambda \in \mathsf{R}^m$ is the vector of Lagrange multipliers. We assume throughout that z^* is a strict local solution of (1.1). When a constraint qualification holds at z^* (see discussion below), first-order necessary conditions imply that there exists a vector

$\lambda^* \in \mathsf{R}^m$ such that

$$\mathcal{L}_z(z^*, \lambda^*) = 0, \quad g(z^*) \le 0, \quad \lambda^* \ge 0, \quad (\lambda^*)^T g(z^*) = 0. \tag{2.2}$$

These relations are the well-known Karush–Kuhn–Tucker (KKT) conditions. The following sets play an important role in the remainder of the paper:

$$\mathcal{S}_\lambda = \{\lambda^* \,|\, \lambda^* \text{ satisfies (2.2)}\}, \tag{2.3a}$$

$$\mathcal{S} = \{z^*\} \times \mathcal{S}_\lambda. \tag{2.3b}$$

We can write the conditions (2.2) alternatively as

$$\begin{bmatrix} \nabla\phi(z^*) + \nabla g(z^*)\lambda^* \\ g(z^*) \end{bmatrix} \in \begin{bmatrix} 0 \\ N(\lambda^*) \end{bmatrix}, \tag{2.4}$$

where $N(\lambda)$ is the set defined by

$$N(\lambda) \stackrel{\text{def}}{=} \begin{cases} \{y \,|\, y \le 0 \text{ and } y^T\lambda = 0\} & \text{if } \lambda \ge 0, \\ \emptyset & \text{otherwise.} \end{cases} \tag{2.5}$$

The *active set* at z^* is defined by

$$\mathcal{B} = \{i = 1, 2, \ldots, m \,|\, g_i(z^*) = 0\}. \tag{2.6}$$

For any optimal multiplier $\lambda^* \in \mathcal{S}_\lambda$, we define the set $\mathcal{B}_+(\lambda^*)$ to be the "support" of λ^*, that is,

$$\mathcal{B}_+(\lambda^*) = \{i \in \mathcal{B} \,|\, \lambda_i^* > 0\}.$$

We define $\mathcal{B}_+$ (without argument) as

$$\mathcal{B}_+ \stackrel{\text{def}}{=} \cup_{\lambda^* \in \mathcal{S}_\lambda} \mathcal{B}_+(\lambda^*) \tag{2.7}$$

and denote its complement in $\mathcal{B}$ by $\mathcal{B}_0$, that is,

$$\mathcal{B}_0 \stackrel{\text{def}}{=} \mathcal{B} \backslash \mathcal{B}_+.$$

Note that $\mathcal{B}_0$ is the set of indices $i \in \mathcal{B}$ such that $\lambda_i^* = 0$ for all $\lambda^* \in \mathcal{S}_\lambda$. The SC condition for the set $\mathcal{S}$ (which we use only sparingly in this paper) is that

$$\mathcal{B}_0 = \emptyset. \tag{2.8}$$

At some points in the paper, we use a condition that Fischer [8] calls *weak complementarity* (WCC), namely that

$$\text{Range}[\nabla g_i(z^*)]_{i \in \mathcal{B}_+(\lambda^*)} = \text{Range}[\nabla g_i(z^*)]_{i \in \mathcal{B}_+} \qquad \text{for all } \lambda^* \in \mathcal{S}_\lambda. \tag{2.9}$$

Despite its name, WCC is not weaker than SC; neither condition implies the other.

In section 4, we define the term *strict working set* to be, roughly speaking, the set of indices in $i \in \{1, 2, \ldots, m\}$ for which the Lagrange multipliers λ_i of the (possibly inexact) QP subproblem are strictly positive.

We assume throughout that the MFCQ holds at z^* [15]. That is,

$$\nabla g_{\mathcal{B}}(z^*)^T y < 0 \quad \text{for some } y \in \mathsf{R}^n, \tag{2.10}$$

where $\nabla g_{\mathcal{B}}(\cdot)$ is the $n \times |\mathcal{B}|$ matrix whose rows $\nabla g_i(\cdot)$, $i \in \mathcal{B}$, are the gradients of the functions g_i, $i \in \mathcal{B}$.

The general smoothness and first-order assumption that we make throughout the paper is as follows.

Assumption 1. The functions $\phi(\cdot)$ and $g(\cdot)$ are twice Lipschitz continuously differentiable in an open neighborhood of z^*, and the first-order condition (2.2) is satisfied at z^*.

The following result concerning boundedness of the optimal multiplier set $\mathcal{S}_\lambda$ is often used in the analysis of later sections.

LEMMA 2.1 (Gauvin [9]). *Suppose that Assumption* 1 *holds. Then* $\mathcal{S}_\lambda$ *defined in* (2.3a) *is bounded if and only if the MFCQ* (2.10) *is satisfied.*

Since $\mathcal{S}_\lambda$ is defined by the linear conditions $\nabla\phi(z^*) + \nabla g(z^*)\lambda^* $ and $\lambda^* \geq 0$, it is closed and convex. Therefore under the conditions of Lemma 2.1, it is also compact.

We use the notation $\delta(\cdot)$ to denote Euclidean distances from the primal, dual, and primal-dual optimal sets, according to context. Specifically, we define

$$\delta(z) \stackrel{\text{def}}{=} \|z - z^*\|, \quad \delta(\lambda) \stackrel{\text{def}}{=} \operatorname{dist}(\lambda, \mathcal{S}_\lambda), \quad \delta(z,\lambda) \stackrel{\text{def}}{=} \operatorname{dist}((z,\lambda), \mathcal{S}). \tag{2.11}$$

We also use $P(\lambda)$ to denote the projection of λ onto $\mathcal{S}_\lambda$; that is, we have $P(\lambda) \in \mathcal{S}_\lambda$ and $\|P(\lambda) - \lambda\| = \operatorname{dist}(\lambda, \mathcal{S}_\lambda)$. Note that from (2.11) we have $\delta(z)^2 + \delta(\lambda)^2 = \delta(z,\lambda)^2$, and therefore

$$\delta(z) \leq \delta(z,\lambda), \qquad \delta(\lambda) \leq \delta(z,\lambda). \tag{2.12}$$

For further analysis of these errors, we use $\mathcal{B}$ and $\mathcal{B}_+$ to define a direction set $\mathcal{T}$ as

$$\mathcal{T} = \left\{ w \mid \begin{array}{ll} \nabla g_i(z^*)^T w = 0 & \text{for } i \in \mathcal{B}_+ \\ \nabla g_i(z^*)^T w \leq 0 & \text{for } i \in \mathcal{B}_0 \end{array} \right\}. \tag{2.13}$$

We define the primal error by

$$e(z) \stackrel{\text{def}}{=} z - z^*,$$

and decompose it as

$$e(z) = e_{\mathcal{T}}(z) + e_{\mathcal{N}}(z), \tag{2.14}$$

where $e_{\mathcal{T}}(z)$ is the projection of $e(z)$ onto the cone $\mathcal{T}$ and $e_{\mathcal{N}}(z)$ is the remainder (which is, of course, normal to $\mathcal{T}$ at $e_{\mathcal{T}}(z)$). In fact, there are coefficients ξ_i, $i \in \mathcal{B}$ (not necessarily unique), such that

$$e_{\mathcal{N}}(z) = \sum_{i \in \mathcal{B}_+} \xi_i \nabla g_i(z^*) + \sum_{i \in \mathcal{B}_0} \xi_i \nabla g_i(z^*), \qquad \xi_i \geq 0 \text{ for } i \in \mathcal{B}_0. \tag{2.15}$$

Since $\mathcal{T}$ is a cone, it is easy to see that $e_{\mathcal{T}}(\cdot)$ and $e_{\mathcal{N}}(\cdot)$ are continuous in their arguments and that

$$e_{\mathcal{N}}(\alpha z) = \alpha e_{\mathcal{N}}(z), \qquad e_{\mathcal{T}}(\alpha z) = \alpha e_{\mathcal{T}}(z) \qquad \text{for all } \alpha \geq 0.$$

Moreover, since $e_{\mathcal{N}}(z)^T e_{\mathcal{T}}(z) = 0$, we have

$$\|e_{\mathcal{N}}(z)\|^2 + \|e_{\mathcal{T}}(z)\|^2 = \|e(z)\|^2 = \delta(z)^2 \leq \delta(z,\lambda)^2$$

and therefore

$$\|e_{\mathcal{N}}(z)\| \leq \delta(z), \qquad \|e_{\mathcal{T}}(z)\| \leq \delta(z). \tag{2.16}$$

We use order notation in the following (standard) way: If a matrix, vector, or scalar quantity M is a function of another matrix, vector, or scalar quantity A, we write $M = O(\|A\|)$ if there is a constant β such that $\|M\| \leq \beta\|A\|$ for all $\|A\|$ sufficiently small. We write $M = \Omega(\|A\|)$ if there is a constant β such that $\|M\| \geq \beta^{-1}\|A\|$ for all $\|A\|$ sufficiently small, and $M = \Theta(\|A\|)$ if both $M = O(\|A\|)$ and $M = \Omega(\|A\|)$. We write $M = o(\|A\|)$ if, for all sequences $\{A_k\}$ with $\|A_k\| \to 0$, the corresponding sequence $\{M_k\}$ satisfies $\|M_k\|/\|A_k\| \to 0$.

If r is a vector and $\mathcal{A}$ is an index set, we use $r_{\mathcal{A}}$ to denote the subvector consisting of components r_i, $i \in \mathcal{A}$.

3. Second-order conditions. The presence of degeneracy allows for a variety of second-order sufficient conditions, all of which can be expected to hold for a wide range of problems and all of which are useful in investigating the local convergence properties of various algorithms. In this section, we define three such conditions that are needed by algorithms in later sections. We also introduce "extended" variants of the nonlinear programming problem (1.1) that differ from (1.1) only in that just a subset of the constraints is enforced. For some of these subsets, z^* remains a strict local solution satisfying some second-order sufficient condition; such subsets are particularly useful in the context of the algorithm to be discussed in section 8. Finally, we include here several results that relate the conditions introduced in this section to the assumptions of the preceding section.

Second-order sufficient conditions typically assume that there is a positive value $\sigma > 0$ such that the condition

$$w^T \mathcal{L}_{zz}(z^*, \lambda^*) w \geq \sigma \|w\|^2 \tag{3.1}$$

holds for some set of λ^* and w vectors. The three conditions used in this paper are as follows.

Condition 2s.1 (second-order sufficient condition). The condition (3.1) holds for all $\lambda^* \in \mathcal{S}_\lambda$ and all w such that

$$\begin{aligned} \nabla g_i(z^*)^T w &= 0 \quad \text{for all } i \in \mathcal{B}_+, \\ \nabla g_i(z^*)^T w &\leq 0 \quad \text{for all } i \in \mathcal{B}_0; \end{aligned}$$

that is, $w \in \mathcal{T}$.

Condition 2s.2 (strong second-order sufficient condition). The condition (3.1) holds for all $\lambda^* \in \mathcal{S}_\lambda$ and all w such that

$$\nabla g_i(z^*)^T w = 0 \quad \text{for all } i \in \mathcal{B}_+.$$

Condition 2s.3 (locally strong second-order sufficient condition). For each $\lambda^* \in \mathcal{S}_\lambda$, the condition (3.1) holds for all w such that

$$\nabla g_i(z^*)^T w = 0 \quad \text{for all } i \in \mathcal{B}_+(\lambda^*). \tag{3.2}$$

Any of these conditions, in tandem with Assumption 1, is sufficient to guarantee that z^* is a strict local solution of (1.1) (see, for instance, Bertsekas [5, Proposition 3.3.2, Exercise 3.3.7]). The following lemma relates these three conditions with the WCC and SC conditions of section 2.

LEMMA 3.1.

(i) *Condition* 2s.3 $\Rightarrow$ *Condition* 2s.2 $\Rightarrow$ *Condition* 2s.1.

(ii) *If the SC condition* (2.8) *holds, then Conditions* 2s.2 *and* 2s.1 *are identical.*

(iii) *If the WCC condition* (2.9) *holds, then Conditions* 2s.3 *and* 2s.2 *are identical.*

Proof. The proof of (i) is obvious, since the set of vectors w on which (3.1) is required to hold is successively larger as we go from 2s.1 to 2s.2 to 2s.3. Statement (ii) follows immediately from the definition (2.8) of SC. For (iii), note that (2.9) implies that

$$\text{null}\,[\nabla g_i(z^*)]^T_{i\in\mathcal{B}_+(\lambda^*)} \;=\; \text{null}\,[\nabla g_i(z^*)]^T_{i\in\mathcal{B}_+} \qquad \text{for all } \lambda^* \in \mathcal{S}_\lambda,$$

from which the result follows immediately. ∎

For any subset $\tilde{\mathcal{B}} \subseteq \mathcal{B}$ (where $\mathcal{B}$ is the active set defined in (2.6)), we can define the nonlinear program in which just the constraints $i \in \tilde{\mathcal{B}}$ are enforced as follows:

$$\text{NLP}(\tilde{\mathcal{B}}): \qquad \min_z \phi(z) \qquad \text{subject to } g_i(z) \le 0, \text{ all } i \in \tilde{\mathcal{B}}. \tag{3.3}$$

Note that any first-order point $(z^*, \lambda^*_{\tilde{\mathcal{B}}})$ for (3.3) can be extended to a first-order point for (1.1) by simply adding zeros to fill out the components $i \notin \tilde{\mathcal{B}}$. Conversely, we can recover a point $(z^*, \lambda^*_{\tilde{\mathcal{B}}})$ that satisfies the first-order conditions for (3.3) from any solution $(z^*, \lambda^*) \in \mathcal{S}$ of (1.1) for which $\lambda^*_i = 0$, $i \notin \tilde{\mathcal{B}}$, by deleting the (zero) components $i \notin \tilde{\mathcal{B}}$ from λ^*. Note, however, that the vector so obtained does not necessarily satisfy second-order conditions for (3.3); in fact, z^* may not even be a local solution for (3.3).

We now define two sets Φ and $\bar{\Phi}$ made up of subsets $\tilde{\mathcal{B}} \subseteq \mathcal{B}$ as follows:

$$\Phi \stackrel{\text{def}}{=} \{\tilde{\mathcal{B}} \subseteq \mathcal{B} \,|\, z^* \text{ is a local solution of NLP}(\tilde{\mathcal{B}}) \text{ that satisfies Condition 2s.1 applied to NLP}(\tilde{\mathcal{B}})\}, \tag{3.4}$$

$$\bar{\Phi} \stackrel{\text{def}}{=} \{\tilde{\mathcal{B}} \in \Phi \,|\, \text{the optimal Lagrange multiplier for NLP}(\tilde{\mathcal{B}}) \text{ is unique}\}. \tag{3.5}$$

When $\tilde{\mathcal{B}} \in \bar{\Phi}$, we use $\lambda^*(\tilde{\mathcal{B}})$ to denote the unique optimal multiplier for NLP$(\tilde{\mathcal{B}})$, padded out with zeros to length m. Note that $\mathcal{B} \in \Phi$, so that $\Phi \neq \emptyset$. When the SC and nondegeneracy conditions hold at the solution of (1.1), we have $\Phi = \bar{\Phi} = \{\mathcal{B}\}$.

The sets Φ and $\bar{\Phi}$ become particularly relevant in section 8, where we propose an algorithm whose steps are obtained by applying SQP to problems of the form NLP$(\tilde{\mathcal{B}})$. For now, we prove two simple results about the way that these sets are related to each other and to the second-order sufficient conditions.

LEMMA 3.2. *Given some set $\tilde{\mathcal{B}} \in \Phi$, a sufficient condition for $\tilde{\mathcal{B}} \in \bar{\Phi}$ is that the vectors $\{\nabla g_i(z^*),\ i \in \tilde{\mathcal{B}} \cap \mathcal{B}_+\}$ are linearly independent.*

Proof. Since $\tilde{\mathcal{B}} \in \Phi$, there is a vector $\lambda^*_{\tilde{\mathcal{B}}}$ such that

$$\sum_{i\in\tilde{\mathcal{B}}} \lambda^*_i \nabla g_i(z^*) + \nabla\phi(z^*) = 0. \tag{3.6}$$

For the components $i \in \tilde{\mathcal{B}} \cap \mathcal{B}_0$, we must have that $\lambda^*_i = 0$, since otherwise the vector $\lambda^*_{\tilde{\mathcal{B}}}$ could be padded out with zeros to yield a vector $\lambda^* \in \mathcal{S}_\lambda$ with $\lambda^*_i \neq 0$ for $i \in \mathcal{B}_0$, contradicting the definition of $\mathcal{B}_+$. Hence, we can rewrite (3.6) as

$$\sum_{i\in\tilde{\mathcal{B}}\cap\mathcal{B}_+} \lambda^*_i \nabla g_i(z^*) + \nabla\phi(z^*) = 0,$$

and so linear independence of the given vector set implies uniqueness of $\lambda^*_{\tilde{\mathcal{B}}}$. Therefore $\tilde{\mathcal{B}} \in \bar{\Phi}$, as claimed. □

LEMMA 3.3. *Suppose that Condition* 2s.3 *holds. Then Condition* 2s.3 *also holds for NLP($\tilde{\mathcal{B}}$), where $\tilde{\mathcal{B}} \subseteq \mathcal{B}$ is such that there exists a $\lambda^* \in \mathcal{S}_\lambda$ with $\mathcal{B}_+(\lambda^*) \subseteq \tilde{\mathcal{B}}$. In particular, $\tilde{\mathcal{B}} \in \Phi$.*

Proof. Suppose that Condition 2s.3 holds for NLP. For all $\lambda^* \in \mathcal{S}_\lambda$ with $\mathcal{B}_+(\lambda^*) \subseteq \tilde{\mathcal{B}}$, we have by the correspondence between optimal Lagrange multipliers for NLP and NLP($\tilde{\mathcal{B}}$) discussed above that (3.1) holds for all w satisfying (3.2). Hence, Condition 2s.3 holds for the problem NLP($\tilde{\mathcal{B}}$) as well. Since there is at least one vector λ^* with the properties indicated, and since Condition 2s.3 implies Condition 2s.1 by Lemma 3.1(i), we have $\tilde{\mathcal{B}} \in \Phi$. □

4. The iSQP framework. In the best-known form of the SQP algorithm, the following inequality constrained subproblem is solved to obtain the step at each iteration:

$$\begin{aligned} &\min_{\Delta z} \ \Delta z^T \nabla\phi(z) + \tfrac{1}{2}\Delta z^T \mathcal{L}_{zz}(z,\lambda)\Delta z \\ &\text{subject to } g(z) + \nabla g(z)^T \Delta z \le 0, \end{aligned} \tag{4.1}$$

where (z, λ) is the current primal-dual iterate. Denoting the Lagrange multipliers for the constraints in (4.1) by λ^+, we see that the solution Δz satisfies the following KKT conditions (cf. (2.4)):

$$\begin{bmatrix} \mathcal{L}_{zz}(z,\lambda)\Delta z + \nabla\phi(z) + \nabla g(z)\lambda^+ \\ g(z) + \nabla g(z)^T \Delta z \end{bmatrix} \in \begin{bmatrix} 0 \\ N(\lambda^+) \end{bmatrix}. \tag{4.2}$$

We focus our attention, however, on a more general framework that allows for inexactness in the subproblem solution by introducing perturbations into both the objective and constraints of (4.1). We assume only that $(\Delta z, \lambda^+)$ is the solution of

$$\begin{aligned} &\min_{\Delta z} \ \Delta z^T (\nabla\phi(z) + t) + \tfrac{1}{2}\Delta z^T \mathcal{L}_{zz}(z,\lambda)\Delta z \\ &\text{subject to } g(z) + \nabla g(z)^T \Delta z + r \le 0, \end{aligned} \tag{4.3}$$

for some perturbation vectors t and r. The KKT conditions for (4.3) are

$$\begin{bmatrix} \mathcal{L}_{zz}(z,\lambda)\Delta z + \nabla\phi(z) + t + \nabla g(z)\lambda^+ \\ g(z) + \nabla g(z)^T \Delta z + r \end{bmatrix} \in \begin{bmatrix} 0 \\ N(\lambda^+) \end{bmatrix}. \tag{4.4}$$

We introduce further terminology and notation here: Given a primal-dual solution $(\Delta z, \lambda^+)$ to (4.1) or (4.3), the *strict working set* is the set of indices i for which λ_i^+ is strictly positive. We denote this set by $\mathcal{B}(z,\lambda)$ in the case of (4.1) and $\mathcal{B}(z,\lambda;t,r)$ in the case of (4.3).

When (z,λ) is sufficiently close to $\mathcal{S}$, it happens under mild assumptions that the strict working set $\mathcal{B}(z,\lambda;t,r)$ from an iteration of iSQP identifies a superset of $\mathcal{B}_+(\lambda^*)$ for some optimal multiplier λ^*. This result is interesting for its own sake and also in the context of section 8, so we prove it here.

LEMMA 4.1. *Suppose that Assumption* 1, *Condition* 2s.1, *and MFCQ hold. Then there is a threshold value $\bar{\delta}$ such that whenever $\delta(z,\lambda) \le \bar{\delta}$ and $\|(t,r)\| \le \bar{\delta}$, the solution of the iSQP subproblem* (4.3) *yields a strict working set $\mathcal{B}(z,\lambda;t,r)$ such that $\mathcal{B}_+(\lambda^*) \subseteq \mathcal{B}(z,\lambda;t,r)$ for at least one $\lambda^* \in \mathcal{S}_\lambda$.*

Proof. Suppose, on the contrary, that there is a sequence (z^ℓ, λ^ℓ) with $\delta(z^\ell, \lambda^\ell) \downarrow 0$ and a sequence of perturbations (t^ℓ, r^ℓ) with $\|(t^\ell, r^\ell)\| \downarrow 0$ such that the stated property does not hold. That is, taking the active set $\mathcal{B}(z^\ell, \lambda^\ell; t^\ell, r^\ell)$ for the iSQP subproblem, we find that the subvector $\lambda^*_{\mathcal{B}\setminus\mathcal{B}(z^\ell,\lambda^\ell;t^\ell,r^\ell)}$ is nonzero for all $\lambda^* \in \mathcal{S}_\lambda$. By taking a subsequence if necessary, we can assume that $\mathcal{B}(z^\ell, \lambda^\ell; t^\ell, r^\ell) \equiv \hat{\mathcal{B}} \subseteq \mathcal{B}$ for all ℓ. By the compactness of $\mathcal{S}_\lambda$ (Lemma 2.1) and continuity of $\|\lambda^*_{\mathcal{B}\setminus\hat{\mathcal{B}}}\|$ as a function of λ^*, we have that

$$0 < \zeta \stackrel{\text{def}}{=} \min_{\lambda^* \in \mathcal{S}_\lambda} \left\|\lambda^*_{\mathcal{B}\setminus\hat{\mathcal{B}}}\right\|.$$

From Lemma 5.1, we have that the updated multiplier $(\lambda^\ell)^+$ obtained from the iSQP subproblem (4.3) at $(z^\ell, \lambda^\ell, t^\ell, r^\ell)$ satisfies

$$\delta((\lambda^\ell)^+) = O(\delta(z^\ell)) + O(\|(t^\ell, r^\ell)\|) \downarrow 0.$$

Denoting by $P((\lambda^\ell)^+)$ the projection of $(\lambda^\ell)^+$ onto the set $\mathcal{S}_\lambda$, we have that

$$\delta((\lambda^\ell)^+)^2 = \left\|(\lambda^\ell)^+ - P((\lambda^\ell)^+)\right\|^2 \geq \sum_{i \in \mathcal{B}\setminus\hat{\mathcal{B}}} \left[P((\lambda^\ell)^+)_i\right]^2 \geq \zeta^2 > 0,$$

giving a contradiction. ∎

5. Local convergence of iSQP. In this section, we describe the improvement obtained in a single iSQP step (4.3) (alternatively, (4.4)).

First, we apply a result of Robinson [17] to show that small-norm local solutions Δz exist for the inexact subproblem (4.3), provided that (z, λ) is sufficiently close to the solution set $\mathcal{S}$ defined in (2.3b). Our two main results, Theorems 5.2 and 5.3, relate the errors $e_{\mathcal{N}}(z + \Delta z)$ and $\delta(z + \Delta z, \lambda^+)$ at the new iterate to errors at the current point (z, λ). In particular, Theorem 5.3 demonstrates that superlinear convergence of the primal iterate depends critically on stability of the Lagrange multiplier estimates: λ must not change too much from one iteration to the next.

The first result is obtained by applying Robinson's stability results in [17] to the iSQP subproblem (4.3).

LEMMA 5.1. *Suppose that Assumption* 1, *Condition* 2s.1, *and MFCQ hold. Then for all* (z, λ, t, r) *with* $\delta(z, \lambda)$ *and* $\|(t, r)\|$ *sufficiently small, the problem* (4.3) *has a local solution* Δz *near* 0 *that satisfies*

$$\|\Delta z\| + \delta(\lambda^+) = O(\delta(z)) + O(\|(t, r)\|), \tag{5.1}$$

where λ^+ *is the vector of multipliers corresponding to the solution* Δz *of* (4.3).

Proof. For any fixed $\lambda^* \in \mathcal{S}_\lambda$, consider the problem

$$\begin{aligned} \min_{\Delta z} \ & \Delta z^T \nabla\phi(z^*) + \tfrac{1}{2}\Delta z^T \mathcal{L}_{zz}(z^*, \lambda^*)\Delta z \\ & \text{subject to } g(z^*) + \nabla g(z^*)^T \Delta z \leq 0, \end{aligned} \tag{5.2}$$

whose primal-dual solution set is $\{0\} \times \mathcal{S}_\lambda$. MFCQ holds for this problem, since the active set $\mathcal{B}$ is the same as for the nonlinear problem (1.1). It is easy to see that Condition 2s.1 is satisfied as well.

Consider now the parametrized version (4.3) of the problem (5.2), in which the parametrization is defined by the vector $p = (z, \lambda, t, r)$. (We recover (5.2) from (4.3) by setting $p = p_0 = (z^*, \lambda^*, 0, 0)$.) We define the following subsets of R^n:

$\text{stat}(p) = \{\Delta z \,|\, (\Delta z, \lambda^+) \text{ satisfies KKT conditions for (4.3) for some } \lambda^+\}$,
$\text{lsol}(p) = \{\Delta z \,|\, \Delta z \text{ is a local solution of (4.3)}\}$.

By applying Theorem 3.2 of Robinson [17], we find that there is a neighborhood $N_1(\lambda^*)$ of p_0 and a neighborhood $M_1(\lambda^*)$ of 0 such that

$\text{stat}(p) \cap M_1(\lambda^*)$ is continuous in p at p_0;
$\text{lsol}(p) \cap M_1(\lambda^*)$ is nonempty for $p \in N_1(\lambda^*)$ and a subset of $\text{stat}(p) \cap M_1(\lambda^*)$.

Moreover, we have from Theorem 4.2 of Robinson [17] that there is a neighborhood $N_2(\lambda^*)$ of p_0 and a constant $\epsilon(\lambda^*) > 0$ such that for each $p \in N_2(\lambda^*)$ we have that all stationary pairs $(\Delta z, \lambda^+)$ for (4.3) satisfy

$$\begin{aligned}(5.3)\quad &\inf_{\bar\lambda \in \mathcal{S}_\lambda} \|(\Delta z, \lambda^+ - \bar\lambda)\| \\ &\le \epsilon(\lambda^*) \left\| \begin{bmatrix} (\mathcal{L}_{zz}(z,\lambda) - \mathcal{L}_{zz}(z^*,\lambda^*))\Delta z + (\nabla\phi(z) - \nabla\phi(z^*)) \\ \qquad\qquad +(\nabla g(z) - \nabla g(z^*))\lambda^+ + t \\ (g(z) - g(z^*)) + (\nabla g(z) - \nabla g(z^*))^T \Delta z + r \end{bmatrix} \right\|.\end{aligned}$$

For the left-hand side of this expression, it is easy to see that

$$(5.4)\qquad \inf_{\bar\lambda \in \mathcal{S}_\lambda} \|(\Delta z, \lambda^+ - \bar\lambda)\| \ge \frac{1}{\sqrt{2}} \left(\|\Delta z\| + \delta(\lambda^+)\right).$$

For the right-hand side, we have by the Lipschitz continuity of $\nabla^2\phi$ and $\nabla^2 g_i$, $i = 1, 2, \ldots, m$, that

$$\begin{aligned} \|[\mathcal{L}_{zz}(z,\lambda) - \mathcal{L}_{zz}(z^*,\lambda^*)]\Delta z\| &\le C_1 \left(\delta(z) + \|\lambda - \lambda^*\|\right) \|\Delta z\|, \\ \|\nabla\phi(z) - \nabla\phi(z^*)\| &\le C_1 \delta(z), \\ \|g(z) + \nabla g(z)^T \Delta z - g(z^*) - \nabla g(z^*)^T \Delta z\| &\le C_1 \delta(z) \left(1 + \|\Delta z\|\right) \end{aligned}$$

for some constant C_1 (which is, in particular, independent of the multiplier λ^*). Moreover, by the boundedness of $\mathcal{S}_\lambda$, we have after a possible adjustment of C_1 that

$$\left\|\nabla g(z)\lambda^+ - \nabla g(z^*)\lambda^+\right\| \le C_1 \|\lambda^+\| \delta(z) \le C_1 \left(C_2 + \delta(\lambda^+)\right) \delta(z),$$

where C_2 is a constant that bounds the norms of all elements of $\mathcal{S}_\lambda$. By using these expressions together with (5.4) in (5.3), we have

$$\begin{aligned}(5.5)\quad &\|\Delta z\| + \delta(\lambda^+) \\ &\le \epsilon(\lambda^*) C_3 \left[\|\lambda - \lambda^*\| \|\Delta z\| + \delta(z)\|\Delta z\| + \delta(z) + \delta(\lambda^+)\delta(z) + \|(t,r)\|\right],\end{aligned}$$

where C_3 is a constant (in particular, independent of the particular choice of $\lambda^* \in \mathcal{S}_\lambda$). By reducing the size of $N_2(\lambda^*)$ if necessary, we have that

$$(5.6)\qquad \epsilon(\lambda^*) C_3 \left[\|\lambda - \lambda^*\| + \delta(z)\right] \le 1/4 \;\text{ for all } (z, \lambda, t, r) \in N_2(\lambda^*).$$

Thus, by transferring terms involving $\|\Delta z\|$ and $\delta(\lambda^+)$ from the right-hand side to the left-hand side in (5.5), we obtain

$$\begin{aligned} &\|\Delta z\| \left[1 - \epsilon(\lambda^*) C_3 (\|\lambda - \lambda^*\| + \delta(z))\right] \\ &+ \delta(\lambda^+) \left[1 - \epsilon(\lambda^*) C_3 \delta(z)\right] \le \epsilon(\lambda^*) C_3 \left[\delta(z) + \|(t,r)\|\right], \end{aligned}$$

so that from (5.6) we have that

$$(5.7)\qquad \|\Delta z\| + \delta(\lambda^+) \le 2\epsilon(\lambda^*) C_3 \left[\delta(z) + \|(t,r)\|\right].$$

Note that (5.7) holds for the *fixed* choice of λ^* in $\mathcal{S}_\lambda$, and only for (z, λ, t, r) within the neighborhood $N_2(\lambda^*)$ of $(z^*, \lambda^*, 0, 0)$. Since

$$\{N_2(\lambda^*) \,|\, \lambda^* \in \mathcal{S}_\lambda\}$$

is a cover of the set

$$\{z^*\} \times \mathcal{S}_\lambda \times \{0\} \times \{0\}, \tag{5.8}$$

we have by the compactness of $\mathcal{S}_\lambda$ (and hence of (5.8)) that there is a finite subcover, say

$$\{N_2(\lambda_l^*) \,|\, \lambda_l^* \in \mathcal{S}_\lambda,\ l = 1, 2, \ldots, L\}.$$

Note that the set

$$\cup_{l=1,2,\ldots,L} N_2(\lambda_l^*)$$

is a neighborhood of (5.8).

By setting

$$C_4 = 2C_3 \max_{l=1,2,\ldots,L} \epsilon(\lambda_l^*),$$

we have from (5.7) that

$$\|\Delta z\| + \delta(\lambda^+) \le C_4 \left[\delta(z) + \|(t, r)\|\right] \tag{5.9}$$

whenever

$$(z, \lambda, t, r) \in \cup_{l=1,2,\ldots,L} N_2(\lambda_l^*). \tag{5.10}$$

We can choose $\bar{\delta}$ sufficiently small that

$$\delta(z, \lambda) \le \bar{\delta},\ \|(t, r)\| \le \bar{\delta} \ \Rightarrow\ (z, \lambda, t, r) \in \cup_{l=1,2,\ldots,L} N_2(\lambda_l^*),$$

so that (5.9) holds for all (z, λ) with $\delta(z, \lambda) \le \bar{\delta}$ and $\|(t, r)\| \le \bar{\delta}$. ∎

In subsequent discussions, we use the term "iSQP" to describe the inexact SQP procedure in which each iteration consists of obtaining a solution to the problem (4.3) from the current iterate (z, λ) and then setting

$$(z, \lambda) \leftarrow (z + \Delta z, \lambda^+), \tag{5.11}$$

where $(\Delta z, \lambda^+)$ is a primal-dual solution of (4.3) that satisfies (5.1).

The next result shows that while the iSQP step may not give a "superlinear" decrease in distance to the solution set, it does reduce the error substantially in the $e_{\mathcal{N}}(\cdot)$ component of the primal error vector. (This result explains an observation made while doing computational experiments for an earlier paper [18]. It is similar to Lemma 3.12 of Fischer [8], though the latter result assumes WCC (2.9), which is not needed below.)

THEOREM 5.2. *Suppose that Assumption* 1*, Condition* 2s.1*, and MFCQ hold. For all* (z, λ) *with* $\delta(z, \lambda)$ *and* $\|(t, r)\|$ *sufficiently small, the new iterate generated by the iSQP algorithm satisfies*

$$\|e_{\mathcal{N}}(z + \Delta z)\| = O(\delta(z)^2) + O(\|t\|^2) + O(\|r\|). \tag{5.12}$$

Moreover, we have

$$g_{\mathcal{B}_+}(z+\Delta z) = O(\delta(z)^2) + O(\|t\|^2) + O(\|r\|). \tag{5.13}$$

Proof. Let $\{(z^k, \lambda^k, t^k, r^k)\}$ be any sequence with

$$\delta(z^k, \lambda^k) \to 0, \qquad \|(t^k, r^k)\| \to 0,$$

and let $(\Delta z^k, \lambda^{k+})$ be the primal-dual solution to (4.3) obtained when $(z, \lambda) = (z^k, \lambda^k)$ and $(t, r) = (t^k, r^k)$. Denote the corresponding sequence of strict working sets for the iSQP subproblem (4.3) by $\mathcal{B}^k$. From Lemma 5.1, we have $\Delta z^k \to 0$, so since $g(z^k) \to g(z^*)$ and $r^k \to 0$, none of the inactive indices $j \notin \mathcal{B}$ can be active in the subproblem for k sufficiently large. We can therefore assume without loss of generality that $\mathcal{B}^k \subseteq \mathcal{B}$.

Note first that, from (5.1) and $\|e(z)\| = \delta(z)$, we have

$$\|e(z^k + \Delta z^k)\| \le \|e(z^k)\| + \|\Delta z^k\| = O(\delta(z^k)) + O(\|(t^k, r^k)\|). \tag{5.14}$$

For all active indices $i \in \mathcal{B}^k$, we have from Taylor's theorem together with (5.14) and (4.3) that

$$\begin{aligned}
\nabla g_i(z^*)^T e(z^k + \Delta z^k) &= g_i(z^k + \Delta z^k) - g_i(z^*) + O(\delta(z^k)^2) + O(\|(t^k, r^k)\|^2) \\
&= g_i(z^k) + \nabla g_i(z^k)^T \Delta z^k + O(\delta(z^k)^2) + O(\|(t^k, r^k)\|^2) \\
&= -r_i^k + O(\delta(z^k)^2) + O(\|(t^k, r^k)\|^2) \\
&= \tilde{r}_i^k, \qquad (5.15)
\end{aligned}$$

where

$$\tilde{r}_i^k = O(\delta(z^k)^2) + O(\|t^k\|^2) + O(\|r^k\|). \tag{5.16}$$

Meanwhile, for $i \in \mathcal{B} \backslash \mathcal{B}^k$, we have from (4.2) and (5.1) that

$$\begin{aligned}
\nabla g_i(z^*)^T e(z^k + \Delta z^k) &= g_i(z^k) + \nabla g_i(z^k)^T \Delta z^k + O(\delta(z^k)^2) + O(\|(t^k, r^k)\|^2) \\
&\le -r_i^k + O(\delta(z^k)^2) + O(\|(t^k, r^k)\|^2) \\
&\le \tilde{r}_i^k, \qquad (5.17)
\end{aligned}$$

where the estimate (5.16) holds once again.

Since $(\Delta z^k, \lambda^{k+})$ is the solution to (4.3), then, by boundedness of $\mathcal{S}_\lambda$, we have from (4.4) and (5.1) that

$$\begin{aligned}
&\nabla \phi(z^k) + \mathcal{L}_{zz}(z^k, \lambda^k) \Delta z^k + \sum_{i \in \mathcal{B}^k} \lambda_i^{k+} \nabla g_i(z^k) + t^k = 0 \\
&\Rightarrow \nabla \phi(z^*) + \sum_{i \in \mathcal{B}^k} \lambda_i^{k+} \nabla g_i(z^*) = O(\delta(z^k)) + O(\|(t^k, r^k)\|).
\end{aligned}$$

By the definition (2.7) of $\mathcal{B}_+$, there is a $\lambda^* \in \mathcal{S}_\lambda$ such that

$$\nabla \phi(z^*) + \sum_{i \in \mathcal{B}_+} \lambda_i^* \nabla g_i(z^*) = 0, \qquad \lambda^*_{\mathcal{B}_+} > 0. \tag{5.18}$$

(We can construct λ^* by taking $\lambda^{(i)} \in \mathcal{S}_\lambda$ with $\lambda_i^{(i)} > 0$ for each $i \in \mathcal{B}_+$, according to the definition (2.7), and setting $\lambda^* = \sum_{i\in\mathcal{B}_+} \lambda^{(i)}/|\mathcal{B}_+|$.) By combining the last two equations, we obtain

$$\sum_{i\in\mathcal{B}_+\setminus\mathcal{B}^k} \lambda_i^* \nabla g_i(z^*) = \sum_{i\in\mathcal{B}^k} (\lambda_i^{k+} - \lambda_i^*)\nabla g_i(z^*) + O(\delta(z^k)) + O(\|(t^k, r^k)\|).$$

By taking inner products of both sides with $e(z^k + \Delta z^k)$ and using (5.14), we have by (5.15), (5.16), and boundedness of $\mathcal{S}_\lambda$ that

$$\sum_{i\in\mathcal{B}_+\setminus\mathcal{B}^k} \lambda_i^* \nabla g_i(z^*)^T e(z^k + \Delta z^k) = O(\delta(z^k)^2) + O(\|t^k\|^2) + O(\|r^k\|).$$

Since $\lambda^*_{\mathcal{B}_+} > 0$, and since by (5.17) none of the terms $\nabla g_i(z^*)^T e(z^k+\Delta z^k)$, $i \in \mathcal{B}_+\setminus\mathcal{B}^k$, can be larger than a small positive number of the size indicated in (5.16), we have that

$$\nabla g_i(z^*)^T e(z^k + \Delta z^k) = \tilde{r}_i^k \quad \text{for all } i \in \mathcal{B}_+\setminus\mathcal{B}^k,$$

where the values of $\tilde{r}_i^k$ may have been adjusted from (5.17) but still satisfy the estimate (5.16). Hence, for the indices $i \in \mathcal{B}_+\setminus\mathcal{B}^k$, we can replace the inequality by an equality in (5.17). By combining this observation with (5.15) and (5.17), we find that

$$\nabla g_{\mathcal{B}_+}(z^*)^T e(z^k + \Delta z^k) = \tilde{r}^k_{\mathcal{B}_+}, \tag{5.19a}$$

$$\nabla g_{\mathcal{B}_0}(z^*)^T e(z^k + \Delta z^k) \le \tilde{r}^k_{\mathcal{B}_0}, \tag{5.19b}$$

where

$$\|\tilde{r}^k_{\mathcal{B}}\| = O(\delta(z^k)^2) + O(\|t^k\|^2) + O(\|r^k\|). \tag{5.20}$$

Consider now the partitioning of $e(z^k+\Delta z^k)$ into its $e_{\mathcal{N}}(\cdot)$ and $e_{\mathcal{T}}(\cdot)$ components, as in (2.14). From (2.13), we see that the $e_{\mathcal{T}}$ component is obtained by solving

$$\begin{aligned} &\min_{e_{\mathcal{T}}} \frac{1}{2}\|e_{\mathcal{T}} - e(z^k + \Delta z^k)\|^2 \\ &\text{subject to } \nabla g_{\mathcal{B}_+}(z^*)^T e_{\mathcal{T}} = 0,\ \nabla g_{\mathcal{B}_0}(z^*)^T e_{\mathcal{T}} \le 0. \end{aligned} \tag{5.21}$$

The problem (5.21) is a feasible and strictly convex problem, so it has a unique solution. Also consider the following perturbation:

$$\begin{aligned} &\min_{e_{\mathcal{T}}} \frac{1}{2}\|e_{\mathcal{T}} - e(z^k + \Delta z^k)\|^2 \\ &\text{subject to } \nabla g_{\mathcal{B}_+}(z^*)^T e_{\mathcal{T}} = \tilde{r}^k_{\mathcal{B}_+},\ \nabla g_{\mathcal{B}_0}(z^*)^T e_{\mathcal{T}} \le \tilde{r}^k_{\mathcal{B}_0}, \end{aligned} \tag{5.22}$$

for which the (unique) solution is $e(z^k + \Delta z^k)$, because of (5.19). By applying Lemma B.1, we have that the solutions of (5.21) and (5.22) are related as follows:

$$\|e_{\mathcal{T}}(z^k + \Delta z^k) - e(z^k + \Delta z^k)\| = O(\|\tilde{r}^k_{\mathcal{B}}\|) = O(\delta(z^k)^2) + O(\|t^k\|^2) + O(\|r^k\|).$$

Since $e_{\mathcal{N}}(\cdot) = e(\cdot) - e_{\mathcal{T}}(\cdot)$, the result (5.12) follows immediately.

The second part of the result follows readily from a Taylor series argument, together with (5.19a), (5.14), and the estimate of $\|\tilde{r}^k\|$. ∎

We are now ready to prove the result about local convergence of the iSQP algorithm.

THEOREM 5.3. *Suppose that Assumption* 1*, Condition* 2s.1*, and the MFCQ condition hold. Suppose that a new iterate* $(z+\Delta z, \lambda^+)$ *is generated by the iSQP algorithm from the point* (z,λ)*. Then for all* (z,λ,t,r) *with* $\delta(z,\lambda)$ *and* $\|(t,r)\|$ *sufficiently small, we have*

$$\delta(z+\Delta z,\lambda^+) = \|\lambda-\lambda^+\|O(\delta(z)) + O(\delta(z)^2) + O(\|(t,r)\|). \tag{5.23}$$

For the special case in which $g(\cdot)$ *is linear, we have*

$$\delta(z+\Delta z,\lambda^+) = O(\delta(z)^2) + O(\|(t,r)\|). \tag{5.24}$$

Proof. Consider the problem

$$\begin{bmatrix} \nabla\phi(\tilde z) + \nabla g(\tilde z)\tilde\lambda + \tilde t \\ g(\tilde z) + \tilde r \end{bmatrix} \in \begin{bmatrix} 0 \\ N(\tilde\lambda) \end{bmatrix}, \tag{5.25}$$

where

$$\tilde t = \mathcal{L}_{zz}(z,\lambda)\Delta z + \nabla\phi(z) + t + \nabla g(z)\lambda^+ - \nabla\phi(z+\Delta z) - \nabla g(z+\Delta z)\lambda^+, \tag{5.26a}$$

$$\tilde r = g(z) + \nabla g(z)^T\Delta z + r - g(z+\Delta z). \tag{5.26b}$$

By (4.4), a solution of (5.25) is simply $(\tilde z,\tilde\lambda) = (z+\Delta z,\lambda^+)$. Viewing (5.25) as a perturbed version of (2.4), we can apply Corollary 4.3 of Robinson [17] to deduce that

$$\delta(z+\Delta z,\lambda^+) = O(\|(\tilde t,\tilde r)\|). \tag{5.27}$$

By using the assumed smoothness of ϕ and g, we have

$$\begin{aligned}\tilde t &= \nabla^2\phi(z)\Delta z + \sum_{i=1}^m \lambda_i\nabla^2 g_i(z)\Delta z - [\nabla\phi(z+\Delta z) - \nabla\phi(z)] \\ &\qquad -[\nabla g(z+\Delta z) - \nabla g(z)]\lambda^+ + t \\ &= \sum_{i=1}^m (\lambda_i - \lambda_i^+)\nabla^2 g_i(z)\Delta z + O(\|\Delta z\|^2) + t,\end{aligned}$$

so by using the boundedness of the sets containing λ and λ^+ and the estimate (5.1), we obtain

$$\begin{aligned}\|\tilde t\| &\le \|\lambda-\lambda^+\|\,[O(\delta(z)) + O(\|(t,r)\|)] + O(\delta(z)^2) + O(\|(t,r)\|^2) + O(\|t\|) \\ &= \|\lambda-\lambda^+\|O(\delta(z)) + O(\delta(z)^2) + O(\|(t,r)\|).\end{aligned} \tag{5.28}$$

For $\tilde r$, we have

$$\tilde r = O(\|\Delta z\|^2) + r = O(\delta(z)^2) + O(\|(t,r)\|). \tag{5.29}$$

The result (5.23) is immediate from (5.27), (5.28), and (5.29).

The second result (5.24) also is immediate if we observe that the term containing $\lambda-\lambda^+$ vanishes from $\tilde t$ when $g(\cdot)$ is linear. ∎

6. sSQP. Superlinear convergence of the sSQP method has been discussed by Wright [18] and Hager [11]. We show here that this method can be placed in the iSQP framework and that the superlinear convergence result therefore can be derived from Theorem 5.3.

The sSQP algorithm is derived by applying proximal point ideas to the SQP subproblem (4.1). Specifically, it adds a term to the objective for (4.1) that penalizes the step from λ to λ^+. From a current primal-dual iterate (z, λ), we find a local solution of the following minimax subproblem for $(\Delta z, \lambda^+)$ such that $(\Delta z, \lambda^+ - \lambda)$ is small:

$$\begin{aligned} \min_{\Delta z} \max_{\lambda^+ \geq 0} \; & \Delta z^T \nabla \phi(z) + \tfrac{1}{2} \Delta z^T \mathcal{L}_{zz}(z, \lambda) \Delta z \\ & + (\lambda^+)^T [g(z) + \nabla g(z)^T \Delta z] - \tfrac{1}{2} \mu \|\lambda^+ - \lambda\|^2, \end{aligned} \tag{6.1}$$

where μ is a positive parameter that may be varied from one iteration to the next. The first-order conditions for $(\Delta z, \lambda^+)$ to solve (6.1) are

$$\begin{bmatrix} \mathcal{L}_{zz}(z, \lambda) \Delta z + \nabla \phi(z) + \nabla g(z) \lambda^+ \\ g(z) + \nabla g(z)^T \Delta z - \mu(\lambda^+ - \lambda) \end{bmatrix} \in \begin{bmatrix} 0 \\ N(\lambda^+) \end{bmatrix}. \tag{6.2}$$

It is easy to show that for $\delta(z, \lambda)$ sufficiently small, any solution to (6.2) with $\|\Delta z\|$ small must have $\lambda_i^+ = 0$ for $i \notin \mathcal{B}$. For such indices i, we have

$$g_i(z) + \nabla g_i(z)^T \Delta z - \mu(\lambda_i^+ - \lambda_i) \leq g_i(z) + \nabla g_i(z)^T \Delta z + \mu \lambda_i \leq (1/2) g_i(z^*) < 0,$$

when the second inequality holds whenever $\delta(z, \lambda)$ and $\|\Delta z\|$ are sufficiently small. By complementarity, it follows that $\lambda_i^+ = 0$, as claimed. Therefore we can asymptotically drop the inactive constraints from consideration. Denoting by $\tilde{\mathcal{B}} \subseteq \mathcal{B}$ the subset of active indices in (6.2) (so that $\lambda^+_{\mathcal{B} \setminus \tilde{\mathcal{B}}} = 0$), we have by partitioning indices that

$$\begin{bmatrix} \mathcal{L}_{zz}(z, \lambda) \Delta z + \nabla \phi(z) + \nabla g_{\tilde{\mathcal{B}}}(z) \lambda^+_{\tilde{\mathcal{B}}} \\ g_{\tilde{\mathcal{B}}}(z) + \nabla g_{\tilde{\mathcal{B}}}(z)^T \Delta z - \mu(\lambda^+_{\tilde{\mathcal{B}}} - \lambda_{\tilde{\mathcal{B}}}) \\ g_{\mathcal{B} \setminus \tilde{\mathcal{B}}}(z) + \nabla g_{\mathcal{B} \setminus \tilde{\mathcal{B}}}(z)^T \Delta z + \mu \lambda_{\mathcal{B} \setminus \tilde{\mathcal{B}}} \end{bmatrix} \in \begin{bmatrix} 0 \\ N(\lambda^+_{\tilde{\mathcal{B}}}) \\ N(\lambda^+_{\mathcal{B} \setminus \tilde{\mathcal{B}}}) \end{bmatrix}.$$

Since $\mu \lambda_{\mathcal{B} \setminus \tilde{\mathcal{B}}} \geq 0$, the pair $(\Delta z, \lambda^+)$ that solves this system also satisfies

$$\begin{bmatrix} \mathcal{L}_{zz}(z, \lambda) \Delta z + \nabla \phi(z) + \nabla g_{\tilde{\mathcal{B}}}(z) \lambda^+_{\tilde{\mathcal{B}}} \\ g_{\tilde{\mathcal{B}}}(z) + \nabla g_{\tilde{\mathcal{B}}}(z)^T \Delta z - \mu(\lambda^+_{\tilde{\mathcal{B}}} - \lambda_{\tilde{\mathcal{B}}}) \\ g_{\mathcal{B} \setminus \tilde{\mathcal{B}}}(z) + \nabla g_{\mathcal{B} \setminus \tilde{\mathcal{B}}}(z)^T \Delta z \end{bmatrix} \in \begin{bmatrix} 0 \\ N(\lambda^+_{\tilde{\mathcal{B}}}) \\ N(\lambda^+_{\mathcal{B} \setminus \tilde{\mathcal{B}}}) \end{bmatrix}. \tag{6.3}$$

Hence, we can view sSQP as a special case of (4.4) in which we have

$$t = 0, \quad r_{\tilde{\mathcal{B}}} = -\mu(\lambda^+_{\tilde{\mathcal{B}}} - \lambda_{\tilde{\mathcal{B}}}), \quad r_{\{1, \ldots, m\} \setminus \tilde{\mathcal{B}}} = 0. \tag{6.4}$$

There is no circular logic here in the choice of λ^+. If we fix λ^+ at its optimal value from (6.1) and fix t and r in (4.4) at the values in (6.4), then the same $(\Delta z, \lambda^+)$ that solves (6.1) (and (6.3)) will solve (4.4).

A form of the sSQP algorithm was proposed earlier by Bartholomew-Biggs [3]. The basic steps generated by Bartholomew-Biggs's algorithm have the form indicated above (except that a quasi-Newton approximation is used in place of the actual Hessian of the Lagrangian). However, there are numerous modifications that place the

algorithm in the global convergence framework of that author's REQP algorithm [1, 2]. For instance, the multiplier estimates λ are not necessarily updated at each iteration, and successive values of μ are chosen by heuristics rather than by an estimate of the distance to the solution set, as is the case in [18]. The focus of Bartholomew-Biggs's work is somewhat complementary to that of Wright and Hager, since the latter concerns itself with local convergence issues (for the case in which MFCQ is satisfied), while the former focuses on global convergence. The sSQP approach is also closely related to a variant of the method of multipliers/augmented Lagrangian algorithm (see Bertsekas [4] and [5, section 4.2]) in which just one Newton step is applied to the augmented Lagrangian function at each iteration, and in which the parameter μ is decreased to zero. Indeed, such a variant is given for the case of equality constrained problems by Bertsekas [4, p. 240], who points out its superlinear local convergence properties (for the case in which the active constraint gradients are linearly independent).

Superlinear convergence of the sSQP algorithm can be proved if the stabilization parameter μ is related appropriately to the distance $\delta(z, \lambda)$ from (z, λ) to the solution set $\mathcal{S}$. Such an estimate is readily available; we show in the appendix (Theorem A.1) that

$$\eta(z,\lambda) \stackrel{\text{def}}{=} \left\| \begin{bmatrix} \mathcal{L}_z(z,\lambda) \\ \min(\lambda, -g(z)) \end{bmatrix} \right\| = \Theta(\delta(z,\lambda)), \tag{6.5}$$

where the "min" operation applies componentwise to the argument vectors.

Suppose now that we choose μ to satisfy

$$\mu = \eta(z,\lambda)^{\tau}, \tag{6.6}$$

where $\tau \in (0, 1)$. From (6.3), since $\tilde{\mathcal{B}}$ denotes the active constraints in (6.2), we have that

$$\begin{bmatrix} \mathcal{L}_{zz}(z,\lambda) & \nabla g_{\tilde{\mathcal{B}}}(z) \\ \nabla g_{\tilde{\mathcal{B}}}(z)^T & -\mu I \end{bmatrix} \begin{bmatrix} \Delta z \\ \lambda^+_{\tilde{\mathcal{B}}} - \lambda_{\tilde{\mathcal{B}}} \end{bmatrix} = \begin{bmatrix} -\nabla\phi(z) - \nabla g_{\tilde{\mathcal{B}}}(z)\lambda_{\tilde{\mathcal{B}}} \\ -g_{\tilde{\mathcal{B}}}(z) \end{bmatrix}. \tag{6.7}$$

From [11, equation (33)], we have that

$$\|\lambda^+ - \lambda\| = O(\delta(z,\lambda)). \tag{6.8}$$

Therefore if we define (t, r) as in (6.4), we have from (6.5), (6.6), and (6.8) that

$$(t, r) = O(\delta(z,\lambda)^{1+\tau}). \tag{6.9}$$

By substituting (6.8) and (6.9) into (5.23), we obtain

$$\delta(z + \Delta z, \lambda^+) = O(\delta(z,\lambda))O(\delta(z)) + O(\delta(z,\lambda)^{1+\tau}) = O(\delta(z,\lambda)^{1+\tau}).$$

Hence, the convergence rate of sSQP can be derived by placing it in the framework of iSQP.

The same result can be derived from the results of Hager [11]. The following result is a simple consequence of [11, Theorem 1] (restated with a minor correction by Wright [19, Theorem 8]).

THEOREM 6.1. *Suppose that Assumption* 1, *Condition* 2s.3, *and the MFCQ condition hold. Suppose, too, that* μ *defined by* (6.6) *is used as the stabilization parameter at each iteration of sSQP. Then there exists a positive threshold* $\bar{\delta}$ *such that for any*

(z, λ) with $\delta(z, \lambda) \leq \bar{\delta}$ there exists a local solution $(\Delta z, \lambda^+)$ of the sSQP subproblem (6.1) *such that*

$$\delta(z + \Delta z, \lambda^+) = O(\delta(z, \lambda)^{1+\tau}). \tag{6.10}$$

Proof. From Condition 2s.3, we have for each $\lambda^* \in \mathcal{S}_\lambda$ that $w^T \mathcal{L}_{zz}(z^*, \lambda^*) w \geq \sigma \|w\|^2$ for some $\sigma > 0$ and all w with $\nabla g_{\mathcal{B}_+(\lambda^*)}(z^*)^T w = 0$. Moreover, the choice (6.6) ensures that we have

$$\sigma_0 \delta(z, \lambda) \leq \mu \leq \sigma_1,$$

for $\delta(z, \lambda)$ sufficiently small, where σ_0 and σ_1 are constants defined in Theorem 1 of Hager [11], with σ_0 "sufficiently large." By applying Hager's result, we find that there is a neighborhood $\mathcal{U}(\lambda^*)$ of (z^*, λ^*) such that if $(z, \lambda) \in \mathcal{U}(\lambda^*)$, then the sSQP subproblem (6.1) yields a local solution such that (6.10) is satisfied. Note that the set

$$\{\mathcal{U}(\lambda^*) \,|\, \lambda^* \in \mathcal{S}_\lambda\}$$

forms a cover of $\mathcal{S}$. By compactness, we can select a finite subcover

$$\{\mathcal{U}(\lambda_1^*), \mathcal{U}(\lambda_2^*), \ldots, \mathcal{U}(\lambda_p^*)\}$$

for some $\lambda_1^*, \lambda_2^*, \ldots, \lambda_p^* \in \mathcal{S}_\lambda$. By choosing $\bar{\delta}$ positive but small enough that

$$\delta(z, \lambda) \leq \bar{\delta} \;\; \Rightarrow \;\; (z, \lambda) \in \mathcal{U}(\lambda_1^*) \cup \mathcal{U}(\lambda_2^*) \cup \cdots \cup \mathcal{U}(\lambda_p^*),$$

we obtain the desired result. ∎

In Wright [18], it was shown that if the initial estimate λ^0 is not too close to the boundary of $\mathcal{S}_\lambda$ (in the sense that $\lambda_i^0 \geq \xi$ for some $\xi > 0$ and all $i \in \mathcal{B}$), then all steps are obtained from a system of the form (6.7) with $\tilde{\mathcal{B}} = \mathcal{B}$. Moreover, we can set $\tau = 1$ in (6.6) (yielding a quadratic rate of convergence), and we need assume only that the weaker Condition 2s.1 holds. Implementation of such an approach would not be difficult, since it requires only a reliable way to estimate the active set $\mathcal{B}$, along with solution of a subproblem to adjust $\lambda_{\mathcal{B}}$ so that all components of this vector are sufficiently positive.

7. Fischer's method. Fischer's method, as described in the paper [8], generates steps Δz in the primal variables by solving a standard SQP subproblem. The Lagrange multiplier estimate obtained from this subproblem is discarded, and an auxiliary subproblem similar to the SQP subproblem is solved to obtain the multiplier estimate corresponding to the updated value of z. Superlinear convergence of the resulting algorithm is proved in [8], under assumptions that we discuss later.

Fischer's method can be described in terms of the iSQP framework of section 4 as analyzed in section 5. We can show that the primal step Δz generated by this method can be embedded in a primal-dual solution $(\Delta z, \tilde{\lambda}^+)$ to an iSQP subproblem of the form (4.3), so that Theorem 5.3 applies. Superlinear convergence of the primal iterates then follows from the fact that the difference between $\tilde{\lambda}^+$ and Fischer's specific Lagrange multiplier estimate $\hat{\lambda}$ has magnitude $O(\delta(z))$. Superlinear convergence of Fischer's Lagrange multiplier estimates $\hat{\lambda}$ to the set $\mathcal{S}_\lambda$ follows from the fact that $\delta(\hat{\lambda}) = O(\delta(z))$.

A single step of Fischer's algorithm proceeds as follows. Given the primal iterate z, the following subproblem is solved to find the pair $(d, \hat{\lambda})$:

$$\begin{bmatrix} \nabla \phi(z) + d + \nabla g(z) \hat{\lambda} \\ g(z) + \nabla g(z)^T d \end{bmatrix} \in \begin{bmatrix} 0 \\ N(\hat{\lambda}) \end{bmatrix}. \tag{7.1}$$

The primal component d is discarded, and $\hat{\lambda}$ is adopted as the Lagrange multiplier estimate. The next primal step is then obtained by solving the SQP subproblem (4.2) from the point $(z, \hat{\lambda})$, that is,

$$\begin{bmatrix} \mathcal{L}_{zz}(z,\hat{\lambda})\Delta z + \nabla\phi(z) + \nabla g(z)\lambda^+ \\ g(z) + \nabla g(z)^T \Delta z \end{bmatrix} \in \begin{bmatrix} 0 \\ N(\lambda^+) \end{bmatrix}. \tag{7.2}$$

The dual component λ^+ is now discarded (indeed, there is no need to calculate it at all), and this iteration is complete. The next iteration begins by solving (7.1) again, with $z + \Delta z$ replacing z, to obtain the new multiplier estimate $\hat{\lambda}^+$.

Note that in the auxiliary problem (7.1), $(d, \hat{\lambda})$ is the primal-dual solution of the problem

$$\min_d \ \frac{1}{2} d^T d + d^T \nabla\phi(z) \quad \text{subject to } \nabla g(z)^T d + g(z) \le 0$$

(see Fischer [8, p. 13]). More tellingly, we can view (7.1) as a perturbation of the problem

$$\begin{bmatrix} \nabla\phi(z^*) + d + \nabla g(z^*)\hat{\lambda} \\ g(z^*) + \nabla g(z^*)^T d \end{bmatrix} \in \begin{bmatrix} 0 \\ N(\hat{\lambda}) \end{bmatrix}, \tag{7.3}$$

in which z has been replaced by z^*. Noting that the solution set for (7.3) is $(d, \hat{\lambda}) \in 0 \times \mathcal{S}_\lambda$, we can again apply Robinson's results from [17] (and, in particular, [17, Corollary 4.3]) to obtain the estimate

$$\|d\| + \delta(\hat{\lambda}) = O(\delta(z)) \tag{7.4}$$

for all solutions $(d, \hat{\lambda})$ of (7.1).

Fischer [8, Theorem 3.13] shows that under certain assumptions (discussed below), the primal component Δz for the solution of (7.2) is also the solution of the following iSQP subproblem:

$$\begin{bmatrix} \mathcal{L}_{zz}(z,\hat{\lambda})\Delta z + \nabla\phi(z) + \nabla g(z)\tilde{\lambda}^+ + t \\ g(z) + \nabla g(z)^T \Delta z + r \end{bmatrix} \in \begin{bmatrix} 0 \\ N(\tilde{\lambda}^+) \end{bmatrix}, \tag{7.5}$$

where the perturbation vectors t and r and the multiplier estimates $\tilde{\lambda}^+$ satisfy

$$t = 0, \qquad \|r\| = O(\delta(z)^2), \qquad \|\tilde{\lambda}^+ - \hat{\lambda}\| = O(\delta(z)). \tag{7.6}$$

Hence, Theorem 5.3 can be applied to deduce that

$$\delta(z + \Delta z, \tilde{\lambda}^+) = \|\tilde{\lambda}^+ - \hat{\lambda}\| O(\delta(z)) + O(\delta(z)^2) + O(\|(t,r)\|) = O(\delta(z)^2). \tag{7.7}$$

Since $\delta(z + \Delta z) \le \delta(z + \Delta z, \tilde{\lambda}^+)$, this expression implies Q-quadratic convergence in the primal iterates. Q-quadratic convergence of the primal-dual iterates follows from (7.4). Note that the multipliers $\tilde{\lambda}^+$ are never calculated explicitly by the algorithm.

The assumptions needed to prove Theorem 3.13 in [8] include the WCC condition (2.9), the MFCQ condition (2.10), the second-order sufficient Condition 2s.1, and the following constant-rank condition:

$$\nabla g_{\mathcal{B}_+}(z) \text{ has constant rank for all } z \text{ near } z^*. \tag{7.8}$$

8. SQP with strict working sets. Although the preceding two sections show that modified versions of SQP algorithms converge superlinearly on degenerate problems (under certain assumptions), practical experience shows that standard SQP strategies usually encounter little trouble with problems of this type. Frequently, the strict working sets $\mathcal{B}^k$ for the QP subproblems settle down to a constant set in the neighborhood of the solution, and the Lagrange multiplier estimates often approach a unique limit.

As we saw in Theorem 5.3, superlinear convergence depends critically on stabilization of the Lagrange multiplier estimates λ^k. Such stabilization is guaranteed to occur if the strict working sets $\mathcal{B}^k$ eventually become subsets of some fixed set $\tilde{\mathcal{B}} \subseteq \mathcal{B}$ with the following properties:

(i) there is a unique multiplier $\lambda^*(\tilde{\mathcal{B}}) \in \mathcal{S}_\lambda$ such that $\lambda_i^*(\tilde{\mathcal{B}}) = 0$ for $i \notin \tilde{\mathcal{B}}$; and
(ii) the nonlinear program NLP($\tilde{\mathcal{B}}$) obtained by dropping the constraints $i \notin \tilde{\mathcal{B}}$ from (1.1) still has a minimizer at z^* that satisfies Condition 2s.1.

If these properties hold, the only possible limit for the sequence of Lagrange multiplier estimates λ^k is the unique vector $\lambda^*(\tilde{\mathcal{B}})$ defined in (i). Recall from definition (3.5) that $\bar{\Phi}$ contains precisely those subsets of $\mathcal{B}$ with properties (i) and (ii).

The code SNOPT [10] is a recent implementation of SQP that appears to exhibit fast local convergence on most degenerate problems of the type we consider in this paper. Rather than our idea of a strict working set, SNOPT uses the slightly different concept of a *working set* $\mathcal{W}^k$ associated with each iteration k, with the properties that equality holds for the ith linearized constraint if $i \in \mathcal{W}^k$; the multiplier estimates λ_i^k are zero for $i \notin \mathcal{W}^k$; and the gradients $\nabla g_i(x^k)$, $i \in \mathcal{W}^k$, are linearly independent. We conjecture that the good behavior of SNOPT is due to the fact that the working sets $\mathcal{W}^k$ eventually become subsets of a set $\tilde{\mathcal{B}}$ with the properties (i) and (ii) above. Features of SNOPT that promote this behavior, besides the maintenance of linear independence already mentioned, include

(a) the use of warm starts; that is, the working set $\mathcal{W}^k$ from the QP subproblem at iteration k is used as a starting estimate of the working set $\mathcal{W}^{k+1}$ at iteration $k+1$; and
(b) the fact that it allows constraints *not* in the working set ($i \notin \mathcal{W}^k$) to be violated by small tolerances.

Typical behavior of an algorithm with these properties is as follows. Because of linear independence of the gradients $\nabla g_i(x^k)$, $i \in \mathcal{W}^k$, a sufficiently advanced iterate will produce a working set $\mathcal{W}^k$ with the property (i). Iterate $k+1$ then uses $\mathcal{W}^k$ as a starting guess and solves a QP that takes just the constraints $i \in \mathcal{W}^k$ into account. It finds a solution with this working set in which the "ignored" constraints $i \notin \mathcal{W}^k$ are violated by only small amounts, if at all, making this QP solution an acceptable approximation to the true SQP step. It then sets $\mathcal{W}^{k+1} = \mathcal{W}^k$, or possibly drops the indices that have become inactive in the subproblem. The new working set usually retains the property (i), and subsequent iterations will stay with this set or some subset thereof, forcing the Lagrange multipliers to converge to a unique limit.

Unfortunately, a rigorous theoretical result concerning the SQP behavior just discussed does not seem to be possible. Instead, we propose here a formal algorithm called SQPsws (for "SQP with strict working sets") that is motivated by the informal procedure with warm starts and tolerances just described. We show that if, near the solution, a strict working set with the properties (i), (ii) is encountered at some iteration, then Algorithm SQPsws converges superlinearly thereafter, even if the strict working sets at later iterations fail to have one of these properties. While a strict

working set of this type is likely to be identified in most practical situations, we can prove a rigorous convergence result only under conditions similar to those of the preceding two sections.

The key features of Algorithm SQPsws are its use of a *stack* of candidate warm-start strict working sets (instead of just the working set from the previous QP subproblem) and its toleration of a specific level of infeasibility for constraints outside the strict working set, which has the effect of allowing the Lagrange multipliers to stabilize, thereby leading to superlinear convergence. Specifically, the tolerances require that violation of the constraints outside the strict working set be no more than $\eta(z^k, \lambda^k)^{1+\tau}$, where the quantity $\eta(\cdot,\cdot)$ defined in (6.5) measures distance from (z^k, λ^k) to the solution set $\mathcal{S}$, and τ is a parameter in the range $(0, 1)$.

The stack of strict working sets maintained by Algorithm SQPsws has the form

$$\texttt{top} \to \hat{\mathcal{B}}_s \to \hat{\mathcal{B}}_{s-1} \to \cdots \to \hat{\mathcal{B}}_1 \to \hat{\mathcal{B}}_0 = \{1, 2, \ldots, m\},$$

where s is a counter of stack size, the top element $\hat{\mathcal{B}}_s$ is the strict working set $\mathcal{B}^{k-1}$ from the previous iteration, and

$$\hat{\mathcal{B}}_s \subset \hat{\mathcal{B}}_{s-1} \subset \cdots \subset \hat{\mathcal{B}}_1 \subset \hat{\mathcal{B}}_0,$$

where all inclusions are strict. The index sets $\hat{\mathcal{B}}_{s-1}, \ldots, \hat{\mathcal{B}}_1$ are all strict working sets from previous iterations of the algorithm. Elements of the stack are popped and discarded if the solution to the subproblem (8.1) fails to meet the prescribed tolerances for the ignored constraints. As a last resort, if the stack is popped down to its last element $\hat{\mathcal{B}}^0$, the full SQP subproblem (4.1) is solved. In any case, the step produced by each iteration of the algorithm fits the iSQP framework (4.3), so the theory developed in section 4 can be applied.

ALGORITHM SQPSWS.

choose $\tau \in (0, 1)$ and set (z^0, λ^0);
set $k \leftarrow 0$, $s \leftarrow 0$, $\hat{\mathcal{B}}_0 \leftarrow \{1, 2, \ldots, m\}$;
repeat
  set $\mu_k \leftarrow \eta(z^k, \lambda^k)$, `isqpsol` $\leftarrow$ `false`;
  while `not isqpsol`
    $\bar{\mathcal{B}} \leftarrow \hat{\mathcal{B}}_s$;
    solve the SQP subproblem for the constraint set $\bar{\mathcal{B}}$:

(8.1)
$$\min_{\Delta z} \Delta z^T \nabla \phi(z^k) + \tfrac{1}{2} \Delta z^T \mathcal{L}_{zz}(z^k, \lambda^k) \Delta z$$
$$\text{subject to } g_i(z^k) + \nabla g_i(z^k)^T \Delta z \le 0, \text{ all } i \in \bar{\mathcal{B}},$$

    denoting its strict working set by $\mathcal{B}^k$ and its primal-dual solution by $(\Delta z, \tilde{\lambda}_{\bar{\mathcal{B}}})$;
    if $g_i(z^k) + \nabla g_i(z^k)^T \Delta z \le \mu_k^{1+\tau}$ for all $i \notin \bar{\mathcal{B}}$,
      `isqpsol` $\leftarrow$ `true`;
    else
      $s \leftarrow s - 1$;
  end while
  if $\mathcal{B}^k \neq \bar{\mathcal{B}}$
    $s \leftarrow s + 1$, $\hat{\mathcal{B}}_s \leftarrow \mathcal{B}^k$;
  $z^{k+1} \leftarrow z^k + \Delta z$, $\lambda^{k+1} \leftarrow (\tilde{\lambda}_{\bar{\mathcal{B}}}, 0)$;
  $k \leftarrow k + 1$;
until convergence.

Our analysis of Algorithm SQPsws requires a number of technical results, most of which pertain to the adequacy of the chosen subset $\bar{\mathcal{B}}$ of constraints in (8.1) in defining an approximate solution of the full subproblem (4.1), and to the progress that the resulting step makes toward the solution of the original problem (1.1).

We start by formalizing some of the ideas mentioned at the start of this section concerning variants of the nonlinear programming problem (1.1) and the iSQP subproblem (4.3) in which some of the constraints are ignored. In (3.3), we defined the extended nonlinear programming problem NLP($\tilde{\mathcal{B}}$) in which just a subset $\tilde{\mathcal{B}} \subseteq \mathcal{B}$ is enforced. We now define the extended iSQP subproblem corresponding to $\tilde{\mathcal{B}}$ and NLP($\tilde{\mathcal{B}}$) as follows:

$$\begin{aligned} &\min_{\Delta z} \ \Delta z^T(\nabla\phi(z)+t) + \tfrac{1}{2}\Delta z^T \mathcal{L}_{zz}(z,\lambda)\Delta z \\ &\text{subject to } g_i(z) + \nabla g_i(z)^T \Delta z + r_i \le 0, \ \ i \in \tilde{\mathcal{B}}. \end{aligned} \tag{8.2}$$

The KKT conditions for (8.2) are

$$\begin{bmatrix} \mathcal{L}_{zz}(z,\lambda)\Delta z + \nabla\phi(z) + t + \nabla g_{\tilde{\mathcal{B}}}(z)\lambda^+_{\tilde{\mathcal{B}}} \\ g_{\tilde{\mathcal{B}}}(z) + \nabla g_{\tilde{\mathcal{B}}}(z)^T \Delta z + r_{\tilde{\mathcal{B}}} \end{bmatrix} \in \begin{bmatrix} 0 \\ N(\lambda^+_{\tilde{\mathcal{B}}}) \end{bmatrix}. \tag{8.3}$$

Note that (8.2) is truly an iSQP subproblem for (3.3) only if the Lagrangian $\mathcal{L}(z,\lambda)$ does not contain terms in its summation for indices i outside of the set $\tilde{\mathcal{B}}$, that is, only if $\lambda_i = 0$ for all $i \notin \tilde{\mathcal{B}}$. For generality, however, we allow $\lambda_{\{1,\ldots,m\}\setminus\tilde{\mathcal{B}}} \neq 0$ in some of the results below.

Our first technical result is a simple result based on Hoffman's lemma concerning the nearness of a given vector $\lambda \in \mathsf{R}^m$ (with property $\lambda_i = 0$ for all $i \notin \tilde{\mathcal{B}}$) to the unique optimal multiplier $\lambda^*(\tilde{\mathcal{B}})$ of NLP($\tilde{\mathcal{B}}$) for some $\tilde{\mathcal{B}} \in \bar{\Phi}$.

LEMMA 8.1. *There is a constant $\beta \ge 1$ such that the following statement holds. For all $\lambda \in \mathsf{R}^m$ with the property that $\lambda_i = 0$ for all $i \notin \tilde{\mathcal{B}}$, for some $\tilde{\mathcal{B}} \in \bar{\Phi}$, we have*

$$\|\lambda - \lambda^*(\tilde{\mathcal{B}})\| = \|\lambda_{\tilde{\mathcal{B}}} - \lambda^*_{\tilde{\mathcal{B}}}(\tilde{\mathcal{B}})\| \le \beta\delta(\lambda),$$

where, as always, $\delta(\lambda)$ denotes the distance from λ to the optimal multiplier set $\mathcal{S}_\lambda$ of the original problem (1.1).

Proof. Denoting by $P(\lambda)$ the closest vector in $\mathcal{S}_\lambda$ to λ, we have that

$$\delta(\lambda)^2 = \|\lambda - P(\lambda)\|^2 = \sum_{i\in\tilde{\mathcal{B}}} [\lambda_i - P(\lambda)_i]^2 + \sum_{i\in\mathcal{B}\setminus\tilde{\mathcal{B}}} P(\lambda)_i^2,$$

implying that

$$\left\| P(\lambda)_{\mathcal{B}\setminus\tilde{\mathcal{B}}} \right\| \le \delta(\lambda). \tag{8.4}$$

Note that $P(\lambda)$ satisfies the following system of linear equalities and inequalities:

$$\sum_{i\in\mathcal{B}} \nabla g_i(z^*) P(\lambda)_i = \nabla\phi(z^*), \qquad P(\lambda) \ge 0, \qquad P(\lambda)_i = 0, \ i \notin \mathcal{B}. \tag{8.5}$$

The following system, on the other hand, has the unique solution $\bar{\lambda} = \lambda^*(\tilde{\mathcal{B}})$:

$$\sum_{i\in\mathcal{B}} \nabla g_i(z^*) \bar{\lambda}_i = \nabla\phi(z^*), \qquad \bar{\lambda} \ge 0, \qquad \bar{\lambda}_i = 0, \ i \notin \tilde{\mathcal{B}}. \tag{8.6}$$

Because of (8.5), we have that $P(\lambda)$ violates (8.6) only in that possibly $P(\lambda)_i > 0$ for some $i \in \mathcal{B}\backslash\tilde{\mathcal{B}}$. Hence, by Hoffman's lemma [13] and the uniqueness of $\lambda^*(\tilde{\mathcal{B}})$, there is a positive quantity $\beta(\tilde{\mathcal{B}})$ such that

$$\|\lambda^*(\tilde{\mathcal{B}}) - P(\lambda)\| \leq \beta(\tilde{\mathcal{B}}) \left\| P(\lambda)_{\mathcal{B}\backslash\tilde{\mathcal{B}}} \right\|.$$

By choosing

$$\beta = \max_{\tilde{\mathcal{B}} \in \bar{\Phi}} \beta(\tilde{\mathcal{B}}) + 1,$$

combining this expression with (8.4), and using $\|P(\lambda) - \lambda\| = \delta(\lambda)$, we have that

$$\|\lambda - \lambda^*(\tilde{\mathcal{B}})\| \leq \|\lambda - P(\lambda)\| + \|P(\lambda) - \lambda^*(\tilde{\mathcal{B}})\| \leq \beta\delta(\lambda),$$

giving the result. □

We now modify two of the results of section 5 to apply to those extended problems (8.2) for which z^* satisfies Condition 2s.1, that is, $\tilde{\mathcal{B}} \in \bar{\Phi}$. The combination of these two results—Lemma 5.1 and Theorem 5.3—with Lemma 8.1 yields some powerful estimates.

LEMMA 8.2. *Suppose that Assumption* 1, *Condition* 2s.1, *and MFCQ hold. Then there exists a threshold value* $\bar{\delta} > 0$ *with the following property. If* $\delta(z, \lambda) \leq \bar{\delta}$ *and* $\|(t, r_{\tilde{\mathcal{B}}})\| \leq \bar{\delta}$ *and* $\lambda_i = 0$ *for all* $i \notin \tilde{\mathcal{B}}$, *for some* $\tilde{\mathcal{B}} \in \bar{\Phi}$, *then the extended iSQP subproblem* (8.2) *has at least one solution* $(\Delta z, \lambda^+)$, *and for all such solutions we have*

$$\|\Delta z\| + \|\lambda^+_{\tilde{\mathcal{B}}} - \lambda^*_{\tilde{\mathcal{B}}}(\tilde{\mathcal{B}})\| = O(\|z - z^*\|) + O(\|(t, r_{\tilde{\mathcal{B}}})\|), \tag{8.7}$$

where $\lambda^*_{\tilde{\mathcal{B}}}(\tilde{\mathcal{B}})$ *is the (unique) optimal Lagrange multiplier for* NLP$(\tilde{\mathcal{B}})$ (3.3). *Moreover, we have that*

$$\begin{aligned} \|z + \Delta z - z^*\| + \|\lambda^+_{\tilde{\mathcal{B}}} - \lambda^*_{\tilde{\mathcal{B}}}(\tilde{\mathcal{B}})\| &\leq \|\lambda_{\tilde{\mathcal{B}}} - \lambda^*_{\tilde{\mathcal{B}}}(\tilde{\mathcal{B}})\| O(\delta(z)) + O(\delta(z)^2) + O(\|(t, r_{\tilde{\mathcal{B}}})\|) \\ &= O(\delta(\lambda)\delta(z)) + O(\delta(z)^2) + O(\|(t, r_{\tilde{\mathcal{B}}})\|) \\ &= O(\delta(z, \lambda)^2) + O(\|(t, r_{\tilde{\mathcal{B}}})\|). \end{aligned} \tag{8.8}$$

Proof. For a given $\tilde{\mathcal{B}} \in \bar{\Phi}$, we obtain by applying Lemma 5.1 to (8.3) that there is a threshold $\bar{\delta}(\tilde{\mathcal{B}})$ such that (8.7) holds whenever

$$\left\| \begin{bmatrix} z - z^* \\ \lambda_{\tilde{\mathcal{B}}} - \lambda^*_{\tilde{\mathcal{B}}}(\tilde{\mathcal{B}}) \end{bmatrix} \right\| \leq \bar{\delta}(\tilde{\mathcal{B}}), \quad \|(t, r_{\tilde{\mathcal{B}}})\| \leq \bar{\delta}(\tilde{\mathcal{B}}). \tag{8.9}$$

(Note that we can write the distance of $\lambda^+_{\tilde{\mathcal{B}}}$ to the dual solution set for (3.3) explicitly as $\|\lambda^+_{\tilde{\mathcal{B}}} - \lambda^*_{\tilde{\mathcal{B}}}(\tilde{\mathcal{B}})\|$, since this set contains just the single element $\lambda^*_{\tilde{\mathcal{B}}}(\tilde{\mathcal{B}})$.) By Lemma 8.1, we have for all (z, λ) with $\lambda_i = 0$, $i \notin \tilde{\mathcal{B}}$, that

$$\left\| \begin{bmatrix} z - z^* \\ \lambda_{\tilde{\mathcal{B}}} - \lambda^*_{\tilde{\mathcal{B}}}(\tilde{\mathcal{B}}) \end{bmatrix} \right\| \leq \|z - z^*\| + \beta\delta(\lambda) \leq 2\beta\delta(z, \lambda),$$

where the last inequality follows from $\beta \geq 1$. It follows that both bounds in (8.9) are satisfied if we set

$$\bar{\delta} = \frac{1}{2\beta} \min_{\tilde{\mathcal{B}} \in \bar{\Phi}} \bar{\delta}(\tilde{\mathcal{B}}).$$

For (8.8), we obtain by applying Theorem 5.3 to the extended iSQP problem (8.2) that

$$\|z + \Delta z - z^*\| + \|\lambda^+_{\tilde{\mathcal{B}}} - \lambda^*_{\tilde{\mathcal{B}}}(\tilde{\mathcal{B}})\| \leq \|\lambda^+_{\tilde{\mathcal{B}}} - \lambda_{\tilde{\mathcal{B}}}\| O(\delta(z)) + O(\delta(z)^2) + O(\|(t, r_{\tilde{\mathcal{B}}})\|).$$

By applying the triangle inequality to the term $\|\lambda^+_{\tilde{\mathcal{B}}} - \lambda_{\tilde{\mathcal{B}}}\|$, reducing $\bar{\delta}$ if necessary to ensure that the $O(\delta(z))$ term is smaller than $1/2$, and rearranging, we obtain that

$$\|z+\Delta z-z^*\|+(1/2)\|\lambda^+_{\tilde{\mathcal{B}}} - \lambda^*_{\tilde{\mathcal{B}}}(\tilde{\mathcal{B}})\| \leq \|\lambda_{\tilde{\mathcal{B}}} - \lambda^*_{\tilde{\mathcal{B}}}(\tilde{\mathcal{B}})\| O(\delta(z)) + O(\delta(z)^2) + O(\|(t, r_{\tilde{\mathcal{B}}})\|),$$

yielding the first inequality in (8.8). The second relation follows immediately from Lemma 8.1, while the third follows from (2.12). ∎

We now are in a position to prove our main convergence result for Algorithm SQPsws. We show that if a strict working set $\mathcal{B}^k$ from $\bar{\Phi}$ enters the stack at some sufficiently advanced iterate, then it remains in the stack and the algorithm converges superlinearly.

THEOREM 8.3. *Suppose that Assumption* 1*, Condition* 2s.1*, and MFCQ hold. Then there exists a positive threshold $\hat{\delta}$ such that if at some iteration $\bar{k}$ we have $\delta(z^{\bar{k}}, \lambda^{\bar{k}}) \leq \hat{\delta}$, and if there is an index set $\tilde{\mathcal{B}} \in \bar{\Phi}$ present in the stack at the start of iteration $\bar{k}$, then $\tilde{\mathcal{B}}$ remains in the stack at all subsequent iterations, and Algorithm* SQPsws *converges superlinearly with Q-order $1 + \tau$.*

Proof. Consider all subproblems (8.1) arising for Algorithm SQPsws in which the conditions $g_i(z^k) + \nabla g_i(z^k)^T \Delta z \leq \mu_k^{1+\tau}$ hold for all $i \notin \bar{\mathcal{B}}$. These subproblems have the form of (4.3), where $t = 0$ and r is a vector whose elements do not exceed $\eta(z^k, \lambda^k)^{1+\tau}$ in magnitude. Because of (6.5), we can choose $\hat{\delta}$ small enough that $\delta(z^k, \lambda^k) \leq \hat{\delta}$ implies that

$$\delta(z^k, \lambda^k) \leq \bar{\delta}, \quad \|(0, r)\| \leq \bar{\delta},$$

where $\bar{\delta}$ is the threshold value such that the assumptions of Lemma 5.1 and Theorems 5.2 and 5.3 are satisfied when $\delta(z, \lambda) \leq \bar{\delta}$ and $\|(t, r)\| \leq \bar{\delta}$. We reduce $\hat{\delta}$ if necessary so that the conditions of Lemma 8.2 are satisfied by (z, λ), $t = 0$, and $r_{\mathcal{B}} = O(\eta(z, \lambda)^{1+\tau})$ whenever $\delta(z^k, \lambda^k) \leq \hat{\delta}$. Further assumptions on the size of $\hat{\delta}$ are made in the course of the proof.

The main part of our proof is to show that, at any iterate k for which

(8.10a) $\tilde{\mathcal{B}}$ is present in the stack at the start of iteration k and

(8.10b) $$\delta(z^k, \lambda^k) \leq \hat{\delta},$$

we have that

(8.11a) $\tilde{\mathcal{B}}$ is present in the stack at the end of iteration k,

(8.11b) $$\lambda^{k+1}_{\{1,\ldots,m\}\setminus\tilde{\mathcal{B}}} = 0, \text{ and}$$

(8.11c) $$\delta(z^{k+1}, \lambda^{k+1}) = O(\delta(z^k, \lambda^k)^{1+\tau}) \leq \delta(z^k, \lambda^k).$$

By definition, the premise (8.10) is satisfied at iteration $\bar{k}$, and so from (8.11a) and (8.11c) it holds for all subsequent iterations. Hence, (8.11c) implies superlinear convergence.

Suppose that (8.10) holds for some k. Since $\tilde{\mathcal{B}}$ is present in the stack at the start of this iteration, the active set $\mathcal{B}^{k-1}$ from the subproblem (8.1) at the previous iteration must be such that $\mathcal{B}^{k-1} \subseteq \tilde{\mathcal{B}}$. In particular, we have that $\lambda^k_i = 0$ for all $i \notin \tilde{\mathcal{B}}$.

At the end of iteration k, the set $\tilde{\mathcal{B}}$ can have disappeared from the stack only if it was tried and rejected in (8.1), that is, if the solution to (8.1) obtained with $\bar{\mathcal{B}} = \tilde{\mathcal{B}}$ had

$$g_i(z^k) + \nabla g_i(z^k)^T \Delta z^k > \mu_k^{1+\tau} \qquad \text{for some } i \notin \tilde{\mathcal{B}}. \tag{8.12}$$

Because of our choice of $\hat{\delta}$, we can apply Lemma 8.2 to (8.1) by setting $\bar{\mathcal{B}} = \tilde{\mathcal{B}}$, $(z, \lambda) = (z^k, \lambda^k)$, and $(t, r_{\tilde{\mathcal{B}}}) = 0$. We obtain from (8.8) that

$$\|z^k + \Delta z^k - z^*\| = O(\delta(z^k, \lambda^k)^2),$$

while from (8.7), we have

$$\|\Delta z^k\| = O(\delta(z^k)) = O(\delta(z^k, \lambda^k)). \tag{8.13}$$

Hence, because of $g_{\mathcal{B}}(z^*) = 0$, we have that

$$\begin{aligned} g_i(z^k) + \nabla g_i(z^k)^T \Delta z^k &= g_i(z^k + \Delta z^k) + O(\|\Delta z^k\|^2) \\ &= O(\|z^k + \Delta z^k - z^*\|) + O(\|\Delta z^k\|^2) \\ &= O(\delta(z^k, \lambda^k)^2) \qquad \text{for all } i \in \mathcal{B}. \end{aligned}$$

Since from (6.5) we have that $\mu_k = \Theta(\delta(z^k, \lambda^k))$, the condition (8.12) cannot hold for any $i \in \mathcal{B}$. Neither can it hold for any constraint for the NLP, because by choosing $\hat{\delta}$ small enough, we have from (8.13) that

$$g_i(z^k) + \nabla g_i(z^k)^T \Delta z^k \le (1/2) g_i(z^*) < 0 \qquad \text{for all } i \notin \mathcal{B}.$$

Hence, the violation (8.12) does not occur, so the set $\tilde{\mathcal{B}}$ will not be popped from the stack.

Since $\tilde{\mathcal{B}}$ remains in the stack, we must have $\mathcal{B}^k \subseteq \tilde{\mathcal{B}}$, so that (8.11b) holds.

Since $\mathcal{B}^k \subseteq \tilde{\mathcal{B}}$, we have that the primal-dual solution of (8.1) with $\bar{\mathcal{B}} = \mathcal{B}^k$ is an approximate solution to the extended iSQP subproblem (8.2) with

$$t = 0, \qquad r_{\mathcal{B}^k} = 0, \qquad 0 \le -r_i \le \mu_k^{1+\tau} \quad \text{for } i \in \tilde{\mathcal{B}} \backslash \mathcal{B}^k.$$

Because of our assumptions on $\hat{\delta}$, Lemma 8.2 applies to this situation, and we obtain from (8.8) and the estimates above that

$$\|z^k + \Delta z^k - z^*\| + \|\lambda_{\tilde{\mathcal{B}}}^{k+1} - \lambda_{\tilde{\mathcal{B}}}^*(\tilde{\mathcal{B}})\| = O(\mu_k^{1+\tau}) = O(\delta(z^k, \lambda^k)^{1+\tau}).$$

Since

$$\delta(z^k + \Delta z^k, \lambda^{k+1}) \le \|z^k + \Delta z^k - z^*\| + \|\lambda_{\tilde{\mathcal{B}}}^{k+1} - \lambda_{\tilde{\mathcal{B}}}^*(\tilde{\mathcal{B}})\|,$$

the first relation in (8.11c) follows. The second relation in (8.11c) follows from a choice of sufficiently small $\hat{\delta}$. □

It seems reasonable to expect a set from $\bar{\Phi}$ to enter the stack at some sufficiently advanced iteration in most nonpathological cases. The strict working set $\mathcal{B}^k$ (for (z^k, λ^k) close to $\mathcal{S}$) is likely to belong at least to Φ by the following argument: The solution of (8.1) satisfies at least the second-order necessary conditions for the quadratic subproblem in which just the constraints $i \in \mathcal{B}^k$ are enforced. Since we know

from Lemma 4.1 that there exists at least one $\lambda^* \in \mathcal{S}_\lambda$ with $\mathcal{B}_+(\lambda^*) \subseteq \mathcal{B}^k$, we can reasonably expect that at least second-order necessary conditions are satisfied at z^* for NLP($\mathcal{B}^k$) as well. Hence, we would have $\mathcal{B}^k \notin \Phi$ only if the second-order conditions for NLP($\mathcal{B}^k$) are necessary but not sufficient—an uncommon scenario. Moreover, we can expect $\mathcal{B}^k$ to belong to the more restricted set $\bar{\Phi}$ because, as mentioned above, active-set solvers for the quadratic subproblem typically ensure that the working constraint Jacobian $\nabla g_{\mathcal{B}^k}(z^k)$ has full rank. (This property can be assured in any case by the simple procedure below.) In fact, our toleration of small violations in the nonenforced constraints $i \notin \mathcal{B}^k$ tends to discourage even nearly dependent active constraint sets. Hence, the optimal Lagrange multiplier vector corresponding to $\mathcal{B}^k$ will be unique unless $\nabla g_{\mathcal{B}^k}$ loses rank between z^k and z^*—another uncommon occurrence.

The preceding paragraph suggests that we can prove that the conditions of Theorem 8.3 are satisfied if we make a few additional assumptions. One such assumption—full rank of $\nabla g_{\mathcal{B}^k}(z^k)$—can be guaranteed by applying a procedure based on the following observations to remove some indices from $\mathcal{B}^k$ if necessary. Any solution $(\Delta z^k, \lambda^{k+1})$ of (8.1) satisfies the system

$$\sum_{i \in \mathcal{B}^k} \nabla g_i(z^k) \lambda_i^{k+1} = -\nabla \phi(z^k) - \mathcal{L}_{zz}(z^k, \lambda^k) \Delta z^k. \tag{8.14}$$

If $\nabla g_{\mathcal{B}^k}(z^k)$ does not have full rank, there is a vector $\Delta\lambda \neq 0$ such that

$$\sum_{i \in \mathcal{B}^k} \nabla g_i(z^k) \Delta\lambda_i = 0.$$

Because of the MFCQ condition (2.10), we have for z^k sufficiently close to z^* that at least one component of $\Delta\lambda$ is negative. Since by (8.14) the vector $(\Delta z^k, \lambda^{k+1} + \alpha\Delta\lambda)$ is a primal-dual solution of (8.1) for all α such that $\lambda^{k+1} + \alpha\Delta\lambda \geq 0$, we can choose α to reduce at least one component of $\lambda^{k+1} + \alpha\Delta\lambda$ to zero. By applying this procedure repeatedly as needed, we can arrive at a revised strict working set $\mathcal{B}^k$ with the desired property. In fact, by allowing a small violation of the equality in (8.14)—a violation $t = O(\mu_k^{1+\tau})$ that stays within the iSQP framework (4.4) and hence retains the stated convergence rate—we can remove even nearly dependent constraints from the active set $\mathcal{B}^k$, thus increasing the likelihood that $\mathcal{B}^k$ belongs to $\bar{\Phi}$.

COROLLARY 8.4. *Suppose that Assumption* 1*, Condition* 2s.3*, and MFCQ are satisfied and that the constant-rank constraint qualification condition of Janin* [14] *holds; that is, there is an open neighborhood of z^* such that for any subset $\hat{\mathcal{B}}$ of $\mathcal{B}$, the matrix $\nabla g_{\hat{\mathcal{B}}}(z)$ has constant rank for all z in this neighborhood. Then there exists a positive threshold $\bar{\delta}$ such that if $\delta(z^k, \lambda^k) \leq \bar{\delta}$ for some k, Algorithm* SQPsws*, with the modification above to ensure full rank of $\nabla g_{\mathcal{B}^k}(z^k)$, converges superlinearly.*

Proof. We prove the result by showing that if $\delta(z^k, \lambda^k) \leq \bar{\delta}$ for some point (z^k, λ^k), then solution of an iSQP subproblem at this point will yield an active set $\mathcal{B}^k$ for which $\mathcal{B}^k \in \bar{\Phi}$. Since we know that *every* iteration of Algorithm SQPsws takes a step that fits the iSQP framework (4.3), it follows that $\mathcal{B}^k$ will appear at the top of the stack at the end of iteration k. Hence, superlinear convergence follows from Theorem 8.3.

From Lemma 4.1, we have that the strict working set $\mathcal{B}^k$ generated by iteration k is such that there exists at least one $\lambda^* \in \mathcal{S}_\lambda$ with $\mathcal{B}_+(\lambda^*) \subseteq \mathcal{B}^k$. Hence, by Lemma 3.3, we have that $\mathcal{B}^k \in \Phi$. Since by our discussion above, the active constraint Jacobian $\nabla g_{\mathcal{B}^k}(z^k)$ has full rank, and since the constant rank condition holds, we have that $\nabla g_{\mathcal{B}^k}(z^*)$ has full rank also. Therefore by Lemma 3.2, we have $\mathcal{B}^k \in \bar{\Phi}$. ∎

Note that the constant-rank condition assumed here is stronger than the corresponding condition (7.8) used by Fischer [8].

Appendix A. Estimating the distance to the optimal set. An estimate of the distance from the current point (z, λ) to the primal-dual optimal set $\mathcal{S}$ is a critical ingredient in the modifications to the SQP algorithm discussed above. We show here that the easily computed quantity $\eta(z, \lambda)$ (6.5) is a satisfactory estimate in a neighborhood of $\mathcal{S}$.

THEOREM A.1. *Suppose that Assumption* 1*, Condition* 2s.1*, and the MFCQ condition are satisfied. Then if* $\lambda \geq 0$*, we have that*

$$\eta(z,\lambda) \stackrel{\text{def}}{=} \left\| \begin{bmatrix} \mathcal{L}_z(z,\lambda) \\ \min(\lambda, -g(z)) \end{bmatrix} \right\| = \Theta(\delta(z,\lambda)).$$

Proof. We start with the easy part of the proof, which is to show that $\eta(z, \lambda) = O(\delta(z, \lambda))$. By the assumed smoothness of ϕ and g, we have

$$\mathcal{L}_z(z,\lambda) = \mathcal{L}_z(z,\lambda) - \mathcal{L}_z(z^*, P(\lambda)) = O(\delta(z,\lambda)). \tag{A.1}$$

For $\delta(z, \lambda)$ sufficiently small, we have $0 \leq \lambda_i < -g_i(z)$ for all $i \notin \mathcal{B}$, and therefore

$$i \notin \mathcal{B} \;\Rightarrow\; 0 \leq \min(\lambda_i, -g_i(z)) = \lambda_i = |\lambda_i - P(\lambda)_i| \leq \delta(\lambda). \tag{A.2}$$

For the active indices $i \in \mathcal{B}$, we have from $\lambda_i \geq 0$ that

$$i \in \mathcal{B} \;\Rightarrow\; |\min(\lambda_i, -g_i(z))| \leq |g_i(z)| \leq |g_i(z) - g_i(z^*)| = O(\delta(z)). \tag{A.3}$$

The result $\eta(z, \lambda) = O(\delta(z, \lambda))$ follows from the estimates (A.1), (A.2), and (A.3).

The more difficult part of the proof is to show that $\delta(z, \lambda) = O(\eta(z, \lambda))$. Our main theoretical tool is again Theorem 4.2 of Robinson [17].

We first define the vectors $v \in \mathsf{R}^m$ and $\omega \in \mathsf{R}^m$ as follows:

$$v_i = \begin{cases} -g_i(z) & \text{if } -g_i(z) < \lambda_i, \\ 0 & \text{otherwise,} \end{cases} \qquad i = 1, 2, \ldots, m, \tag{A.4}$$

$$\omega_i = \begin{cases} \lambda_i & \text{if } -g_i(z) \geq \lambda_i, \\ 0 & \text{otherwise,} \end{cases} \qquad i = 1, 2, \ldots, m. \tag{A.5}$$

For each $i = 1, 2, \ldots, m$, we have that either $g_i(z) + v_i = 0$ or $\lambda_i - \omega_i = 0$. We have also that $g(z) + v \leq 0$ and $\lambda - \omega \geq 0$, and therefore, by definition of $N(\cdot)$ in (2.5), we have that

$$g(z) + v \in N(\lambda - \omega). \tag{A.6}$$

Note too that v and ω are complementary; that is,

$$v \geq 0, \quad \omega \geq 0, \quad v^T\omega = 0. \tag{A.7}$$

In fact, we have that

$$v + \omega = \min(-g(z), \lambda),$$

and so from (A.7), (A.2), and (A.3) we obtain that

$$\|v\|^2 + \|\omega\|^2 = \|v + \omega\|^2 = \|\min(-g(z), \lambda)\|^2 = O(\delta(z,\lambda)^2). \tag{A.8}$$

We therefore have the following estimates for v and w:

$$\|v\| \leq O(\delta(z,\lambda)), \qquad \|\omega\| \leq O(\delta(z,\lambda)). \tag{A.9}$$

We now define perturbed variants of the objective function ϕ and constraint function g as follows:

$$\hat{\phi}(\hat{z};\hat{u},\hat{v}) \stackrel{\text{def}}{=} \phi(\hat{z}) - \hat{z}^T\hat{u},$$
$$\hat{g}(\hat{z};\hat{u},\hat{v}) \stackrel{\text{def}}{=} g(\hat{z}) + \hat{v},$$

where $(\hat{u},\hat{v})$ is the perturbation vector. Note that $\hat{\phi}(\cdot;0,0) = \phi(\cdot)$ and $\hat{g}(\cdot;0,0) = g(\cdot)$. It is not difficult to show, with the help of (A.6), that $(\hat{z},\hat{\lambda}) = (z,\lambda-\omega)$ is a primal-dual solution of the following perturbed version of (1.1):

$$\min_{\hat{z}} \hat{\phi}(\hat{z};\hat{u},\hat{v}) \qquad \text{subject to } \hat{g}(\hat{z};\hat{u},\hat{v}) \leq 0,$$

where

$$\hat{u} = \mathcal{L}_z(z,\lambda-\omega), \qquad \hat{v} = v.$$

Both perturbations are small. For $\hat{u}$, we have from (A.1) and (A.9) that

$$\begin{aligned} \|\hat{u}\| &= \|\mathcal{L}_z(z,\lambda-\omega)\| \\ &\leq \|\mathcal{L}_z(z,\lambda)\| + \left\|\sum_{i=1}^m \omega_i \nabla g_i(z)\right\| \leq O(\delta(z,\lambda)) + O(\|\omega\|) = O(\delta(z,\lambda)), \end{aligned} \tag{A.10}$$

while for $\hat{v}$, we have immediately from (A.9) that $\|\hat{v}\| = \|v\| = O(\delta(z,\lambda))$. Hence, $(z,\lambda-\omega)$ is the solution of a slightly perturbed nonlinear program, where the size of the perturbation is uniformly small for (z,λ) near $\mathcal{S}$, so we can apply Theorem 4.2 of Robinson [17].

By the first inequality in the cited theorem, we have that

$$\begin{aligned} \delta(z,\lambda-\omega) &= O\left(\operatorname{dist}\left(0, \begin{bmatrix} \mathcal{L}_z(z,\lambda-\omega) \\ g(z) \end{bmatrix} - \begin{bmatrix} 0 \\ N(\lambda-\omega) \end{bmatrix}\right)\right) \\ &= O\left(\|\mathcal{L}_z(z,\lambda-\omega)\| + \operatorname{dist}(g(z), N(\lambda-\omega))\right). \end{aligned} \tag{A.11}$$

For the first term, we have as in (A.10) that

$$\|\mathcal{L}_z(z,\lambda-\omega)\| = O(\|\mathcal{L}_z(z,\lambda)\|) + O(\|\omega\|).$$

For the second term, we have by application of the triangle inequality and (A.6) that

$$\operatorname{dist}(g(z), N(\lambda-\omega)) \leq \|v\| + \operatorname{dist}(g(z)+v, N(\lambda-\omega)) = \|v\|.$$

By substituting these estimates into (A.11), we obtain

$$\delta(z,\lambda-\omega) = O\left(\|\mathcal{L}_z(z,\lambda)\| + \|v\| + \|\omega\|\right). \tag{A.12}$$

By applying the triangle inequality again, we obtain from (A.8) and (A.12) that

$$\begin{aligned} \delta(z,\lambda) &\leq \delta(z,\lambda-\omega) + \|\omega\| \\ &= O\left(\|\mathcal{L}_z(z,\lambda)\| + \|v\| + \|\omega\|\right) \\ &= O\left(\|\mathcal{L}_z(z,\lambda)\| + \|\min(-g(z),\lambda)\|\right) \\ &= O(\eta(z,\lambda)), \end{aligned}$$

as required. □

The estimate (6.5) was proposed by several other authors independently of this paper. Facchinei, Fischer, and Kanzow propose the same estimate in a revised version of their paper [7]. Hager and Gowda [12, Theorems 1 and 3] propose a more general measure, which reduces to (6.5) when $\lambda \geq 0$ and does not require the MFCQ condition to hold.

Appendix B. Perturbation analysis of a convex program. We consider the following convex quadratic program:

$$\min_x \tfrac{1}{2} x^T Q x + c^T x \quad \text{subject to } Ax = b,\ Cx \leq d, \tag{B.1}$$

where Q is symmetric positive definite. Suppose the constraints satisfy the following property:

$$Aw = 0, \quad Cw < 0 \quad \text{for some vector } w. \tag{B.2}$$

If in addition we were to assume that the rows of A were linearly independent, these constraints would satisfy the MFCQ. We have the following result.

LEMMA B.1. *Consider the problem* (B.1), *where we take Q, A, and C to be fixed, while c, b, and d are allowed to vary. Assume that* (B.2) *holds. Then*

(i) (B.1) *has at most one solution $x(c, b, d)$ for any vector triple (c, b, d), and if in addition the rows of A are linearly independent, it has exactly one solution;*

(ii) *if the solution exists for two vector triples (c, b, d) and (c', b', d'), the following Lipschitz continuity property is satisfied:*

$$\|x(c, b, d) - x(c', b', d')\| \leq L \|(c, b, d) - (c', b', d')\|, \tag{B.3}$$

where the constant L depends only on Q, A, and C.

Proof. Because the objective function is strictly convex and the feasible region is convex polyhedral, a unique solution will exist whenever the feasible region is nonempty. When the MFCQ is satisfied, the feasible set is in fact nonempty for all b and d. Therefore, (i) is true.

The proof of (ii) is similar to that of Proposition 7.5.9 and Corollary 7.5.10 of Cottle, Pang, and Stone [6], so we omit the details. □

Acknowledgments. I thank Michael Wagner, Mihai Anitescu, Alex Shapiro, Philip Gill, and Andreas Fischer for many interesting discussions on this topic. Thanks also to Michael for carrying out numerical experiments during his summer at Argonne in 1997. I am also grateful to two referees for their comments on the first version and their pointers to related literature, which improved this version of the paper considerably.

REFERENCES

[1] M. C. BARTHOLOMEW-BIGGS, *Recursive quadratic programming methods for nonlinear constraints*, in Nonlinear Optimization 1981, M. J. D. Powell, ed., Academic Press, London, 1982, pp. 213–221.

[2] M. C. BARTHOLOMEW-BIGGS, *A Globally Convergent Version of REQP for Constrained Minimization*, Technical Report 168, Hatfield Polytechnic, Hatfield, UK, 1986.

[3] M. C. BARTHOLOMEW-BIGGS, *Recursive quadratic programming methods based on the augmented Lagrangian*, Math. Programming Stud., 31 (1987), pp. 21–41.

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

[5] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed., Athena Scientific, Belmont, MA, 1999.
[6] R. W. Cottle, J.-S. Pang, and R. E. Stone, *The Linear Complementarity Problem*, Academic Press, San Diego, 1992.
[7] F. Facchinei, A. Fischer, and C. Kanzow, *On the accurate identification of active constraints*, SIAM J. Optim., 9 (1998), pp. 14–32.
[8] A. Fischer, *Modified Wilson method for nonlinear programs with nonunique multipliers*, Math. Oper. Res., 24 (1999), pp. 699–727.
[9] J. Gauvin, *A necessary and sufficient regularity condition to have bounded multipliers in nonconvex programming*, Math. Program., 12 (1977), pp. 136–138.
[10] P. E. Gill, W. Murray, and M. Saunders, *SNOPT: An SQP Algorithm for Large-Scale Constrained Optimization*, Report NA 97-2, Department of Mathematics, University of California, San Diego, 1997.
[11] W. W. Hager, *Stabilized sequential quadratic programming*, Comput. Optim. Appl., 12 (1999), pp. 253–273.
[12] W. W. Hager and M. S. Gowda, *Stability in the presence of degeneracy and error estimation*, Math. Program. Ser. A, 85 (1999), pp. 181–192.
[13] A. J. Hoffman, *On approximate solutions of systems of linear inequalities*, J. Res. Nat. Bur. Standards, 49 (1952), pp. 263–265.
[14] R. Janin, *Directional derivative of the marginal function in nonlinear programming*, Math. Programming Stud., (1984), pp. 110–126.
[15] 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.
[16] D. Ralph and S. J. Wright, *Superlinear convergence of an interior-point method despite dependent constraints*, Math. Oper. Res., 25 (2000), pp. 179–194.
[17] S. M. Robinson, *Generalized equations and their solutions. Part* II*: Applications to nonlinear programming*, Math. Programming Stud., 19 (1982), pp. 200–221.
[18] S. J. Wright, *Superlinear convergence of a stabilized SQP method to a degenerate solution*, Comput. Optim. Appl., 11 (1998), pp. 253–275.
[19] S. J. Wright *Constraint identification and algorithm stabilization for degenerate nonlinear programs*, Math. Program., to appear.
[20] S. J. Wright and D. Ralph, *A superlinear infeasible-interior-point algorithm for monotone nonlinear complementarity problems*, Math. Oper. Res., 21 (1996), pp. 815–838.

SIAM J. OPTIM.
Vol. 13, No. 2, pp. 498–507

THE STEINER TRAVELING SALESMAN POLYTOPE AND RELATED POLYHEDRA*

MOURAD BAÏOU† AND ALI RIDHA MAHJOUB‡

Abstract. In this paper we consider an extended formulation of the Steiner traveling salesman problem, that is, when variables are associated with both the edges and the nodes of the graph. We give a complete linear description of the associated polytope when the underlying graph is series-parallel. By projecting this polytope onto the edge variables, we obtain a characterization of the Steiner traveling salesman polytope in the same class of graphs. Both descriptions yield polynomial time (cutting plane) algorithms for the corresponding problems in that class of graphs.

Key words. Steiner traveling salesman problem, polyhedral combinatorics

AMS subject classifications. 05C85, 90C27

PII. S1052623400322287

1. Introduction.

A cycle of a graph G is called *simple* if no node is incident to more than two of its edges. Given a graph $G = (V, E)$, a weight vector $w \in \mathbb{R}^{|E|}$ associated with the edges of G, and a subset of distinguished nodes $T \subseteq V$, called *terminals*, the *Steiner traveling salesman problem* (StSP) is the problem of finding a minimum weight simple cycle of G spanning T. Such a cycle is called a *Steiner tour*. The nodes not in T are called *Steiner nodes*. Given a weight vector $c \in \mathbb{R}^{|V|}$ associated with the nodes of G (in addition to the edge weights), and a root vertex $r \in V$, the *r-traveling salesman problem* (r-TSP) is to find a simple cycle containing r and whose total weight of both nodes and edges is minimized. Such a cycle is called an *r-tour*. An r-tour will be called *trivial* if it is reduced to the node r. The r-TSP is also called the *extended formulation* of the StSP.

In this paper we give a complete description, in $\mathbb{R}^{|E|+|V|}$, of the polytope associated with the solutions to the r-TSP in the class of series-parallel graphs. By projecting this polytope onto $\mathbb{R}^{|E|}$, we obtain a complete characterization of the polytope associated with the solutions to the StSP in the same class of graphs. This yields polynomial cutting plane algorithms to solve both the r-TSP and the StSP in that class of graphs.

The StSP and r-TSP are both NP-hard. They contain as a special case the well-known traveling salesman problem (TSP). The TSP has been shown to be polynomial in special classes of graphs. In [9], Cornuéjols, Fonlupt, and Naddef consider the graphical Steiner TSP, that is, when the Steiner tour can go through a node more than once. They give a linear time algorithm for this problem on series-parallel graphs. Their algorithm is an extension of an algorithm of Ratliff and Rosenthal [25] for graphs that model rectangular warehouses (a particular class of series-parallel graphs).

*Received by the editors January 20, 2000; accepted for publication (in revised form) February 10, 2002; published electronically October 1, 2002. This research was supported by FONDAP Matematicas Aplicadas.

http://www.siam.org/journals/siopt/13-2/32228.html

†Universidad de Chile, Departamento de Ingenieria Matematica and Centro de Modelamiento Matematico, CNRS, Santiago, Chile. Current address: CUST, Université de Clermont II and Laboratoire d'Econométrie, CNRS, Ecole Polytechnique, 1 rue Descartes, 75005 Paris, France (baiou@cust.univ-bpclermont.fr).

‡Laboratoire LIMOS, CNRS, Université de Clermont II, Complexe Scientifique des Cézeaux, 63177 Aubière Cedex, France (Ridha.Mahjoub@math.univ-bpclermont.fr).

Let $G = (V, E)$ be a graph. If $F \subseteq E$, $U \subseteq V$, then $(x^F, y^U) \in \mathbb{R}^{|E|+|V|}$ denotes the *incidence vector* of the subgraph (U, F) of G, i.e., $x^F(e) = 1$ if $e \in F$ and 0 otherwise, and $y^U(v) = 1$ if $v \in U$ and 0 otherwise. The *r-traveling* (resp., *Steiner traveling*) *salesman polytope* of G, denoted by r-TSP(G) (resp., StSP(G, T)), is the convex hull of the incidence vectors of the r-tours (resp., Steiner tours) of G, i.e.,

$$r\text{-TSP}(G) = \text{conv}\{(x^F, y^U) \in \mathbb{R}^{|E|+|V|} |\ (U, F) \text{ is an } r\text{-tour of } G\},$$
$$\text{StSP}(G, T) = \text{conv}\{x^F \in \mathbb{R}^{|E|} |\ F \subseteq E \text{ is a Steiner tour}\}.$$

Let TSP(G) denote the polytope associated with the TSP.

To the best of our knowledge, neither the r-TSP(G) nor the StSP(G, T) has been considered in the literature. However, the traveling salesman polytope, TSP(G), has been one of the most attractive subjects in polyhedral combinatorics in the past three decades [20], [21]. In particular, several classes of facet defining inequalities of TSP(G) have been identified, and efficient separation algorithms have been devised.

Complete descriptions of the TSP(G) have been obtained for some classes of graphs. Cornuéjols, Naddef, and Pulleyblank [8] describe the TSP(G) for Halin graphs. In [3], Barahona and Grötschel characterize the TSP(G) for graphs not contractible to $K_5 \setminus \{e\}$. A complete description of a minimal system of inequalities defining TSP(G), when G is complete, is known for graphs having no more than 8 nodes. Norman [24] describes the TSP(G) for complete graphs on 6 nodes. Boyd and Cunningham [5] give that description for graphs on 7 nodes, and Christof, Jünger, and Reinelt [10] give a description of the TSP(G) for graphs on 8 nodes.

A graph $G = (V, E)$ is said to be *k-edge connected* (for k fixed) if, for any pair of nodes $i, j \in V$, there are at least k edge-disjoint paths from i to j. Given weights on the edges of G and a set of terminals $T \subseteq V$, the *Steiner 2-edge connected subgraph problem* is the problem of finding a minimum 2-edge connected subgraph of G, spanning T. This problem is closely related to the StSP. In fact, as is pointed out in [13], when $T = V$, the problem of determining if a graph $G = (V, E)$ contains a Steiner tour (Hamiltonian cycle) can be reduced to the Steiner 2-edge connected subgraph problem. The relation between the two problems has been widely investigated in the metric case, that is, when the underlying graph $G = (V, E)$ is complete and the weight function satisfies the triangle inequalities (i.e., $w(e_1) \leq w(e_2) + w(e_3)$ for every three edges e_1, e_2, e_3 defining a triangle in G). In particular, Monma, Munson, and Pulleyblank [23] showed that $\tau \leq \frac{4}{3}Q_2$ when $T = V$, where τ is the weight of an optimal Steiner tour and Q_2 is the weight of an optimal 2-edge connected subgraph. Then it follows that the value τ' of an optimal solution of the classical linear relaxation of the TSP(G) provides a lower bound on τ. Cunningham (see [23]) shows that τ' also provides a lower bound on Q_2. Further structural properties and worst case analysis are given in Frederickson and Ja'Ja' [15], Bienstock, Brickell, and Monma [4], and Goemans and Bertsimas [17].

Given a graph $G = (V, E)$ with weights on its edges and a set of terminals $S \subseteq V$, the *Steiner tree problem* is to find a minimum weight tree in G which spans S. This problem, which is known to be NP-hard, is closely related to the StSP. Although a polynomial time algorithm in series-parallel graphs is known for this problem, still we do not have a complete description of the associated polytope in that class of graphs. In [16], Goemans gives an extended formulation for that problem and characterizes the associated polytope when the graph is series-parallel. By projecting that polytope onto the edge variables, he also obtains a large class of facet-defining inequalities for the Steiner tree polytope. For more details on the polyhedral aspect of that problem, see [6], [7], [22], and [11].

In the next section we present an integer programming formulation of the r-TSP and give some basic properties of the relaxation of our formulation. In section 3, we prove that the linear inequalities in our formulation are sufficient to completely characterize the r-TSP(G) when G is series-parallel. In section 4, we give a complete description of the StSP(G,T) in series-parallel graphs; this is done by projecting r-TSP(G) on the edge variables. The remainder of this section is devoted to more definitions and notations.

The graphs we consider are finite, undirected, and connected and may have multiple edges and loops. We denote a graph by $G=(V,E)$, where V is the *node set* and E is the *edge set* of G. If e is an edge with endnodes u and v, then we write $e=uv$.

A graph G is said to be *contractible* to a graph H if H may be obtained from G by a sequence of elementary removals and contractions of edges. A contraction consists of identifying a pair of adjacent vertices, preserving all other vertices, and preserving all other adjacencies between vertices. A graph is called *series-parallel* [12] if it is not contractible to K_4 (the complete graph on four nodes). Clearly, series-parallel graphs have the following property.

REMARK 1. *If G is a series-parallel graph contractible to a graph H, then H is series-parallel.*

Given a graph $G=(V,E)$ and a node subset $W\subseteq V$ of G, the set of edges having one endnode in W and the other in $V\setminus W$ is called a *cut* of G and denoted by $\delta(W)$. If $v\in V$ is a node of G, then we write $\delta(v)$ for the cut $\delta(\{v\})$. We denote by $G(W)$ the subgraph of G induced by W, and by $E(W)$ its edges. For $W,W'\subseteq V$, (W,W') denotes the set of edges having one endnode in W and the other in W'. If $W\subseteq V$, we let $\overline{W}=V\setminus W$. Given a constraint $ax\geq\alpha$, a^T, $x\in\mathbb{R}^n$, and a solution $x^*\in\mathbb{R}^n$, we will say that $ax\geq\alpha$ is *tight* for x^* if $ax^*=\alpha$.

2. The polytope r-TSP(G). Let $G=(V,E)$ be a graph and $r\in V$ a root vertex. Let $x(e)$, $y(v)$ be variables associated with each edge e and node v. For any subset of edges $F\subseteq E$, we let $x(F)=\sum_{e\in F}x(e)$.

The r-TSP can then be formulated as the following integer program:

$$\text{Minimize } \sum_{e\in E} w(e)x(e)+\sum_{v\in V} c(v)y(v)$$

subject to

$$
\begin{aligned}
&(1) && x(\delta(W))\geq 2y(v) && \text{for all } W\subset V,\ |\overline{W}|\geq 2,\ r\in W,\ v\in\overline{W},\\
&(2) && x(\delta(r))\leq 2y(r), &&\\
&(3) && x(\delta(v))=2y(v) && \text{for all } v\in V\setminus\{r\},\\
&(4) && x(e)\leq y(v) && \text{for all } v\in V,\ e\in\delta(v),\\
&(5) && y(v)\leq 1 && \text{for all } v\in V,\\
&(6) && x(e)\geq 0 && \text{for all } e\in E,\\
&(7) && x(e),y(v)\in\mathbb{N} && \text{for all } e\in E,\ v\in V.
\end{aligned}
$$

Constraints (1) and (3) will be called *generalized cut constraints*. A generalized cut constraint is associated with a cut $\delta(W)$ and a node $v\in\overline{W}$. The pair $(\delta(W),v)$ will be called a *generalized cut*. A generalized cut will be called *tight* for a solution (x,y) if the corresponding constraint is tight for (x,y). Notice that the generalized cuts $(\delta(W),v)$ with $\overline{W}=\{v\}$ (equations (3)) are tight for all solutions of $H(G)$. The case where $|\overline{W}|\geq 2$ will be specified if necessary. Inequalities (5) and (6) are called *trivial inequalities*. Inequalities (4) combined with the trivial inequalities (5) imply

that if $x(e) = 1$ for some $e \in \delta(v)$, then $y(v) = 1$. Let $H(G)$ denote the polytope defined by inequalities (1)–(6). We have the following.

THEOREM 2. *If G is series-parallel, then r-TSP$(G) = H(G)$.*

The proof of this theorem will be given in the following section. In what follows we are going to discuss some properties of the solutions of $H(G)$, which will be useful in the rest of the paper.

LEMMA 3. *Let $(x, y) \in \mathbb{R}^{|E|+|V|}$ be a solution of $H(G)$ such that $x(e) > 0$ for all $e \in E$. If $(\delta(W), v)$ is a generalized cut tight for (x, y), then $G(\overline{W})$ is connected.*

Proof. This is clear if $\overline{W} = \{v\}$. So suppose that $|\overline{W}| \geq 2$, and let us assume, on the contrary, that there is a partition $\overline{W}_1, \overline{W}_2$ of $\overline{W}$ such that $(\overline{W}_1, \overline{W}_2) = \emptyset$. Without loss of generality, we may suppose that $v \in \overline{W}_1$. Since G is connected, it follows that $(W, \overline{W}_1) \neq \emptyset \neq (W, \overline{W}_2)$. By our hypothesis, we have $x(W, \overline{W}_2) > 0$. As $(\delta(W), v)$ is tight for (x, y), it follows that $x(\delta(W)) = x(W, \overline{W}_1) + x(W, \overline{W}_2) = 2y(v)$. This implies that $x(\delta(W \cup \overline{W}_2)) = x(W, \overline{W}_1) < 2y(v)$, and thus the generalized cut $(\delta(W \cup \overline{W}_2), v)$ is violated by (x, y). But this contradicts the fact that $(x, y) \in H(G)$. ∎

LEMMA 4. *Let $(x, y) \in H(G)$, and let $(\delta(W), v)$ and $(\delta(W'), v')$ be two generalized cuts tight for (x, y). Then the following hold:*

(i) *If $v \in \overline{W \cup W'}$, then $(\delta(W \cap W'), v')$ and $(\delta(W \cup W'), v)$ are both generalized cuts tight for (x, y).*

(ii) *If $v \in W' \setminus W$ and $v' \in W \setminus W'$, then $(\delta(\overline{W' \setminus W}), v)$ and $(\delta(\overline{W \setminus W'}), v')$ are both generalized cuts tight for (x, y).*

Proof. The proof follows from the submodularity of the cuts, that is,

$$x(\delta(W)) + x(\delta(W')) \geq x(\delta(W \cap W')) + x(\delta(W \cup W')) \text{ for any } W, W' \subset V. \quad \blacksquare$$

3. Proof of Theorem 2. Let $G = (V, E)$ be a graph and $T \subseteq V$ a set of terminals. A *Steiner 2-edge connected subgraph* of G is a 2-edge connected subgraph of G spanning T. Denote by STECP(G, T) the convex hull of the incidence vectors of the Steiner 2-edge connected subgraphs of G, and let $P(G, T)$ be the polytope given by the following linear inequalities:

$$0 \leq x(e) \leq 1 \qquad \text{for all } e \in E, \tag{8}$$

$$x(\delta(W)) \geq 2 \qquad \text{for all } W \subseteq V,\ T \neq W \bigcap T \neq \emptyset, \tag{9}$$

$$x(\delta(W)) \geq 2x(e) \qquad \text{for all } W \subseteq V,\ T \subseteq W,\ e \notin E(W). \tag{10}$$

Inequalities (9) and (10) are called *Steiner* and *left-Steiner cut inequalities*, respectively. In [2], Baïou and Mahjoub state the following.

THEOREM 5. *If G is series-parallel, then STECP$(G, T) = P(G, T)$.*

For a complete proof of this theorem, see [1]. In what follows we are going to use that description to prove Theorem 2.

The proof of Theorem 2 is by induction on the number of edges. The theorem is trivially true for a graph with no more than two edges. Suppose it is true for any series-parallel graph with no more than m edges and suppose that G contains exactly $m+1$ edges. Let us assume, on the contrary, that r-TSP$(G, S) \neq H(G)$, and let (x, y) be a fractional extreme point of $H(G)$. We have the following lemmas.

LEMMA 6. *$x(e)$ and $y(v)$ are positive for all $e \in E$ and $v \in V$.*

Proof. By inequalities (4) it suffices to prove that $x(e) > 0$ for all $e \in E$. If e_0 is an edge such that $x(e_0) = 0$, then let $x' \in \mathbb{R}^{|E|-1}$ be given by $x'(e) = x(e)$ for all $e \in E \setminus \{e_0\}$. Clearly, (x', y) belongs to $H(G')$, where G' is the graph obtained from G by deleting e. Moreover (x', y) is an extreme point of $H(G')$. Since (x', y) is fractional and G' is series-parallel, we have a contradiction. ∎

LEMMA 7. *If $(\delta(W), v)$ is a generalized cut tight for (x, y) with $|\overline{W}| \geq 2$, then $y(v) = 1$.*

Proof. Suppose, on the contrary, that $y(v) < 1$. Suppose that $|W|$ is minimum. That is, for every generalized tight cut $(\delta(W'), w')$ with $|\overline{W'}| \geq 2$ and $|W'| < |W|$, we have $y(w') = 1$. Now remark that by constraints (1)

$$y(v) \geq y(v') \quad \text{for all } v' \in \overline{W}. \tag{11}$$

Let $G' = (V', E')$ be the graph obtained from G by contracting $\overline{W}$, and denote by $\bar{w}$ the node resulting from this contraction. By Lemma 3 together with Remark 1, it follows that G' is series-parallel. Let x' be the restriction of x on E', and $y' \in \mathbb{R}^{|W|+1}$ such that $y'(u) = y(u)$ if $u \in W$ and $y'(\bar{w}) = y(v)$.

It is easy to see that (x', y') is a solution of $H(G')$. As G' is series-parallel and $|E'| < |E|$, by the induction hypothesis, $H(G')$ is integral. In consequence, (x', y') can be written as a convex combination of (integral) extreme points of $H(G')$. Thus there are t extreme points of $H(G')$, $(x'_1, y'_1), \ldots, (x'_t, y'_t)$ and $\lambda_1, \ldots, \lambda_t \geq 0$, such that

$$(x', y') = \sum_{i=1}^{t} \lambda_i (x'_i, y'_i), \qquad \sum_{i=1}^{t} \lambda_i = 1.$$

Since $y'(\bar{w}) = y(v) < 1$, there must exist a solution among $(x'_1, y'_1), \ldots, (x'_t, y'_t)$, say (x'_1, y'_1) such that $y'_1(\bar{w}) = 0$. By equality (3) associated with $\bar{w}$, it follows that $x'_1(\delta(W)) = x'_1(\delta(\bar{w})) = 2y'_1(\bar{w})) = 0$. Let $(x^*, y^*) \in \mathbb{R}^{|E|+|V|}$ be the solution such that

$$x^*(e) = \begin{cases} x'_1(e) & \text{if } e \in E(W), \\ 0 & \text{otherwise,} \end{cases} \qquad y^*(v) = \begin{cases} y'_1(v) & \text{if } v \in W, \\ 0 & \text{otherwise.} \end{cases}$$

In what follows we are going to show that every constraint of $H(G)$ that is tight for (x, y) is also tight for (x^*, y^*). Since $(x, y) \neq (x^*, y^*)$, this contradicts the extremality of (x, y).

First, it can be easily seen that every inequality among (2)–(6) that is tight for (x, y) is also tight for (x^*, y^*). So let us consider a generalized cut $(\delta(W'), v')$ tight for (x, y) with $|\overline{W'}| \geq 2$. Suppose first that $W' \subseteq W$.

If $v' \in W$, then $(\delta(W'), v')$ is also a generalized cut in G', and thus it is tight for both (x', y') and (x'_1, y'_1). Hence $x^*(\delta(W')) = x'_1(\delta(W')) = 2y'_1(v') = 2y^*(v')$.

If $v' \in \overline{W}$, then $2y(v) \leq x(\delta(W')) = 2y(v')$. By (11) this implies that $y(v') = y(v) = y'(\bar{w})$. Thus $(\delta(W'), \bar{w})$ is a generalized cut in G' tight for (x', y'), and hence $x^*(\delta(W')) = x'_1(\delta(W')) = 2y'_1(\bar{w}) = 0 = 2y^*(v')$. Now if $W \subseteq W'$, by the definition of (x^*, y^*), we have $x^*(\delta(W')) = 2y^*(v') = 0$. Thus we can suppose that $W \setminus W' \neq \emptyset \neq W' \setminus W$. We consider two cases.

Case 1. $v \in \overline{W \cup W'}$. From Lemma 4(i) we have that $(\delta(W \cap W'), v')$ is a generalized cut tight for (x, y). Since $(W \cap W') \subset W$, it follows from above that $(\delta(W \cap W'), v')$ is also tight for (x^*, y^*) and thus $x^*(\delta(W')) = x^*(\delta(W \cap W')) = 2y^*(v')$.

Case 2. $v \in W' \setminus W$. Then $v' \notin \overline{W \cup W'}$; otherwise by Lemma 4(i), by exchanging v' and v, $(\delta(W \cap W'), v)$ would be a generalized cut tight for (x, y), which contradicts the minimality of $|W|$. Thus suppose that $v' \in W \setminus W'$. By Lemma 4(ii), $(\delta(\overline{W \setminus W'}), v')$ is a generalized cut tight for (x, y). Since $(\delta(\overline{W \setminus W'}), v')$ is also a generalized cut in G', it is also tight for (x', y') and hence for (x'_1, y'_1). Thus

$$x^*(\delta(W')) = x'_1(\delta(\overline{W \setminus W'})) = 2y'_1(v') = 2y^*(v'). \qquad \square$$

Now, let

$$T = \{v \in V : \ y(v) = 1\}.$$

LEMMA 8. $|T| \geq 2$.

Proof. Assume the contrary, that is, $|T| \leq 1$. Then by inequalities (4), it follows that $x(e) < 1$ for all $e \in E$. And by the inequality of type (1) corresponding to $W = \{r\}$ together with inequality (2), it follows that $y(v) < 1$ for all $v \in V \setminus \{r\}$. If $T = \emptyset$, then we also have $y(r) < 1$. In consequence, if we consider the solution $(0, 0)$, we will have that all the constraints of $H(G)$ that are tight for (x, y) are also tight for that solution. But this contradicts the extremality of (x, y).

Now let us assume that $|T| = 1$. Hence $y(r) = 1$ (and $y(v) < 1$ for all $v \in V \setminus \{r\}$).

If $x(\delta(r)) < 2$, then, by considering the incidence vector of the trivial r-tour, we will also have a solution that satisfies with equality all the constraints of $H(G)$ tight for (x, y), which again yields a contradiction.

So suppose that $x(\delta(r)) = 2$. Since $0 < y(v) < 1$ for all $v \in V \setminus \{r\}$, by Lemma 7 no inequality (1) is tight for (x, y).

Claim. No inequality (4) is tight for (x, y).

Proof of the claim. As (x, y) is an extreme point of $H(G)$ and $0 < x(e) < 1$ for all $e \in E$, it follows that there is a set of pairs $(e_1, v_1), \ldots, (e_l, v_l)$, $e_i \in \delta(v_i)$ for $i = 1, \ldots, l$, such that (x, y) is the unique solution of the system

$$\text{(L)} \begin{cases} y(r) = 1, & \\ x(\delta(v)) = 2y(v) & \text{for all } v \in V, \\ x(e_i) = y(v_i) & \text{for } i = 1, \ldots, l. \end{cases}$$

Let $f = uv \in E$. Suppose that $x(f) = y(u)$. Let $G' = (V', E')$ be the graph obtained from G by contracting f. Let x' be the restriction of x on E' and $y' \in \mathbb{R}^{|V'|}$ such that $y'(w) = y(w)$ if $w \in V' \setminus \{w_0\}$ and $y'(w_0) = y(v)$, where w_0 is the node of V' that arises from the contraction of f. Now as $x(f) < 1$ and $y(r) = 1$, we have $u \neq r$. If $v = r$, we let w_0 be the root vertex in G'. Note that $x'(\delta(w_0)) = 2y'(w_0)$. It can be, in fact, easily seen that $(x', y') \in H(G')$. In what follows we will show that (x', y') is also an extreme point of $H(G')$. Indeed, if this is not the case, as by the induction hypothesis $H(G')$ is integral, there are integral extreme points $(x^1, y^1), \ldots, (x^k, y^k)$ and scalars $\lambda_1, \ldots, \lambda_k \geq 0$ such that

$$(x', y') = \sum_{i=1}^{k} \lambda_i (x^i, y^i), \qquad \sum_{i=1}^{k} \lambda_i = 1.$$

Note that any constraint of $H(G')$ tight for (x', y') is also tight for (x^i, y^i) for $i = 1, \ldots, k$. We distinguish two cases.

Case 1. $v \neq r$. Thus $y'(w_0) < 1$. In consequence, there must exist one of the extreme points (x^i, y^i), say (x^1, y^1), such that $y^1(w_0) = 0$. Since $x^1(\delta(w_0)) = 2y^1(w_0)$ and $x^1(e) \geq 0$ for all $e \in E$, it follows that $x^1(e) = 0$ for all $e \in \delta(w_0)$. Consider the solution $(\bar{x}, \bar{y}) \in \mathbb{R}^{|E|+|V|}$ given by

$$\bar{x}(e) = \begin{cases} x^1(e) & \text{if } e \in E \setminus (\delta(u) \cup \delta(v)), \\ 0 & \text{otherwise,} \end{cases} \qquad \bar{y}(w) = \begin{cases} y^1(w) & \text{if } w \in V \setminus \{u, v\}, \\ 0 & \text{otherwise.} \end{cases}$$

We have that $(\bar{x}, \bar{y})$ is a solution of (L). In fact, clearly equalities (3) as well as the equality $x(\delta(r)) = 2y(r)$ are satisfied by $(\bar{x}, \bar{y})$. Moreover, as $y'(r) = 1$, we

have $\bar{y}(r) = y^1(r) = 1$. Now for a pair (e_i, v_i), as $\bar{x}(e) = 0$ for $e \in \delta(u) \cup \delta(v)$ and $\bar{y}(u) = \bar{y}(v) = 0$, it follows that the corresponding inequality is satisfied with equality if $v_i \in \{u, v\}$. If $v_i \in V \setminus \{u, v\}$, as $x'(e_i) = y'(v_i)$, we should have $x^1(e_i) = y^1(v_i)$. Hence $(\bar{x}, \bar{y})$ satisfies system (L). As $(\bar{x}, \bar{y}) \neq (x, y)$, this is a contradiction.

Case 2. $v = r$. Thus $y'(w_0) = 1$. Since $x'(\delta(u) \setminus \{f\}) < 1$, there must exist one of the extreme points (x^i, y^i), say (x^1, y^1), such that $x^1(\delta(u) \setminus \{f\}) = 0$. Consider the solution $(\bar{x}, \bar{y}) \in \mathbb{R}^{|E|+|V|}$ given by

$$\bar{x}(e) = \begin{cases} 0 & \text{if } e = f, \\ x^1(e) & \text{otherwise,} \end{cases} \qquad \bar{y}(w) = \begin{cases} y^1(w) & \text{if } w \in V \setminus \{u, v\}, \\ 0 & \text{if } w = u, \\ 1 & \text{if } w = v. \end{cases}$$

As above, one can easily verify that $(\bar{x}, \bar{y})$ is a solution of (L), which is again a contradiction and this ends the proof of the claim.

From the claim above and Lemma 6, it follows that the only inequalities tight for (x, y) are

$$(\mathrm{L}') \begin{cases} y(r) = 1, \\ x(\delta(v)) = 2y(v) \quad \text{for all } v \in V. \end{cases}$$

Now we claim that G contains at least one nontrivial r-tour. In fact, as $x(\delta(r)) = 2$ and $x(e) \leq 1$ for all $e \in E$, r is adjacent to at least two nodes. If there is no nontrivial tour, then G must contain a cut $\delta(S)$ separating r and one of its neighbors, say u, such that $\delta(S) = \{ru\}$. On the other hand, we have $x(ru) \leq x(\delta(u)) = 2y(u)$. If $|\bar{S}| \geq 2$, as $y(u) < 1$, by Lemma 7 it follows that $x(ru) = x(\delta(S)) \neq 2y(u)$. Hence $x(\delta(S)) < 2y(u)$, a contradiction. Now suppose that $|\bar{S}| = 1$, that is, $\bar{S} = \{u\}$. Thus $x(ru) = x(\delta(u)) = 2y(u)$. However, by inequalities (4) one should have $x(ru) \leq y(u)$. Since $y(u) > 0$, this is also impossible.

Now the incidence vector of any nontrivial r-tour verifies equalities (L′). This yields a contradiction with the fact that (x, y) is an extreme point, which finishes the proof of our lemma. □

In what follows, we will show that the projection of (x, y) onto $\mathbb{R}^{|E|}$, i.e., x, is an extreme point of $P(G, T)$. It is clear that every constraint of $P(G, T)$ can be obtained from some linear combination of constraints of $H(G)$. Thus $x \in P(G, T)$. Now to prove that x is an extreme point of $P(G, T)$, it suffices to display a system of equalities from $P(G, T)$, where x is the unique solution.

If there exists an inequality of type (4) that is tight for (x, y) with $y(v) = 1$, then this equality corresponds to an inequality $x(e) \leq 1$ of $P(G, T)$ that is tight for x. Denote such equalities by (8′). Let $(\delta(W), v)$ be a generalized cut tight for (x, y). Then by Lemma 7 we have $y(v) = 1$ and hence $v \in T$. Thus the equation yielded by $(\delta(W), v)$ corresponds to the Steiner cut inequality $x(\delta(W)) \geq 2$ of $P(G, T)$ that is tight for x. Let us denote by (9′) such equalities.

Now consider an equality of type (3). If $y(v) = 1$ for $v \neq r$, then, as before, this equality corresponds to a Steiner cut of $P(G, T)$ that is tight for x. If inequality (2) is tight for (x, y)—that is, $x(\delta(r)) = 2$—then, by Lemma 8, $|T| \geq 2$, and $r \in T$, this equality also corresponds to a Steiner cut tight for x in $P(G, T)$. We will also denote these equalities by (9′). If $y(v) < 1$ and there exists $e \in \delta(v)$ such that $x(e) = y(v)$, then this yields a left-Steiner cut $x(\delta(v)) \geq 2x(e)$ tight for x. We let (10′) be the set of these equalities. Let (S) be the system of equalities defined by (8′), (9′), and (10′).

We claim that x is the unique solution of (S). Indeed, if there is a further solution x' of (S), then by considering $y' \in \mathbb{R}^{|V|}$ such that $y'(v) = \frac{1}{2}x'(\delta(v))$ for all $v \in V$,

the solution (x', y') would verify with equality all the constraints tight for (x, y). As $x' \neq x$ and (x, y) is an extreme point of $H(G)$, this is impossible.

Now since the equalities of (S) all come from inequalities of $P(G, T)$, x is an extreme point of $P(G, T)$. Since x is fractional and G is series-parallel, this contradicts Theorem 5.

4. The polytope StSP(G, T). Let $G = (V, E)$ be a graph and $T \subseteq V$ a set of terminals. Let $N = V \setminus T$ be the set of Steiner vertices.

Let $P_{E,N}(G) \subseteq \mathbb{R}^{|E|+|N|}$ be the polytope obtained from $H(G)$ by selecting a root vertex $r \in T$ and setting $y(v) = 1$ for all $v \in T$. Thus $P_{E,N}(G)$ is given by the following system:

$$x(\delta(W)) \geq 2y(v) \quad \text{for all } W \subset V,\ T \subseteq W,\ v \notin W, \tag{12}$$
$$x(\delta(v)) \leq 2y(v) \quad \text{for all } v \in N, \tag{13}$$
$$x(e) \leq y(v) \quad \text{for all } v \in N,\ e \in \delta(v), \tag{14}$$
$$y(v) \leq 1 \quad \text{for all } v \in N, \tag{15}$$
$$x(\delta(W)) \geq 2 \quad \text{for all } W \subseteq V,\ T \neq W \bigcap T \neq \emptyset, \tag{16}$$
$$x(\delta(v)) = 2 \quad \text{for all } v \in T \setminus \{r\}, \tag{17}$$
$$x(\delta(r)) \leq 2, \tag{18}$$
$$x(e) \leq 1 \quad \text{for all } e \in \delta(v),\ v \in T, \tag{19}$$
$$x(e) \geq 0 \quad \text{for all } e \in E. \tag{20}$$

As $P_{E,N}(G)$ is a face of $H(G)$, by Theorem 2, we have the following.

COROLLARY 9. *$P_{E,N}(G)$ is integral if G is series-parallel.*

Now, to describe the polytope $\mathrm{StSP}(G, T)$, we are going to project onto the subspace of the edge variables. To do this we use Fourier–Motzkin elimination [26] to eliminate the node variables $y(v)$ from $P_{E,N}(G)$. For every node $v \in N$, we will combine inequalities containing $+y(v)$ with the ones containing $-y(v)$ as follows:

- By combining inequalities (12) and (13), we obtain the inequalities

$$x(\delta(W)) \geq x(\delta(v)) \quad \text{for all } W \subset V,\ T \subseteq W,\ v \notin W; \tag{21}$$

- combining inequalities (12) and (14), we obtain the left-Steiner cut inequalities (10);
- combining inequalities (13) and (15), we obtain

$$x(\delta(v)) \leq 2 \quad \text{for all } v \in N; \tag{22}$$

- and finally, the combination of inequalities (14) and (15) gives the inequalities $x(e) \leq 1$ for all $e \in \delta(v)$, $v \in N$. This, together with inequalities (19), yields

$$x(e) \leq 1 \quad \text{for all } e \in E. \tag{23}$$

LEMMA 10. *The left-Steiner cut inequalities* (10), $x(\delta(W)) \geq 2x(e)$, *with* $|\overline{W}| \geq 2$, *are redundant for StSP(G, T).*

Proof. As $e \notin E(W)$, there is a node, say v, of e that belongs to $\overline{W}$. By inequality (21) associated with W and v together with the left-Steiner cut associated with $\delta(V \setminus \{v\})$ and the edge e, we have

$$x(\delta(W)) \geq x(\delta(v)) = x(\delta(V \setminus \{v\})) \geq 2x(e). \qquad \square$$

By Lemma 10, the left-Steiner cut inequalities that may be essential in the description of $\mathrm{StSP}(G, T)$ can be written as follows:

$$x(\delta(v)) \geq 2x(e) \quad \text{for all } v \in N,\ e \in \delta(v). \tag{24}$$

Now from the development above and Corollary 9 we obtain the following result.

THEOREM 11. *If G is series-parallel, then inequalities* (16)–(18), (20), *and* (21)–(24) *completely describe StSP(G,T).*

5. Concluding remarks. We have studied an extended formulation of the StSP and have given a complete linear description of the associated polytope when the underlying graph is series-parallel. By projecting this polytope onto the edge variables space, we have obtained a description of the Steiner traveling salesman polytope in that class of graphs.

It would be interesting to have such a description for the graphical Steiner traveling salesman polyhedron in that class of graphs. A complete characterization of that polyhedron in series-parallel graphs is, unfortunately, still unknown even when $T = V$. In fact, as shown by Cornuéjols, Fonlupt, and Naddef [9], the traveling salesman polyhedron in this case may contain constraints which do not come from cuts. In [14], Fonlupt and Naddef characterize the graphs for which the graphical traveling salesman polyhedron is given by the nonnegativity and the cut constraints.

Given a graph $G = (V, E)$ and two nodes u, v of V, let $G_{u,v}$ be the graph obtained from G by identifying u and v. Let w be the node resulting from the identification of u and v. Let $P_{u,v}(G)$ be the polytope, the extreme points of which are the incidence vectors of the paths of G between u and v, different from uv (if $uv \in E$). Clearly, $P_{u,v}(G) = \mathrm{StSP}(G_{u,v}, \{w\})$. Thus Theorem 11 provides at the same time a description of $P_{u,v}(G)$ when G is series-parallel and $uv \in E$.

We conclude by mentioning that, as inequalities (1), (16), and (21) can be separated in polynomial time, by the ellipsoid method [18], Theorems 2 and 11 provide polynomial cutting plane algorithms for both the r-TSP and StSP problems on series-parallel graphs. These are, to the best of our knowledge, the first polynomial time algorithms for these problems in that class of graphs.

Acknowledgments. We are grateful to an anonymous referee for his comments that permitted us to improve the presentation and considerably shorten the proof of the main result. Part of this work was done while the second author was visiting Departamento de Ingenieria Matematica of Universidad de Chile in April, 1999. The financial support is greatly appreciated.

REFERENCES

[1] M. BAÏOU, *Le problème du sous-graphe Steiner 2-arête connexe: Approche polyédrale*, Ph.D. dissertation, N 1639, Université de Rennes 1, Rennes, France, 1996.

[2] M. BAÏOU AND A. R. MAHJOUB, *Steiner 2-edge connected subgraph polytopes on series-parallel graphs*, SIAM J. Discrete Math., 10 (1997), pp. 505–514.

[3] F. BARAHONA AND M. GRÖTSCHEL, *The Traveling Salesman Problem for Graphs Not Contractible to $K_5 - \{e\}$*, Technical Report 77, Mathematiches Institut, Universität Augsburg, Augsburg, Germany, 1985.

[4] D. BIENSTOCK, E. F. BRICKELL, AND C. L. MONMA, *On the structure of minimum-weight k-connected spanning networks*, SIAM J. Discrete Math., 3 (1990), pp. 320–329.

[5] S. C. BOYD AND W. R. CUNNINGHAM, *Small traveling salesman polytopes*, Math. Oper. Res., 16 (1991), pp. 259–271.

[6] S. CHOPRA AND M. R. RAO, *The Steiner tree problem* I: *Formulations, compositions and extension of facets*, Math. Program., 64 (1994), pp. 209–229.

[7] S. CHOPRA AND M. R. RAO, *The Steiner tree problem* II: *Properties and classes of facets*, Math. Program., 64 (1994), pp. 231–246.

[8] G. CORNUÉJOLS, D. NADDEF, AND W. R. PULLEYBLANK, *Halin graphs and the traveling salesman problem*, Math. Program., 26 (1983), pp. 287–294.

[9] G. Cornuéjols, J. Fonlupt, and D. Naddef, *The traveling salesman problem on a graph and some related integer polyhedra*, Math. Program., 33 (1985), pp. 1–27.
[10] T. Christof, M. Jünger, and G. Reinelt, *A complete description of the traveling salesman polytope on* 8 *nodes*, Oper. Res. Lett., 10 (1991), pp. 497–500.
[11] M. Didi Biha, H. Kerivin, and A. R. Mahjoub, *Steiner trees and polyhedra*, Discrete Appl. Math., 112 (2001), pp. 101–120.
[12] R. J. Duffin, *Topology of series-parallel networks*, J. Math. Anal. Appl., 10 (1965), pp. 303–318.
[13] K. P. Eswaran and R. E. Tarjan, *Augmentation problems*, SIAM J. Comput., 5 (1976), pp. 653–665.
[14] J. Fonlupt and D. Naddef, *The traveling salesman problem in graphs with some excluded minors*, Math. Program., 53 (1992), pp. 147–172.
[15] G. N. Frederickson and J. Ja'Ja', *On the relationship between the biconnectivity augmentations and traveling salesman problem*, Theoret. Comput. Sci., 13 (1982), pp. 189–201.
[16] M. X. Goemans, *The Steiner tree polytope and related polyhedra*, Math. Program., 63 (1994), pp. 157–182.
[17] M. X. Goemans and D. J. Bertsimas, *Survivable networks, linear programming and the parsimonious property*, Math. Program., 60 (1993), pp. 145–166.
[18] M. Grötschel, L. Lovász, and A. Schrijver, *The ellipsoid method and its consequences in combinatorial optimization*, Combinatorica, 1 (1981), pp. 70–89.
[19] M. Jünger and W. R. Pulleyblank, *New primal and dual matching heuristics*, Algorithmica, 13 (1995), pp. 357–380.
[20] M. Jünger, G. Reinelt, and G. Rinaldi, *The traveling salesman problem*, in Network Models, M. O. Ball, T. L. Magnanti, C. L. Monma, and G. L. Nemhauser, eds., Handbooks Oper. Res. Management Sci., North–Holland, Amsterdam, 1996, pp. 225–330.
[21] E. L. Lawler, J. K. Lenstra, A. H. G. Rinnooy Kan, and D. B. Shmoys, eds., *The Traveling Salesman Problem*, John Wiley, New York, 1985.
[22] F. Margot, A. Prodon, and Th.-M. Liebling, *Tree polytope on* 2*-trees*, Math. Program., 63 (1994), pp. 183–191.
[23] C. L. Monma, B. S. Munson, and W. R. Pulleyblank, *Minimum-weight two connected spanning networks*, Math. Program., 46 (1990), pp. 153–171.
[24] R. Z. Norman, *On the convex polyhedra of the symmetric traveling salesman problem (abstract)*, Bull. Amer. Math. Soc., 61 (1955), p. 559.
[25] D. M. Ratliff and A. S. Rosenthal, *Order-picking in a rectangular warehouse: A solvable case of the traveling salesman problem*, Oper. Res., 31 (1983), pp. 507–521.
[26] A. Schrijver, *Theory of Linear and Integer Programming*, John Wiley, New York, 1986.

SIAM J. OPTIM.
Vol. 13, No. 2, pp. 508–519

ADDITIVE BOUNDING, WORST-CASE ANALYSIS, AND THE BREAKPOINT MEDIAN PROBLEM*

ALBERTO CAPRARA†

Abstract. We analyze the worst-case performance of a simple algorithm for the *breakpoint median problem* (BMP), a well-known problem in computational biology. BMP is the special case of the min-cost traveling salesman problem on a complete graph $G = (V, E)$, in which the edge cost vector $c \in \mathbb{R}^E$ has the form $1 - x^*$, with x^* a convex combination of the incidence vectors of the Hamiltonian circuits of G. The performance guarantee shown is 5/3, which improves on the previously known guarantee of 2. We also consider the *signed* variant of BMP and prove that a similar approach yields a performance guarantee of 3/2 (again improving over the previously known 2). Our proofs are based on formulating the problem as a suitable integer linear program and then defining a feasible dual solution for the associated linear programming relaxation in two phases, in a so-called *additive bounding* fashion.

Key words. breakpoint median problem, approximation algorithm, linear programming, additive bounding, traveling salesman problem, perfect matching

AMS subject classifications. 68Q25, 68R10, 05C45

PII. S1052623401384849

1. Introduction.

We analyze the worst-case performance of a simple algorithm for a problem currently very popular in computational molecular biology, namely a (1-)*median* problem in which one wants to find a point closest to a given set of points in a finite but exponentially large metric space. Our analysis will yield the first improvement over a general approximation algorithm that achieves a performance guarantee of 2. Our proofs are based on formulating the problem as a suitable *integer linear program* (ILP) and then defining a feasible dual solution for the associated *linear programming* (LP) relaxation in two phases. This two- (or more) phase approach to deriving feasible dual solutions, when applied to the practical solution of ILPs, goes under the name of *additive bounding*; see Fischetti and Toth [4].

1.1. The breakpoint median problem.

The *breakpoint median problem* (BMP) was introduced by Sankoff and Blanchette [11, 12] as a model for finding the genome that is closest to a given set of genomes, and it is widely used by the methods that solve the fundamental problem of reconstructing evolutionary trees for several species based on their genomic sequence; see Sankoff, Sundaram, and Kececioglu [14], Sankoff and Blanchette [11, 12], Blanchette, Bourque, and Sankoff [1], Sankoff, Bryant, Denault, Lang, and Burger [13], Moret, Wyman, Bader, Warnow, and Yan [8], and Moret, Wang, Warnow, and Wyman [7]. In particular, all these methods iteratively find the "best" genome to assign to a node of the tree, once the genomes associated with the neighbors of the node have been fixed. If the measure of distance between genomes is the breakpoint distance (as is almost always the case), then the problem of finding the "best" genome is a BMP. For this reason, BMP has received considerable attention from computational biologists in the last few years.

*Received by the editors February 12, 2001; accepted for publication (in revised form) April 25, 2002; published electronically October 1, 2002. This work was partially supported by MIUR and CNR, Italy.

http://www.siam.org/journals/siopt/13-2/38484.html

†DEIS, Università di Bologna, Viale Risorgimento 2, I-40136 Bologna, Italy (acaprara@deis.unibo.it).

We will first illustrate the problem for the model in which gene orientation is unknown, also called the *unsigned* BMP. In the last section, we will consider the case in which this orientation is known, the *signed* BMP. Moreover, we will focus on the case of *circular* genomes, which makes notation easier. It is easy to verify that the same results hold for the case of *linear* genomes.

Assuming that all given genomes contain the same genes, each genome (species) can be represented as a *circular permutation* of the numbers from 1 to n, where n is the number of genes, i.e., a circular sequence $\pi_1\ \pi_2\ \ldots\ \pi_n\ \pi_1\ \ldots$, where, for each $j = 1, \ldots, n$, there exists an i such that $\pi_i = j$. In what follows, we will use the term *permutation* implicitly to refer to a circular permutation. The elements of the permutation represent *genes* within the genome. A *breakpoint* of π^1 with respect to π^2 is a pair $\pi^1_i\ \pi^1_{i+1}$ (with $i \in \{1, \ldots, n\}$ and indices intended modulo n) of consecutive elements in π^1 that are not consecutive in π^2, i.e., such that neither $\pi^1_i\ \pi^1_{i+1}$ nor $\pi^1_{i+1}\ \pi^1_i$ appear consecutively in π^2. In the breakpoint model, the *distance* (called also *breakpoint distance*) between two permutations π^1, π^2 is the number of breakpoints of π^1 with respect to π^2, denoted by $b(\pi^1, \pi^2)$. It is easy to verify that $b(\pi^1, \pi^2) = b(\pi^2, \pi^1)$ and that the breakpoint distance satisfies the triangle inequality.

Given q permutations $\pi^1, \pi^2, \ldots, \pi^q$, BMP calls for a permutation μ at minimum overall breakpoint distance from the given permutations, i.e., such that $\sum_{k=1}^{q} b(\pi^k, \mu)$ is minimized. It is known (and formally proved in the next subsection) that BMP is a special case of the *traveling salesman problem* (TSP). The problem is trivial for $q \leq 2$, but $\mathcal{NP}$-hard for any $q \geq 3$, as shown by Pe'er and Shamir [9] and Bryant [2]. For general q, it is possible to show that the problem is $\mathcal{APX}$-hard [15], i.e., it does not have a polynomial time approximation scheme unless $\mathcal{P}$=$\mathcal{NP}$. General results about finding the median element in a metric space (see Gusfield [5]) show that the approximation achieved by taking as a solution the best permutation among $\pi^1, \pi^2, \ldots, \pi^q$ yields an approximation guarantee of $2 - \frac{2}{q}$. In particular, this yields a trivial $\frac{4}{3}$ approximation algorithm when $q = 3$. Actually, for $q = 3$, Pe'er and Shamir [10] presented a $\frac{7}{6}$ approximation algorithm for the signed case. In this paper, we will improve on the trivial 2 approximation for general q, showing a $\frac{5}{3}$ and a $\frac{3}{2}$ approximation algorithm for the unsigned and signed cases, respectively.

1.2. A graph theoretic representation. We start with a description of the notation used in what follows. Given an undirected graph $G = (V, E)$ and a node set $S \subset V$, we let $\delta(S)$ be the set of edges with one endpoint in S and the other in $V \setminus S$, and $E(S)$ the set of edges with both endpoints in S. Moreover, for $i \in V$, we will write $\delta(i)$ instead of $\delta(\{i\})$. We let $\mathcal{S}$ denote the collection of the nontrivial subsets of V, i.e., $\mathcal{S} = \{S \subset V : \emptyset \neq S \neq V\}$. We let a *cycle* C be defined by its set of edges, i.e., $C \subseteq E$. The set of vertices visited by C is denoted by $V(C)$. A *Hamiltonian cycle* (*tour* for short) is a cycle that visits each node of G once. We let $\mathcal{C}$ denote the family of all cycles of G which are *not* tours, and $\mathcal{T}$ denote the collection of all tours of G. Given a tour $T \in \mathcal{T}$, we let $\chi^T \in \{0, 1\}^E$ denote the *incidence vector* of T, i.e., $\chi^T_e = 1$ if $e \in T$, $\chi^T_e = 0$ if $e \notin T$. For an edge set $S \subseteq E$, let $c(E) := \sum_{e \in S} c_e$.

Throughout the paper we will adopt the following graph theoretic representation of the problem; see [11, 9, 3]. We will work on a complete undirected graph $G = (V, E)$ on n nodes $1, \ldots, n$. Any (circular) permutation on n elements $\pi_1\ \pi_2\ \ldots\ \pi_n\ \pi_1\ \ldots$ is naturally represented in G by the *Hamiltonian cycle* (*tour* for short), which visits nodes $\pi_1, \pi_2, \ldots, \pi_n$ in this order. (Clearly, there is a bijection between (circular) permutations and tours of G.) Given two tours T_1, T_2, the breakpoint distance between the associated permutations is given by $n - |T_1 \cap T_2|$. Accordingly, given q tours of

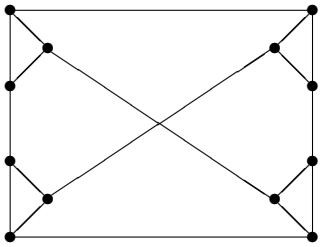


FIG. 1.1. *A graph for which the* 5/3 *ratio of the approximation algorithm is tight. The edges drawn have cost* 1/3*, whereas the edges not drawn have cost* 1.

G, say $T_1, T_2, \ldots, T_q$, BMP calls for another tour T such that $qn - \sum_{k=1}^{q} |T_k \cap T|$ is minimized.

We can express the BMP objective function by associating costs with the edges of G. In particular, for $e \in E$, the cost of an edge is defined as q minus the number of tours among $T_1, \ldots, T_q$ that contain edge e. In fact, in order to simplify the notation in the remainder of the paper, we define the *(normalized) cost* of edge e as

$$c_e := \frac{q - |\{k \in \{1, \ldots, q\} : e \in T_k\}|}{q}. \tag{1.1}$$

Clearly, BMP calls for a cheapest tour with respect to these costs, the actual value of the BMP solution being the cost of the tour multiplied by q. In what follows, we will refer to the optimal BMP solution value corresponding to these costs, keeping in mind that the actual solution value is scaled by q. Note that we have $c = 1 - \sum_{k=1}^{q} \frac{1}{q} \chi^{T_k}$. Hence, by possibly allowing some tours among $T_1, \ldots, T_q$ to be equal, the cost vectors $c \in \mathbb{R}^E$ defined by (1.1) coincide with the vectors of the form $1 - x^*$, with $x^* \in \mathbb{R}^E$ a *convex combination* of the incidence vectors of the tours of G, i.e., $x^* = \sum_{T \in \mathcal{T}} \lambda_T \chi^T$ with $\sum_{T \in \mathcal{T}} \lambda_T = 1$ and $\lambda_T \geq 0$ for $T \in \mathcal{T}$.

Given a complete undirected graph $G = (V, E)$ with edge costs c_e, $e \in E$, TSP calls for a tour of G of minimum cost. The above discussion shows that BMP is a special case of the TSP in which edges have a special cost structure. We stress that these costs *do not satisfy the triangle inequality*; i.e., the approximation results for this latter case do not extend to BMP. For instance, in the example of Figure 1.1, there are a few triples i, j, k such that $c_{ij} = c_{jk} = \frac{1}{3}$ and $c_{ik} = 1$. The most popular ILP formulation of TSP is the following (see, e.g., [6]):

$$\min \sum_{e \in E} c_e x_e, \tag{1.2}$$

$$\sum_{e \in \delta(i)} x_e = 2, \quad i \in V, \tag{1.3}$$

$$\sum_{e \in \delta(S)} x_e \geq 2, \quad S \in \mathcal{S}, \tag{1.4}$$

$$x_e \leq 1, \quad e \in E, \tag{1.5}$$

$$x_e \geq 0, \quad e \in E, \tag{1.6}$$

$$x_e \quad \text{integer}, \quad e \in E. \tag{1.7}$$

It is easy to show that constraints (1.4) are equivalent to the following:

$$\sum_{e\in E(S)} x_e \le |S| - 1, \quad S \in \mathcal{S}. \tag{1.8}$$

The convex combinations of incidence vectors of tours coincide with the vectors in the *convex hull* H of the solutions of (1.3)–(1.7), i.e., the BMP edge costs are given by $1 - x^*$ with $x^* \in H$. In fact, the results shown in this paper hold for the more general case in which costs are given by $1 - x^*$ and x^* satisfies (1.5), (1.6) and the following relaxation of (1.3) and (1.8):

$$\sum_{e\in\delta(i)} x_e \le 2, \quad i \in V, \tag{1.9}$$

$$\sum_{e\in C} x_e \le |C| - 1, \quad C \in \mathcal{C}. \tag{1.10}$$

The basic properties of the edge costs that will be used within the analysis, derived immediately from the above requirements, are the following.

PROPERTY 1. For each $e \in E$, $0 \le c_e \le 1$.

PROPERTY 2. For each $i \in V$ and $S \subseteq \delta(i)$, $c(S) \ge |S| - 2$.

PROPERTY 3. For each cycle C which is not Hamiltonian, $c(C) \ge 1$.

1.3. Additive bounding. Additive bounding [4] is a general methodology for combinatorial optimization problems. Here we will consider the main ideas of the method *only* when applied to an LP of the form

$$\begin{aligned} &\min cx, \\ Ax &\ge b, \\ Cx &\ge d, \\ x &\ge 0, \end{aligned} \tag{1.11}$$

which is typically the LP relaxation of some ILP and where the constraint set has been partitioned into two parts for convenience of illustration. The corresponding *dual LP* reads

$$\begin{aligned} &\max by + dz, \\ A^T y + C^T z &\le c, \\ y, z &\ge 0. \end{aligned} \tag{1.12}$$

The main idea of additive bounding, applied in the context of LPs, is to first *fix* $z = 0$ and determine a feasible solution $\bar{y}$ of the *restricted dual*

$$\begin{aligned} &\max by, \\ A^T y &\le c, \\ y &\ge 0, \end{aligned} \tag{1.13}$$

by some combinatorial method. Then, for the given $\bar{y}$, one finds a feasible solution $\bar{z}$ of the *residual dual*

$$\begin{aligned} &b\bar{y} + \max dz, \\ C^T z &\le c - A^T \bar{y}, \\ z &\ge 0, \end{aligned} \tag{1.14}$$

again by some combinatorial method. The final solution is $\bar{y}, \bar{z}$, yielding the lower bound $b\bar{y} + d\bar{z}$ on (1.11). Overall, additive bounding for LPs is essentially a *dual heuristic* working in two phases.

2. The approximation algorithm and its analysis. The approximation algorithm that we propose for BMP is very simple and has already been proposed for the general TSP in several contexts. We first find a 2-*matching* R of minimum cost in G; i.e., R is a set of edges of minimum cost such that every node is incident with exactly two edges in the set. If R turns out to be a tour, it is our (optimal) solution S. Otherwise, we delete the most expensive edge from every cycle in R (whose length is at least 3) and add arbitrarily chosen edges in $E \setminus R$ to get a tour S. Let c^* denote the cost of the optimal solution. The purpose of this section is to show the following result.

THEOREM 2.1. $c(S) \leq \frac{5}{3}c^*$.

The ratio is tight, as shown by the example in Figure 1.1, given by the complete graph on 12 nodes in which the edges depicted have cost $\frac{1}{3}$ and the remaining edges have cost 1. (It is not difficult to show that $1 - c$ is a convex combination of the incidence vectors of tours.) An optimal 2-matching solution is given by the four triangles and has cost 4, as does the optimal tour. After removal of an arbitrary edge from each triangle, the heuristic above may add four edges of cost 1, returning a solution of value $\frac{20}{3}$.

Clearly, $c(R)$ is a valid lower bound on c^*. The proof of Theorem 2.1 is made nontrivial by the fact that there are examples in which c^* is arbitrarily close to $2c(R)$; i.e., no approximation guarantee better than 2 can be shown by using only $c(R)$ as a lower bound.

Before giving a formal proof of Theorem 2.1, we outline the main ideas of the proof. We will apply the additive bounding ideas by considering in subsection 2.1 a suitable LP relaxation of TSP analogous to (1.2)–(1.6). Then, starting from R, we define in subsection 2.2 a heuristic solution $\bar{y}$ of a suitable restricted dual, corresponding to an LP relaxation of 2-matching. In the classical additive bounding approach, one would find a dual solution of value $c(R)$. Nevertheless, for our purposes, we will consider a solution of value $\leq c(R)$, where inequality may be strict. In subsection 2.3, we then find a convenient solution $\bar{z}$ of the residual dual problem. Letting $a(R)$ denote the lower bound associated with $\bar{y}, \bar{z}$, we will conclude the proof by showing that $c(S) \leq \frac{5}{3}\max\{c(R), a(R)\}$.

We stress that what we will do to derive lower bound $a(R)$ will be *simply* to show a feasible solution of the dual of an LP relaxation of our problem. However, the only intuitive interpretation that we can give for the derivation of this dual solution is the additive bounding framework, without which the origin of this solution would be completely unclear.

2.1. An alternative LP relaxation. Our analysis will use the following LP relaxation of the TSP, equivalent to (1.2)–(1.6), in which we will use variables x_{ij} and x_{ji} for each edge $ij \in E$. The LP relaxation is

$$\min \sum_{i \in V} \sum_{j \in V \setminus \{i\}} \tfrac{1}{2} c_{ij} x_{ij}, \tag{2.1}$$

$$x_{ij} - x_{ji} = 0, \quad i \in V,\ j \in V \setminus \{i\}, \tag{2.2}$$

$$\sum_{j \in V \setminus \{i\}} x_{ij} = 2, \quad i \in V, \tag{2.3}$$

$$\sum_{i \in S} \sum_{j \in V \setminus S} x_{ij} \geq 2, \quad S \in \mathcal{S}, \tag{2.4}$$

$$x_{ij} \leq 1, \quad i \in V,\ j \in V \setminus \{i\}, \tag{2.5}$$

$$x_{ij} \geq 0, \quad i \in V,\, j \in V \setminus \{i\}. \tag{2.6}$$

Graphically, one may imagine that each edge $ij \in E$ is replaced by two antiparallel arcs (i,j) and (j,i), with associated binary variables x_{ij} and x_{ji}. Constraints (2.2) require that if an arc is selected in the solution, so is its antiparallel counterpart. Moreover, equations (2.3) require that exactly two arcs in the solution leave each node $i \in V$, and inequalities (2.4) that at least two arcs in the solution go from nodes in S to nodes in $V \setminus S$ for each $S \in \mathcal{S}$.

In our analysis, we will *never* consider constraints (2.2) (fixing the associated dual variable to 0). After removal of these constraints, the dual reads

$$\begin{array}{rrcl} \max \displaystyle\sum_{i\in V} 2y_i + \sum_{S\in\mathcal{S}} 2w_S - \sum_{i\in V}\sum_{j\in V\setminus\{i\}} u_{ij}, & & & \\ y_i + \displaystyle\sum_{S\in\mathcal{S}: i\in S, j\in V\setminus S} w_S - u_{ij} & \leq & \frac{1}{2}c_{ij}, & i \in V,\, j \in V\setminus\{i\}, \\ w_S & \geq & 0, & S \in \mathcal{S}, \\ u_{ij} & \geq & 0, & i \in V,\, j \in V\setminus\{i\}. \end{array} \tag{2.7}$$

2.2. The restricted dual. In the restricted dual, defined from (2.7), we will fix $w_S = 0$ for $S \in \mathcal{S}$, obtaining

$$\begin{array}{rrcl} \max \sum_{i\in V} 2y_i - \displaystyle\sum_{i\in V}\sum_{j\in V\setminus\{i\}} u_{ij}, & & & \\ y_i - u_{ij} & \leq & \frac{1}{2}c_{ij}, & i \in V,\, j \in V\setminus\{i\}, \\ u_{ij} & \geq & 0, & i \in V,\, j \in V\setminus\{i\}. \end{array} \tag{2.8}$$

Note that, with constraints (2.2) and (2.4) removed (fixing w_S to 0 is equivalent to removing (2.4)), LP relaxation (2.1)–(2.6) is also a relaxation of the 2-matching problem. Here, our aim is to define a solution of (2.8) that has a convenient expression and whose value is "close" to $c(R)$.

DEFINITION 2.2. *For each node $i \in V$, let*

$$\bar{y}_i := \frac{c_e}{2}, \tag{2.9}$$

where e is the most expensive edge in R incident with i,

$$\bar{u}_{ij} := \max\{0, \bar{y}_i - \tfrac{1}{2}c_{ij}\} \quad \textit{for } j \in V \setminus \{i\}, \tag{2.10}$$

and

$$r(i) := 2\bar{y}_i - \sum_{j\in V\setminus\{i\}} \bar{u}_{ij}. \tag{2.11}$$

It is immediate to check that the solution $\bar{y}, \bar{u}$ defined by (2.9) and (2.10) is feasible for (2.8) and has value $\sum_{i\in V} r(i)$. We now compute a convenient lower bound on $r(i)$.

LEMMA 2.3. *For each node $i \in V$ let ij, ik be the two edges in R incident with i such that $c_{ij} \geq c_{ik}$. Then,*

$$r(i) \geq \begin{cases} \dfrac{c_{ij} + c_{ik}}{2} & \textit{if } 2c_{ij} + c_{ik} < 1, \\ \dfrac{1 - c_{ij}}{2} & \textit{if } 2c_{ij} + c_{ik} \geq 1. \end{cases} \tag{2.12}$$

Proof. Note that $\bar{u}_{ij} = 0$ and $\bar{u}_{ik} = \frac{c_{ij}-c_{ik}}{2}$. If $2c_{ij} + c_{ik} < 1$, by Property 2, the cost of each edge $it \notin R$ is $c_{it} \geq 1 - c_{ij} - c_{ik} > c_{ij}$. Hence, $\bar{u}_{it} = \max\{0, \bar{y}_i - \frac{c_{it}}{2}\} = 0$ for all $t \neq j, k$, implying

$$r(i) = 2 \cdot \frac{c_{ij}}{2} - \frac{c_{ij} - c_{ik}}{2} = \frac{c_{ij} + c_{ik}}{2}.$$

Consider now the case $2c_{ij} + c_{ik} \geq 1$. In this case, there may be one or more edges $it \notin R$ such that $\bar{u}_{it} > 0$. Let S be the subset of these edges, i.e., $S := \{it : it \notin R$ and $\bar{u}_{it} > 0\}$ and $s := |S|$. From (2.10) we have

$$\begin{aligned} r(i) = 2\bar{y}_i - \bar{u}_{ik} - \sum_{it \in S} \bar{u}_{it} &= 2 \cdot \frac{c_{ij}}{2} - \frac{c_{ij} - c_{ik}}{2} - \sum_{it \in S} \frac{c_{ij} - c_{jt}}{2} \\ &= \frac{1}{2}\left(c_{ik} - (s-1)c_{ij} + \sum_{it \in S} c_{it}\right). \end{aligned}$$

If $s = 0$, we have

$$\frac{1}{2}\left(c_{ik} - (s-1)c_{ij} + \sum_{it \in S} c_{it}\right) = \frac{c_{ij} + c_{ik}}{2} \geq \frac{1 - c_{ij}}{2}$$

as $2c_{ij} + c_{ik} \geq 1$. Otherwise, $s \geq 1$, and noting that $ij, ik \notin S$ since $ij, ik \in R$, Property 2 implies

$$\sum_{it \in S \cup \{ij,ik\}} c_{it} \leq |S \cup \{ij, ik\}| - 2 = s,$$

i.e., $\sum_{it \in S} c_{it} \geq s - c_{ij} - c_{ik}$. Therefore,

$$\frac{1}{2}\left(c_{ik} - (s-1)c_{ij} + \sum_{it \in S} c_{it}\right) \geq \frac{s(1 - c_{ij})}{2} \geq \frac{1 - c_{ij}}{2}. \qquad \square$$

2.3. The residual dual and the worst-case ratio. Consider the dual solution $\bar{y}, \bar{u}$ given in Definition 2.2. Keeping these values fixed, we have the following residual dual:

$$\begin{aligned} &\sum_{i \in V} 2\bar{y}_i - \sum_{i \in V} \sum_{j \in V \setminus \{i\}} \bar{u}_{ij} + \max \sum_{S \in \mathcal{S}} 2w_S, \\ &\sum_{S \in \mathcal{S}: i \in S, j \in V \setminus S} w_S \leq \frac{1}{2}c_{ij} - \bar{y}_i + \bar{u}_{ij}, \quad i \in V,\, j \in V \setminus \{i\}, \\ &w_S \geq 0, \quad S \in \mathcal{S}. \end{aligned} \tag{2.13}$$

Let $C_1, C_2, \ldots, C_p$ be the cycles in R. For $\ell = 1, \ldots, p$ define $\beta_\ell := \sum_{e \in C_\ell} c_e$ and $\gamma_\ell := \max_{e \in C_\ell} c_e$. Note that $\beta_\ell \geq 1$ by Property 3. We have the following.

LEMMA 2.4. *For $\ell = 1, \ldots, p$, $i \in V(C_\ell)$, and $j \in V \setminus V(C_\ell)$,*

$$\tfrac{1}{2}c_{ij} - \bar{y}_i + \bar{u}_{ij} \geq \begin{cases} 0 & \text{if } \gamma_\ell \geq \frac{1}{3}, \\ \dfrac{1 - 3\gamma_\ell}{2} & \text{if } \gamma_\ell < \frac{1}{3}. \end{cases} \tag{2.14}$$

Proof. Nonnegativity follows immediately from dual feasibility. Moreover, if $\gamma_\ell < \frac{1}{3}$, then $\bar{y}_i \le \frac{1}{2}\gamma_\ell$, $c_{ij} \ge 1 - 2\gamma_\ell$ by Property 2, and $\bar{y}_i - \frac{1}{2}c_{ij} \le \frac{3}{2}\gamma_\ell - \frac{1}{2} < 0$, implying $\bar{u}_{ij} = 0$ by (2.10). ∎

We can therefore define the following feasible solution of (2.13).

DEFINITION 2.5. *For $S \in \mathcal{S}$, define*

$$\bar{w}_S := \begin{cases} \dfrac{1-3\gamma_\ell}{2} & \text{if } S = V(C_\ell) \text{ for some } \ell \in \{1,\dots,p\} \text{ such that } \gamma_\ell < \frac{1}{3}, \\ 0, & \text{otherwise.} \end{cases} \tag{2.15}$$

The lower bound associated with the dual solution $\bar{y}, \bar{w}, \bar{u}$ of (2.7) defined by (2.9), (2.10), and (2.15) is given by (recalling the definition of $r(i)$)

$$a(R) := \sum_{S\in\mathcal{S}} 2\bar{w}_S + \sum_{i\in V} 2\bar{y}_i - \sum_{(i,j)\in A} \bar{u}_{ij} = \sum_{\ell=1}^{p} \left(2\bar{w}_{V(C_\ell)} + \sum_{i\in V(C_\ell)} r(i) \right). \tag{2.16}$$

We now derive a convenient lower bound on $2\bar{w}_{V(C_\ell)} + \sum_{i\in V(C_\ell)} r(i)$ as a function of β_ℓ and γ_ℓ, $\ell = 1,\dots,p$.

LEMMA 2.6. *For $\ell = 1,\dots,p$,*

$$2\bar{w}_{V(C_\ell)} + \sum_{i\in V(C_\ell)} r(i) \ge \frac{3}{2}(1-\gamma_\ell). \tag{2.17}$$

Proof. As before, for a given node $i \in V(C_\ell)$, we let ij and ik denote the two edges in R incident with i such that $c_{ij} \ge c_{ik}$.

If $\gamma_\ell < \frac{1}{3}$, we have $c_e < \frac{1}{3}$ for all $e \in C_\ell$, i.e., $2c_e + c_f < 1$ for all $e, f \in C_\ell$. By Lemma 2.3, this implies $r(i) \ge \frac{c_{ij}+c_{ik}}{2}$ for all $i \in V(C_\ell)$, i.e., $\sum_{i\in V(C_\ell)} r(i) = \sum_{e\in C_\ell} c_e \ge \beta_\ell$. Moreover, $2\bar{w}_{V(C_\ell)} = 1 - 3\gamma_\ell$, i.e.,

$$2\bar{w}_{V(C_\ell)} + \sum_{i\in V(C_\ell)} r(i) \ge \beta_\ell + 1 - 3\gamma_\ell \ge \frac{3}{2}(1-\gamma_\ell),$$

as $\gamma_\ell < \frac{1}{3}$ and $\beta_\ell \ge 1$. For the rest of the proof, we will consider the case $\gamma_\ell \ge \frac{1}{3}$.

For the given i, we partition $V(C_\ell)$ into $V_1 := \{i \in V(C_\ell) : r(i) = \frac{c_{ij}+c_{ik}}{2}\}$ and $V_2 := V(C_\ell) \setminus V_1$. Note that $\frac{1-\gamma_\ell}{2} \le \frac{1-c_e}{2}$ for all $e \in C_\ell$. Hence, if $|V_2| \ge 3$,

$$\sum_{i\in V(C_\ell)} r(i) \ge \sum_{i\in V_2} r(i) \ge \sum_{i\in V_2} \frac{1-c_{ij}}{2} \ge |V_2| \frac{1-\gamma_\ell}{2} \ge \frac{3}{2}(1-\gamma_\ell).$$

On the other hand, if $|V_2| = 0$, we have (as in the case of $\gamma_\ell < \frac{1}{3}$) $\sum_{i\in V(C_\ell)} r(i) \ge \beta_\ell \ge 1$, whereas $\frac{3}{2}(1-\gamma_\ell) \le 1$ as $\gamma_\ell \ge \frac{1}{3}$. Hence, the two cases left are $|V_2| = 2$ and $|V_2| = 1$. If $|V_2| = 1$, say $V_2 = \{i_1\}$, where i_1j_1 and i_1k_1 are the two edges in R incident with i_1 such that $c_{i_1j_1} \ge c_{i_1k_1}$, we have

$$\begin{aligned} \sum_{i\in V(C_\ell)} r(i) &\ge \left(\sum_{i\in V(C_\ell)\setminus\{i_1\}} \frac{c_{ij}+c_{ik}}{2} \right) + \frac{1-c_{i_1j_1}}{2} \\ &= \beta_\ell - \frac{c_{i_1j_1}+c_{i_1k_1}}{2} + \frac{1-c_{i_1j_1}}{2} \ge \frac{3}{2}(1-\gamma_\ell), \end{aligned}$$

as $\frac{1-c_{i_1j_1}}{2} \geq \frac{1-\gamma_\ell}{2}$ and $\beta_\ell - \frac{c_{i_1j_1}+c_{i_1k_1}}{2} \geq 1-\gamma_\ell$. Finally, if $|V_2| = 2$, say $V_2 = \{i_1, i_2\}$, where i_1j_1, i_1k_1 and i_2j_2, i_2k_2 are the edges in R incident, respectively, with i_1 and i_2 such that $c_{i_1j_1} \geq c_{i_1k_1}$ and $c_{i_2j_2} \geq c_{i_2k_2}$, we have

$$\sum_{i \in V(C_\ell)} r(i) = \beta_\ell - \frac{c_{i_1j_1} + c_{i_1k_1}}{2} - \frac{c_{i_2j_2} + c_{i_2k_2}}{2} + \frac{1 - c_{i_1j_1}}{2} + \frac{1 - c_{i_2j_2}}{2} \geq \frac{3}{2}(1 - \gamma_\ell),$$

as $\frac{1-c_{i_1j_1}}{2} \geq \frac{1-\gamma_\ell}{2}$, $\frac{1-c_{i_2j_2}}{2} \geq \frac{1-\gamma_\ell}{2}$, and $\beta_\ell - \frac{c_{i_1j_1}+c_{i_1k_1}}{2} - \frac{c_{i_2j_2}+c_{i_2j_2}}{2} \geq \frac{1-\gamma_\ell}{2}$, since the latter is equivalent to $2\beta_\ell \geq (c_{i_1j_1} + c_{i_1k_1} + c_{i_2j_2} + c_{i_2k_2} - \gamma_\ell) + 1$, implied by $\beta_\ell \geq 1$ and $\beta_\ell \geq c_{i_1j_1} + c_{i_1k_1} + c_{i_2j_2} + c_{i_2k_2} - \gamma_\ell$, which holds since $\gamma_\ell := \max_{e \in C_\ell} c_e$, $|V(C_\ell)| \geq 3$, and $|\{i_1j_1, i_1k_1, i_2j_2, i_2k_2\}| \geq 3$. □

From (2.16) and (2.17) we have $a(R) \geq \sum_{\ell=1}^{p} \frac{3}{2}(1 - \gamma_\ell)$, implying

$$c^* \geq \sum_{\ell=1}^{p} \tfrac{3}{2}(1 - \gamma_\ell).$$

Combining this with

$$c^* \geq c(R) = \sum_{\ell=1}^{p} \beta_\ell$$

and with

$$c(S) \leq \sum_{\ell=1}^{p} (\beta_\ell - \gamma_\ell + 1),$$

where S is the heuristic solution found by our approximation algorithm, we get

$$c(S) \leq \sum_{\ell=1}^{p} \beta_\ell + \sum_{\ell=1}^{p} (1 - \gamma_\ell) \leq c^* + \tfrac{2}{3}c^* \leq \tfrac{5}{3}c^*,$$

proving Theorem 2.1.

3. The signed BMP. In this section we consider the case of BMP arising when the orientation of the genes within the given genomes is known. We will use an approach similar to the one in the previous section to achieve an approximation ratio of $\frac{3}{2}$. As the proofs are absolutely analogous to those of the previous section, in this section we will state only the main lemmas without giving proofs.

When gene orientation is known, genomes are represented by *(circular) signed permutations* of the numbers from 1 to n, i.e., cyclic sequences $\pi_1\ \pi_2\ \ldots\ \pi_n\ \pi_1\ \ldots$ where, for each $j = 1, \ldots, n$, there exists an i such that $\pi_i = \pm j$. A *breakpoint* of π^1 with respect to π^2 is now a pair $\pi^1_i\ \pi^1_{i+1}$ of consecutive elements in π^1 that are not consecutive in π^2, i.e., such that neither $\pi^1_i\ \pi^1_{i+1}$ nor $-\pi^1_{i+1}\ -\pi^1_i$ appear consecutively in this order in π^2.

The graph representing the problem now has $2n$ nodes $1, \ldots, 2n$. Following the terminology in [9], the *base (perfect) matching* B is given by edge set $\{(2i-1, 2i) : i = 1, \ldots, n\}$. A *Hamiltonian (perfect) matching* of G is a perfect matching M such that $M \cup B$ is a tour of G. Any signed permutation on n elements corresponds to a Hamiltonian matching and vice versa. In particular, given a signed permutation $\pi_1\ \pi_2\ \ldots\ \pi_n\ \pi_1\ \ldots$, the associated Hamiltonian matching is defined by

$$M := \{(2|\pi_i| - \nu(\pi_i), 2|\pi_{i+1}| - 1 + \nu(\pi_{i+1})) : i \in \{1, \ldots, n\}\},$$

where the indices of the vertices must be taken modulo $2n$ and $\nu(\pi_i) := 0$ if $\pi_i \geq 0$, $\nu(\pi_i) := 1$ if $\pi_i < 0$. This is a bijection [9, 3], i.e., its inverse yields the signed permutation associated with a given Hamiltonian matching. Now, given two permutations π^1, π^2 and the associated Hamiltonian matchings M_1, M_2, it is easy to check that $\pi^1_i\ \pi^1_{i+1}$ is a breakpoint of π^1 with respect to π^2 if and only if edge $e = (2|\pi_i| - \nu(\pi_i), 2|\pi_{i+1}| - 1 + \nu(\pi_{i+1})) \notin M_2$ (noting that $e \in M_1$); for further details, see [9, 3]. Hence, given two Hamiltonian matchings M_1, M_2, the breakpoint distance between the associated permutations is given by $n - |M_1 \cap M_2|$. Accordingly, given q Hamiltonian matchings of G, say $M_1, M_2, \ldots, M_q$, BMP calls for a Hamiltonian matching M such that $qn - \sum_{k=1}^{q} |M_k \cap M|$ is minimized.

Let $E = \{ij : 1 \leq i, j \leq 2n, i \neq j\} \setminus B$ be the set of edges which are contained in some Hamiltonian matching. For $e \in E$, the (normalized) cost of e is defined by

$$c_e := \frac{q - |\{k \in \{1, \ldots, q\} : e \in M_k\}|}{q},$$

and BMP calls for the cheapest Hamiltonian matching.

The counterpart of a cycle that is not a tour in the previous section is now called a *half-cycle*, which is a matching of G that contains strictly fewer than n edges (i.e., it is not perfect) and forms a cycle with some of the edges of the base matching B—formally, a half-cycle is an edge set $H \subset E$ for which there exists $A \subseteq B$ with $|A| = |H|$ such that $H \cup A$ is a cycle. Note that each perfect matching $P \subset E$ that is not Hamiltonian is partitioned into half-cycles, each corresponding to a cycle of $P \cup B$.

The approximation algorithm finds a perfect matching of minimum cost in G, say R, removes from R the most expensive edge in each half-cycle, and adds edges in an arbitrary way so as to get a Hamiltonian matching S. The result illustrated in this section is the following theorem.

THEOREM 3.1. $c(S) \leq \frac{3}{2} c^*$.

The properties of the edge costs that we will use within this section are Property 1 and the following.

PROPERTY 4. For each $i \in V$ and $S \subseteq \delta(i)$, $c(S) \geq |S| - 1$.

PROPERTY 5. For each half-cycle H, $c(H) \geq 1$.

Again, $c(R)$ is a valid lower bound on the cost of the optimal BMP solution, and the optimal solution value c^* can be as large as $2c(R)$. In order to get an alternative lower bound, we use the following LP relaxation. Let $\mathcal{S}$ now denote the collection of nontrivial subsets $S \subset V$ such that no edge in B has one endpoint in S and the other in $V \setminus S$,

$$\min \sum_{i \in V} \sum_{j \in V \setminus \{i\} : ij \in E} \tfrac{1}{2} c_{ij} x_{ij}, \tag{3.1}$$

$$x_{ij} - x_{ji} = 0, \quad i \in V,\ j \in V \setminus \{i\} : ij \in E, \tag{3.2}$$

$$\sum_{j \in V \setminus \{i\} : ij \in E} x_{ij} = 1, \quad i \in V, \tag{3.3}$$

$$\sum_{i \in S} \sum_{j \in V \setminus S} x_{ij} \geq 2, \quad S \in \mathcal{S}, \tag{3.4}$$

$$x_{ij} \geq 0, \quad i \in V,\ j \in V \setminus \{i\} : ij \in E, \tag{3.5}$$

whose dual, after removal of constraints (3.2), reads

$$\begin{array}{rrcl} & \max \sum_{i\in V} y_i, & & \\ y_i + \sum_{S\in\mathcal{S}: i\in S, j\in V\setminus S} w_S & \leq & \frac{1}{2}c_{ij}, & i\in V,\ j\in V\setminus\{i\} : ij\in E, \\ w_S & \geq & 0, & S\in\mathcal{S}. \end{array} \tag{3.6}$$

The restricted dual is obtained by fixing $w_S = 0$ for $S \in \mathcal{S}$, and its feasible solution is given in the folowing definition.

DEFINITION 3.2. *For each node $i \in V$, let*

$$\bar{y}_i := \begin{cases} \dfrac{c_e}{2} & \text{if } c_e < \frac{1}{2}, \\ \dfrac{1-c_e}{2} & \text{if } c_e \geq \frac{1}{2}, \end{cases} \tag{3.7}$$

where e is the edge in R incident with i and

$$r(i) := \bar{y}_i. \tag{3.8}$$

It is not difficult to check that $\bar{y}$ is a feasible solution of (3.6) (with w_S fixed to 0). The following lemma is a restatement of the above definition and is given only to show the analogy to Lemma 2.3.

LEMMA 3.3. *For each node $i \in V$, let ij be the edge in R incident with i. Then,*

$$r(i) \geq \begin{cases} \dfrac{c_{ij}}{2} & \text{if } c_{ij} < \frac{1}{2}, \\ \dfrac{1-c_{ij}}{2} & \text{if } c_{ij} \geq \frac{1}{2}. \end{cases} \tag{3.9}$$

Consider the set of half-cycles in which R is partitioned, say $H_1, \ldots, H_p$. For $\ell = 1, \ldots, p$ let $V(H_\ell)$ denote the set of nodes visited by cycle C_ℓ, and define $\beta_\ell := \sum_{e\in H_\ell} c_e$ and $\gamma_\ell := \max_{e\in H_\ell} c_e$. Note that $\beta_\ell \geq 1$ by Property 5. We have the following.

LEMMA 3.4. *For $\ell = 1, \ldots, p$, $i \in V(H_\ell)$, and $j \in V \setminus V(H_\ell)$,*

$$\tfrac{1}{2}c_{ij} - \bar{y}_i \geq \begin{cases} 0 & \text{if } \gamma_\ell \geq \frac{1}{2}, \\ \dfrac{1-2\gamma_\ell}{2} & \text{if } \gamma_\ell < \frac{1}{2}. \end{cases} \tag{3.10}$$

Accordingly, the feasible solution of the restricted dual is the following definition.

DEFINITION 3.5. *For $S \in \mathcal{S}$, define*

$$\bar{w}_S := \begin{cases} \dfrac{1-2\gamma_\ell}{2} & \text{if } S = V(H_\ell) \text{ for some } \ell \in \{1,\ldots,p\} \text{ such that } \gamma_\ell < \frac{1}{2}, \\ 0, & \text{otherwise.} \end{cases} \tag{3.11}$$

Hence, our lower bound is

$$a(R) := \sum_{\ell=1}^{p} \left(2\bar{w}_{V(H_\ell)} + \sum_{i\in V(H_\ell)} r(i) \right), \tag{3.12}$$

and we have the following result.

LEMMA 3.6. *For* $\ell = 1, \ldots, p$,

$$2\bar{w}_{V(H_\ell)} + \sum_{i \in V(H_\ell)} r(i) \geq 2(1 - \gamma_\ell). \tag{3.13}$$

The proof of Theorem 3.1 follows by $c^* \geq a(R) \geq \sum_{\ell=1}^{p} 2(1-\gamma_\ell)$, $c^* \geq c(R) = \sum_{\ell=1}^{p} \beta_\ell$, and $c(S) \leq \sum_{\ell=1}^{p} (\beta_\ell + 1 - \gamma_\ell)$.

Acknowledgments. I am grateful to Paola Bonizzoni, Itzik Pe'er, and Luca Trevisan for helpful discussions on the subject. Moreover, I would like to thank an anonymous referee for his/her helpful comments.

REFERENCES

[1] M. BLANCHETTE, G. BOURQUE, AND D. SANKOFF, *Breakpoint phylogenies*, in Proceedings of Genome Informatics 1997, S. Miyano and T. Takagi, eds., Universal Academy Press, Tokyo, 1997, pp. 25–34.

[2] D. BRYANT, *The complexity of the breakpoint median problem*, J. Comput. Biol., to appear.

[3] A. CAPRARA, *Formulations and hardness of multiple sorting by reversals*, in Proceedings of the Third Annual International Conference on Computational Molecular Biology (RECOMB'99), ACM Press, New York, 1999, pp. 84–93.

[4] M. FISCHETTI AND P. TOTH, *An additive bounding procedure for combinatorial optimization problems*, Oper. Res., 37 (1989), pp. 319–328.

[5] D. GUSFIELD, *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*, Cambridge University Press, Cambridge, UK, 1997.

[6] M. JÜNGER, G. REINELT, AND G. RINALDI, *The traveling salesman problem*, in Network Models, M. Ball, T. Magnanti, C. Monma, and G. Nemhauser, eds., Handb. Oper. Res. Management Sci. 7, Elsevier, New York, 1995, pp. 225–330.

[7] B.M.E. MORET, L.-S. WANG, T. WARNOW, AND S.K. WYMAN, *New approaches to phylogeny reconstruction from gene order data*, Bioinformatics, 17 (2001), pp. 165S–173S.

[8] B.M.E. MORET, S.K. WYMAN, D.A. BADER, T. WARNOW, AND M. YAN, *A new implementation and detailed study of breakpoint analysis*, in Proceedings of the Sixth Pacific Symposium on Biocomputing (PSB 2001), World Scientific, River Edge, NJ, 2001, pp. 583–594.

[9] I. PE'ER AND R. SHAMIR, *The Median Problems for Breakpoints are $\mathcal{NP}$-Complete*, ECCC Report 71, University of Trier, Trier, Germany, 1998; available at http://www.eccc.uni-trier.de/.

[10] I. PE'ER AND R. SHAMIR, *Approximation algorithms for the median problem in the breakpoint model*, in Comparative Genomics, D. Sankoff and J.H. Nadeau, eds., Kluwer, Dordrecht, The Netherlands, 2000, pp. 225–241.

[11] D. SANKOFF AND M. BLANCHETTE, *The median problem for breakpoints in comparative genomics*, in Computing and Combinatorics, Proceedings of COCOON '97, T. Jiang and D.T. Lee, eds., Lecture Notes in Comput. Sci. 1276, Springer-Verlag, New York, 1997, pp. 251–263.

[12] D. SANKOFF AND M. BLANCHETTE, *Multiple genome rearrangement and breakpoint phylogeny*, J. Comput. Biol., 5 (1998), pp. 555–570.

[13] D. SANKOFF, D. BRYANT, M. DENAULT, B.F. LANG, G. BURGER, *Early eukaryote evolution based on mitochondrial gene order breakpoints*, J. Comput. Biol., 7 (2000), pp. 521–536.

[14] D. SANKOFF, G. SUNDARAM AND J. KECECIOGLU, *Steiner points in the space of genome rearrangements*, Internat. J. Found. Comput. Sci., 7 (1996), pp. 1–9.

[15] L. TREVISAN, *personal communication*, 2000.

SIAM J. OPTIM.
Vol. 13, No. 2, pp. 520–534

SUBDIFFERENTIAL CONDITIONS FOR CALMNESS OF CONVEX CONSTRAINTS*

R. HENRION† AND A. JOURANI‡

Abstract. We study subdifferential conditions of the calmness property for multifunctions representing convex constraint systems in a Banach space. Extending earlier work in finite dimensions [R. Henrion and J. Outrata, *J. Math. Anal. Appl.*, 258 (2001), pp. 110–130], we show that, in contrast to the stronger Aubin property of a multifunction (or metric regularity of its inverse), calmness can be ensured by corresponding weaker constraint qualifications, which are based only on boundaries of subdifferentials and normal cones rather than on the full objects. Most of the results can be immediately interpreted in the context of error bounds.

Key words. calmness, multifunctions, convex systems, error bounds, constraint qualifications

AMS subject classifications. 90C31, 26E25, 49J52

PII. S1052623401386071

1. Introduction. Following [24, p. 399], a multifunction $M : Y \rightrightarrows X$ between metric spaces X, Y is *calm* at some point $(\bar{y}, \bar{x})$ of its graph if there exist neighborhoods $\mathcal{V}, \mathcal{U}$ of $\bar{y}, \bar{x}$, respectively, and some $L > 0$ such that the corresponding distance functions satisfy

$$d(x, M(\bar{y})) \leq L d(y, \bar{y}) \qquad \forall x \in M(y) \cap \mathcal{U},\ \forall y \in \mathcal{V}. \tag{1.1}$$

With $\mathcal{U} := X$, calmness reduces to the upper Lipschitz property of multifunctions, introduced by Robinson [23]. Obviously, calmness is also weaker than the well-known Aubin property of multifunctions

$$d(x, M(y')) \leq L d(y, y') \qquad \forall x \in M(y) \cap \mathcal{U},\ \forall y, y' \in \mathcal{V}. \tag{1.2}$$

(In particular, $M(y) = \emptyset$ for y close to but different from $\bar{y}$ is possible under calmness but violates the Aubin property.) Calmness plays a key role in many issues of mathematical programming like optimality conditions, error bounds, or stability of solutions. The focus of this paper will be on multifunctions defined by convex systems in a Banach space X like

$$M(y) := \{x \in C \mid f(x) \leq y\} \quad \text{or } M(y) := \{x \in X \mid d(x, C) + d(x, D) \leq y\} \quad (y \in \mathbb{R}), \tag{1.3}$$

where $C, D \subseteq X$ are closed, convex subsets and f is convex. Of course, writing down the calmness property (1.1) for the first system considered in (1.3) immediately yields the existence of a local error bound for f with respect to the set C. Hence all results obtained for this first part have an immediate link to the context of error bounds, which are extensively studied in the literature (e.g., [5], [15], [14], [16], [17], [19]). The aim of this paper is to derive dual (i.e., formulated in terms of the subdifferential and normal cone) conditions for calmness of the systems in (1.3) which are weaker than

*Received by the editors March 2, 2001; accepted for publication (in revised form) March 16, 2002; published electronically October 1, 2002.
http://www.siam.org/journals/siopt/13-2/38607.html
†Weierstrass Institute, Mohrenstrasse 39, 10117 Berlin, Germany (henrion@wias-berlin.de).
‡Université de Bourgogne, Département de Mathématiques, Analyse Appliquée et Optimisation, BP 47870, 21078 Dijon, France (jourani@u-bourgogne.fr).

the well-known Slater-type conditions implying the stronger Aubin property (1.2) of M (or, equivalently, the metric regularity of M^{-1}).

If M is a polyhedral multifunction, then it is automatically calm (see [23]). Apart from this special class, certain conditions have to hold true in order to ensure calmness, and it seems natural to characterize these conditions in terms of well-known objects from nonsmooth analysis such as (co-) derivatives, (sub-) differentials, or tangent or normal cones. Similar characterizations have been successfully established for the stronger Aubin property. In finite dimensions, for instance, (1.2) is equivalent to each of the following two conditions, described by Mordukhovich [18] and Aubin (see, e.g., [1] and [6, Corollary 1.19] for necessity), respectively:

$$D^*M(\bar{y},\bar{x})(0) = \{0\}, \tag{1.4}$$

$$\exists \alpha, \beta > 0 : \; B(0,1) \subseteq D_- M^{-1}(x,y)(B(0,\alpha)) \quad \forall x, y \in \operatorname{Gph} M \cap B((\bar{x},\bar{y}),\beta). \tag{1.5}$$

Here, D^* and D_- refer to Mordukhovich's coderivative and to Aubin's contingent derivative, respectively, while B refers to appropriate closed balls. As coderivatives relate to normal cones while derivatives are associated with tangent cones, the first criterion above is of dual nature and the second one is of primal nature. The question arises of whether the criteria above can be modified appropriately to characterize the weaker calmness property (1.1) rather than (1.2). A primal criterion of calmness was found in [9, Proposition 2.1] (sufficiency) and [10, Proposition 4.1] (necessity):

$$DM(\bar{y},\bar{x})(0) = \{0\}. \tag{1.6}$$

Note that (1.6) immediately enforces the isolatedness of $\bar{x}$ in $M(\bar{y})$ because a sequence $x_n \to \bar{x}$, $x_n \in M(\bar{y})$, $x_n \neq \bar{x}$ would generate a nontrivial tangent vector $(0,\xi)$ to $\operatorname{Gph} M$ at $(\bar{y},\bar{x})$, whence a contradiction $0 \neq \xi \in DM(\bar{y},\bar{x})(0)$ to (1.6). Consequently, a reduced version of calmness (also called calmness on selections) is equivalently characterized by (1.6). A dual characterization of calmness in the broader sense of (1.1) was given in [7] for the special case of finite-dimensional multifunctions

$$M(y) := \{x \in C | g(x) + y \in D\},$$

where $C \subseteq \mathbb{R}^p, D \subseteq \mathbb{R}^m$ are closed subsets and $g : \mathbb{R}^p \to \mathbb{R}^m$ is locally Lipschitz. In this special case, Mordukhovich's criterion (1.4) for the Aubin property takes the form

$$\bigcup_{y^* \in N_D(g(\bar{x})) \setminus \{0\}} D^*g(\bar{x})(y^*) \cap (-N_C(\bar{x})) = \emptyset,$$

where N refers to Mordukhovich's normal cone. It was shown in [7] that under mild assumptions, calmness is implied by the weaker condition

$$\bigcup_{y^* \in N_D(g(\bar{x})) \setminus \{0\}} D^*g(\bar{x})(y^*) \cap (-\operatorname{bd} N_C(\bar{x})) = \emptyset,$$

where "bd" refers to the topological boundary. Hence, reducing Lipschitzian stability to upper Lipschitzian stability corresponds to a transition from certain geometric objects to their boundaries. This fact becomes most evident for the simple case of one single inequality $g(x) + y \leq 0$ (i.e., $D = \mathbb{R}_-$): if g (as a function) and C (as a set) are regular in the sense of Clarke, then calmness of M holds true at some point

$(0, \bar{x})$ with $g(\bar{x}) = 0$, provided that $\operatorname{bd} \partial g(\bar{x}) \cap (-\operatorname{bd} N_C(\bar{x})) = \emptyset$ (see Theorem 4.2 in [7]). Here, "∂" refers to either Mordukhovich's or Clarke's subdifferential (which coincide due to regularity). This last constraint qualification can be opposed again to the corresponding criterion of the Aubin property, which now takes the form $\partial g(\bar{x}) \cap (-N_C(\bar{x})) = \emptyset$. For absent abstract constraints ($C = \mathbb{R}^p$) the calmness condition reduces to $0 \notin \operatorname{bd} \partial g(\bar{x})$ (as opposed to the condition $0 \notin \partial g(\bar{x})$, which ensures the stronger Aubin property).

The aim of this paper is to study possible infinite-dimensional extensions of the previous results. For the first system in (1.3), a counterexample will show that, even for Clarke-regular data, the mentioned constraint qualification $\operatorname{bd} \partial g(\bar{x}) \cap (-\operatorname{bd} N_C(\bar{x})) = \emptyset$ no longer implies calmness in a Banach space setting. It does, however, for convex data, and in this case it can even be weakened again. This gives an improvement even for the finite-dimensional case. Therefore, the focus of the paper is on convex constraint systems.

2. Notation. Throughout this paper, X will denote some Banach space, and X^* its dual endowed with the strong topology. In these spaces, $B(\alpha, \beta)$ and $B^*(\alpha, \beta)$ are the closed balls around α with radius β, whereas $B^0(\alpha, \beta)$ refers to the corresponding open ball in X. By i_S we denote the indicator function of a closed set $S \subseteq X$, and by $\operatorname{epi} f$ the epigraph of some function $f : X \to \mathbb{R} \cup \{\infty\}$. $N(S; x), \partial f$, and $\partial^\infty f$ refer to the normal cone to S at some $x \in S$ and to the usual and singular subdifferentials of f, respectively, all in the sense of convex analysis. In contrast, ∂^c represents Clarke's subdifferential. "bd" and "int" are the topological boundary and interior. For a multifunction $M : X \rightrightarrows Y$ between Banach spaces,

$$\begin{aligned} \operatorname{Gph} M &= \{(x, y) \in X \times Y | y \in M(x)\}, \\ \operatorname{range} M &= \{y \in Y | \, \exists x \in X, y \in M(x)\}, \\ M^{-1} &: Y \rightrightarrows X, \qquad M^{-1}(y) = \{x \in X | y \in M(x)\} \end{aligned}$$

denote its graph, its range, and its inverse, respectively.

3. Convex constraint systems with a perturbed inequality. In this section, we consider constraint systems involving a fixed abstract constraint set and an inequality which is subject to perturbations. More precisely, we are interested in the calmness property (1.1) of the multifunction

$$M(y) := \{x \in C \mid f(x) \leq y\} \quad (y \in \mathbb{R}), \tag{3.1}$$

where C is a closed, convex subset of some Banach space X and f is a convex, lower semicontinuous function. First, we state an auxiliary result. Recall from [2] that a set $S \subseteq X$ is *compactly epi-Lipschitzian* at some $x^0 \in S$ if there exist a norm-compact set K and a constant $r > 0$ such that

$$S \cap B(x^0, r) + B(0, tr) \subseteq S - tK \quad \forall t \in (0, r).$$

LEMMA 3.1. *For C and f as introduced above, the sum rule*

$$\partial(f + i_C)(\bar{x}) \subseteq \partial f(\bar{x}) + N(C; \bar{x})$$

applies if the following constraint qualification is satisfied:

$$\left.\begin{array}{l} \partial^\infty f(\bar{x}) \cap -N(C; \bar{x}) = \{0\} \quad \textit{and} \\ C \ \textit{or} \ \operatorname{epi} f \ \textit{is compactly epi-Lipschitzian at} \ \bar{x}. \end{array}\right\} \quad (\mathrm{CQ}^*)$$

Proof. Define two closed and convex subsets of $X \times \mathbb{R}$ by $D_1 = \operatorname{epi} f$ and $D_2 = C \times \mathbb{R}$. The first part of (CQ*) implies that

$$N(D_1; (\bar{x}; f(\bar{x}))) \cap -N(D_2; (\bar{x}; f(\bar{x}))) = \{0\}.$$

Along with the second part of (CQ*), this last relation is sufficient for the intersection rule

$$N(D_1 \cap D_2; (\bar{x}; f(\bar{x}))) \subseteq N(D_1; (\bar{x}; f(\bar{x}))) + N(D_2; (\bar{x}; f(\bar{x})))$$

(see [11, Corollary 4.5]). Now, let $x^* \in \partial(f + i_C)(\bar{x})$ be arbitrarily given, i.e., $\langle x^*, x - \bar{x}\rangle \leq f(x) - f(\bar{x})$ for all $x \in C$. Consequently,

$$\langle (x^*, -1), (x, t) - (\bar{x}, f(\bar{x}))\rangle \leq 0 \quad \forall x \in C, \forall t \geq f(x).$$

In other words, $(x^*, -1) \in N(D_1 \cap D_2; (\bar{x}; f(\bar{x})))$, and the above intersection rule ensures that $(x^*, -1) = (y^*, r) + (z^*, t)$ for certain $(y^*, r) \in N(D_1; (\bar{x}; f(\bar{x})))$ and $(z^*, t) \in N(D_2; (\bar{x}; f(\bar{x})))$. By definition of D_2, one gets $t = 0$ and $z^* \in N(C; \bar{x})$. It results that $r = -1$; hence $y^* \in \partial f(\bar{x})$ by definition of D_1. Summarizing, $x^* \in \partial f(\bar{x}) + N(C; \bar{x})$, as we wanted to show. □

Remark 3.2. The constraint qualification (CQ*) in Lemma 3.1 is always satisfied if the convex function f is continuous at $\bar{x}$ or if $\bar{x}$ is an interior point of C. The second part of (CQ*) holds true whenever X is finite-dimensional or the convex set C has nonempty interior.

THEOREM 3.3. *With the setting introduced above, the multifunction M in* (3.1) *is calm at a point $(0, \bar{x}) \in \operatorname{Gph} M$ of its graph if* one *of the following conditions is satisfied:*

$$f(\bar{x}) < 0, \tag{3.2}$$

$$\operatorname{bd} \partial f(\bar{x}) \cap -\operatorname{bd} N(C; \bar{x}) \neq \partial f(\bar{x}) \cap -N(C; \bar{x}), \tag{3.3}$$

$$\operatorname{bd} \partial f(\bar{x}) \cap -\operatorname{bd} N(C; \bar{x}) = \emptyset, \quad \textit{and} \quad (\mathrm{CQ}^*). \tag{3.4}$$

Proof. From $(0, \bar{x}) \in \operatorname{Gph} M$ it follows that $\bar{x} \in C$ and $f(\bar{x}) \leq 0$. In case of (3.2), it follows that

$$0 \in \operatorname{int} [f(\bar{x}), \infty) \subseteq \operatorname{int} \operatorname{range} M^{-1}. \tag{3.5}$$

Since M has a closed and convex graph, this last relation implies the metric regularity of M^{-1} at $(\bar{x}, 0)$ by the Robinson–Ursescu theorem (see [21], [25]). However, the metric regularity of M^{-1} at $(\bar{x}, 0)$ is equivalent to M having the Aubin property at $(0, \bar{x})$ (cf. [3], [20], [24]), which in turn implies the calmness of M at $(0, \bar{x})$. Hence, in what follows we assume that $f(\bar{x}) = 0$. Suppose next that (3.3) is satisfied. Then, since both $\partial f(\bar{x})$ and $-N(C; \bar{x})$ are (strongly) closed in X^*, it holds that

$$\operatorname{int} \partial f(\bar{x}) \cap -N(C; \bar{x}) \neq \emptyset \qquad \text{or} \qquad \partial f(\bar{x}) \cap -\operatorname{int} N(C; \bar{x}) \neq \emptyset. \tag{3.6}$$

If the first condition of (3.6) holds, then choose $x^* \in \operatorname{int} \partial f(\bar{x}) \cap -N(C; \bar{x})$. Accordingly, there exists some $\alpha > 0$ such that $B^*(x^*; \alpha) \subseteq \partial f(\bar{x})$. In other words,

$$\langle x^* + \alpha p^*, x - \bar{x}\rangle \leq f(x) - f(\bar{x}) = f(x) \quad \forall p^* \in B^*(0; 1), \forall x \in X.$$

It follows that

$$\langle p^*, x - \bar{x}\rangle \leq \alpha^{-1}(f(x) - \langle x^*, x - \bar{x}\rangle) \leq \alpha^{-1} f(x) \quad \forall p^* \in B^*(0; 1), \forall x \in C,$$

since $x^* \in -N(C;\bar{x})$. Consequently,

$$\|x-\bar{x}\| \le \alpha^{-1} f(x) \quad \forall x \in C \qquad \text{and} \qquad f(x) \ge 0 \quad \forall x \in C, \tag{3.7}$$

and thus the desired calmness property of M follows (with $\mathcal{U} := X$ and $\mathcal{V} := \mathbb{R}$ in (1.1)):

$$d(x, M(0)) \le \|x-\bar{x}\| \le \alpha^{-1} y = \alpha^{-1} d(y,0) \quad \forall y \in \mathbb{R}, \forall x \in M(y).$$

If the second condition of (3.6) holds true, then choose $x^* \in \partial f(\bar{x}) \cap -\operatorname{int} N(C;\bar{x})$. Now, there is some $\alpha > 0$ such that $B^*(x^*;\alpha) \subseteq -N(C;\bar{x})$; hence

$$\langle x^* - \alpha p^*, x-\bar{x}\rangle \ge 0 \quad \text{or} \quad \langle p^*, x-\bar{x}\rangle \le \alpha^{-1}\langle x^*, x-\bar{x}\rangle \quad \forall p^* \in B^*(0;1), \forall x \in C.$$

Due to $x^* \in \partial f(\bar{x})$, this yields $\|x-\bar{x}\| \le \alpha^{-1}\langle x^*, x-\bar{x}\rangle \le \alpha^{-1} f(x)$ for all $x \in C$. In this way, we end up once more at relation (3.7) and, hence, at the calmness of M at $(0,\bar{x})$, as above.

Finally, assume that (3.4) holds. If $0 \in \operatorname{int} \partial f(\bar{x})$, then—because of $0 \in \partial f(\bar{x}) \cap -N(C;\bar{x})$—(3.3) is satisfied and calmness of M follows as shown before. Suppose that $0 \in \operatorname{bd} \partial f(\bar{x})$. When $N(C;\bar{x}) = X^*$, calmness of M follows again from (3.3). In the opposite case, $N(C;\bar{x}) \ne X^*$, it always holds that $0 \in -\operatorname{bd} N(C;\bar{x})$, which gives a contradiction to (3.4). It remains to check the case of

$$0 \notin \partial f(\bar{x}). \tag{3.8}$$

Then, one has

$$\partial f(\bar{x}) \cap -N(C;\bar{x}) = \emptyset \quad \text{or} \quad \partial f(\bar{x}) \subseteq -\operatorname{int} N(C;\bar{x}). \tag{3.9}$$

To verify this statement, assume that neither of the two conditions is satisfied. Then, there exist $x_1^*, x_2^* \in \partial f(\bar{x})$ such that $x_1^* \in -N(C;\bar{x})$ and $x_2^* \notin -\operatorname{int} N(C;\bar{x})$. The convexity of $\partial f(\bar{x})$ and $-N(C;\bar{x})$ guarantees the existence of some x^* (on the line segment $[x_1^*, x_2^*]$) such that $x^* \in \partial f(\bar{x}) \cap -\operatorname{bd} N(C;\bar{x})$. By the cone property of $N(C;\bar{x})$, one has that $tx^* \in -\operatorname{bd} N(C;\bar{x})$ for all $t > 0$. Due to the closedness of $\partial f(\bar{x})$, there must be some $t^* > 0$ such that $t^* x^* \notin \partial f(\bar{x})$ (otherwise we have a contradiction with (3.8)). But then, since $x^* \in \partial f(\bar{x})$, there must exist some $\hat{t} > 0$ such that $\hat{t} x^* \in \operatorname{bd} \partial f(\bar{x})$. At the same time, $\hat{t} x^* \in -\operatorname{bd} N(C;\bar{x})$, whence a contradiction to (3.4), and (3.9) must hold true.

Now, the first case of (3.9) implies the existence of some $x' \in C$ such that $f(x') < 0$ (Slater's condition). Indeed, negating Slater's condition means that $\bar{x}$ is a minimum of f over C or, equivalently, a free minimum of the lower semicontinuous function $f + i_C$. Consequently,

$$0 \in \partial(f + i_C)(\bar{x}) \subseteq \partial f(\bar{x}) + N(C;\bar{x}),$$

where we have applied Lemma 3.1. However, the obtained relation contradicts the first case of (3.9). Hence, Slater's condition is satisfied, and one has (3.5) with $\bar{x}$ replaced by x'. Consequently, the calmness of M at $\bar{x}$ follows as in the lines below (3.5).

Concerning the second case of (3.9), assume first that $\partial f(\bar{x}) = \emptyset$. Then, we are back to the first case of (3.9) already considered. Finally, if $\partial f(\bar{x}) \ne \emptyset$, then the second case of (3.9), along with (3.4), yields (3.3), and the calmness of M at $(0,\bar{x})$ follows again. ∎

For missing abstract constraints, a much simpler characterization of calmness can be derived from Theorem 3.3, as follows.

COROLLARY 3.4. *Let X be a Banach space, and $f : X \to \mathbb{R} \cup \{\infty\}$ a convex, lower semicontinuous function. Then, the multifunction $M(y) := f^{-1}(-\infty, y]$ is calm at a point $(0, \bar{x})$ with $f(\bar{x}) \le 0$ if*

$$f(\bar{x}) < 0 \quad \text{or} \quad 0 \notin \operatorname{bd} \partial f(\bar{x}). \tag{3.10}$$

Proof. The first condition of (3.10) coincides with (3.2); thus it suffices to consider the second condition of (3.10). Evidently, in the setting of (3.1), we have $C = X$; hence $N(C; \bar{x}) = \operatorname{bd} N(C; \bar{x}) = \{0\}$. Along with $0 \notin \operatorname{bd} \partial f(\bar{x})$, this provides

$$\operatorname{bd} \partial f(\bar{x}) \cap -\operatorname{bd} N(C; \bar{x}) = \emptyset;$$

hence (3.4) is satisfied. (Note that (CQ*) is trivially satisfied in the context of this corollary; see Remark 3.2.) □

Note that in the setting of Corollary 3.4 we have the following implications:

$$(3.3) \Longrightarrow 0 \in \operatorname{int} \partial f(\bar{x}) \Longrightarrow (3.4) \Longrightarrow (3.10).$$

Hence, in contrast to the alternative of conditions (3.3) and (3.4) in Theorem 3.3, there is no point in considering (3.3) here in addition to (3.10). In the general setting of Theorem 3.3, however, it is no longer true that (3.3) implies (3.4), as can be seen from the second part of Example 3.6 below.

Remark 3.5. For finite-dimensional X, condition (3.4)—with the convex subdifferential replaced by Clarke—was shown in [7] to be sufficient for calmness of the multifunction M if f is locally Lipschitzian and both f and C are regular in the sense of Clarke. Theorem 3.3 demonstrates that this condition can be weakened to "(3.3) or (3.4)" in the convex case even if X is infinite-dimensional. More precisely, one has the following structure of constraint qualifications here (assuming that f is continuous at $\bar{x} \in C$ and $f(\bar{x}) = 0$):

$$\begin{array}{ccccc} \partial f(\bar{x}) \cap -N(C; \bar{x}) = \emptyset & \Longrightarrow & (3.4) & \Longrightarrow & (3.3) \text{ or } (3.4) \\ \Updownarrow & & & & \\ \text{Slater's condition} & & & & \Downarrow \\ \Updownarrow & & & & \\ \text{Aubin property of } M \text{ at } (0, \bar{x}) & & & & \text{calmness of } M \text{ at } (0, \bar{x}). \end{array} \tag{3.11}$$

In this diagram, we mean by Slater's condition the existence of some $x^* \in C$ such that $f(x^*) < 0$ (which is equivalent to the Aubin property of M at $(0, \bar{x})$ or to the metric regularity of M^{-1} at $(\bar{x}, 0)$ by the Robinson–Ursescu theorem).

We continue with some examples.

EXAMPLE 3.6. *The three constraint qualifications considered in Remark* 3.5 *are strictly different. Setting, for instance, $f(x) = |x|$, $C = \mathbb{R}$, $\bar{x} = 0$, Slater's condition is obviously violated (and also $0 \in \partial f(\bar{x}) \cap -N(C; \bar{x}) \neq \emptyset$), whereas* (3.4) *holds true:*

$$\operatorname{bd} \partial f(\bar{x}) \cap -\operatorname{bd} N(C; \bar{x}) = \{-1, 1\} \cap \{0\} = \emptyset.$$

Indeed, M is calm at $(0, \bar{x})$ but fails to have the Aubin property there. Another example is $f(x) = f(x_1, x_2) = \|x\|$, $C = \{(x_1, x_2) \mid x_1 \ge 0\}$. Then, at $\bar{x} = (0, 0)$, one has

$$\begin{aligned} \operatorname{bd} \partial f(\bar{x}) \cap -\operatorname{bd} N(C; \bar{x}) &= \{(x_1, x_2) \mid x_1^2 + x_2^2 = 1,\ x_1 \ge 0, x_2 = 0\} = \{(1, 0)\}, \\ \partial f(\bar{x}) \cap -N(C; \bar{x}) &= \{(x_1, x_2) \mid x_1^2 + x_2^2 \le 1,\ x_1 \ge 0, x_2 = 0\} \\ &= \operatorname{conv} \{(0, 0), (1, 0)\}. \end{aligned}$$

Hence, (3.4) *is violated here, whereas* (3.3) *is satisfied, and thus Theorem* 3.3 *ensures the calmness of* M *at* $(0,\bar{x})$*. Again,* M *fails to have the Aubin property.*

The following example demonstrates that Theorem 3.3 provides a sufficient but not a necessary condition for the calmness of the multifunction M considered there.

EXAMPLE 3.7. *Let* $X = C = \mathbb{R}$, $\bar{x} = 0$*, and* $f(x) = \max\{x, 0\}$*. Then,* $(0,\bar{x}) \in \operatorname{Gph} M$*,* $f(\bar{x}) = 0$*, and* $M(0) = \mathbb{R}_-$*. One has* $M(y) = \emptyset$ *for* $y < 0$*, and* $M(y) = (-\infty, y]$ *for* $y \geq 0$*; hence* $d(x, M(0)) \leq d(y, 0)$ *for all* $y \in \mathbb{R}$ *and all* $x \in M(y)$*. This means calmness of* M *at* $(0,\bar{x})$*. On the other hand, since* $\partial f(\bar{x}) = [0, 1]$*,* (3.10) *is violated, which implies the violation of both* (3.4) *and* (3.3).

Note that, in the last example, M was a polyhedral multifunction; hence it seems that one cannot recover by Theorem 3.3 Robinson's result mentioned in the introduction. However, this will be possible after some modification, following the ideas of [12].

The next example requires some technical work. It illustrates the limitation of Theorem 3.3 to convex data. In finite dimensions, the condition "$f(\bar{x}) < 0$ or $0 \notin \operatorname{bd} \partial^c f(\bar{x})$" (i.e., (3.10) with the convex subdifferential replaced by Clarke's) was found in [7] to ensure calmness of the multifunctions (3.1) without abstract constraints (i.e., $C = X$) as long as f is regular at $\bar{x}$ in the sense of Clarke. This is no longer true in infinite dimensions unless the data are restricted to be convex as in Corollary 3.4.

EXAMPLE 3.8. *For* $k \in \mathbb{N}$*, let* $\tau_k \in (0, k^{-2})$ *be the unique solution of* $\tau + k\sqrt{\tau} = 1$*. Define the sequence of real functions*

$$\varphi_k(\tau) := \begin{cases} |\tau|(1 - k\sqrt{|\tau|}) & \text{if } \tau \in [-\tau_k, \tau_k], \\ \tau_k^2 & \text{if } |\tau| \geq \tau_k. \end{cases}$$

Elementary analysis shows that each φ_k *is (globally) Lipschitz continuous with modulus* 1 *and regular at zero in the sense of Clarke. (Close to the origin, each* φ_k *can be represented as the maximum of two* $\mathcal{C}^1$*- functions.) Furthermore,*

$$\varphi_k \geq 0, \quad \varphi_k(\tau) = 0 \Longleftrightarrow \tau = 0, \quad \text{and} \quad \varphi_k(\tau_k) = \tau_k^2 \quad \forall k \in \mathbb{N}, \forall \tau \in \mathbb{R}. \tag{3.12}$$

Now, let $X = l^1$*, and define* $f : X \to \mathbb{R}$ *by* $f(x) := \sum_{k=1}^{\infty} \varphi_k(x_k)$*. Evidently,* $f(0) = 0$ *by* (3.12)*. Since* $\varphi_k(\tau) \leq \tau_k^2 \leq k^{-4}$ *for all* $\tau \in \mathbb{R}$ *and all* $k \in \mathbb{N}$*,* f *is well defined. For arbitrary* $x, y \in X$*, one has*

$$\begin{aligned} |f(x) - f(y)| &= \left| \sum_{k=1}^{\infty} (\varphi_k(x_k) - \varphi_k(y_k)) \right| \leq \sum_{k=1}^{\infty} |\varphi_k(x_k) - \varphi_k(y_k)| \\ &\leq \sum_{k=1}^{\infty} |x_k - y_k| = \|x - y\|_1; \end{aligned}$$

hence f *is (globally) Lipschitz continuous with modulus* 1*.*

Next, we calculate Clarke's directional derivative $f^0(0; h)$ *of* f *at zero in arbitrary direction* $h \in X$*. By definition (see* [4]*), one has*

$$\begin{aligned} f^0(0; h) &= \limsup_{t \downarrow 0, x \to 0} \frac{f(x + th) - f(x)}{t} = \lim_{n \to \infty} \frac{f(x^{(n)} + t^{(n)}h) - f(x^{(n)})}{t^{(n)}} \\ &= \lim_{n \to \infty} \sum_{k=1}^{\infty} \frac{\varphi_k(x_k^{(n)} + t^{(n)}h_k) - \varphi_k(x_k^{(n)})}{t^{(n)}}, \end{aligned} \tag{3.13}$$

where $x^{(n)} \to 0$ and $t^{(n)} \downarrow 0$ are suitable sequences realizing the above limsup as a limit. Now, we fix an arbitrary $k' \in \mathbb{N}$. Assume that there exist $\varepsilon > 0$ and $n_0 \in \mathbb{N}$ such that

$$\frac{\varphi_{k'}(x_{k'}^{(n)} + t^{(n)}h_{k'}) - \varphi_{k'}(x_{k'}^{(n)})}{t^{(n)}} \leq \frac{\varphi_{k'}(t^{(n)}h_{k'})}{t^{(n)}} - \varepsilon \quad \forall n \geq n_0. \tag{3.14}$$

In order to lead (3.14) *to a contradiction, define a sequence $\tilde{x}^{(n)} \in X$ by*

$$\tilde{x}_k^{(n)} := \begin{cases} x_k^{(n)}, & k \neq k', \\ 0, & k = k', \end{cases} \qquad \forall k, n \in \mathbb{N}.$$

It follows that $\tilde{x}^{(n)} \to 0$ and, in view of (3.12),

$$\begin{aligned} \frac{f(\tilde{x}^{(n)} + t^{(n)}h) - f(\tilde{x}^{(n)})}{t^{(n)}} &= \sum_{k=1, k\neq k'}^{\infty} \frac{\varphi_k(x_k^{(n)} + t^{(n)}h_k) - \varphi_k(x_k^{(n)})}{t^{(n)}} + \frac{\varphi_{k'}(t^{(n)}h_{k'})}{t^{(n)}} \\ &\geq \sum_{k=1}^{\infty} \frac{\varphi_k(x_k^{(n)} + t^{(n)}h_k) - \varphi_k(x_k^{(n)})}{t^{(n)}} + \varepsilon \\ &= \frac{f(x^{(n)} + t^{(n)}h) - f(x^{(n)})}{t^{(n)}} + \varepsilon \end{aligned}$$

for $n \geq n_0$, whence the contradiction with (3.13),

$$\limsup_{t\downarrow 0, x\to 0} \frac{f(x+th) - f(x)}{t} \geq f^0(0;h) + \varepsilon.$$

Therefore, we may negate (3.14) *in order to obtain a subsequence symbolized by the index $m(n)$ such that*

$$\begin{aligned} &\liminf_{n\to\infty} \frac{\varphi_{k'}(x_{k'}^{(m(n))} + t^{(m(n))}h_{k'}) - \varphi_{k'}(x_{k'}^{(m(n))})}{t^{(m(n))}} \\ &\quad \geq \lim_{n\to\infty} \frac{\varphi_{k'}(t^{(m(n))}h_{k'})}{t^{(m(n))}} = d\varphi_{k'}(0;h_{k'}) = \varphi_{k'}^0(0;h_{k'}) \\ &\quad \geq \limsup_{n\to\infty} \frac{\varphi_{k'}(x_{k'}^{(m(n))} + t^{(m(n))}h_{k'}) - \varphi_{k'}(x_{k'}^{(m(n))})}{t^{(m(n))}}, \end{aligned} \tag{3.15}$$

where "$d\varphi_{k'}$" refers to the usual directional derivative, which, by the already stated regularity of $\varphi_{k'}$ in the sense of Clarke, exists and coincides with $\varphi_{k'}^0$. From the definition of $\varphi_{k'}$, one calculates $d\varphi_{k'}(0;h_{k'}) = |h_{k'}|$. Since k' was arbitrarily fixed, (3.15) *provides*

$$\lim_{n\to\infty} \frac{\varphi_k(x_k^{(m(n))} + t^{(m(n))}h_k) - \varphi_k(x_k^{(m(n))})}{t^{(m(n))}} = |h_k| \quad \forall k \in \mathbb{N}.$$

This finally allows us to interchange limit and summation in the last term of (3.13) *(upon passing to the subsequence $m(n)$ there too):*

$$f^0(0;h) = \sum_{k=1}^{\infty} |h_k| = \|h\|_1 \qquad \forall h \in X.$$

Consequently, $\partial^c f(0) = B_1$, where ∂^c denotes Clarke's subdifferential and B_1 is the unit ball in X.

Next, we verify that f is regular at 0 in the sense of Clarke. To this aim, we calculate its usual directional derivative at 0 in arbitrary direction h. Since for each sequence $t^{(n)} \downarrow 0$ it holds that

$$\lim_{n\to\infty} \frac{\varphi_k(t^{(n)}\tilde{h})}{t^{(n)}} = d\varphi_k(0;\tilde{h}) = |\tilde{h}| \quad \forall \tilde{h} \in \mathbb{R}, \forall k \in \mathbb{N},$$

one may interchange limit and summation once more:

$$\|h\|_1 = \sum_{k=1}^{\infty} \lim_{n\to\infty} \frac{\varphi_k(t^{(n)}h_k)}{t^{(n)}} = \lim_{n\to\infty} \sum_{k=1}^{\infty} \frac{\varphi_k(t^{(n)}h_k)}{t^{(n)}} = \lim_{n\to\infty} \frac{f(t^{(n)}h) - f(0)}{t^{(n)}}.$$

As $t^{(n)} \downarrow 0$ was arbitrary, it follows that $df(0;h) = \|h\|_1 = f^0(0;h)$; hence f is regular in the sense of Clarke.

Finally, we consider the multivalued mapping $M : \mathbb{R} \rightrightarrows X$ defined by $M(t) := \{x \in X \mid f(x) \le t\}$. This is exactly the setting of (3.1) *with abstract constraints missing ($X = C$). By the definition of f and* (3.12)*, one has*

$$f(x) \ge 0 \quad \forall x \in X \qquad \text{and} \qquad f(x) = 0 \Longleftrightarrow x = 0.$$

Hence, $M(0) = \{0\}$. Define a sequence $z^{(n)} = (0, \dots, 0, \tau_n, 0, 0, \dots) \in X$, with τ_n at position n. Then, again by (3.12)*,*

$$d(z^{(n)}, M(0)) = \|z^{(n)}\|_1 = \tau_n \qquad \text{and} \qquad f(z^{(n)}) = \varphi_n(\tau_n) = \tau_n^2 \quad \forall n \in \mathbb{N}.$$

Setting $y^{(n)} := f(z^{(n)})$, we have constructed sequences $z^{(n)}, y^{(n)}$ such that $z^{(n)} \in M(y^{(n)})$, $z^{(n)} \to 0$, $y^{(n)} \to 0$ (because of $\tau_n \in (0, n^{-2})$). From here, we derive that M fails to be calm at $(0,0)$:

$$d(z^{(n)}, M(0)) = \tau_n^{-1} f(z^{(n)}) = \tau_n^{-1} d(f(z^{(n)}), 0) \ge n^2 d(f(z^{(n)}), 0)$$

(again by $\tau_n \in (0, n^{-2})$), which contradicts (1.1)*. On the other hand, we have seen that $\partial^c f(0) = B_1$; hence $0 \in \operatorname{int} \partial^c f(0)$, and the constraint qualification "$f(\bar{x}) < 0$ or $0 \notin \operatorname{bd} \partial^c f(\bar{x})$"—which was sufficient for calmness in the regular, finite-dimensional and in the convex, infinite-dimensional cases—is evidently satisfied. However, the same constraint qualification (to which the conditions* (3.4) *and* (3.3) *reduce when $C = X$) does not imply calmness in the regular, infinite-dimensional case, as was shown in this example.*

The next result is an immediate application of Theorem 3.3 to the characterization of calmness for nonstructured multifunctions.

COROLLARY 3.9. *Let X be a Banach space, Y a metric space, $M : X \rightrightarrows Y$ a multifunction with closed values, and $(\bar{x}, \bar{y}) \in \operatorname{Gph} M$. Assume further that*

(1) *the distance function $d(\bar{y}, M(\cdot))$ is convex and lower semicontinuous in a neighborhood of $\bar{x}$;*
(2) $0 \notin \operatorname{bd} \partial d(\bar{y}, M(\cdot))(\bar{x})$.

Then M^{-1} is calm at $(\bar{y}, \bar{x})$.

Proof. Corollary 3.4 immediately provides calmness at $(\bar{y}, \bar{x})$ of the multifunction $P : \mathbb{R} \rightrightarrows X$ defined by

$$P(t) := \{x \in X \mid d(\bar{y}, M(x)) \le t\}.$$

This means the existence of some $L > 0$, $\varepsilon > 0$ such that

$$d(x, P(0)) \le L|t| \quad \forall t \in (-\varepsilon, \varepsilon),\ \forall x \in B^0(\bar{x}; \varepsilon) \cap P(t).$$

Since $P(0) = M^{-1}(\bar{y})$ and $M^{-1}(y) \subseteq P(d(\bar{y}, y))$ for all $y \in Y$, the calmness of M^{-1} at $(\bar{y}, \bar{x})$ follows:

$$d(x, M^{-1}(\bar{y})) = d(x, P(0)) \le L d(\bar{y}, y) \quad \forall y \in B^0(\bar{y}; \varepsilon),\ \forall x \in B^0(\bar{x}; \varepsilon) \cap M^{-1}(y). \qquad \square$$

Note that Corollary 3.9(2) is far removed from being necessary for calmness or even the stronger Aubin property.

EXAMPLE 3.10. *Consider* $M(x) := [x, \infty)$ *at* $(0, 0) \in \operatorname{Gph} M$. *Since* $d(0, M(x)) = \max\{0, x\}$, *Corollary* 3.9(1) *is satisfied, whereas condition* (2) *is violated. On the other hand, the inverse multifunction* $M^{-1}(y) = \{x | x \le y\}$ *is easily seen to satisfy the Aubin property* (1.2) *and, hence, calmness at* $(0, 0)$.

At the end of this section we want briefly to compare our conditions for the calmness of system (3.1) with similar conditions which were obtained in the context of error bounds. First, recall that the calmness of (3.1) is equivalent to the existence of a local error bound. A rigorous comparison is difficult because the obtained conditions may differ by many features (e.g., local vs. global error bounds, primal vs. dual conditions, finite- vs. infinite-dimensional spaces, point vs. neighborhood conditions). However, one could at least try to reduce all these conditions to a simple common setting, where $C = X$ is finite-dimensional and f is convex and finite-valued. As far as dual conditions for error bounds are concerned, they usually come down to just Slater's condition in dual form, "$0 \notin \partial f(\bar{x})$" in that situation (see, e.g., [14, condition (ACQ11)], [15, Section 3, Corollary 2(b)], or [5, Theorem 1]). Slater's condition, however, is much stronger than our condition (3.10), as was shown in Example 3.6 (see also (3.11)). A primal condition for calmness proposed in [17, Theorem 13] is

$$0 \in \operatorname{int}(f(C) + \mathbb{R}_+).$$

However, in our setting, with $C = X$, this relation obviously reduces to Slater's condition in primal form: "$\exists x^* : f(x^*) < 0$." Hence, the same remarks as above apply with respect to condition (3.10). A mixed primal/dual condition was derived in [15, Theorem 1] for finite dimensions:

$$\exists \gamma > 0 : f'(\bar{x}; d) \ge \gamma^{-1} \|d\| \quad \forall \bar{x} \in f^{-1}(0),\ \forall d \in N(f^{-1}(-\infty, 0]; \bar{x}). \tag{3.16}$$

Here, f' refers to the directional derivative of f. It is elementary to verify that in the special setting considered here ($C = X$), (3.10) implies (3.16). In particular, (3.16) could be applied in Example 3.7, where (3.10) failed. On the other hand, (3.16) is not a point condition by relying on the whole solution set $f^{-1}(-\infty, 0]$. This could make its verification in general problems less convenient than that of (3.10), which is sufficient at least for local error bounds. A similar comparison holds true for a nonsmooth Abadie's constraint qualification formulated in [19].

4. Calmness of the intersection of two sets. In this section, we turn to the calmness property with respect to two sets. To this aim, let $C, D \subseteq X$ be closed, convex subsets such that $\bar{x} \in C \cap D$. We want to characterize the calmness of the multivalued mapping $Q : \mathbb{R} \rightrightarrows X$ defined by

$$Q(t) := \{x \in X \mid d(x, C) + d(x, D) \le t\}$$

at the point $(0,\bar{x}) \in \operatorname{Gph} Q$.

LEMMA 4.1. *Q is calm at $(0,\bar{x}) \in \operatorname{Gph} Q$, provided that*

$$\operatorname{int} N(D;\bar{x}) \cap -N(C;\bar{x}) \neq \emptyset. \tag{4.1}$$

Proof. Choose $x^* \in \operatorname{int} N(D;\bar{x}) \cap -N(C;\bar{x})$. From $x^* \in \operatorname{int} N(D;\bar{x})$, it follows, similarly to the proof of Theorem 3.3, that there exists some $\alpha > 0$ such that

$$\alpha\|x-\bar{x}\| + \langle x^*, x-\bar{x}\rangle \leq 0 \quad \forall x \in D.$$

Hence, $\bar{x}$ is a minimizer of the function $\langle x^*, \bar{x}-\cdot\rangle - \alpha\|\cdot-\bar{x}\|$ on the set D. Now, using a well-known penalization argument, which appeals to the Lipschitz constant of the function involved, it follows that there exists some $\varepsilon > 0$ such that

$$\langle x^*, \bar{x}-x\rangle - \alpha\|x-\bar{x}\| + (\|x^*\|+\alpha)d(x,D) \geq 0 \quad \forall x \in B(\bar{x};\varepsilon),$$

whence, by $x^* \in -N(C;\bar{x})$,

$$-\alpha\|x-\bar{x}\| + (\|x^*\|+\alpha)d(x,D) \geq 0 \quad \forall x \in B(\bar{x};\varepsilon) \cap C.$$

In other words, $\bar{x}$ is a local minimizer of the function $-\alpha\|\cdot-\bar{x}\| + (\|x^*\|+\alpha)d(\cdot,D)$ on the set C. Now, upon repeating the same penalization argument, one arrives at

$$-\alpha\|x-\bar{x}\| + (\|x^*\|+\alpha)d(x,D) + (\|x^*\|+2\alpha)d(x,C) \geq 0 \quad \forall x \in B(\bar{x};\varepsilon')$$

for some $\varepsilon' > 0$. This, however, is the desired calmness property

$$d(x,Q(0)) \leq \|x-\bar{x}\| \leq \alpha^{-1}(\|x^*\|+2\alpha)(d(x,D)+d(x,C)) \leq \alpha^{-1}(\|x^*\|+2\alpha)|t|,$$

which holds true for all $t \in \mathbb{R}$ and all $x \in B(\bar{x};\varepsilon') \cap Q(t)$. ∎

Next, we need an auxiliary result, which is of independent interest.

LEMMA 4.2. *If one of the sets C or D is compactly epi-Lipschitzian in a neighborhood of $\bar{x}$, then*

$$N(D;\bar{x}) \cap -N(C;\bar{x}) = \{0\} \Longleftrightarrow 0 \in \operatorname{int}(D - C \cap B(\bar{x},1)).$$

Proof. ($\Longrightarrow$) For symmetry reasons, one may take, e.g., D to be compactly epi-Lipschitzian in a neighborhood of $\bar{x}$. Assume that

$$0 \notin \operatorname{int}(D - C \cap B(\bar{x},1)) = \operatorname{int}\overline{(D - C \cap B(\bar{x},1))}$$

(the equality follows from [22, Lemma 1]). Accordingly, there exists a sequence $b_n \to 0$ with

$$b_n \notin \overline{D - C \cap B(\bar{x},1))}.$$

The separation theorem provides a corresponding sequence $x_n^* \in X^*$ such that $\|x_n^*\| = 1$ and

$$\langle x_n^*, b_n\rangle \leq \langle x_n^*, d-\bar{x}\rangle \quad \forall d \in D, \quad \langle x_n^*, b_n\rangle \leq \langle x_n^*, \bar{x}-c\rangle \quad \forall c \in C \cap B(\bar{x},1). \tag{4.2}$$

The first relation of (4.2) yields that $\langle x_n^*, \bar{x}\rangle \leq \inf_{d\in D}\langle x_n^*, d\rangle + \|b_n\|$. Now Ekeland's variational principle provides a sequence $d_n \in D$ such that

$$\|d_n - \bar{x}\| \leq \sqrt{\|b_n\|} \quad \text{and} \quad \langle x_n^*, d_n\rangle \leq \langle x_n^*, d\rangle + \sqrt{\|b_n\|}\,\|d_n - d\| \quad \forall d \in D. \tag{4.3}$$

The second relation of (4.3) entails that $-x_n^* \in N(D;d_n)+B^*(0,\sqrt{\|b_n\|})$; hence there are sequences $z_n^* \in N(D;d_n)$ and b_n^* with $\|b_n^*\| \le \sqrt{\|b_n\|}$ such that $z_n^* + x_n^* + b_n^* = 0$. In particular, $\|z_n^*\| \to 1$. Thus, the sequence z_n^* is bounded and, hence, there exists a weak* convergent subnet $z_\lambda^* \rightharpoonup_{w^*} z^*$. Now, since $z_\lambda^* \in N(D;d_\lambda)$, this last convergence, $d_\lambda \to \bar{x}$ (see the first relation of (4.3)), and the very definition of the normal cone to convex sets yield that $z^* \in N(D;\bar{x})$. Now, the assumed property of D being compactly epi-Lipschitzian in a neighborhood $\mathcal{V}_{\bar{x}}$ of $\bar{x}$ results in the inclusion

$$N(D;x) \subseteq \left\{ x^* \mid \|x^*\| \le \max_{i=1,\dots,k} \langle x^*, h_i\rangle \right\} \quad \forall x \in \mathcal{V}_{\bar{x}} \cap D$$

for certain $h_i \in X$ $(i = 1,\dots,k)$. From $d_\lambda \to \bar{x}$, one derives that

$$\max_{i=1,\dots,k} \langle z_\lambda^*, h_i\rangle \ge \|z_\lambda^*\|.$$

Consequently, $z^* \ne 0$. On the other hand, we also have that $x_\lambda^* = -z_\lambda^* - b_\lambda^* \rightharpoonup_{w^*} -z^*$, which together with the second part of (4.2) provides

$$\langle -z^*, \bar{x}-c\rangle \leftharpoonup_{w^*} \langle x_\lambda^*, \bar{x}-c\rangle \ge \langle x_\lambda^*, b_\lambda\rangle \to 0 \quad \forall c \in C \cap B(\bar{x},1),$$

whence $z^* \in -N(C;\bar{x})$. Summarizing, there is some $z^* \ne 0$ with $z^* \in N(D;\bar{x}) \cap -N(C;\bar{x})$. This contradicts our assumption.

($\Longleftarrow$) Choose an arbitrary $x^* \in N(D;\bar{x}) \cap -N(C;\bar{x})$. Then,

$$\langle x^*, d-\bar{x}\rangle \le 0 \quad \forall d \in D \quad \text{and} \quad \langle x^*, \bar{x}-c\rangle \le 0 \quad \forall c \in C.$$

In other words, $\langle x^*, d-c\rangle \le 0$ for all $d \in D$ and all $c \in C$. However, since by assumption $0 \in \operatorname{int}(D-C)$, it results that $x^* = 0$, as we wanted to show. ∎

THEOREM 4.3. *Let one of the sets C or D be compactly epi-Lipschitzian at $\bar{x}$. Then, Q is calm at $(0,\bar{x})$ under the following condition:*

$$\operatorname{bd} N(D;\bar{x}) \cap -\operatorname{bd} N(C;\bar{x}) = \{0\}. \tag{4.4}$$

Proof. For the case in which $N(D;\bar{x}) \cap -N(C;\bar{x}) = \{0\}$, Lemma 4.2 ensures that $0 \in \operatorname{int}(D-C)$. Since $D-C$ equals the range of the multifunction $M : X \rightrightarrows X$ defined by

$$M(x) = \begin{cases} -x + D, & x \in C, \\ \emptyset, & x \notin C, \end{cases}$$

we have $0 \in \operatorname{int} \operatorname{range} M$, and the Robinson–Ursescu theorem yields the metric regularity of M at the point $(\bar{x},0)$ of its graph. This property means the existence of $L, \varepsilon > 0$ such that

$$d(x, M^{-1}(y)) \le L\, d(y, M(x)) \quad \forall x \in B(\bar{x},\varepsilon),\ \forall y \in B(0,\varepsilon).$$

With $M^{-1}(y) = C \cap (D-y)$ and fixing $y := 0$, one arrives at

$$d(x, C\cap D) \le L\, d(x,D) \quad \forall x \in B(\bar{x},\varepsilon) \cap C; \text{ hence}$$

$$d(x, C\cap D) \le (L+1)\,(d(x,D) + d(x,C)) \quad \forall x \in B(\bar{x},\varepsilon).$$

This, of course, is the calmness of the multifunction Q at $(0,\bar{x})$.

Otherwise ($N(D;\bar{x}) \cap -N(C;\bar{x}) \ne \{0\}$), (4.4) implies that

$$\operatorname{int} N(D;\bar{x}) \cap -N(C;\bar{x}) \ne \emptyset \quad \text{or} \quad N(D;\bar{x}) \cap -\operatorname{int} N(C;\bar{x}) \ne \emptyset.$$

In both cases, Lemma 4.1 yields the desired result. ∎

5. The differentiable nonconvex case. In this section we briefly return to the constraint system (3.1), with a convex closed subset $C \subseteq X$ as before, but with a (strictly) differentiable function f. Theorem 3.3 has shown that, in the completely convex case (C and f), each of the constraint qualifications (3.4), (3.3) is sufficient for the calmness of (3.1). On the other hand, we know by Example 3.8 that neither of the two conditions ensures calmness if f is just regular in the sense of Clarke. Since, in that example, f was nondifferentiable, the question arises of whether a positive result can be expected in the smooth case. The answer is affirmative even for a finite number of inequalities.

THEOREM 5.1. *Consider a multifunction $M : \mathbb{R}^m \rightrightarrows X$ defined by*

$$M(y) := \{x \in C \mid f(x) \leq y\} \quad (y \in \mathbb{R}^m),$$

where $C \subseteq X$ is convex and closed and $f : X \to \mathbb{R}^m$ is strictly differentiable. Then, the constraint qualification

$$\operatorname{conv}\{\nabla f_i(\bar{x})\}_{i \in I(\bar{x})} \cap -\operatorname{bd} N(C; \bar{x}) = \emptyset \tag{5.1}$$

implies the calmness of M at $(0, \bar{x}) \in \operatorname{Gph} M$. Here, f_i denote the components of f, and $I(x) = \{i \in \{1, \ldots, m\} | f_i(x) = 0\}$ refers to the set of active indices.

Proof. Assume first that $\operatorname{conv}\{\nabla f_i(\bar{x})\}_{i \in I(\bar{x})} \cap -N(C; \bar{x}) = \emptyset$. Then, the strict differentiability assumption on f allows us to apply Theorem 2.4 in [13] in order to derive the metric regularity of M^{-1} at $(\bar{x}, 0)$, which is equivalent to the Aubin property of M at $(0, \bar{x})$ and, hence, implies calmness of M at $(0, \bar{x})$. In the opposite case, (5.1) guarantees the existence of some $x^* \in \operatorname{conv}\{\nabla f_i(\bar{x})\}_{i \in I(\bar{x})} \cap -\operatorname{int} N(C; \bar{x})$. Accordingly, there exist $\lambda_i \geq 0$ ($i \in I(\bar{x})$) with $\sum_{i \in I(\bar{x})} \lambda_i = 1$ as well as $\varepsilon > 0$ such that

$$x^* = \sum_{i \in I(\bar{x})} \lambda_i \nabla f_i(\bar{x}) \quad \text{and} \quad \varepsilon \|x - \bar{x}\| \leq \langle x^*, x - \bar{x} \rangle \quad \forall x \in C.$$

Due to the differentiability assumption on f and to the finiteness of $I(\bar{x})$, there is some $\eta > 0$ such that

$$f_i(x) - f_i(\bar{x}) \geq \langle \nabla f_i(\bar{x}), x - \bar{x} \rangle - \frac{\varepsilon}{2} \|x - \bar{x}\| \quad \forall x \in B(\bar{x}, \eta),\ \forall i \in I(\bar{x}).$$

Using the fact that $f_i(\bar{x}) = 0$ for $i \in I(\bar{x})$, it holds for all $x \in C \cap B(\bar{x}, \eta)$ that

$$\max_{i \in I(\bar{x})} f_i(x) \geq \sum_{i \in I(\bar{x})} \lambda_i f_i(x) \geq \sum_{i \in I(\bar{x})} \lambda_i \langle \nabla f_i(\bar{x}), x - \bar{x} \rangle - \frac{\varepsilon}{2} \|x - \bar{x}\| \geq \frac{\varepsilon}{2} \|x - \bar{x}\|.$$

Measuring, without loss of generality, the distance in $\mathbb{R}^m$ with respect to the maximum norm, one has for all $x \in M(y) \cap B(\bar{x}, \eta)$ and all $y \in \mathbb{R}^m$

$$d(x, M(0)) \leq \|x - \bar{x}\| \leq \frac{2}{\varepsilon} \max_{i \in I(\bar{x})} f_i(x) \leq \frac{2}{\varepsilon} \max_{i=1,\ldots,m} |y_i| = \frac{2}{\varepsilon} d(y, 0).$$

This, however, is the calmness of M at $(0, \bar{x})$. ∎

The last result shows that the ideas of the completely convex case can be extended to differentiable inequalities. With a single inequality which is differentiable *and* convex, (5.1) reduces to (3.4) (without the need of the additional constraint qualification (CQ*)). One might ask about an alternative condition in the sense of (3.3) for the

differentiable case as well. However, the closedness of the normal cone immediately provides that the differentiable formulation of (3.3) implies (3.4); hence the two conditions are not independent as in the convex (nonsmooth) setting. Finally, we note that for finite-dimensional X, (5.1) can be weakened to the condition

$$\operatorname{bd}\operatorname{conv}\{\nabla f_i(\bar{x})\}_{i\in I(\bar{x})} \cap -\operatorname{bd} N(C;\bar{x}) = \emptyset$$

(see [7, Theorem 9]). In infinite dimensions, the interior of the convex hull involved is empty; hence this last relation is equivalent to (5.1).

6. Conclusion. The dual conditions for calmness formulated in this paper (in particular, (3.3), (3.4), (3.10), and (5.1)) are weaker than the usual Slater-type characterizations, which ensure the stronger Aubin property (or metric regularity) of the considered systems. These conditions can be immediately applied to various issues in mathematical programming such as error bounds, optimality conditions, weak sharp minima (see also [8]), or the stability of solutions under perturbations.

Acknowledgments. This paper originated from the authors' mutual visits at the University of Bourgogne, Dijon, and Weierstrass Institute Berlin, respectively. The authors express their gratitude to both institutions for their support.

REFERENCES

[1] J.-P. Aubin and I. Ekeland, *Applied Nonlinear Analysis*, Wiley, New York, 1984.
[2] J.M. Borwein and H.M. Strojwas, *Tangential approximations*, Nonlinear Anal., 9 (1985), pp. 1347–1366.
[3] J.M. Borwein and D.M. Zhuang, *Verifiable necessary and sufficient conditions for regularity of set-valued and single-valued maps*, J. Math. Anal. Appl., 134 (1988), pp. 441–459.
[4] F. Clarke, *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983.
[5] S. Deng, *Global error bounds for convex inequality systems in Banach spaces*, SIAM J. Control. Optim., 36 (1998), pp. 1240–1249.
[6] R. Henrion, *The Approximate Subdifferential and Parametric Optimization*, Habilitation Thesis, Humboldt University, Berlin, 1997.
[7] R. Henrion and J. Outrata, *A subdifferential criterion for calmness of multifunctions*, J. Math. Anal. Appl., 258 (2001), pp. 110–130.
[8] R. Henrion, A. Jourani and J. Outrata, *On the calmness of a class of multifunctions*, SIAM J. Optim., 13 (2002), pp. 603–618.
[9] A.J. King and R.T. Rockafellar, *Sensitivity analysis for nonsmooth generalized equations*, Math. Programming, 55 (1992), pp. 193–212.
[10] A.B. Levy, *Implicit multifunction theorems for the sensitivity analysis of variational conditions*, Math. Programming, 74 (1996), pp. 333–350.
[11] A. Jourani, *Intersection formulae and the marginal function in Banach spaces*, J. Math. Anal. Appl., 192 (1995), pp. 867–891.
[12] A. Jourani, *Hoffman's error bound, local controllability, and sensitivity analysis*, SIAM J. Control. Optim., 38 (2000), pp. 947–970.
[13] A. Jourani and L. Thibault, *Metric regularity for strongly compactly Lipschitzian mappings*, Nonlinear Anal., 24 (1995), pp. 229–240.
[14] D. Klatte and W. Li, *Asymptotic constraint qualifications and global error bounds for convex inequalities*, Math. Program., 84 (1999), pp. 137–160.
[15] A.S. Lewis and J.-S. Pang, *Error bounds for convex inequality systems*, in Generalized Convexity, Generalized Monotonicity: Recent Results, J.P. Crouzeix, J.E. Martinez-Legaz, and M. Volle, eds., Kluwer, Dordrecht, the Netherlands, 1997, pp. 75–100.
[16] W. Li, *Abadie's constraint qualification, metric regularity, and error bounds for differentiable convex inequalities*, SIAM J. Optim., 7 (1997), pp. 966–978.
[17] W. Li and I. Singer, *Global error bounds for convex multifunctions and applications*, Math. Oper. Res., 23 (1998), pp. 443–462.
[18] B.S. Mordukhovich, *Complete characterization of openness, metric regularity and Lipschitzian properties of multifunctions*, Trans. Amer. Math. Soc., 340 (1993), pp. 1–35.

[19] J.-S. Pang, *Error bounds in mathematical programming,* Math. Programming, 79 (1997), pp. 299–332.
[20] J.-P. Penot, *Metric regularity, openness and Lipschitzian behavior of multifunctions*, Nonlinear Anal., 13 (1989), pp. 629–643.
[21] S.M. Robinson, *Normed convex processes*, Trans. Amer. Math. Soc., 174 (1972), pp. 127–140.
[22] S.M. Robinson, *Regularity and stability for convex multivalued functions*, Math. Oper. Res., 1 (1976), pp. 130–143.
[23] S.M. Robinson, *Some continuity properties of polyhedral multifunctions*, Math. Program. Studies, 14 (1981), pp. 206–214.
[24] R.T. Rockafellar and R.J-B. Wets, *Variational Analysis*, Springer, New York, 1997.
[25] C. Ursescu, *Multifunctions with closed, convex graph*, Czechoslovak Math. J., 25 (1975), pp. 438–441.

SIAM J. OPTIM.
Vol. 13, No. 2, pp. 535–560

ROBUST SOLUTIONS OF UNCERTAIN QUADRATIC AND CONIC-QUADRATIC PROBLEMS*

A. BEN-TAL†, A. NEMIROVSKI†, AND C. ROOS‡

Dedicated to Jochem Zowe on the occasion of his 60*th birthday.*

Abstract. We consider a conic-quadratic (and in particular a quadratically constrained) optimization problem with uncertain data, known only to reside in some uncertainty set $\mathcal{U}$. The robust counterpart of such a problem leads usually to an NP-hard semidefinite problem; this is the case, for example, when $\mathcal{U}$ is given as the intersection of ellipsoids or as an n-dimensional box. For these cases we build a single, explicit semidefinite program, which approximates the NP-hard robust counterpart, and we derive an estimate on the quality of the approximation, which is essentially independent of the dimensions of the underlying conic-quadratic problem.

Key words. semidefinite relaxation of NP-hard problems, (conic) quadratic programming, robust optimization

AMS subject classification. 90C05

PII. S1052623401392354

1. Introduction.

A conic problem is an optimization problem of the form

$$\min_{x\in\mathbb{R}^n} \left\{c^T x \ : \ Ax - b \in \mathcal{K}\right\}, \tag{CP}$$

where $\mathcal{K} \subseteq \mathbb{R}^N$ is a closed pointed convex cone with nonempty interior. The *data* associated with (CP) is the triple (A, b, c), with $A \in \mathbb{R}^{N\times n}$, $b \in \mathbb{R}^N$, and $c \in \mathbb{R}^n$.

1.1. Uncertainty in conic problems.

When the data (A, b) associated with the constraint is *uncertain*[1] and is only known to belong to some *uncertainty set* $\mathcal{U}$, we speak about an *uncertain conic problem*, which is in fact a *family* of conic problems:

$$\left\{\min_{x\in\mathbb{R}^n} \left\{c^T x \ : \ Ax - b \in \mathcal{K}\right\} \ : \ (A, b) \in \mathcal{U}\right\}. \tag{UCP}$$

The *robust optimization* (RO-) methodology, developed in [1, 2, 3, 4], associates with (UCP) a single deterministic convex problem, the so-called *robust counterpart* (RC):

$$\min_{x\in\mathbb{R}^n} \left\{c^T x \ : \ Ax - b \in \mathcal{K} \quad \forall\, (A, b) \in \mathcal{U}\right\}. \tag{RC}$$

*Received by the editors July 16, 2001; accepted for publication (in revised form) February 12, 2002; published electronically October 1, 2002.

http://www.siam.org/journals/siopt/13-2/39235.html.

†Faculty of Industrial Engineering and Management, Technion-Israel Institute of Technology, 32000 Haifa, Israel (morbt@ie.technion.ac.il, nemirovs@ie.technion.ac.il). The research for this paper was done while the first author spent a sabbatical as visiting professor at TU Delft, with the support of TU Delft and the Dutch Organization of Scientific Research (NWO).

‡Faculty of Information Technology and Systems, Delft University of Technology, P.O. Box 5031, 2600 GA Delft, The Netherlands (C.Roos@its.tudelft.nl).

[1]Without loss of generality we assume that the objective function is *certain*. Indeed, if c is uncertain, we can use the following equivalent formulation:

$$\min_{x\in\mathbb{R}^n} \Big\{t \ : \ Ax - b \in \mathcal{K}, \quad c^T x - t \le 0\Big\},$$

which is a conic problem with a certain objective function.

A feasible/optimal solution of (RC) is called a *robust feasible/optimal* solution of (UCP). The importance of these solutions is motivated and illustrated in [1, 2, 3, 4]. Of course, a crucial issue regarding the usefulness and applicability of the RO-methodology is the extent of the computational effort needed to solve problems such as (RC). At first glance, this looks hopeless, as (RC) is a *semi-infinite* conic problem. Nevertheless, for $\mathcal{K} = \mathbb{R}^N_+$ (the nonnegative orthant), i.e., when (CP) is a linear programming problem, the first two authors have shown [3] that for a very wide class of uncertainty sets $\mathcal{U}$ the resulting (RC)-problem is tractable (i.e., can be solved in time polynomial in the dimensions n, N of (CP)). This is also the case for conic-quadratic problems, i.e., when $\mathcal{K}$ is the Lorentz cone L^N:

$$L^N = \left\{ x \in \mathbb{R}^N \ : \ x_N \geq \sqrt{x_1^2 + \cdots + x_{N-1}^2} \right\},$$

provided that the uncertainty set is an ellipsoid (see [2]); the corresponding results are restated below in Theorems 2.1 and 3.2.

In this paper we deal with (conic) quadratic problems for which the uncertainty sets $\mathcal{U}$ are more general. In particular, we are interested in the case in which $\mathcal{U}$ is given as the intersection of *several* ellipsoids (we call this case the "$\cap$-ellipsoid" case). The situation is then severely aggravated: problem (RC) becomes NP-hard (see [2] and also section 2.2 below).

The goal of this paper is to build *approximate robust counterparts* for the above NP-hard problems, which are computationally tractable and for which a concise statement can be given on the quality of the approximation.

1.2. Approximate robust counterparts. The approximation scheme we use is of the *lift-and-project* type. Specifically, let the uncertainty set $\mathcal{U}$ be given as

$$\mathcal{U} = \left(A^0, b^0\right) + W,$$

where (A^0, b^0) is a *nominal* data vector and W is a compact convex set, symmetric with respect to the origin. (W is interpreted as the *perturbation set.*) Our aim is to approximate the set $\mathcal{X}$ of robust feasible solutions:

$$\mathcal{X} = \left\{ x \in \mathbb{R}^n \ : \ Ax - b \in K \quad \forall (A, b) \in \left(A^0, b^0\right) + W \right\}.$$

Towards this aim, we augment the vector x by an additional vector u and look at the following set $\mathcal{R}$, which is given by conic constraints:

$$\mathcal{R} := \{(x, u) \ : \ Px + Qu + r \in \hat{K}\}$$

in terms of some matrices P and Q, a vector r, and a closed convex pointed nonempty cone $\hat{K}$ with nonempty interior.

DEFINITION 1.1. *We say that $\mathcal{R}$ is an* approximate robust counterpart *of $\mathcal{X}$ if the projection of $\mathcal{R}$ onto the plane of x-variables, i.e., the set $\hat{\mathcal{R}} \subseteq \mathbb{R}^n$ given by*

$$\hat{\mathcal{R}} := \{x \ : \ Px + Qu + r \in \hat{K} \text{ for some } u\},$$

is contained in $\mathcal{X}$:

$$\hat{\mathcal{R}} \subseteq \mathcal{X}. \tag{1.1}$$

1.3. Level of conservativeness. Next we introduce a measure, called the *level of conservativeness*, for the proximity of $\hat{\mathcal{R}}$ to $\mathcal{X}$. To this end, let us look at an uncertainty set

$$\mathcal{U}_\rho = \{(A^0, b^0) + \rho W\}, \quad \rho \geq 1.$$

Compared to the original uncertainty set $\mathcal{U} = \mathcal{U}_1$, the perturbations in $\mathcal{U}_\rho$ are increased by a factor ρ. The set of robust feasible solutions corresponding to $\mathcal{U}_\rho$ is

$$\mathcal{X}_\rho := \{x \in \mathbb{R}^n \ : \ Ax - b \in K \quad \forall\, (A, b) \in \mathcal{U}_\rho\}.$$

Clearly, $\mathcal{X}_1 = \mathcal{X}$. As ρ increases from 1, the set $\mathcal{X}_\rho$ shrinks, and eventually we will have

$$\mathcal{X}_\rho \subseteq \hat{\mathcal{R}}. \tag{1.2}$$

The smallest ρ for which this occurs,

$$\rho^* = \inf_{\rho \geq 1} \{\rho \ : \ \mathcal{X}_\rho \subseteq \hat{\mathcal{R}}\}, \tag{1.3}$$

is called the *level of conservativeness* of the approximate counterpart $\mathcal{R}$. Thus we have

$$\mathcal{X}_{\rho^*} \subseteq \hat{\mathcal{R}} \subseteq \mathcal{X}.$$

The implications of this concept are twofold:

(i) If $x \in \hat{\mathcal{R}}$, i.e., if x *can* be augmented to a solution $(x, u) \in \mathcal{R}$, then x *is* a robust feasible solution of problem (CP). This follows from relation (1.1).

(ii) If $x \notin \hat{\mathcal{R}}$, i.e., if x *cannot* be augmented to a solution $(x, u) \in \mathcal{R}$, then x is *not* a robust feasible solution of problem (CP) if its uncertainty set $\mathcal{U}$ is increased to $\mathcal{U}_\rho$, with $\rho \geq \rho^*$.

In real-world applications, the level of uncertainty (the size of vectors in the perturbation set W) is not something that can be specified precisely by the decision maker; it is more likely that it will be specified *up to a factor of order* 1. Thus, for problems for which the level of conservativeness itself is of order 1, the approximate robust counterpart can be as meaningful as the *true* robust counterpart. The main results of the paper show that for *conic-quadratic problems* under "$\cap$-ellipsoid" uncertainty, this is indeed the case: we derive an explicit semidefinite program which is an approximate robust counterpart of the uncertain conic-quadratic problem and whose level of conservativeness is a constant, essentially independent of the dimensions n, N of (CP). The profound implication of these results is that the NP-hardness, associated with uncertain conic-quadratic problems, can be circumvented, and a computationally tractable tool is at hand, capable of producing robust solutions to these difficult problems.

1.4. Organization of the paper. We start in the next section by considering the case of an uncertain convex quadratically constrained problem. The more general case of conic-quadratic problems is considered in section 3. In both cases we first recall the results already known for simple-ellipsoid uncertainty from [2]. The main results concern the cases of $\cap$-ellipsoid uncertainty, and, as a special case of it, box uncertainty. In the box uncertainty case, we present robust counterparts for which

the level of conservativeness is bounded above by a constant, namely, $\pi/2$. The robust counterparts presented in the $\cap$-ellipsoid cases have level of conservativeness at most

$$\left(2\log\left(6\sum_{k=1}^{K}\operatorname{rank}Q_k\right)\right)^{\frac{1}{2}}. \tag{1.4}$$

The matrices Q_k in this expression are symmetric positive semidefinite matrices, of the same order L, and it will always be assumed that their sum is positive definite. Note that the expression under the logarithm in (1.4) may be as large as $6KL$. In many applications, however, it is likely to be much smaller.

2. Approximate robust counterparts of uncertain quadratically constrained problems. A generic convex quadratically constrained problem has the form

$$\min_{x\in\mathbb{R}^n}\left\{c_0^Tx\ :\ x^TA_i^TA_ix\le 2b_i^Tx+c_i,\quad i=1,\dots,m\right\}, \tag{QC}$$

where the matrices A_i have size $m_i\times n$. Note that (QC) can be written as a conic-quadratic problem:[2]

$$\min_{x\in\mathbb{R}^n}\left\{c_0^Tx\ :\ \left\|\begin{pmatrix}A_ix\\ \frac{1}{2}\left(1-2b_i^Tx-c_i\right)\end{pmatrix}\right\|\le\tfrac{1}{2}\left(1+2b_i^Tx+c_i\right),\quad i=1,\dots,m\right\}.$$

However, we shall treat the direct formulation (QC). An uncertain (QC)-problem corresponds to the case in which the data $\{(A_i,b_i,c_i)\ :\ i=1,\dots,m\}$ of the problem is uncertain. To model the uncertainty, we use *uncertainty sets* $\mathcal{U}_i$, and we assume

$$(A_i,b_i,c_i)\in\mathcal{U}_i,\quad i=1,\dots,m, \tag{2.1}$$

where the uncertainty set $\mathcal{U}_i$ associated with the ith constraint is given as the intersection of ellipsoids.

In order to construct the robust counterpart (RC) of problem (QC), we should be able to construct the robust counterpart of a single uncertain quadratic constraint

$$x^TA^TAx\le 2b^Tx+c\quad\forall\,(A,b,c)\in\mathcal{U}_\rho, \tag{UQC}$$

where $\mathcal{U}_\rho$ is the intersection of K ellipsoids; i.e., it is described as

$$\mathcal{U}_\rho=\left\{(A,b,c)=\left(A^0,b^0,c^0\right)+\sum_{\ell=1}^{L}y_\ell\left(A^\ell,b^\ell,c^\ell\right)\ :\ y\in\rho V\right\},$$

where V is the intersection of K ellipsoids,

$$V=\left\{y\in\mathbb{R}^L\ :\ y^TQ_ky\le 1,\quad k=1,\dots,K\right\},$$

and where each $Q_k\succeq 0$. As stated above, we make the generic assumption that $\sum_{k=1}^{K}Q_k\succ 0$.

[2]When not further specified, $\|.\|$ always denotes the 2-norm $\|.\|_2$.

2.1. Simple ellipsoidal uncertainty. In this case we look for the robust counterpart of the convex quadratic constraint

$$x^T A^T A x \le 2b^T x + c \quad \forall\, (A,b,c) \in \mathcal{U}_{\text{simple}}, \tag{2.2}$$

where

$$\mathcal{U}_{\text{simple}} = \left\{ (A,b,c) = \left(A^0,b^0,c^0\right) + \sum_{\ell=1}^{L} y_\ell \left(A^\ell,b^\ell,c^\ell\right) \ : \ \|y\|^2 \le 1 \right\},$$

with

$$A^\ell \in \mathbb{R}^{M\times n}, \quad b^\ell \in \mathbb{R}^n, \quad c^\ell \in \mathbb{R}, \qquad \ell = 0,\dots,L.$$

This is a special case of (UQC), where $K = 1$ and Q_1 is the identity matrix. This case has been considered already in [2, 6], where the following result is proved.

THEOREM 2.1. *A vector $x \in \mathbb{R}^n$ is a solution of (2.2) if and only if for some $\lambda \in \mathbb{R}$ the pair (x,λ) is a solution of the following linear matrix inequality (LMI):*

$$\left[\begin{array}{c|ccc|c} c^0 + 2x^T b^0 - \lambda & \frac{1}{2}c^1 + x^T b^1 & \cdots & \frac{1}{2}c^L + x^T b^L & \left(A^0 x\right)^T \\ \hline \frac{1}{2}c^1 + x^T b^1 & \lambda & & & \left(A^1 x\right)^T \\ \vdots & & \ddots & & \vdots \\ \frac{1}{2}c^L + x^T b^L & & & \lambda & \left(A^L x\right)^T \\ \hline A^0 x & A^1 x & \cdots & A^L x & I_M \end{array}\right] \succeq 0.$$

Fundamental in the proof of this result is the so-called S-lemma (see, e.g., [5]).

LEMMA 2.2 (S-lemma). *Let P and Q be symmetric matrices of the same order, and assume that $y^T P y > 0$ for some vector y. Then the implication*

$$z^T P z \ge 0 \quad \Rightarrow \quad z^T Q z \ge 0$$

is valid if and only if $Q \succeq \lambda P$ for some $\lambda \ge 0$.

2.2. Intersection-of-ellipsoids uncertainty. In this case we consider the robust feasible set for (UQC):

$$\mathcal{X}_\rho = \left\{ x \ : \ x^T A^T A x \le 2b^T x + c \quad \forall\, (A,b,c) \in \mathcal{U}_\rho \right\},$$

where

$$\mathcal{U}_\rho = \left\{ (A,b,c) = \left(A^0,b^0,c^0\right) + \rho \sum_{\ell=1}^{L} y_\ell \left(A^\ell,b^\ell,c^\ell\right) \ : \ y^T Q_k y \le 1, \quad k = 1,\dots,K \right\}.$$

Note that the robust counterpart of (UQC) with the $\cap$-ellipsoid uncertainty $\mathcal{U}_\rho$ is, in general, NP-hard. In fact, the associated *analysis* problem "given x, check whether it is robust feasible" is already NP-hard. To support our claim, we observe that the

analysis problem in question is at least as difficult as the problem of maximizing a positive definite quadratic form over the unit cube:

$$\text{(MAXQ)} \qquad \text{given } Q \succ 0 \text{ and } q \in \mathbb{R}, \text{ check whether } \max_{y:|y_\ell|\le 1} y^T Q y \le q;$$

the latter problem is known to be NP-hard.[3] Indeed, given data Q, q of (MAXQ) with a $K \times K$ matrix Q, let us find a $K \times K$ matrix D such that $D^T D = Q$ and associate with Q, q the following uncertainty set for (UQC):

$$\mathcal{U}_1 = \left\{ (A,b,c) = (0_{K\times K}, 0_{K\times 1}, q) + \sum_{\ell=1}^{K} y_\ell \left(A^\ell, 0_{K\times 1}, 0\right) \ : \ y_\ell^2 \le 1, \quad \ell = 1,\dots,K \right\},$$

where the first column of A_ℓ equals the ℓth column of D, and the remaining columns in A_ℓ are zero, $\ell = 1,\dots,K$. With this setup, one has $(A,b,c) = (Dy, 0_{(K-1)\times K}, 0_{K\times 1}, q)$. If $x = (1,0,0,\dots,0)^T \in \mathbb{R}^K$, then $Ax = Dy$, and hence checking whether x is robust feasible for (UQC) is exactly the same as checking whether $y^T Q y = \|Dy\|^2$ is $\le q$ for all y with $|y_\ell| \le 1$; thus, the NP-hard problem (MAXQ) is reducible to the analysis problem for (UQC) with a pretty simple $\cap$-ellipsoid uncertainty ("box uncertainty": $L = K$ and Q_k is the diagonal matrix with the only nonzero diagonal entry, equal to 1, in the cell (k,k)).

The NP-hardness of the robust counterpart of (UQC) in the presence of $\cap$-ellipsoid uncertainty motivates our current goal—to build a tractable approximate robust counterpart. We introduce some more convenient notations:

$$a[x] = A^0 x, \quad c[x] = 2x^T b^0 + c^0, \quad A_\rho[x] = \rho\left(A^1 x, \dots, A^L x\right),$$

$$b_\rho[x] = \rho \begin{bmatrix} x^T b^1 \\ \vdots \\ x^T b^L \end{bmatrix}, \qquad d_\rho = \tfrac{1}{2}\rho \begin{bmatrix} c^1 \\ \vdots \\ c^L \end{bmatrix}.$$

Then one may easily verify that $x \in \mathcal{X}_\rho$ holds if and only if

$$y^T Q_k y \le 1,\ k = 1,\dots,K \Rightarrow (a[x] + A_\rho[x]y)^T (a[x] + A_\rho[x]y) \le 2\left(b_\rho[x] + d_\rho\right)^T y + c[x].$$

The last inequality can be rewritten as

$$y^T A_\rho[x]^T A_\rho[x] y + 2y^T \left(A_\rho[x]^T a[x] - b_\rho[x] - d_\rho\right) \le c[x] - a[x]^T a[x].$$

Hence we obtain that $x \in \mathcal{X}_\rho$ holds if and only if

$$\begin{array}{c} y^T Q_k y \le 1,\ k = 1,\dots,K \quad \Rightarrow \\ y^T A_\rho[x]^T A_\rho[x] y + 2y^T \left(A_\rho[x]^T a[x] - b_\rho[x] - d_\rho\right) \le c[x] - a[x]^T a[x]. \end{array} \tag{2.3}$$

Observe that if y satisfies $y^T Q_k y \le 1$, then so does $-y$. Hence, $x \in \mathcal{X}_\rho$ holds if and only if

$$\begin{array}{c} y^T Q_k y \le 1,\ k = 1,\dots,K \quad \Rightarrow \\ y^T A_\rho[x]^T A_\rho[x] y \pm 2y^T \left(A_\rho[x]^T a[x] - b_\rho[x] - d_\rho\right) \le c[x] - a[x]^T a[x]. \end{array}$$

[3] From MAXCUT-related studies it is known that it is NP-hard even to approximate the maximum of a positive definite quadratic form over the unit cube within relative accuracy like 5% [7].

Therefore, we may replace the implication by

$$t^2 \le 1,\ y^T Q_k y \le 1,\ \ k = 1, \dots, K \quad \Rightarrow$$
$$y^T A_\rho[x]^T A_\rho[x] y + 2ty^T \left(A_\rho[x]^T a[x] - b_\rho[x] - d_\rho\right) \le c[x] - a[x]^T a[x].$$

This implication certainly holds if there exist $\lambda_k \ge 0$, $k = 1, \dots, K$, such that for all t and for all y

$$\sum_{k=1}^{K} \lambda_k y^T Q_k y + \left(c[x] - a[x]^T a[x] - \sum_{k=1}^{K} \lambda_k\right) t^2 \ge y^T A_\rho[x]^T A_\rho[x] y + 2ty^T \left(A_\rho[x]^T a[x] - b_\rho[x] - d_\rho\right).$$

We can equivalently express the last condition in a more concise form:

$$\begin{bmatrix} t \\ y \end{bmatrix}^T \begin{bmatrix} c[x] - a[x]^T a[x] - \sum_{k=1}^K \lambda_k & \left(A_\rho[x]^T a[x] - b_\rho[x] - d_\rho\right)^T \\ A_\rho[x]^T a[x] - b_\rho[x] - d_\rho & \sum_{k=1}^K \lambda_k Q_k - A_\rho[x]^T A_\rho[x] \end{bmatrix} \begin{bmatrix} t \\ y \end{bmatrix} \ge 0.$$

In other words, $x \in \mathcal{X}_\rho$ certainly holds if

$$\exists \lambda \ge 0 \ \text{ s.t. } \begin{bmatrix} c[x] - a[x]^T a[x] - \sum_{k=1}^K \lambda_k & \left(A_\rho[x]^T a[x] - b_\rho[x] - d_\rho\right)^T \\ A_\rho[x]^T a[x] - b_\rho[x] - d_\rho & \sum_{k=1}^K \lambda_k Q_k - A_\rho[x]^T A_\rho[x] \end{bmatrix} \succeq 0,$$

which can be rewritten as

$$\exists \lambda \ge 0 \ \text{ s.t. } \begin{bmatrix} c[x] - \sum_{k=1}^K \lambda_k & \left(-b_\rho[x] - d_\rho\right)^T \\ -b_\rho[x] - d_\rho & \sum_{k=1}^K \lambda_k Q_k \end{bmatrix} \succeq \begin{bmatrix} a[x]^T \\ -A_\rho[x]^T \end{bmatrix} \begin{bmatrix} a[x] & -A_\rho[x] \end{bmatrix}.$$

By the Schur complement lemma the latter is equivalent to

$$\exists \lambda \ge 0 \ \text{ s.t. } \begin{bmatrix} c[x] - \sum_{k=1}^K \lambda_k & \left(-b_\rho[x] - d_\rho\right)^T & a[x]^T \\ -b_\rho[x] - d_\rho & \sum_{k=1}^K \lambda_k Q_k & -A_\rho[x]^T \\ a[x] & -A_\rho[x] & I_M \end{bmatrix} \succeq 0.$$

Thus we have proved the following theorem.

THEOREM 2.3. *The set $\mathcal{R}_\rho$ of (x, λ) satisfying $\lambda \ge 0$ and*

$$\begin{bmatrix} c[x] - \sum_{k=1}^K \lambda_k & \left(-b_\rho[x] - d_\rho\right)^T & a[x]^T \\ -b_\rho[x] - d_\rho & \sum_{k=1}^K \lambda_k Q_k & -A_\rho[x]^T \\ a[x] & -A_\rho[x] & I_M \end{bmatrix} \succeq 0 \tag{2.4}$$

is an approximate robust counterpart of the set $\mathcal{X}_\rho$ of robust feasible solutions of (UQC).

Unlike the case in which $\mathcal{U}$ is a single ellipsoid, in the general case of $\cap$-ellipsoids we can no longer use the S-lemma (Lemma 2.2) to get an equivalence between the LMI (2.4) and the uncertain quadratic inequality (UQC). Thus another fundamental tool is needed, and this is offered by our so-called approximate S-lemma (cf. Lemma

A.6 in the appendix). With this tool we are able to derive the main results of this paper: Theorem 2.4, which follows, and Theorem 3.5 in the next subsection.

THEOREM 2.4. *The level of conservativeness of the approximate robust counterpart $\mathcal{R}$ (as given by Theorem 2.3) of the set $\mathcal{X}$ is at most*

$$\tilde{\rho} := \left(2\log\left(6\sum_{k=1}^{K}\operatorname{rank} Q_k\right)\right)^{\frac{1}{2}}. \tag{2.5}$$

Proof. We have to show that when x cannot be extended to a solution (x,λ) of (2.4), then there exists $\zeta_* \in \mathbb{R}^n$ such that

$$\zeta_*^T Q_k \zeta_* \le 1, \quad k = 1,\dots,K, \tag{2.6}$$

and

$$\tilde{\rho}^2 \zeta_*^T A_\rho[x]^T A_\rho[x]\zeta_* + 2\tilde{\rho}\zeta_*^T \left(A_\rho[x]^T a[x] - b_\rho[x] - d\right) \ge c[x] - a[x]^T a[x]. \tag{2.7}$$

The proof is based on Lemma A.6, which can be seen as an "approximate S-lemma." Using the notation of this lemma, let

$$R = \left[\begin{array}{c|c} 0 & \left(A_\rho[x]^T a[x] - b_\rho[x] - d\right)^T \\ \hline A_\rho[x]^T a[x] - b_\rho[x] - d & A_\rho[x]^T A_\rho[x] \end{array}\right],$$

$$R_0 = \left[\begin{array}{c|c} 1 & 0^T \\ \hline 0 & 0 \end{array}\right], \quad R_k = \left[\begin{array}{c|c} 0 & 0^T \\ \hline 0 & Q_k \end{array}\right],$$

and $r_0 = 1$. Note that $R_1,\dots,R_K$ are positive semidefinite, and, due to our generic assumption on the Q_k's,

$$R_0 + \sum_{k=1}^{K} R_k = \left[\begin{array}{c|c} 1 & 0^T \\ \hline 0 & \sum_{k=1}^{K} Q_k \end{array}\right] \succ 0.$$

Moreover, R_0 is dyadic and $r_0 = 1 > 0$. We are therefore in the situation in Lemma A.6 where R_0 is dyadic and $r_0 > 0$. Hence the estimate (2.5) is valid. We proceed by distinguishing two cases.

Case I. We assume in this case that there exist $\lambda_0, \dots, \lambda_K \ge 0$ such that

$$R \preceq \sum_{k=0}^{K} \lambda_k R_k, \tag{2.8}$$

$$\sum_{k=0}^{K} \lambda_k \le c[x] - a[x]^T a[x]. \tag{2.9}$$

Since the LMI (2.4) was shown to imply (2.3), our assumption that x cannot be extended to a solution of (2.4) implies that x cannot be extended to a solution of (2.3). On the other hand, by (2.8),

$$(t, y^T) R \begin{pmatrix} t \\ y \end{pmatrix} \le \sum_{k=0}^{K} \lambda_k (t, y^T) R_k \begin{pmatrix} t \\ y \end{pmatrix} \quad \forall t, y.$$

Hence, using the definition of R and R_k, with $t = 1$,

$$y^T A_\rho[x]^T A_\rho[x] y + 2y^T \left(A_\rho[x]^T a[x] - b_\rho[x] - d \right) \le \lambda_0 + \sum_{k=1}^{K} \lambda_k y^T Q_k y \le \sum_{k=0}^{K} \lambda_k$$

whenever $y^T Q_k y \le 1$, $k = 1, \ldots, K$. Therefore, by (2.9),

$$y^T A_\rho[x]^T A_\rho[x] y + 2y^T \left(A_\rho[x]^T a[x] - b_\rho[x] - d \right) \le c[x] - a[x]^T a[x],$$

showing that x *is* a solution of (2.3). Due to this contradiction, Case I cannot occur.

Case II. In this case there do not exist $\lambda_0, \ldots, \lambda_K \ge 0$ such that (2.8) and (2.9) hold. Hence, every feasible solution of problem (SDP) (in Lemma A.6) has objective value greater than $c[x] - a[x]^T a[x]$. Thus we have

$$\text{SDP} > c[x] - a[x]^T a[x]. \tag{2.10}$$

By Lemma A.6, there exists $y_* = (t_*, \eta_*)$ such that

$$y_*{}^T R_0 y_* = t_*^2 \le r_0 = 1, \tag{2.11}$$

$$y_*{}^T R_k y_* = \eta_*{}^T Q_k \eta_* \le \tilde{\rho}^2, \quad k = 1, \ldots, K, \tag{2.12}$$

$$\begin{aligned} y_*{}^T R y_* &= \eta_*^T A_\rho[x]^T A_\rho[x] \eta_* + 2t_* \eta_*^T \left(A_\rho[x]^T a[x] - b_\rho[x] - d \right) = \text{SDP} \\ &> c[x] - a[x]^T a[x], \end{aligned} \tag{2.13}$$

by (2.10). Setting $\bar{\eta} = \tilde{\rho}^{-1} \eta_*$, the last three relations become

$$\begin{cases} |t_*| \le 1, \\ \bar{\eta}^T Q_k \bar{\eta} \le 1, \quad k = 1, \ldots, K, \\ \tilde{\rho}^2 \bar{\eta}^T A_\rho[x]^T A_\rho[x] \bar{\eta} + 2\tilde{\rho} \bar{\eta}^T t_* \left(A_\rho[x]^T a[x] - b_\rho[x] - d \right) > c[x] - a[x]^T a[x]. \end{cases} \tag{2.14}$$

It is easily seen that if $(t_*, \bar{\eta})$ is a solution of (2.14), then either $\zeta_* = \bar{\eta}$ or $\zeta_* = -\bar{\eta}$ is a solution of (2.6)–(2.7).

This completes the proof of Theorem 2.4. ∎

2.3. Box uncertainty.

THEOREM 2.5. *Consider the uncertain quadratic constraint* (UQC), *where the uncertainty set is the "box"*

$$\mathcal{U}_\rho = \left\{ (A, b, c) = \left(A^0, b^0, c^0 \right) + \rho \sum_{\ell=1}^{L} y_\ell \left(A^\ell, b^\ell, c^\ell \right) \ : \ |y_\ell| \le 1, \quad \ell = 1, \ldots, L \right\}. \tag{2.15}$$

Then

(i) *the set* $\mathcal{R}_\rho$ *of* (x, λ) *satisfying* $\lambda \ge 0$ *and*

$$\begin{bmatrix} c[x] - \sum_{\ell=1}^{L} \lambda_\ell & (-b_\rho[x] - d)^T & a[x]^T \\ -b_\rho[x] - d & \operatorname{diag}(\lambda) & -A_\rho[x]^T \\ a[x] & -A_\rho[x] & I_M \end{bmatrix} \succeq 0 \tag{2.16}$$

is an approximate robust counterpart of the set $\mathcal{X}_\rho$ *of robust feasible solutions of* (UQC), *and*

(ii) *the level of conservativeness* Ω *of* $\mathcal{R}$ *is at most*

$$\Omega \le \frac{\pi}{2}. \tag{2.17}$$

Proof. Part (i) of the theorem is a special case of Theorem 2.3, with $K = L$ and where each Q_k is equal to a diagonal matrix whose only nonzero element is a 1 in the kth position of the diagonal. Thus it remains to prove part (ii). This proof will proceed in two steps. In Step 1 we build an approximate robust counterpart $\hat{\mathcal{R}}$ of (UQC), which is seemingly different from the $\mathcal{R}$ given in part (i) of the theorem, and we prove that the level of conservativeness of $\hat{\mathcal{R}}$ is $\pi/2$. In step 2 we demonstrate that $\hat{\mathcal{R}}$ is in fact equivalent to $\mathcal{R}$.

Step I. (Construction of $\hat{\mathcal{R}}$). The quadratic constraint

$$x^T A^T A x \le 2b^T x + c$$

is equivalent, by the Schur complement lemma, to the LMI

$$\begin{bmatrix} 2b^T x + c & (Ax)^T \\ Ax & I \end{bmatrix} \succeq 0.$$

Thus the robust feasible set of (UQC), $\mathcal{X}_\rho$, corresponding to the uncertainty set $\mathcal{U}_\rho$ in (2.15), is given by

$$\mathcal{X}_\rho = \left\{ x : \begin{bmatrix} 2x^T b^0 + c^0 \left(A^0 x\right)^T \\ A^0 x \qquad\qquad I \end{bmatrix} + \rho \sum_{\ell=1}^{L} y_\ell \begin{bmatrix} 2x^T b^\ell + c^\ell \left(A^\ell x\right)^T \\ A^\ell x \qquad\qquad 0 \end{bmatrix} \succeq 0,\ \|y\|_\infty \le 1 \right\}.$$

An evident sufficient condition for a vector x to belong to $\mathcal{X}_\rho$ is the possibility of extending x by L matrix variables $X^1, \ldots, X^L$, which together satisfy the following LMIs:

$$\begin{aligned} X^\ell &\succeq \pm\rho \begin{bmatrix} 2x^T b^\ell + c^\ell & \left(A^\ell x\right)^T \\ A^\ell x & 0 \end{bmatrix} \equiv \tilde{A}_\ell[x], \quad \ell = 1, \ldots, L, \\ \begin{bmatrix} 2x^T b^0 + c^0 & \left(A^0 x\right)^T \\ A^0 x & I \end{bmatrix} &\succeq \textstyle\sum_{\ell=1}^{L} X^\ell. \end{aligned} \tag{2.18}$$

The system (2.18) is the aforementioned approximate robust counterpart of $\hat{\mathcal{R}}$. The fact that the level of conservativeness of $\mathcal{R}$ is at most $\pi/2$ is then a direct consequence of [4, Theorem 4.4.1, p. 190]. In using the latter, note that $\operatorname{rank} \tilde{A}_\ell[x] = 2$.

Step II. ($\hat{\mathcal{R}}$ is equivalent to $\mathcal{R}$). The equivalence is shown in two parts:

II.1. If $(x, \lambda_1, \ldots, \lambda_K)$ is a solution of (2.16), then x can be extended to a solution $\left(x, X^1, \ldots, X^L\right)$ of (2.18).

II.2. If $\left(x, X^1, \ldots, X^L\right)$ a solution of (2.18), then for some $\lambda \ge 0$, $(x, \lambda_1, \ldots, \lambda_K)$ is a solution of (2.16).

The proofs of both parts rely on the following lemma, whose proof depends on Lemma A.8 in the appendix.

LEMMA 2.6. *Let $c \in \mathbb{R}$, $d \in \mathbb{R}^m$, and*

$$P = \begin{bmatrix} 2c & d^T \\ d & 0 \end{bmatrix}.$$

Then

(i) *for every* $\lambda > 0$ *the matrix*

$$Y[\lambda, P] := \begin{bmatrix} \lambda + \frac{c^2}{\lambda} & \frac{cd^T}{\lambda} \\ \frac{cd}{\lambda} & \frac{dd^T}{\lambda} \end{bmatrix} = \begin{bmatrix} \lambda & 0 \\ 0 & 0 \end{bmatrix} + \frac{1}{\lambda} \begin{bmatrix} c \\ d \end{bmatrix} \begin{bmatrix} c \\ d \end{bmatrix}^T \tag{2.19}$$

belongs to the set

$$\mathcal{L}[P, -P] = \{X \ : \ X \succeq P, X \succeq -P\}; \tag{2.20}$$

(ii) *if* $P \neq 0$, *then for every* $X \in \mathcal{L}[P, -P]$ *there exists* $\lambda > 0$ *such that* $X \succeq Y[\lambda, P]$.

Proof. Setting

$$a = (1, 0, \ldots, 0)^T, \qquad b = (c, d_1, \ldots, d_m)^T,$$

one has

$$ab^T + ba^T = \begin{pmatrix} 2c & d^T \\ d & 0 \end{pmatrix}, \quad \lambda aa^T + \frac{1}{\lambda} bb^T = \begin{bmatrix} \lambda & 0 \\ 0 & 0 \end{bmatrix} + \frac{1}{\lambda} \begin{bmatrix} c \\ d \end{bmatrix} \begin{bmatrix} c \\ d \end{bmatrix}^T = Y[\lambda, P].$$

Hence, Lemma 2.6 immediately follows from Lemma A.8. □

Proof of II.1. For our case of box uncertainty we have $(Q_k)_{kk} = 1$ and $(Q_k)_{ij} = 0$ ($i \neq k$ or $j \neq k$), and so the system (2.16) reduces (by the Schur complement lemma) to

$$\begin{bmatrix} c[x] - \sum_{\ell=1}^{L} \lambda_\ell & a[x]^T \\ a[x] & I_M \end{bmatrix} \succeq \sum_{\ell,\, \lambda_\ell > 0} \frac{1}{\lambda_\ell} \left(f_\ell[x] f_\ell[x]^T \right), \tag{2.21}$$

where $\lambda_\ell \geq 0$, $\ell = 1, \ldots, L$, with $\lambda_\ell = 0 \Rightarrow f_\ell[x] = 0$, and $f_1[x], \ldots, f_L[x]$ are the columns of the matrix

$$\begin{bmatrix} (b_\rho[x] + d)^T \\ A_\rho[x]^T \end{bmatrix},$$

i.e.,

$$f_\ell[x] = \begin{bmatrix} \rho x^T b^\ell + \frac{\rho}{2} c^\ell \\ \rho A^\ell x \end{bmatrix} = \rho \begin{bmatrix} x^T b^\ell + \frac{1}{2} c^\ell \\ A^\ell x \end{bmatrix}. \tag{2.22}$$

We rewrite (2.21) as

$$\begin{bmatrix} c[x] & a[x]^T \\ a[x] & I_M \end{bmatrix} \succeq \sum_{\ell,\, \lambda_\ell > 0} \left(\begin{bmatrix} \lambda_\ell & 0 \\ 0 & 0 \end{bmatrix} + \frac{1}{\lambda_\ell} \left(f_\ell[x] f_\ell[x]^T \right) \right),$$

or, more explicitly,

$$\begin{bmatrix} 2x^T b^0 + c^0 & \left(A^0 x\right)^T \\ A^0 x & I_M \end{bmatrix} \succeq \sum_{\ell,\, \lambda_\ell > 0} \left(\begin{bmatrix} \lambda_\ell & 0 \\ 0 & 0 \end{bmatrix} + \frac{1}{\lambda_\ell} \left(f_\ell[x] f_\ell[x]^T \right) \right). \tag{2.23}$$

Note that the matrix under the sum has the form of the matrix $Y[\lambda, P]$ in (2.19). Hence, denoting this matrix as X_ℓ whenever $\lambda_\ell > 0$, we may conclude from Lemma 2.6 that

$$\lambda_\ell > 0 \Rightarrow X_\ell \succeq \pm\rho \begin{bmatrix} 2x^T b^\ell + c^\ell & \left(A^\ell x\right)^T \\ A^\ell x & 0 \end{bmatrix}. \tag{2.24}$$

Setting $X_\ell = 0$ whenever $\lambda_\ell = 0$, and using (2.23), we ensure that x, $X^1, \ldots, X^L$ is a solution of (2.18). This proves II.1.

Proof of II.2. Assume that x can be extended to a solution x, $X^1, \ldots, X^L$ of (2.18). By Lemma 2.6, for those ℓ's for which

$$\rho \begin{bmatrix} 2x^T b^\ell + c^\ell & \left(A^\ell x\right)^T \\ A^\ell x & 0 \end{bmatrix} \neq 0 \tag{2.25}$$

there exist $\lambda_\ell > 0$ such that

$$X^\ell \succeq Y^\ell := \begin{bmatrix} \lambda_\ell & 0 \\ 0 & 0 \end{bmatrix} + \frac{1}{\lambda_\ell}\left(f_\ell[x] f_\ell[x]^T\right) \succeq \rho \begin{bmatrix} 2x^T b^\ell + c^\ell & \left(A^\ell x\right)^T \\ A^\ell x & 0 \end{bmatrix}, \tag{2.26}$$

where the vectors $f_\ell[x]$ are as defined by (2.22). Setting $Y^\ell = 0$ and $\lambda_\ell = 0$ whenever the left-hand side in (2.25) vanishes, it follows that x, $Y^1, \ldots, Y^L$ is a feasible solution of (2.18), which in turn implies

$$\begin{bmatrix} 2x^T b^0 + c^0 & \left(A^0 x\right)^T \\ A^0 x & I \end{bmatrix} \succeq \sum_{\ell=1}^{L} Y^\ell = \sum_{\ell,\, \lambda_\ell > 0} \left(\begin{bmatrix} \lambda_\ell & 0 \\ 0 & 0 \end{bmatrix} + \frac{1}{\lambda_\ell}\left(f_\ell[x] f_\ell[x]^T\right) \right).$$

Via the Schur complement lemma (note that $\lambda_\ell = 0 \Rightarrow f_\ell[x] = 0$), the latter LMI shows that $(x, \lambda_1, \ldots, \lambda_L)$ is a feasible solution of (2.4). This completes the proof of II.2, and thus of Theorem 2.5. □

3. Robust solutions of uncertain conic-quadratic problems. An uncertain conic-quadratic problem (CQP) has the form

$$\min_{x \in \mathbb{R}^n} \left\{ c^T x \;:\; \left\| A^i x + b_i \right\| \leq a_i^T x + \beta_i, \quad i = 1, \ldots, m \right\}, \tag{CQP}$$

where the data (A^i, b_i, a_i, β_i) is uncertain and is only known to belong to some uncertainty sets $\mathcal{U}_i$,

$$\left(A^i, b_i, a_i, \beta_i\right) \in \mathcal{U}_i, \quad i = 1, \ldots, m.$$

The crucial step in building a robust counterpart for (CQP) is the ability to build a robust counterpart for a single constraint, i.e., the set of solutions $x \in \mathbb{R}^n$ of the semi-infinite inequality system

$$\|Ax + b\| \leq a^T x + \beta, \quad (A, b, a, \beta) \in \mathcal{U}_\rho. \tag{3.1}$$

Here, we deal with the situation in which the uncertainty affecting (3.1) is *sidewise*, i.e., the uncertainty affecting the right-hand side in (3.1) is independent of that affecting the left-hand side. More specifically,

$$\mathcal{U}_\rho = \mathcal{U}_\rho^L \times \mathcal{U}_\rho^R, \tag{3.2}$$

where

$$\mathcal{U}_\rho^L = \left\{ (A,b) = \left(A^0, b^0\right) + \sum_{\ell=1}^{L} y_\ell \left(A^\ell, b^\ell\right) \ : \ y \in \rho \mathcal{V}^L \right\}, \tag{3.3}$$

$$\mathcal{U}_\rho^R = \left\{ (a,\beta) = \left(a^0, \beta^0\right) + \sum_{r=1}^{R} \xi_r \left(a^r, \beta^r\right) \ : \ \xi \in \rho \mathcal{V}^R \right\}. \tag{3.4}$$

The sets $\mathcal{V}^L$ and $\mathcal{V}^R$ are convex *perturbation sets*, and $\rho > 0$ is a parameter expressing the magnitude of the perturbation.

As before, the specific form of $\mathcal{V}^L$ is an *intersection of ellipsoids*, i.e., $\mathcal{V}^L = \mathcal{V}_K^L$, where

$$\mathcal{V}_K^L = \left\{ y \in \mathbb{R}^L \ : \ y^T Q_k y \le 1, \, k = 1, \ldots, K \right\}, \tag{3.5}$$

with

$$Q_k \succeq 0 \quad \text{and} \quad \sum_{k=1}^{K} Q_k \succ 0. \tag{3.6}$$

The form (3.5) includes two important special cases, namely,

- *simple ellipsoidal uncertainty* ($K = 1$),
- *box uncertainty* ($K = L$, $(Q_k)_{kk} = 1$, and $(Q_k)_{ij} = 0$ ($i \ne k$ or $j \ne k$) for $k = 1, \ldots, K$).

For the right-hand side perturbation set $\mathcal{V}^R$ we allow a much more general geometry: $\mathcal{V}^R$ is assumed to be bounded, containing zero, and *semidefinite representable* (sdr), i.e., it can be represented as the projection of a set described by LMIs:

$$\mathcal{V}^R = \left\{ \zeta \in \mathbb{R}^R \ : \ \exists u \in \mathbb{R}^S \ : \ P(\zeta) + Q(u) - T \succeq 0 \right\} \tag{3.7}$$

for some symmetric matrix T, and symmetric matrices $P(\zeta), Q(u)$, which depend linearly on their respective arguments. Specifically,

$$P(\zeta) = \sum_{r=1}^{R} \zeta_r P_r, \qquad Q(u) = \sum_{s=1}^{S} u_s Q_s, \tag{3.8}$$

where $P_r (r = 1, \ldots, R)$ and $Q_s (s = 1, \ldots, S)$ are symmetric matrices, and it is assumed that

$$\exists \bar{\zeta}, \, \bar{u} \ : \ P(\bar{\zeta}) + Q(\bar{u}) \succ T. \tag{3.9}$$

It is well known that sdr-sets include $\cap$-ellipsoids and many more [4, Lecture 4].

The sidewise uncertainty assumption implies the following fact: x is robust feasible for (3.1) if and only if there exists a τ such that

$$\|Ax + b\| \le \tau, \tag{3.10}$$

$$\tau \le a^T x + \beta. \tag{3.11}$$

This fact allows us to handle (3.1) by treating (3.10) and (3.11) separately. We start with (3.11).

THEOREM 3.1. *A pair* (x, τ) *satisfies* (3.11), *where* $\mathcal{U}_\rho^R$ *is given by* (3.4), (3.7), *and* (3.8), *if and only if, for some symmetric matrix* V, *the triple* (x, τ, V) *is a solution of the following system of LMIs:*

$$x^T a^0 + \beta^0 + \text{Tr}\,(TV) \geq \tau, \tag{3.12}$$

$$\text{Tr}\,(VP_r) = \rho\left(x^T a^r + \beta^r\right), \quad r = 1, \ldots, R, \tag{3.13}$$

$$\text{Tr}\,(VQ_s) = 0, \quad s = 1, \ldots, S, \tag{3.14}$$

$$V \succeq 0. \tag{3.15}$$

Proof. The pair (x, τ) satisfies (3.11) if and only if

$$\tau \leq x^T \left(a^0 + \rho \sum_{r=1}^R \xi_r a^r\right) + \beta^0 + \rho \sum_{r=1}^R \xi_r \beta^r \quad \forall \xi \in \mathcal{V}^R,$$

which is equivalent to

$$\tau - x^T a^0 - \beta^0 \leq \inf_{\xi,u} \left\{ \sum_{r=1}^R \xi_r \rho\left(x^T a^r + \beta^r\right) \;:\; P(\xi) + Q(u) \succeq T \right\}. \tag{3.16}$$

The problem on the right-hand side (rhs) of (3.16) is a semidefinite problem:

$$\inf_{\xi,u} \left\{ \xi^T \gamma_r[x] \;:\; P(\xi) + Q(u) \succeq T \right\}, \tag{P}$$

where

$$\gamma_r[x] = \rho\left(x^T a^r + \beta^r\right).$$

The dual problem of (P) is the semidefinite problem

$$\sup_V \{\text{Tr}\,(TV) \;:\; P^*(V) = \gamma_r[x],\, Q^*(V) = 0,\, V \succeq 0\}, \tag{D}$$

where P^* and Q^* are the respective adjoints of P and Q, as given in (3.8). Thus

$$P^*(V) \in \mathbb{R}^R, \qquad P^*(V)_r = \text{Tr}\,(VP_r), \quad r = 1, \ldots, R,$$

$$Q^*(V) \in \mathbb{R}^S, \qquad Q^*(V)_s = \text{Tr}\,(VQ_s), \quad s = 1, \ldots, S.$$

By assumption (3.9), problem (P) is strictly feasible, and due to the assumption that the set $\mathcal{V}^R$ is bounded, the objective value of (P) is bounded from below. Hence, by SDP duality theory (see, e.g., [4]), problem (D) has an optimal solution and $\inf$ (P) $=$ $\max$ (D), i.e., there exists a V such that

$$\begin{aligned} \text{rhs of (3.16)} &= \inf(\text{P}) = \text{Tr}\,(TV), \\ \text{Tr}\,(VP_r) &= \gamma_r[x] = \rho\left(x^T a^r + \beta^r\right), \quad r = 1, \ldots, R, \\ \text{Tr}\,(VQ_s) &= 0, \quad s = 1, \ldots, S, \\ V &\succeq 0. \end{aligned} \tag{3.17}$$

Now (3.16) and (3.17) show that (x, τ, V) indeed satisfies (3.12)–(3.15). ∎

We now turn to the condition (3.10) with the uncertainty set $\mathcal{U}_\rho^L$ given by (3.3), (3.5), and (3.6). For a general perturbation set as given in (3.5), with $K > 1$, the

verification of (3.10) is an NP-hard problem (see [2]). Therefore we shall derive an approximate robust counterpart (Theorem 3.3 below). For the simple ellipsoidal case ($K = 1$) an exact robust counterpart is given by the following result of [2].

THEOREM 3.2. *Consider the condition* (3.10)*, where* $\mathcal{U}_\rho^L$ *is given by* (3.3) *and* $\mathcal{V}^L$ *is the ellipsoid* $\mathcal{V}_1^L$ *(see* (3.5)*). Then a pair* (x, τ) *satisfies* (3.10) *if and only if there exists some* $\lambda_1 \geq 0$ *such that the triple* (x, τ, λ_1) *satisfies the following LMI:*

$$\begin{bmatrix} \tau - \lambda_1 & 0 & a[x]^T \\ 0 & \lambda_1 Q_1 & \rho A[x]^T \\ a[x] & \rho A[x] & \tau I_M \end{bmatrix} \succeq 0, \tag{3.18}$$

where

$$a[x] = A^0 x + b^0 \tag{3.19}$$

$$A[x] = \left(A^1 x + b^1, \ldots, A^L x + b^L\right). \tag{3.20}$$

□

For the general ∩-ellipsoids case ($K > 1$), the following theorem gives an approximate robust counterpart of (3.10).

THEOREM 3.3. *The set* $\mathcal{S}_L$ *of triples* $(x, \tau, \lambda) \in \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^K$ *satisfying the LMI*

$$\begin{bmatrix} \tau - \sum_{k=1}^K \lambda_k & 0 & a[x]^T \\ 0 & \sum_{k=1}^K \lambda_k Q_k & \rho A[x]^T \\ a[x] & \rho A[x] & \tau I_M \end{bmatrix} \succeq 0, \quad \lambda \geq 0, \tag{3.21}$$

with $a[x]$ *and* $A[x]$ *as given by* (3.19)–(3.20)*, is an approximate robust counterpart of the set of pairs* (x, τ) *satisfying* (3.10)*, under the uncertainty set* $\mathcal{U}_\rho^L$ *given by* (3.3) *and* (3.5).

Proof. We have to show that if (x, τ, λ) solves (3.21), then (x, τ) solves (3.10). Now (3.21) is equivalent to the following three conditions:

(i)
$$Y := \begin{bmatrix} \mu & 0 & a[x]^T \\ 0 & \sum_{k=1}^K \lambda_k Q_k & \rho A[x]^T \\ a[x] & \rho A[x] & \tau I_M \end{bmatrix} \succeq 0,$$

(ii) $\mu \geq 0$, $\lambda \geq 0$,

(iii) $\mu + \sum_{k=1}^K \lambda_k \leq \tau$.

Condition (i) implies that for every $y \in \mathbb{R}^L$ and $t \in \mathbb{R}$

$$\begin{bmatrix} t & y^T & 0 \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} \mu & 0 & a[x]^T \\ 0 & \sum_{k=1}^K \lambda_k Q_k & \rho A[x]^T \\ a[x] & \rho A[x] & \tau I_M \end{bmatrix} \begin{bmatrix} t & y^T & 0 \\ 0 & 0 & I \end{bmatrix}^T \succeq 0,$$

which is equivalent to

$$\left[\begin{array}{cc|c} \begin{bmatrix} t & y^T \end{bmatrix} \begin{bmatrix} \mu & 0 \\ 0 & \sum_{k=1}^K \lambda_k Q_k \end{bmatrix} \begin{bmatrix} t \\ y \end{bmatrix} & & \begin{bmatrix} t & y^T \end{bmatrix} \begin{bmatrix} a[x]^T \\ \rho A[x]^T \end{bmatrix} \\ \hline \multicolumn{2}{c|}{\begin{bmatrix} a[x] & \rho A[x] \end{bmatrix} \begin{bmatrix} t \\ y \end{bmatrix}} & \tau I \end{array}\right] \succeq 0.$$

By the Schur complement lemma, the latter is equivalent to

(i′) $\tau(\mu t^2 + \sum_{k=1}^K \lambda_k y^T Q_k y) \geq \|ta[x] + \rho A[x]y\|^2$.

Therefore, conditions (i)–(iii) reduce to (i′), (ii), and (iii). From these conditions it follows that if (y, t) are chosen such that

$$t^2 \leq 1, \quad y^T Q_k y \leq 1, \qquad k = 1, \ldots, K, \tag{3.22}$$

then

$$\mu t^2 + \sum_{k=1}^K \lambda_k y^T Q_k y \leq \mu + \sum_{k=1}^K \lambda_k \leq \tau, \tag{3.23}$$

and, since $\tau \geq 0$, from (3.23) and (i′),

$$\tau^2 \geq \|ta[x] + \rho A[x]y\|^2 \quad \forall (y, t) \text{ satisfying (3.22)}. \tag{3.24}$$

In particular, for $t = 1$ we get

$$\tau \geq \|a[x] + \rho A[x]y\| \quad \forall y \text{ satisfying (3.22)}. \tag{3.25}$$

Substituting into (3.25) the expression for $a[x]$ and $A[x]$ (see (3.19)–(3.20)), (3.25) becomes explicitly

$$\tau \geq \left\| A^0 x + b^0 + \rho \sum_{\ell=1}^L \lambda_\ell \left(A^\ell x + b^\ell \right) \right\| \quad \forall y \text{ s.t. } y^T Q_k y \leq 1, \ k = 1, \ldots, K. \tag{3.26}$$

Finally, (3.26) is precisely condition (3.10) for $\mathcal{U}_\rho^L$ given by (3.3) and $\mathcal{V}^L = \mathcal{V}_K^L$ given by (3.5). □

Combining the results of Theorems 3.1 and 3.3, we obtain the following result.

COROLLARY 3.4. *The set* $\mathcal{S}$ *of tuples*

$$(x, \tau, \lambda, V) \ \textit{satisfying} \ (3.12)\text{–}(3.15) \ \textit{and} \ (3.21)$$

is an approximate robust counterpart of the uncertain conic-quadratic constraint (3.1), *where the uncertainty set* $\mathcal{U}_\rho$ *is given by* (3.2)–(3.9).

The level of conservativeness of $\mathcal{S}$ can be estimated in a way very similar to that used in the case of uncertain quadratic constraints (see the proofs of Theorem 2.4 and Theorem 2.5), and the result is in fact similar.

THEOREM 3.5. (i) *For the case of* $\cap$*-ellipsoidal uncertainty, with* $K > 1$, *the level of conservativeness* Ω *of the approximate robust counterpart* $\mathcal{S}$ *in Corollary* 3.4 *is at most*

$$\tilde{\Omega} := \left(2 \log \left(6 \sum_{k=1}^K \operatorname{rank} Q_k \right) \right)^{\frac{1}{2}}.$$

(ii) *For the special case of box uncertainty, one has*

$$\Omega \leq \frac{\pi}{2}. \qquad \square$$

Appendix. Some technical lemmas.

LEMMA A.1. *Let* $x = (x_1, \ldots, x_n)$ *and* $\xi = (\xi_1, \ldots, \xi_n) \in \mathbb{R}^n$. *If* $\|x\|_2 = 1$ *and the coordinates* ξ_i *of* ξ *are independently identically distributed random variables with*

$$\Pr(\xi_i = 1) = \Pr(\xi_i = -1) = \frac{1}{2},$$

then one has

$$\Pr\left(\left|\xi^T x\right| \leq 1\right) \geq \frac{1}{3}. \tag{A.1}$$

Proof.[4] Without loss of generality we may assume that $x \geq 0$ and

$$x_1 \geq x_2 \geq \cdots \geq x_n \geq 0.$$

We define $\theta = x_1$ and

$$s_0 = 0, \quad s_k = s_k(\xi) = \sum_{i=1}^{k} x_i \xi_i, \qquad k = 1, 2, \ldots, n.$$

Then (A.1) is equivalent to $\Pr(|s_n| \leq 1) \geq \frac{1}{3}$. To prove this, we define the following events:

$$A_0 = \{\xi \ : \ |s_j| \leq 1 - \theta, \ j = 1, \ldots, n\},$$

$$A_k = \{\xi \ : \ |s_j| \leq 1 - \theta, \ j = 1, \ldots, k-1, \text{ and } |s_k| > 1 - \theta\}.$$

Note that the events $A_0, A_1, \ldots, A_n$ form a partition of the probability space.

Assuming $A_k \neq \emptyset$, we proceed by deriving a lower bound on the probability that $|s_n| \leq 1$ occurs, namely:

$$\Pr(|s_n| \leq 1 \mid A_k) \geq p(\theta) := \frac{1}{2}\left(1 - \frac{1 - \theta^2}{(2 - \theta)^2}\right). \tag{A.2}$$

For $k = 0$ this is evident, since the left-hand side is then equal to 1. So let $k \geq 1$ and let us fix a realization $\xi \in A_k$. Then we have

$$1 - \theta < |s_k| \leq 1. \tag{A.3}$$

Indeed, the left-hand side of (A.3) follows from the definition of A_k, and the right-hand side from

$$|s_k| = |s_{k-1} + s_k - s_{k-1}| \leq |s_{k-1}| + |s_k - s_{k-1}| \leq 1 - \theta + x_k \leq 1 - \theta + \theta = 1.$$

Because of (A.3) we have the following implication:

$$0 \geq (s_n - s_k)\ \operatorname{sign}(s_k) \geq -2 + \theta \quad \Rightarrow \quad |s_n| \leq 1. \tag{A.4}$$

Indeed, if $s_k \geq 0$, then $1 - \theta < s_k \leq 1$ and $0 \geq s_n - s_k \geq -2 + \theta$ imply that $s_n \leq s_k \leq 1$ and $s_n \geq s_k + \theta - 2 > 1 - \theta + \theta - 2 = -1$, and if $s_k \leq 0$, then $1 - \theta < -s_k \leq 1$ and

[4] This proof is mainly due to P. van der Wal, Delft University of Technology (private communication).

$0 \geq -s_n + s_k \geq -2+\theta$ imply $s_n \geq s_k \geq -1$ and also $s_n \leq s_k - \theta + 2 < \theta - 1 - \theta + 2 = 1$. So in both cases one has $|s_n| \leq 1$, proving (A.4). Hence we may write

$$\begin{aligned}
\Pr\left(|s_n| \leq 1 \mid A_k\right) &\geq \Pr\left(0 \geq (s_n - s_k)\ \operatorname{sign}(s_k) \geq -2 + \theta \mid A_k\right) \\
&\geq \frac{1}{2} \Pr\left(|s_n - s_k| \leq 2 - \theta\right), \quad \text{by symmetry,} \\
&\geq \frac{1}{2}\left(1 - \frac{\operatorname{Var}(s_n - s_k)}{(2-\theta)^2}\right), \quad \text{by the Chebyshev inequality,} \\
&= \frac{1}{2}\left(1 - \frac{\sum_{j=k+1}^{n} x_k^2}{(2-\theta)^2}\right) \geq \frac{1}{2}\left(1 - \frac{1-\theta^2}{(2-\theta)^2}\right) = p(\theta).
\end{aligned}$$

Thus we have proved (A.2). Since $(A_0, A_1, \ldots, A_n)$ is a partition of the probability space, it follows that

$$\Pr\left(|s_n| \leq 1\right) \geq p(\theta) \geq \min_{0 \leq \theta \leq 1} \frac{1}{2}\left(1 - \frac{1-\theta^2}{(2-\theta)^2}\right) = p\left(\frac{1}{2}\right) = \frac{1}{3}.$$

This proves the lemma. ∎

Based on numerical experiments, we believe that Lemma A.1 can be improved as stated in the conjecture below.

CONJECTURE A.2. *Let x and ξ be as defined in Lemma* A.1. *Then*

$$\Pr\left(\left|\xi^T x\right| \leq 1\right) \geq \frac{1}{2}.$$

LEMMA A.3. *Let* $\operatorname{rank} B = k$, $B \succeq 0$, *and let ξ be as defined in Lemma* A.1. *Then*

$$\Pr\left(\xi^T B \xi \geq \alpha \operatorname{Tr} B\right) \leq 2ke^{-\frac{\alpha}{2}} \quad \forall \alpha > 0.$$

Proof. Writing

$$B = \sum_{i=1}^{k} b_i b_i^T,$$

the statement in the lemma can be rewritten as

$$\Pr\left(\sum_{i=1}^{k} \left(b_i^T \xi\right)^2 \geq \alpha \sum_{i=1}^{k} \|b_i\|^2\right) \leq 2ke^{-\frac{\alpha}{2}} \quad \forall \alpha > 0.$$

We have

$$\begin{aligned}
p_k := \Pr\left(\sum_{i=1}^{k} \left(b_i^T \xi\right)^2 \geq \alpha \sum_{i=1}^{k} \|b_i\|^2\right) &= \Pr\left(\sum_{i=1}^{k} \left(\left(b_i^T \xi\right)^2 - \alpha \|b_i\|^2\right) \geq 0\right) \\
\leq \Pr\left(\max_i \left(\left(b_i^T \xi\right)^2 - \alpha \|b_i\|^2\right) \geq 0\right) &\leq \sum_{i=1}^{k} \Pr\left(\left(b_i^T \xi\right)^2 \geq \alpha \|b_i\|^2\right) \\
= \sum_{i=1}^{k} \Pr\left(\left|b_i^T \xi\right| \geq \sqrt{\alpha}\, \|b_i\|\right) &= 2 \sum_{i=1}^{k} \Pr\left(b_i^T \xi \geq \sqrt{\alpha}\, \|b_i\|\right).
\end{aligned}$$

For any random variable y with distribution F, we have for any $\rho \geq 0$,

$$\mathcal{E}\left(e^{\rho y}\right) = \int_{-\infty}^{0} e^{\rho y} dF(y) + \int_{0}^{\infty} e^{\rho y} dF(y) \geq 0 + \int_{0}^{\infty} dF(y) = \Pr\left(y \geq 0\right).$$

Hence,

$$\Pr\left(b_i^T \xi \geq \sqrt{\alpha}\, \|b_i\|\right) \leq \mathcal{E}\left(e^{\rho(b_i^T \xi - \sqrt{\alpha}\|b_i\|)}\right) = \mathcal{E}\left(e^{\rho\,(b_i^T \xi)}\right) e^{-\rho\,\sqrt{\alpha}\,\|b_i\|}.$$

Furthermore, using the inequality $\cosh t \leq e^{\frac{1}{2}t^2}$, we have

$$\mathcal{E}\left(e^{\rho(b_i^T \xi)}\right) = \prod_{j=1}^{n} \mathcal{E}(e^{\rho b_{ij} \xi_j}) = \prod_{j=1}^{n} \cosh\left(\rho b_{ij}\right) \leq \prod_{j=1}^{n} e^{\frac{1}{2}\rho^2 b_{ij}^2} = e^{\frac{1}{2}\rho^2 \|b_i\|^2}.$$

Substitution gives

$$\Pr\left(b_i^T \xi \geq \sqrt{\alpha}\, \|b_i\|\right) \leq e^{\frac{1}{2}\rho^2 \|b_i\|^2 - \rho\sqrt{\alpha}\|b_i\|}, \quad \rho \geq 0.$$

The right-hand side is minimal if $\rho = \sqrt{\alpha}/\|b_i\|$. Thus we obtain

$$\Pr\left(b_i^T \xi \geq \sqrt{\alpha}\, \|b_i\|\right) \leq e^{-\frac{\alpha}{2}}.$$

From this we derive the inequality

$$p_k \leq 2 \sum_{i=1}^{k} \Pr\left(b_i^T \xi \geq \sqrt{\alpha}\, \|b_i\|\right) \leq 2 \sum_{i=1}^{k} e^{\frac{-\alpha}{2}} = 2ke^{-\frac{\alpha}{2}},$$

which completes the proof of the lemma. □

LEMMA A.4. *Let B denote a symmetric $n \times n$ matrix and ξ be as defined in Lemma* A.1. *Then*

$$\Pr\left(\xi^T B \xi \leq \operatorname{Tr} B\right) > \frac{1}{8n^2}. \tag{A.5}$$

Proof. Consider the random variable

$$\gamma := \sum_{i<j} \xi_i \xi_j A_{ij} = \frac{1}{2}(\xi^T B \xi - \operatorname{Tr} B).$$

Then (A.5) is equivalent to

$$\omega := \Pr\left(\gamma \leq 0\right) > \frac{1}{8n^2}. \tag{A.6}$$

Let $\mu(dt)$ be the distribution of γ, and let

$$I_\ell = \int_{-\infty}^{\infty} |t|^\ell \, \mu(dt), \quad \ell = 1, 2, \ldots.$$

Since $\mathcal{E}\left(\gamma\right) = 0$, we have

$$\int_{t \leq 0} |t|\, \mu(dt) = \int_{t \geq 0} |t|\, \mu(dt).$$

Hence

$$I_1 = 2\int_{t\le 0} |t|\,\mu(dt) = 2\int_{t\le 0} |t|\,\frac{\mu(dt)}{\Pr(\gamma\le 0)} \times \Pr(\gamma\le 0)$$
$$\le 2\left(\int_{t\le 0} t^2\mu(dt)\right)^{\frac{1}{2}} \sqrt{\Pr(\gamma\le 0)} \le 2\omega^{\frac{1}{2}} I_2^{\frac{1}{2}}.$$

Further,

$$I_2 = \int_{-\infty}^{\infty} t^2\mu(dt) = \int_{-\infty}^{\infty} |t|^{\frac{1}{2}}\,|t|^{\frac{3}{2}}\,\mu(dt) \le I_1^{\frac{1}{2}} I_2^{\frac{1}{2}} \le \sqrt{2}\,\omega^{\frac{1}{4}} I_2^{\frac{1}{4}} I_3^{\frac{1}{2}}.$$

Thus it follows that

$$\omega \ge \frac{I_2^3}{16 I_3^2}. \tag{A.7}$$

Also

$$I_2 = \mathcal{E}\left(\gamma^2\right) = \mathcal{E}\left(\sum_{i<j,\,k<\ell} \xi_i\xi_j\xi_k\xi_\ell A_{ij}A_{k\ell}\right) = \sum_{i<j} A_{ij}^2$$

and

$$I_3 \le \mathcal{E}\left(\left|\sum_{i<j}\xi_i\xi_j A_{ij}\right|^3\right) \le \mathcal{E}\left(\left(\sum_{i<j}\xi_i\xi_j A_{ij}\right)^2\left|\sum_{i<j}\xi_i\xi_j A_{ij}\right|\right)$$
$$\le \left(\sum_{i<j} A_{ij}^2\right)\sum_{i<j}|A_{ij}| \le \left(\sum_{i<j} A_{ij}^2\right)\sqrt{\frac{n(n-1)}{2}}\sqrt{\sum_{i<j} A_{ij}^2}.$$

The last inequality uses that $\sum_{i=1}^k |\alpha_i| \le \sqrt{k}\sqrt{\sum_{i=1}^k \alpha_i^2}$. Putting the above estimates for I_2 and I_3 into (A.7), we get

$$\omega \ge \frac{1}{16}\,\frac{2}{n(n-1)} > \frac{1}{8n^2},$$

and hence the lemma is proved. □

CONJECTURE A.5. *Let B and ξ be as defined in Lemma* A.4. *Then*

$$\Pr\left(\xi^T B\xi \le \operatorname{Tr} B\right) \ge \frac{1}{4}.$$

LEMMA A.6 (approximate S-lemma). *Let $R, R_0, R_1, \ldots, R_K$ be symmetric $n\times n$ matrices such that*

$$R_1, \ldots, R_K \succeq 0, \tag{A.8}$$

and assume that

$$\exists \lambda_0, \lambda_1, \ldots, \lambda_K \ge 0 \quad s.t. \quad \sum_{k=0}^{K} \lambda_k R_k \succ 0. \tag{A.9}$$

Consider the following quadratically constrained quadratic program,

$$QCQ = \max_{y\in\mathbb{R}^n} \left\{ y^T R y \ : \ y^T R_0 y \le r_0,\ y^T R_k y \le 1,\ k = 1,\dots,K \right\} \tag{A.10}$$

and the semidefinite optimization problem

$$SDP = \min_{\mu_0,\mu_1,\dots,\mu_K} \left\{ r_0\mu_0 + \sum_{k=1}^K \mu_k \ : \ \sum_{k=0}^K \mu_k R_k \succeq R,\ \mu \ge 0 \right\}. \tag{A.11}$$

Then

(i) *If problem* (A.10) *is feasible, then problem* (A.11) *is bounded below and*

$$SDP \ge QCQ. \tag{A.12}$$

Moreover, there exist $y_* \in \mathbb{R}^n$ *such that*

$$y_*{}^T R y_* = SDP, \tag{A.13}$$
$$y_*{}^T R_0 y_* \le r_0, \tag{A.14}$$
$$y_*{}^T R_k y_* \le \tilde{\rho}^2, \quad k = 1,\dots,K, \tag{A.15}$$

where (cf. (2.5)*)*

$$\tilde{\rho} := \left(2\log\left(6\sum_{k=1}^K \operatorname{rank} R_k \right) \right)^{\frac{1}{2}}$$

if R_0 *is a dyadic matrix, and*

$$\tilde{\rho} = \left(2\log\left(16n^2\sum_{k=1}^K \operatorname{rank} R_k \right) \right)^{\frac{1}{2}} \tag{A.16}$$

otherwise.

(ii) *If*

$$r_0 > 0, \tag{A.17}$$

then (A.10) *is feasible, problem* (A.11) *is solvable, and*

$$0 \le QCQ \le SDP \le \tilde{\rho}^2 QCQ. \tag{A.18}$$

Remark A.7. We claim that the usual S-lemma (cf. Lemma 2.2) can be obtained as a corollary of Lemma A.6. The "if" part of the S-lemma being evident, we focus below on the "only if" part.

1^0. Observe that it suffices to prove the following statement:

(!) *Assume that the set* $\{z : z^T P z > 0\}$ *is nonempty and that*

$$z \ne 0, \quad z^T P z \ge 0 \Rightarrow z^T Q z > 0. \tag{A.19}$$

Then

$$\exists \lambda \ge 0 : \quad Q \succeq \lambda P. \tag{A.20}$$

Indeed, let P, Q be such that

$$\{z : z^T P z > 0\} \neq \emptyset \quad \text{and} \quad z^T P z \geq 0 \Rightarrow z^T Q z \geq 0.$$

Then the pair $(P, Q + \epsilon I)$ for $\epsilon > 0$ clearly satisfies the premise in (!). Believing in (!), we therefore conclude that for every $\epsilon > 0$ there exists $\lambda(\epsilon) \geq 0$ such that $Q + \epsilon I \succeq \lambda(\epsilon) P$. As $\epsilon \to +0$, $\lambda(\epsilon)$ remains bounded due to $\lambda(\epsilon)\bar{z}^T P \bar{z} \leq \bar{z}^T (Q + \epsilon I)\bar{z}$, where $\bar{z}$ is such that $\bar{z}^T P \bar{z} > 0$. Since $\lambda(\epsilon) \geq 0$ remains bounded as $\epsilon \to +0$, there exists an accumulation point $\lambda \geq 0$ of $\lambda(\epsilon)$ as $\epsilon \to +0$; since $Q + \epsilon I \succeq \lambda(\epsilon) P$, one clearly has $Q \succeq \lambda P$, as required.

2^0. To prove (!), assume that the premise in (!) holds true, and observe that then the optimal value $QCQ(\epsilon)$ in the optimization problem

$$\max_x \left\{-x^T Q x : -x^T P x \leq 1,\ \epsilon x^T x \leq 1\right\} \tag{A.21}$$

remains bounded as $\epsilon \to +0$. Indeed, otherwise there clearly would exist a sequence of vectors x_i, $\|x_i\| \to \infty$ as $i \to \infty$, such that $x_i^T P x_i \geq -1$ and $x_i^T Q x_i \to -\infty$ as $i \to \infty$. By evident reasons, this would imply the existence of a unit vector $\bar{x}$ such that $\bar{x}^T P \bar{x} \geq 0$ and $\bar{x}^T Q \bar{x} \leq 0$, which would contradict (A.19). Now, the data

$$R = -Q, \quad R_0 = -P, \quad R_1 = \epsilon I, \quad r_0 = 1, \quad K = 1$$

clearly satisfy the premises (A.8) and (A.9) of Lemma A.6, and with these data, (A.10) coincides with (A.21). Since $r_0 = 1 > 0$, part (ii) of Lemma A.6 applies. Thus problem (A.11) is solvable and (A.18) holds. Hence, for every $\epsilon > 0$ there exist $\mu_0(\epsilon) \geq 0$ and $\mu_1(\epsilon) \geq 0$ such that

$$-\mu_0(\epsilon) P + \mu_1(\epsilon)\epsilon I \succeq -Q, \qquad \mu_0(\epsilon) + \mu_1(\epsilon) \leq \bar{\rho}^2 QCQ(\epsilon).$$

Since $QCQ(\epsilon)$ remains bounded as $\epsilon \to 0$, so are $\mu_0(\epsilon)$, $\mu_1(\epsilon)$; therefore there exists an accumulation point $(\mu_1 \geq 0, \mu_2 \geq 0)$ of $(\mu_0(\epsilon), \mu_1(\epsilon))$ as $\epsilon \to +0$, and $\lambda = \mu_1$ clearly satisfies the conclusion in (A.20). □

Proof. Notice that problem (A.11) is the semidefinite dual of

$$RQCQ = \max_{X \succeq 0} \{\operatorname{Tr} RX \ :\ \operatorname{Tr} R_0 X \leq r_o,\ \operatorname{Tr} R_k X \leq 1,\ k = 1, \ldots, K\}. \tag{A.22}$$

The latter problem is the standard semidefinite relaxation of the quadratically constrained quadratic problem (A.10), so we have

$$RQCQ \geq QCQ. \tag{A.23}$$

In part (i) of the lemma, (A.10) is assumed to be feasible; hence (A.22) is feasible as well, and hence, by weak duality, between problem (A.11) and its dual (A.22), problem (A.11) is bounded below. Now assumption (A.9) ensures that (A.11) is strictly feasible; thus from semidefinite duality theory, problem (A.22) is solvable and

$$SDP = RQCQ. \tag{A.24}$$

By (A.23) and (A.24), $SDP \geq QCQ$, which completes the proof of the first part of claim (i) in the lemma. To prove the second part, we first simplify the system (A.13)–(A.15). Letting

X_* denote an optimal solution of problem (A.22),

we introduce

$$\hat{R} = X_*^{\frac{1}{2}} R X_*^{\frac{1}{2}}. \tag{A.25}$$

Let

$$\hat{R} = U\tilde{R}U^T \qquad (U^TU = I,\ \tilde{R} = \operatorname{diag}(r_1, \ldots, r_n)) \tag{A.26}$$

be the eigenvalue decomposition of $\hat{R}$. Choosing

$$y_* = X_*^{\frac{1}{2}} U u, \quad u \in \mathbb{R}^n, \tag{A.27}$$

we have

$$y_*^T R y_* = u^T U^T X_*^{\frac{1}{2}} R X_*^{\frac{1}{2}} U u = u^T U^T \hat{R} U u = u^T \tilde{R} u = \sum_{i=1}^n r_i u_i^2.$$

Also

$$SDP = RQCQ = \operatorname{Tr} R X_* = \operatorname{Tr} \hat{R} = \operatorname{Tr} \tilde{R} = \sum_{i=1}^n r_i,$$

and thus (A.13) is equivalent to

$$\text{(a)} \qquad \sum_{i=1}^n r_i u_i^2 = \sum_{i=1}^n r_i.$$

Now, defining

$$\hat{R}_k = X_*^{\frac{1}{2}} R_k X_*^{\frac{1}{2}}, \quad \tilde{R}_k = U^T \hat{R}_k U, \qquad k = 0, 1, \ldots, K,$$

and using (A.27), we obtain

$$y_*^T R_k y_* = u^T U^T X_*^{\frac{1}{2}} R_k X_*^{\frac{1}{2}} U u = u^T U^T \hat{R}_k U u = u^T \tilde{R}_k u. \tag{A.28}$$

Since X_* solves RQCQ,

$$r_0 \geq \operatorname{Tr} R_0 X_* = \operatorname{Tr} \hat{R}_0 = \operatorname{Tr} \tilde{R}_0 \tag{A.29}$$

and

$$1 \geq \operatorname{Tr} R_k X_* = \operatorname{Tr} \hat{R}_k = \operatorname{Tr} \tilde{R}_k, \quad k = 1, \ldots, K. \tag{A.30}$$

From (A.28) and (A.29) we see that (A.14) holds if

$$\text{(b)} \qquad u^T \tilde{R}_0 u \leq \operatorname{Tr} \tilde{R}_0,$$

and from (A.28) and (A.30), relation (A.15) holds if

$$\text{(c)} \qquad u^T \tilde{R}_k u \leq \tilde{\rho}^2 \operatorname{Tr} \tilde{R}_k, \quad k = 1, \ldots, K.$$

We conclude that if there exists a $\bar{u}$ satisfying

$$\sum_{i=1}^{n} r_i \bar{u}_i^2 = \sum_{i=1}^{n} r_i, \tag{A.31}$$

$$\bar{u}^T \tilde{R}_0 \bar{u} \le \operatorname{Tr} \tilde{R}_0, \tag{A.32}$$

$$\bar{u}^T \tilde{R}_k \bar{u} \le \tilde{\rho}^2 \operatorname{Tr} \tilde{R}_k, \quad k = 1, \ldots, K, \tag{A.33}$$

then $y_* = X_*^{\frac{1}{2}} U \bar{u}$ satisfies (A.13)–(A.15). Note that (A.31) is automatically satisfied if $\bar{u}$ is a ± 1-vector. Thus it suffices to show that (A.32) and (A.33) can be satisfied by a ± 1-vector $\bar{u}$.

Let us pretend for a moment that the vector $\bar{u}$ is a *random* ± 1-vector such that $\Pr(\bar{u}_i = 1) = \Pr(\bar{u}_i = -1) = \frac{1}{2}$ for each i. Let B denote the event that $\bar{u}$ satisfies (A.32), and C_k the event that $\bar{u}^T \tilde{R}_k \bar{u} \le \tilde{\rho}^2 \operatorname{Tr} \tilde{R}_k$, and $C = \cap_k C_k$, i.e., C denotes the event that $\bar{u}$ satisfies (A.33). Then we only need to show that

$$\Pr(B \cap C) > 0. \tag{A.34}$$

Since

$$B \subseteq (B \cup C) \cap C^c,$$

where ^{c} refers to the complement of the event, we may write

$$\begin{aligned} \Pr(B \cup C) &\ge \Pr(B) - \Pr(C^c) = \Pr(B) - \Pr\left((\cap_k C_k)^c\right) \\ &= \Pr(B) - \Pr(\cup_k C_k^c) \ge \Pr(B) - \sum_{k=1}^{K} \Pr(C_k^c). \end{aligned}$$

Hence, (A.34) will certainly hold if for some $p_0 > 0$,

$$\Pr(B) > p_0, \tag{A.35}$$

$$\sum_{k=1}^{K} \Pr(C_k^c) \le p_0. \tag{A.36}$$

We first consider the case in which R_0 is dyadic. Then $\hat{R}_0$ and $\tilde{R}_0$ are also dyadic, and hence we may write, for a suitable vector b,

$$\tilde{R}_0 = b b^T.$$

Then condition (A.32) is equivalent to

$$\bar{u}^T \frac{b}{\|b\|} \le 1.$$

Hence, in the dyadic case, (A.35) is equivalent to

$$\Pr\left(\left|u^T x\right| \le 1\right) > p_0, \tag{A.37}$$

where x is the unit vector $b/\|b\|$. By Lemma A.1 this inequality certainly holds if $p_0 = \frac{1}{3}$. On the other hand, by Lemma A.3, for each k,

$$\Pr(C_k^c) = \Pr\left(\bar{u}^T \tilde{R}_k \bar{u} > \tilde{\rho}^2 \operatorname{Tr} \tilde{R}_k\right) \le 2 (\operatorname{rank} \tilde{R}_k) e^{-\frac{\tilde{\rho}^2}{2}}.$$

Since $\operatorname{rank} \tilde{R}_k \le \operatorname{rank} R_k$, we obtain

$$\sum_{k=1}^{K} \Pr\left(C_k^c\right) \le 2\, e^{-\frac{\tilde{\rho}^2}{2}} \sum_{k=1}^{K} \operatorname{rank} R_k,$$

and so inequalities (A.35) and (A.36) will hold if $p_0 = \frac{1}{3}$ and $\tilde{\rho}$ is such that

$$2\, e^{-\frac{\tilde{\rho}^2}{2}} \sum_{k=1}^{K} \operatorname{rank} R_k = \frac{1}{3} = p_0. \tag{A.38}$$

One may easily verify that the value of $\tilde{\rho}$ as given by (2.5) is indeed the solution of (A.38). Thus the proof is complete for the case in which R_0 is dyadic.

We finally consider the general case, where R_0 is an arbitrary symmetric matrix. Then we apply Lemma A.4, which gives that (A.35) holds for $p_0 = 1/(8n^2)$. Then solving $\tilde{\rho}$ from (A.38) with this value of p_0, we get the value given in (A.16).

To complete the proof of the lemma we need only to prove $SDP \le \tilde{\rho}^2 QCQ$, the last inequality in (A.18). For this, let y_* satisfy (A.13)–(A.15). Then, since $\tilde{\rho} > 1$, the vector

$$\bar{y} = \frac{y_*}{\tilde{\rho}}$$

is feasible for problem (A.10). Therefore, using (A.13),

$$QCQ \ge \bar{y}^T R \bar{y} = \frac{1}{\tilde{\rho}^2} y_*^T R y_* = \frac{SDP}{\tilde{\rho}^2},$$

and hence the proof is complete. □

LEMMA A.8. *Let $a, b \in \mathbb{R}^n$ be two nonzero vectors and X a symmetric $n \times n$ matrix. Then*

$$X \succeq \pm\left(ab^T + ba^T\right) \tag{A.39}$$

holds if and only if

$$\exists \rho > 0 \quad s.t. \quad X \succeq \rho a a^T + \frac{1}{\rho} b b^T. \tag{A.40}$$

Proof. Suppose (A.40) holds. Then, for arbitrary $y \in \mathbb{R}^n$ one has

$$\begin{aligned} y^T X y &\ge y^T \left(\rho a a^T + \frac{1}{\rho} b b^T\right) y = \rho \left(a^T y\right)^2 + \frac{1}{\rho} \left(b^T y\right)^2 \\ &\ge 2 \left|a^T y\right| \left|b^T y\right| \ge \left|y^T \left(ab^T + ba^T\right) y\right|; \end{aligned}$$

hence (A.39) follows. On the other hand, if (A.40) does not hold, then the system

$$X \succeq \rho a a^T + \mu b b^T, \quad \begin{pmatrix} \rho & 1 \\ 1 & \mu \end{pmatrix} \succeq 0 \tag{A.41}$$

does not have a solution (ρ, μ). This implies that the optimal value p^* of the semidefinite optimization problem

$$p^* = \min_{t,\rho,\mu} \left\{ t : tI + X \succeq \rho a a^T + \mu b b^T, \quad \begin{pmatrix} \rho & 1 \\ 1 & \mu \end{pmatrix} \succeq 0 \right\} \tag{SDP}$$

is positive. Clearly (SDP) is strictly feasible and bounded below. Hence its dual problem (SDD),

(SDD)

$$\max_{\substack{v,\,v_1,\,v_2,\\ U\,\succeq\,0}} \left\{ -\mathrm{Tr}\,(UX) - 2v : \mathrm{Tr}\,(U) = 1, \begin{pmatrix} v_1 & v \\ v & v_2 \end{pmatrix} \succeq 0,\ v_1 = b^T U b,\ v_2 = a^T U a \right\},$$

is solvable and has the same optimal value $p^* > 0$. A feasible solution (U, v, v_1, v_2) to (SDD) satisfies

$$|v| \le \sqrt{v_1 v_2} = \sqrt{(a^T U a)\,(b^T U b)}.$$

Hence, $p^* > 0$ implies the existence of $U \succeq 0$ such that

$$2\sqrt{(a^T U a)\,(b^T U b)} > \mathrm{Tr}\,(UX). \tag{A.42}$$

Let

$$\bar{a} = U^{\frac{1}{2}} a, \quad \bar{b} = U^{\frac{1}{2}} b, \quad \bar{X} = U^{\frac{1}{2}} X U^{\frac{1}{2}}.$$

Then (A.42) can be rewritten as

$$\mathrm{Tr}\,\bar{X} < 2\sqrt{(\bar{a}^T \bar{a})\,(\bar{b}^T \bar{b})} = 2\,\|\bar{a}\|\,\|\bar{b}\|. \tag{A.43}$$

Now suppose that X satisfies (A.39). Then it follows that

$$\bar{X} \succeq \pm\left(\bar{a}\bar{b}^T + \bar{b}\bar{a}^T\right).$$

Define $Q = \bar{a}\bar{b}^T + \bar{b}\bar{a}^T$, and let $\lambda(Q)$ be the vector of eigenvalues of Q. It then follows that

$$\mathrm{Tr}\,\bar{X} \ge \|\lambda(Q)\|_1 = \left\|\lambda(\bar{a}\bar{b}^T + \bar{b}\bar{a}^T)\right\|_1 = 2\,\|\bar{a}\|\,\|\bar{b}\|.$$

This contradicts (A.43). Hence the proof is complete. □

REFERENCES

[1] A. Ben-Tal and A. Nemirovski, *Robust truss topology design via semidefinite programming,* SIAM J. Optim., 7 (1997), pp. 991–1016.

[2] A. Ben-Tal and A. Nemirovski, *Robust convex optimization,* Math. Oper. Res., 23 (1998), pp. 769–805.

[3] A. Ben-Tal and A. Nemirovski, *Robust solutions of uncertain linear programs,* Oper. Res. Lett., 25 (1999), pp. 1–13.

[4] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization. Analysis, Algorithms, and Engineering Applications*, MPS/SIAM Ser. Optim. MP02, SIAM, Philadelphia, 2001.

[5] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory,* SIAM Stud. Appl. Math. 15, SIAM, Philadelphia, 1994.

[6] 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.

[7] J. Håstad, *Some optimal in-approximability results,* in Proceedings of the 29th ACM Symposium on Theory of Computing, El Paso, TX, 1997, SIGACT, ACM, New York, 1997, pp. 1–10.

SIAM J. OPTIM.
Vol. 13, No. 2, pp. 561–587

SHAPE OPTIMIZATION IN CONTACT PROBLEMS WITH COULOMB FRICTION*

P. BEREMLIJSKI†, J. HASLINGER‡, M. KOČVARA§, AND J. OUTRATA¶

Dedicated to Jochem Zowe on the occasion of his sixtieth birthday

Abstract. The paper deals with a discretized problem of the shape optimization of elastic bodies in unilateral contact. The aim is to extend existing results to the case of contact problems following the Coulomb friction law. Mathematical modelling of the Coulomb friction problem leads to a quasi-variational inequality. It is shown that for small coefficients of friction the discretized problem with Coulomb friction has a unique solution and that this solution is Lipschitzian as a function of a control variable describing the shape of the elastic body.

The shape optimization problem belongs to a class of so-called mathematical programs with equilibrium constraints (MPECs). The uniqueness of the equilibria for fixed controls enables us to apply the so-called implicit programming approach. Its main idea consists of minimizing a nonsmooth composite function generated by the objective and the (single-valued) control-state mapping. In this paper, the control-state mapping is much more complicated than in most MPECs solved in the literature so far, and the generalization of the relevant results is by no means straightforward. Numerical examples illustrate the efficiency and reliability of the suggested approach.

Key words. shape optimization, contact problems, Coulomb friction, mathematical programs with equilibrium constraints

AMS subject classifications. 49Q10, 74M10, 74S05

PII. S1052623401395061

Introduction. Shape optimization is a branch of optimal control theory in which control variables are related to the geometry of considered structures. From our daily experience we know that the geometry of a structure is one of the decisive factors determining its properties. The goal of shape optimization is to find "the best possible" geometry of a structure in order to enhance some desired properties. Special attention is paid to shape optimization of structures governed by *variational inequalities*. It is well known that optimal control problems with state relations represented by variational inequalities are generally *nonsmooth*, in view of a possible nondifferentiability of the respective *control-state* mapping. This fact not only makes the analysis more difficult, but also complicates the numerical computation. In particular, the possible nondifferentiability of minimized functions restricts the choice of available minimization methods.

The present paper deals with a particular problem of so-called *contact shape optimization*, i.e., optimization of the geometry of a system of deformable bodies that are

*Received by the editors September 12, 2001; accepted for publication (in revised form) February 27, 2002; published electronically October 1, 2002. This work was supported by grant A1075707 of the Czech Academy of Sciences (JH,MK,JO), grant 101/01/0538 of the Grant Agency of the Czech Republic (PB,JH), and BMBF project 03ZOM3ER (MK).

http://www.siam.org/journals/siopt/13-2/39506.html

†Faculty of Electrical Engineering, Technical University of Ostrava, 17. listopadu 15, 708 33 Ostrava-Poruba, Czech Republic (petr.beremlijski@vsb.cz).

‡Department of Metal Physics, Charles University, Ke Karlovu 3, 121 16 Praha 2, Czech Republic (haslin@met.mff.cuni.cz).

§Institute of Applied Mathematics, University of Erlangen, Martensstrasse 3, 91058 Erlangen, Germany (kocvara@am.uni-erlangen.de). The contributions of this author were made while on leave from the Czech Academy of Sciences.

¶Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Pod vodárenskou věží 4, 182 08 Praha 8, Czech Republic (outrata@utia.cas.cz).

in mutual contact and subject to external forces. The mathematical model describing the equilibrium state of such a system is represented by *differential inclusions*, whose complexity depends on physical phenomena involved in the model, such as the influence of friction on contact parts of the boundaries. In [9], shape optimization of elastic bodies in unilateral contact was analyzed with and without given friction. The aim of the current paper is to extend those results to the case of friction that obeys the classic Coulomb law. We confine ourselves to the simplest case, namely, the static form of the Coulomb law of friction, which may not always be relevant from the mechanical point of view. To get a more realistic model, one has to use the quasi-static formulation, e.g., one involving the rate of change of the displacement field u (for details, see [20]). The static case is, however, important from the computational point of view since appropriate discretizations of rate dependent models lead to a sequence of static ones. In contrast to the frictionless case or to a model with given friction, both of which are described by classical variational inequalities, the mathematical model of Coulomb friction leads to *a quasi-variational inequality* (see [7]), making the mathematical analysis and the numerical realization substantially more involved. For mechanical aspects of contact shape optimization, see [11].

The subject of this paper is the shape optimization of *discretized contact problems* with Coulomb friction, provided that the coefficient of friction is sufficiently small. To simplify our presentation, we focus on the so-called *Signorini problem*, i.e., on the contact problem for one elastic body unilaterally supported by a rigid foundation. The discretization of the state problem is based on a mixed finite element formulation of the Signorini problem with given friction, i.e., on the formulation in terms of displacements and normal contact stresses which are equal to Lagrange multipliers associated with the unilateral constraints. This formulation is used to define a mapping Φ associating the respective Lagrange multipliers in the aforementioned mixed finite element formulation with a given slip bound. Solutions to contact problems with Coulomb friction are finally defined as fixed points of Φ in an appropriate set. It is well known that for the coefficients of Coulomb friction that are small enough, the mapping Φ has a unique fixed point or, equivalently, the discretized contact problem with Coulomb friction has a unique solution (see [5]).

The shape optimization problem belongs to a class of so-called *mathematical programs with equilibrium constraints* (MPECs), which have been intensively studied especially in the past fifteen years; see, e.g., [13]. The uniqueness of the equilibria for fixed controls enables us to apply an effective method belonging to the useful and reliable *implicit programming approach*. This method is described in detail in [18], but in connection with substantially simpler equilibria. Its main idea consists of analyzing a composite function generated by the discretized objective and the discretized (single-valued) control-state mapping. Subsequently, this composite function is minimized by a suitable nonsmooth minimization algorithm, e.g., by a bundle method. In this paper, however, the control-state mapping is much more complicated than in most MPECs solved in [18] and other works, and the generalization of the relevant results is by no means straightforward.

The paper is organized as follows: Section 1 collects known results related to finite element approximations, which are needed in section 2. Additionally, we provide the reader with some basic notions from nonsmooth analysis, which are extensively used in sections 3 and 4.

In section 2 we first present a mixed finite element formulation of the Signorini problem with given friction, which is used to define the Coulomb friction model. We prove that the discretized contact problem with Coulomb friction is solvable for *any*

positive value of the coefficient $\mathcal{F}$ of Coulomb friction and is uniquely solvable for $\mathcal{F}$ sufficiently small. We also prove that the bounds on $\mathcal{F}$ ensuring the uniqueness of the solution are uniform with respect to the discretized control (design) variables. The rest of the paper is restricted solely to this case. The main result of section 2 states that the control-state mapping is Lipschitzian on a set of admissible discretized control variables.

Section 3 concerns the sensitivity analysis of this mapping. We show that it is piecewise C^1 and compute an upper approximation of its Clarke generalized Jacobian. Moreover, we apply a suitable chain rule and clarify how to obtain "subgradient information" needed in the chosen numerical solver.

Section 4 presents numerical tests. They illustrate the efficiency and reliability of the suggested approach as well as the relevance of the obtained results from the mechanical point of view.

The following notation is employed: x^i is the ith component of a vector $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{E}$ is the unit matrix, and $\mathbb{R}^n_+$ is the nonnegative orthant of $\mathbb{R}^n$. For an $[m \times n]$ matrix $\boldsymbol{A}$ and index sets $I \subset \{1, 2, \ldots, m\}$, $J \subset \{1, 2, \ldots, n\}$, $\boldsymbol{A}_{I,J}$ denotes the submatrix of $\boldsymbol{A}$ with rows and columns specified by I and J, respectively. Further, $\boldsymbol{A}_I$ and ${}_I\boldsymbol{A}$ are submatrices of $\boldsymbol{A}$ with rows and columns, respectively, specified by I. Similarly, for a vector $\in \mathbb{R}^n$, ${}_I$ is the subvector composed from the components δ^i, $i \in I$. For a function f of two variables x, y, $\partial_x f(x, y)$ denotes its partial subdifferential with respect to x. If f is differentiable, $\nabla_x f(x, y)$ denotes its partial gradient. In certain cases, we will prefer the notation $\nabla_i f(x, y)$, $i = 1, 2$, for the gradient of f with respect to the ith variable. For a finite set $\mathcal{I}$, $|\mathcal{I}|$ denotes its cardinality. Let $P_1(T)$ be a space of polynomials on a set $T \subset \mathbb{R}^n$ of degree ≤ 1.

Furthermore, in agreement with the standard literature on linear elasticity [16], $u = (u_1, u_2)$ denotes the displacement field, $\varepsilon(u) = (\varepsilon_{ij}(u))_{i,j=1,2}$ stands for the linearized strain tensor with components $\varepsilon_{ij}(u(x)) = \frac{1}{2}(\frac{\partial u(x)_i}{\partial x_j} + \frac{\partial u(x)_j}{\partial x_i})$, and $\tau(u) = (\tau_{ij}(u))_{i,j=1,2}$ is the stress tensor related to $\varepsilon(u)$ by means of the linear Hooke's law.

1. Preliminaries. The first part of this preliminary section collects results from the finite element theory used in section 2 to the discretization of our problem.

Let $\widehat{\Omega} = (0, a) \times (0, b)$, $a, b \in \mathbb{R}_+$, be a rectangle whose boundary $\partial\widehat{\Omega}$ is split into three nonempty nonoverlapping parts $\widehat{\Gamma}$, $\widehat{\Gamma}_u$, and $\widehat{\Gamma}_P$, where $\widehat{\Gamma} = (0, a) \times \{0\}$. Let $\widehat{\mathcal{T}}_h$ be a uniform triangulation of $\overline{\widehat{\Omega}}$, whose nodes $\{\widehat{A}_{ij}\}$ form a rectangular grid, $\widehat{A}_{ij} = (\hat{x}_i, \hat{y}_j)$, $\hat{x}_i = \frac{a}{N} i$, $\hat{y}_j = \frac{b}{M} j$, $i = 0, \ldots, N$, $j = 0, \ldots, M$, where N, M are positive integers. The symbol h stands for the norm of $\widehat{\mathcal{T}}_h$.

We shall define the following finite element spaces in $\widehat{\Omega}$:

$$\widehat{X}_h = \{\hat{v}_h \in C(\overline{\widehat{\Omega}}) \mid \hat{v}_h|_{\widehat{T}} \in P_1(\widehat{T}) \quad \forall\, \widehat{T} \in \widehat{\mathcal{T}}_h\},$$
$$\widehat{V}_h = \{\hat{v}_h \in \widehat{X}_h \mid \hat{v}_h = 0 \text{ on } \widehat{\Gamma}_u\},$$
$$\widehat{X}_{h0} = \{\hat{v}_h \in \widehat{X}_h \mid \hat{v}_h = 0 \text{ on } \widehat{\Gamma}_u \cup \widehat{\Gamma}\}.$$

Let $\widehat{\Delta}_h$ be a partitioning of $[0, a]$ realized by the nodes of $\widehat{\mathcal{T}}_h$ lying on $\overline{\widehat{\Gamma}}$. By U^h_{ad} we denote the set of discretized design variables defined as follows:

$$U^h_{ad} = \{\alpha_h \in C([0, a]) \mid \alpha_h \text{ is piecewise linear on } \widehat{\Delta}_h,\ 0 \leq \alpha_h \leq C_0 \text{ in } [0, a]\}$$

for some $C_0 \in (0, b/2)$.

With any $\alpha_h \in U_{ad}^h$ we will associate the following polygonal domain $\Omega(\alpha_h)$:

$$\Omega(\alpha_h) = \{(x_1, x_2) \in \mathbb{R}^2 \mid \alpha_h(x_1) < x_2 < b;\, x_1 \in (0, a)\}.$$

The system of all $\Omega(\alpha_h)$, $\alpha_h \in U_{ad}^h$, will be denoted by $\mathcal{O}_h$. Any $\Omega(\alpha_h) \in \mathcal{O}_h$ will be considered as an image $F_{\alpha_h}(\widehat{\Omega})$, where $F_{\alpha_h} : \mathbb{R}^2 \to \mathbb{R}^2$ is a mapping of the form $F_{\alpha_h} = A_{ij}\hat{\varphi}_{ij}$, $A_{ij} = (\hat{x}_i, \alpha_h(\hat{x}_i) + j(b - \alpha_h(\hat{x}_i))/M) \in \mathbb{R}^2$, $i = 0, \dots, N$, $j = 0, \dots, M$, and $\hat{\varphi}_{ij}$ is the Courant basis function of $\widehat{X}_h$ associated with the node $\widehat{A}_{ij}$ of $\widehat{\mathcal{T}}_h$. By means of F_{α_h} we define the partitioning of $\partial\Omega(\alpha_h)$ into $\Gamma(\alpha_h)$, $\Gamma_u(\alpha_h)$, and $\Gamma_P(\alpha_h)$ as follows: $\Gamma(\alpha_h) = F_{\alpha_h}(\widehat{\Gamma})$, $\Gamma_u(\alpha_h) = F_{\alpha_h}(\widehat{\Gamma}_u)$, $\Gamma_P(\alpha_h) = F_{\alpha_h}(\widehat{\Gamma}_P)$. The triangulation $\mathcal{T}_h(\alpha_h)$ of $\overline{\Omega}(\alpha_h)$ is a "deformation" of $\widehat{\mathcal{T}}_h$ by means of F_{α_h}.

The triangulations $\widehat{\mathcal{T}}_h$ and $\mathcal{T}_h(\alpha_h)$ are *topologically equivalent*; i.e., for any $\alpha_h \in U_{ad}^h$ the triangulation $\mathcal{T}_h(\alpha_h)$ has the same number of the nodes as $\widehat{\mathcal{T}}_h$, and the nodes of $\mathcal{T}_h(\alpha_h)$ keep the same neighbors as $\widehat{\mathcal{T}}_h$. In addition, the position of all A_{ij} depends *continuously* on variations of $\alpha_h \in U_{ad}^h$.

For any $\Omega(\alpha_h) \in \mathcal{O}_h$ we define the following spaces:

$$\begin{aligned}
X_h(\alpha_h) &= \{v_h \in C(\overline{\Omega}(\alpha_h)) \mid v_h|_T \in P_1(T) \quad \forall\, T \in \mathcal{T}_h(\alpha_h)\},\\
V_h(\alpha_h) &= \{v_h \in X_h(\alpha_h) \mid v_h = 0 \text{ on } \Gamma_u(\alpha_h)\},\\
X_{h0}(\alpha_h) &= \{v_h \in X_h(\alpha_h) \mid v_h = 0 \text{ on } \Gamma_u(\alpha_h) \cup \Gamma(\alpha_h)\},\\
\boldsymbol{V}_h(\alpha_h) &= V_h(\alpha_h) \times V_h(\alpha_h).
\end{aligned}$$

It is readily seen that

$$v_h \in \{X_h(\alpha_h), V_h(\alpha_h), X_{h0}(\alpha_h)\} \quad \text{iff} \quad \hat{v}_h = v_h \circ F_{\alpha_h} \in \{\widehat{X}_h, \widehat{V}_h, \widehat{X}_{h0}\}. \tag{1.1}$$

By means of (1.1), the one-to-one correspondence $X_h(\alpha_h) \longleftrightarrow \widehat{X}_h$, $V_h(\alpha_h) \longleftrightarrow \widehat{V}_h$, $X_{h0}(\alpha_h) \longleftrightarrow \widehat{X}_{h0}$ is established.

Convention. *In what follows, the symbol "ˆ" above a function v_h defined in $\Omega(\alpha_h)$ denotes its "transport" on $\widehat{\Omega}$ by means of* (1.1).

We now introduce trace spaces on $\widehat{\Gamma}$ and $\Gamma(\alpha_h)$, $\alpha_h \in U_{ad}^h$. Denote by $\widehat{\mathcal{V}}_h$, $\mathcal{V}_h(\alpha_h)$ the spaces of restrictions of functions from $\widehat{V}_h$, $V_h(\alpha_h)$ to $\widehat{\Gamma}$, $\Gamma(\alpha_h)$, respectively:

$$\widehat{\mathcal{V}}_h = \widehat{V}_h|_{\widehat{\Gamma}}, \qquad \mathcal{V}_h(\alpha_h) = V_h(\alpha_h)|_{\Gamma(\alpha_h)}.$$

It is readily seen that

$$\varphi_h \in \mathcal{V}_h(\alpha_h) \quad \text{iff} \quad \hat{\varphi}_h = \varphi_h \circ F_{\alpha_h}|_{\widehat{\Gamma}} \in \widehat{\mathcal{V}}_h.$$

The trace spaces on $\widehat{\Gamma}$, $\Gamma(\alpha_h)$ will be equipped with the following norms:

$$\|\hat{\varphi}_h\|_{h,\widehat{\Gamma}} := \inf_{\substack{\hat{v}_h \in \widehat{V}_h \\ \hat{v}_h = \hat{\varphi}_h \text{ on } \widehat{\Gamma}}} |\hat{v}_h|_{1,\widehat{\Omega}}, \qquad \|\varphi_h\|_{h,\alpha_h} := \inf_{\substack{v_h \in V_h(\alpha_h) \\ v_h = \varphi_h \text{ on } \Gamma(\alpha_h)}} |v_h|_{1,\Omega(\alpha_h)}, \tag{1.2}$$

where $|\cdot|_{1,\widehat{\Omega}}$, $|\cdot|_{1,\Omega(\alpha_h)}$ stand for the $H^1(\widehat{\Omega})$, $H^1(\Omega(\alpha_h))$-seminorms, respectively. The following assertion follows directly from the above definitions.

Proposition 1.1. *Let $\hat{\varphi}_h \in \widehat{\mathcal{V}}_h$. Then*

$$\|\hat{\varphi}_h\|_{h,\widehat{\Gamma}} = |\hat{u}_h|_{1,\widehat{\Omega}}, \qquad \|\varphi_h\|_{h,\alpha_h} = |z_h|_{1,\Omega(\alpha_h)},$$

where $\hat{u}_h \in \widehat{V}_h$, $z_h \in V_h(\alpha_h)$ *are solutions of the following discretized nonhomogeneous Dirichlet boundary value problems:*

$$\int_{\widehat{\Omega}} \nabla \hat{u}_h \cdot \nabla \hat{v}_h d\hat{x} = 0 \qquad \forall \hat{v}_h \in \widehat{X}_{h0},\ \hat{u}_h = \hat{\varphi}_h \ \text{on } \widehat{\Gamma},$$
$$\int_{\Omega(\alpha_h)} \nabla z_h \cdot \nabla v_h dx = 0 \quad \forall v_h \in X_{h0},\ z_h = \varphi_h \ \text{on } \Gamma(\alpha_h),$$

respectively.

It is easy to show that the norms $\|\cdot\|_{h,\widehat{\Gamma}}$, $\|\cdot\|_{h,\alpha_h}$ are uniformly equivalent with respect to $\alpha_h \in U_{ad}^h$ as follows from the following result.

PROPOSITION 1.2. *There exist constants* $c_1, c_2 > 0$ *such that*

$$c_1 \|\varphi_h\|_{h,\alpha_h} \le \|\hat{\varphi}_h\|_{h,\widehat{\Gamma}} \le c_2 \|\varphi_h\|_{h,\alpha_h} \tag{1.3}$$

holds for any $\varphi_h \in \mathcal{V}_h(\alpha_h)$ *and any* $\alpha_h \in U_{ad}^h$.

Remark 1.1. Let us consider a regular system $\{\widehat{\mathcal{T}}_h\}$ of triangulations of $\overline{\widehat{\Omega}}$, $h \to 0_+$. Then the system $\{\mathcal{T}_h(\alpha_h), \alpha_h \in U_{ad}^h\}$, $h \to 0_+$, is uniformly regular with respect to $\alpha_h \in U_{ad}^h$ and $h \to 0_+$, and the constants c_1, c_2 in (1.3) *do not depend* on $h \to 0_+$.

In what follows, we shall suppose that $\{\widehat{\mathcal{T}}_h\}$, $h \to 0_+$ *is a regular system of triangulations of* $\overline{\widehat{\Omega}}$.

Next we introduce a dual space of $\widehat{\mathcal{V}}_h$. Let L_κ be a finite-dimensional space in duality with $\widehat{\mathcal{V}}_h$, and let $\langle \cdot, \cdot \rangle$ be the corresponding duality pairing. In what follows, we shall suppose that L_κ is chosen in such a way that the following condition is satisfied:

$$(S) \qquad \mu_\kappa \in L_\kappa :\ (\langle \mu_\kappa, \hat{\varphi}_h \rangle = 0 \quad \forall \hat{\varphi}_h \in \widehat{\mathcal{V}}_h) \Longrightarrow \mu_\kappa = 0.$$

If so, then

$$\|\mu_\kappa\|_{-h,\widehat{\Gamma}} := \sup_{\substack{\hat{\varphi}_h \in \widehat{\mathcal{V}}_h \\ \hat{\varphi}_h \neq 0}} \frac{\langle \mu_\kappa, \hat{\varphi}_h \rangle}{\|\hat{\varphi}_h\|_{h,\widehat{\Gamma}}} \tag{1.4}$$

defines a norm in L_κ. A more suitable expression for $\|\cdot\|_{-h,\widehat{\Gamma}}$ gives the next result (see [3]).

PROPOSITION 1.3. *It holds that*

$$\|\mu_\kappa\|_{-h,\widehat{\Gamma}} = |\widehat{w}_h|_{1,\widehat{\Omega}} = \sup_{\substack{\hat{v}_h \in \widehat{V}_h \\ \hat{v}_h \neq 0}} \frac{\langle \mu_\kappa, \hat{v}_h \rangle}{|\hat{v}_h|_{1,\widehat{\Omega}}},$$

where $\widehat{w}_h \in \widehat{V}_h$ *is the solution of the discretized Neumann problem in* $\widehat{\Omega}$*:*

$$\int_{\widehat{\Omega}} \nabla \widehat{w}_h \cdot \nabla \hat{v}_h d\hat{x} = \langle \mu_\kappa, \hat{v}_h \rangle \quad \forall \hat{v}_h \in \widehat{V}_h. \tag{1.5}$$

Let $\mu_\kappa \in L_\kappa$ and $\varphi_h \in \mathcal{V}_h(\alpha_h)$. We define

$$\langle \mu_\kappa, \varphi_h \rangle_{\alpha_h} := \langle \mu_\kappa, \hat{\varphi}_h \rangle$$

and

$$\|\mu_\kappa\|_{-h,\alpha_h} := \sup_{\substack{v_h \in V_h(\alpha_h) \\ v_h \neq 0}} \frac{\langle \mu_\kappa, v_h \rangle_{\alpha_h}}{|v_h|_{1,\Omega(\alpha_h)}}. \tag{1.6}$$

By virtue of (S), (1.6) defines *a shape dependent* norm in L_κ. However, it is easy to show that $\|\cdot\|_{-h,\widehat{\Gamma}}$ and $\|\cdot\|_{-h,\alpha_h}$ are equivalent uniformly with respect to $\alpha_h \in U_{ad}^h$. This follows from the next proposition.

PROPOSITION 1.4. *It holds that*

$$\frac{1}{c_2}\|\mu_\kappa\|_{-h,\alpha_h} \leq \|\mu_\kappa\|_{-h,\widehat{\Gamma}} \leq \frac{1}{c_1}\|\mu_\kappa\|_{-h,\alpha_h} \tag{1.7}$$

for any $\mu_\kappa \in L_\kappa$ and any $\alpha_h \in U_{ad}^h$, where c_1, c_2 are the same constants as in Proposition 1.2.

We close this part with an extension-type result. By $r_h\hat{\varphi}_h \in \widehat{V}_h$ we denote the extension of $\hat{\varphi}_h \in \widehat{\mathcal{V}}_h$ from $\widehat{\Gamma}$ into $\widehat{\Omega}$ having the smallest support:

$$r_h\hat{\varphi}_h = \hat{\varphi}_h \quad \text{on } \overline{\widehat{\Gamma}}, \qquad r_h\hat{\varphi}_h(\widehat{A}_{ij}) = 0 \quad \forall \widehat{A}_{ij} \notin \overline{\widehat{\Gamma}}.$$

PROPOSITION 1.5 (see [22]). *There exists a constant $c_0 > 0$ independent of h such that*

$$\|r_h\hat{\varphi}_h\|_{1,\widehat{\Omega}} \leq c_0 h^{-1/2}\|\hat{\varphi}_h\|_{0,\widehat{\Gamma}} \tag{1.8}$$

holds for any $\hat{\varphi}_h \in \widehat{\mathcal{V}}_h$.

In the rest of this section we list, for the reader's convenience, those basic concepts from nonsmooth analysis which are essential for the whole development in sections 3 and 4.

Let $F[\mathbb{R}^n \to \mathbb{R}^m]$ be Lipschitzian near a point $x \in \mathbb{R}^n$. The *generalized Jacobian* of F at x, denoted $\partial F(x)$, is a subset of $\mathbb{R}^{m\times n}$ given by

$$\partial F(x) = \operatorname{conv}\left\{ \lim_{i\to\infty} \nabla F(x_i) \mid x_i \to x,\ x_i \notin \Omega_F \right\},$$

where Ω_F is the set of points at which F is not differentiable, and $\nabla F(x_i)$ denotes the standard Jacobian of F at x_i. For $m = 1$ the generalized Jacobian amounts to a set of row vectors whose transpose is the Clarke subdifferential of F at x, denoted also by $\partial F(x)$. These objects enjoy a rich calculus, which is thoroughly investigated, e.g., in [1]. In section 3 we will employ a chain rule from this calculus.

In connection with MPECs, one frequently encounters the following mappings.

DEFINITION 1.6. *Let U be an open subset of $\mathbb{R}^n$. A function $F[U \to \mathbb{R}^m]$ is called a* PC^1-function *if it is continuous and if for every $x_0 \in U$ there exists an open neighborhood $\mathcal{O} \subset U$ and a finite number of continuously differentiable functions $F_i[\mathcal{O} \to \mathbb{R}^m]$, $i = 1, 2, \ldots, k$, such that for every $x \in \mathcal{O}$ one has $F(x) \in \{F_1(x), \ldots, F_k(x)\}$.*

The single functions F_i are called *selections* or *pieces* for F at x_0, and the set

$$I_F(x_0) := \{i \in \{1, 2 \ldots, k\} \mid F(x_0) = F_i(x_0)\}$$

is the *active index set* at x_0. The selections $F_i, i \in I_F(x_0)$, are called *active selections (pieces)* for F at x_0. A selection F_i is *essentially active* at x_0, provided

$$x_0 \in \operatorname{cl}\left(\operatorname{int}\{z \in \mathcal{O} \mid F(z) = F_i(z)\}\right).$$

The respective index set of essentially active selections at x_0 is denoted by $I_F^e(x_0)$ and plays an important role in connection with the generalized Jacobian of F at x_0. Indeed, in [23] it was proved that every PC^1-function is locally Lipschitzian and

$$\partial F(x_0) = \operatorname{conv}\{\nabla F_i(x_0) \mid i \in I_F^e(x_0)\}.$$

To ensure the convergence of so-called bundle methods for nonsmooth optimization in case of nonconvex objectives [24], one usually requires a slightly weakened variant of the following property introduced by Mifflin in [14].

DEFINITION 1.7. *We say that* $f[\mathbb{R}^n \to \mathbb{R}]$ *is* semismooth *at* x *if* f *is Lipschitzian near* x *and the limit*

$$\lim_{\substack{g\in\partial f(x+th')\\ h'\to h,\, t\downarrow 0}} \{\langle g,\, h'\rangle\}$$

exists for all $h \in \mathbb{R}^n$.

This property can be extended in a straightforward way to vector-valued maps [21], and it is not difficult to show that a PC^1-function $F[U \to \mathbb{R}^m]$ is semismooth at each $x \in U$.

2. Shape optimization in discretized contact problems with Coulomb friction. Let us consider a *plane elastic body* represented by a domain $\Omega(\alpha_h)$, $\alpha_h \in U_{ad}^h$ (defined in the previous section), which is unilaterally supported along $\Gamma(\alpha_h)$ by the half-plane $\mathbb{R}^2_-$. On $\Gamma(\alpha_h)$ the following conditions will be prescribed:

$$\left.\begin{aligned} &u_2(x_1,\alpha_h(x_1)) \ge -\alpha_h(x_1), \quad T_2(x_1) \ge 0,\\ &(u_2(x_1,\alpha_h(x_1)) + \alpha_h)T_2(x_1) = 0 \quad \forall x_1 \in (0,a), \end{aligned}\right\} \tag{2.1}$$

$$\left.\begin{aligned} &|T_1(x_1)| \le \mathcal{F}T_2(x_1),\\ &(T_1u_1 + \mathcal{F}T_2|u_1|)(x_1) = 0 \quad \forall x_1 \in (0,a), \end{aligned}\right\} \tag{2.2}$$

where $u = (u_1, u_2)$ denotes the displacement field, $T(x_1) = (T_1(x_1), T_2(x_1))$ stands for the stress vector at a point $(x_1, \alpha_h(x_1)) \in \Gamma(\alpha_h)$, and $\mathcal{F}$ is the coefficient of Coulomb friction. The complementarity conditions (2.1) express the fact that the body cannot penetrate into the rigid foundation $\mathbb{R}^2_-$, that only compression may occur, and that no contact induces zero pressure (the last equation in (2.1)). The set of conditions (2.2) is the mathematical expression of the classical Coulomb law of friction.

The body is subject to body forces $F = (F_1, F_2)$ and surface tractions $P = (P_1, P_2)$ on the part $\Gamma_P(\alpha_h)$. The goal is to find an equilibrium state. For the variational formulation and the mathematical analysis of this problem, we refer to [15]. The discretization of the contact problem with Coulomb friction will be based on a mixed finite element formulation of contact problems with given friction, by means of which Coulomb friction is incorporated into the mathematical model. For a detailed analysis, see [7].

Let $\boldsymbol{V}_h(\alpha_h)$ and L_κ be the same as before. In addition, we shall suppose that L_κ is in duality also with $\widehat{X}_h|_{\widehat{\Gamma}}$. If so, one can define the value $\langle \alpha_h, \mu_\kappa\rangle$ for $\alpha_h \in U_{ad}^h \subseteq \widehat{X}_h|_{\widehat{\Gamma}}$ and $\mu_\kappa \in L_\kappa$. By $\Lambda_\kappa \subset L_\kappa$ we denote the cone of positive functionals:

$$\mu_\kappa \in \Lambda_\kappa \Longrightarrow \langle \mu_\kappa, \hat{\varphi}_h\rangle \ge 0 \qquad \forall \hat{\varphi}_h \in \widehat{X}_h|_{\widehat{\Gamma}},\ \hat{\varphi}_h \ge 0 \text{ on } \widehat{\Gamma}.$$

Let $\alpha_h \in U_{ad}^h$, $g_\kappa \in \Lambda_\kappa$ be given. By a *discretized mixed finite element formulation* of a contact problem with *given friction* g_κ we call the following problem:

$$(\mathcal{P}(\alpha_h, g_\kappa))_h \qquad \left.\begin{array}{l} \text{Find } (u_h, \lambda_\kappa) \in \boldsymbol{V}_h(\alpha_h) \times \Lambda_\kappa \text{ such that} \\ a_{\alpha_h}(u_h, v_h - u_h) - \langle v_{h2} - u_{h2}, \lambda_\kappa \rangle_{\alpha_h} + \langle \mathcal{F} g_\kappa, \Pi_h(|\hat{v}_{h1}| - |\hat{u}_{h1}|) \rangle \\ \qquad\qquad \geq \ell_{\alpha_h}(v_h - u_h) \quad \forall v_h \in \boldsymbol{V}_h(\alpha_h), \\ \langle u_{h2}, \mu_\kappa - \lambda_\kappa \rangle_{\alpha_h} \geq -\langle \alpha_h, \mu_\kappa - \lambda_\kappa \rangle \quad \forall \mu_\kappa \in \Lambda_\kappa, \end{array}\right\}$$

where[1]

- $a_{\alpha_h}(u_h, v_h) := \int_{\Omega(\alpha_h)} \tau_{ij}(u_h)\varepsilon_{ij}(v_h)dx = (\mathcal{C}\varepsilon(u_h), \varepsilon(v_h))_{0,\Omega(\alpha_h)}$;
- $\ell_{\alpha_h}(v_h) := (F, v_h)_{0,\Omega(\alpha_h)} + (P, v_h)_{0,\Gamma_P(\alpha_h)}$;
- $\mathcal{C} \in \mathcal{L}(\mathbb{R}^{2\times 2}_{\rm sym}, \mathbb{R}^{2\times 2}_{\rm sym})$ is a linear mapping of the space of (2×2) symmetric matrices into itself, defining a linear Hooke's law: $\tau = \mathcal{C}\varepsilon$;
- $\Pi_h \in \mathcal{L}(C(\overline{\widehat{\Gamma}}),\ \widehat{\mathcal{V}}_h)$ is a piecewise linear Lagrange interpolation operator on $\widehat{\Gamma}$.

We make the following assumptions:

$$F \in (L^2(\widehat{\Omega}))^2, \qquad P \in (L^2(\widehat{\Gamma}_P))^2; \tag{2.3}$$

$$\mathcal{C} \text{ is constant in } \widehat{\Omega} \text{ and satisfies the usual symmetry and ellipticity conditions (see [16])}; \tag{2.4}$$

$$\mathcal{F} > 0 \text{ is constant on } \widehat{\Gamma}. \tag{2.5}$$

Proposition 2.1 (see [7]). *Let the assumptions* (S) *of section* 1 *and* (2.3)–(2.5) *be satisfied. Then* $(\mathcal{P}(\alpha_h, g_\kappa))_h$ *has a unique solution* (u_h, λ_κ) *for any* $(\alpha_h, g_\kappa) \in U_{ad}^h \times \Lambda_\kappa$.

Remark 2.1. The solution of $(\mathcal{P}(\alpha_h, g_\kappa))_h$ depends on α_h and g_κ. For the sake of simplicity of notation, we use the abbreviated form (u_h, λ_κ) instead of $(u_h(\alpha_h, g_\kappa), \lambda_\kappa(\alpha_h, g_\kappa))$.

We now give an interpretation of the solution to $(\mathcal{P}(\alpha_h, g_\kappa))_h$. To this end, let us introduce

$$\boldsymbol{K}_{h\kappa}(\alpha_h) = \{v_h \in \boldsymbol{V}_h(\alpha_h) \mid \langle v_{h2}, \mu_\kappa \rangle_{\alpha_h} \geq -\langle \alpha_h, \mu_\kappa \rangle \quad \forall \mu_\kappa \in \Lambda_\kappa\}.$$

$\boldsymbol{K}_{h\kappa}(\alpha_h)$ is a closed convex subset of $\boldsymbol{V}_h(\alpha_h)$ approximating the nonpenetration condition on $\Gamma(\alpha_h)$. From the last inequality in $(\mathcal{P}(\alpha_h, g_\kappa))_h$ it follows that $u_h \in \boldsymbol{K}_{h\kappa}(\alpha_h)$. Restricting the choice of test functions to $v_h \in \boldsymbol{K}_{h\kappa}(\alpha_h)$, it is easy to see that the duality term $-\langle v_{h2} - u_{h2}, \lambda_\kappa \rangle_{\alpha_h}$ is nonpositive for any $v_h \in \boldsymbol{K}_{h\kappa}(\alpha_h)$ so that it can be omitted in the first inequality in $(\mathcal{P}(\alpha_h, g_\kappa))_h$. Therefore $u_h \in \boldsymbol{K}_{h\kappa}(\alpha_h)$ solves the variational inequality

$$a_{\alpha_h}(u_h, v_h - u_h) + \mathcal{F}\langle g_\kappa, \Pi_h(|\hat{v}_{h1}| - |\hat{u}_{h1}|) \rangle \geq \ell_{\alpha_h}(v_h - u_h) \quad \forall v_h \in \boldsymbol{K}_{h\kappa}(\alpha_h), \tag{2.6}$$

while $\lambda_\kappa \in \Lambda_\kappa$ is the Lagrange multiplier releasing the constraint $u_h \in \boldsymbol{K}_{h\kappa}(\alpha_h)$.

We now present the algebraic form of $(\mathcal{P}(\alpha_h, g_\kappa))_h$, which will be used in the analysis that follows. Let $\{\varphi_i\}_{i=1}^n$, $\{\hat{\xi}_i\}_{i=1}^m$, $\{\hat{\pi}_i\}_{i=1}^p$, $\{\hat{\omega}_i\}_{i=1}^d$ be basis functions of $\boldsymbol{V}_h(\alpha_h)$, L_κ, $\widehat{\mathcal{V}}_h$, and $\widehat{X}_h|_{\widehat{\Gamma}}$. (Observe that the dimension of $\boldsymbol{V}_h(\alpha_h)$ does not depend on $\alpha_h \in U_{ad}^h$!)

[1] We use the summation convention and standard notations of linear elasticity.

We shall suppose that $\hat{\xi}_i \in \Lambda_\kappa$ for $i = 1, \dots, m$. As usual, $\boldsymbol{V}_h(\alpha_h)$, Λ_κ, $\widehat{\mathcal{V}}_h$, and $\widehat{X}_h|_{\widehat{\Gamma}}$ are isometrically isomorphic with $\mathbb{R}^n$, $\mathbb{R}^m_+$, $\mathbb{R}^p$, and $\mathbb{R}^d$, respectively. Since $\alpha_h \in U^h_{ad} \subseteq \widehat{X}_h|_{\widehat{\Gamma}}$, it can be identified with a vector $\in \mathbb{R}^d$ whose components are given by the nodal values of α_h, and the set U^h_{ad} itself can be identified with a compact subset $\mathcal{U} \subseteq \mathbb{R}^d$. Let us note that $d = p + \operatorname{card}(\overline{\Gamma}_u(\alpha_h) \cap \overline{\Gamma}(\alpha_h))$.

Let $\in \mathcal{U}$, $\boldsymbol{g} \in \mathbb{R}^m_+$ be given. The algebraic form of $(\mathcal{P}(\alpha_h, g_\kappa))_h$ reads as follows:

$$(\mathcal{P}(\ ,\boldsymbol{g}))\quad \left.\begin{array}{l} \text{Find } (\boldsymbol{u}, \breve{\ }) \in \mathbb{R}^n \times \mathbb{R}^m_+ \text{ such that} \\ (\boldsymbol{A}(\ )\boldsymbol{u}, \boldsymbol{v}-\boldsymbol{u})_{\mathbb{R}^n} - (\boldsymbol{B}(\boldsymbol{v}_\nu - \boldsymbol{u}_\nu), \breve{\ })_{\mathbb{R}^m} + \mathcal{F}(\boldsymbol{g}, \boldsymbol{B}(|\boldsymbol{v}_\tau| - |\boldsymbol{u}_\tau|))_{\mathbb{R}^m} \\ \qquad\qquad \geq (\boldsymbol{\ell}(\ ), \boldsymbol{v}-\boldsymbol{u})_{\mathbb{R}^n} \quad \forall \boldsymbol{v} \in \mathbb{R}^n, \\ (\boldsymbol{B}\boldsymbol{u}_\nu, \bar{\ } - \breve{\ })_{\mathbb{R}^m} \geq -(\widetilde{\boldsymbol{B}}\ \ , \bar{\ } - \breve{\ })_{\mathbb{R}^m} \quad \forall \bar{\ } \in \mathbb{R}^m_+, \end{array}\right\}$$

where $\boldsymbol{A}(\ ) \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$, $\boldsymbol{B} \in \mathcal{L}(\mathbb{R}^p, \mathbb{R}^m)$, and $\widetilde{\boldsymbol{B}} \in \mathcal{L}(\mathbb{R}^d, \mathbb{R}^m)$ are matrices with the following elements:

$$\begin{array}{ll} a_{ij}(\ ) = a_{\alpha_h}(\varphi_i, \varphi_j), & i,j = 1, \dots, n, \\ b_{kl} = \langle \hat{\xi}_k, \hat{\pi}_l \rangle, & k = 1, \dots, m;\ l = 1, \dots, p, \\ \tilde{b}_{kt} = \langle \hat{\xi}_k, \hat{\omega}_t \rangle, & k = 1, \dots, m;\ t = 1, \dots, d. \end{array}$$

The symbols $\boldsymbol{v}_\tau$ and $\boldsymbol{v}_\nu \in \mathbb{R}^p$ stand for subvectors made of those components of $\boldsymbol{v} \in \mathbb{R}^n$ which correspond to the tangential and normal displacements, respectively, at the contact nodes, and $|\boldsymbol{z}| := (|z_1|, \dots, |z_p|)$ for any $\boldsymbol{z} \in \mathbb{R}^p$. Finally, the components of $\boldsymbol{\ell}(\ )$ are given by

$$\ell^i(\ ) = (F, \varphi_i)_{0,\Omega(\alpha_h)} + (P, \varphi_i)_{0,\Gamma_P(\alpha_h)}.$$

Remark 2.2. From the construction of $\boldsymbol{V}_h(\alpha_h)$, $\alpha_h \in U^h_{ad}$, it easily follows that the mappings $\boldsymbol{A}: \ \to \boldsymbol{A}(\ )$, $\boldsymbol{\ell}: \ \to \boldsymbol{\ell}(\ )$ are Lipschitzian in $\mathcal{U}$.

We now present two types of L_κ satisfying condition (S) of section 1.

Type I. Let $\mathcal{N} = \{C_i\}_{i=1}^p$ be the set of all contact nodes of $\mathcal{T}_h(\alpha_h)$, i.e., $C_i = (\hat{x}_i, \alpha_h(\hat{x}_i))$, where $\{\hat{x}_i\}$ are those nodes of $\widehat{\Delta}_h$ such that $C_i \in \overline{\Gamma}(\alpha_h) \backslash \overline{\Gamma}_u(\alpha_h)$. With any $C_i \in \mathcal{N}$ the Dirac distribution δ_i will be associated: $\langle \delta_i, \hat{\varphi} \rangle = \hat{\varphi}(C_i)$ for all $\hat{\varphi} \in C(\overline{\widehat{\Gamma}})$. We define $L_\kappa = \{\delta_1, \dots, \delta_p\}$ and $\Lambda_\kappa = \{\mu_\kappa \in L_\kappa \mid \mu_\kappa = \bar{\ } \odot\ , \bar{\ } \in \mathbb{R}^p_+\}$, where $= (\delta_1, \dots, \delta_p)$ and $\bar{\ } \odot\ := \mu^i \delta_i$. It is readily seen that (S) is satisfied and

$$\boldsymbol{K}_{h\kappa}(\alpha_h) := \boldsymbol{K}_h(\alpha_h) = \{v_h \in \boldsymbol{V}_h(\alpha_h) \mid v_{h2}(C_i) \geq -\alpha_h(\hat{x}_i),\ C_i \in \mathcal{N},\ \forall i = 1, \dots, p\}$$

is the inner approximation of the set of kinematically admissible displacements.

Type II. Let $\widehat{\Delta}_\kappa$ be another partition of $[0, a]$ into q segments S_i, $i = 1, \dots, q$, generally different from $\widehat{\Delta}_h$, and let χ_i be the characteristic function of S_i, $i = 1, \dots, q$. We define

$$\begin{aligned} L_\kappa &= \{\mu_\kappa \in L^2(\widehat{\Gamma}) \mid \mu_\kappa = \mu^i \chi_i,\ \bar{\ } \in \mathbb{R}^q\}, \\ \Lambda_\kappa &= \{\mu_\kappa \in L_\kappa \mid \mu_\kappa \geq 0 \text{ a.e. on } \widehat{\Gamma}\}, \\ \langle \mu_\kappa, \hat{\varphi}_h \rangle &:= \int_0^a \mu_\kappa \hat{\varphi}_h dx_1 \quad \forall \hat{\varphi}_h \in \widehat{X}_h|_{\widehat{\Gamma}}. \end{aligned}$$

One can show (see [8]) that for this type of L_κ, condition (S) is satisfied, provided that the ratio $\max_i |S_i| / \max_{\widehat{T} \in \widehat{\mathcal{T}}_h} \operatorname{diam} \widehat{T}$ is "sufficiently large." In this case,

$$\boldsymbol{K}_{h\kappa}(\alpha_h) = \left\{ v_h \in \boldsymbol{V}_h(\alpha_h) \mid \int_{S_i} v_{h2}(x_1, \alpha_h(x_1)) dx_1 \geq -\int_{S_i} \alpha_h(x_1) dx_1\ \forall i = 1, \dots, q \right\}$$

is the external approximation of the set of kinematically admissible displacements.

Proposition 2.1 enables us to define the mapping $\Phi_{h\kappa} : U_{ad}^h \times \Lambda_\kappa \to \Lambda_\kappa$ by

$$\Phi_{h\kappa}(\alpha_h, g_\kappa) = \lambda_\kappa, \quad (\alpha_h, g_\kappa) \in U_{ad}^h \times \Lambda_\kappa,$$

where λ_κ is the second component of the solution to $(\mathcal{P}(\alpha_h, g_\kappa))_h$.

In what follows, we shall examine basic properties of $\Phi_{h\kappa}$. We start with the following result.

PROPOSITION 2.2. *The solutions (u_h, λ_κ) of $(\mathcal{P}(\alpha_h, g_\kappa))_h$ are uniformly bounded with respect to $(\alpha_h, g_\kappa) \in U_{ad}^h \times \Lambda_\kappa$ and $\mathcal{F} \in \mathbb{R}^1_+$: There exists a constant $c > 0$ such that*

$$\begin{aligned} &\|u_h\|_{1,\Omega(\alpha_h)} \le c, \\ &\|\lambda_\kappa\|_{-h,\widehat{\Gamma}} \le c \quad \forall (\alpha_h, g_\kappa) \in U_{ad}^h \times \Lambda_\kappa, \quad \forall \mathcal{F} \in \mathbb{R}^1_+. \end{aligned} \tag{2.7}$$

Proof. Inserting $0 \in \boldsymbol{K}_{h\kappa}(\alpha_h)$ into (2.6), we obtain

$$a_{\alpha_h}(u_h, u_h) \le \ell_{\alpha_h}(u_h) - \mathcal{F}\langle g_\kappa, \Pi_h(|\hat{u}_{h1}|)\rangle \le \ell_{\alpha_h}(u_h) \le \overline{c}\|u_h\|_{1,\Omega(\alpha_h)}, \tag{2.8}$$

where $\overline{c} := \overline{c}(\|F\|_{0,\widehat{\Omega}}, \|P\|_{0,\widehat{\Gamma}_p})$ is a positive constant depending solely on the indicated parameters. The energy norm on the left-hand side of (2.8) can be bounded from below by using the Korn's inequality:

$$\beta\|u_h\|^2_{1,\Omega(\alpha_h)} \le a_{\alpha_h}(u_h, u_h), \tag{2.9}$$

where $\beta > 0$ is a constant which can be chosen independently of $\alpha_h \in U_{ad}^h$, owing to the fact that $\mathcal{O}_h$ is the system of domains satisfying the uniform cone property (see [9]). From (2.8) and (2.9), $(2.7)_1$ follows.

Let $v_h \in \boldsymbol{V}_h(\alpha_h)$ be of the form $v_h = (u_{h1}, u_{h2} \pm w_h)$, $w_h \in V_h(\alpha_h)$, and define $z_h := v_h - u_h = (0, \pm w_h)$. Substitution of such v_h into $(\mathcal{P}(\alpha_h, g_\kappa))_h$ yields

$$\begin{aligned} \langle w_h, \lambda_\kappa\rangle_{\alpha_h} &= a_{\alpha_h}(u_h, z_h) - \ell_{\alpha_h}(z_h) \\ &\le c\{\|u_h\|_{1,\Omega(\alpha_h)}\ |w_h|_{1,\Omega(\alpha_h)} + |w_h|_{1,\Omega(\alpha_h)}\}, \end{aligned} \tag{2.10}$$

where $c > 0$ is a constant not depending on $(\alpha_h, g_\kappa) \in U_{ad}^h \times \Lambda_\kappa$ and on $\mathcal{F} \in \mathbb{R}^1_+$. (We have used the fact that the constant in the Friedrichs inequality and the norm of the trace mapping from $V_h(\alpha_h)$ to $\mathcal{V}_h(\alpha_h)$ can be chosen independently of $\alpha_h \in U_{ad}^h$.) From (2.10) and $(2.7)_1$ we obtain

$$\|\lambda_\kappa\|_{-h,\alpha_h} = \sup_{\substack{w_h \in V_h(\alpha_h) \\ w_h \neq 0}} \frac{\langle w_h, \lambda_\kappa\rangle_{\alpha_h}}{|w_h|_{1,\Omega(\alpha_h)}} \le c\,. \tag{2.11}$$

This and Proposition 1.4 result in $(2.7)_2$. ∎

Remark 2.3. The constant c in (2.7) depends generally on the discretization parameter h. Two additional assumptions—namely, a slope-type condition imposed on $\alpha_h \in U_{ad}^h$, which is uniform with respect to h, and the Ladyzhenskaya–Babuška–Brezzi condition satisfied by the couple $\{\widehat{V}_h, \Lambda_\kappa\}$—ensure that (2.7) is uniform also with respect to $h \to 0_+$. In this case, the mesh dependent norm $\|\cdot\|_{-h,\widehat{\Gamma}}$ is replaced by the classical dual norm $\|\cdot\|_{-1/2,\widehat{\Gamma}}$.

COROLLARY 2.3. *There exists an $r_0 > 0$ independent of $(\alpha_h, g_\kappa) \in U_{ad}^h \times \Lambda_\kappa$ and $\mathcal{F} \in \mathbb{R}^1_+$ such that $\Phi_{h\kappa}(U_{ad}^h \times \Lambda_\kappa) \subset \Lambda_\kappa^{r_0} = \{\mu_\kappa \in \Lambda_\kappa |\ \|\mu_\kappa\|_{-h,\widehat{\Gamma}} \le r_0\}$. In the*

situation described in Remark 2.3, *the radius* r_0 *can also be chosen independently of* $h \to 0_+$.

PROPOSITION 2.4. *The mapping* $\Phi_{h\kappa}$ *is continuous in* $U_{ad}^h \times \Lambda_\kappa$.

Proof. We use the algebraic form of $(\mathcal{P}(\alpha_h, g_\kappa))_h$. Let $\{(\ _k, \boldsymbol{g}_k)\}$, $(\ _k, \boldsymbol{g}_k) \in \mathcal{U} \times \mathbb{R}^m_+$, be a convergent sequence: $\ _k \to\ \ \in \mathcal{U}$, $\boldsymbol{g}_k \to \boldsymbol{g} \in \mathbb{R}^m_+$, $k \to \infty$. Denote by $(\boldsymbol{u}_k, \breve{}_k)$ the solution of $(\mathcal{P}(\ _k, \boldsymbol{g}_k))$. From Proposition 2.2 we know that $\{(\boldsymbol{u}_k, \breve{}_k)\}$ is bounded. Therefore one can pass to an appropriate subsequence tending to $(\boldsymbol{u}, \breve{}) \in \mathbb{R}^n \times \mathbb{R}^m_+$. It is easy to show that $(\boldsymbol{u}, \breve{}\,)$ solves the limit problem $(\mathcal{P}(\ , \boldsymbol{g}))$, making use of Remark 2.2. Since $(\mathcal{P}(\ , \boldsymbol{g}))$ has a unique solution, the whole sequence $\{(\boldsymbol{u}_k, \breve{}_k)\}$ tends to $(\boldsymbol{u}, \breve{}\,)$. ∎

With this result at hand, we have the following.

PROPOSITION 2.5. *For any* $\alpha_h \in U_{ad}^h$ *and any* $\mathcal{F} \in \mathbb{R}^1_+$ *there exists a fixed point of* $\Phi_{h\kappa}(\alpha_h, \cdot)$ *in* $\Lambda_\kappa^{r_0}$.

Proof. The proof follows from Corollary 2.3, Proposition 2.4, and the Brouwer fixed-point theorem. ∎

The importance of fixed points of $\Phi_{h\kappa}(\alpha_h, \cdot)$ follows from the next definition.

DEFINITION 2.6. *By a solution to the discretized contact problem with Coulomb friction we mean any solution to* $(\mathcal{P}(\alpha_h, g_\kappa))_h$, *where* $g_\kappa = \Phi_{h\kappa}(\alpha_h, g_\kappa)$, $g_\kappa \in \Lambda_\kappa$.

Denote by $\mathcal{G}$ a subset of $U_{ad}^h \times \Lambda_\kappa$ defined as follows:

$$(\alpha_h, g_\kappa) \in \mathcal{G} \Longleftrightarrow \alpha_h \in U_{ad}^h, \qquad g_\kappa \text{ is a fixed point of } \Phi_{h\kappa}(\alpha_h, \cdot) \text{ in } \Lambda_\kappa^{r_0}.$$

Using the same approach as in Proposition 2.4, one can prove the next result.

PROPOSITION 2.7. $\mathcal{G}$ *is a compact subset of* $U_{ad}^h \times \Lambda_\kappa^{r_0}$.

Next we shall study under which conditions the mapping $\Phi_{h\kappa}(\alpha_h, \cdot)$ is contractive in $\Lambda_\kappa^{r_0}$.

PROPOSITION 2.8. *There exists some* $\mathcal{F}_0 > 0$ *such that for any* $\mathcal{F} \in (0, \mathcal{F}_0)$ *the mapping* $\Phi_{h\kappa}(\alpha_h, \cdot)$ *is contractive in* $\Lambda_\kappa^{r_0}$ *uniformly with respect to* $\alpha_h \in U_{ad}^h$:

$$\exists q \in (0,1): \ \|\Phi_{h\kappa}(\alpha_h, g_\kappa) - \Phi_{h\kappa}(\alpha_h, \overline{g}_\kappa)\|_{-h,\widehat{\Gamma}} \le q \|g_\kappa - \overline{g}_\kappa\|_{-h,\widehat{\Gamma}}$$

holds for any $\alpha_h \in U_{ad}^h$ *and* $g_\kappa, \overline{g}_\kappa \in \Lambda_\kappa^{r_0}$.

Proof. Let (u_h, λ_κ), $(\overline{u}_h, \overline{\lambda}_\kappa)$ be solutions to $(\mathcal{P}(\alpha_h, g_\kappa))_h$, $(\mathcal{P}(\alpha_h, \overline{g}_\kappa))_h$, respectively, $g_\kappa, \overline{g}_\kappa \in \Lambda_\kappa^{r_0}$. From (2.6) it follows that

$$\begin{aligned} a_{\alpha_h}(u_h, v_h - u_h) + \mathcal{F}\langle g_\kappa, \Pi_h(|\hat{v}_{h1}| - |\hat{u}_{h1}|)\rangle &\ge \ell_{\alpha_h}(v_h - u_h), \\ a_{\alpha_h}(\overline{u}_h, v_h - \overline{u}_h) + \mathcal{F}\langle \overline{g}_\kappa, \Pi_h(|\hat{v}_{h1}| - |\hat{\overline{u}}_{h1}|)\rangle &\ge \ell_{\alpha_h}(v_h - \overline{u}_h) \end{aligned}$$

holds for any $v_h \in \boldsymbol{K}_{h\kappa}(\alpha_h)$. Inserting $v_h := \overline{u}_h$ and u_h into the first and second inequality, respectively, and summing them up, we obtain

$$\begin{aligned} \beta \|u_h - \overline{u}_h\|^2_{1,\Omega(\alpha_h)} &\le a_{\alpha_h}(u_h - \overline{u}_h, u_h - \overline{u}_h) \\ &\le \mathcal{F}\langle g_\kappa - \overline{g}_\kappa, \Pi_h(|\hat{\overline{u}}_{h1}| - |\hat{u}_{h1}|)\rangle \\ &\le \mathcal{F} \|g_\kappa - \overline{g}_\kappa\|_{-h,\widehat{\Gamma}} \|\widehat{w}_h\|_{h,\widehat{\Gamma}}, \end{aligned} \tag{2.12}$$

where for brevity of notation we have defined $\widehat{w}_h := \Pi_h(|\hat{\overline{u}}_{h1}| - |\hat{u}_{h1}|) \in \widehat{\mathcal{V}}_h$. From Proposition 1.1 it follows that $\|\widehat{w}_h\|_{h,\widehat{\Gamma}} = |\hat{u}_h|_{1,\widehat{\Omega}}$, where $\hat{u}_h \in \widehat{V}_h$ solves

$$\int_{\widehat{\Omega}} \nabla \hat{u}_h \cdot \nabla \hat{v}_h \, d\hat{x} = 0 \qquad \forall \hat{v}_h \in \widehat{X}_{h0}, \ \ \hat{u}_h = \widehat{w}_h \text{ on } \widehat{\Gamma}\,. \tag{2.13}$$

The solution $\hat{u}_h$ can be written as

$$\hat{u}_h = \hat{z}_h + r_h \widehat{w}_h, \quad \hat{z}_h \in \widehat{X}_{h0}, \tag{2.14}$$

where $r_h \widehat{w}_h \in \widehat{V}_h$ is the extension of $\widehat{w}_h$ from $\widehat{\Gamma}$ into $\widehat{\Omega}$ satisfying (1.8). This, (2.13), and (2.14) entail that

$$\|\hat{u}_h\|_{1,\widehat{\Omega}} \le c_0 h^{-1/2} \|\widehat{w}_h\|_{0,\widehat{\Gamma}}. \tag{2.15}$$

Π_h being the piecewise linear Lagrange interpolation operator preserves the monotonicity,

$$|w_h| = |\Pi_h(|\hat{\overline{u}}_{h1}| - |\hat{u}_{h1}|)| \le \Pi_h(|\hat{\overline{u}}_{h1} - \hat{u}_{h1}|) \quad \text{on } \widehat{\Gamma},$$

implying that

$$\|\widehat{w}_h\|_{0,\widehat{\Gamma}} \le \|\Pi_h(|\hat{\overline{u}}_{h1} - \hat{u}_{h1}|)\|_{0,\widehat{\Gamma}} \le c\|u_h - \overline{u}_h\|_{1,\Omega(\alpha_h)},$$

where $c > 0$ does not depend on $\alpha_h \in U_{ad}^h$ and $h > 0$. This, together with (2.12) and (2.15), yields

$$\|u_h - \overline{u}_h\|_{1,\Omega(\alpha_h)} \le c\mathcal{F} h^{-1/2} \|g_\kappa - \overline{g}_\kappa\|_{-h,\widehat{\Gamma}}. \tag{2.16}$$

As in (2.10), we have

$$\langle w_h, \lambda_\kappa - \overline{\lambda}_\kappa \rangle_{\alpha_h} = a_{\alpha_h}(\overline{u}_h - u_h, z_h)$$

for any $z_h = (0, \pm w_h)$, $w_h \in V_h(\alpha_h)$. Hence

$$\begin{aligned} c_1 \|\lambda_\kappa - \overline{\lambda}_\kappa\|_{-h,\widehat{\Gamma}} &\le \|\lambda_\kappa - \overline{\lambda}_\kappa\|_{-h,\alpha_h} \\ &\le c\|\overline{u}_h - u_h\|_{1,\Omega(\alpha_h)} \\ &\le c\mathcal{F} h^{-1/2} \|g_\kappa - \overline{g}_\kappa\|_{-h,\widehat{\Gamma}}, \end{aligned}$$

taking into account Proposition 1.4 and (2.16). If $q := c\mathcal{F} h^{-1/2}/c_1 < 1$, the mapping $\Phi_{h\kappa}(\alpha_h, \cdot)$ is contractive in $\Lambda_\kappa^{r_0}$ uniformly with respect to U_{ad}^h, and, since c, c_1 do not depend on $h > 0$, in view of the regularity of $\{\widehat{\mathcal{T}}_h\}, h \to 0+$, one has $\mathcal{F}_0 = O(h^{1/2})$. □

COROLLARY 2.9. *For any $\alpha_h \in U_{ad}^h$ and $\mathcal{F} \in (0, \mathcal{F}_0)$, the fixed point of $\Phi_{h\kappa}(\alpha_h, \cdot)$ is unique in $\Lambda_h^{r_0}$ and can be revealed by the method of successive approximations.*

Remark 2.4. Observe that the contractivity of $\Phi_{h\kappa}(\alpha_h, \cdot)$ is *mesh dependent*, and this dependency *cannot* be removed.

PROPOSITION 2.10. *The mapping $\Phi_{h\kappa}(\cdot, g_\kappa)$ is Lipschitzian in U_{ad}^h uniformly with respect to $g_\kappa \in \Lambda_\kappa^{r_0}$ and $\mathcal{F} \in \mathbb{R}_+^1$:*

$$\begin{gathered} \exists c > 0 \text{ such that } \forall \alpha_h, \overline{\alpha}_h \in U_{ad}^h,\ \forall g_\kappa \in \Lambda_\kappa^{r_0},\ \forall \mathcal{F} \in \mathbb{R}_+^1 \\ \|\Phi_{h\kappa}(\alpha_h, g_\kappa) - \Phi_{h\kappa}(\overline{\alpha}_h, g_\kappa)\|_{-h,\widehat{\Gamma}} \le c\|\alpha_h - \overline{\alpha}_h\|_{h,\widehat{\Gamma}}. \end{gathered}$$

Proof. We use the algebraic form of (2.6). The set $\boldsymbol{K}_{h\kappa}(\alpha_h)$ can be identified with the convex subset $\mathcal{K}(\ \ )$ of $\mathbb{R}^n$ given by

$$\mathcal{K}(\ \ ) = \{\boldsymbol{v} \in \mathbb{R}^n \mid \boldsymbol{B}\boldsymbol{v}_\nu + \widetilde{\boldsymbol{B}} \quad \ge \boldsymbol{0}\},$$

with the same meaning of symbols as in $(\mathcal{P}(\ ,\boldsymbol{g}))$. Further, let $\mathcal{K} := \mathcal{K}(\mathbf{0})$. If $(\boldsymbol{u}(\ ), \check{}(\ ))$ solves $(\mathcal{P}(\ ,\boldsymbol{g}))$, then $\boldsymbol{u}(\ ) \in \mathcal{K}(\ )$ and

$$(\boldsymbol{A}(\ )\boldsymbol{u}(\ ),\ \boldsymbol{v}-\boldsymbol{u}(\ ))_{\mathbb{R}^n} + \mathcal{F}(\boldsymbol{g}, \boldsymbol{B}(|\boldsymbol{v}_\tau| - |\boldsymbol{u}_\tau(\ )|))_{\mathbb{R}^m} \geq (\boldsymbol{\ell}(\ ), \boldsymbol{v}-\boldsymbol{u}(\ ))_{\mathbb{R}^n} \qquad \forall \boldsymbol{v} \in \mathcal{K}(\ ). \tag{2.17}$$

Let $\ _c \in \mathbb{R}^p$ be such that $\boldsymbol{B}\ _c = \widetilde{\boldsymbol{B}}\ $, and denote by $\underset{\sim}{\ } \in \mathbb{R}^n$ the vector such that $\underset{\sim}{\ }^\nu = \ _c$ and the remaining components are equal to zero. Any vector $\boldsymbol{v} \in \mathcal{K}(\ )$ can be written as $\boldsymbol{v} = \boldsymbol{z} - \underset{\sim}{\ }$, $\boldsymbol{z} \in \mathcal{K}$, and, in particular, $\boldsymbol{u}(\ ) = \boldsymbol{U}(\ ) - \underset{\sim}{\ }$, $\boldsymbol{U}(\ ) \in \mathcal{K}$. Inserting these expressions into (2.17), we obtain the inequality for $\boldsymbol{U}(\ )$:

$$(\boldsymbol{A}(\ )\boldsymbol{U}(\ ),\ \boldsymbol{z}-\boldsymbol{U}(\ ))_{\mathbb{R}^n} + \mathcal{F}(\boldsymbol{g}, \boldsymbol{B}(|\boldsymbol{z}_\tau| - |\boldsymbol{U}_\tau(\ )|))_{\mathbb{R}^m} \geq (\boldsymbol{A}(\ )\underset{\sim}{\ } + \boldsymbol{\ell}(\ ), \boldsymbol{z}-\boldsymbol{U}(\ ))_{\mathbb{R}^n} \qquad \forall \boldsymbol{z} \in \mathcal{K}, \tag{2.18}$$

using that $\boldsymbol{v}_\tau = \boldsymbol{z}_\tau$ and $\boldsymbol{u}_\tau(\ ) = \boldsymbol{U}_\tau(\ )$. A similar inequality can be written for another design vector $\mathsf{fi} \in \mathcal{U}$:

$$(\boldsymbol{A}(\mathsf{fi})\boldsymbol{U}(\mathsf{fi}),\ \boldsymbol{z}-\boldsymbol{U}(\mathsf{fi}))_{\mathbb{R}^n} + \mathcal{F}(\boldsymbol{g}, \boldsymbol{B}(|\boldsymbol{z}_\tau| - |\boldsymbol{U}_\tau(\mathsf{fi})|))_{\mathbb{R}^m} \geq (\boldsymbol{A}(\mathsf{fi})\underset{\sim}{\mathsf{fi}} + \boldsymbol{\ell}(\mathsf{fi}), \boldsymbol{z}-\boldsymbol{U}(\mathsf{fi}))_{\mathbb{R}^n} \qquad \forall \boldsymbol{z} \in \mathcal{K}. \tag{2.19}$$

Substituting $\boldsymbol{z} := \boldsymbol{U}(\mathsf{fi})$ and $\boldsymbol{U}(\ )$ into (2.18) and (2.19), respectively, and summing both inequalities, we obtain

$$(\boldsymbol{A}(\ )\boldsymbol{U}(\ ) - \boldsymbol{A}(\mathsf{fi})\boldsymbol{U}(\mathsf{fi}), \boldsymbol{U}(\mathsf{fi}) - \boldsymbol{U}(\ ))_{\mathbb{R}^n} \geq (\boldsymbol{A}(\ )\underset{\sim}{\ } + \boldsymbol{\ell}(\ ) - \boldsymbol{A}(\mathsf{fi})\underset{\sim}{\mathsf{fi}} - \boldsymbol{\ell}(\mathsf{fi}), \boldsymbol{U}(\mathsf{fi}) - \boldsymbol{U}(\ ))_{\mathbb{R}^n}. \tag{2.20}$$

Adding and subtracting the vector $\boldsymbol{A}(\ )\boldsymbol{U}(\mathsf{fi})$ to/from the left of (2.20), we arrive at

$$\begin{aligned}(\boldsymbol{A}(\ )(\boldsymbol{U}(\mathsf{fi}) - \boldsymbol{U}(\ )), \boldsymbol{U}(\mathsf{fi}) - \boldsymbol{U}(\ ))_{\mathbb{R}^n} &\leq ((\boldsymbol{A}(\ ) - \boldsymbol{A}(\mathsf{fi}))\boldsymbol{U}(\mathsf{fi}), \boldsymbol{U}(\mathsf{fi}) - \boldsymbol{U}(\ ))_{\mathbb{R}^n} \\ &\quad + (\boldsymbol{A}(\mathsf{fi})\underset{\sim}{\mathsf{fi}} + \boldsymbol{\ell}(\mathsf{fi}) - \boldsymbol{A}(\ )\underset{\sim}{\ } - \boldsymbol{\ell}(\ ), \boldsymbol{U}(\mathsf{fi}) - \boldsymbol{U}(\ ))_{\mathbb{R}^n}.\end{aligned} \tag{2.21}$$

Since $\boldsymbol{A}(\ )$ is positive definite uniformly in $\mathcal{U}$ (from the Korn's inequality and definition of $V_h(\alpha_h)$), it follows from (2.21) that there exists a positive constant m_0 such that

$$\begin{aligned}m_0 \|\boldsymbol{U}(\mathsf{fi}) - \boldsymbol{U}(\ )\|^2 &\leq \|\boldsymbol{A}(\ ) - \boldsymbol{A}(\mathsf{fi})\|\,\|\boldsymbol{U}(\mathsf{fi})\|\,\|\boldsymbol{U}(\mathsf{fi}) - \boldsymbol{U}(\ )\| \\ &\quad + \{\|\boldsymbol{A}(\mathsf{fi})\underset{\sim}{\mathsf{fi}} - \boldsymbol{A}(\ )\underset{\sim}{\ }\| + \|\boldsymbol{\ell}(\mathsf{fi}) - \boldsymbol{\ell}(\ )\|\}\|\boldsymbol{U}(\mathsf{fi}) - \boldsymbol{U}(\ )\|,\end{aligned}$$

where m_0 does not depend on $\ ,\mathsf{fi} \in \mathcal{U}$. Thus one can find constants $c, c_1 > 0$ that do not depend on $\mathcal{F} \in \mathbb{R}^1_+$, $\ ,\mathsf{fi} \in \mathcal{U}$, and $\boldsymbol{g} \in \mathbb{R}^m_+$ such that

$$\begin{aligned}\|\boldsymbol{U}(\mathsf{fi}) - \boldsymbol{U}(\ )\| &\leq c\{\|\boldsymbol{A}(\ ) - \boldsymbol{A}(\mathsf{fi})\| + \|\boldsymbol{\ell}(\ ) - \boldsymbol{\ell}(\mathsf{fi})\| \\ &\qquad + \|\boldsymbol{A}(\mathsf{fi})\underset{\sim}{\mathsf{fi}} - \boldsymbol{A}(\ )\underset{\sim}{\ }\|\} \\ &\leq c_1\|\ - \mathsf{fi}\|,\end{aligned} \tag{2.22}$$

taking into account Remark 2.2 and the fact that $\|\boldsymbol{U}(\mathsf{fi})\|$ is bounded on $\mathcal{U}$. Consequently,

$$\|\boldsymbol{u}(\mathsf{fi}) - \boldsymbol{u}(\quad)\| \le \|\boldsymbol{U}(\mathsf{fi}) - \boldsymbol{U}(\quad)\| + \|\underset{\sim}{\mathsf{fi}} - \underset{\sim}{\quad}\| \le c\|\quad - \mathsf{fi}\|. \tag{2.23}$$

Let $\boldsymbol{v} = \boldsymbol{u}(\quad) \pm \boldsymbol{z}$, where $\boldsymbol{z} = (\boldsymbol{0}, \boldsymbol{z}_\nu) \in \mathbb{R}^n$, meaning that $\boldsymbol{z}$ is a vector with nonzero components only at the positions corresponding to the normal displacements at the contact nodes. Its substitution into the first inequality in $(\mathcal{P}(\quad, \boldsymbol{g}))$ yields

$$(\boldsymbol{B}\boldsymbol{z}_\nu, \breve{}(\quad))_{\mathbb{R}^m} = (\boldsymbol{\ell}(\quad), \boldsymbol{z})_{\mathbb{R}^n} - (\boldsymbol{A}(\quad)\boldsymbol{u}(\quad), \boldsymbol{z})_{\mathbb{R}^n}. \tag{2.24}$$

The same holds for the design variable $\mathsf{fi} \in \mathcal{U}$:

$$(\boldsymbol{B}\boldsymbol{z}_\nu, \breve{}(\mathsf{fi}))_{\mathbb{R}^m} = (\boldsymbol{\ell}(\mathsf{fi}), \boldsymbol{z})_{\mathbb{R}^n} - (\boldsymbol{A}(\mathsf{fi})\boldsymbol{u}(\mathsf{fi}), \boldsymbol{z})_{\mathbb{R}^n}. \tag{2.25}$$

From (2.24) and (2.25) we obtain

$$\begin{aligned}\|\breve{}(\quad) - \breve{}(\mathsf{fi})\| &= \sup_{\boldsymbol{z}\in\mathbb{R}^n} \frac{(\breve{}(\quad) - \breve{}(\mathsf{fi}), \boldsymbol{B}\boldsymbol{z}_\nu)_{\mathbb{R}^m}}{\|\boldsymbol{z}\|} \\ &\le \|\boldsymbol{\ell}(\quad) - \boldsymbol{\ell}(\mathsf{fi})\| + \|\boldsymbol{A}(\quad)\boldsymbol{u}(\quad) - \boldsymbol{A}(\mathsf{fi})\boldsymbol{u}(\mathsf{fi})\| \\ &\le c\|\quad - \mathsf{fi}\|,\end{aligned}$$

where $c > 0$ is a constant which does not depend on $\mathcal{F}$ and $\boldsymbol{g} \in \mathbb{R}^m_+$, making use of condition (S) of section 1, (2.23), and Remark 2.2. □

The main result of this section is the following.

PROPOSITION 2.11. *Let $\mathcal{F} \in (0, \mathcal{F}_0)$, where $\mathcal{F}_0$ is the same as in Proposition* 2.8. *Then the mapping $\lambda_\kappa : U^h_{ad} \to \Lambda^{r_0}_\kappa$ associating with any $\alpha_h \in U^h_{ad}$ the unique fixed point $\lambda_\kappa(\alpha_h)$ of $\Phi_{h\kappa}(\alpha_h, \cdot)$ is Lipschitzian in U^h_{ad}.*

Proof. Since $\Phi_{h\kappa}(\alpha_h, \cdot)$ is contractive in $\Lambda^{r_0}_\kappa$ for any $\mathcal{F} \in (0, \mathcal{F}_0)$, $\alpha_h \in U^h_{ad}$, the method of successive approximations is convergent for any $\alpha_h \in U^h_{ad}$ and any initial approximation from $\Lambda^{r_0}_\kappa$. Let $\alpha_h,\ \overline{\alpha}_h \in U^h_{ad}$, and $g^*_\kappa \in \Lambda^{r_0}_\kappa$ be given, and consider the iterative process:

$$\begin{aligned}&g^{(k+1)}_\kappa = \Phi_{h\kappa}(\alpha_h, g^{(k)}_\kappa), \qquad \overline{g}^{(k+1)}_\kappa = \Phi_{h\kappa}(\overline{\alpha}_h, \overline{g}^{(k)}_\kappa), \quad k = 0, 1, \dots, \\ &\text{with } g^{(0)}_\kappa = \overline{g}^{(0)}_\kappa = g^*_\kappa.\end{aligned}$$

Then

$$\begin{aligned}\|g^{(1)}_\kappa - \overline{g}^{(1)}_\kappa\|_{-h,\widehat{\Gamma}} &= \|\Phi_{h\kappa}(\alpha_h, g^*_\kappa) - \Phi_{h\kappa}(\overline{\alpha}_h, g^*_\kappa)\|_{-h,\widehat{\Gamma}} \\ &\le c\|\alpha_h - \overline{\alpha}_h\|_{h,\widehat{\Gamma}},\end{aligned} \tag{2.26}$$

where $c > 0$ is a constant which does not depend on the choice of g^*_κ, as follows from Proposition 2.10. Similarly,

$$\begin{aligned}\|g^{(2)}_\kappa - \overline{g}^{(2)}_\kappa\|_{-h,\widehat{\Gamma}} &= \|\Phi_{h\kappa}(\alpha_h, g^{(1)}_\kappa) - \Phi_{h\kappa}(\overline{\alpha}_h, \overline{g}^{(1)}_\kappa)\|_{-h,\widehat{\Gamma}} \\ &\le \|\Phi_{h\kappa}(\alpha_h, g^{(1)}_\kappa) - \Phi_{h\kappa}(\overline{\alpha}_h, g^{(1)}_\kappa)\|_{-h,\widehat{\Gamma}} \\ &\quad + \|\Phi_{h\kappa}(\overline{\alpha}_h, g^{(1)}_\kappa) - \Phi_{h\kappa}(\overline{\alpha}_h, \overline{g}^{(1)}_\kappa)\|_{-h,\widehat{\Gamma}} \\ &\le c\|\alpha_h - \overline{\alpha}_h\|_{h,\widehat{\Gamma}} + q\|g^{(1)}_\kappa - \overline{g}^{(1)}_\kappa\| \\ &\le (c + cq)\|\alpha_h - \overline{\alpha}_h\|_{h,\widehat{\Gamma}},\end{aligned}$$

where $q \in (0,1)$, as follows from Proposition 2.8 and (2.26). By induction we obtain that

$$\|g_\kappa^{(k+1)} - \overline{g}_\kappa^{(k+1)}\|_{-h,\widehat{\Gamma}} \le (c + cq + \cdots + cq^k)\|\alpha_h - \overline{\alpha}_h\|_{h,\widehat{\Gamma}} \tag{2.27}$$

holds for any k. Passing to the limit with $k \to \infty$, we arrive at

$$\|\lambda_\kappa(\alpha_h) - \lambda_\kappa(\overline{\alpha}_h)\|_{-h,\widehat{\Gamma}} \le \frac{c}{1-q}\|\alpha_h - \overline{\alpha}_h\|_{h,\widehat{\Gamma}}. \qquad \square$$

COROLLARY 2.12. *From Proposition* 2.11 *and Remark* 2.2 *it follows that the mapping* $u_h : U_{ad}^h \to \boldsymbol{V}_h(\alpha_h)$ *associating with any* $\alpha_h \in U_{ad}^h$ *the displacement field* u_h *corresponding to the fixed point* $\lambda_\kappa(\alpha_h)$ *is also Lipschitzian in* U_{ad}^h. *Thus the control-state mapping* $\mathcal{S} : U_{ad}^h \to \boldsymbol{V}_h(\alpha_h) \times \Lambda_\kappa^{r_0}$, $\mathcal{S}(\alpha_h) = (u_h(\alpha_h), \lambda_\kappa(\alpha_h))$ *is Lipschitzian in* U_{ad}^h.

In the rest of this paper we restrict ourselves to the case in which $\mathcal{S}$ is Lipschitzian in U_{ad}^h. Let $J : U_{ad}^h \times \boldsymbol{V}_h(\alpha_h) \times \Lambda_\kappa \to \mathbb{R}^1$ be a cost functional, and define the following optimal shape design problem:

$$(\mathcal{P}) \qquad \left.\begin{array}{l} \text{Find } \alpha_h^* \in U_{ad}^h \text{ such that} \\ J(\alpha_h^*, \mathcal{S}(\alpha_h^*)) \le J(\alpha_h, \mathcal{S}(\alpha_h)) \quad \forall \alpha_h \in U_{ad}^h. \end{array}\right\}$$

If J is lower semicontinuous in $U_{ad}^h \times \boldsymbol{V}_h(\alpha_h) \times \Lambda_\kappa$, then $(\mathcal{P})$ has a solution.

3. Sensitivity analysis. In the previous section we analyzed the mappings λ_κ and u_h and found conditions under which these mappings are single-valued and Lipschitzian in U_{ad}^h. From now on, we shall concentrate solely on the algebraic formulation of contact problems with Coulomb friction. In particular, we will pay attention to the computation of the Clarke generalized Jacobians of these maps at a reference control, say $\bar{\ } \in \mathcal{U}$. For the sake of simplicity, we will work with the "reduced" displacement field $\boldsymbol{u} = (\boldsymbol{u}_\tau, \boldsymbol{u}_\nu)$. Recall that $\boldsymbol{u}_\tau$, $\boldsymbol{u}_\nu$, respectively, stand for subvectors of $\boldsymbol{u}$ that correspond to the tangential and normal displacements at the contact nodes. Let $\boldsymbol{A}(\ )$ be the stiffness matrix and $\boldsymbol{\ell}(\ )$ the right-hand side vector introduced in the previous section. We introduce the partitioning (to simplify the notation, we skip the argument alpha)

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{ii} & \boldsymbol{A}_{ic} \\ \boldsymbol{A}_{ci} & \boldsymbol{A}_{cc} \end{bmatrix}, \qquad \boldsymbol{\ell} = \begin{bmatrix} \boldsymbol{\ell}_i \\ \boldsymbol{\ell}_c \end{bmatrix},$$

corresponding to the nodes on the contact boundary (subscript c) and the others (subscript i). We further introduce the restriction of $\boldsymbol{A}$ and $\boldsymbol{\ell}$ on the contact boundary by elimination of the noncontact nodes:

$$\boldsymbol{A}_{cont} = \boldsymbol{A}_{cc} - \boldsymbol{A}_{ci}\boldsymbol{A}_{ii}^{-1}\boldsymbol{A}_{ic}, \qquad \boldsymbol{\ell}_{cont} = \boldsymbol{\ell}_c - \boldsymbol{A}_{ci}\boldsymbol{A}_{ii}^{-1}\boldsymbol{\ell}_i.$$

Finally, we use the partitioning of $\boldsymbol{A}_{cont}$ and $\boldsymbol{\ell}_{cont}$ to the tangential and normal components, corresponding to vectors $\boldsymbol{u}_\tau$ and $\boldsymbol{u}_\nu$:

$$\boldsymbol{A}_{cont} = \begin{bmatrix} \boldsymbol{A}_{\tau\tau} & \boldsymbol{A}_{\tau\nu} \\ \boldsymbol{A}_{\nu\tau} & \boldsymbol{A}_{\nu\nu} \end{bmatrix}, \qquad \boldsymbol{\ell}_{cont} = \begin{bmatrix} \boldsymbol{\ell}_\tau \\ \boldsymbol{\ell}_\nu \end{bmatrix}.$$

We choose Λ_κ of the type I so that $\breve{\ }$ belongs to the same space as the vectors $\boldsymbol{u}_\tau$, $\boldsymbol{u}_\nu$, namely $\mathbb{R}^p$, and $\boldsymbol{B}$ is the unit matrix; then $\in \mathbb{R}^p$. Let $\mathcal{S}[\mathbb{R}^p \to \mathbb{R}^{3p}]$ denote the map which associates with any $\in \mathcal{U}$ the triple $(\boldsymbol{u}_\tau, \boldsymbol{u}_\nu, \breve{\ })$, the solution of the contact problem with Coulomb friction. Evidently, by virtue of Proposition 2.11 and Corollary 2.12, $\mathcal{S}$ is Lipschitzian in $\mathcal{U}$.

The next proposition shows that the contact problem with Coulomb friction can be written in a compact form of a *generalized equation*. This will be used for the stability and sensitivity investigations in the rest of this section.

PROPOSITION 3.1. *Let $\in \mathcal{U}$ be given. The triple $(\boldsymbol{u}_\tau, \boldsymbol{u}_\nu, \breve{\ }) \in \mathbb{R}^{2p} \times \mathbb{R}^p_+$ is (a part of) the solution of the discretized contact problem with Coulomb friction (in the sense of Definition* 2.6*) if and only if it solves the following generalized equation:*

$$\left.\begin{aligned} &\mathbf{0} \in \boldsymbol{A}_{\tau\tau}(\ \ )\boldsymbol{u}_\tau + \boldsymbol{A}_{\tau\nu}(\ \ )\boldsymbol{u}_\nu - \boldsymbol{\ell}_\tau(\ \ ) + \boldsymbol{Q}(\boldsymbol{u}_\tau, \breve{\ }),\\ &\mathbf{0} = \boldsymbol{A}_{\nu\tau}(\ \ )\boldsymbol{u}_\tau + \boldsymbol{A}_{\nu\nu}(\ \ )\boldsymbol{u}_\nu - \boldsymbol{\ell}_\nu(\ \ ) - \breve{\ },\\ &\mathbf{0} \in \boldsymbol{u}_\nu + \ \ + N_{\mathbb{R}^p_+}(\breve{\ }), \end{aligned}\right\} \tag{3.1}$$

where

$$\boldsymbol{Q}(\boldsymbol{u}_\tau, \breve{\ }) = \partial_{\boldsymbol{u}_\tau} j(\boldsymbol{u}_\tau, \breve{\ }), \qquad j(\boldsymbol{u}_\tau, \breve{\ }) = \mathcal{F}\sum_{i=1}^{p} \lambda^i |u_\tau^i|,$$

and $N_{\mathbb{R}^p_+}(\cdot)$ is the standard normal-cone mapping in the sense of convex analysis.

Proof. Observe that the algebraic form of the contact problem with given friction $\boldsymbol{g}$ can be written as the following generalized system of equations:

$$\begin{aligned} &\mathbf{0} \in \boldsymbol{A}_{\tau\tau}(\ \ )\boldsymbol{u}_\tau + \boldsymbol{A}_{\tau\nu}(\ \ )\boldsymbol{u}_\nu - \boldsymbol{\ell}_\tau(\ \ ) + \boldsymbol{Q}(\boldsymbol{u}_\tau, \boldsymbol{g}),\\ &\mathbf{0} = \boldsymbol{A}_{\nu\tau}(\ \ )\boldsymbol{u}_\tau + \boldsymbol{A}_{\nu\nu}(\ \ )\boldsymbol{u}_\nu - \boldsymbol{\ell}_\nu(\ \ ) - \breve{\ },\\ &\mathbf{0} \in \boldsymbol{u}_\nu + \ \ + N_{\mathbb{R}^p_+}(\breve{\ }). \end{aligned}$$

By substituting $\boldsymbol{g} = \breve{\ }$, we immediately get the assertion of the proposition. ∎

Clearly, the multivalued part $\boldsymbol{Q}$ of the first line in (3.1) equals the Cartesian product

$$\bigtimes_{i=1}^{p} \mathcal{F}\lambda^i \partial|u_\tau^i|,$$

and for a fixed $i_0 \in \{1, 2, \ldots, p\}$ the respective set attains the form

$$\mathcal{F}\lambda^{i_0}\partial|u_\tau^{i_0}| = \begin{cases} \mathcal{F}\lambda^{i_0} & \text{if } u_\tau^{i_0} > 0,\\ -\mathcal{F}\lambda^{i_0} & \text{if } u_\tau^{i_0} < 0,\\ [-\mathcal{F}\lambda^{i_0}, \mathcal{F}\lambda^{i_0}] & \text{if } u_\tau^{i_0} = 0. \end{cases}$$

Let $\boldsymbol{A}_\tau(\ \ ) := [\boldsymbol{A}_{\tau\tau}(\ \ ), \boldsymbol{A}_{\tau\nu}(\ \ )]$, and let $(\bar{\ }, \overline{\boldsymbol{u}}_\tau, \overline{\boldsymbol{u}}_\nu, \bar{\breve{\ }})$ be a reference point satisfying the generalized equation (3.1). With this point we associate the following index sets which play a crucial role in the further development:

$$
\begin{aligned}
K_+(\overline{\ },\overline{\boldsymbol{u}}_\tau,\overline{\boldsymbol{u}}_\nu,\overline{\ }) &:= \{i\in\{1,2,\dots,p\} \mid \overline{u}^i_\tau > 0\},\\
K_-(\overline{\ },\overline{\boldsymbol{u}}_\tau,\overline{\boldsymbol{u}}_\nu,\overline{\ }) &:= \{i\in\{1,2,\dots,p\} \mid \overline{u}^i_\tau < 0\},\\
K_{0+}(\overline{\ },\overline{\boldsymbol{u}}_\tau,\overline{\boldsymbol{u}}_\nu,\overline{\ }) &:= \{i\in\{1,2,\dots,p\} \mid \overline{u}^i_\tau = 0,\ (-\boldsymbol{A}_\tau(\overline{\ })\overline{\boldsymbol{u}}+\boldsymbol{\ell}_\tau(\overline{\ }))^i = \mathcal{F}\overline{\lambda}^i\},\\
K_{0-}(\overline{\ },\overline{\boldsymbol{u}}_\tau,\overline{\boldsymbol{u}}_\nu,\overline{\ }) &:= \{i\in\{1,2,\dots,p\} \mid \overline{u}^i_\tau = 0,\ (-\boldsymbol{A}_\tau(\overline{\ })\overline{\boldsymbol{u}}+\boldsymbol{\ell}_\tau(\overline{\ }))^i = -\mathcal{F}\overline{\lambda}^i\},\\
K_{00}(\overline{\ },\overline{\boldsymbol{u}}_\tau,\overline{\boldsymbol{u}}_\nu,\overline{\ }) &:= \{i\in\{1,2,\dots,p\} \mid \overline{u}^i_\tau = 0,\ -\mathcal{F}\overline{\lambda}^i < (-\boldsymbol{A}_\tau(\overline{\ })\overline{\boldsymbol{u}}+\boldsymbol{\ell}_\tau(\overline{\ }))^i < \mathcal{F}\overline{\lambda}^i\},\\
M(\overline{\ },\overline{\boldsymbol{u}}_\tau,\overline{\boldsymbol{u}}_\nu,\overline{\ }) &:= \{i\in\{1,2,\dots,p\} \mid \overline{u}^i_\nu + \overline{\alpha}^i > 0\},\\
I_+(\overline{\ },\overline{\boldsymbol{u}}_\tau,\overline{\boldsymbol{u}}_\nu,\overline{\ }) &:= \{i\in\{1,2,\dots,p\} \mid \overline{\lambda}^i > 0\},\\
I_0(\overline{\ },\overline{\boldsymbol{u}}_\tau,\overline{\boldsymbol{u}}_\nu,\overline{\ }) &:= \{i\in\{1,2,\dots,p\} \mid \overline{u}^i_\nu + \overline{\alpha}^i = 0,\ \overline{\lambda}^i = 0\}.
\end{aligned}
$$

Remark 3.1. We will often use the above index sets as vector and matrix subscripts and skip the arguments.

From the results of the preceding section it follows that, if $\ $ is sufficiently close to $\overline{\ }$, one has for $(\boldsymbol{u}_\tau,\boldsymbol{u}_\nu,\check{\ }) = \mathcal{S}(\ )$ that

$$
\begin{gathered}
K_+(\ ,\boldsymbol{u}_\tau,\boldsymbol{u}_\nu,\check{\ }) \supset K_+(\overline{\ },\overline{\boldsymbol{u}}_\tau,\overline{\boldsymbol{u}}_\nu,\overline{\ }), \qquad K_-(\ ,\boldsymbol{u}_\tau,\boldsymbol{u}_\nu,\check{\ }) \supset K_-(\overline{\ },\overline{\boldsymbol{u}}_\tau,\overline{\boldsymbol{u}}_\nu,\overline{\ }),\\
K_{00}(\ ,\boldsymbol{u}_\tau,\boldsymbol{u}_\nu,\check{\ }) \supset K_{00}(\overline{\ },\overline{\boldsymbol{u}}_\tau,\overline{\boldsymbol{u}}_\nu,\overline{\ }),\\
M(\ ,\boldsymbol{u}_\tau,\boldsymbol{u}_\nu,\check{\ }) \supset M(\overline{\ },\overline{\boldsymbol{u}}_\tau,\overline{\boldsymbol{u}}_\nu,\overline{\ }), \qquad I_+(\ ,\boldsymbol{u}_\tau,\boldsymbol{u}_\nu,\check{\ }) \supset I_+(\overline{\ },\overline{\boldsymbol{u}}_\tau,\overline{\boldsymbol{u}}_\nu,\overline{\ }).
\end{gathered}
$$

Therefore

$$
\begin{gathered}
K_{0+}(\ ,\boldsymbol{u}_\tau,\boldsymbol{u}_\nu,\check{\ }) \subset K_{0+}(\overline{\ },\overline{\boldsymbol{u}}_\tau,\overline{\boldsymbol{u}}_\nu,\overline{\ }), \qquad K_{0-}(\ ,\boldsymbol{u}_\tau,\boldsymbol{u}_\nu,\check{\ }) \subset K_{0-}(\overline{\ },\overline{\boldsymbol{u}}_\tau,\overline{\boldsymbol{u}}_\nu,\overline{\ }),\\
I_0(\ ,\boldsymbol{u}_\tau,\boldsymbol{u}_\nu,\check{\ }) \subset I_0(\overline{\ },\overline{\boldsymbol{u}}_\tau,\overline{\boldsymbol{u}}_\nu,\overline{\ }).
\end{gathered}
$$

These inclusions imply that there is a neighborhood of $(\overline{\ },\overline{\boldsymbol{u}}_\tau,\overline{\boldsymbol{u}}_\nu,\overline{\ })$ where for each $i\in K_{0+}(\overline{\ },\overline{\boldsymbol{u}}_\tau,\overline{\boldsymbol{u}}_\nu,\overline{\ })$ one of the following two possibilities occurs:

$$
\left.
\begin{aligned}
u^i_\tau \ge 0, &\qquad (-\boldsymbol{A}_\tau(\ )\boldsymbol{u}+\boldsymbol{\ell}_\tau(\ ))^i = \mathcal{F}\lambda^i,\\
u^i_\tau = 0, &\qquad (-\boldsymbol{A}_\tau(\ )\boldsymbol{u}+\boldsymbol{\ell}_\tau(\ ))^i \le \mathcal{F}\lambda^i.
\end{aligned}
\right\} \tag{3.2}
$$

Further, for each $i\in K_{0-}(\overline{\ },\overline{\boldsymbol{u}}_\tau,\overline{\boldsymbol{u}}_\nu,\overline{\ })$ one of the following two possibilities occurs:

$$
\left.
\begin{aligned}
u^i_\tau \le 0, &\qquad (-\boldsymbol{A}_\tau(\ )\boldsymbol{u}+\boldsymbol{\ell}_\tau(\ ))^i = -\mathcal{F}\lambda^i,\\
u^i_\tau = 0, &\qquad (-\boldsymbol{A}_\tau(\ )\boldsymbol{u}+\boldsymbol{\ell}_\tau(\ ))^i \ge -\mathcal{F}\lambda^i.
\end{aligned}
\right\} \tag{3.3}
$$

Analogously, there exists a neighborhood of $(\overline{\ },\overline{\boldsymbol{u}}_\tau,\overline{\boldsymbol{u}}_\nu,\overline{\ })$ where for each $i\in I_0(\overline{\ },\overline{\boldsymbol{u}}_\tau,\overline{\boldsymbol{u}}_\nu,\overline{\ })$ one of the following two possibilities occurs:

$$
\left.
\begin{aligned}
\lambda^i \ge 0, &\qquad u^i_\nu + \alpha^i = 0,\\
\lambda^i = 0, &\qquad u^i_\nu + \alpha^i \ge 0.
\end{aligned}
\right\} \tag{3.4}
$$

In this way, we get a decomposition of $K_{0+}(\overline{\ },\overline{\boldsymbol{u}}_\tau,\overline{\boldsymbol{u}}_\nu,\overline{\ })$ into two subsets, say K_1, K_2, defined by relations (3.2); a decomposition of $K_{0-}(\overline{\ },\overline{\boldsymbol{u}}_\tau,\overline{\boldsymbol{u}}_\nu,\overline{\ })$ into two subsets, say K_3, K_4, defined by relations (3.3); and a decomposition of $I_0(\overline{\ },\overline{\boldsymbol{u}}_\tau,\overline{\boldsymbol{u}}_\nu,\overline{\ })$ into two subsets, say J_1, J_2, defined by relations (3.4). In the further development, we will especially make use of the equality constraints appearing in these decompositions. More precisely, we will neglect all inequalities in (3.2)–(3.4) and obtain from (3.1)

the equation system (3.5) below, which is linear in variables $\boldsymbol{u}_\tau$, $\boldsymbol{u}_\nu$, and $\breve{\ }$ and nonlinear in the design variable . Furthermore, to simplify the notation, we set $\beta := K_+ \cup K_1 \cup K_- \cup K_3$, $\gamma := K_{00} \cup K_2 \cup K_4$ and denote by $\boldsymbol{D}$ a $p \times p$ diagonal matrix given by

$$d_{ii} = \begin{cases} \mathcal{F} & \text{for } i \in K_+ \cup K_1, \\ -\mathcal{F} & \text{for } i \in K_- \cup K_3, \\ 0 & \text{otherwise.} \end{cases}$$

The announced system attains the form

$$\left.\begin{aligned} \mathbf{0} &= (\boldsymbol{A}_{\tau\tau}(\ \ ))_\beta \boldsymbol{u}_\tau + (\boldsymbol{A}_{\tau\nu}(\ \ ))_\beta \boldsymbol{u}_\nu - (\boldsymbol{\ell}_\tau(\ \ ))_\beta + \boldsymbol{D}_\beta \breve{\ }, \\ \mathbf{0} &= \boldsymbol{A}_{\nu\tau}(\ \ )\boldsymbol{u}_\tau + \boldsymbol{A}_{\nu\nu}(\ \ )\boldsymbol{u}_\nu - \boldsymbol{\ell}_\nu(\ \ ) - \breve{\ }, \\ 0 &= u_\tau^i \quad \text{for } i \in \gamma, \\ 0 &= u_\nu^i + \alpha^i \quad \text{for } i \in I_+ \cup J_1, \\ 0 &= \lambda^i \quad \text{for } i \in M \cup J_2. \end{aligned}\right\} \tag{3.5}$$

To (3.5) one could apply the classic implicit function theorem, provided that the regularity assumption is fulfilled. This is clarified in the next statement.

PROPOSITION 3.2. *Consider system* (3.5) *around the reference point* $(\bar{\ }, \overline{\boldsymbol{u}}_\tau, \overline{\boldsymbol{u}}_\nu, \breve{\ })$. *Assume that the matrix*

$$\begin{bmatrix} \boldsymbol{A}_{\tau\tau}(\bar{\ }) & \boldsymbol{A}_{\tau\nu}(\bar{\ }) \\ \boldsymbol{A}_{\nu\tau}(\bar{\ }) & \boldsymbol{A}_{\nu\nu}(\bar{\ }) \end{bmatrix}$$

is positive definite. Then for each choice of the sets K_1, K_2, K_3, K_4, J_1, J_2 *there exist a neighborhood* $\mathcal{O}$ *of* $\bar{\ }$ *and a continuously differentiable operator* $\mathcal{S}^*[\mathcal{O} \to \mathbb{R}^{3p}]$, *defined implicitly by* (3.5), *such that* $(\overline{\boldsymbol{u}}_\tau, \overline{\boldsymbol{u}}_\nu, \breve{\ }) = \mathcal{S}^*(\bar{\ })$ *and the points* $(\ \ , \mathcal{S}^*(\ \ ))$ *satisfy* (3.5) *for all* $\in \mathcal{O}$, *whenever the friction coefficient* $\mathcal{F}$ *is sufficiently small.*

Proof. Let the index sets K_1, K_2, K_3, K_4, J_1, J_2 be chosen arbitrarily. We first show that $(\overline{\boldsymbol{u}}_\tau, \overline{\boldsymbol{u}}_\nu, \breve{\ }) \in \mathcal{S}^*(\bar{\ })$. It follows directly from (3.1) that $(\overline{\boldsymbol{u}}_\tau, \overline{\boldsymbol{u}}_\nu, \breve{\ })$ fulfills all equations of (3.5) except the first one. To see that the first one also holds true, consider a vector $\boldsymbol{\emptyset} \in \boldsymbol{Q}(\overline{\boldsymbol{u}}_\tau, \breve{\ })$. Then one has

$$\chi^i = d_{ii}\overline{\lambda}^i \quad \text{for } i \in \beta \qquad \text{(no sum)},$$

or, in other words,

$$\boldsymbol{\emptyset}_\beta = \boldsymbol{D}_\beta \breve{\ }.$$

Therefore $(\overline{\boldsymbol{u}}_\tau, \overline{\boldsymbol{u}}_\nu, \breve{\ })$ fulfills the whole system (3.5).

It remains to show that $\mathcal{S}^*$ is in fact a differentiable single-valued map on a neighborhood of $\bar{\ }$. This will follow from the classic implicit function theorem, provided that we succeed in verifying the respective regularity assumption.

Ignoring those components of $(\boldsymbol{u}_\tau, \boldsymbol{u}_\nu, \breve{\ })$ that are equal to zero, the reduced partial Jacobian of the right-hand side of (3.5) at $(\bar{\ }, \overline{\boldsymbol{u}}_\tau, \overline{\boldsymbol{u}}_\nu, \breve{\ })$ with respect to the remaining components of $(\boldsymbol{u}_\tau, \boldsymbol{u}_\nu, \breve{\ })$ is

$$\boldsymbol{\Pi} = \begin{bmatrix} (\boldsymbol{A}_{\tau\tau}(\bar{\ }))_{\beta,\beta} & (\boldsymbol{A}_{\tau\nu}(\bar{\ }))_\beta & \boldsymbol{D}_{\beta,(I_+\cup J_1)} \\ {}_\beta(\boldsymbol{A}_{\nu\tau}(\bar{\ })) & \boldsymbol{A}_{\nu\nu}(\bar{\ }) & -\boldsymbol{E}^T_{I_+\cup J_1} \\ \mathbf{0} & \boldsymbol{E}_{I_+\cup J_1} & \mathbf{0} \end{bmatrix}. \tag{3.6}$$

If $\boldsymbol{D}_{\beta,(I_+\cup J_1)}$ were a zero matrix, the matrix (3.6) would be nonsingular by the assumptions imposed. Hence, $\boldsymbol{\Pi}$ is a sum of a nonsingular matrix and the matrix

$$\overline{\boldsymbol{\Pi}} = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{D}_{\beta,(I_+\cup J_1)} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}.$$

The only nonzero entries of $\overline{\boldsymbol{\Pi}}$ are equal to $\pm\mathcal{F}$, which can be chosen arbitrarily small. Thus we can use the well-known perturbation lemma [17] and conclude that $\boldsymbol{\Pi}$ is indeed nonsingular. Since our choice of K_1, K_2, K_3, K_4, J_1, J_2 was arbitrary, the statement follows directly from the classic implicit function theorem. □

From the above statement we infer that $\mathcal{S}$ is a PC^1-map whose active pieces at the reference control $\bar{\ }$ are given by combinations of index sets K_1, K_2, K_3, K_4, J_1, J_2. Of course, not each of these active pieces is essentially active at $\bar{\ }$ (besides the trivial situation of $K_{0+} = K_{0-} = I_0 = \emptyset$). To test the essential activeness of a piece, one could verify the existence of a sequence $\ _i \to \bar{\ }$ such that the images $\mathcal{S}^*(\ _i)$ fulfill the respective inequalities in (3.2)–(3.4) with strict inequality signs. This can sometimes, but not always (see [18, Example 7.3]), be performed by tools of first-order analysis. Since the search for an essentially active piece can be really time-consuming, in our computations we will use an upper approximation of $\partial\mathcal{S}(\bar{\ })$ provided by the next statement. To furnish it, we introduce a new index set $\mathbb{L}$ in such a way that there is a one-to-one correspondence between the indices of $\mathbb{L}$ and the possible combinations of the index sets K_1, K_2, K_3, K_4, J_1, J_2. By $\boldsymbol{\Pi}_i$, $i \in \mathbb{L}$, we then denote that matrix of the type (3.6) determined by the index sets K_1, K_2, K_3, K_4, J_1, J_2 associated with i. Furthermore, with $i \in \mathbb{L}$ we associate also the (reduced) partial Jacobians $\boldsymbol{\Xi}_i$ of the right-hand side of (3.5) at $(\bar{\ }, \overline{\boldsymbol{u}}_\tau, \overline{\boldsymbol{u}}_\nu, \breve{\ })$ with respect to the (independent) variable . These matrices attain the form

$$\boldsymbol{\Xi}_i = \begin{bmatrix} \nabla\ \ ((\boldsymbol{A}_{\tau\tau}(\bar{\ }))_{\beta,\beta}(\overline{\boldsymbol{u}}_\tau)_\beta) + \nabla\ \ ((\boldsymbol{A}_{\tau\nu}(\bar{\ }))_\beta \overline{\boldsymbol{u}}_\nu) - \nabla(\boldsymbol{\ell}_\tau(\bar{\ }))_\beta \\ \nabla\ \ (_\beta(\boldsymbol{A}_{\nu\tau}(\bar{\ }))(\overline{\boldsymbol{u}}_\tau)_\beta) + \nabla\ \ (\boldsymbol{A}_{\nu\nu}(\bar{\ })\overline{\boldsymbol{u}}_\nu) - \nabla\boldsymbol{\ell}_\nu(\bar{\ }) \\ \boldsymbol{E}_{I_+\cup J_1} \end{bmatrix}. \tag{3.7}$$

Clearly, under the condition of Proposition 3.2, for each $i \in \mathbb{L}$ the matrix $-\boldsymbol{\Pi}_i^{-1}\boldsymbol{\Xi}_i$ is equal to the derivative of the map

$$\mapsto ((\boldsymbol{u}_\tau)_\beta, \boldsymbol{u}_\nu, \breve{\ }_{I_+\cup J_1})$$

defined by (3.5) and computed at the reference control $\bar{\ }$. The derivative of the complete map $\mapsto (\boldsymbol{u}_\tau, \boldsymbol{u}_\nu, \breve{\ })$ at $\bar{\ }$, corresponding to $i \in \mathbb{L}$, can thus be obtained from $-\boldsymbol{\Pi}_i^{-1}\boldsymbol{\Xi}_i$ by inserting zero rows for the derivatives of u_τ^j, $j \in \gamma$, and λ^l, $l \in M \cup J_2$. Let us denote such "completed" $[3p \times p]$ matrices by $\boldsymbol{R}_i$, $i \in \mathbb{L}$.

PROPOSITION 3.3. *Under the assumptions of Proposition* 3.2, *one has*

$$\partial\mathcal{S}(\bar{\ }) \subset \operatorname{conv}\{\boldsymbol{R}_i \mid i \in \mathbb{L}\}. \tag{3.8}$$

Proof. The proof follows readily from the above analysis and [23, Proposition A.4.1]. □

In the real computations, however, it is not necessary to evaluate matrices from $\partial\mathcal{S}(\bar{\ })$ explicitly. Our aim is to compute just one subgradient of the composite map $\Theta(\ ) := \mathcal{J}(\ , \mathcal{S}(\ ))$, where $\mathcal{J}[\mathbb{R}^p \times \mathbb{R}^{3p} \to \mathbb{R}]$ is the (discretized) objective in our shape optimization problem.

THEOREM 3.4. *Assume that $\mathcal{J}$ is continuously differentiable and all assumptions from Proposition* 2.11 *and Corollary* 2.12 *are fulfilled. Let $\bar{\ } \in \mathcal{U}$ be given, and suppose that for $i \in \mathbb{L}$*

$$\boldsymbol{R}_i \in \partial \mathcal{S}(\bar{\ }).$$

Finally, let $\boldsymbol{p}_i$ be the (unique) solution of the adjoint equation

$$\boldsymbol{\Pi}_i^T \boldsymbol{p}_i + (\nabla_2 \mathcal{J}(\bar{\ }, \mathcal{S}(\bar{\ })))_i = \boldsymbol{0}, \tag{3.9}$$

where $\boldsymbol{\Pi}_i$ denotes the matrix (3.6) *for which the index sets K_1, K_2, K_3, K_4, J_1, J_2 are specified by i, and $(\nabla_2 \mathcal{J}(\bar{\ }, \mathcal{S}(\bar{\ })))_i$ denotes the subvector of $\nabla_2 \mathcal{J}(\bar{\ }, \mathcal{S}(\bar{\ }))$ in which the components corresponding to the partial derivatives with respect to u_τ^j, $j \in \gamma$, and with respect to λ^j, $j \in M \cup J_2$, were omitted. Then one has*

$$\boldsymbol{\xi}_i = \nabla_1 \mathcal{J}(\bar{\ }, \mathcal{S}(\bar{\ })) + \boldsymbol{\Xi}_i^T \boldsymbol{p}_i \in \partial \Theta(\bar{\ }).$$

Proof. It suffices to apply [1, Theorem 2.3.10] in the same way as was done in [18]. □

The above analysis enables one to apply a bundle method of nonsmooth optimization to the numerical solution of the discretized shape optimization problem with equilibrium governed by (3.1). A detailed description of the resulting procedure together with a number of test examples is given in the next section.

4. Numerical results. The results of the previous sections will now be used for computation of numerical examples. We assume that the friction coefficient $\mathcal{F}$ is small enough so that the solution of the contact problem with Coulomb friction is unique. Then we can use the implicit programming approach [18] to solve the shape optimization problem $(\mathcal{P})$. Further, we assume that the cost functional $\mathcal{J}$ is locally continuously differentiable, so that, by Proposition 2.11 and Corollary 2.12, the composite map $\Theta(\cdot)$ is locally Lipschitzian. Moreover, since $\mathcal{S}$ is a PC^1-function, Θ is semismooth as a composition of two semismooth mappings [14]. This implies that for the minimization of Θ we have a choice between two classes of algorithms: derivative-free methods (like pattern-search methods, genetic algorithms, etc.) and methods that use (sub)gradient information (bundle, ellipsoid, cutting-plane, and other methods). Since the subgradient information is available in our case, we opted for the second class, and more specifically, for the most robust method, namely, the bundle algorithm. The particular code of choice was the BT code [25] based on the bundle-truss algorithm of Schramm and Zowe [24]. In every step of the iteration process, this code needs the function value $\Theta(\ \ )$ and one (arbitrary) Clarke subgradient of Θ at $\ \ $.

Assume that the cost functional $\mathcal{J}$ is continuously differentiable. Then we can apply Theorem 3.4 to compute a subgradient $\boldsymbol{\xi}$ from the generalized Jacobian $\partial\Theta(\ \ )$. We should specify only the decomposition of the sets K_{0+}, K_{0-}, and I_0. We have chosen

$$K_1 = K_3 = \emptyset, \quad K_2 = K_{0+}, \quad K_4 = K_{0-}, \quad J_1 = \emptyset, \quad J_2 = I_0.$$

To compute a function value $\mathcal{J}(\ \ , \mathcal{S}(\ \ ))$, we have to solve the state problem, i.e., the Signorini problem with Coulomb friction formulated as a fixed-point problem. For that, we use the splitting variant of the fixed-point method introduced in [6]. This is basically the method of successive approximations, where at each step we solve the contact problem with given friction. The iterative process then updates

the given friction g_κ. The problem with given friction is solved using the so-called *reciprocal variational formulation* (see [10, 12]), where the variables are the contact stresses $\breve{} = (\breve{}_\tau, \breve{}_\nu)$ on contact boundaries. This formulation leads to a quadratic programming problem with simple box constraints imposed on normal and tangential contact stresses, namely, to the minimization of a quadratic function $\mathcal{Q}$ on $\mathbb{R}^p_{[-1,1]} \times \mathbb{R}^p_-$. For the solution of this quadratic program, we use the so-called splitting technique, a version of the Gauss–Seidel algorithm: Instead of minimizing a function of $2p$ variables $(\breve{}_\tau, \breve{}_\nu)$, the process is split into two separate minimizations with respect to the normal $(\breve{}_\nu)$ and tangential $(\breve{}_\tau)$ contact stress, respectively, keeping the remaining stress component fixed. At each step we thus minimize just a function of p variables. The convergence of this algorithm for contact problems with given friction was shown in [4]. To solve the problem with Coulomb friction, we use an extended version of this algorithm, introduced in [6]. In a sense, we combine the successive approximation scheme with the splitting technique in one iterative algorithm.

For the solution of practical examples, we slightly modify the shape optimization problem $(\mathcal{P})$. The purpose of this modification is to work with a relatively small number of control variables and, at the same time, to get a smooth shape of the optimal boundary. Therefore, the contact boundary Γ is modelled by a Bezier curve of order d, and the design variable is a vector of its control points. The Bezier curve $F_\alpha(x)$ of order d in $[0, a]$ is generated by a vector as

$$F_\alpha(x) = \sum_{i=0}^{d} \alpha^i \beta_d^i(x), \qquad \beta_d^i(x) = \frac{1}{a^d} \binom{d}{i} x^i (a-x)^{d-i}, \quad x \in [0, a],$$

where d is the dimension of . The end points of a Bezier curve are identical with the first and last control point. The curve itself lies in the convex envelope of the control points. This means that any upper and lower bounds on the control points are automatically satisfied for the curve too.

The modified shape optimization problem is defined as follows:

$$(\mathcal{P}_\mathcal{B}) \qquad \left.\begin{array}{l} \text{minimize } \mathcal{J}(\ , \mathcal{S}(\ )) \\ \text{subject to } \ \in \mathcal{U}, \end{array}\right\}$$

where $\mathcal{U}$ is given by

$$\mathcal{U} = \left\{ \ \in \mathbb{R}^d \mid 0 \le \alpha^i \le C_0, \ i = 0, 1, \ldots, d; \right.$$
$$\left. |\alpha^{i+1} - \alpha^i| \le \frac{C_1}{d}, \ i = 0, 1, \ldots, d-1; \quad \sum_{i=0}^{d} \alpha^i = C_2(d+1) \right\}$$

and C_0, C_1, C_2 are given positive constants. The first set of constraints guarantees that $|F_\alpha(x)| \le C_0$ for all $x \in [0, a]$. The second constraint set takes care of the smoothness of the optimal shape. It is well known that if the control points satisfy this condition, then $|F'_\alpha(x)| \le C_1$ for all $x \in [0, a]$. The equality constraint is added to control the volume of the domain by the control points of the Bezier curve. The number $(ab - C_2)$ equals the area of $\Omega(\ )$. Thus the equality constraint (working with the control points) has a physical meaning of preserving the weight of the structure.

We will present the results of two examples solved by the proposed implicit programming technique combined with the BT code. In both examples we use the same

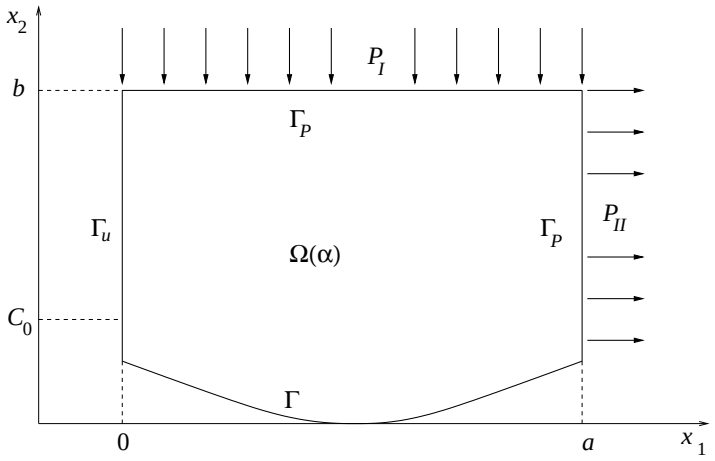


FIG. 4.1. *The elastic body and applied loads.*

data and change only the cost function $\mathcal{J}$. The shape of the elastic body $\Omega(\quad)$, $\quad \in \mathcal{U}$, is defined through a Bezier curve F_α as follows:

$$\Omega(\quad) = \{(x_1, x_2) \in \mathbb{R}^2 \mid x_1 \in (0, a),\ F_\alpha(x_1) < x_2 < b\};$$

see Figure 4.1. This figure also shows the distribution of external loads on the boundary Γ_P, given as $P_I = -80 \cdot 10^6 (\mathrm{N/m^2})$, $P_{II} = 50 \cdot 10^6 (\mathrm{N/m^2})$. Further, Γ_u is the part of the boundary with prescribed Dirichlet condition where both displacements are fixed to zero.

The set of admissible designs $\mathcal{U}$ is determined by the choice $a = 2$, $b = 1$, $C_0 = 0.75$, $C_1 = 0.5$, and $C_2 = 0.1$. The examples were solved with the Young modulus $E = 1$ GPa, the Poisson constant $\sigma = 0.3$, and the friction coefficient $\mathcal{F} = 0.25$. In both examples, we discretized $\Omega(\quad)$ by a regular 29×9 mesh; i.e., we have 281 nodes and 562 unknowns in the state problem. The dimension of the control vector , generating the Bezier curve and defining $\Omega(\quad)$, was set to 20.

Example 4.1. In the first example we try to smooth down peaks of the normal contact stress distribution. To cut the peaks, we should minimize the infinity norm. The objective function $\mathcal{J}$, however, must be continuously differentiable, so we will use the (p power of) p-norm with $p = 4$. (Higher values of p cause difficulties in the BT code, due to ill-conditioning.) The shape optimization problem then reads as follows:

$$\begin{aligned} &\text{minimize } \| \breve{} \ _\nu \|_4^4 \\ &\text{subject to} \qquad \in \mathcal{U}. \end{aligned}$$

In Figure 4.2 we present the initial shape and its deformation. Figure 4.3 shows the optimal shape and its deformation under given loads. Finally, Figure 4.4 compares the contact normal stresses for the initial (left) and optimal (right) shapes, respectively. The decrease in the peak stress is quite significant.

To see the importance of proper modelling of contact problems, let us compute the same example, now *without friction*. Figure 4.5 shows the optimal shape before and after deformation in this case. We can see that the optimal shape indeed differs from that computed with Coulomb friction. Further, in Figure 4.6 we compare contact stresses for the optimal design from Figure 4.5 computed by models without friction

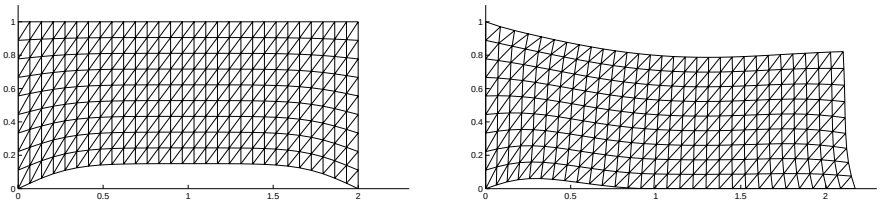


FIG. 4.2. *Example* 4.1, *initial design.*

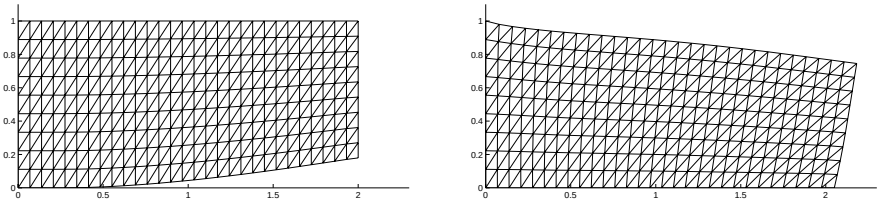


FIG. 4.3. *Example* 4.1, *optimal design.*

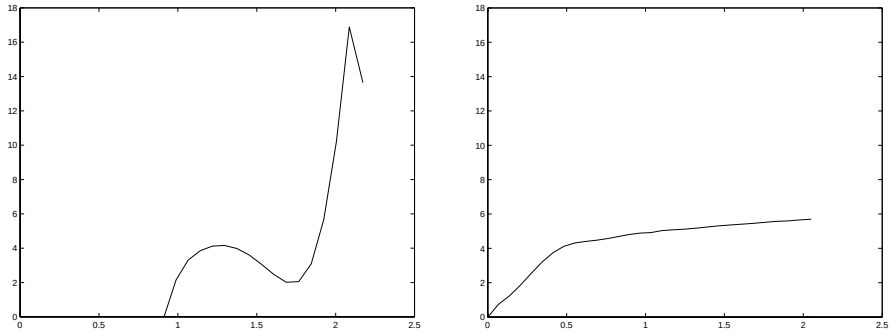


FIG. 4.4. *Example* 4.1, *normal stress for initial (left) and optimal (right) design.*

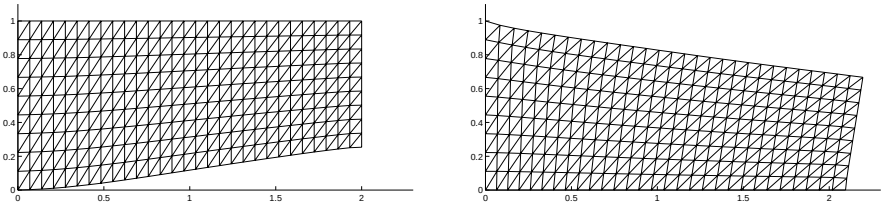


FIG. 4.5. *Example* 4.1, *optimal design for the problem without friction.*

(left-hand figure) and with Coulomb friction (right-hand figure). The peak stress in the left-hand figure is 5.7 (about the same as in Figure 4.4 (right)), while it is 8.5 in the right-hand figure. This shows that it does not make much sense to replace the Coulomb friction problem by a (much simpler) model without friction to get an "approximate optimal design," as is often done in the engineering praxis. This will be even more significant in the following example.

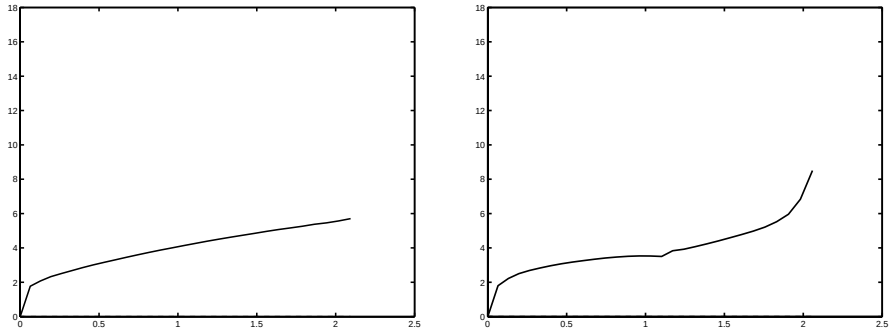


FIG. 4.6. *Example* 4.1, *stress distribution computed without (left) and with (right) friction for optimal design without friction.*

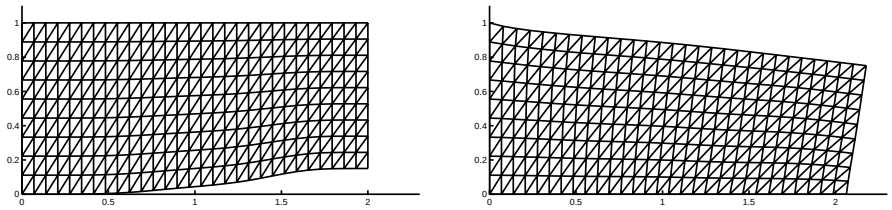


FIG. 4.7. *Example* 4.2, *initial design.*

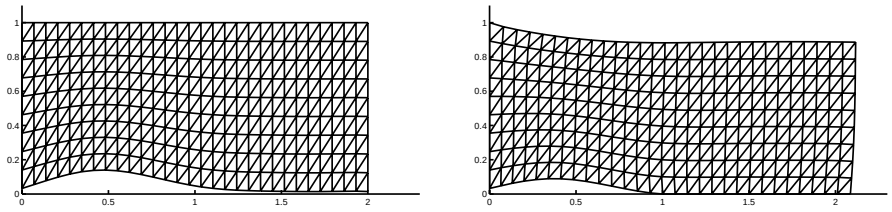


FIG. 4.8. *Example* 4.2, *optimal design.*

Example 4.2. Here we try to identify the contact normal stress with a prescribed value. The shape optimization can be written as

$$\begin{aligned} &\text{minimize } \|\overline{\sigma}_\nu - \sigma_\nu\|_2^2 \\ &\text{subject to} \qquad \in \mathcal{U}, \end{aligned}$$

where $\overline{\sigma}_\nu$ is a vector of prescribed normal stresses. This vector was chosen to model a step function, depicted in Figure 4.9 by the dashed line.

The initial design and its deformation under load are presented in Figure 4.7, while Figure 4.8 shows the optimal design before and after deformation. Finally, Figure 4.9 compares the contact normal stresses with the prescribed values. While the initial contact stresses are far from the prescribed values, the stresses for the optimal shape follow the step function very closely.

We again solved this example for the case without friction. Figure 4.10 presents the optimal shape before and after deformation. Again, the optimal shape differs significantly from that computed with Coulomb friction, just as the normal stresses

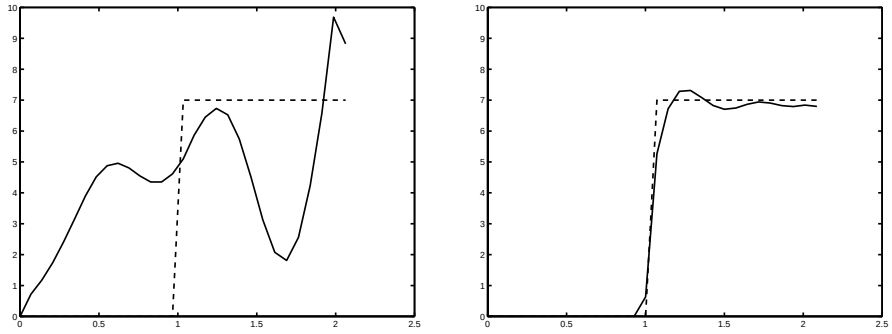


FIG. 4.9. *Example* 4.2, *normal stress for initial (left) and optimal (right) design.*

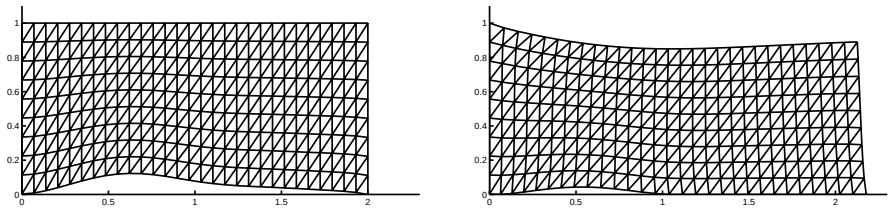


FIG. 4.10. *Example* 4.2, *optimal design for the problem without friction.*

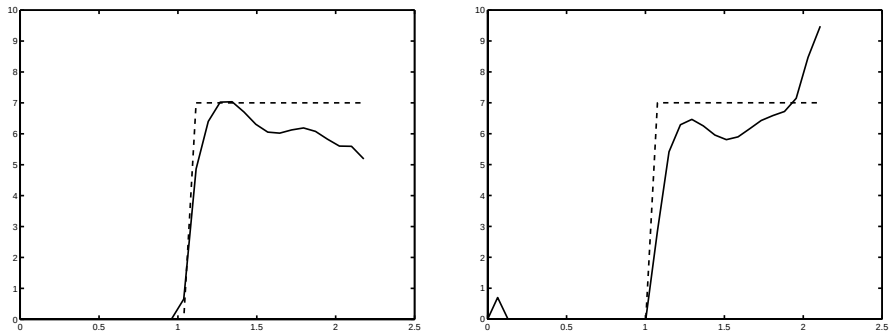


FIG. 4.11. *Example* 4.2, *stress distribution computed without (left) and with (right) friction for optimal design without friction.*

computed for the optimal design from Figure 4.10 but by a model with Coulomb friction (Figure 4.11 (right)).

As soon as $|\mathbb{L}| \geq 1$, an arbitrary choice of an index from $\mathbb{L}$ does not necessarily lead to a subgradient of the composite objective Θ. As thoroughly analyzed in [18] (in a slightly less general context), the verification of the correctness of the subgradient information is a time-consuming procedure that can hardly be performed during the iteration process. Therefore it is reasonable to ask whether for such an arbitrary $i \in \mathbb{L}$ the approximate solution, provided by the bundle algorithm, really approximates a Clarke stationary point of the solved problem. To answer this question, one can test whether the obtained solution satisfies first-order necessary optimality conditions of the Clarke or Mordukhovich type (see, e.g., [19]). Such conditions were not specialized

for the MPEC investigated in this paper, but, using the results of the preceding section, their derivation does not present any serious difficulties. A plausible alternative is to recompute the problem by a completely different approach, e.g., by a derivative-free algorithm; this way was used in the case of the results presented in this section. If no such testing is possible, we know from [2] that the applied approach leads to points that are stationary in a weaker sense (in comparison with Clarke or Mordukhovich). This holds for equilibria governed by standard variational inequalities, but we dare to conjecture that an analogous statement can be proved in our case as well. Indeed, in both cases we face the minimization of a PC^1-function, whose essentially active pieces cannot always be identified.

To conclude, from the user's point of view it is not so important to know what type of stationary point our procedure approximates but to get a relative decrease (improvement) of the objective with respect to the initial design. Our test examples can be viewed also in this way.

Acknowledgement. The authors would like to thank the anonymous referees for their constructive comments.

REFERENCES

[1] F. F. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley & Sons, New York, 1983.
[2] S. Dempe and J. F. Bard, *A bundle trust region algorithm for bilinear bilevel programming*, J. Optim. Theory Appl., (2001), 110 (2001), pp. 265–288.
[3] V. Girault and P. A. Raviart, *Finite Element Approximation of the Navier-Stokes Equations*, Lecture Notes in Math. 749, Springer, Berlin, 1979.
[4] R. Glowinski, *Numerical Methods for Nonlinear Variational Problems*, Springer-Verlag, New York, 1984.
[5] J. Haslinger, *Approximation of the Signorini problem with friction, obeying Coulomb law*, Math. Methods Appl. Sci., 5 (1983), pp. 422–437.
[6] J. Haslinger, Z. Dostál, and R. Kučera, *Signorini problem with a given friction based on the reciprocal variational formulation*, in Nonsmooth/Nonconvex Mechanics: Modeling, Analysis and Numerical Methods, D. Y. Gao, R. W. Ogden, and G. E. Stavroulakis, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2001, pp. 141–172.
[7] J. Haslinger, I. Hlaváček, and J. Nečas, *Numerical methods for unilateral problems in solid mechanics*, in Handbook of Numerical Analysis, Vol. IV, P. G. Ciarlet and J. L. Lions, eds., North–Holland, Amsterdam, 1996, pp. 313–485.
[8] J. Haslinger and I. Hlaváček, *Approximation of the Signorini problem with friction by a mixed finite element method*, J. Math. Anal. Appl., 86 (1982), pp. 99–122.
[9] J. Haslinger and P. Neittaanmäki, *Finite Element Approximation for Optimal Shape, Material and Topology Design*, 2nd ed., John Wiley & Sons, Chichester, U.K., 1996.
[10] 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.
[11] D. Hilding, A. Klarbring, and J. Petersson, *Optimization of structures in unilateral contact*, ASME Applied Mechanics Rev., 52 (1999), pp. 139–160.
[12] N. Kikuchi and J. T. Oden, *Contact Problems in Elasticity: A Study of Variational Inequalities and Finite Element Methods*, SIAM, Philadelphia, 1988.
[13] Z.-Q. Luo, J.-S. Pang, and D. Ralph, *Mathematical Programs with Equilibrium Constraints*, Cambridge University Press, Cambridge, U.K., 1996.
[14] R. Mifflin, *Semismooth and semiconvex functions in constrained optimization*, SIAM J. Control Optim., 15 (1977), pp. 959–972.
[15] J. Nečas, J. Jarušek, and J. Haslinger, *On the solution of the variational inequality to the Signorini-problem with small friction*, Boll. Unione Mat. Ital. Sez. B Artic. Ric. Mat., 17 (1980), pp. 796–811.
[16] J. Nečas and I. Hlaváček, *Mathematical Theory of Elastic and Elasto-Plastic Bodies: An Introduction*, Elsevier, Amsterdam, 1981.
[17] J. M. Ortega and W. C. Rheinboldt, *Iterative Solutions of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[18] J. Outrata, M. Kočvara, and J. Zowe, *Nonsmooth Approach to Optimization Problems with Equilibrium Constraints: Theory, Applications and Numerical Results*, Kluwer Academic Publishers, Dordrecht, Boston, London, 1998.

[19] J. V. Outrata, *A generalized mathematical program with equilibrium constraints*, SIAM J. Control Optim., 38 (2000), pp. 1623–1638.

[20] P. D. Panagiotopoulos, *Inequality problems in mechanics and applications, convex and nonconvex energy functions*, Birkhäuser-Verlag, Basel, 1985.

[21] L. Qi and J. Sun, *A nonsmooth version of Newton's method*, Math. Programming, 58 (1993), pp. 353–368.

[22] A. Quarteroni and A. Valli, *Numerical Approximation of Partial Differential Equations*, Springer, Berlin, 1994.

[23] S. Scholtes, *Introduction to Piecewise Differential Equations*, Habilitation Thesis, Institut für Statistik und Mathematische Wirtschaftstheorie, Universität Karlsruhe, Karlsruhe, Germany, 1994.

[24] H. Schramm and J. Zowe, *A version of the bundle idea for minimizing a nonsmooth function: Conceptual idea, convergence analysis, numerical results*, SIAM J. Optim., 2 (1992), pp. 121–152.

[25] J. Zowe, M. Kočvara, J. Outrata, and H. Schramm, *Bundle Trust Methods: Fortran Codes for Nonsmooth Optimization. User's Guide*, Preprint 259, Inst. Appl. Math., University of Erlangen, Erlangen, Germany, 2000.

SIAM J. OPTIM.
Vol. 13, No. 2, pp. 588–602

A NEWTON METHOD FOR SHAPE-PRESERVING SPLINE INTERPOLATION*

ASEN L. DONTCHEV†, HOU-DUO QI‡, LIQUN QI§, AND HONGXIA YIN¶

This work is dedicated to Professor Jochem Zowe

Abstract. In 1986, Irvine, Marin, and Smith proposed a Newton-type method for shape-preserving interpolation and, based on numerical experience, conjectured its quadratic convergence. In this paper, we prove local quadratic convergence of their method by viewing it as a semismooth Newton method. We also present a modification of the method which has global quadratic convergence. Numerical examples illustrate the results.

Key words. shape-preserving interpolation, splines, semismooth equation, Newton's method, quadratic convergence

AMS subject classifications. 41A29, 65D15, 49J52, 90C25

PII. S1052623401393128

1. Introduction.

Given nodes $a = t_1 < t_2 < \cdots < t_{N+2} = b$ and values $y_i = f(t_i), i = 1, \ldots, N+2, N \geq 3$, of an unknown function $f : [a, b] \to \mathbf{R}$, the standard interpolation problem consists of finding a function s from a given set S of interpolants such that $s(t_i) = y_i$, $i = 1, \ldots, N+2$. When S is the set of twice continuously differentiable piecewise cubic polynomials across t_i, we deal with cubic spline interpolation. The problem of cubic spline interpolation can be viewed in various ways; the closest to this paper is the classical Holladay variational characterization, according to which the natural cubic interpolating spline can be defined as the unique solution of the following optimization problem:

$$\min \|f''\|_2 \quad \text{subject to} \quad f(t_i) = y_i,\ i = 1, \ldots, N+2, \tag{1}$$

where $\|\cdot\|$ denotes the norm of $L^2[a, b]$. With a simple transformation, this problem can be written as a nearest point problem in $L^2[a, b]$: find the projection of the origin on the intersection of the hyperplanes

$$\left\{ u \in L^2[a, b] \mid \int_a^b u(t) B_i(t) dt = d_i, \quad i = 1, \ldots, N \right\},$$

where B_i are the piecewise linear normalized B-splines with support $[t_i, t_{i+2}]$ and d_i are the second divided differences.

*Received by the editors July 29, 2001; accepted for publication (in revised form) March 26, 2002; published electronically October 1, 2002.

http://www.siam.org/journals/siopt/13-2/39312.html

†Mathematical Reviews, Ann Arbor, MI 48107 (ald@ams.org).

‡School of Mathematics, University of New South Wales, Sydney 2052, NSW, Australia (hdqi@maths.unsw.edu.au). The research of this author was supported by the Australian Research Council.

§Department of Applied Mathematics, Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong (maqilq@polyu.edu.hk). This author's work is supported by the Hong Kong Research Grant Council, grant PolyU 5296/02P.

¶Department of Mathematics, Graduate School, Chinese Academy of Sciences, P.O. Box 3908, Beijing 100039, People's Republic of China (hxyin@maths.unsw.edu.au). This author's work was done while the author was visiting the University of New South Wales, Australia, and was supported by the Australian Research Council.

Since the mid '80s, after the ground-breaking paper of Micchelli et al. [15], the attention of a number of researchers has been attracted to spline interpolation problems with constraints. For example, if we add to problem (1) the additional constraint $f'' \geq 0$, we obtain a convex interpolation problem; provided that the data are "convex," then a convex interpolant "preserves the shape" of the data. If we add the constraint $f' \geq 0$, we obtain a monotone interpolation problem. Central to our analysis here is a subsequent paper by Irvine, Marin, and Smith [11], who rigorously defined the problem of shape-preserving spline interpolation and laid the groundwork for its numerical analysis. In particular, they proposed a Newton-type method and, based on numerical examples, conjectured its fast (quadratic) theoretical convergence. In the present paper we prove this conjecture.

We approach the problem of Irvine, Marin, and Smith [11] in a new way, by using recent advances in optimization. It is now well understood that, in general, the traditional methods based on standard calculus may not work for optimization problems with constraints; however, such problems can be reformulated as nonsmooth problems that need special treatment. The corresponding theory emerged already in the '70s, championed by the works of R. T. Rockafellar and his collaborators, and is now becoming a standard tool for more and more theoretical and practical problems. The present paper is an example of how nonsmooth analysis can be applied to solve a problem from numerical analysis that hasn't been solved for quite a while.

Before stating the problem of shape-preserving interpolation that we consider in this paper, we briefly review the result of nonsmooth analysis which provides the basis for this work.

For a locally Lipschitz continuous function $G : \mathbf{R}^n \to \mathbf{R}^n$, the generalized Jacobian $\partial G(x)$ of G at x in the sense of Clarke [2] is the convex hull of all limits obtained along sequences on which G is differentiable:

$$\partial G(x) = \operatorname{co}\left\{ \lim_{x^j \to x} \nabla G(x^j) \mid G \text{ is differentiable at } x^j \in \mathbf{R}^n \right\}.$$

The generalized Newton method for the (nonsmooth) equation $G(x) = 0$ has the following form:

$$x^{k+1} = x^k - V_k^{-1} G(x^k), \quad V_k \in \partial G(x^k). \tag{2}$$

A function $G : \mathbf{R}^n \to \mathbf{R}^m$ is strongly semismooth at x if it is locally Lipschitz and directionally differentiable at x, and for all $h \to 0$ and $V \in \partial G(x + h)$ one has $G(x + h) - G(x) - Vh = O(\|h\|^2)$.

The local convergence of the generalized Newton method for strongly semismooth equations is summarized in the following fundamental result, which is a direct generalization of the classical theorem of quadratic convergence of the Newton method.

THEOREM 1.1 (see [16, Theorem 3.2]). *Let $G : \mathbf{R}^n \to \mathbf{R}^n$ be strongly semismooth at x^* and let $G(x^*) = 0$. Assume that all elements V of the generalized Jacobian $\partial G(x^*)$ are nonsingular matrices. Then every sequence generated by the method* (2) *is q-quadratically convergent to x^*, provided that the starting point x^0 is sufficiently close to x^*.*

In the remaining part of the introduction we review the method of Irvine, Marin, and Smith [11] for shape-preserving cubic spline interpolation and also briefly discuss the contents of this paper. Let $\{(t_i, y_i)\}_1^{N+2}$ be given interpolation data and let $d_i, i = 1, 2, \ldots, N$, be the associated second divided differences. Throughout the

paper we assume that $d_i \neq 0$ for all $i = 1, \ldots, N$; we will discuss this assumption later. Define the following subsets $\Omega_i, i = 1, 2, 3$, of $[a, b]$:

$$\begin{aligned}
\Omega_1 &:= \{[t_i, t_{i+1}] |\ d_{i-1} > 0 \text{ and } d_i > 0\},\\
\Omega_2 &:= \{[t_i, t_{i+1}] |\ d_{i-1} < 0 \text{ and } d_i < 0\},\\
\Omega_3 &:= \{[t_i, t_{i+1}] |\ d_{i-1} d_i < 0\}.
\end{aligned}$$

Also, let

$$[t_1, t_2] \subset \begin{cases} \Omega_1 & \text{if } d_1 > 0,\\ \Omega_2 & \text{if } d_1 < 0, \end{cases} \qquad [t_{N+1}, t_{N+2}] \subset \begin{cases} \Omega_1 & \text{if } d_N > 0,\\ \Omega_2 & \text{if } d_N < 0. \end{cases}$$

The problem of shape-preserving interpolation as stated by Micchelli et al. [15] is as follows:

$$\text{minimize } \|f''\|_2 \tag{3}$$

$$\begin{aligned}
\text{subject to } \ & f(t_i) = y_i, \quad i = 1, 2, \ldots, N+2,\\
& f''(t) \geq 0, \quad t \in \Omega_1, \qquad f''(t) \leq 0, \quad t \in \Omega_2,\\
& f \in W^{2,2}[a, b].
\end{aligned}$$

Here $W^{2,2}[a, b]$ denotes the Sobolev space of functions with absolutely continuous first derivatives and second derivatives in $L^2[a, b]$. The inequality constraint on the set Ω_1 (resp., Ω_2) means that the interpolant preserves the convexity (resp., concavity) of the data; for more details, see [11, p. 137].

Micchelli et al. [15, Theorem 4.3] showed that the solution of the problem (3) exists and is unique, and its second derivative has the following form:

$$\begin{aligned}
f''(t) = & \left(\sum_{i=1}^{N} \lambda_i B_i(t) \right)_+ \mathcal{X}_{\Omega_1}(t) - \left(\sum_{i=1}^{N} \lambda_i B_i(t) \right)_- \mathcal{X}_{\Omega_2}(t)\\
& + \left(\sum_{i=1}^{N} \lambda_i B_i(t) \right) \mathcal{X}_{\Omega_3}(t),
\end{aligned} \tag{4}$$

where $\lambda = (\lambda_1, \ldots, \lambda_N)^T$ is a vector in $\mathbf{R}^N$, $a_+ = \max\{0, a\}$, $(a)_- = (-a)_+$, and $\mathcal{X}_\Omega$ is the characteristic function of the set Ω. This result can also be deduced, as shown first in [4], from duality in optimization; specifically, here λ is the vector of the Lagrange multipliers associated with the equality (interpolation) constraints. For more on duality in this context, see the discussion in our previous paper [5]. In short, the optimality condition of the problem dual to (3) has the form of the nonlinear equation

$$F(\lambda) = d, \tag{5}$$

where $d = (d_1, \ldots, d_N)^T$ and the vector function $F : \mathbf{R}^N \to \mathbf{R}^N$ has components

$$F_i(\lambda) = \int_{[t_i, t_{i+2}] \cap \Omega_1} \left(\sum_{l=1}^{N} \lambda_l B_l(t) \right)_+ B_i(t) dt - \int_{[t_i, t_{i+2}] \cap \Omega_2} \left(\sum_{l=1}^{N} \lambda_l B_l(t) \right)_- B_i(t) dt$$

$$+\int_{[t_i,t_{i+2}]\cap\Omega_3}\left(\sum_{l=1}^{N}\lambda_l B_l(t)\right)B_i(t)dt,\quad i=1,2,\ldots,N. \tag{6}$$

Irvine, Marin, and Smith [11] proposed the following method for solving equation (5): Given $\lambda^0\in\mathbf{R}^N$, λ^{k+1} is a solution of the linear system

$$M(\lambda^k)(\lambda^{k+1}-\lambda^k)=-F(\lambda^k)+d, \tag{7}$$

where $M(\lambda)\in\mathbf{R}^{N\times N}$ is the tridiagonal symmetric matrix with components

$$(M(\lambda))_{ij}=\int_a^b P(\lambda,t)B_i(t)B_j(t)dt.$$

Here

$$P(\lambda,t):=\left(\sum_{l=1}^{N}\lambda_l B_l(t)\right)_+^0\mathcal{X}_{\Omega_1}(t)+\left(\sum_{l=1}^{N}\lambda_l B_l(t)\right)_-^0\mathcal{X}_{\Omega_2}(t)+\mathcal{X}_{\Omega_3}(t), \tag{8}$$

where

$$(\tau)_+^0:=\begin{cases}1 & \text{if } \tau>0,\\ 0 & \text{otherwise,}\end{cases}\qquad (\tau)_-^0:=(-\tau)_+^0.$$

Since the matrix M resembles the Jacobian of F (which may not exist for some λ, and then M is a kind of "directional Jacobian," more precisely, as we will see later, an element of the generalized Jacobian), the method (7) has been named the Newton method. It was also observed in [11] that the Newton-type iteration (7) reduces to $M(\lambda^k)\lambda^{k+1}=d$; that is, no evaluations of the function F are needed during iterations.

In our previous paper [5], we considered the problem of convex spline interpolation, that is, with $\Omega_1=[a,b]$, and proved local superlinear convergence of the corresponding version of the Newton method (7). In a subsequent paper [6], by a more detailed analysis of the geometry of the dual problem, we obtained local quadratic convergence of the Newton method, again for convex interpolation. In this paper, we consider the shape-preserving interpolation problem originally stated in Irvine, Marin, and Smith [11] and prove their conjecture that the method is locally quadratically convergent. As a side result, we observe that the solution of the problem considered is Lipschitz continuous with respect to the interpolation values. In section 3 we give a modification of the method which has global quadratic convergence. Results of extensive numerical experiments are presented in section 4.

As for related results, the conjecture of Irvine, Marin, and Smith [11] was proved in [1] under an additional condition which turned out to be equivalent to smoothness of the function F in (5). Also, a positive answer to this conjecture without additional assumptions was announced in [10], but a proof was never made available to us.

2. Local quadratic convergence. For notational convenience, we introduce a "dummy" node t_0 with corresponding $\lambda_0=0$ and $B_0(t)=0$; then, for every i, the sum $\sum_{l=1}^{N}\lambda_l B_l(t)$ restricted to $[t_i,t_{i+1}]$ has the form $\lambda_{i-1}B_{i-1}(t)+\lambda_i B_i(t)$. Our first result concerns continuity and differentiability properties of the function F defined in (6).

LEMMA 2.1. *The function F with components defined in* (6) *is strongly semismooth.*

Proof. The claim is merely an extension of [6, Proposition 2.4], where it is proved that the functions

$$\int_{t_i}^{t_{i+1}} \left(\sum_{l=1}^{N} \lambda_l B_l(t)\right)_+ B_i(t)dt, \qquad \int_{t_{i+1}}^{t_{i+2}} \left(\sum_{l=1}^{N} \lambda_l B_l(t)\right)_+ B_i(t)dt,$$

and

$$\int_{t_i}^{t_{i+2}} \left(\sum_{l=1}^{N} \lambda_l B_l(t)\right)_+ B_i(t)dt$$

are strongly semismooth. Hence the function

$$\int_{[t_i,t_{i+2}]\cap\Omega_1} \left(\sum_{l=1}^{N} \lambda_l B_l(t)\right)_+ B_i(t)dt$$

is strongly semismooth by noticing that

$$[t_i, t_{i+2}] \cap \Omega_1 \in \{[t_i, t_{i+1}],\ [t_{i+1}, t_{i+2}],\ [t_i, t_{i+2}],\ \emptyset\}.$$

We note that the function

$$\int_{[t_i,t_{i+2}]\cap\Omega_3} \left(\sum_{l=1}^{N} \lambda_l B_l(t)\right) B_i(t)dt$$

is linear and therefore is strongly semismooth. Since either $[t_i, t_{i+2}] \cap \Omega_1 = \emptyset$ or $[t_i, t_{i+2}] \cap \Omega_2 = \emptyset$, F_i is given either by

$$F_i(\lambda) = \int_{[t_i,t_{i+2}]\cap\Omega_1} \left(\sum_{l=1}^{N} \lambda_l B_l(t)\right)_+ B_i(t)dt + \int_{[t_i,t_{i+2}]\cap\Omega_3} \left(\sum_{l=1}^{N} \lambda_l B_l(t)\right) B_i(t)dt \tag{9}$$

or by

$$F_i(\lambda) = -\int_{[t_i,t_{i+2}]\cap\Omega_2} \left(\sum_{l=1}^{N} \lambda_l B_l(t)\right)_- B_i(t)dt + \int_{[t_i,t_{i+2}]\cap\Omega_3} \left(\sum_{l=1}^{N} \lambda_l B_l(t)\right) B_i(t)dt. \tag{10}$$

A composite of strongly semismooth functions is strongly semismooth [8, Theorem 19]. Hence the function F_i by (9) is strongly semismooth. If F_i is given by (10), then

$$F_i(\lambda) = -\int_{[t_i,t_{i+2}]\cap\Omega_2} \left(-\sum_{l=1}^{N} \lambda_l B_l(t)\right)_+ B_i(t)dt + \int_{[t_i,t_{i+2}]\cap\Omega_3} \left(\sum_{l=1}^{N} \lambda_l B_l(t)\right) B_i(t)dt.$$

Again from [8, Theorem 19], the first part of F_i is strongly semismooth, which in turn implies the strong semismoothness of F_i. We conclude that F is strongly semismooth since each component of F is strongly semismooth. ∎

If the integral over $[a,b]$ of the piecewise linear function $(\sum_{l=1}^{N}\lambda_l B_l(t))_+$ in λ were piecewise smooth, then one would automatically obtain that F is strongly semismooth. Furthermore, in this case quadratic convergence of the Newton method would follow directly from [13]. The following example of dimension 2 shows that such an argument does not work. Let

$$f(\lambda_1,\lambda_2)=\int_0^1((1-t)\lambda_1+t\lambda_2)_+\,dt.$$

Direct calculation shows that f is continuously differentiable everywhere except at the origin $(0,0)$. A result due to Rockafellar [17] says that any function from $\mathbf{R}^n$ to $\mathbf{R}$ with $n\geq 2$, which is continuously differentiable everywhere but one point, could not be piecewise smooth. Hence the function above is not piecewise smooth.

In order to apply Theorem 1.1, we next prove that $M(\lambda)\in\partial F(\lambda)$ for any $\lambda\in\mathbf{R}^N$ and that V is nonsingular for any $V\in\partial F(\lambda^*)$, where λ^* is the unique solution of (5).

LEMMA 2.2. *For any* $\lambda\in\mathbf{R}^N$, $M(\lambda)\in\partial F(\lambda)$.

Proof. Let $\lambda\in\mathbf{R}^N$ be arbitrarily chosen (but fixed) and let

$$T(\lambda):=\left\{t\in\Omega_1\cup\Omega_2\mid\sum_{l=1}^{N}\lambda_l B_l(t)=0\right\},\qquad \bar{T}(\lambda):=(\Omega_1\cup\Omega_2)\setminus T(\lambda).$$

Suppose $[t_i,t_{i+1}]\subset\Omega_1\cup\Omega_2$ for some i. Due to the form of B_i, the restriction of $(\sum_{l=1}^{N}\lambda_l B_l(t))$ to $[t_i,t_{i+1}]$ becomes $(\lambda_{i-1}B_{i-1}(t)+\lambda_i B_i(t))$, i.e.,

$$\sum_{l=1}^{N}\lambda_l B_l(t)\,\big|_{[t_i,t_{i+1}]}=\lambda_{i-1}B_{i-1}(t)+\lambda_i B_i(t).$$

Then

$$T(\lambda)\,\big|_{[t_i,t_{i+1}]}=\begin{cases}[t_i,t_{i+1}] & \text{if } \lambda_{i-1}=\lambda_i=0,\\ t_i^* & \text{otherwise,}\end{cases}\tag{11}$$

where t_i^* is a point in $[t_i,t_{i+1}]$. Hence $T(\lambda)$ contains closed intervals of the form $[t_i,t_{i+1}]$ and finitely many isolated points. For $i=1,\dots,N$, define

$$F_i^-(\xi):=\int_{T(\lambda)\cap\Omega_1}\left(\sum_{l=1}^{N}\xi_l B_l(t)\right)_+B_i(t)dt-\int_{T(\lambda)\cap\Omega_2}\left(\sum_{l=1}^{N}\xi_l B_l(t)\right)_-B_i(t)dt,$$

$$F_i^+(\xi):=\int_{\bar{T}(\lambda)\cap\Omega_1}\left(\sum_{l=1}^{N}\xi_l B_l(t)\right)_+B_i(t)dt-\int_{\bar{T}(\lambda)\cap\Omega_2}\left(\sum_{l=1}^{N}\xi_l B_l(t)\right)_-B_i(t)dt$$
$$+\int_{\Omega_3}\left(\sum_{l=1}^{N}\xi_l B_l(t)\right)B_i(t)dt,$$

and let $F^-(\xi):=(F_1^-(\xi),\dots,F_N^-(\xi))^T$, $F^+(\xi):=(F_1^+(\xi),\dots,F_N^+(\xi))^T$. Then for any $\xi\in\mathbf{R}^N$, we have

$$F(\xi)=F^-(\xi)+F^+(\xi),$$

and it follows from (11) that F^+ is continuously differentiable in a neighborhood of λ, say $U(\lambda)$. From the definition of the generalized Jacobian we obtain that for any $\xi\in U(\lambda)$,

$$\partial F(\xi)=\partial F^-(\xi)+\nabla F^+(\xi),\tag{12}$$

where $\nabla F^+(\xi)$ is the Jacobian of F^+ at $\xi \in U(\lambda)$ given by

$$\left(\nabla F^+(\xi)\right)_{ij} = \int_{\bar{T}(\lambda)} \left[\left(\sum_{l=1}^N \xi_l B_l(t) \right)_+^0 \mathcal{X}_{\Omega_1}(t) + \left(\sum_{l=1}^N \xi_l B_l(t) \right)_-^0 \mathcal{X}_{\Omega_2}(t) \right] B_i(t) B_j(t) dt$$

$$+ \int_a^b B_i(t) B_j(t) \mathcal{X}_{\Omega_3}(t) dt. \tag{13}$$

Since

$$\sum_{l=1}^N \lambda_l B_l(t) = 0 \qquad \text{for all } \ t \in T(\lambda),$$

(13) becomes

$$\left(\nabla F^+(\lambda)\right)_{ij} = \int_a^b P(\lambda, t) B_i(t) B_j(t) dt. \tag{14}$$

We will next prove that every element in $\partial F^-(\lambda)$ is positive semidefinite. In particular, the zero matrix belongs to $\partial F^-(\lambda)$. Define $\theta : \mathbf{R}^N \to \mathbf{R}$ as

$$\theta(\xi) := \frac{1}{2} \int_{T(\lambda)\cap\Omega_1} \left(\sum_{l=1}^N \xi_l B_l(t) \right)_+^2 dt + \frac{1}{2} \int_{T(\lambda)\cap\Omega_2} \left(\sum_{l=1}^N \xi_l B_l(t) \right)_-^2 dt.$$

The function θ is a continuously differentiable convex function, and its gradient is equal to $F^-(\xi)$. Then the positive semidefiniteness of the elements of $\partial F^-(\lambda)$ follows from the fact that any matrix in the generalized Jacobian of the gradient of a convex function must be symmetric and positive semidefinite. Because isolated points make no contribution to $\theta(\xi)$, we assume without loss of generality that $T(\lambda)$ contains only intervals of the form $[t_i, t_{i+1}]$. Let

$$\mathcal{I}_1 := \{i \in \{1, \ldots, N\} |\ [t_i, t_{i+1}] \subset T(\lambda) \cap \Omega_1\},$$
$$\mathcal{I}_2 := \{i \in \{1, \ldots, N\} |\ [t_i, t_{i+1}] \subset T(\lambda) \cap \Omega_2\}.$$

Then

$$\theta(\xi) = \frac{1}{2} \sum_{i\in\mathcal{I}_1} \int_{t_i}^{t_{i+1}} (\xi_{i-1} B_{i-1}(t) + \xi_i B_i(t))_+^2 dt + \frac{1}{2} \sum_{i\in\mathcal{I}_2} \int_{t_i}^{t_{i+1}} (\xi_{i-1} B_{i-1}(t) + \xi_i B_i(t))_-^2 dt.$$

Now define $e = (e_1, \ldots, e_N)^T$ by

$$e_{i-1} = e_i = 1 \quad \text{for } \ i \in \mathcal{I}_1, \qquad e_{i-1} = e_i = -1 \quad \text{for } \ i \in \mathcal{I}_2,$$

and zero for the remaining components. We note that e is well defined since for any $i \in \{1, \ldots, N\}$, $[t_i, t_{i+2}] \cap \Omega_1 = \emptyset$ or $[t_i, t_{i+2}] \cap \Omega_2 = \emptyset$. Then $F^-(\lambda - \tau e)$ is differentiable for all $\tau > 0$ because

$$\sum_{l=1}^N (\lambda - \tau e)_l B_l(t) \begin{cases} < 0 & \text{for } \ t \in T(\lambda) \cap \Omega_1 \text{ and } \tau > 0, \\ > 0 & \text{for } \ t \in T(\lambda) \cap \Omega_2 \text{ and } \tau > 0. \end{cases}$$

Hence

$$\lim_{\tau \to 0} \nabla F^-(\lambda - \tau e) = 0 \in \partial F^-(\lambda).$$

We are ready to complete the proof of the lemma. From (14) we have $\nabla F^+(\lambda) = M(\lambda)$. Since the zero matrix belongs to $\partial F^-(\lambda)$, we get $M(\lambda) \in \partial F(\lambda)$ from (12). ∎

If λ^* is the solution of (5), we are able to show a stronger result about the generalized Jacobian of F at λ^*.

LEMMA 2.3. *If λ^* is the solution of* (5), *then every element of $\partial F(\lambda^*)$ is positive definite.*

Proof. We have already shown in the preceding proof that

$$\partial F(\lambda^*) = \partial F^-(\lambda^*) + \nabla F^+(\lambda^*),$$

and every element in $\partial F^-(\lambda^*)$ is positive semidefinite. Thus, it is sufficient to prove that $\nabla F^+(\lambda^*)$ is positive definite; that is, $M(\lambda^*)$ is positive definite. We use a result from [11, p. 138] which says that if $P(\lambda)$ does not vanish identically on any $[t_i, t_{i+2}]$, $i = 1, \ldots, N$, then $M(\lambda)$ is positive definite. On the contrary, suppose that $P(\lambda^*)$ vanishes on, say, $[t_i, t_{i+2}]$. Then $[t_i, t_{i+2}] \cap \Omega_3 = \emptyset$ and

$$0 \neq d_i = F_i(\lambda^*) = \int_{[t_i, t_{i+2}] \cap \Omega_1} \left(\sum_{l=1}^{N} \lambda_l^* B_l(t) \right)_+ B_i(t) dt - \int_{[t_i, t_{i+2}] \cap \Omega_2} \left(\sum_{l=1}^{N} \lambda_l^* B_l(t) \right)_- B_i(t) dt = 0.$$

The obtained contradiction completes the proof. ∎

By combining the above lemmas and applying Theorem 1.1, we obtain the main result of this paper which settles the question posed in [11].

THEOREM 2.4. *Let λ^* be the solution of* (5), *and let all second divided differences d_i be nonzero. Then the method* (7) *is well defined, and the sequence generated by this method converges quadratically to λ^* if the starting point λ^0 is sufficiently close to λ^*.*

Proof. The method (7) is a particular case of the generalized Newton method (2) for (5) inasmuch as $M(\lambda) \in \partial F(\lambda)$ (Lemma 2.2). Moreover, F is strongly semismooth at λ^* (Lemma 2.1), and every element in $\partial F(\lambda^*)$ is nonsingular (Lemma 2.3). Hence all conditions in Theorem 1.1 are satisfied, and we obtain the claim. ∎

Remark 2.5. As a side result, from Lemma 2.3 and the Clarke inverse function theorem [2, Theorem 7.1.1], we obtain that the solution of the problem (3) is a Lipschitz continuous function of the interpolation values y_i. Indeed, since the generalized Jacobian $\partial F(\lambda^*)$ is nonsingular, where λ^* is the optimal multiplier associated with the solution f^*, the map F^{-1} is, locally around $d^* = F(\lambda^*)$, single-valued and Lipschitz continuous. Thus for d close to d^* there exists a unique solution $\lambda(d)$ to (5), and the function $d \mapsto \lambda(d)$ is Lipschitz continuous. It remains to observe that d is linear in y and, from (4), f'' is a Lipschitz continuous function of λ in the supremum norm of $C[a, b]$. Thus the mapping "interpolation values $y \mapsto$ solution of (3)" is a Lipschitz continuous function from $y \in \mathbf{R}^{N+2}$ to the space $C^2[a, b]$ equipped with the supremum norm. This result could be further strengthened with respect to differentiability of the solution, but we shall not go into this here.

3. Global convergence. In this section we give a damped version of algorithm (7) by using the following merit function:

$$L(\lambda) = \frac{1}{2}\int_a^b \left(\sum_{l=1}^N \lambda_l B_l(t)\right)_+^2 \mathcal{X}_{\Omega_1}(t)dt + \frac{1}{2}\int_a^b \left(\sum_{l=1}^N \lambda_l B_l(t)\right)_-^2 \mathcal{X}_{\Omega_2}(t)dt + \frac{1}{2}\int_a^b \left(\sum_{l=1}^N \lambda_l B_l(t)\right)^2 \mathcal{X}_{\Omega_3}(t)dt - \sum_{l=1}^N \lambda_l d_l. \tag{15}$$

From the very definition, this function is convex and continuously differentiable, with $\nabla L(\lambda) = F(\lambda) - d$.

Recall that a function $\varphi : \mathbf{R}^N \to \mathbf{R}$ is *coercive* (also called inf-compact) if for every $c \in \mathbf{R}$ its level set

$$\mathcal{L}_\varphi(c) = \{x \in \mathbf{R}^N \mid \varphi(x) \le c\}$$

is bounded. In the proposition below we will show that the function L in (15) is coercive. To begin with, we define three index sets

$$\begin{aligned}\mathcal{I}_+ &:= \{i \in \{1,\dots,N\} |\ [t_i, t_{i+1}] \subset \Omega_1\},\\ \mathcal{I}_- &:= \{i \in \{1,\dots,N\} |\ [t_i, t_{i+1}] \subset \Omega_2\},\\ \mathcal{I}_0 &:= \{i \in \{1,\dots,N\} |\ [t_i, t_{i+1}] \subset \Omega_3\}\end{aligned}$$

and associate with them the following function:

$$\hat{L}(\lambda) := \frac{1}{2}\sum_{i\in\mathcal{I}_+}\int_{t_i}^{t_{i+1}} \left(\sum_{l=1}^N \lambda_l B_l(t)\right)_+^2 dt + \frac{1}{2}\sum_{i\in\mathcal{I}_-}\int_{t_i}^{t_{i+1}} \left(\sum_{l=1}^N \lambda_l B_l(t)\right)_-^2 dt + \frac{1}{2}\sum_{i\in\mathcal{I}_0}\int_{t_i}^{t_{i+1}} \left(\sum_{l=1}^N \lambda_l B_l(t)\right)^2 dt - \sum_{l=1}^N \lambda_l d_l.$$

Observe that, from the definition of the sets $\Omega_i, i = 1, 2, 3$, for any $i \in \{1,\dots,N\}$, we have $[t_i, t_{i+2}] \cap \Omega_1 = \emptyset$ or $[t_i, t_{i+2}] \cap \Omega_2 = \emptyset$. For a fixed i this implies

$$i \in \mathcal{I}_+ \implies \begin{cases} i-1 \in \mathcal{I}_+ \ \text{ or } \ i-1 \in \mathcal{I}_0,\\ i+1 \in \mathcal{I}_+ \ \text{ or } \ i+1 \in \mathcal{I}_0\end{cases} \tag{16}$$

and

$$i \in \mathcal{I}_- \implies \begin{cases} i-1 \in \mathcal{I}_- \ \text{ or } \ i-1 \in \mathcal{I}_0,\\ i+1 \in \mathcal{I}_- \ \text{ or } \ i+1 \in \mathcal{I}_0.\end{cases} \tag{17}$$

Also, observe that

$$\hat{L}(\lambda) = \frac{1}{2}\int_a^{t_{N+1}} \left(\sum_{l=1}^N \lambda_l B_l(t)\right)_+^2 \mathcal{X}_{\Omega_1}(t)dt + \frac{1}{2}\int_a^{t_{N+1}} \left(\sum_{l=1}^N \lambda_l B_l(t)\right)_-^2 \mathcal{X}_{\Omega_2}(t)dt + \frac{1}{2}\int_a^{t_{N+1}} \left(\sum_{l=1}^N \lambda_l B_l(t)\right)^2 \mathcal{X}_{\Omega_3}(t)dt - \sum_{l=1}^N \lambda_l d_l \le L(\lambda).$$

Thus, if we show the coercivity of $\hat{L}$, the coercivity of L will follow. In the proposition below we use the index set

$$\bar{\mathcal{I}}_0 := \{1, \ldots, N\} \setminus \cup_{i \in \mathcal{I}_0} \{i-1, i\}$$

and the following four sets in $\mathbf{R}^N$:

$$\begin{aligned} V_0 &:= \{v \in \mathbf{R}^N \mid v_{i-1} = v_i = 0 \text{ for all } i \in \mathcal{I}_0\}, \\ V_+ &:= \{v \in \mathbf{R}^N \mid v_i \le 0 \text{ for all } i \in \mathcal{I}_+ \cap \bar{\mathcal{I}}_0\}, \\ V_- &:= \{v \in \mathbf{R}^N \mid v_i \ge 0 \text{ for all } i \in \mathcal{I}_- \cap \bar{\mathcal{I}}_0\}, \qquad V := V_0 \cap V_+ \cap V_- \,. \end{aligned}$$

PROPOSITION 3.1. *The function L is coercive.*

Proof. In view of the above, it is sufficient to prove that the level sets

$$\mathcal{L}(c) := \{\lambda \in \mathbf{R}^N \mid \hat{L}(\lambda) \le c\}$$

are bounded for every $c \in \mathbf{R}$. Note that, for every $c \in \mathbf{R}$, the set $\mathcal{L}(c)$ is closed and convex. Assume on the contrary that $\mathcal{L}(c_0)$ is unbounded for some $c_0 \in \mathbf{R}$ and let, without loss of generality, $c_0 > 0$. We first show that there exists a vector $s \in \mathbf{R}^N$, $s \neq 0$, such that $\beta s \in \mathcal{L}(c_0)$ for every $\beta \ge 0$. Suppose that for every $s \in \mathbf{R}^N$ there exists $\beta_s \ge 0$ such that $\beta_s s \notin \mathcal{L}(c_0)$. From the convexity of $\mathcal{L}(c_0)$ and $0 \in \mathcal{L}(c_0)$, it follows that $\beta s \notin \mathcal{L}(c_0)$ whenever $\beta \ge \beta_s$. Let

$$\beta(s) := \max\{\beta \mid \ \beta \ge 0, \ \beta s \in \mathcal{L}(c_0)\}.$$

Then $\beta(s) < \infty$ since $\mathcal{L}(c_0)$ is closed and $\beta(\cdot)$ is an upper semicontinuous function over $\mathbf{R}^N$. Then

$$\beta^* := \sup\{\beta(s) : \ \|s\| = 1\} < \infty.$$

Hence $\mathcal{L}(c_0)$ is contained in a ball centered at the origin with radius $\beta^* + 1$. This contradiction establishes the existence of a vector $s \in \mathbf{R}^N$, $s \neq 0$, such that $\beta s \in \mathcal{L}(c_0)$ for all $\beta \ge 0$. Now for such s we define

$$\begin{aligned} \kappa(\beta) := \hat{L}(\beta s) = \frac{1}{2} \sum_{i \in \mathcal{I}_+} \int_{t_i}^{t_{i+1}} \beta^2 \left(\sum_{l=1}^{N} s_l B_l(t) \right)_+^2 dt + \frac{1}{2} \sum_{i \in \mathcal{I}_-} \int_{t_i}^{t_{i+1}} \beta^2 \left(\sum_{l=1}^{N} s_l B_l(t) \right)_-^2 dt \\ + \frac{1}{2} \sum_{i \in \mathcal{I}_0} \int_{t_i}^{t_{i+1}} \beta^2 \left(\sum_{i=1}^{N} s_l B_l(t) \right)^2 dt - \beta \sum_{l=1}^{N} s_l d_l. \end{aligned}$$

A more explicit form of $\kappa(\beta)$ is

$$\begin{aligned} \kappa(\beta) = \frac{1}{2} \sum_{i \in \mathcal{I}_+} \int_{t_i}^{t_{i+1}} \beta^2 (s_{i-1} B_{i-1} + s_i B_i)_+^2 dt + \frac{1}{2} \sum_{i \in \mathcal{I}_-} \int_{t_i}^{t_{i+1}} \beta^2 (s_{i-1} B_{i-1} + s_i B_i)_-^2 dt \\ + \frac{1}{2} \sum_{i \in \mathcal{I}_0} \int_{t_i}^{t_{i+1}} \beta^2 (s_{i-1} B_{i-1} + s_i B_i)^2 dt - \beta \sum_{l=1}^{N} s_l d_l. \end{aligned}$$

Now we consider the following cases.

Case 1. $s \in V$. Consider three subcases corresponding to the three quadratic terms of $\kappa(\beta)$, respectively.

Subcase 1.1. $i \in \mathcal{I}_0$. By the definition of V_0, we have $s_{i-1} = 0, s_i = 0$.

Subcase 1.2. $i \in \mathcal{I}_+$. It follows from (16) that $(i-1) \in \mathcal{I}_+$ or $(i-1) \in \mathcal{I}_0$, and $(i+1) \in \mathcal{I}_+$ or $(i+1) \in \mathcal{I}_0$. In particular, we have from $s \in V$ and the definitions of V_0 and V_+ that

$$i-1 \in \mathcal{I}_0 \implies \begin{cases} s_{i-1} = 0, & \\ i \in \mathcal{I}_+ \cap \bar{\mathcal{I}}_0 \Longrightarrow s_i \le 0 & \text{if } i+1 \in \mathcal{I}_+, \\ s_i = 0 & \text{if } i+1 \in \mathcal{I}_0 \end{cases}$$

and

$$i-1 \in \mathcal{I}_+ \implies \begin{cases} i-1 \in \mathcal{I}_+ \cap \bar{\mathcal{I}}_0 \Longrightarrow s_{i-1} \le 0, & \\ i \in \mathcal{I}_+ \cap \bar{\mathcal{I}}_0 \Longrightarrow s_i \le 0 & \text{if } i+1 \in \mathcal{I}_+, \\ s_i = 0 & \text{if } i+1 \in \mathcal{I}_0. \end{cases}$$

Hence for this subcase we have $s_{i-1} \le 0, s_i \le 0$.

Subcase 1.3. $i \in \mathcal{I}_-$. Then it follows from (17) that $(i-1) \in \mathcal{I}_-$ or $(i-1) \in \mathcal{I}_0$ and $(i+1) \in \mathcal{I}_-$ or $(i+1) \in \mathcal{I}_0$. In particular, we have again from $s \in V$ and the definitions of V_0 and V_- that

$$i-1 \in \mathcal{I}_0 \implies \begin{cases} s_{i-1} = 0, & \\ i \in \mathcal{I}_- \cap \bar{\mathcal{I}}_0 \Longrightarrow s_i \ge 0 & \text{if } i+1 \in \mathcal{I}_-, \\ s_i = 0 & \text{if } i+1 \in \mathcal{I}_0 \end{cases}$$

and

$$i-1 \in \mathcal{I}_- \implies \begin{cases} i-1 \in \mathcal{I}_- \cap \bar{\mathcal{I}}_0 \Longrightarrow s_{i-1} \ge 0, & \\ i \in \mathcal{I}_- \cap \bar{\mathcal{I}}_0 \Longrightarrow s_i \ge 0 & \text{if } i+1 \in \mathcal{I}_-, \\ s_i = 0 & \text{if } i+1 \in \mathcal{I}_0. \end{cases}$$

Hence for this case we have $s_{i-1} \ge 0, s_i \ge 0$.

It follows from the three subcases that the first three terms of $\kappa(\beta)$ (the quadratic part) vanish. Taking $s \in V$ into account, we have

$$\kappa(\beta) = -\beta \sum_{l=1}^{N} s_l d_l = -\beta \sum_{l \in \mathcal{I}_+ \cap \bar{\mathcal{I}}_0} s_l d_l - \beta \sum_{l \in \mathcal{I}_- \cap \bar{\mathcal{I}}_0} s_l d_l.$$

Note that $d_l > 0$, $s_l \le 0$ for any $l \in \mathcal{I}_+ \cap \bar{\mathcal{I}}_0$, and $d_l < 0$, $s_l \ge 0$ for any $l \in \mathcal{I}_- \cap \bar{\mathcal{I}}_0$. Hence the fact that there exists at least one $s_l \ne 0$ (this l must belong to $\mathcal{I}_+ \cap \bar{\mathcal{I}}_0$ or $\mathcal{I}_- \cap \bar{\mathcal{I}}_0$) implies $\kappa(\beta) \to +\infty$ as $\beta \to +\infty$, contradicting $\hat{L}(\beta s) \le c_0$.

Case 2. $s \notin V$.

From the analysis of Case 1, for each i, at least one of the conditions $s_{i-1}s_i = 0$ for $i \in \mathcal{I}_0$, $(s_{i-1} \le 0, s_i \le 0)$ for $i \in \mathcal{I}_+$, and $(s_{i-1} \ge 0, s_i \ge 0)$ for $i \in \mathcal{I}_-$ is violated. Hence

$$r := \frac{1}{2} \sum_{i \in \mathcal{I}_+} \int_{t_i}^{t_{i+1}} (s_{i-1}B_{i-1}(t) + s_iB_i(t))_+^2 dt + \frac{1}{2} \sum_{i \in \mathcal{I}_-} \int_{t_i}^{t_{i+1}} (s_{i-1}B_{i-1}(t) + s_iB_i(t))_-^2 dt$$
$$+ \frac{1}{2} \sum_{i \in \mathcal{I}_0} \int_{t_i}^{t_{i+1}} (s_{i-1}B_{i-1}(t) + s_iB_i(t))^2 dt > 0.$$

Then, $\kappa(\beta) = r\beta^2 - \beta \sum_{l=1}^{N} s_l d_l \to +\infty$ as $\beta \to +\infty$, contradicting $\hat{L}(\beta s) \le c_0$. This completes the proof. □

Since $L(\lambda)$ is convex and coercive and $\nabla L(\lambda) = F(\lambda) - d$, finding a solution of (5) is equivalent to solving the following unconstrained optimization problem:

$$\min_{\lambda \in \mathbf{R}^N} \; L(\lambda). \tag{18}$$

Now we apply the following damped Newton method to the problem (18), which uses the Newton direction given by (7).

ALGORITHM 3.2.

(S.0) *Choose* $\lambda^0 \in \mathbf{R}^N$, $\rho \in (0,1)$, $\sigma \in (0, 1/2)$, *and tolerance* $tol > 0$. $k := 0$.

(S.1) *If* $\varepsilon_k = \|F(\lambda^k) - d\| \le tol$, *then stop. Otherwise, go to* (S.2).

(S.2) *Let* s^k *be a solution of the linear system*

$$(M(\lambda^k) + \varepsilon_k I)s = -\nabla L(\lambda^k). \tag{19}$$

(S.3) *Choose* m_k *as the smallest nonnegative integer* m *satisfying*

$$L(\lambda^k + \rho^m s^k) - L(\lambda^k) \le \sigma \rho^m \nabla L(\lambda^k)^T s^k. \tag{20}$$

(S.4) *Set* $\lambda^{k+1} = \lambda^k + \rho^{m_k} s^k$, $k := k + 1$; *return to step* (S.1).

Assume that tol $= 0$ and Algorithm 3.2 never stops at (S.1) (otherwise, λ^k would be the solution of (5)). The matrix $M(\lambda^k)$ is always positive semidefinite because $M(\lambda^k) \in \partial F(\lambda^k)$, F is monotone, and every element of the generalized Jacobian of the monotone function is positive semidefinite [12, Proposition 2.3(a)]. Hence $M(\lambda^k) + \varepsilon_k I$ is always positive definite for $\varepsilon_k > 0$, and therefore the linear system (19) is uniquely solvable and $s^k \neq 0$. Moreover,

$$(s^k)^T \nabla L(\lambda^k) = -(s^k)^T (M(\lambda^k) + \varepsilon_k I) s^k \le -\varepsilon_k \|s^k\|^2 < 0;$$

that is, s^k provides a descent direction for the function L. Hence the line search criterion (20) is always satisfied for some integer m. Since L is coercive, the sequence generated by the algorithm is bounded and therefore converges quadratically to the solution of (18). The proof of the latter is in line with the standard argument in these circumstances. Specifically, since locally the unit steplength is accepted, our algorithm eventually reduces to the following iteration:

$$M(\lambda^k)s^k = -(F(\lambda^k) - d) + r^k, \qquad \lambda^{k+1} = \lambda^k + s^k,$$

where $r^k = -\varepsilon_k s^k$ is the residual which measures the inaccuracy in the Newton equation

$$M(\lambda^k)\Delta\lambda^k = -(F(\lambda^k) - d).$$

Using the uniform nonsingularity of $M(\lambda^k)$ near solution λ^*, it is easy to see that

$$s^k = O(\|F(\lambda^k) - d\|).$$

According to [3, Theorem 2.2], the accuracy $\|r^k\| = O(\|F(\lambda^k) - d\|^2)$ is sufficient for the local quadratic convergence of the inexact Newton method. Since $\varepsilon_k = \|F(\lambda^k) - d\|$, we have

$$\|r^k\| = \varepsilon_k \|s^k\| = O(\|F(\lambda^k) - d\|^2).$$

For more discussion of the inexact Newton method, we refer to [3, 7, 14].

Summarizing, we have the following theorem.

THEOREM 3.3. *Let the sequence* $\{\lambda^k\}$ *be generated by Algorithm* 3.2 *starting from an arbitrary* $\lambda^0 \in \mathbf{R}^N$. *Then the sequence* $\{\lambda^k\}$ *converges quadratically to the solution* λ^*.

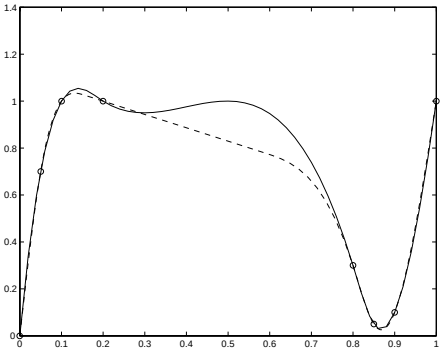


FIG. 1. *Example* 4.1.

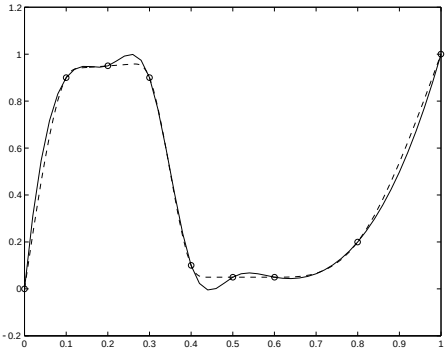


FIG. 2. *Example* 4.2.

4. Numerical results. In this section, we report on some numerical experience with Algorithm 3.2 and demonstrate its global convergence from arbitrary starting points. The typical starting point in shape-preserving algorithms is sign(d), the sign vector of d; see [11]. We report results with the starting point $\mathbf{e}$, the vector of all ones in $\mathbf{R}^N$, which is commonly selected as a starting point in algorithms for convex best interpolations; see [11, 6]. We also test the influence on Algorithm 3.2 of the standing assumption $d_i \neq 0$, $i = 1, \ldots, N$.

We implemented Algorithm 3.2 in MATLAB and tested it on a DEC George Server 8200 with the termination criterion $\|F(\lambda^k) - d\| \leq \text{tol}$ and the following values of the parameters: $\rho = 0.5$, $\sigma = 0.1$, $\text{tol} = 10^{-12}$. In our implementation, $\varepsilon_k = \min\{\delta, \|F(\lambda^k) - d\|\}$ with $\delta = 0.01$. The integrals involved are evaluated exactly using Simpson's rule. The testing problems are collected from the literature and are described in details as follows.

Example 4.1. This problem is from [11] and has the following data:

$$\begin{array}{rcccccccc} t_i = & 0.0 & 0.05 & 0.1 & 0.2 & 0.8 & 0.85 & 0.9 & 1.0. \\ y_i = & 0.0 & 0.7 & 1.0 & 1.0 & 0.3 & 0.05 & 0.1 & 1.0. \end{array}$$

Example 4.2. This problem is again from [11] and has the following data:

$$\begin{array}{rccccccccc} t_i = & 0.0 & 0.1 & 0.2 & 0.3 & 0.4 & 0.5 & 0.6 & 0.8 & 1.0. \\ y_i = & 0.0 & 0.9 & 0.95 & 0.9 & 0.1 & 0.05 & 0.05 & 0.2 & 1.0. \end{array}$$

Example 4.3. This problem is from [9] and has the following data:

$$\begin{array}{rcccccccc} t_i = & 0 & 4 & 6 & 10 & 12 & 14 & 18 & 20. \\ y_i = & 3 & 4 & 9 & 10 & 9 & 5 & 4 & 3. \end{array}$$

Example 4.4. This problem is from [4]: $t_1 = 0, t_2 = 0.1, t_3 = 0.4, t_5 = 0.8, t_6 = 1$, $t_7 = 1.166, t_8 = 1.333, t_9 = 1.5, t_{10} = 1.666$. $y_i = 1/((0.05+t_i)(1.05-t_i)), i = 1, \ldots, 4$, $y_5 = 10, y_6 = 5, y_7 = y_8 = y_9 = 4, y_{10} = 10$.

In Figures 1–5, the dashed line is for the resulting shape-preserving cubic spline (using the data obtained with the starting point $\lambda^0 = \text{sign}(d)$); the solid line is for the natural spline (using the MATLAB SPLINE function), and "o" stands for the original given data. In Table 1 for results of the numerical experiments we use the following notation:

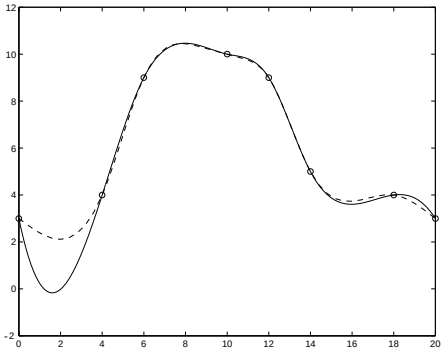


FIG. 3. *Example* 4.3.

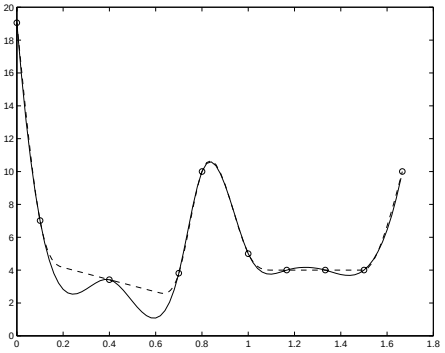


FIG. 4. *Example* 4.4.

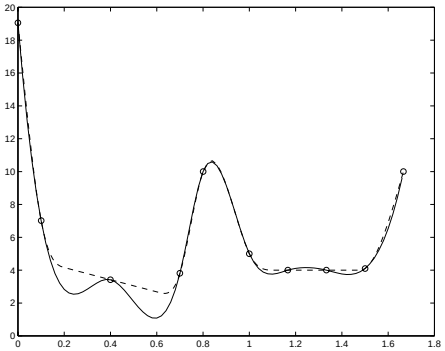


FIG. 5. *Example* 4.4 ($y_9 = 4.1$).

TABLE 1
Numerical results with Algorithm 3.2.

Problem	λ^0	It	Nf	$\|F(\lambda^f) - d\|$
4.1	$\mathbf{e}$	11	17	8.57e-15
	sign(d)	9	10	1.06e-14
4.2	$\mathbf{e}$	11	15	3.02e-14
	sign(d)	10	11	1.03e-14
4.3	$\mathbf{e}$	8	11	4.59e-16
	sign(d)	7	8	2.91e-16
4.4	$\mathbf{e}$	30	31	1.43e-01
($y_9 = 4$)	sign(d)	30	31	1.43e-01
4.4	$\mathbf{e}$	24	44	2.39e-13
($y_9 = 4.1$)	sign(d)	23	39	1.95e-13
4.4	$\mathbf{e}$	12	13	1.01e-13
($y_9 = 5$)	sign(d)	12	13	1.43e-13

Problem: name of the test problem.
λ^0: starting point.
It : number of iterations.
Nf : number of evaluations of the function $f(\lambda)$.
$\|F(\lambda^f) - d\|$: value of $\|F(\lambda) - d\|$ at the last iteration.

From Table 1, we observe that Algorithm 3.2 converges rapidly to the solution

from both starting points for all problems except Example 4.4 ($y_9 = 4$), to which the algorithm within 30 iterations failed to produce an approximate solution meeting the required accuracy. A close look at the example shows that $d_7 = 0$, which violates our theoretical assumption $d_i \neq 0, i = 1, \ldots, N$. To avoid such a degeneracy in Example 4.4, we increase the value y_9 from 4 to 4.1; Algorithm 3.2 now finds an approximate solution within accuracy 10^{-13}, but using a relatively large number of Newton steps (≥ 20). When we further increase the value y_9 to 5, the number of Newton steps needed for the assumed tolerance is reduced considerably. These observations indicate that how far away from zero each divided difference is may make a big difference in the numerical performance of the algorithm. This is perhaps related to a property that can be regarded as conditioning. The problem is, however, nonsmooth, and here we are entering a new territory.

REFERENCES

[1] L.-E. Andersson and T. Elfving, *An algorithm for constrained interpolation,* SIAM J. Sci. Statist. Comput., 8 (1987), pp. 1012–1025.

[2] F. H. Clarke, *Optimization and Nonsmooth Analysis,* John Wiley and Sons, New York, 1983; reprinted by SIAM, Philadelphia, 1990.

[3] T. De Luca, F. Facchinei, and C. Kanzow, *A theoretical and numerical comparison of some semismooth algorithms for complementarity problems,* Comput. Optim. Appl., 16 (2000), pp. 173–205.

[4] A. L. Dontchev and B. D. Kalchev, *Duality and well-posedness in convex interpolation,* Numer. Funct. Anal. Optim., 10 (1989), pp. 673–689.

[5] A. L. Dontchev, H.-D. Qi, and L. Qi, *Convergence of Newton's method for convex best interpolation,* Numer. Math., 87 (2001), pp. 435–456.

[6] A. L. Dontchev, H.-D. Qi, and L. Qi, *Quadratic convergence of Newton's method for convex interpolation and smoothing,* Constr. Approx., to appear.

[7] F. Facchinei, A. Fischer, and C. Kanzow, *Inexact Newton methods for semismooth equations with applications to variational inequality problems,* in Nonlinear Optimization and Applications, G. Di Pillo and F. Giannessi, eds., Plenum Press, New York, 1996, pp. 125–139.

[8] A. Fischer, *Solution of monotone complementarity problems with locally Lipschitzian functions,* Math. Program., 76 (1997), pp. 513–532.

[9] S. Fredenhagen, H. J. Oberle, and G. Opfer, *On the construction of optimal monotone cubic spline interpolations,* J. Approx. Theory, 96 (1999), pp. 182–201.

[10] G. L. Iliev, *Numerical methods under interpolation with restrictions and their convergence method of Newton,* C. R. Acad. Bulgare Sci., 40 (1987), pp. 37–40.

[11] L. D. Irvine, S. P. Marin, and P. W. Smith, *Constrained interpolation and smoothing,* Constr. Approx., 2 (1986), pp. 129–151.

[12] H. Jiang and L. Qi, *Local uniqueness and Newton-type methods for nonsmooth variational inequalities,* J. Math. Anal. Appl., 196 (1995), pp. 314–331.

[13] M. Kojima and S. Shindo, *Extension of Newton and quasi-Newton methods to systems of PC^1 equations*, J. Oper. Res. Soc. Japan, 29 (1986), pp. 352–375.

[14] J. M. Martínez and L. Qi, *Inexact Newton methods for solving nonsmooth equations,* J. Comput. Appl. Math., 60 (1995), pp. 127–145.

[15] C. A. Micchelli, P. W. Smith, J. Swetits, and J. D. Ward, *Constrained L_p approximation,* Constr. Approx., 1 (1985), pp. 93–102.

[16] L. Qi and J. Sun, *A nonsmooth version of Newton's method,* Math. Program., 58 (1993), pp. 353–367.

[17] R. T. Rockafellar, *Some properties of piecewise smooth functions,* Comput. Optim. Appl., to appear.

SIAM J. OPTIM.
Vol. 13, No. 2, pp. 603–618

ON THE CALMNESS OF A CLASS OF MULTIFUNCTIONS*

RENÉ HENRION†, ABDERRAHIM JOURANI‡, AND JIŘI OUTRATA§

Dedicated to Jochem Zowe on the occasion of his sixtieth birthday

Abstract. The paper deals with the calmness of a class of multifunctions in finite dimensions. Its first part is devoted to various conditions for calmness, which are derived in terms of coderivatives and subdifferentials. The second part demonstrates the importance of calmness in several areas of nonsmooth analysis. In particular, we focus on nonsmooth calculus and solution stability in mathematical programming and in equilibrium problems. The derived conditions find a number of applications there.

Key words. calmness, multifunctions, constraint qualifications, nonsmooth calculus, solution stability, equilibrium problems, weak sharp minima

AMS subject classifications. 90C31, 26E25, 49J52

PII. S1052623401395553

1. Introduction.

The concept of calmness plays a key role in the analysis of Lipschitz properties for multifunctions. It is closely related to issues from optimization theory like nondegenerate multiplier rules (e.g., [10], [2], [4]), existence of error bounds (e.g., [5], [18], [24]), or sensitivity analysis of generalized equations (e.g., [13], [17]). The aim of this paper is to provide subdifferential conditions for ensuring the calmness of constraint systems in finite dimensions and to consider calmness in the context of different applications like nonsmooth calculus or solutions to parametric optimization or equilibrium problems.

We start by recalling some of the prominent Lipschitz properties formulated for multifunctions. Let $M : Y \rightrightarrows X$ be a multifunction between metric spaces. M is said to have the *Aubin property* around some $(\bar{y}, \bar{x}) \in \operatorname{Gph} M$ (graph of M) if there exist neighborhoods $\mathcal{V}$ and $\mathcal{U}$ of $\bar{y}$ and $\bar{x}$ as well as some $L > 0$ such that

$$d(x, M(y_2)) \leq L d(y_1, y_2) \quad \forall y_1, y_2 \in \mathcal{V},\ \forall x \in M(y_1) \cap \mathcal{U}.$$

It is well known that M has the Aubin property around $(\bar{y}, \bar{x})$ if and only if its inverse M^{-1} is metrically regular around $(\bar{x}, \bar{y})$ (e.g., [27, Theorem 9.43]). Fixing one of the y-parameters as $\bar{y}$ in the definition of the Aubin property yields the calmness of M at $(\bar{y}, \bar{x})$:

$$d(x, M(\bar{y})) \leq L d(y, \bar{y}) \quad \forall y \in \mathcal{V},\ \forall x \in M(y) \cap \mathcal{U}.$$

Obviously, the Aubin property implies calmness, whereas the converse is not true (e.g., $M(y) = \{x | x^2 \geq y\}$ at $(0, 0)$). If one may choose $\mathcal{U} = X$ in this last definition, then the calmness becomes the slightly stronger local upper Lipschitz property introduced in [25].

*Received by the editors September 25, 2001; accepted for publication January 17, 2002; published electronically October 1, 2002.
http://www.siam.org/journals/siopt/13-2/39555.html
†Weierstrass Institute Berlin, 10117 Berlin, Germany (henrion@wias-berlin.de).
‡Département de Mathématiques, Université de Bourgogne, 21078 Dijon, France (Abderrahim.Jourani@u-bourgogne.fr).
§Institute of Information Theory and Automation, 18208 Prague, Czech Republic (outrata@utia.cas.cz). The research of this author was supported by grant 201/00/0080 of the Grant Agency of the Czech Republic.

A restricted version of calmness, namely, *calmness on selections,* has been studied in the context of sensitivity analysis for generalized equations [13], [15], [6]. Here it is required that $\mathcal{U} \cap M(\bar{y}) = \{\bar{x}\}$ in the general definition of calmness, i.e., $\bar{x}$ is isolated in $M(\bar{y})$. Such an assumption is relevant, for instance, when analyzing solutions to nonlinear optimization problems. Moreover, one may even further restrict calmness by combining it with the local uniqueness of M at $(\bar{y}, \bar{x})$. Then, locally around $(\bar{y}, \bar{x})$, M is just a usual function satisfying the condition

$$d(M(y), M(\bar{y})) \leq Ld(y, \bar{y}).$$

This situation was studied, for instance, in [16].

For the purpose of verifying the Lipschitz properties of multifunctions, it is useful to have suitable criteria from nonsmooth calculus. Such criteria have proven to be particularly efficient in finite dimensions. For instance, X and Y being finite-dimensional, the Aubin property of a closed graph multifunction M is equivalent to the condition (see [21])

$$D^*M(\bar{y}, \bar{x})(0) = \{0\}. \tag{1.1}$$

Here, D^* refers to Mordukhovich's coderivative (see section 2). This is a dual criterion that relies on a normal cone construction to the graph of M. Similar dual conditions were given in [20, Theorem 5.4] for a property related to but different from calmness.

An equivalent primal criterion for the Aubin property can be formulated in terms of the contingent derivative D, based on the contingent cone to $\operatorname{Gph} M$ (see [1, Theorem 4, p. 431] for sufficiency in arbitrary Banach spaces and, e.g., [7, Corollary 1.19] for necessity in the case of finite-dimensional X):

$$\exists\, \alpha > 0, \beta > 0 :\ B(0,1) \subseteq [D\,M(y,x)]^{-1}(B(0,\alpha)) \quad \forall (y,x) \in \operatorname{Gph} M \cap B((\bar{y},\bar{x}),\beta).$$

Here, $B(z, r)$ refers to the closed ball around z with radius r. As far as corresponding criteria for calmness are concerned, the following primal condition was found to be sufficient in [13, Proposition 2.1] and necessary in [15, Proposition 4.1] for calmness on selections in finite dimensions:

$$DM(\bar{y}, \bar{x})(0) = \{0\}. \tag{1.2}$$

Note that this condition immediately enforces the isolatedness of $\bar{x}$ in $M(\bar{y})$ because a sequence $x_n \to \bar{x}$, $x_n \in M(\bar{y}), x_n \neq \bar{x}$ would generate a nontrivial tangent vector $(0, \xi)$ to $\operatorname{Gph} M$ at $(\bar{y}, \bar{x})$, whence a contradiction $0 \neq \xi \in DM(\bar{y}, \bar{x})(0)$ to the above condition.

Calmness in the broader sense introduced above is closely related to the regularity concept of Ioffe studied in [10], [11], even in a Banach space setting. In fact, in [11] a sufficient condition for calmness has been derived for multifunctions of the type

$$M(y) = \{x \in C | g(x) = y\} \tag{1.3}$$

in terms of Clarke's subdifferential. Another sufficient condition for calmness in the broader sense was given in [8] for multifunctions of the type

$$M(y) = \{x \in C | g(x) + y \in D\}, \tag{1.4}$$

where $g : \mathbb{R}^k \to \mathbb{R}^m$ is locally Lipschitz and $C \subseteq \mathbb{R}^k$, $D \subseteq \mathbb{R}^m$ are closed. It was shown there that under mild assumptions the calmness of M is implied by the condition

$$\bigcup_{y^* \in N_D(g(\bar{x})) \setminus \{0\}} D^*g(\bar{x})(y^*) \cap (-\operatorname{bd} N_C(\bar{x})) = \emptyset, \tag{1.5}$$

where "bd" refers to the topological boundary. Recalling that the criterion (1.1) for the Aubin property reduces in the special setting of (1.4) to the sufficient condition

$$\bigcup_{y^* \in N_D(g(\bar{x}))\setminus\{0\}} D^* g(\bar{x})(y^*) \cap (-N_C(\bar{x})) = \emptyset, \tag{1.6}$$

the reduction from the stronger Aubin to the weaker calmness property in (1.4) is reflected by a transition from a normal cone to its boundary in the criteria (1.5) and (1.6), respectively. Under some additional regularity assumptions, one may even pass to the boundary in the left part of (1.6). In [9], attempts were made to extend these ideas to the infinite-dimensional case, but it seems to be difficult to pass beyond convex or differentiable structures in this framework. For instance, if f is a locally Lipschitz function, regular in the sense of Clarke and satisfying $f(0) = 0$, then the condition $0 \notin \operatorname{bd} \partial f(0)$ guarantees calmness of the parametric inequality $f(x) \leq y$ at $(0, 0)$ as long as either f is defined on a finite-dimensional space [8, Theorem 4.2] or f is convex on a Banach space [9, Corollary 3.4]. In contrast, one may construct a locally Lipschitz f defined on the sequence space l^1 which is Clarke regular and nonconvex such that the mentioned condition is satisfied but calmness fails to hold.

The paper is organized as follows: first, subdifferential criteria for calmness in finite dimensions are developed which extend those given in [8]. In particular, the multifunction M in (1.4) gets the more general form $M(y) = S(y) \cap C$, with a purely parametric contribution by S and a nonparametric contribution by C. In a second part, calmness (as a condition by itself or implied by the previously derived subdifferential criteria) is studied in several applications like nonsmooth calculus, stability of solutions to nonsmooth optimization problems, and equilibrium problems.

2. Notation and basic concepts. In the following, we denote by $\partial f(x)$ and $N_C(x)$, respectively, the subdifferential of a function f at some x and the normal cone to some closed set C at some $x \in C$, both in the sense of Mordukhovich. In contrast, $T_C(x)$ refers to the contingent cone. Note that if f is regular in the sense of Clarke, then $\partial f(x)$ coincides with Clarke's subdifferential. Similarly, if C is a regular set at x, then $T_C(x)$ and $N_C(x)$ coincide with Clarke's tangent and normal cone, respectively. In that case it also holds true that each one of these cones is the (negative) polar cone of the other. With a multifunction $Z : \mathbb{R}^p \rightrightarrows \mathbb{R}^k$ and some $(\bar{u}, \bar{v}) \in \operatorname{Gph} Z$ we associate Mordukhovich's coderivative $D^* Z(\bar{u}, \bar{v}) : \mathbb{R}^k \rightrightarrows \mathbb{R}^p$ defined by

$$D^* Z(\bar{u}, \bar{v})(v^*) = \{u^* \in \mathbb{R}^p | (u^*, -v^*) \in N_{\operatorname{Gph} Z}(\bar{u}, \bar{v})\}.$$

If Z is single-valued, we simply write $D^* Z(\bar{u})$ instead of $D^* Z(\bar{u}, Z(\bar{u}))$. For single-valued, locally Lipschitz mappings Z it holds that

$$D^* Z(\bar{u})(v^*) = \partial \langle v^*, Z \rangle (\bar{u}).$$

For a detailed presentation of these concepts, we refer to [20], [22], [27] and [4].

By $B(x, r)$, $\mathbb{B}$, and $\mathbb{S}$ we shall denote a closed ball centered at x with radius r, the closed unit ball, and the unit sphere in corresponding spaces, respectively. By $d(x, C)$ we denote the point-to-set distance between x and C induced by a corresponding norm on $\mathbb{R}^n$, whereas $d_C^e(x)$ represents the particular case of the Euclidean distance function.

A basic concept which we shall use in the derivation of subdifferential criteria for calmness is *semismoothness* as introduced in [19].

DEFINITION 2.1. *A function $\psi : \mathbb{R}^k \to \mathbb{R}$ is called semismooth at $\bar{x} \in \mathbb{R}^k$ if it is locally Lipschitz around $\bar{x}$ and the following property holds true: for each $d \in \mathbb{R}^k$ and for any sequences $t_n \downarrow 0$, $d_n \to d$, $x_n^* \in \partial\psi(\bar{x} + t_n d_n)$, the limit $\lim_{n\to\infty} \langle x_n^*, d\rangle$ exists.*

It has to be noted that in the original definition of [19], the corresponding property was required for Clarke's subdifferential of ψ. However, exploiting the well-known fact that Clarke's subdifferential is the closed convex hull of Mordukhovich's, it easily follows that both definitions of semismoothness are equivalent. As a consequence of Definition 2.1, a semismooth function ψ has a conventional directional derivative $\psi'(\bar{x}; h)$ at $\bar{x}$ in direction d which coincides with the common limit in Definition 2.1.

As with Clarke regularity, semismoothness of functions can be carried over to sets.

DEFINITION 2.2. *A set $A \subseteq \mathbb{R}^k$ is called semismooth at $\bar{x} \in \mathrm{cl}A$ if for any sequence $x_n \to \bar{x}$ with $x_n \in A$ and $\|x_n - \bar{x}\|^{-1}(x_n - \bar{x}) \to d$ it holds that $\langle x_n^*, d\rangle \to 0$ for all selections of subgradients $x_n^* \in \partial d_A^e(x_n)$.*

If A is closed and d_A^e is semismooth in the sense of Definition 2.1, then A is semismooth in the sense of Definition 2.2 (see [8, Proposition 2.4]).

3. Subdifferential characterization of calmness. We start with an auxiliary result which is crucial for passing to the boundary of the normal cone in (1.5) and in the corresponding generalization we have in mind.

PROPOSITION 3.1. *Let $C \subseteq \mathbb{R}^k$ be regular (in the sense of Clarke) and semismooth at $\bar{x} \in C$. Consider a sequence $x_n \to \bar{x}$ such that $x_n \in C$ and $\|x_n - \bar{x}\|^{-1}(x_n - \bar{x}) \to h$ with $\|h\| = 1$. Then each accumulation point x^* of a sequence $x_n^* \in \partial d_C^e(x_n)$ belongs to $\mathrm{bd}\, N_C(\bar{x})$.*

Proof. By virtue of the semismoothness of C at $\bar{x}$, one has $\langle x^*, h\rangle = 0$. From $\partial d_C^e(x_n) \subseteq N_C(x_n)$ and from the closedness of the mapping $N_C(\cdot)$, it follows that $x^* \in N_C(\bar{x})$. By construction, $h \in T_C(\bar{x})$; hence regularity of C at $\bar{x}$ implies that $\langle y^*, h\rangle \le 0$ for all $y^* \in N_C(\bar{x})$. For arbitrary $\varepsilon > 0$, one has $\langle x^* + \varepsilon h, h\rangle = \varepsilon > 0$, whence $x^* + \varepsilon h \notin N_C(\bar{x})$. Along with $x^* \in N_C(\bar{x})$, this means that $x^* \in \mathrm{bd}\, N_C(\bar{x})$. □

Consider now a multifunction $M : \mathbb{R}^p \rightrightarrows \mathbb{R}^k$ defined as the intersection $M(y) = S(y) \cap C$, where $S : \mathbb{R}^p \rightrightarrows \mathbb{R}^k$ is a multifunction with closed graph and $C \subseteq \mathbb{R}^k$ is closed. As a consequence, M has closed graph as well. The following theorem is the main result of this section.

THEOREM 3.2. *Consider some $(\bar{y}, \bar{x}) \in \mathrm{Gph}\, M$. Assume that C is regular and semismooth at $\bar{x}$. If for all $y^* \in \mathbb{R}^p$ it holds that*

$$D^*S^{-1}(\bar{x}, \bar{y})(y^*) \cap -\mathrm{bd}\, N_C(\bar{x}) = \begin{cases} \emptyset & \text{or} \\ \{0\} & \text{if } y^* = 0, \end{cases} \tag{3.1}$$

then M is calm at $(\bar{y}, \bar{x})$. (Note that the case $D^*S^{-1}(\bar{x}, \bar{y})(0) \cap -\mathrm{bd}\, N_C(\bar{x}) = \emptyset$ is formally included in (3.1).)

Proof. Assume by contradiction that M is not calm at $(\bar{y}, \bar{x})$. By definition, there exist sequences $x_n \to \bar{x}$, $y_n \to \bar{y}$, $x_n \in M(y_n)$ such that $d(x_n, M(\bar{y})) > n\|y_n - \bar{y}\|$. Now, set $h(y, x) := \|y - \bar{y}\|$ so that each pair (y_n, x_n) is an ε-minimizer of $h(y, x)$ over $\mathrm{Gph}\, M$ with $\varepsilon = \|y_n - \bar{y}\|$. The application of the Ekeland variational principle with ε and $\lambda := n\varepsilon$ to the minimization of h over $\mathrm{Gph}\, M$ yields for each n the existence of a pair $(\tilde{y}_n, \tilde{x}_n) \in \mathrm{Gph}\, M$ such that for all $(y, x) \in \mathrm{Gph}\, M$

$$\|(\tilde{y}_n, \tilde{x}_n) - (y_n, x_n)\| \le n\|y_n - \bar{y}\|, \tag{3.2}$$

$$\|\tilde{y}_n - \bar{y}\| \leq \|y - \bar{y}\| + n^{-1}\|(y,x) - (\tilde{y}_n, \tilde{x}_n)\|. \tag{3.3}$$

From (3.2) we infer that

$$\|(\tilde{y}_n, \tilde{x}_n) - (\bar{y}, \bar{x})\| \leq n\|y_n - \bar{y}\| + \|(y_n, x_n) - (\bar{y}, \bar{x})\| < d(x_n, M(\bar{y})) + \|(y_n, x_n) - (\bar{y}, \bar{x})\|$$

so that $(\tilde{y}_n, \tilde{x}_n) \to (\bar{y}, \bar{x})$. Furthermore, $\tilde{y}_n \neq \bar{y}$ and $\tilde{x}_n \neq \bar{x}$, because otherwise $\tilde{x}_n \in M(\bar{y})$, whence the contradiction

$$n\|y_n - \bar{y}\| < d(x_n, M(\bar{y})) \leq \|x_n - \tilde{x}_n\| \leq n\|y_n - \bar{y}\|,$$

using (3.2). Now, (3.3) means that $(\tilde{y}_n, \tilde{x}_n)$ is a (global) solution of the problem

$$\min\{\|y - \bar{y}\| + n^{-1}\|(y,x) - (\tilde{y}_n, \tilde{x}_n)\| \,|\, (y,x) \in \operatorname{Gph} M\}. \tag{3.4}$$

Since $\operatorname{Gph} M = \operatorname{Gph} S \cap (\mathbb{R}^p \times C)$, it follows that exactly one of the following cases occurs (with $\mathbb{S}$ denoting the unit sphere):

$$\{0\} = N_{\operatorname{Gph} S}(\tilde{y}_n, \tilde{x}_n) \cap [\{0\} \times (-N_C(\tilde{x}_n))], \tag{3.5}$$

$$\exists\, \xi_n \in \mathbb{S} \cap N_{\operatorname{Gph} S}(\tilde{y}_n, \tilde{x}_n) \cap [\{0\} \times (-N_C(\tilde{x}_n))]. \tag{3.6}$$

At least one of these two cases must apply for infinitely many n. Suppose first that this is true for (3.5). Without loss of generality, we assume that (3.5) is valid for all n. Then (see [27, Theorem 6.4.2])

$$N_{\operatorname{Gph} M}(\tilde{y}_n, \tilde{x}_n) \subseteq N_{\operatorname{Gph} S}(\tilde{y}_n, \tilde{x}_n) + [\{0\} \times N_C(\tilde{x}_n)].$$

Application of the necessary optimality conditions to the solution $(\tilde{y}_n, \tilde{x}_n)$ of the constrained problem (3.4) then yields

$$0 \in [\mathbb{S}_y \times \{0\}] + n^{-1}\mathbb{B} + N_{\operatorname{Gph} S}(\tilde{y}_n, \tilde{x}_n) + [\{0\} \times N_C(\tilde{x}_n)],$$

where $\mathbb{S}_y$ refers to the unit sphere in $\mathbb{R}^p$ (and occurs due to $\tilde{y}_n \neq \bar{y}$) and $\mathbb{B}$ is the unit ball in $\mathbb{R}^p \times \mathbb{R}^k$. Without loss of generality, $\mathbb{B}$ is taken with respect to the maximum norm; hence $\mathbb{B} = \mathbb{B}_y \times \mathbb{B}_x$. Accordingly, there exist $(y_n^*, z_n^*) \in N_{\operatorname{Gph} S}(\tilde{y}_n, \tilde{x}_n)$ and $x_n^* \in -N_C(\tilde{x}_n)$ such that

$$0 \in \mathbb{S}_y + n^{-1}\mathbb{B}_y + y_n^* \quad \text{and} \quad \|x_n^* - z_n^*\| \leq n^{-1}.$$

By the boundedness of y_n^* we may assume that $y_n^* \to y^* \in \mathbb{S}_y$.

If $\{x_n^*\}$ is unbounded, then for $\hat{x}_n^* := \|x_n^*\|^{-1} x_n^*$ we may assume that $\hat{x}_n^* \to x^*$ for some $x^* \in \mathbb{S}_x$. Furthermore, $\hat{x}_n^* \in -N_C(\tilde{x}_n)$ and

$$d^e_{N_{\operatorname{Gph} S}(\tilde{y}_n, \tilde{x}_n)}(\|x_n^*\|^{-1} y_n^*, \hat{x}_n^*) \leq d^e_{N_{\operatorname{Gph} S}(\tilde{y}_n, \tilde{x}_n)}(y_n^*, x_n^*) \leq \rho \|x_n^* - z_n^*\| \leq \rho n^{-1},$$

where d^e denotes the Euclidean distance function and $\rho > 0$ is some modulus relating the Euclidean and maximum norms. Since, without loss of generality, $\|x_n^*\|^{-1} y_n^* \to 0$, the closedness of the coderivative mapping implies that $x^* \in D^*S^{-1}(\bar{x}, \bar{y})(0)$. On the other hand, $\hat{x}_n^* \in -N_C(\tilde{x}_n) \cap \mathbb{B}_x = -\partial d_C^e(\tilde{x}_n)$ (see [27, Example 8.5.3]). Recalling that $\tilde{x}_n \neq \bar{x}$ and $\tilde{x}_n \in C$, Proposition 3.1 provides that $x^* \in -\operatorname{bd} N_C(\bar{x})$, whence the contradiction $x^* \in \mathbb{S}_x \cap D^*S^{-1}(\bar{x}, \bar{y})(0) \cap -\operatorname{bd} N_C(\bar{x})$ with (3.1).

Assuming that $\{x_n^*\}$ is bounded instead, one has without loss of generality that

$$x_n^* \to x^* \in D^*S^{-1}(\bar{x}, \bar{y})(y^*) \cap -N_C(\bar{x})$$

(again by closedness of the coderivative and of the normal cone mapping). Due to $\tilde{x}_n \neq \bar{x}$ we have that $T_C(\bar{x}) \neq \{0\}$, whence $N_C(\bar{x}) \neq \mathbb{R}^k$ and $0 \in -\mathrm{bd}\, N_C(\bar{x})$. Now, the case $x^* = 0$ leads to an immediate contradiction with (3.1) due to $y^* \neq 0$. If $x^* \neq 0$, then set

$$\hat{x}_n^* := \|x_n^*\|^{-1} x_n^* \to \hat{x}^* := \|x^*\|^{-1} x^*,$$

as before. Invoking Proposition 3.1 in the same way as above, one arrives at $\hat{x}^* \in \mathbb{S}_x \cap D^*S^{-1}(\bar{x},\bar{y})(\|x^*\|^{-1} y^*) \cap -\mathrm{bd}\, N_C(\bar{x})$ by positive homogeneity of the coderivative mapping. This again is a contradiction with (3.1).

Finally, suppose instead that (3.6) applies for infinitely many n. Again, we do not relabel the corresponding subsequence. Then, defining $\xi_n = (\xi_n^y, \xi_n^x)$, we may assume without loss of generality that $\xi_n = (0, \xi_n^x) \to (0, \xi^x)$, where $\xi_n^x, \xi^x \in \mathbb{S}_x$ and, according to (3.6),

$$\xi_n^x \in D^*S^{-1}(\tilde{x}_n, \tilde{y}_n)(0) \cap -N_C(\tilde{x}_n).$$

Consequently, $\xi_n^x \in -\partial d_C^e(\tilde{x}_n)$, and we may invoke Proposition 3.1 again to obtain that $\xi^x \in -\mathrm{bd}\, N_C(\bar{x})$. Summarizing, we arrive at the contradiction

$$\xi^x \in D^*S^{-1}(\bar{x},\bar{y})(0) \cap -\mathrm{bd}\, N_C(\bar{x})$$

with (3.1). □

REMARK 3.3. *The assumptions of (Clarke-) regularity and semismoothness for C in Theorem* 3.2 *are completely independent (see Example* 3.5 *in* [8]*). Their joint validity is guaranteed for a sufficiently broad class of closed sets, like convex sets or sets defined by $\mathcal{C}^1$-inequalities and satisfying the Mangasarian–Fromovitz constraint qualification (cf. Lemma* 3.6 *in* [8]*).*

Now, we specialize the above theorem to the parametrized constraint system $x \in C,\ g(x,y) \in D$, where $g : \mathbb{R}^k \times \mathbb{R}^p \to \mathbb{R}^m$ is locally Lipschitz and $C \subseteq \mathbb{R}^k$, $D \subseteq \mathbb{R}^m$ are closed. We associate with this system the multifunction $M : \mathbb{R}^p \rightrightarrows \mathbb{R}^k$ defined by

$$M(y) := \{x \in C \,|\, g(x,y) \in D\}. \tag{3.7}$$

COROLLARY 3.4. *In* (3.7)*, let $(\bar{y},\bar{x}) \in \mathrm{Gph}\, M$ and C be regular and semismooth at $\bar{x}$. Further, assume the qualification condition*

$$\bigcup_{y^* \in N_D(g(\bar{x},\bar{y}))\setminus\{0\}} [\partial\langle y^*, g\rangle(\bar{x},\bar{y})]_x \cap -\mathrm{bd}\, N_C(\bar{x}) = \emptyset, \tag{3.8}$$

where $[\,]_x$ denotes projection onto the x-component. Then M is calm at $(\bar{y},\bar{x})$.

Proof. The case in which $0 \notin \mathrm{bd}\, N_C(\bar{x})$ is trivial, so assume that $0 \in \mathrm{bd}\, N_C(\bar{x})$. Consider the map $S : \mathbb{R}^p \rightrightarrows \mathbb{R}^k$ defined by

$$S(y) := \{x \in \mathbb{R}^k \mid g(x,y) \in D\}.$$

To compute $D^*S^{-1}(\bar{x},\bar{y})$, we invoke a result from [22]. Since $0 \in \mathrm{bd}\, N_C(\bar{x})$, (3.8) yields in particular the implication

$$D^*g(\bar{x},\bar{y})(v^*) = 0, \qquad v^* \in N_D(g(\bar{x},\bar{y})) \Longrightarrow v^* = 0.$$

This is, however, the qualification condition from [22, Theorem 6.10], and so one has for each $v^* \in \mathbb{R}^p$ the inclusion

$$\begin{aligned} D^*S^{-1}(\bar{x},\bar{y})(v^*) &\subseteq \{x^* \in \mathbb{R}^k \mid (x^*,-v^*) \in \partial\langle y^*, g\rangle(\bar{x},\bar{y}),\ y^* \in N_D(g(\bar{x},\bar{y}))\} \\ &\subseteq \{x^* \in [\partial\langle y^*, g\rangle(\bar{x},\bar{y})]_x \mid y^* \in N_D(g(\bar{x},\bar{y}))\}. \end{aligned} \tag{3.9}$$

Let us write (3.8) in the form

$$[\partial\langle y^*, g\rangle(\bar{x},\bar{y})]_x \cap -\mathrm{bd}\, N_C(\bar{x}) \neq \emptyset, \qquad y^* \in N_D(g(\bar{x},\bar{y})) \Longrightarrow y^* = 0. \tag{3.10}$$

By combining (3.9) and (3.10), one obtains that

$$D^*S^{-1}(\bar{x},\bar{y})(0) \cap -\mathrm{bd}\, N_C(\bar{x}) = \{0\}$$

and

$$D^*S^{-1}(\bar{x},\bar{y})(v^*) \cap -\mathrm{bd}\, N_C(\bar{x}) \neq \emptyset \Longrightarrow v^* = 0.$$

These two conditions amount, however, to (3.1), and thus Theorem 3.2 can be applied to finish the proof. □

The following example illustrates the application of Theorem 3.2 in the specific situation of Corollary 3.4.

EXAMPLE 3.5. *Define M in* (3.7) *by $C = \{(x_1,x_2) | x_2 \geq |x_1|\}$, $D = \mathbb{R}_-$, $g(x,y) = \min\{x_1,x_2\} - y$. Then, all data assumptions of Theorem* 3.2 *are satisfied at $(\bar{x}_1,\bar{x}_2,\bar{y}) = (0,0,0) \in \mathrm{Gph}M$, and also* (3.1) *holds true:*

$$\begin{aligned} \bigcup_{y^* \in N_D(g(\bar{x},\bar{y}))\setminus\{0\}} [\partial\langle y^*, g\rangle(\bar{x},\bar{y})]_x \cap -\mathrm{bd}\, N_C(\bar{x}) &= \bigcup_{y^*>0} y^* \partial \min\{\cdot,\cdot\}(0,0) \cap \mathrm{bd}\, C \\ = \{(x_1,x_2) | x_1 + x_2 > 0,\ x_1 \cdot x_2 = 0\} \cap \mathrm{Gph}\,|\cdot| &= \emptyset. \end{aligned}$$

Consequently, the calmness of M in (3.7) *can be derived. Note that the stronger criterion* (1.6) *ensuring the Aubin property of M fails to apply here due to*

$$\{(x_1,x_2) | x_1 + x_2 > 0, x_1 \cdot x_2 = 0\} \cap -N_C(\bar{x}) = \{(0,x_2) | x_2 > 0\} \neq \emptyset.$$

At the same time, the contingent derivative criterion (1.2) *for calmness on selections does not apply either, due to $M(0) = \{(x_1,x_2) | x_2 \geq -x_1 \geq 0\}$ not being single-valued.*

The following theorem provides a calmness result for the system (1.4) of functional constraints with canonical perturbations. In contrast to Theorem 3.2, no regularity or semismoothness assumption on C will be made. Rather, the regularity assumption is shifted to the perturbed part of the constraints.

THEOREM 3.6. *In* (1.4) *let g be Lipschitz near $\bar{x} \in M(0)$, and D be regular at $g(\bar{x})$. Further assume that the function $\langle y^*, g\rangle(\cdot)$ is regular at $\bar{x}$ for all $y^* \in \partial d_D^e(g(\bar{x}))$ and that the qualification condition*

$$\mathrm{int} \bigcup_{y^* \in N_D(g(\bar{x})) \cap \mathbb{B}} \partial\langle y^*, g\rangle(\bar{x}) \cap -[T_C(\bar{x})]^0 \neq \emptyset \tag{3.11}$$

holds true. Then M is calm at $(0,\bar{x})$.

Proof. Consider the composite function $\pi(x) = d^e_D(g(\bar{x}))$, which is evidently Lipschitz near $\bar{x}$ and for which one has $\pi(\bar{x}) = 0$. From [27, Theorem 10.49] we know that under our assumptions π is even regular at $\bar{x}$ and

$$\partial\pi(\bar{x}) = \bigcup_{y^* \in N_D(g(\bar{x})) \cap \mathbb{B}} \partial\langle y^*, g\rangle(\bar{x}). \tag{3.12}$$

From (3.11) and (3.12) we infer the existence of some $\tilde{z}^* \in -[T_C(\bar{x})]^0$ and of some $\alpha > 0$ such that $B(\tilde{z}^*, \alpha) \subseteq \partial\pi(\bar{x})$. Then, regularity of π at $\bar{x}$ implies that $\langle \tilde{z}^* + \alpha p^*, h\rangle \leq \pi'(\bar{x}; h)$ for all $p^* \in \mathbb{B}$ and all $h \in \mathbb{R}^k$, where $\pi'(\bar{x}; h)$ refers to the conventional directional derivative of π taken at $\bar{x}$ in direction h. Consequently,

$$\alpha\,\langle p^*, h\rangle \leq \pi'(\bar{x}; h) - \langle \tilde{z}^*, h\rangle \leq \pi'(\bar{x}; h) \qquad \forall p^* \in \mathbb{B},\ \forall h \in T_C(\bar{x}).$$

For arbitrary $h \in T_C(\bar{x}) \cap \mathbb{S}$ we set $p^* := h$ and derive from the last relation that

$$\pi'(\bar{x}; h) \geq \alpha > 0 \quad \forall h \in T_C(\bar{x}) \cap \mathbb{S}. \tag{3.13}$$

Assume that M fails to be calm at $(0,\bar{x})$. Then, as in the proof of Theorem 3.2, there exist sequences $x_n \to \bar{x}$, $y_n \to 0$, $x_n \in M(y_n)$ such that $d(x_n, M(0)) > n\|y_n\|$. From here we deduce that $x_n \neq \bar{x}$, $x_n \in C$, and $\|x_n - \bar{x}\| > n(\pi(x_n) - \pi(\bar{x}))$ for all n. This amounts to $\|x_n - \bar{x}\|^{-1}(\pi(x_n) - \pi(\bar{x})) < n^{-1}$. It suffices now to pass to an appropriate subsequence $\{x_{n'}\}$ such that $\|x_{n'} - \bar{x}\|^{-1}(x_{n'} - \bar{x}) \to h$ for some $h \in T_C(\bar{x}) \cap \mathbb{S}$. Local Lipschitz continuity of π yields that $\pi'(\bar{x}; h) = 0$, which contradicts (3.13) and thus proves the calmness of M at $(0,\bar{x})$. ∎

REMARK 3.7. *From* (3.13) *it immediately follows that* (3.11) *implies not only the calmness of M at $(0,\bar{x})$ but also the isolatedness of $\bar{x}$ in $M(0)$, i.e., $\mathcal{U} \cap M(0) = \{\bar{x}\}$ for some neighborhood $\mathcal{U}$ of $\bar{x}$.*

Example 3.5 shows that the last remark does not apply to the setting of Theorem 3.2 or Corollary 3.4, where no regularity assumptions are made with respect to S or g.

4. Calmness in applications.

4.1. Nonsmooth calculus. As shown, e.g., in [2], [4], [28], calmness plays an important role in deriving optimality conditions and in construction of local Lipschitz error bounds. It enables us, among other things, to replace the constraint system

$$g(x) \in D, \quad x \in C, \tag{4.1}$$

by a more easily tractable constraint

$$(y, x) \in \operatorname{Gph} M,$$

where M is given by (3.7), and the new variable y enters the objective via a suitable penalty term. Clearly, the feasible set given by (4.1) amounts to $M(0)$. For the evaluation of the normal cone to $M(0)$ at a given point $\bar{x}$, one usually employs various constraint qualifications. A prominent place is occupied by the *Mangasarian–Fromovitz constraint qualification,* which in case of (4.1) becomes (1.6). Condition (1.6) ensures the Aubin property of M around $(0,\bar{x})$ and, a fortiori, the inclusion

$$N_{M(0)}(\bar{x}) \subset \bigcup_{y^* \in N_D(g(\bar{x}))} D^*g(\bar{x})\,(y^*) + N_C(\bar{x}). \tag{4.2}$$

It turns out, however, that the calmness of M at $(0,\bar{x})$ also implies (4.2), and therefore, at least in some cases, condition (1.6) can be weakened.

THEOREM 4.1. *Consider the multifunction M given by* (1.4) *and a pair* $(0,\bar{x}) \in \operatorname{Gph} M$. *Assume that g is Lipschitz near $\bar{x}$ and that M is calm at $(0,\bar{x})$. Then inclusion* (4.2) *holds true.*

Proof. We start with the observation that (see [22, Theorem 6.10])

$$N_{\operatorname{Gph} M}(0,\bar{x}) \subset \{(y^*,x^*) | y^* \in N_D(g(\bar{x})),\ x^* \in \partial\langle y^*, g\rangle(\bar{x}) + N_C(\bar{x})\}. \tag{4.3}$$

Let L be the modulus of calmness of M at $(0,\bar{x})$. We claim that

$$\forall x^* \in \partial d^e_{M(0)}(\bar{x}) \quad \exists y^* \in L\mathbb{B} :\ (y^*,x^*) \in N_{\operatorname{Gph} M}(0,\bar{x}). \tag{4.4}$$

To see this, note that $x^* \in \partial d^e_{M(0)}(\bar{x})$ means the existence of sequences $x_n \to \bar{x}$ $(x_n \in M(0))$, $r_n \downarrow 0$, $x^*_n \to x^*$, and $\varepsilon_n \downarrow 0$ such that

$$d^e_{M(0)}(x) - d^e_{M(0)}(\bar{x}) \geq \langle x^*_n, x - x_n\rangle - \varepsilon_n\|x - x_n\| \qquad \forall x \in B(x_n, r_n).$$

Since M is calm at $(0,\bar{x})$, along with $L > 0$ there exists some $r > 0$ such that

$$d^e_{M(0)}(x) \leq L\|y\| \qquad \forall x \in B(\bar{x},r) \cap M(y),\ \forall y \in B(0,r). \tag{4.5}$$

This implies that

$$\begin{aligned} &L\|y\| - \langle x^*_n, x - x_n\rangle + \varepsilon_n\|x - x_n\| \geq 0 \\ &\forall (y,x) \in \operatorname{Gph} M \cap (B(0,r) \times B(x_n, r_n)) \end{aligned} \tag{4.6}$$

for sufficiently large n. The function of (y,x) on the left-hand side of (4.6) attains a constrained minimum at $(0,x_n)$. According to Proposition 4.3.4 in [4], the function

$$L\|y\| - \langle x^*_n, x - x_n\rangle + \varepsilon_n\|x - x_n\| + K d^e_{\operatorname{Gph} M}(y,x)$$

attains an unconstrained local minimum at $(0,x_n)$ for sufficiently large penalty parameter K. The respective optimality conditions imply the existence of some $y^*_n \in L\mathbb{B}$ such that

$$0 \in \{-y^*_n\} \times (\{-x^*_n\} + \varepsilon_n\mathbb{B}) + N_{\operatorname{Gph} M}(0,x_n).$$

We now let n tend to infinity and, passing to a subsequence $\{y^*_{n'}\}$, establish the existence of a limit vector $y^* \in L\mathbb{B}$ such that $(y^*,x^*) \in N_{\operatorname{Gph} M}(0,\bar{x})$. This proves (4.4). It remains to observe that for each $\xi \in N_{M(0)}(\bar{x})$ there is some $x^* \in \partial d^e_{M(0)}(\bar{x}) = N_{M(0)}(\bar{x}) \cap \mathbb{B}$ such that $\xi = \|\xi\|\, x^*$. Since g is Lipschitz, the result follows from (4.3) and (4.4). ∎

COROLLARY 4.2. *In* (1.4), *let $k = m$ and $\bar{x} \in C \cap D$. Assume that the map*

$$\tilde{M}(y) := \{x \in C \,|\, x + y \in D\}$$

is calm at $(0,\bar{x})$. Then one has

$$N_{C\cap D}(\bar{x}) \subset N_C(\bar{x}) + N_D(\bar{x}). \tag{4.7}$$

Proof. It suffices to specialize the statement of Theorem 4.1 for g being the identity mapping. ∎

REMARK 4.3. *The calmness of* $\tilde{M}$ *at* $(0,\bar{x})$ *is closely related to the so-called metric inequality for the sets* C, D *at* $\bar{x}$ [12], *which also implies inclusion* (4.7).

In the literature (e.g., [20], [22]) one usually requires the qualification condition

$$N_D(\bar{x}) \cap -N_C(\bar{x}) = \{0\} \tag{4.8}$$

to ensure the validity of inclusion (4.7). However, condition (4.8) implies the Aubin property of $\tilde{M}$ around $(0,\bar{x})$ and is thus clearly more demanding than the calmness required in Corollary 4.2.

By combining Theorem 3.2 and the above corollary, we immediately conclude that, to ensure inclusion (4.7), it suffices to replace (4.8) by a weaker condition

$$N_D(\bar{x}) \cap -\mathrm{bd}\, N_C(\bar{x}) = \{0\} \tag{4.9}$$

whenever C is regular and semismooth at $\bar{x}$. Moreover, as observed by Kruger [14], condition (4.9) alone (without regularity or semismoothness assumptions) implies inclusion (4.7). The respective statement can be formulated even for a general mapping M permitting noncanonical perturbations.

PROPOSITION 4.4 (adapted from [14]). *Consider the map* M *given by* (3.7), *where* g *is Lipschitz around a reference pair* $(\bar{y},\bar{x}) \in \mathrm{Gph}\, M$ *and* C, D *are closed subsets of the respective spaces. Assume that* (3.8) *is fulfilled. Then either* M *possesses the Aubin property around* $(\bar{y},\bar{x})$ *or*

$$\bigcup_{y^* \in N_D(g(\bar{x},\bar{y}))\setminus\{0\}} [\partial\langle y^*, g\rangle(\bar{x},\bar{y})]_x + N_C(\bar{x}) = \mathbb{R}^p. \tag{4.10}$$

Proof. If

$$\bigcup_{y^* \in N_D(g(\bar{x},\bar{y}))\setminus\{0\}} [\partial\langle y^*, g\rangle(\bar{x},\bar{y})]_x \cap -N_C(\bar{x}) = \emptyset, \tag{4.11}$$

then it follows from [22, Theorem 6.10] that

$$\begin{aligned} D^*M(\bar{y},\bar{x})\,(x^*) \subset \{y^* \in \mathbb{R}^m \mid (y^*,-x^*) \in D^*g(\bar{x},\bar{y}) \circ N_D(g(\bar{x},\bar{y})) \\ + (0 \times N_C(\bar{x}))\}. \end{aligned} \tag{4.12}$$

Combining (4.11) and (4.12) provides $D^*M(\bar{y},\bar{x})\,(0) = \{0\}$, whence the Aubin property of M at $(\bar{y},\bar{x})$ (see (1.1)). According to (3.8), assume therefore that

$$\bigcup_{y^* \in N_D(g(\bar{x},\bar{y}))\setminus\{0\}} [\partial\langle y^*, g\rangle(\bar{x},\bar{y})]_x \cap -\mathrm{int}\, N_C(\bar{x}) \neq \emptyset. \tag{4.13}$$

Then

$$\exists y^* \in N_D(g(\bar{x},\bar{y})) \setminus \{0\},\ \exists x^* \in [\partial\langle y^*, g\rangle(\bar{x},\bar{y})]_x,\ \exists \alpha > 0 : \quad B(x^*,\alpha) \subset -N_C(\bar{x}).$$

This implies for each $p^* \in B(0,\alpha)$ that

$$p^* \in \bigcup_{y^* \in N_D(g(\bar{x},\bar{y}))\setminus\{0\}} [\partial\langle y^*, g\rangle(\bar{x},\bar{y})]_x + N_C(\bar{x}).$$

Now, the result follows. ∎

COROLLARY 4.5. *Let $C, D \subseteq \mathbb{R}^k$ be arbitrary closed sets with $\bar{x} \in C \cap D$. Then* (4.9) *ensures inclusion* (4.7).

Proof. Apply Proposition 4.4 with $g(x, y) := x$. □

According to the proof of Proposition 4.4, the difference between (3.8) and the classical Mangasarian–Fromovitz constraint qualification (4.11) reduces to the case (4.13), for which the argument from Remark 3.7 implies the local isolatedness of the feasible points of $M(0)$ (under the additional assumptions of Theorem 3.6). This fact is easily interpreted for mathematical programs of the form

$$\min\{f(x)|x \in M(0)\}. \tag{4.14}$$

Evidently, isolated points of $M(0)$ are automatically local minima; hence, in this context (3.8) goes beyond the Mangasarian–Fromovitz constraint qualification as a condition providing nondegenerate Lagrange multipliers, in that it identifies local minima given by isolated feasible points.

Another observation is the following: Since polyhedral mappings are automatically calm (cf. [26]), we derive from Theorem 4.1 that a nonsmooth calculus rule like (4.2) can be obtained under no constraint qualifications for polyhedral data.

4.2. First-order growth (weak sharp minima), local uniqueness, and stability of solutions. Consider the problem

$$(P) \qquad \min\{f(x)|x \in C\},$$

where $f : \mathbb{R}^k \to R$ is a continuous function and $C \subseteq \mathbb{R}^k$ a closed subset. Denote the solution set of (P) by S. Recall the following definition.

DEFINITION 4.6. *In (P), the objective function f is said to satisfy a first-order growth condition if there exist a constant $c > 0$ and a neighborhood $\mathcal{N}$ of S such that*

$$f(x) \geq f_* + cd(x, S) \qquad \forall x \in C \cap \mathcal{N},$$

where $f_* = \inf\{f(x)|x \in C\}$. Equivalently, f is said to have a set S of weak sharp minima with respect to $C \cap \mathcal{N}$ (cf. [3]*).*

LEMMA 4.7. *Let the solution set S of (P) be nonempty and bounded, and suppose that the multifunction $M(y) := \{x \in C|f(x) \leq y\}$ is calm on $\{f_*\} \times S$ (i.e., calm at all (f_*, x) with $x \in S$). Then, f satisfies a first-order growth condition in (P).*

Proof. Fix an arbitrary $x^0 \in S$. Obviously, $f(x^0) = f_*$; hence the calmness of M at $(f(x^0), x^0)$ implies the existence of $\varepsilon, \delta, L > 0$ such that

$$d(x, M(f(x^0))) \leq L|y - f(x^0)| \quad \forall y : |y - f(x^0)| < \delta,\ \forall x \in M(y) \cap B(x^0, \varepsilon).$$

Choose $\varepsilon > 0$ small enough to meet $|f(x) - f(x^0)| < \delta$ for all $x \in B(x^0, \varepsilon)$. Now, one may put $y := f(x)$ in the above estimation and derive from $M(f(x^0)) = S$ that

$$d(x, S) \leq L|f(x) - f(x^0)| \qquad \forall x \in C \cap B(x^0, \varepsilon).$$

From $f(x) \geq f(x^0)$ for all $x \in C$, it follows that

$$f(x) \geq f_* + L^{-1}d(x, S) \qquad \forall x \in C \cap B(x^0, \varepsilon).$$

By our assumptions, S is compact. Hence, a finite number of $x^i \in S$, $\varepsilon_i > 0$, and $L_i > 0$ exists such that $S \subseteq \cup_i B(x^i, \varepsilon_i)$ and

$$f(x) \geq f_* + L_i^{-1}d(x, S) \qquad \forall x \in C \cap B(x^{\forall i}, \varepsilon_i).$$

This, however, implies that f satisfies a first-order growth condition with $c := (\max L_i)^{-1}$ and $\mathcal{N} := \cup_i B(x^i, \varepsilon_i)$. □

COROLLARY 4.8. *In (P) let f be locally Lipschitz and C be regular and semismooth. Then f satisfies a first-order growth condition if the solution set S is nonempty and bounded and, moreover, the condition*

$$\partial f(x) \cap -\mathrm{bd}\, N_C(x) = \emptyset \qquad \forall x \in S$$

holds true.

Proof. Combine Lemma 4.7 with Corollary 3.4 (setting $g(x, y) := f(x) + y$ and $D := \mathbb{R}_-$ there). □

A consequence of the constraint qualification in the last corollary is that solutions are locally isolated, as described in the following.

PROPOSITION 4.9. *Let $f : \mathbb{R}^k \to \mathbb{R}$ be Lipschitz near $\bar{x} \in S$, and $C \subseteq \mathbb{R}^k$ be regular at $\bar{x}$. If, in addition, C* or *f is semismooth at $\bar{x}$, then the condition $\partial f(\bar{x}) \cap -\mathrm{bd}\, N_C(\bar{x}) = \emptyset$ entails that $\mathcal{U} \cap S = \{\bar{x}\}$ for some neighborhood $\mathcal{U}$ of $\bar{x}$.*

Proof. Assume, by contradiction, that $x_n \to \bar{x}$ for some sequence $x_n \in S \backslash \{\bar{x}\}$. Then, without loss of generality, $\|x_n - \bar{x}\|^{-1}(x_n - \bar{x}) \to h \in T_C(\bar{x})$. On the other hand, as $x_n \in S$, it follows that $f(x_n) = f(\bar{x})$ and $0 \in \partial f(x_n) + N_C(x_n)$. Accordingly, we may extract a sequence $y_n^* \in \partial f(x_n) \cap -N_C(x_n)$. This sequence is bounded because f is Lipschitz around $\bar{x}$. Hence, without loss of generality, $y_n^* \to y^*$ for some $y^* \in \partial f(\bar{x}) \cap -N_C(\bar{x})$. We claim that $y^* \in -\mathrm{bd}\, N_C(\bar{x})$, whence a contradiction to the assumed condition $\partial f(\bar{x}) \cap -\mathrm{bd}\, N_C(\bar{x}) = \emptyset$. Indeed, if C is semismooth at $\bar{x}$, this is an immediate consequence of Proposition 3.1. In the opposite case, the semismoothness of f at $\bar{x}$ provides that

$$\langle y_n^*, h\rangle \to \langle y^*, h\rangle = f'(\bar{x}; h) = \lim_{n\to\infty} \|x_n - \bar{x}\|^{-1} (f(x_n) - f(\bar{x})) = 0.$$

Now the same reasoning as in the proof of Proposition 3.1 allows us to derive that $y^* \in -\mathrm{bd}\, N_C(\bar{x})$. □

Evidently, Proposition 4.9 may be taken as a subdifferential condition for the local uniqueness of solutions. Now we are in a position to state a subdifferential condition for upper Lipschitz stability of solution sets. Consider the parametric optimization problem

$$P(y) \qquad \min\{f(x) | g(x) \le y\},$$

where $f : \mathbb{R}^k \to \mathbb{R}$ and $g : \mathbb{R}^k \to \mathbb{R}^m$ are locally Lipschitz, and $M(y)$ and $S(y)$ denote the parameter-dependent sets of feasible points and solutions, respectively. The set of active indices at x in the relation $g(x) \le y$ will be denoted by $I(x)$.

THEOREM 4.10. *Let $S(0)$ be nonempty and bounded, and assume the following conditions to hold true for all $x \in S(0)$:*

(1) *All components g_i of g are regular and semismooth at x.*

(2) $\partial f(x) \cap -\mathrm{bd}\, N_{g^{-1}(\mathbb{R}^m_-)}(x) = \emptyset$.

(3) $0 \notin \mathrm{bd}\, \mathrm{conv}\, \{\partial g_i(x)\} | i \in I(x)\}$ *("conv"= convex hull).*

Then, there exist some neighborhood $\mathcal{U}$ of $S(0)$ and constants $\varepsilon, L > 0$ such that

$$d(x, S(0)) \le L\, \|y\| \qquad \forall y \in B(0, \varepsilon),\ \forall x \in \mathcal{U} \cap S(y).$$

Proof. We shall show that S is calm at $(0, x)$ for all $x \in S(0)$ and that $S(0)$ consists just of isolated points. Given this fact, our compactness assumption ensures

that $S(0)$ will consist of only finitely many points, say $S(0) = \{x^1, \dots, x^N\}$. The calmness property then means the existence of constants $L_i, \varepsilon_i, \delta_i$ such that

$$d(x, S(0)) \le L_i \|y\| \qquad \forall y \in B(0, \varepsilon_i),\ \forall x \in B(x^i, \delta_i) \cap S(y)\ (i = 1, \dots, N).$$

Setting $L := \max L_i$, $\varepsilon := \min \varepsilon_i$, and $\mathcal{U} := \cup B(x^i, \delta_i)$, the assertion of the theorem follows.

In order to prove the stated facts, let $\bar{x} \in S(0)$ be arbitrarily given. Note that our constraint system $M(y) = \{x | g(x) \le y\}$ is a special case of (1.4) with $D := \mathbb{R}^m_-$ and $C := \mathbb{R}^k$. It is easily checked that assumption (1) implies the setting considered in Theorem 3.6. Indeed, regularity of the g_i implies regularity of any function $\sum_{i=1}^m y_i^* g_i$ with $y_i^* \ge 0$; hence $\langle y^*, g\rangle$ is regular at $\bar{x}$ for all

$$y^* \in \partial d_D^e(g(\bar{x})) = N_D(g(\bar{x})) \cap \mathbb{B} = \{y^* \in \mathbb{R}^m_+ | \, \|y^*\| \le 1,\ y_i^* = 0\, (i \notin I(\bar{x}))\},$$

as required in Theorem 3.6.

Suppose first that $0 \in \operatorname{int} H$, where $H := \operatorname{conv} \{\partial g_i(\bar{x})\} | i \in I(\bar{x})\}$. By regularity of the g_i, the subdifferentials $\partial g_i(\bar{x})$ are convex; hence

$$H = \left\{ \sum_{i \in I(\bar{x})} y_i^* \partial g_i(\bar{x}) \,\middle|\, \sum_{i \in I(\bar{x})} y_i^* = 1,\ y_i^* \ge 0 \right\}.$$

Therefore

$$H \subseteq \bigcup_{y^* \in N_D(g(\bar{x})) \cap \mathbb{B}} \partial \langle y^*, g\rangle\, (\bar{x}),$$

which along with $[T_C(\bar{x})]^0 = \{0\}$ implies that (3.11) holds. Hence, by Remark 3.7, $M(0)$ is locally isolated at $\bar{x}$. Then, $S(0)$ is isolated at $\bar{x}$ as well due to $S(0) \subseteq M(0)$. Furthermore, Theorem 3.6 allows us to derive the calmness of M at $(0, \bar{x})$, i.e.,

$$d(x, M(0)) \le L \|y\| \qquad \forall y \in B(0, \varepsilon),\ \forall x \in \mathcal{V} \cap M(y)$$

for some neighborhood $\mathcal{V}$ of $\bar{x}$ and some $\varepsilon, L > 0$. Choosing $\mathcal{V}$ small enough to meet $d(x, S(0)) = \|x - \bar{x}\|$ (by the local isolatedness of $S(0)$), one may conclude that

$$d(x, S(0)) \le d(x, M(0)) \le L \|y\| \qquad \forall y \in B(0, \varepsilon),\ \forall x \in \mathcal{V} \cap S(y),$$

where we used once more that $S(y) \subseteq M(y)$. This, however, is calmness of S at $\bar{x}$.

In the opposite case, $0 \notin \operatorname{int} H$, assumption (3) entails that $0 \notin H$. This condition along with assumption (1) implies the regularity and semismoothness of the set $g^{-1}(\mathbb{R}^m_-)$ at $\bar{x}$ (see [8, Lemma 3.6]). Then, in view of our assumptions, Proposition 4.9 may be invoked to show the local isolatedness of $S(0)$ at $\bar{x}$ again. Furthermore, the condition $0 \notin H$ is nothing but the Mangasarian–Fromovitz constraint qualification for a finite set of locally Lipschitz inequalities. It is well known that then the constraint mapping M has even the Aubin property around $(0, \bar{x})$, which is stronger than calmness. Hence, exactly the same argument as in the previous case can be applied to derive the calmness of S at $(0, \bar{x})$. □

Concerning the first assumption in Theorem 4.10, an analogous statement to that of Remark 3.3 applies. In particular, convex and $\mathcal{C}^1$-functions are regular and semismooth (even a maximum of such functions).

The next example illustrates the application of Theorem 4.10 in a smooth setting and, along the way, demonstrates how the upper Lipschitz stability of solutions can be established despite violation of the Mangasarian–Fromovitz constraint qualification.

EXAMPLE 4.11. *Consider the parametric optimization problem*

$$\min\{(x_1 - 1/2)^2 | - x_1 - x_2 \le y_1;\ x_2 \le y_2;\ x_1(1 - x_1) - x_2 \le y_3\}.$$

Then $S(0) = \{x^a, x^b\}$ *with* $x^a = (0, 0)$, $x^b = (1, 0)$. *Obviously,* $S(0)$ *is nonempty and bounded, and the constraint functions satisfy assumption* (1) *of Theorem* 4.10 *by smoothness. At* x^a *all unperturbed constraints are binding; hence the set* H *from assumption* (3) *is given as the convex hull of the three gradients:*

$$H = \operatorname{conv}\{(-1, -1), (0, 1), (1, -1)\}.$$

Obviously, $0 \in \operatorname{int} H$*; hence the Mangasarian–Fromovitz constraint qualification is violated at* x^a*. In contrast, the condition* $0 \notin \operatorname{bd} H$ *of assumption* (3) *is fulfilled. Furthermore,* $0 \in \operatorname{int} H$ *implies that the unperturbed constraint set* $M(0) = g^{-1}(\mathbb{R}^3_-)$ *is locally isolated at* x^a *(see the proof of Theorem* 4.10*). Therefore,* $N_{g^{-1}(\mathbb{R}^3_-)}(x^a) = \mathbb{R}^2$*, and assumption* (2) *holds trivially. Concerning* x^b*, only the second and third constraint are binding, so* $H = \operatorname{conv}\{(0, 1), (-1, -1)\}$ *and* $0 \notin H$*. Again, assumption* (3) *is satisfied. Moreover,* $N_{g^{-1}(\mathbb{R}^3_-)}(x^b)$ *is the convex cone generated by the two active gradients* $(0, 1)$ *and* $(-1, -1)$*, so its negative boundary is* $(\mathbb{R}_+ \cdot (0, -1)) \cup (\mathbb{R}_+ \cdot (1, 1))$*. Again, assumption* (2) *is fulfilled. Summarizing, the upper Lipschitz behavior of solutions to the above parametric problem can be derived.*

4.3. Equilibrium mappings. In [23] and [6] the authors study various stability properties of parametrized equilibria governed by the generalized equations

$$0 \in f(x, y) + Q(x), \tag{4.15}$$

where $x \in \mathbb{R}^k$ is the decision variable, $y \in \mathbb{R}^p$ is the parameter, $f : \mathbb{R}^k \times \mathbb{R}^p \to \mathbb{R}^k$ is continuously differentiable, and $Q : \mathbb{R}^k \rightrightarrows \mathbb{R}^k$ is a closed-valued multifunction. If one considers an optimization problem with (4.15) as a constraint, and an additional abstract constraint $(x, y) \in C$, then it is important to verify the calmness of the mapping $H : \mathbb{R}^k \rightrightarrows \mathbb{R}^k \times \mathbb{R}^p$ defined by

$$H(z) := \{(x, y) \in C | z \in f(x, y) + Q(x)\}.$$

H can easily be converted to the form (3.7), and so Corollary 3.4 can be applied. In fact, this procedure is illustrated in [8] by a parameterized equilibrium governed by a nonlinear complementarity problem. In this section we concentrate on a different mapping associated with parameterized equilibria, namely, the intersection

$$\Theta(y) := S(y) \cap C,$$

where S is the so-called solution mapping defined by

$$S(y) = \{x \in \mathbb{R}^k \,|\, 0 \in f(x, y) + Q(x, y)\}, \tag{4.16}$$

and C is a closed subset of $\mathbb{R}^k$ specifying the feasible decision variables. In (4.16) we admit that Q also depends on the parameter y, which extends the class of considered equilibria. Calmness of S (with Q depending only on x) has been investigated in [6],

but in the narrower sense of calmness on selections (see the introduction) where, for a reference pair $(\bar{y},\bar{x})$, one requires $\bar{x}$ to be an isolated point of $S(\bar{y})$.

The mapping S can be written in the form $S(y)=\{x\in\mathbb{R}^k | g(x,y)\in D\}$, where $g(x,y)=(x,y,-f(x,y))^T$ and $D=\operatorname{Gph} Q$. Therefore, Θ has exactly the structure of the multifunction M in (3.7), and we immediately obtain the following statement from Corollary 3.4.

THEOREM 4.12. *Let C be regular and semismooth at $\bar{x}\in\Theta(\bar{y})$. Further assume that the qualification condition*

$$\left.\begin{array}{r} 0\in w-(\nabla_x f(\bar{x},\bar{y}))^T z+\operatorname{bd} N_C(\bar{x}),\\ (w,v,z)\in N_{\operatorname{Gph} Q}(\bar{x},\bar{y},-f(\bar{x},\bar{y}))\end{array}\right\}\quad \textit{implies}\quad \left\{\begin{array}{l} w=0,\\ v=0,\\ z=0,\end{array}\right. \tag{4.17}$$

holds true. Then Θ is calm at $(\bar{y},\bar{x})$.

If Q depends just on x, then $g(x,y)=(x,-f(x,y))^T$, and the qualification condition (4.17) reduces to

$$\left.\begin{array}{r} 0\in w-(\nabla_x f(\bar{x},\bar{y}))^T z+\operatorname{bd} N_C(\bar{x}),\\ (w,z)\in N_{\operatorname{Gph} Q}(\bar{x},-f(\bar{x},\bar{y}))\end{array}\right\}\quad \text{implies}\quad \left\{\begin{array}{l} w=0,\\ z=0.\end{array}\right. \tag{4.18}$$

The following example shows that the qualification conditions (4.17), (4.18) may well be violated even when Θ is calm at $(\bar{y},\bar{x})$.

EXAMPLE 4.13. *In* (4.16) *let $k=p=1$, $f\equiv 0$, and*

$$Q(x,y)=\partial\varphi(x)+N_{y+\mathbb{R}_-}(x),\qquad \varphi(x)=\left\{\begin{array}{lll} -x & \textit{for} & x\le 0,\\ 0 & \textit{for} & x>0.\end{array}\right.$$

Clearly,

$$S(y)=\left\{\begin{array}{ll} y & \textit{for } y\le 0,\\ [0,y] & \textit{otherwise.}\end{array}\right.$$

Let $(\bar{y},\bar{x})=(0,0)$. It is easily seen that with $C=\mathbb{R}_+$ or $C=\mathbb{R}_-$ the mapping Θ is calm at $(\bar{y},\bar{x})$. Nevertheless, condition (4.17) *is not fulfilled.*

The reason for the failure of (4.17) in the last example is that this condition works with a too large upper approximation of $D^*S(\bar{y},\bar{x})$. In such cases it makes sense to directly apply Theorem 3.2: In Example 4.13 one calculates

$$D^*S^{-1}(\bar{y},\bar{x})\,(y^*)=\left\{\begin{array}{lll} y^* & \text{if} & y^*\neq 0,\\ \mathbb{R}_- & \text{if} & y^*=0.\end{array}\right.$$

Both for $C=\mathbb{R}_+$ and $C=\mathbb{R}_-$, it is easily verified that (3.1) holds true, and hence, calmness of Θ can be derived. Observe that this result could not be obtained when considering the whole cone $N_C(\bar{x})$ instead of its boundary.

REMARK 4.14. *The calmness of Θ in the above example follows directly from its polyhedral nature. Nevertheless, it illustrates well the need to weaken the standard criteria ensuring the Aubin property when analyzing calmness.*

Acknowledgments. The authors gratefully acknowledge support by the Weierstrass Institute Berlin, where this paper was prepared during the second and third authors' stay as visiting guests. They are also indebted to A. Kruger (Minsk) for helpful discussions on the subject of the paper, and to both referees for their valuable comments and suggestions.

REFERENCES

[1] J.-P. Aubin and I. Ekeland, *Applied Nonlinear Analysis*, Wiley, New York, 1984.

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

[3] J.V. Burke and M.C. Ferris, *Weak sharp minima in mathematical programming*, SIAM J. Control Optim., 31 (1993), pp. 1340–1359.

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

[5] S. Deng, *Global error bounds for convex inequality systems in Banach spaces*, SIAM J. Control. Optim., 36 (1998), pp. 1240–1249.

[6] A.L. Dontchev and R.T. Rockafellar, *Ample Parameterization of Variational Inclusions*, SIAM J. Optim., 12 (2001), 170–187.

[7] R. Henrion, *The Approximate Subdifferential and Parametric Optimization*, Habilitation thesis, Humboldt University, Berlin, Germany, 1997.

[8] R. Henrion and J. Outrata, *A subdifferential condition for calmness of multifunctions*, J. Math. Anal. Appl., 258 (2001), pp. 110–130.

[9] R. Henrion and A. Jourani, *Subdifferential systems for calmness for convex systems*, SIAM J. Optim., 13 (2002), pp. 520–534.

[10] A.D. Ioffe, *Necessary and sufficient conditions for a local minimum,* 1: *A reduction theorem and first order conditions,* SIAM J. Control Optim., 17 (1979), pp. 245–250.

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

[12] A.D. Ioffe, *Approximate subdifferential and applications, Part* 3, Mathematika, 36 (1989), pp. 1–38.

[13] A.J. King and R.T. Rockafellar, *Sensitivity analysis for nonsmooth generalized equations*, Math. Programming, 55 (1992), pp. 193–212.

[14] A. Kruger, *private communication.*

[15] A.B. Levy, *Implicit multifunction theorems for the sensitivity analysis of variational conditions*, Math. Programming, 74 (1996), pp. 333–350.

[16] A.B. Levy, *Calm minima in parameterized finite-dimensional optimization*, SIAM J. Optim., 11 (2000), pp. 160–178.

[17] A.B. Levy, *Solution sensitivity from general principles*, SIAM J. Control Optim., 40 (2001), pp. 1–38.

[18] W. Li and I. Singer, *Global error bounds for convex multifunctions and applications*, Math. Oper. Res., 23 (1998), pp. 443–462.

[19] R. Mifflin, *Semismooth and semiconvex functions in constrained optimization*, SIAM J. Control Optim., 15 (1977), pp. 959–972.

[20] B.S. Mordukhovich, *Approximation Methods in Problems of Optimization and Control*, Nauka, Moscow, 1988 (in Russian: Wiley-Interscience English translation to appear).

[21] B.S. Mordukhovich, *Complete characterization of openness, metric regularity and Lipschitz properties of multifunctions*, Trans. Amer. Math. Soc., 340 (1993), pp. 1–35.

[22] B.S. Mordukhovich, *Generalized differential calculus for nonsmooth and set-valued mappings*, J. Math. Anal. Appl., 183 (1994), pp. 250–288.

[23] B.S. Mordukhovich, *Stability theory for parametric generalized equations and variational inequalities via nonsmooth analysis*, Trans. Amer. Math. Soc., 343 (1994), pp. 609–655.

[24] J.-S. Pang, *Error bounds in mathematical programming*, Math. Programming, 79 (1997), pp. 299–332.

[25] S.M. Robinson, *Generalized equations and their solutions, Part* I: *Basic theory*, Math. Programming Study, 10 (1979), pp. 128–141.

[26] S.M. Robinson, *Some continuity properties of polyhedral multifunctions*, Math. Programming Study, 14 (1981), pp. 206–214.

[27] R.T. Rockafellar and R. Wets, *Variational Analysis*, Springer-Verlag, New York, 1999.

[28] 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. 13, No. 2, pp. 619–633

CONSTRAINED MINIMA AND LIPSCHITZIAN PENALTIES IN METRIC SPACES*

DIETHARD KLATTE† AND BERND KUMMER‡

This paper is dedicated to our colleague and friend Jochem Zowe on the occasion of his 60th birthday

Abstract. It is well known that a local minimizer of a constrained optimization problem with Lipschitzian objective is a free local minimizer of an assigned penalty function if the constraints satisfy an appropriate regularity condition. We use an upper Lipschitz property (L1) as regularity concept and present locally Lipschitz penalty functions defined on the whole space for arbitrary constraint maps of this type. We give conditions under which the maximum of the penalties of finitely many multifunctions is a valid penalty function for the intersection of these multifunctions.

Further, the same statements will be derived under other regularity assumptions, namely, for calm or pseudo-Lipschitz constraints which violate (L1), by showing that some submapping of a calm map always has property (L1) and possesses (locally) the same penalties. In this way our penalizations induce in a unified manner, via known properties of free local minimizers for Lipschitz functions only, primal and dual necessary conditions for these basic notions of regularity.

Key words. constrained optimization, exact penalty functions, upper Lipschitz property, calmness, pseudoregularity

AMS subject classifications. 49J52, 90C48, 58C06

PII. S105262340139625X

1. Introduction.

Given metric spaces X and Y, our basic model is the optimization problem

$$\min f(x) \quad \text{s.t.} \quad x \in X^0, \tag{1}$$

where $X^0 \neq \emptyset$ is a fixed subset of X and $f : X \to \mathbb{R}$ is locally Lipschitz, along with any parametric embedding of (1)

$$\min f(x) \quad \text{s.t.} \quad x \in S(y) \tag{2}$$

such that $S : Y \rightrightarrows X$ is a multifunction and $X^0 \subset S(y^0)$. Here, y^0 is any fixed element of Y, and no particular structure on S is required. For instance, $S(y)$ may be a solution set of a generalized equation

$$z^0 \in H(x, y), \quad \text{where } H : X \times Y \rightrightarrows Z,$$

or one may assume that X, Y are Banach spaces, $y^0 = 0$, and

$$S(y) = \{x \in X \mid g(x) \in y + K\}, \qquad X^0 = S(0), \tag{3}$$

where a function $g : X \to Y$ and a nonempty set $K \subset Y$ are given.

The model (3) is of fundamental importance and has been studied extensively and under various assumptions in the optimization literature; for basic results we refer,

*Received by the editors October 8, 2001; accepted for publication (in revised form) April 11, 2002; published electronically October 1, 2002.
http://www.siam.org/journals/siopt/13-2/39625.html
†Institut für Operations Research, Universität Zürich, Moussonstrasse 15, CH-8044 Zürich, Switzerland (klatte@ior.unizh.ch).
‡Institut für Mathematik, Humboldt–Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany (kummer@mathematik.hu-berlin.de).

e.g., to [39, 17, 40, 21, 20, 25, 41, 2, 4, 10]. These studies tell us that, under various Lipschitz-type conditions on the map S near (y^0, X^0) or near (y^0, x^0), the following basic statement holds:

(4) If x^0 is a local minimizer of (1), then x^0 locally minimizes a function $P(x) = f(x) + R_S(x)$ w.r.t. $x \in X$,

where R_S depends on the (analytical) description of S and can be defined by means of Lagrange multipliers or directly by some *(exact) penalty term*, e.g., for (3),

$$R_S(x) = \alpha \operatorname{dist}(g(x), K), \quad \alpha \text{ large.} \tag{5}$$

Clearly, R_S is not unique and may be more or less complicated. In particular, its continuity depends on the continuity of g. An alternative classical approach to exact penalization consists in defining a (globally Lipschitz) penalty term depending only on X^0:

$$R_S(x) = \alpha \operatorname{dist}(x, X^0), \quad \alpha \text{ large.} \tag{6}$$

The regularity notion of *calmness* allows a connection between $\operatorname{dist}(x, X^0)$ and S for specific mappings S in problem (1)–(2); see, e.g., [13, 10, 6]. As a basic reference to these penalization techniques of constrained optimization, we refer to the survey [7], which also gives a historical overview including the basic early work (e.g., [15, 47, 13, 18, 20, 16]) devoted to this subject.

The main purpose of the present paper is to demonstrate how one and the same simple construction of penalty terms (which are everywhere defined and globally Lipschitz) can be applied under different regularity conditions (locally upper Lipschitz, pseudo-Lipschitz, calm) and for arbitrary constraint maps S, which may be defined even discontinuously or via multifunctions.

The required regularity properties are defined and discussed in section 2. In section 3, we show that under Robinson's *upper Lipschitz condition* (L1) for S, there is a special graph-distance function R_S satisfying (4) and locally Lipschitz on X even if S in (2) has been arbitrarily defined (see Lemmas 3.1, 3.2).

Moreover, for *calm* intersection maps $\Sigma(y, z) = S(y) \cap T(z)$ (which describe two constraints), the maximum function $\max\{R_S, R_T\}$ at the place of R_Σ satisfies (4) again; i.e., x^0 minimizes $f(x) + \max\{R_S, R_T\}$ whenever S^{-1} is pseudo-Lipschitz (Theorem 3.5). The latter turns out to be a trivial condition for equations $g(x) = y$ or "inequalities" (3) if g is locally Lipschitz, since $S^{-1}(x) = \{y \in Y \mid y \in g(x) - K\}$ is only a Lipschitzian translation of a fixed set. In addition, we demonstrate that this pseudo-Lipschitz condition permits us to replace the *mapping* T by the *fixed set* $T(z^0)$ for showing calmness of Σ.

In section 4, we illustrate some ideas for deriving dual optimality conditions from Lipschitzian penalties.

Finally, we show in section 5 that (and how) R_S can be constructed in the same way under other Lipschitz assumptions, namely, for *calm* or *pseudo-Lipschitz* mappings S, though S does not necessarily satisfy the initial supposition (L1) in these cases. For the construction in Theorem 5.1, we assign to S an appropriate submapping $\Gamma \subset S$ that is again upper Lipschitz and defines locally the same residuals $R_\Gamma = R_S$. This construction is one reason why we allow X^0 to be an arbitrary subset of $S(y^0)$ in (2). Another reason is to emphasize the point that only the behavior of f and S near $X^0 := S(y^0) \cap (x^0 + \varepsilon B)$ plays a role in characterizing a local minimizer x^0 of f on $S(y^0)$.

To indicate that our penalties and the key interrelations do not depend on any linear structure of the assigned spaces, we use metric spaces, although, of course, dual conditions cannot be obtained in this general setting.

2. Preliminaries.

Lipschitz conditions for the constraint map. Let X and Y be metric spaces, $S : Y \rightrightarrows X$, and $(y^0, x^0) \in \operatorname{gph} S$. Further, let $\emptyset \neq X^0 \subset S(y^0)$. We write $X^0 + \varepsilon B_X := \{x \mid \operatorname{dist}(x, X^0) \leq \varepsilon\}$ with the usual point-to-set distance $\operatorname{dist}(x, X^0) = \inf_{x' \in X^0} d(x, x')$, based on the metric d of X. The expression $y^0 + \varepsilon B_Y$ is defined analogously.

In what follows, we need three Lipschitz properties of S. The map S is called

(L1) *locally upper Lipschitz at* (y^0, X^0) if there are positive constants L and ε such that

$$S(y) \cap (X^0 + \varepsilon B_X) \subset X^0 + L d(y, y^0) B_X \quad \forall y \in y^0 + \varepsilon B_Y; \tag{7}$$

(L2) *pseudo-Lipschitz at* (y^0, x^0) if there are positive constants L and ε such that

$$S(y) \cap (x^0 + \varepsilon B_X) \subset S(y') + L d(y, y') B_X \quad \forall y, y' \in y^0 + \varepsilon B_Y; \tag{8}$$

(L3) *calm at* (y^0, x^0) if there are positive constants L and ε such that

$$S(y) \cap (x^0 + \varepsilon B_X) \subset S(y^0) + L d(y, y^0) B_X \quad \forall y \in y^0 + \varepsilon B_Y. \tag{9}$$

In every case we call L a *rank* of the related Lipschitz property.

If S is a function, then (L2) simply claims the Lipschitz continuity of S on some neighborhood of y^0. Trivially, if $x^0 \in X^0$, then (L1) implies (L3) since $X^0 \subset S(y^0)$. Further, with $y' = y^0$, condition (L2) implies (L3). However, the implications (L2) $\Rightarrow$ (L1), (L3) $\Rightarrow$ (L1), and (L3) $\Rightarrow$ (L2) are not valid in general (cf. Examples 1, 2 below).

Property (L1) has been taken from Robinson [38]. There, X^0 coincided with $S(y^0)$ or with some closed, connected component of $S(y^0)$. In fact, these are the most important situations for applications of (L1). Nevertheless, we allow X^0 to be an arbitrary subset of $S(y^0)$, in view of our construction of the mapping Γ in Theorem 5.1 below; see the remark at the end of the introductory section.

For $X^0 = \{x^0\}$, (L1) was called locally upper Lipschitz property at (y^0, x^0) in [14]. In this case, x^0 is necessarily isolated in $S(y^0)$.

Property (L2) (also called the Aubin property in various papers) is a basic stability condition (cf., e.g., [1, 43]), and the calmness property (L3) has been applied and investigated, e.g., in [10, 42, 6, 7, 43], for deriving optimality conditions.

It is a consequence of Theorem 5.1 below that calmness can be used in a way similar to the upper Lipschitz property (L1) for the definition of exact penalties. An interesting recent calmness condition for multifunctions can be found in [19]. It uses a so-called semismoothness property [30] and can be applied to the models in [34] and many models in [28].

It is well known from Robinson's [36] work that a finite-dimensional constraint map

$$S(y, z) = \{x \mid g(x) \leq y,\ h(x) = z\}, \tag{10}$$

with $(g, h) \in C^1(\mathbb{R}^n, \mathbb{R}^{m+k})$, is pseudo-Lipschitz at (y^0, z^0, x^0) iff the *Mangasarian–Fromovitz condition* (MFCQ) [29] is satisfied:

(MFCQ) $Dh(x^0)$ has full rank and there is some u such that $Dh(x^0)u = 0$ and $g(x^0) + Dg(x^0)u < y^0$.

In the setting (3) with closed convex set K and continuously Fréchet differentiable function g, the pseudo-Lipschitz property of S is characterized by *Robinson's constraint qualification*; see, e.g., [37, 11] for details.

Example 1 (pseudo-Lipschitz but not locally upper Lipschitz). Let $s(y) = 1 + \sqrt{|y|}$ and $S(y)$ be the interval $[-s(y), s(y)]$ for real y.

Then, if $\emptyset \neq X^0 \subset S(0)$, the mapping S is not locally upper Lipschitz at $(0, X^0)$ because, for each set $U = X^0 + \varepsilon B$ (for some fixed $\varepsilon > 0$) and any $L > 0$, one finds points $x(y) \in S(y) \cap U$ such that $\operatorname{dist}(x(y), X^0) > L|y|$ and $|y| < 1/L$. Further, S is not calm at the point $(0, 1)$. On the other hand, S is pseudo-Lipschitz (hence also calm) at each point $(0, x^0)$, $x^0 \in \operatorname{int} S(0)$.

Example 2 (the inverse of Dirichlet's function). For the real function

$$h(x) = 0 \text{ if } x \text{ is rational}, \qquad h(x) = 1 \text{ otherwise},$$

the inverse multifunction h^{-1} is calm at $(y^0, x^0) = (0, 0)$ and locally upper Lipschitz at $(0, h^{-1}(0))$ since $\operatorname{cl} h^{-1}(0) = \mathbb{R}$. The mapping $S(y) = \{x \mid h(x) \geq y\}$ is even pseudo-Lipschitz at $(0, 0)$, since $h(x) = 1 \geq y$ holds for all irrational x and all y near 0.

The latter example indicates that the usual construction of penalties for calm equations, $P(x) = f(x) + \alpha\|h(x)\|$, may lead to terrible auxiliary functions P, whereas our subsequent definition of P via (13) always generates a locally Lipschitz function.

3. Characterizing the upper Lipschitz property by functions. As in section 2, let X, Y be metric spaces, $S : Y \rightrightarrows X$, $y^0 \in Y$, and $\emptyset \neq X^0 \subset S(y^0)$. Further, let $p : X \to \mathbb{R}$. We call p *Lipschitzian increasing* near X^0 if $p \equiv 0$ on X^0 and there are constants $c > 0$, $\delta > 0$ such that

$$p(x) \geq c \operatorname{dist}(x, X^0) \quad \text{whenever} \quad \operatorname{dist}(x, X^0) < \delta. \tag{11}$$

3.1. Describing functions. We say that p *is a describing function for* S *near* (y^0, X^0), briefly $p \stackrel{\circ}{=} S(y^0, X^0)$, if the statement

$$\begin{aligned} &S \text{ is locally upper Lipschitz at } (y^0, X^0) \\ &\Leftrightarrow\ p \text{ is Lipschitzian increasing near } X^0 \end{aligned}$$

holds true.

Notice that this definition tacitly assumes that a describing function is defined in terms of the multifunction S or its describing data. For example, let $h : X \to Y$ be any function and $S = h^{-1}$. Then one easily sees that $p(x) = d(h(x), y^0)$ fulfills $p \stackrel{\circ}{=} S(y^0, X^0)$ for each fixed pair y^0 and $\emptyset \neq X^0 \subset S(y^0)$. On the other hand, for any real C^1 function h, the function $p(x) = d(h(x), y^0)^2$ increases too slowly near X^0 and does not satisfy $p \stackrel{\circ}{=} S(y^0, X^0)$.

Generally, for *getting* $p \stackrel{\circ}{=} S(y^0, X^0)$, the function p must be assigned to the map S near (y^0, X^0) in some reasonable way. However, if we already *know* that $p \stackrel{\circ}{=} S(y^0, X^0)$ holds true, then checking the locally upper Lipschitz property at (y^0, X^0) is reduced to the often simpler question of whether the function p is Lipschitzian increasing near X^0.

In addition, property (L1) for feasible-set maps S now allows the derivation of optimality conditions for constrained minimization in terms of free local minimizers of an auxiliary function involving p. This is stated in the next lemma, which applies basically the same arguments as the related propositions in [10, 6] and many other papers for calm constraints and is a consequence of the following basic argument: If

x^0 locally minimizes a locally Lipschitz function f on X^0, then x^0 locally minimizes the function $f(x) + k\text{dist}\,(x, X^0)$ on X for sufficiently large k. Hence, for p satisfying $p(x^0) = 0$ and (11), x^0 is also a (free) local minimizer of

$$P(x) := f(x) + \alpha p(x) \tag{12}$$

if $\alpha \geq k/c$.

We shall prove the following lemma only for completeness and in order to demonstrate that additional assumptions like the closeness or compactness of X^0, which we cannot guarantee for our application in Theorem 5.1 below, are indeed not needed in this context.

LEMMA 3.1 (local minima under upper Lipschitzian constraints). *Let S be locally upper Lipschitz at (y^0, X^0) with rank L, and let $f : X \to \mathbb{R}$ be Lipschitz near x^0 with rank K. Further, let x^0 be a local minimizer of f on X^0 and $p \stackrel{\circ}{=} S(y^0, X^0)$. Then x^0 is a local minimizer of P in* (12) *whenever $\alpha > Kc^{-1}$ with c from* (11).

Proof. Let $\mu > 0$ and $U = X^0 + \varepsilon B_X$ be the set in (7). Given $x \in U$, select some $\pi_x \in X^0$ with

$$d(x, \pi_x) \leq \text{dist}\,(x, X^0) + \mu.$$

Then, $d(x, \pi_x) \leq d(x, x^0) + \mu$. For $d(x, x^0) < \delta$ and small δ and μ, we know that $d(\pi_x, x^0) \leq d(\pi_x, x) + d(x, x^0)$ is small enough to apply the Lipschitz estimate $f(x) \geq f(\pi_x) - Kd(x, \pi_x)$ and $f(\pi_x) \geq f(x^0)$. Further, since $p \stackrel{\circ}{=} S(y^0, X^0)$, we have $-p(x) \leq -c\ \text{dist}\,(x, X^0)$. Therefore,

$$\begin{aligned} f(x) &\geq f(\pi_x) - Kd(x, \pi_x) \\ &\geq f(x^0) - Kd(x, \pi_x) \\ &\geq f(x^0) - K\ [\text{dist}\,(x, X^0) + \mu] \\ &\geq f(x^0) - K\ c^{-1}p(x) - K\mu. \end{aligned}$$

After passing to the limit $\mu \downarrow 0$, we obtain the assertion since

$$P(x) \geq f(x) + \alpha p(x) \geq P(x^0) = f(x^0) \quad \text{if} \quad \alpha > Kc^{-1}. \qquad \square$$

Having the penalization (12) in mind, the structure and continuity of possible "candidates" p are interesting issues. Often, a describing function can be defined quite naturally.

For instance, given the usual system of equations and inequalities (10), with $(y^0, z^0) = (0, 0)$ and $X^0 = S(0, 0)$, one may set

$$p(x) = \|h(x)\| + \max_i\{0, g_i(x)\}.$$

As long as h and g are locally Lipschitz, then so is p, and one obtains a locally Lipschitz penalty term $R_S(x) = \alpha\ p(x)$.

However, if h and g are not locally Lipschitz or if S is an arbitrary multifunction, the construction of a locally Lipschitz function $p \stackrel{\circ}{=} S(y^0, X^0)$ is desirable (for analytical reasons) and less obvious. For S, y^0 and $X^0 = S(y^0)$ from Example 2, one easily finds two describing functions, namely, $p_1(x) = h(x)$ and $p_2(x) \equiv 0$.

By the next basic lemma, there is always a describing *Lipschitz* function p, globally defined with rank 1 and not depending on X^0. We assume that the metric in product spaces $Y \times X$ is defined as

$$d((y, x),\ (y', x')) = \max\{d_Y(y, y'),\ d_X(x, x')\}.$$

LEMMA 3.2 (describing Lipschitz function). *Given any multifunction*

$$S : Y \rightrightarrows X \quad \textit{and} \quad \emptyset \neq X^0 \subset S(y^0),$$

the distance function

$$p_S(x) = \operatorname{dist}((y^0, x), \operatorname{gph} S) \quad (\leq \operatorname{dist}(x, X^0)) \tag{13}$$

satisfies $p_S \stackrel{\circ}{=} S(y^0, X^0)$. *Thus* S *is locally upper Lipschitz at* (y^0, X^0) *iff* p_S *is Lipschitzian increasing near* X^0.

Proof. For simplicity, we write $d(\cdot,\cdot)$ both for $d_X(\cdot,\cdot)$ and $d_Y(\cdot,\cdot)$. Evidently, p_S vanishes on X^0. Let p_S be Lipschitzian increasing near X^0. Then S is locally upper Lipschitz with rank $L = c^{-1}$, since, for $\operatorname{dist}(x, X^0) < \delta$,

$$x \in S(y) \Rightarrow c\, d(x, X^0) \leq p_S(x) \leq d((y^0, x),\ (y, x)) = d(y^0, y).$$

Conversely, suppose p_S is *not* Lipschitzian increasing near X^0. Then for each $\varepsilon > 0$ there is some x such that

$$\operatorname{dist}(x, X^0) < \varepsilon \quad \text{and} \quad \operatorname{dist}((y^0, x), \operatorname{gph} S) = p_S(x) < \varepsilon \operatorname{dist}(x, X^0).$$

Select any $(y_t, x_t) \in \operatorname{gph} S$ with

$$d((y^0, x), (y_t, x_t)) = \max\{d(x, x_t), d(y^0, y_t)\} < t := \varepsilon \operatorname{dist}(x, X^0).$$

Then,

$$\begin{aligned} &\varepsilon + t > \operatorname{dist}(x_t, X^0) \geq \operatorname{dist}(x, X^0) - t = (1 - \varepsilon) \operatorname{dist}(x, X^0) \\ &\text{and} \quad d(y^0, y_t) < \varepsilon \operatorname{dist}(x, X^0). \end{aligned}$$

Thus, both

$$\operatorname{dist}(x_t, X^0) < \varepsilon + t \quad \text{and} \quad \frac{d(y^0, y_t)}{\operatorname{dist}(x_t, X^0)} < \frac{\varepsilon}{(1 - \varepsilon)}$$

vanish (as $\varepsilon \downarrow 0$); thus S is not locally upper Lipschitz at (y^0, X^0). ∎

It is trivial but useful to note that functions $p \stackrel{\circ}{=} S(y^0, X^0)$ may be replaced, in Lemma 3.1, by any function p^+ satisfying $p^+ \geq p$ and $p^+(x^0) = 0$. Applying the function $p = p_S$ of Lemma 3.2, the new objective P turns out to be even locally Lipschitz.

3.2. More examples of describing functions. The distance function p_S of (13) will play a crucial role in the remainder of this paper. Before we proceed, however, we provide further examples of useful describing functions in this section.

3.2.1. Cone constraints. Let Y be a linear normed space, X be a metric space, $g : X \to Y$, $K \subset Y$ be a convex cone, $\eta \in \operatorname{int} K \backslash \{0\}$, and $S(y) = \{x \in X \mid g(x) \in y + K\}$.

LEMMA 3.3 (cone constraints). *Let* $\emptyset \neq X^0 \subset S(0)$ *and* $X_r = X^0 + rB$. *Then, if* g *is Lipschitz on* X_β *for some* $\beta > 0$, *the function*

$$p(x) = \inf\{\lambda > 0 \mid g(x) + \lambda\eta \in K\} \tag{14}$$

satisfies

$$c_1\, p_S(x) \leq p(x) \leq c_2\, p_S(x) \quad \forall x \in X_r \tag{15}$$

for certain constants $0 < c_1 \le c_2$ *and* $r = \beta/3$. *Hence* $p \triangleq S(0, X^0)$.

Proof. Let L_g be some Lipschitz rank of g on X_β and $\eta + \alpha B_Y \subset K$, $\alpha > 0$. Then, one obtains, for all $\lambda > 0$ and $x \in X_\beta$,

$$\lambda\eta + g(x) \in K \quad \text{if } \|g(x)\| \le \lambda\alpha.$$

Hence $p(x) \le \alpha^{-1}\|g(x)\|$ and

$$\begin{aligned} p_S(x) &= \operatorname{dist}((0_Y, x), \operatorname{gph} S) \\ &\le \inf_{\lambda > p(x)} \operatorname{dist}((0_Y, x), (\lambda\eta, x)) \\ &= p(x)\|\eta\|. \end{aligned} \tag{16}$$

Next, fix any $r \in (0, \frac{1}{2}\beta)$. We verify

$$p(x) \le (1 + L_g)\alpha^{-1} p_S(x) \quad \forall x \in X_r. \tag{17}$$

Since $(0_Y, X^0) \subset \operatorname{gph} S$, the inequality $p_S(x) \le \operatorname{dist}(x, X^0) \le r < \frac{1}{2}\beta$ holds. Thus one finds some ε satisfying $p_S(x) < \varepsilon < \frac{1}{2}\beta$ as well as some (y', x') such that

$$g(x') \in y' + K, \qquad \|y'\| < \varepsilon, \qquad \text{and} \quad d(x', x) < \varepsilon.$$

From $g(x') - y' \in K$ we conclude (by adding points of a convex cone) that

$$g(x') - y' + \lambda\eta + \lambda\alpha B_Y \subset K \quad \forall \lambda > 0.$$

Therefore, the inclusion

$$g(x) + \lambda\eta \in K \tag{18}$$

holds whenever

$$g(x) \in g(x') - y' + \lambda\alpha B_Y.$$

Because of $\|y'\| < \varepsilon$, the latter can be guaranteed by $\|g(x) - g(x')\| + \varepsilon \le \lambda\alpha$. Now, by the choice of r, x and x' belong to X_β and satisfy $\|g(x) - g(x')\| \le L_g\varepsilon$; hence (18) is ensured whenever $(L_g + 1)\varepsilon \le \lambda\alpha$. Considering $\inf \lambda$, this yields

$$p(x) \le \alpha^{-1}(1 + L_g)p_S(x) \quad \text{via} \quad \varepsilon \to p_S(x).$$

The assertion now follows from (16), (17), and Lemma 3.2. ∎

Notice that we did not need to require that the convex cone K be closed. Also, if X is a linear normed space and g is linear and continuous, then one easily shows that p is convex. In addition, p is bounded on some neighborhood of $x \in \operatorname{int} X_r$ due to (15). So it is also locally Lipschitz on $\operatorname{int} X_r$. Needless to say, p is simpler than p_S from a computational point of view.

For $Y = \mathbb{R}^m$, $K = \{y \in Y \mid y_i \le 0 \ \forall i\}$, and $\eta = -(1, \dots, 1)$, one obtains the usual penalty term $p(x) = \max_i\{0, g_i(x)\}$.

If Y is the space of real, symmetric (n, n) matrices, $K \subset Y$ the cone of all positive semidefinite symmetric (n, n) matrices, and $\eta = E$, one has $p(x) = \inf\{\lambda > 0 \mid g(x) + \lambda E \in K\}$. Thus $-p(x)$ is the smallest eigenvalue of $g(x)$, provided that $g(x) \notin K$.

3.2.2. Cone constraints and equations. Let

$$\Sigma(y,z) = S(y) \cap T(z),$$

where S satisfies the assumptions of Lemma 3.3, $h : X \to Z$ sends X into a linear normed space Z, $T(z) = h^{-1}(z)$, and $\emptyset \neq X^0 \subset \Sigma(0_{Y\times Z})$.

If we write Σ as a cone constraint with the cone $K' = (K, \{0_Z\})$ in the product space, the interior of K' is empty and Lemma 3.3 cannot be applied. In addition, the describing distance function p_Σ according to Lemma 3.2 satisfies only

$$\begin{aligned} p_\Sigma(x) &= \inf_{(y',z',x')\in \operatorname{gph}\Sigma} \max\{d((0_Y,x),(y',x')),\ d((0_Z,x),(z',x'))\} \\ &\geq \max\Big\{ \inf_{(y',x')\in \operatorname{gph} S} d((0_Y,x),(y',x')),\ \inf_{(z'',x'')\in \operatorname{gph} T} d((0_Z,x),(z'',x''))\Big\} \\ &= \max\{p_S(x),\ p_T(x)\}. \end{aligned}$$

Thus we know, by the previous statements, only that the maximum function

$$q(x) = \max\{p_S(x),\ p_T(x)\}, \qquad T = h^{-1}, \tag{19}$$

fulfills

$$p_\Sigma(x) \ \geq\ q(x), \tag{20}$$

and that q is Lipschitzian increasing near X^0 iff so is

$$Q(x) = \max\{p(x),\ \|h(x)\|\}, \qquad p \text{ from Lemma 3.3.}$$

However, due to the gap between $p_\Sigma(x)$ and $q(x)$, the function p_Σ may Lipschitzian increase near X^0 while q does not. (Then S and h^{-1} violate (L1), but Σ does not.) In this situation, q and Q are not describing functions for Σ.

On the other hand, the maximum q turns out to be a describing function under all classical regularity assumptions that ensure, as in the subsequent lemma, that Σ is pseudo-Lipschitzian (or only calm) at $(0_Y, 0_Z, x^0)$.

LEMMA 3.4 (the max-form under calmness). *Suppose that X, Y, Z are Banach spaces, $g, h \in C^1$, $Dh(x^0)X = Z$, some u satisfies $Dh(x^0)u = 0$ and $g(x^0) + Dg(x^0)u \in \operatorname{int} K$, and $x^0 \in X^0 \subset \Sigma(0_Y, 0_Z)$. Moreover, let X^0 be contained in a sufficiently small (by diameter) neighborhood Ω of x^0. Then, q in* (19) *is a describing function for Σ near $(0_Y, 0_Z, X^0)$.*

Proof. Our suppositions are nothing but well-known regularity conditions for optimization problems in Banach spaces (cf. [37, 36, 48]), which ensure that the map Σ is pseudo-Lipschitz at $(0_Y, 0_Z, x^0)$; for the pseudo-Lipschitz property of more general intersection maps, we refer to [26]. Thus the lemma will follow from our Theorem 3.5 below because $T^{-1} = h$ is locally Lipschitz. □

3.2.3. Arbitrary intersections. More generally, let X, Y, Z be metric spaces, $S : Y \rightrightarrows X$, $T : Z \rightrightarrows X$, and $\Sigma(y,z) = S(y) \cap T(z)$.

THEOREM 3.5 (the max-form for intersections). *Let $x^0 \in X^0 \subset \Sigma(y^0, z^0)$, Σ be calm at (y^0, z^0, x^0), and T^{-1} be pseudo-Lipschitz at (x^0, z^0). Moreover, suppose that X^0 is contained in a sufficiently small (by diameter) neighborhood Ω of x^0. Then, the function*

$$q(x) = \max\{p_S(x),\ p_T(x)\} \tag{21}$$

fulfills $q \doteq \Sigma(y^0, z^0, X^0)$.

Proof. The inequality (20) follows as above without any assumptions. We estimate q in the opposite direction. First notice that q is Lipschitz, so $q(x) \downarrow 0$ as $x \to x^0$. Consequently, for sufficiently small neighborhoods Ω we find arbitrarily small $\delta > q(x)$. Now, for x near $X^0 \subset \Omega$ and (small) $\delta > q(x)$, there are (by definition of p_S and p_T) points $(y', x') \in \operatorname{gph} S$ and $(z'', x'') \in \operatorname{gph} T$ such that

$$\max\{d(x', x), d(y', y^0)\} < \delta \quad \text{and} \quad \max\{d(x'', x), d(z'', z^0)\} < \delta. \tag{22}$$

Next we apply that T^{-1} is pseudo-Lipschitz at (x^0, z^0), say with rank K. Since $z'' \in T^{-1}(x'')$, there exists, for small δ and Ω, some $z' \in T^{-1}(x')$ satisfying

$$d(z', z'') \le K d(x'', x') \le 2K\delta. \tag{23}$$

We thus obtain $(y', z', x') \in \operatorname{gph} \Sigma$ and

$$d((y', z', x'), (y^0, z^0, x^0)) \le \max\{\delta,\ \delta + 2K\delta,\ \delta + d(x, x^0)\}.$$

Therefore, since (y', z', x') is close to (y^0, z^0, x^0), we may use the calmness of Σ, say with rank L at (y^0, z^0, x^0). By (22) and (23), this ensures the existence of some $\xi \in \Sigma(y^0, z^0)$ such that

$$d(\xi, x') \le L \max\{d(y', y^0), d(z', z^0)\} \le L(1 + 2K)\delta.$$

Finally, $p_\Sigma(x) \le d(\xi, x)$ implies the upper estimate

$$p_\Sigma(x) \le d(\xi, x) \le d(\xi, x') + d(x', x) \le L(1 + 2K)\delta + \delta$$

and yields (as $\delta \downarrow q(x)$) $p_\Sigma(x) \le (L(1+2K)+1)\, q(x)$. Recalling (20) and Lemma 3.2, the latter tells us that q is a describing function for Σ near (y^0, z^0, X^0) because so is p_Σ. ∎

Notice that neither Lemma 3.4 nor Theorem 3.5 asserts the upper Lipschitz property of Σ at (y^0, z^0, X^0), itself. The relation between the upper and pseudo-Lipschitz properties as well as calmness will be investigated under Theorem 5.1. Next, we inspect the hypothesis that Σ is calm in the previous theorem and reduce the calmness of the intersection of two mappings to the intersection of one mapping with a constant set (a new space X) only.

THEOREM 3.6 (calm intersections). *Let S be calm at (y^0, x^0), T be calm at (z^0, x^0), and T^{-1} be pseudo-Lipschitz at (x^0, z^0). Moreover, let $H(z) = T(z) \cap S(y^0)$ be calm at (z^0, x^0). Then $\Sigma(y, z) = S(y) \cap T(z)$ is calm at (y^0, z^0, x^0).*

Proof. Let $(y, z, x) \in \operatorname{gph} \Sigma$ be close to (y^0, z^0, x^0). Since S and T are calm (say, with rank L), there are $x' \in S(y^0)$ and $x'' \in T(z^0)$ such that

$$\max\{d(x, x'), d(x, x'')\} \le L \max\{d(y, y^0), d(z, z^0)\}.$$

Since T^{-1} is pseudo-Lipschitz (rank K), $z^0 \in T^{-1}(x'')$, and x', x'' are close to x^0, we find z' such that

$$z' \in T^{-1}(x') \quad \text{and} \quad d(z', z^0) \le K d(x', x'').$$

Next observe that $x' \in H(z')$. Therefore, there also exists some $\xi \in H(z^0)$ satisfying

$$d(\xi, x') \le L_H\ d(z', z^0).$$

Using these inequalities, we directly obtain the required Lipschitz estimate

$$\begin{aligned} d(x,\xi) &\le d(x,x') + d(x',\xi) \\ &\le L \max\{d(y,y^0), d(z,z^0)\} + L_H \; d(z',z^0) \\ &\le L \max\{d(y,y^0), d(z,z^0)\} + L_H \; K \; d(x',x'') \\ &\le L \max\{d(y,y^0), d(z,z^0)\} + 2L_H \; K \; L \; \max\{d(y,y^0), d(z,z^0)\}. \end{aligned}$$ □

3.2.4. Fixed set-constraints. Assume that

$$\Sigma(y) = S(y) \cap M \quad \text{and} \quad M \text{ is a fixed, closed subset of } X.$$

Clearly, then, one may study S on the new metric space $X := M$, which leads us to a new function p_S. Nevertheless, let us also regard two usual descriptions of $x \in M$ via functions from the viewpoint of the pseudo-Lipschitz assumption for T^{-1} in the theorem.

(a) If $h(x) = \operatorname{dist}(x, M)$, $Z = \mathbb{R}^+$, $z^0 = 0$, and $T = h^{-1}$, the mapping $T^{-1} = h$ is pseudo-Lipschitz and $p_T(x) = \inf\{\max\{z', d(x', x)\} \mid \operatorname{dist}(x', M) = z'\}$. If S is already calm (w.r.t. the space X), then Theorem 3.6 allows us to study, instead of $S(y) \cap T(z)$, the calmness of the mapping

$$H(z) = S(y^0) \cap T(z) = S(y^0) \cap h^{-1}(z).$$

If H is calm at $(0, x^0)$, then so is $S \cap T$ at $(y^0, 0, x^0)$; hence the original map $\Sigma(y) = S(y) \cap M$ at (y^0, x^0) is also calm. This way, one may replace (for the calmness investigation) the fixed set M by $S(y^0)$, and the mapping S by h^{-1}.

(b) If $h(x) = 0$ for $x \in M$, $h(x) = 1$ otherwise, and Z, z^0, and T are as above, the function $T^{-1} = h$ is discontinuous and it holds that $p_T(x) = \operatorname{dist}(x, M)$ for $\operatorname{dist}(x, M) < 1$. The theorem cannot be applied. Indeed, for small $z > 0$, we would obtain the trivial constant map

$$H(z) = S(y^0) \cap T(z) = \emptyset,$$

which tells us nothing about Σ.

4. Dual optimality conditions. Provided that X and Y are linear normed spaces, now all necessary optimality conditions for free local minimizers x^0 of $P = f + \alpha p$ in (12) induce necessary conditions for the originally constrained problem (1). In particular, if directional derivatives $P'(x^0; u)$ of P at x^0 in direction u exist, then

$$P'(x^0; u) \ge 0 \qquad \forall u \in X \tag{24}$$

must hold. On the other hand,

$$0 \in \partial_g P(x^0) \tag{25}$$

holds for every (generalized) subdifferential ∂_g of P at x^0, at least after restricting f, p to the set $X_\varepsilon = x^0 + \varepsilon B$, $\varepsilon > 0$ sufficiently small.

Let us mention only two basic approaches for obtaining dual conditions in terms of f and p. Various other approaches and more involved results can be found in [39, 17, 40, 21, 27, 46, 20, 3, 25, 41, 2, 4, 35, 10, 45, 31, 8, 44, 5]; for some unification of several approaches, we refer again to [7].

4.1. Dual conditions via directional derivatives. If f and p are directionally differentiable, then $(f+\alpha p)'(x^0;u) = f'(x^0;u) + \alpha p'(x^0;u)$; hence also

$$(f+\alpha p)'(x^0;u) \le f'(x^0;u) + \alpha p'(x^0;u), \tag{26}$$

which is the crucial requirement when the directional derivatives are generalized. Now, (26) and (24) yield a condition for the sum

$$\inf_u \, (f'(x^0;u) + \alpha p'(x^0;u)) \ge 0. \tag{27}$$

Let, in addition, the directional derivatives be *continuous and sublinear* in u (which is evident for locally Lipschitz convex functions).

Then, applying the Hahn–Banach theorem (see, e.g., [22]) to the sublinear function

$$Q(u,v) := f'(x^0;u) + \alpha p'(x^0,v) \quad \text{in the product space } \Pi = X \times X,$$

the supporting functional $L_0(u,v) = 0$ of Q on the subspace $\Pi_0 = \{(u,v) \mid u = v\}$ can be extended to an additive and homogeneous functional $L(u,v) = L_1(u) + L_2(v)$ on Π that supports Q everywhere. Thus,

$$L_1(u) + L_2(u) = 0 \qquad \text{and} \qquad Q(u,v) \ge L_1(u) + L_2(v)$$

hold for all $u, v \in X$. The latter implies (since Q is continuous by assumption) that L_1, L_2 are bounded, and

$$\inf_u \, (f'(x^0;u) - L_1(u)) \; + \; \inf_v \, (\alpha p'(x^0,v) - L_2(v)) \ge 0.$$

Thus one obtains the existence of some $x^* = L_1 \in X^*$ satisfying the (conjugate duality) inequality

$$\inf_u \, (f'(x^0;u) - x^*(u)) \; + \; \inf_u \, (x^*(u) + \alpha p'(x^0;u)) \ge 0.$$

Since the involved directional derivatives are positively homogeneous, the infima are zero, and x^* belongs (by definition) just to the usual, convex subdifferential $\partial f'(x^0;\cdot)(0)$. Similarly, one obtains $-x^* \in \partial(\alpha p')(x^0;\cdot)(0)$.

In other words, after defining a new subdifferential ∂_n for the nonconvex function f at x^0 as

$$\partial_n f(x^0) = \partial f'(x^0;\cdot)(0) \tag{28}$$

(and applying it to αp, too), some $x^* \in X^*$ satisfies the inclusions

$$x^* \in \partial_n f(x^0) \qquad \text{and} \qquad -x^* \in \partial_n(\alpha p)(x^0) = \alpha \partial_n p(x^0),$$

which is the generalized Lagrange condition

$$0 \in \partial_n f(x^0) \; + \; \alpha \partial_n p(x^0), \tag{29}$$

or simply $Df(x^0) + \alpha Dp(x^0) = 0$ for Fréchet differentiable functions.

Recalling the form of q in Theorem 3.5, one sees that directional derivatives of maximum functions play a crucial role in this context. Further, one observes that several concepts of directional derivatives f' may be applied to derive (29) for the subdifferential (28) in the above way, provided that

(i) condition (27) remains valid for local minimizers x^0 of P, and
(ii) the existence of directional derivatives as well as sublinearity and continuity w.r.t. the directions u can be guaranteed.

For locally Lipschitz functions $f : X \to \mathbb{R}$ on linear normed spaces X, these hypotheses are satisfied by Clarke's directional derivatives

$$f^c(x^0;u) = \limsup_{x\to x^0,\ t\downarrow 0} t^{-1}[f(x+tu) - f(x)]$$

and his subdifferential $\partial_c f(x^0)$, which coincides with $\partial_n f(x^0)$ after identifying f' and f^c; cf. [10]. For $X = \mathbb{R}^n$, the equation $\partial_c f(x^0) = \partial f(x^0)$ in terms of generalized Jacobians $\partial f(x^0)$ (cf. [9]) increases the analytical tools for computing the derivatives in question.

4.2. Dual conditions via generalized subdifferentials. Without applying directional derivatives, one may use that (25) holds true for the minimizer x^0 and some generalized subdifferentials ∂_g. Then, provided that a chain rule

$$\partial_g(f+\alpha p)(x^0) \subset \partial_g f(x^0) + \partial_g(\alpha p)(x^0) \tag{30}$$

is valid, one directly obtains (29) w.r.t. the subdifferential under consideration. We refer the reader who is interested in recent results on inclusion (30) for particular subdifferentials to [12, 32, 23, 33, 24]. For the related theory (mainly of certain limiting Fréchet subdifferentials), the Lipschitz property of f and p as well as the fact that X is an Asplund space play an important role.

However, the calculus becomes simpler if we may start with inequality

$$f(x) + \alpha p(x) \geq f(x^0) + \alpha p(x^0) \qquad (x \text{ near } x^0)$$

instead of (25), since now

$$\alpha(p(x) - p(x^0)) \geq -(f(x) - f(x^0)) \qquad (x \text{ near } x^0)$$

directly shows that $x^* \in \partial_g(-f)(x^0)$ yields $x^* \in \alpha\partial_g p(x^0)$ for all subdifferentials ∂_g.

5. Relations between regularity conditions. Having only calmness (or the pseudo-Lipschitz property) of S at (y^0, x^0), our Example 1 indicates that the local upper Lipschitz condition at (y^0, X^0) may fail to hold even if X^0 is replaced by $X^0 \cap (x^0 + \varepsilon B)$, which would be enough to see that x^0 minimizes P locally. So the optimality condition of Lemma 3.1 (which is, in fact, true) needs another proof.

For this reason, we establish a general connection between calm, upper Lipschitz, and pseudo-Lipschitz maps S by verifying that, supposing calmness (L3) at (y^0, x^0), there exists a submapping Γ of S that satisfies (L1) at $(y^0, \Gamma(y^0))$ and (in addition) $x^0 \in S(y^0) \cap (x^0 + \frac{1}{2}\varepsilon B) \subset \Gamma(y^0)$. This submapping, defined as the intersection of $S(y)$ with (open) balls of different radii, will replace S in Lemma 3.1 in such a way that $p_S = p_\Gamma$ still holds on some neighborhood of x^0.

THEOREM 5.1 (selection maps and optimality condition). *Let $S : Y \rightrightarrows X$ be calm at (y^0, x^0) with rank L and constant ε. Then, the submapping*

$$\Gamma(y) = S(y) \cap \{x \mid d(x, x^0) < \varepsilon - Ld(y, y^0)\}$$

has the following properties:

(i) *Γ is locally upper Lipschitz at $(y^0, \Gamma(y^0))$ with rank L and also constant ε;*

(ii) *the functions*

$$p_S(x) = \operatorname{dist}((y^0, x), \operatorname{gph} S) \quad \textit{and} \quad p_\Gamma(x) = \operatorname{dist}((y^0, x), \operatorname{gph} \Gamma)$$

coincide for x near x^0;

(iii) *if $f : X \to \mathbb{R}$ is locally Lipschitz and x^0 (locally) minimizes f on $S(y^0)$, then x^0 is a free local minimizer of $P(x) := f(x) + \alpha p_S(x)$ whenever α is sufficiently large.*

Proof. We set $U = x^0 + \varepsilon B_X$, $V = y^0 + \varepsilon B_Y$.

(i) Let $x \in \Gamma(y)$, $y \in V$. Then we have $d(x, x^0) < \varepsilon - Ld(y, y^0)$, and, due to the calmness of S, there exists $x' \in S(y^0)$ with $d(x', x) \le Ld(y, y^0)$. Moreover, since

$$d(x', x^0) \le d(x', x) + d(x, x^0) < Ld(y, y^0) + (\varepsilon - Ld(y, y^0)) = \varepsilon,$$

we obtain $x' \in \Gamma(y^0)$. Thus $\operatorname{dist}(x, \Gamma(y^0)) \le Ld(y, y^0)$.

(ii) Clearly, $p_S(x^0) = p_\Gamma(x^0)$ is evident, and $p_S \le p_\Gamma$ follows from $\operatorname{gph} \Gamma \subset \operatorname{gph} S$. To verify $p_\Gamma(x) \le p_S(x)$ for x near x^0, we consider any $x \in U$ such that $x \ne x^0$ and

$$(3 + 2L)\, d(x, x^0) < \varepsilon. \tag{31}$$

Let $(y', x') \in \operatorname{gph} S$ realize the distance $p_S(x)$ up to an error $\lambda d(x, x^0)$, $0 < \lambda < 1$. We show that $(y', x') \in \operatorname{gph} \Gamma$. Indeed, since $(y^0, \Gamma(y^0)) \subset \operatorname{gph} S$, we obtain

$$\begin{aligned} \max\{d(x', x), d(y', y^0)\} &\le p_S(x) + \lambda d(x, x^0) \\ &\le \operatorname{dist}(x, \Gamma(y^0)) + \lambda d(x, x^0) \\ &\le (1 + \lambda) d(x, x^0) \\ &< 2d(x, x^0). \end{aligned}$$

Thus, the inequalities

$$d(x', x^0) \le d(x', x) + d(x, x^0) < 3d(x, x^0) \quad \text{and} \quad d(y', y^0) < 2d(x, x^0)$$

hold. From (31), we further have $3d(x, x^0) < \varepsilon - 2Ld(x, x^0)$. Therefore, we can estimate

$$d(x', x^0) < 3d(x, x^0) < \varepsilon - 2Ld(x, x^0) \le \varepsilon - Ld(y', y^0)$$

in order to obtain that $(y', x') \in \operatorname{gph} S$ also belongs to $\operatorname{gph} \Gamma$. The latter yields, by the choice of (y', x'),

$$p_\Gamma(x) \le p_S(x) + \lambda d(x, x^0)$$

as well as $p_\Gamma(x) \le p_S(x)$ via $\lambda \downarrow 0$. Summarizing, $p_\Gamma(x) = p_S(x)$ holds for all x satisfying (31).

(iii) For sufficiently small $\delta > 0$, x^0 minimizes f on $\Gamma(y^0) = S(y^0) \cap (x^0 + \delta B)$. Decreasing ε if necessary, we have $0 < \varepsilon < \delta$, and (i) ensures that Γ is locally upper Lipschitz. Thus x^0 is, by Lemma 3.2 and Lemma 3.1, a local minimizer of $P(x) := f(x) + \alpha p_\Gamma(x)$. The assertion thus follows from (ii). ∎

In the theorem, calmness and the pseudo-Lipschitz property of S may replace each other, and the form of the function P (by applying p_S) remains the same under each of the Lipschitz conditions (L1), (L2), and (L3). Once again, we emphasize that P is locally Lipschitz on X without any hypothesis concerning the metric spaces

X and Y or the structure of S. For linear normed spaces and convex sets $\operatorname{gph} S$, one easily sees that p_S is convex, too.

Finally, we note that the definition of $\Gamma(y)$ via intersection with the open balls

$$B^0(x^0, \varepsilon - Ld(y, y^0)) = \{x \mid d(x, x^0) < \varepsilon - Ld(y, y^0)\}$$

preserves the properties of lower semicontinuity of S; in contrast, the intersection with closed balls

$$B(x^0, \varepsilon - Ld(y, y^0)) = \{x \mid d(x, x^0) \leq \varepsilon - Ld(y, y^0)\}$$

does not (although the theorem remains true by the same arguments). However, if we define $\Gamma(y)$ as the intersection of $S(y)$ with the *fixed* open ball $B^0(x^0, \varepsilon)$, the theorem fails to hold.

Example 3 (intersection with $B^0(x^0, \varepsilon)$). Take the mapping $S : \mathbb{R}^2 \rightrightarrows \mathbb{R}^2$ defined as $S(y) = \{x \mid \|x - y\| \geq \frac{1}{2}\|y\|\}$ with maximum norm, and set $y^0 = (0, \frac{2}{3}\varepsilon)$ and $x^0 = (0, 0) \in S(y^0)$. Notice that S is pseudo-Lipschitz at (y^0, x^0), with rank $L = 1$ and constant ε. Next define $\Gamma(y) = S(y) \cap B^0(x^0, \varepsilon)$. Then, the point $x' = (0, \varepsilon)$ has distance $0 < d \leq \varepsilon$ to the set $\Gamma(y^0)$. On the other hand, $x' \in \Gamma(y(t))$ for all $y(t) = y^0 - t(0, 1)$, $t > 0$. Thus, we observe that Γ is not locally upper Lipschitz (with rank L and constant ε) at $(y^0, \Gamma(y^0))$, since $x' \in \Gamma(y(t)) \cap (\Gamma(y^0) + \varepsilon B)$ and $x' \notin \Gamma(y^0) + Ld(y(t), y^0)B$ for sufficiently small $t > 0$.

Acknowledgments. We are most grateful to the referees for their detailed and constructive comments. Also, we would like to thank Stefan Scholtes for his editorial help and many valuable suggestions.

REFERENCES

[1] J.-P. Aubin and I. Ekeland, *Applied Nonlinear Analysis*, Wiley, New York, 1984.
[2] A. Ben-Israel, A. Ben-Tal, and S. Zlobec, *Optimality in Nonlinear Programming: A Feasible Direction Approach*, Wiley, New York, 1981.
[3] A. Ben-Tal, *Second-order and related extremality conditions in nonlinear programming*, J. Optim. Theory Appl., 31 (1980), pp. 143–165.
[4] A. Ben-Tal and J. Zowe, *A unified theory of first and second order conditions for extremum problems in topological vector spaces*, Math. Program. Study, 19 (1982), pp. 39–76.
[5] J. F. Bonnans and A. Shapiro, *Perturbation Analysis of Optimization Problems*, Springer, New York, 2000.
[6] J. V. Burke, *Calmness and exact penalization*, SIAM J. Control Optim., 29 (1991), pp. 493–497.
[7] J. V. Burke, *An exact penalization viewpoint of constrained optimization*, SIAM J. Control Optim., 29 (1991), pp. 968–998.
[8] R. W. Chaney, *Optimality conditions for piecewise C^2 nonlinear programming*, J. Optim. Theory Appl., 61 (1989), pp. 179–202.
[9] F. H. Clarke, *On the inverse function theorem*, Pacific J. Math., 64 (1976), pp. 97–102.
[10] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983.
[11] R. Cominetti, *Metric regularity, tangent sets and second-order optimality conditions*, Appl. Math. Optim., 21 (1990), pp. 265–287.
[12] R. Deville and E. M. E. Haddad, *The subdifferential of the sum of two functions in Banach spaces.* I. *First order case*, J. Convex Anal., 3 (1996), pp. 295–308.
[13] S. Dolecki and S. Rolewicz, *Exact penalties for local minima*, SIAM J. Control Optim., 17 (1979), pp. 596–606.
[14] A. Dontchev, *Characterizations of Lipschitz stability in optimization*, in Recent Developments in Well-Posed Variational Problems, R. Lucchetti and J. Revalski, eds., Kluwer, Dordrecht, The Netherlands, 1995, pp. 95–116.
[15] I. I. Eremin, *The penalty method in convex programming*, Soviet Math. Dokl., 8 (1966), pp. 459–462.
[16] R. Fletcher, *Practical Methods of Optimization, Vol.* 2: *Constrained Optimization*, Wiley, New York, 1981.

[17] E. G. Golstein, *Theory of Convex Programming*, Transl. Math. Monogr. 36, AMS, Providence, RI, 1972.
[18] S.-P. Han and O. L. Mangasarian, *Exact penalty functions in nonlinear programming*, Math. Programming, 17 (1979), pp. 251–269.
[19] R. Henrion and J. Outrata, *A subdifferential condition for calmness of multifunctions*, J. Math. Anal. Appl., 258 (2001), pp. 110–130.
[20] 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.
[21] A. D. Ioffe and V. M. Tichomirov, *Theory of Extremal Problems*, Nauka, Moscow, 1974 (in Russian).
[22] L. W. Kantorovich and G. P. Akilov, *Funktionalanalysis in normierten Räumen*, Akademie Verlag, Berlin, 1964.
[23] A. Y. Kruger, *Strict* (ε,δ)*-semidifferentials and extremality of sets and functions*, Dokl. Nat. Akad. Nauk Belarus, 44 (2000), pp. 421–424 (in Russian).
[24] A. Y. Kruger, *Strict* (ε,δ)*-subdifferentials and extremality conditions*, Optimization, 51 (2002), pp. 539–554.
[25] A. Y. Kruger and B. S. Mordukhovich, *Extremal points and Euler equations in nonsmooth optimization*, Doklady Akad. Nauk BSSR, 24 (1980), pp. 684–687 (in Russian).
[26] B. Kummer, *Inverse functions of pseudo regular mappings and regularity conditions*, Math. Programming, 88 (2000), pp. 313–339.
[27] E. S. Levitin, A. A. Miljutin, and N. P. Osmolovski, *On conditions for a local minimum in a problem with constraints*, in Mathematical Economics and Functional Analysis, B. S. Mitjagin, ed., Nauka, Moscow, 1974, pp. 139–202 (in Russian).
[28] Z.-Q. Luo, J.-S. Pang, and D. Ralph, *Mathematical Programs with Equilibrium Constraints*, Cambridge University Press, Cambridge, UK, 1996.
[29] 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.
[30] R. Mifflin, *Semismooth and semiconvex functions in constrained optimization*, SIAM J. Control Optim., 15 (1977), pp. 959–972.
[31] B. S. Mordukhovich, *Approximation Methods in Problems of Optimization and Control*, Nauka, Moscow, 1988 (in Russian).
[32] B. S. Mordukhovich and Y. Shao, *Stability of set-valued mappings in infinite dimensions: Point criteria and applications*, SIAM J. Control Optim., 35 (1997), pp. 285–314.
[33] H. V. Ngai and M. Théra, *Metric inequality, subdifferential calculus and applications*, Set-Valued Anal., 9 (2001), pp. 187–216.
[34] J. V. Outrata, *A generalized mathematical program with equilibrium constraints*, SIAM J. Control Optim., 38 (2000), pp. 1623–1638.
[35] J.-P. Penot, *On regularity conditions in mathematical programming*, Math. Program. Study, 19 (1982), pp. 167–199.
[36] S. M. Robinson, *Regularity and stability for convex multivalued functions*, Math. Oper. Res., 1 (1976), pp. 130–143.
[37] S. M. Robinson, *Stability theorems for systems of inequalities, Part* II: *Differentiable nonlinear systems*, SIAM J. Numer. Anal., 13 (1976), pp. 497–513.
[38] S. M. Robinson, *Some continuity properties of polyhedral multifunctions*, Math. Program. Study, 14 (1981), pp. 206–214.
[39] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[40] R. T. Rockafellar, *Conjugate Duality and Optimization*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 16, SIAM, Philadelphia, 1974.
[41] R. T. Rockafellar, *The Theory of Subgradients and its Application to Problems of Optimization. Convex and Nonconvex Functions*, Heldermann, Berlin, 1981.
[42] R. T. Rockafellar, *Extension of subgradient calculus with applications to optimization*, Nonlinear Anal., 9 (1985), pp. 665–698.
[43] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer, Berlin, 1998.
[44] A. Shapiro, *First and second order optimality conditions and perturbation analysis of semi–infinite programming problems*, in Semi–Infinite Programming, R. Reemtsen and J.-J. Rückmann, eds., Kluwer, Boston, Dordrecht, London, 1998, pp. 103–133.
[45] M. Studniarski, *Necessary and sufficient conditions for isolated local minima of nonsmooth functions*, SIAM J. Control Optim., 24 (1986), pp. 1044–1049.
[46] J. Warga, *Necessary conditions without differentiability assumptions in optimal control*, J. Differential Equations, 15 (1975), pp. 13–46.
[47] W. I. Zangwill, *Nonlinear programming via penalty functions*, Management Sci., 13 (1967), pp. 344–358.
[48] J. Zowe and S. Kurcyusz, *Regularity and stability for the mathematical programming problem in Banach spaces*, Appl. Math. Optim., 5 (1979), pp. 42–62.

SIAM J. OPTIM.
Vol. 13, No. 3, pp. 635–659

GLOBAL CONVERGENCE OF A TRUST-REGION SQP-FILTER ALGORITHM FOR GENERAL NONLINEAR PROGRAMMING*

ROGER FLETCHER†, NICHOLAS I. M. GOULD‡, SVEN LEYFFER†, PHILIPPE L. TOINT§, AND ANDREAS WÄCHTER¶

Abstract. A trust-region SQP-filter algorithm of the type introduced by Fletcher and Leyffer [*Math. Program.*, 91 (2002), pp. 239–269] that decomposes the step into its normal and tangential components allows for an approximate solution of the quadratic subproblem and incorporates the safeguarding tests described in Fletcher, Leyffer, and Toint [*On the Global Convergence of an SLP-Filter Algorithm*, Technical Report 98/13, Department of Mathematics, University of Namur, Namur, Belgium, 1998; *On the Global Convergence of a Filter-SQP Algorithm*, Technical Report 00/15, Department of Mathematics, University of Namur, Namur, Belgium, 2000] is considered. It is proved that, under reasonable conditions and for every possible choice of the starting point, the sequence of iterates has at least one first-order critical accumulation point.

Key words. nonlinear optimization, sequential quadratic programming, filter methods, convergence theory

AMS subject classifications. 90C30, 65K05

PII. S1052623499357258

1. Introduction.

We analyze an algorithm for solving optimization problems where a smooth objective function is to be minimized subject to smooth nonlinear constraints. No convexity assumption is made. More formally, we consider the problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & c_{\mathcal{E}}(x) = 0, \\ & c_{\mathcal{I}}(x) \geq 0, \end{array} \tag{1.1}$$

where f is a twice continuously differentiable real valued function of the variables $x \in \mathbb{R}^n$ and $c_{\mathcal{E}}(x)$ and $c_{\mathcal{I}}(x)$ are twice continuously differentiable functions from $\mathbb{R}^n$ into $\mathbb{R}^m$ and from $\mathbb{R}^n$ into $\mathbb{R}^p$, respectively. Let $c(x)^T = (c_{\mathcal{E}}(x)^T \;\; c_{\mathcal{I}}(x)^T)$.

The class of algorithms that we discuss belongs to the class of trust-region methods and, more specifically, to that of *filter methods* introduced by Fletcher and Leyffer [18], in which the use of a penalty function, a common feature of the large majority of the algorithms for constrained optimization, is replaced by the introduction of a so-called filter.

A global convergence theory for an algorithm of this class is proposed by Fletcher, Leyffer, and Toint in [19], in which the objective function is locally approximated by a linear function, leading, at each iteration, to the (exact) solution of a linear program.

*Received by the editors June 14, 1999; accepted for publication (in revised form) March 12, 2002; published electronically November 6, 2002.
http://www.siam.org/journals/siopt/13-3/35725.html

†Department of Mathematics, University of Dundee, Dundee, Scotland (fletcher@maths.dundee.ac.uk, sleyffer@maths.dundee.ac.uk).

‡Computational Science and Engineering Department, Rutherford Appleton Laboratory, Chilton, Oxfordshire OX11 0QX, England (gould@rl.ac.uk).

§Department of Mathematics, University of Namur, Namur, Belgium (philippe.toint@fundp.ac.be).

¶Mathematical Sciences Department, IBM T. J. Watson Research Center, P. O. Box 218, Yorktown Heights, NY 10598. Current address: Department of Chemical Engineering, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213 (andreasw@andrew.cmu.edu).

This algorithm therefore mixes the use of the filter with sequential linear programming (SLP). This approach was generalized by the same authors in [20], where the objective function is approximated by a quadratic model, which results in a sequential quadratic programming (SQP) technique in which each quadratic program must be solved globally. In this paper, we again consider approximating the objective function by a quadratic model, but, at variance with the latter reference, the method discussed here does not require the global solution of the associated nonconvex quadratic programming (QP) subproblem, which is known to be a theoretically difficult process—it is known to be NP hard (see Murty and Kabadi [26]). The algorithm analyzed here also has a different mechanism for deciding on the compatibility of this subproblem and allows for an approximate subproblem solution.

2. A class of trust-region SQP-filter algorithms.

2.1. An approximate SQP framework. SQP methods are iterative. At a given iterate x_k, they implicitly apply Newton's method to solve (a local version of) the first-order necessary optimality conditions by solving the QP subproblem $\mathrm{QP}(x_k)$ given by

$$\begin{array}{ll} \text{minimize} & f_k + \langle g_k, s\rangle + \frac{1}{2}\langle s, H_k s\rangle \\ \text{subject to} & c_{\mathcal{E}}(x_k) + A_{\mathcal{E}}(x_k)s = 0, \\ & c_{\mathcal{I}}(x_k) + A_{\mathcal{I}}(x_k)s \geq 0, \end{array} \tag{2.1}$$

where $f_k = f(x_k)$, $g_k = g(x_k) \stackrel{\text{def}}{=} \nabla_x f(x_k)$, where $A_{\mathcal{E}}(x_k)$ and $A_{\mathcal{I}}(x_k)$ are the Jacobians of the constraint functions $c_{\mathcal{E}}$ and $c_{\mathcal{I}}$ at x_k, and where H_k is a symmetric matrix. We will not immediately be concerned about how H_k is obtained, but we will return to this point in section 3. Assuming that a suitable matrix H_k can be found, the solution of $\mathrm{QP}(x_k)$ then yields a step s_k. If $s_k = 0$, then x_k is first-order critical for problem (1.1).

Unfortunately, due to the locally convergent nature of Newton's iteration, the step s_k may not always be very good. One possible way to cope with this difficulty is to define an appropriate merit function whose value decreases with the goodness of s_k, which is where penalty functions typically play a role. A trust-region or a linesearch method is then applied to minimize this merit function, ensuring global convergence under reasonable assumptions. However, as one of our objectives is to avoid penalty functions (and the need to update the associated penalty parameter), we instead consider a trust-region approach that will not use any penalty function.[1] In such an approach, the objective function of $\mathrm{QP}(x_k)$ is intended to be only of local interest; that is, we restrict the step s_k in the norm to ensure that $x_k + s_k$ remains in a *trust-region* centered at x_k. In other words, we replace $\mathrm{QP}(x_k)$ by the subproblem $\mathrm{TRQP}(x_k, \Delta_k)$ given by

$$\begin{array}{ll} \text{minimize} & m_k(x_k + s) \\ \text{subject to} & c_{\mathcal{E}}(x_k) + A_{\mathcal{E}}(x_k)s = 0, \\ & c_{\mathcal{I}}(x_k) + A_{\mathcal{I}}(x_k)s \geq 0, \\ \text{and} & \|s\| \leq \Delta_k \end{array} \tag{2.2}$$

for some (positive) value of the *trust-region radius* Δ_k, where we have defined

$$m_k(x_k + s) = f_k + \langle g_k, s\rangle + \frac{1}{2}\langle s, H_k s\rangle \tag{2.3}$$

[1] Recently, Wächter and Biegler [31] have proposed a linesearch variant of the ideas described in this paper.

and where $\|\cdot\|$ denotes the Euclidean norm. This latter choice is purely for ease of exposition. We could equally use a family of iteration-dependent norms $\|\cdot\|_k$, so long as we require that all members of the family are uniformly equivalent to the Euclidean norm. The interested reader may verify that all subsequent developments can be adapted to this more general case by introducing the constants implied by this uniform equivalence wherever needed.

Remarkably, most early SQP algorithms assume that an exact local solution of QP(x_k) or TRQP(x_k, Δ_k) is found, although attempts have been made by Dembo and Tulowitzki [8] and Murray and Prieto [25] to design conditions under which an approximate solution of the subproblem is acceptable. We revisit this issue in what follows, and start by noting that the step s_k may be viewed as the sum of two distinct components, a *normal step* n_k, such that $x_k + n_k$ satisfies the constraints of TRQP(x_k, Δ_k), and a *tangential step* t_k, whose purpose is to obtain reduction of the objective function's model while continuing to satisfy those constraints. This framework is therefore similar in spirit to the composite-step SQP methods pioneered by Vardi [30], Byrd, Schnabel, and Shultz [5], and Omojokun [27], and later developed by several authors, including Biegler, Nocedal, and Schmid [1], El-Alem [12, 13], Byrd, Gilbert, and Nocedal [3], Byrd, Hribar, and Nocedal [4], Bielschowsky and Gomes [2], Liu and Yuan [23], and Lalee, Nocedal, and Plantenga [22]. More formally, we write

$$s_k = n_k + t_k \tag{2.4}$$

and assume that

$$c_{\mathcal{E}}(x_k) + A_{\mathcal{E}}(x_k)n_k = 0, \quad c_{\mathcal{I}}(x_k) + A_{\mathcal{I}}(x_k)n_k \geq 0, \tag{2.5}$$

$$\|s_k\| \leq \Delta_k, \tag{2.6}$$

and

$$c_{\mathcal{E}}(x_k) + A_{\mathcal{E}}(x_k)s_k = 0, \quad c_{\mathcal{I}}(x_k) + A_{\mathcal{I}}(x_k)s_k \geq 0. \tag{2.7}$$

Of course, this is a strong assumption, since in particular (2.5) or (2.6)/(2.7) may not have a solution. We shall return to this possibility shortly. Given our assumption, there are many ways to compute n_k and t_k. For instance, we could compute n_k from

$$n_k = P_k[x_k] - x_k, \tag{2.8}$$

where P_k is the orthogonal projector onto the feasible set of QP(x_k). In what follows, we do not make any specific choice for n_k, but we shall make the assumptions that n_k exists when the maximum violation of the nonlinear constraints at the kth iterate $\theta_k \stackrel{\text{def}}{=} \theta(x_k)$, where

$$\theta(x) = \max\left[0, \max_{i\in\mathcal{E}} |c_i(x)|, \max_{i\in\mathcal{I}} -c_i(x)\right] \tag{2.9}$$

is sufficiently small, and that n_k is then reasonably scaled with respect to the values of the constraints. In other words, we assume that

$$n_k \text{ exists and } \|n_k\| \leq \kappa_{\text{usc}}\theta_k \text{ whenever } \theta_k \leq \delta_n \tag{2.10}$$

for some constants $\kappa_{\text{usc}} > 0$ and $\delta_n > 0$. This assumption is also used by Dennis, El-Alem, and Maciel [9] and Dennis and Vicente [11] in the context of problems only

involving equality constraints. We can interpret it in terms of the constraint functions themselves and the geometry of the boundary of the feasible set. For instance, if we define the linearized feasible set at x by

$$\mathcal{L}(x) \stackrel{\text{def}}{=} \{v \in \mathbb{R}^n \mid c_{\mathcal{E}}(x) + A_{\mathcal{E}}(x)(v-x) = 0, \;\; c_{\mathcal{I}}(x) + A_{\mathcal{I}}(x)(v-x) \geq 0\}$$

and assume that, at every limit point x_* of the sequence of iterates, the relative interior of the linearized constraints $\mathrm{ri}\{\mathcal{L}(x_*)\}$ is nonempty and that the active set settles, in that $\mathcal{A}(x_{k_i}) = \mathcal{A}(x_*)$ for sufficiently large k_i with $\lim_i x_{k_i} = x_*$, we know, by applying a continuity argument, that the feasible set of $\mathrm{QP}(x_k)$ is nonempty for such a k, which implies that P_k is well defined and that a normal step n_k of the form (2.8) exists. Furthermore, if the singular values of the Jacobian of constraints active at x_*, $A_{\mathcal{A}(x_*)}(x_*)$, are nonzero, those of $A_{\mathcal{A}(x_*)}(x_k)$ must be bounded away from zero by continuity in a neighborhood of x_*. Since only the constraints active at x_* can be active in a sufficiently small neighborhood of this limit point, this in turn guarantees that (2.10) holds for the normal step

$$-A^T_{\mathcal{A}(x_*)}(x_k)\big[A_{\mathcal{A}(x_*)}(x_k)A^T_{\mathcal{A}(x_*)}(x_k)\big]^{-1} c_{\mathcal{A}(x_*)}(x_k)$$

for all k sufficiently large, provided that the sequence of iterates remains bounded, because this latter assumption ensures that x_k must be arbitrarily close to a least one limit point of the sequence $\{x_k\}$ for such k. Thus we see that (2.10) does not impose conditions on the constraints or the normal step itself that are unduly restrictive.

Having defined the normal step, we are in position to use it if it falls within the trust-region, that is, if $\|n_k\| \leq \Delta_k$. In this case, we write

$$x^{\mathrm{N}}_k = x_k + n_k \tag{2.11}$$

and observe that n_k satisfies the constraints of $\mathrm{TRQP}(x_k, \Delta_k)$ and thus also of $\mathrm{QP}(x_k)$. It is crucial to note, at this stage, that such an n_k may fail to exist because the constraints of $\mathrm{QP}(x_k)$ may be incompatible, in which case P_k is undefined, or because all feasible points for $QP(x_k)$ may lie outside the trust-region.

Let us continue to consider the case where this problem does not arise, and a normal step n_k has been found with $\|n_k\| \leq \Delta_k$. We then have to find a tangential step t_k, starting from x^{N}_k and satisfying (2.6) and (2.7), with the aim of decreasing the value of the objective function. As always in trust-region methods, this is achieved by computing a step that produces a sufficient decrease in m_k, which is to say that we wish $m_k(x^{\mathrm{N}}_k) - m_k(x_k + s_k)$ to be "sufficiently large." Of course, this is only possible if the maximum size of t_k is not too small, which is to say that x^{N}_k is not too close to the trust-region boundary. We formalize this supposition by replacing our condition that $\|n_k\| \leq \Delta_k$ with the stronger requirement that

$$\|n_k\| \leq \kappa_\Delta \Delta_k \min[1, \kappa_\mu \Delta_k^\mu] \tag{2.12}$$

for some $\kappa_\Delta \in (0, 1]$, some $\kappa_\mu > 0$, and some $\mu \in (0, 1)$. If condition (2.12) does not hold, we assume that the computation of t_k is unlikely to produce a satisfactory decrease in m_k, and proceed just as if the feasible set of $\mathrm{TRQP}(x_k, \Delta_k)$ were empty. If n_k can be computed and (2.12) holds, we shall say that $\mathrm{TRQP}(x_k, \Delta_k)$ is *compatible*. In this case at least a sufficient model decrease seems possible, which we state in the form of a familiar Cauchy-point condition. In order to formalize what we mean, we

recall that the feasible set of $\mathrm{QP}(x_k)$ is convex, and we can therefore introduce the measure

$$\chi_k = \left| \min_{\substack{A_{\mathcal{E}}(x_k)t=0 \\ c_{\mathcal{I}}(x_k)+A_{\mathcal{I}}(x_k)(n_k+t)\geq 0 \\ \|t\|\leq 1}} \langle g_k + H_k n_k, t\rangle \right| \tag{2.13}$$

(see Conn et al. [6]), which we will use to deduce first-order criticality for our problem (see Lemma 3.2). Note that this function is zero if the origin is a first-order critical point of the "tangential" problem

$$\begin{array}{ll} \text{minimize} & \langle g_k + H_k n_k, t\rangle + \frac{1}{2}\langle H_k t, t\rangle \\ \text{subject to} & A_{\mathcal{E}}(x_k)t = 0, \\ & c_{\mathcal{I}}(x_k) + A_{\mathcal{I}}(x_k)(n_k + t) \geq 0, \end{array} \tag{2.14}$$

which is, up to the constant term $\frac{1}{2}\langle n_k, H_k n_k\rangle$, equivalent to $\mathrm{QP}(x_k)$ with $s = n_k + t$. Our sufficient decrease condition is then to require that, whenever $\mathrm{TRQP}(x_k, \Delta_k)$ is compatible,

$$m_k(x_k^{\mathrm{N}}) - m_k(x_k^{\mathrm{N}} + t_k) \geq \kappa_{\mathrm{tmd}} \chi_k \min\left[\frac{\chi_k}{\beta_k}, \Delta_k\right] \tag{2.15}$$

for some constant $\kappa_{\mathrm{tmd}} > 0$, where $\beta_k = 1 + \|H_k\|$. We know from Toint [29] and Conn et al. [6] that this condition holds if the model reduction exceeds that which would be obtained at the generalized Cauchy point, that is, the point resulting from a backtracking curvilinear search along the projected gradient path from x_k^{N}, that is,

$$x_k(\alpha) = P_k[x_k^{\mathrm{N}} - \alpha \nabla_x m_k(x_k^{\mathrm{N}})].$$

This technique therefore provides an implementable algorithm for computing a step that satisfies (2.15) (see Gould, Hribar, and Nocedal [21] for an example in the case where $c(x) = c_{\mathcal{E}}(x)$, or Toint [29] and Moré and Toraldo [24] for the case of bound constraints), but, of course, reduction of m_k beyond that imposed by (2.15) is often possible and desirable if fast convergence is sought. Also note that the minimization problem of the right-hand side of (2.13) would reduce to a linear programming problem if we had chosen to use a polyhedral norm in its definition at iteration k.

Let us now return to the case where $\mathrm{TRQP}(x_k, \Delta_k)$ is not compatible, that is, when the feasible set determined by the constraints of $\mathrm{QP}(x_k)$ is empty, or the freedom left to reduce m_k within the trust-region is too small in the sense that (2.12) fails. In this situation, solving $\mathrm{TRQP}(x_k, \Delta_k)$ is most likely pointless, and we must consider an alternative. We base this on the intuitive observation that, if $\theta(x_k)$ is sufficiently small and the true nonlinear constraints are locally compatible, the linearized constraints should also be compatible, since they approximate the nonlinear constraints (locally) correctly. Furthermore, the feasible region for the linearized constraints should then be close enough to x_k for there to be some room to reduce m_k, at least if Δ_k is large enough. If the nonlinear constraints are locally incompatible, we have to find a neighborhood where this is not the case, since the problem (1.1) does not make sense in the current one. We thus rely on a *restoration procedure*, whose aim is to produce a new point $x_k + r_k$ for which $\mathrm{TRQP}(x_k + r_k, \Delta_{k+1})$ is compatible for some $\Delta_{k+1} > 0$—we will actually need another condition which we will discuss shortly.

The idea of the restoration procedure is to (approximately) solve

$$\min_{x\in\mathbb{R}^n} \theta(x), \tag{2.16}$$

perhaps starting from x_k, the current iterate. This is a nonsmooth problem, but we know that there exist methods, possibly of trust-region type (such as that suggested by Yuan [32]), which can be successfully applied to solve it. Thus we will not describe the restoration procedure in detail. Note that we have chosen here to reduce the infinity norm of the constraint violation, but we could equally well consider other norms, such as ℓ_1 or ℓ_2, in which case the methods of Fletcher and Leyffer [17] or of El-Hallabi and Tapia [14] and Dennis, El-Alem, and Williamson [10], respectively, can be considered. Of course, this technique only guarantees convergence to a first-order critical point of the chosen measure of constraint violation, which means that, in fact, the restoration procedure may fail as this critical point may not be feasible for the constraints of (1.1). However, even in this case, the result of the procedure is of interest because it typically produces a local minimizer of $\theta(x)$, or of whatever other measure of constraint violation we choose for the restoration, yielding a point of locally least infeasibility.

There is no easy way to circumvent this drawback, as it is known that finding a feasible point or proving that no such point exists is a global optimization problem and can be as difficult as the optimization problem (1.1) itself. We therefore accept two possible outcomes of the restoration procedure: either the procedure fails in that it does not produce a sequence of iterates converging to feasibility, or a point $x_k + r_k$ is produced such that $\theta(x_k + r_k)$ is as small as we wish. We will shortly see that this is all we need.

2.2. The notion of a filter. Having computed a step $s_k = n_k + t_k$ (or r_k), we still need to decide whether the trial point $x_k + s_k$ (or $x_k + r_k$) is any better than x_k as an approximate solution to our original problem (1.1). We shall use a concept borrowed from multicriteria optimization. We say that a point x_1 *dominates* a point x_2 whenever

$$\theta(x_1) \leq \theta(x_2) \;\text{ and }\; f(x_1) \leq f(x_2).$$

Thus, if iterate x_k dominates iterate x_j, the latter is of no real interest to us since x_k is at least as good as x_j on account of both feasibility and optimality. All we need to do now is to remember iterates that are not dominated by any other iterates using a structure called a filter. A *filter* is a list $\mathcal{F}$ of pairs of the form (θ_i, f_i) such that either

$$\theta_i < \theta_j \;\text{ or }\; f_i < f_j$$

for $i \neq j$. We thus aim to accept a new iterate x_i only if it is not dominated by any other iterate in the filter. In the vocabulary of multicriteria optimization, this amounts to building elements of the efficient frontier associated with the bicriteria problem of reducing infeasibility and the objective function value.

Figure 2.1 illustrates the concept of a filter by showing the pairs (θ_k, f_k) as black dots in the (θ, f)-space. Each such pair is called the (θ, f)-pair associated with x_k. The lines radiating from each (θ, f)-pair indicate that any iterate whose associated (θ, f)-pair occurs above and to the right of that of a given filter point is dominated by this (θ, f)-pair.

While the idea of not accepting dominated trial points is simple and elegant, it needs to be refined a little in order to provide an efficient algorithmic tool. In

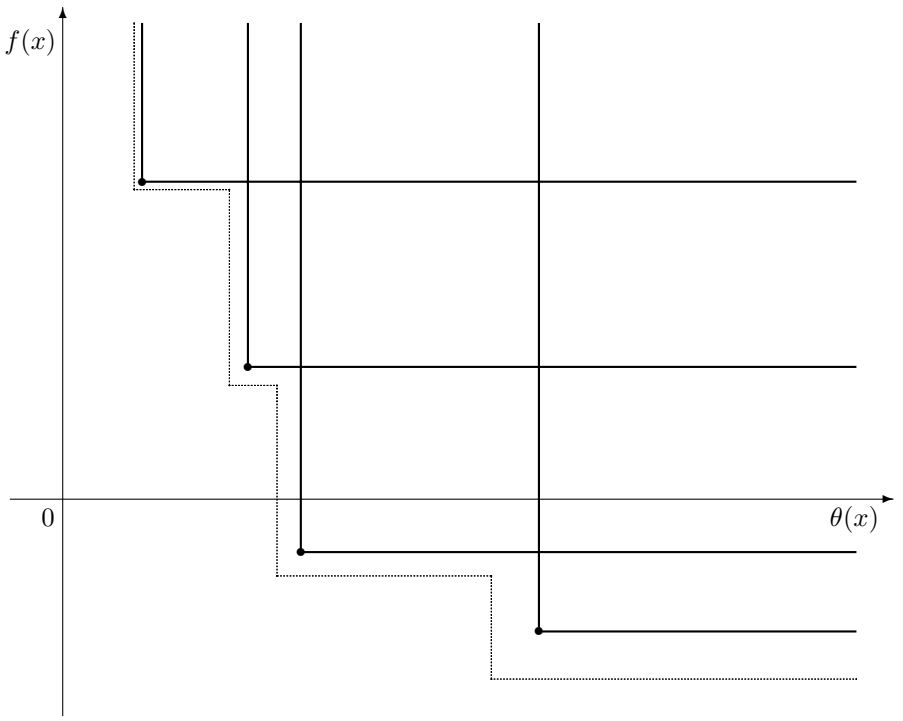


FIG. 2.1. *A filter with four pairs.*

particular, we do not wish to accept $x_k + s_k$ if its (θ, f)-pair is arbitrarily close to that of x_k or that of a point already in the filter. Thus we set a small "margin" around the border of the dominated part of the (θ, f)-space in which we shall also reject trial points. Formally, we say that a point x is *acceptable for the filter* if and only if

$$\theta(x) \leq (1 - \gamma_\theta)\theta_j \ \text{ or } \ f(x) \leq f_j - \gamma_\theta \theta_j \ \text{ for all } \ (\theta_j, f_j) \in \mathcal{F} \tag{2.17}$$

for some $\gamma_\theta \in (0, 1)$. In Figure 2.1, the set of acceptable points corresponds to the set of (θ, f)-pairs below the thin line. We also say that x is "acceptable for the filter and x_k" if (2.17) holds with $\mathcal{F}$ replaced by $\mathcal{F} \cup (\theta_k, f_k)$. We thus consider moving from x_k to $x_k + s_k$ only if $x_k + s_k$ is acceptable for the filter and x_k.

As the algorithm progresses, we may want to *add a* (θ, f)*-pair to the filter*. If an iterate x_k is acceptable for $\mathcal{F}$, we do this by adding the pair (θ_k, f_k) to the filter and by removing from it every other pair (θ_j, f_j) such that both

$$\theta_j \geq \theta_k \ \text{ and } \ f_j - \gamma_\theta \theta_j \geq f_k - \gamma_\theta \theta_k. \tag{2.18}$$

Only entries whose *envelope* is dominated by a new entry are thus removed from the filter. As a consequence, the margin of the filter never decreases, and it can be shown that, for all infinite subsequences of points added to the filter, $\lim \theta_{k_i} = 0$ (see Lemma 3.3). We also refer to this operation as "adding x_k to the filter," although, strictly speaking, it is the (θ, f)-pair which is added.

We conclude this section by noting that, if a point x_k is in the filter or is acceptable for the filter, then any other point x such that

$$\theta(x) \leq (1 - \gamma_\theta)\theta_k \ \text{ and } \ f(x) \leq f_k - \gamma_\theta \theta_k$$

is also acceptable for the filter and x_k.

2.3. An SQP-filter algorithm. We have now discussed the main ingredients of the class of algorithms we wish to consider, and we are now ready to define it formally as Algorithm 2.1 below. A flowchart of the algorithm is given as an appendix; see Figure A.1.

ALGORITHM 2.1: SQP-FILTER ALGORITHM.

Step 0: Initialization. Let an initial point x_0, an initial trust-region radius $\Delta_0 > 0$, and an initial symmetric matrix H_0 be given, as well as constants $0 < \gamma_0 < \gamma_1 \leq 1 \leq \gamma_2$, $0 < \eta_1 \leq \eta_2 < 1$, $\gamma_\theta \in (0,1)$, $\kappa_\theta \in (0,1)$, $\kappa_\Delta \in (0,1]$, $\kappa_\mu > 0$, $\mu \in (0,1)$, $\psi > 1/(1+\mu)$, and $\kappa_{\text{tmd}} \in (0,1]$. Compute $f(x_0)$ and $c(x_0)$. Set $\mathcal{F} = \emptyset$ and $k = 0$.

Step 1: Test for optimality. If $\theta_k = \chi_k = 0$, stop.

Step 2: Ensure compatibility. Attempt to compute a step n_k. If TRQP (x_k, Δ_k) is compatible, go to Step 3. Otherwise, include x_k in the filter and compute a restoration step r_k for which TRQP$(x_k + r_k, \Delta_{k+1})$ is compatible for some $\Delta_{k+1} > 0$, and $x_k + r_k$ is acceptable for the filter. If this proves impossible, stop. Otherwise, define $x_{k+1} = x_k + r_k$ and go to Step 7.

Step 3: Determine a trial step. Compute a step t_k and set $s_k = n_k + t_k$.

Step 4: Tests to accept the trial step.

- Evaluate $c(x_k + s_k)$ and $f(x_k + s_k)$.
- If $x_k + s_k$ is not acceptable for the filter and x_k, set $x_{k+1} = x_k$, choose $\Delta_{k+1} \in [\gamma_0 \Delta_k, \gamma_1 \Delta_k]$, set $n_{k+1} = n_k$, increment k by one, and go to Step 2.
- If

$$m_k(x_k) - m_k(x_k + s_k) \geq \kappa_\theta \theta_k^\psi \tag{2.19}$$

and

$$\rho_k \stackrel{\text{def}}{=} \frac{f(x_k) - f(x_k + s_k)}{m_k(x_k) - m_k(x_k + s_k)} < \eta_1, \tag{2.20}$$

again set $x_{k+1} = x_k$, choose $\Delta_{k+1} \in [\gamma_0 \Delta_k, \gamma_1 \Delta_k]$, set $n_{k+1} = n_k$, increment k by one, and go to Step 2.

Step 5: Test to include the current iterate in the filter. If (2.19) fails, include x_k in the filter $\mathcal{F}$.

Step 6: Move to the new iterate. Set $x_{k+1} = x_k + s_k$ and choose Δ_{k+1} such that

$$\Delta_{k+1} \in [\Delta_k, \gamma_2 \Delta_k] \text{ if } \rho_k \geq \eta_2 \text{ and (2.19) holds.}$$

Step 7: Update the Hessian approximation. Determine H_{k+1}. Increment k by one and go to Step 1.

As in Fletcher and Leyffer [18, 17], one may choose $\psi = 2$. (Note that the choice $\psi = 1$ is always possible because $\mu > 0$.) Reasonable values for the constants might then be

$$\gamma_0 = 0.1, \quad \gamma_1 = 0.5, \quad \gamma_2 = 2, \quad \eta_1 = 0.01, \quad \eta_2 = 0.9,$$
$$\gamma_\theta = 10^{-4}, \quad \kappa_\Delta = 0.7, \quad \kappa_\mu = 100, \quad \mu = 0.01, \quad \kappa_\theta = 10^{-4}, \text{ and } \kappa_{\text{tmd}} = 0.01,$$

but it is too early to know if these are even close to the best possible choices.

Observe first that, by construction, every iterate x_k must be acceptable for the filter at the beginning of iteration k, irrespective of the possibility that it might be added to the filter later. Also note that the restoration step r_k cannot be zero, that is, restoration cannot simply entail enlarging the trust-region radius to ensure (2.12), even if n_k exists. This is because x_k is added to the filter before r_k is computed, and $x_k + r_k$ must be acceptable for the filter which now contains x_k. Also note that the restoration procedure cannot be applied on two successive iterations, since the iterate $x_k + r_k$ produced by the first of these iterations leads to a compatible $\mathrm{TRQP}(x_{k+1}, \Delta_{k+1})$ and is acceptable for the filter.

For the restoration procedure in Step 2 to succeed, we have to evaluate whether $\mathrm{TRQP}(x_k + r_k, \Delta_{k+1})$ is compatible for a suitable value of Δ_{k+1}. This requires that a suitable normal step be computed which successfully passes the test (2.12). Of course, once this is achieved, this normal step may be reused at iteration $k+1$. Thus we shall require that the normal step calculated to verify compatibility of $\mathrm{TRQP}(x_k+r_k, \Delta_{k+1})$ should actually be used as n_{k+1}.

As it stands, the algorithm is not specific about how to choose Δ_{k+1} during a restoration iteration. On one hand, there is an advantage to choosing a large Δ_{k+1}, since this allows a large step and, one hopes, good progress. On the other hand, it may be unwise to choose it to be too large, as this may possibly result in a large number of unsuccessful iterations, during which the radius is reduced, before the algorithm can make any progress. A possible choice might be to restart from the radius obtained during the restoration iteration itself, if it uses a trust-region method. Reasonable alternatives would be to use the average radius observed during past successful iterations, or to apply the internal doubling strategy of Byrd, Schnabel, and Shultz [5] to increase the new radius, or even to consider the technique described by Sartenaer [28]. However, we recognize that numerical experience with the algorithm is too limited at this stage to make definite recommendations.

The role of condition (2.19) may be interpreted as follows. If this condition fails, then one may think that the constraint violation is significant and that one should aim to improve on this situation in the future by inserting the current point into the filter. Fletcher, Leyffer, and Toint [19] use the term of "θ-step" in such circumstances to indicate that the main preoccupation is to improve feasibility. On the other hand, if condition (2.19) holds, then the reduction in the objective function predicted by the model is more significant than the current constraint violation, and it is thus appealing to let the algorithm behave as if it were unconstrained. Fletcher and Leyffer [18] use the term "f-step" to denote the step generated, in order to reflect the dominant role of the objective function f. In this case, it is important that the predicted decrease in the model be realized by the actual decrease in the function, which is why we also require that (2.20) not hold. In particular, if the iterate x_k is feasible, then (2.10) implies that $x_k = x_k^{\mathrm{N}}$, and we obtain that

$$\kappa_\theta \theta_k^\psi = 0 \leq m_k(x_k^{\mathrm{N}}) - m_k(x_k + s_k) = m_k(x_k) - m_k(x_k + s_k). \tag{2.21}$$

As a consequence, the filter mechanism is irrelevant if all iterates are feasible, and the algorithm reduces to a classical unconstrained trust-region method. Another consequence of (2.21) is that *no feasible iterate is ever included in the filter*, which is crucial in allowing finite termination of the restoration procedure. Indeed, if the restoration procedure is required at iteration k of the filter algorithm and produces a sequence of points $\{x_{k,j}\}$ converging to feasibility, there must be an iterate $x_{k,j}$ for

which

$$\theta_{k,j} \stackrel{\text{def}}{=} \theta(x_{k,j}) \leq \min\left[(1-\gamma_\theta)\theta_k^{\min}, \frac{\kappa_\Delta}{\kappa_{\text{usc}}}\Delta_{k+1}\min[1, \kappa_\mu \Delta_{k+1}^{\mu}]\right]$$

for any given $\Delta_{k+1} > 0$, where

$$\theta_k^{\min} = \min_{i \in \mathcal{Z},\, i \leq k} \theta_i > 0$$

and

$$\mathcal{Z} = \{k \mid x_k \text{ is added to the filter}\}.$$

Moreover, $\theta_{k,j}$ must eventually be small enough to ensure, using our assumption on the normal step, the existence of a normal step $n_{k,j}$ from $x_{k,j}$. In other words, the restoration iteration must eventually find an iterate $x_{k,j}$ which is acceptable for the filter and for which the normal step exists and satisfies (2.12), i.e., an iterate x_j which is both acceptable and compatible. As a consequence, the restoration procedure will terminate in a finite number of steps, and the filter algorithm may then proceed. Note that the restoration step may not terminate in a finite number of iterations if we do not assume the existence of the normal step when the constraint violation is small enough, even if this violation converges to zero (see Fletcher, Leyffer, and Toint [19], for an example).

Notice also that (2.19) ensures that the denominator of ρ_k in (2.20) will be strictly positive whenever θ_k is. If $\theta_k = 0$, then $x_k = x_k^{\text{N}}$, and the denominator of (2.20) will be strictly positive unless x_k is a first-order critical point because of (2.15).

The reader may have observed that Step 6 allows a relatively wide choice of the new trust-region radius Δ_{k+1}. While the stated conditions appear to be sufficient for the theory developed below, one must obviously be more specific in practice. For instance, one may wish to distinguish, at this point in the algorithm, the cases where (2.19) fails or holds. If (2.19) fails, the main effect of the current iteration is not to reduce the model (which makes the value of ρ_k essentially irrelevant), but rather to reduce the constraint violation (which is taken care of by inserting the current iterate into the filter at Step 5). In this case, Step 6 imposes no further restriction on Δ_{k+1}. In practice, it may be reasonable not to reduce the trust-region radius, because this might cause too small steps towards feasibility or an unnecessary restoration phase. However, there is no compelling reason to increase the radius either, given the compatibility of TRQP(x_k,Δ_k). A reasonable strategy might then be to choose $\Delta_{k+1} = \Delta_k$. If, on the other hand, (2.19) holds, the emphasis of the iteration is then on reducing the objective function, a case akin to unconstrained minimization. Thus a more detailed rule of the type

$$\Delta_{k+1} \in \begin{cases} [\gamma_1 \Delta_k, \gamma_2 \Delta_k] & \text{if } \rho_k \in [\eta_1, \eta_2), \\ [\Delta_k, \gamma_2 \Delta_k] & \text{if } \rho_k \geq \eta_2 \end{cases}$$

seems reasonable in these circumstances.

Finally, we recognize that (2.15) may be difficult to verify in practice, since it may be expensive to compute x_k^{N} and P_k when the dimension of the problem is large.

3. Convergence to first-order critical points. We now prove that our algorithm generates a globally convergent sequence of iterates. In the following analysis, we concentrate on the case in which the restoration iteration always succeeds. If this

is not the case, then it usually follows that the restoration phase has converged to an approximate solution of the feasibility problem (2.16) and we can conclude that (1.1) is locally inconsistent. For the purpose of our analysis, we shall consider

$$\mathcal{S} = \{k \mid x_{k+1} = x_k + s_k\},$$

the set of (indices of) successful iterations, and

$$\mathcal{R} = \{k \mid n_k \text{ does not satisfy (2.10)} \quad \text{or} \quad \|n_k\| > \kappa_\Delta \Delta_k \min[1, \kappa_\mu \Delta_k^\mu]\},$$

the set of *restoration* iterations. In order to obtain our global convergence result, we will use the following assumptions.

AS1. f and the constraint functions $c_\mathcal{E}$ and $c_\mathcal{I}$ are twice continuously differentiable.

AS2. There exists $\kappa_{\rm umh} > 1$ such that

$$\|H_k\| \leq \kappa_{\rm umh} - 1 \text{ for all } k.$$

AS3. The iterates $\{x_k\}$ remain in a closed, bounded domain $X \subset \mathbb{R}^n$.

If, for example, H_k is chosen as the Hessian of the Lagrangian function

$$\ell(x, y) = f(x) + \langle y_\mathcal{E}, c_\mathcal{E}(x)\rangle + \langle y_\mathcal{I}, c_\mathcal{I}(x)\rangle$$

at x_k, in that

$$H_k = \nabla_{xx} f(x_k) + \sum_{i \in \mathcal{E} \cup \mathcal{I}} [y_k]_i \nabla_{xx} c_i(x_k), \tag{3.1}$$

where $[y_k]_i$ denotes the ith component of the vector of Lagrange multipliers $y_k^T = (y_{\mathcal{E},k}^T \;\; y_{\mathcal{I},k}^T)$, then we see from AS1 and AS3 that AS2 is satisfied when these multipliers remain bounded. The same is true if the Hessian matrices in (3.1) are replaced by bounded approximations.

A first immediate consequence of AS1–AS3 is that there exists a constant $\kappa_{\rm ubh} > 1$ such that, for all k,

$$|f(x_k + s_k) - m_k(x_k + s_k)| \leq \kappa_{\rm ubh} \Delta_k^2. \tag{3.2}$$

A proof of this property, based on Taylor expansion, may be found, for instance, in Toint [29]. A second important consequence of our assumptions is that AS1 and AS3 together directly ensure that, for all k,

$$f^{\rm min} \leq f(x_k) \text{ and } 0 \leq \theta_k \leq \theta^{\rm max} \tag{3.3}$$

for some constants $f^{\rm min}$ and $\theta^{\rm max} > 0$. Thus the part of the (θ, f)-space in which the (θ, f)-pairs associated with the filter iterates lie is restricted to the rectangle

$$\mathcal{A}_0 = [0, \theta^{\rm max}] \times [f^{\rm min}, \infty].$$

We also note the following simple consequence of (2.10) and AS3.

LEMMA 3.1. *Suppose that Algorithm* 2.1 *is applied to problem* (1.1). *Suppose also that* (2.10) *and* AS3 *hold and that*

$$\theta_k \leq \delta_n.$$

Then there exists a constant $\kappa_{lsc} > 0$ independent of k such that

$$\kappa_{lsc}\theta_k \leq \|n_k\|. \tag{3.4}$$

Proof. First define

$$\mathcal{V}_k \stackrel{\rm def}{=} \{j \in \mathcal{E} \mid \theta_k = |c_j(x_k)|\} \bigcup \{j \in \mathcal{I} \mid \theta_k = -c_j(x_k)\},$$

which is the subset of most-violated constraints. From the definitions of θ_k in (2.9) and of the normal step in (2.5) we obtain, using the Cauchy–Schwarz inequality, that

$$\theta_k \leq |\langle \nabla_x c_j(x_k), n_k\rangle| \leq \|\nabla_x c_j(x_k)\| \, \|n_k\| \tag{3.5}$$

for all $j \in \mathcal{V}_k$. But AS3 ensures that there exists a constant $\kappa_{\rm lsc} > 0$ such that

$$\max_{j\in\mathcal{E}\cup\mathcal{I}} \max_{x\in X} \|\nabla_x c_j(x)\| \stackrel{\rm def}{=} \frac{1}{\kappa_{\rm lsc}}.$$

We then obtain the desired conclusion by substituting this bound into (3.5). ∎

Our assumptions and the definition of χ_k in (2.13) ensure that θ_k and χ_k can be used (together) to measure criticality for problem (1.1).

LEMMA 3.2. *Suppose that Algorithm* 2.1 *is applied to problem* (1.1) *and that finite termination does not occur. Suppose also that* AS1 *and* AS3 *hold, and that there exists a subsequence $\{k_i\}$ such that, for any i, $k_i \notin \mathcal{R}$ with*

$$\lim_{i\to\infty} \chi_{k_i} = 0 \quad \textit{and} \quad \lim_{i\to\infty} \theta_{k_i} = 0. \tag{3.6}$$

Then every limit point of the subsequence $\{x_{k_i}\}$ is a first-order critical point for problem (1.1).

Proof. Consider x_*, a limit point of the sequence $\{x_{k_i}\}$, whose existence is ensured by AS3, and assume that $\{k_\ell\} \subseteq \{k_i\}$ is the index set of a subsequence such that $\{x_{k_\ell}\}$ converges to x_*. The fact that $k_\ell \notin \mathcal{R}$ implies that n_{k_ℓ} satisfies (2.10) for sufficiently large ℓ and converges to zero, because $\{\theta_{k_\ell}\}$ converges to zero and the second part of this condition. As a consequence, we deduce from (2.11) that $\{x^{\rm N}_{k_\ell}\}$ also converges to x_*. Since the minimization problem occurring in the definition of χ_{k_ℓ} (in (2.13)) is convex, we then obtain from classical perturbation theory (see, for instance, Fiacco [15, pp. 14–17], AS1, and the first part of (3.6) that

$$\left| \min_{\substack{A_{\mathcal{E}}(x_*)t=0 \\ c_{\mathcal{I}}(x_*)+A_{\mathcal{I}}(x_*)t\geq 0 \\ \|t\|\leq 1}} \langle g_*, t\rangle \right| = 0.$$

This in turn guarantees that x_* is first-order critical for problem (1.1). ∎

We start our analysis by examining what happens when an infinite number of iterates (that is, their (θ, f)-pairs) are added to the filter.

LEMMA 3.3. *Suppose that Algorithm* 2.1 *is applied to problem* (1.1) *and that finite termination does not occur. Suppose that* AS1 *and* AS3 *hold and that $|\mathcal{Z}| = \infty$. Then*

$$\lim_{\substack{k\to\infty \\ k\in\mathcal{Z}}} \theta_k = 0.$$

Proof. Suppose, for the purpose of obtaining a contradiction, that there exists an infinite subsequence $\{k_i\} \subseteq \mathcal{Z}$ such that

$$\theta_{k_i} \geq \epsilon \tag{3.7}$$

for all i and for some $\epsilon > 0$. At each iteration k_i, the (θ, f)-pair associated with x_{k_i}, that is (θ_{k_i}, f_{k_i}), is added to the filter. This means that no other (θ, f)-pair can be added to the filter at a later stage within the square

$$[\theta_{k_i} - \gamma_\theta \epsilon, \theta_{k_i}] \times [f_{k_i} - \gamma_\theta \epsilon, f_{k_i}]$$

or with the intersection of this square with $\mathcal{A}_0$. Note that this holds, even if (θ_{k_i}, f_{k_i}) is later removed from the filter, since the rule for removing entries, (2.18), ensures that the envelope never shrinks. Now observe that the area of each of these squares is $\gamma_\theta^2 \epsilon^2$. As a consequence, the set $\mathcal{A}_0 \cap \{(\theta, f) | f \leq \kappa_f\}$ is completely covered by at most a finite number of such squares, for any choice of $\kappa_f \geq f^{\min}$. Since the pairs (θ_{k_i}, f_{k_i}) keep on being added to the filter, this implies that f_{k_i} tends to infinity when i tends to infinity. Let us assume, without loss of generality, that $f_{k_{i+1}} \geq f_{k_i}$ for all i sufficiently large. But (2.17) and (3.7) then imply that

$$\theta_{k_{i+1}} \leq (1 - \gamma_\theta)\theta_{k_i} \leq \theta_{k_i} - \gamma_\theta \epsilon,$$

and therefore that θ_{k_i} converges to zero, which contradicts (3.7). Hence this latter assumption is imposssible and the conclusion follows. ∎

We next examine the size of the constraint violation before and after an iteration where restoration did not occur.

LEMMA 3.4. *Suppose that Algorithm* 2.1 *is applied to problem* (1.1) *and that finite termination does not occur. Suppose also that* AS1 *and* AS3 *hold, that* $k \notin \mathcal{R}$, *and that* n_k *satisfies* (3.4). *Then*

$$\theta_k \leq \kappa_{ubt} \Delta_k^{1+\mu} \tag{3.8}$$

and

$$\theta(x_k + s_k) \leq \kappa_{ubt} \Delta_k^2 \tag{3.9}$$

for some constant $\kappa_{ubt} \geq 0$.

Proof. Since $k \notin \mathcal{R}$, we have from (3.4) and (2.12) that

$$\kappa_{\text{lsc}} \theta_k \leq \|n_k\| \leq \kappa_\Delta \kappa_\mu \Delta_k^{1+\mu}, \tag{3.10}$$

which gives (3.8). Now, the ith constraint function at $x_k + s_k$ can be expressed as

$$c_i(x_k + s_k) = c_i(x_k) + \langle e_i, A_k s_k \rangle + \frac{1}{2} \langle s_k, \nabla_{xx} c_i(\xi_k) s_k \rangle$$

for $i \in \mathcal{E} \cup \mathcal{I}$, where we have used AS1 and the mean-value theorem and where ξ_k belongs to the segment $[x_k, x_k + s_k]$. Using AS3, we may bound the Hessian of the constraint functions, and we obtain from (2.7), the Cauchy–Schwarz inequality, and (2.6) that

$$|c_i(x_k + s_k)| \leq \frac{1}{2} \max_{x \in X} \|\nabla_{xx} c_i(x)\| \, \|s_k\|^2 \leq \kappa_1 \Delta_k^2$$

if $i \in \mathcal{E}$, or

$$-c_i(x_k + s_k) \leq \frac{1}{2} \max_{x \in X} \|\nabla_{xx} c_i(x)\| \, \|s_k\|^2 \leq \kappa_1 \Delta_k^2$$

if $i \in \mathcal{I}$, where we have defined

$$\kappa_1 \stackrel{\text{def}}{=} \frac{1}{2} \max_{i \in \mathcal{E} \cup \mathcal{I}} \max_{x \in X} \|\nabla_{xx} c_i(x)\|.$$

This gives the desired bound with

$$\kappa_{\text{ubt}} = \max[\kappa_1, \kappa_\Delta \kappa_\mu / \kappa_{\text{lsc}}]. \qquad \square$$

We next assess the model decrease when the trust-region radius is sufficiently small.

LEMMA 3.5. *Suppose that Algorithm* 2.1 *is applied to problem* (1.1) *and that finite termination does not occur. Suppose also that* AS1–AS3, (2.12), *and* (2.15) *hold, that* $k \notin \mathcal{R}$, *that*

$$\chi_k \geq \epsilon \tag{3.11}$$

for some $\epsilon > 0$, *and that*

$$\Delta_k \leq \min \left[\frac{\epsilon}{\kappa_{umh}}, \left(2 \frac{\kappa_{ubg}}{\kappa_{umh} \kappa_\Delta \kappa_\mu} \right)^{\frac{1}{1+\mu}}, \left(\frac{\kappa_{tmd} \epsilon}{4 \kappa_{ubg} \kappa_\Delta \kappa_\mu} \right)^{\frac{1}{\mu}} \right] \stackrel{\text{def}}{=} \delta_m, \tag{3.12}$$

where $\kappa_{ubg} \stackrel{\text{def}}{=} \max_{x \in X} \|\nabla_x f(x)\|$. *Then*

$$m_k(x_k) - m_k(x_k + s_k) \geq \frac{1}{2} \kappa_{tmd} \epsilon \Delta_k.$$

Proof. We first note that, by (2.15), AS2, (3.11), and (3.12),

$$m_k(x_k^{\text{N}}) - m_k(x_k + s_k) \geq \kappa_{\text{tmd}} \chi_k \min \left[\frac{\chi_k}{\kappa_{\text{umh}}}, \Delta_k \right] \geq \kappa_{\text{tmd}} \epsilon \Delta_k. \tag{3.13}$$

Now

$$m_k(x_k^{\text{N}}) = m_k(x_k) + \langle g_k, n_k \rangle + \frac{1}{2} \langle n_k, H_k n_k \rangle,$$

and therefore, using the Cauchy–Schwarz inequality, AS2, (2.12), and (3.12),

$$\begin{aligned} |m_k(x_k) - m_k(x_k^{\text{N}})| &\leq \|n_k\| \, \|g_k\| + \tfrac{1}{2} \|H_k\| \, \|n_k\|^2 \\ &\leq \kappa_{\text{ubg}} \|n_k\| + \tfrac{1}{2} \kappa_{\text{umh}} \|n_k\|^2 \\ &\leq \kappa_{\text{ubg}} \kappa_\Delta \kappa_\mu \Delta_k^{1+\mu} + \tfrac{1}{2} \kappa_{\text{umh}} \kappa_\Delta^2 \kappa_\mu^2 \Delta_k^{2(1+\mu)} \\ &\leq 2 \kappa_{\text{ubg}} \kappa_\Delta \kappa_\mu \Delta_k^{1+\mu} \\ &\leq \tfrac{1}{2} \kappa_{\text{tmd}} \epsilon \Delta_k. \end{aligned}$$

We thus conclude from this last inequality and (3.13) that the desired conclusion holds. $\square$

We continue our analysis by showing, as the reader has grown to expect, that iterations have to be very successful when the trust-region radius is sufficiently small.

LEMMA 3.6. *Suppose that Algorithm* 2.1 *is applied to problem* (1.1) *and that finite termination does not occur. Suppose also that* AS1–AS3, (2.15), *and* (3.11) *hold, that* $k \notin \mathcal{R}$, *and that*

$$\Delta_k \leq \min\left[\delta_m, \frac{(1-\eta_2)\kappa_{tmd}\epsilon}{2\kappa_{ubh}}\right] \stackrel{\text{def}}{=} \delta_\rho. \tag{3.14}$$

Then

$$\rho_k \geq \eta_2.$$

Proof. Using the definition of ρ_k in (2.20), (3.2), Lemma 3.5, and (3.14), we find that

$$|\rho_k - 1| = \frac{|f(x_k+s_k) - m_k(x_k+s_k)|}{|m_k(x_k) - m_k(x_k+s_k)|} \leq \frac{\kappa_{\text{ubh}}\Delta_k^2}{\frac{1}{2}\kappa_{\text{tmd}}\epsilon\Delta_k} \leq 1-\eta_2,$$

from which the conclusion immediately follows. ∎

Note that this proof could easily be extended if the definition of ρ_k in (2.20) were altered to be of the form

$$\rho_k \stackrel{\text{def}}{=} \frac{f(x_k) - f(x_k+s_k) + \Theta_k}{m_k(x_k) - m_k(x_k+s_k)}, \tag{3.15}$$

provided that Θ_k is bounded above by a multiple of Δ_k^2. We will comment in section 4 why such a modification might be of interest (see also section 10.4.3 of Conn, Gould, and Toint [7]).

Now, we also show that the test (2.19) will always be satisfied when the trust-region radius is sufficiently small.

LEMMA 3.7. *Suppose that Algorithm* 2.1 *is applied to problem* (1.1) *and that finite termination does not occur. Suppose also that* AS1–AS3, (2.12), (2.15), *and* (3.11) *hold, that* $k \notin \mathcal{R}$, *that* n_k *satisfies* (3.4), *and that*

$$\Delta_k \leq \min\left[\delta_m, \left(\frac{\kappa_{tmd}\epsilon}{2\kappa_\theta\kappa_{ubt}^\psi}\right)^{\frac{1}{\psi(1+\mu)-1}}\right] \stackrel{\text{def}}{=} \delta_f. \tag{3.16}$$

Then

$$m_k(x_k) - m_k(x_k+s_k) \geq \kappa_\theta\theta_k^\psi.$$

Proof. This directly results from the inequalities

$$\kappa_\theta\theta_k^\psi \leq \kappa_\theta\kappa_{\text{ubt}}^\psi\Delta_k^{\psi(1+\mu)} \leq \frac{1}{2}\kappa_{\text{tmd}}\epsilon\Delta_k \leq m_k(x_k) - m_k(x_k+s_k),$$

where we have successively used Lemma 3.4, (3.16), and Lemma 3.5. ∎

We may also guarantee a decrease in the objective function, large enough to ensure that the trial point is acceptable with respect to the (θ, f)-pair associated with x_k, so long as the constraint violation is itself sufficiently small.

LEMMA 3.8. *Suppose that Algorithm* 2.1 *is applied to problem* (1.1) *and that finite termination does not occur. Suppose also that* AS1–AS3, (2.15), (3.11), *and* (3.14) *hold, that* $k \notin \mathcal{R}$, *that* n_k *satisfies* (3.4), *and that*

$$\theta_k \leq \kappa_{ubt}^{-\frac{1}{\mu}}\left(\frac{\eta_2\kappa_{tmd}\epsilon}{2\gamma_\theta}\right)^{\frac{1+\mu}{\mu}} \stackrel{\text{def}}{=} \delta_\theta. \tag{3.17}$$

Then

$$f(x_k + s_k) \leq f(x_k) - \gamma_\theta \theta_k.$$

Proof. Applying Lemmas 3.4–3.6—which is possible because of (3.11), (3.14), $k \notin \mathcal{R}$, and the fact that n_k satisfies (3.4)—and (3.17), we obtain that

$$\begin{aligned} f(x_k) - f(x_k + s_k) &\geq \eta_2[m_k(x_k) - m_k(x_k + s_k)] \\ &\geq \tfrac{1}{2}\eta_2 \kappa_{\rm tmd} \epsilon \Delta_k \\ &\geq \tfrac{1}{2}\eta_2 \kappa_{\rm tmd} \epsilon \left(\frac{\theta_k}{\kappa_{\rm ubt}}\right)^{\frac{1}{1+\mu}} \\ &\geq \gamma_\theta \theta_k, \end{aligned}$$

and the desired inequality follows. ∎

We now establish that if the trust-region radius and the constraint violation are both small at a noncritical iterate x_k, TRQP(x_k, Δ_k) must be compatible.

LEMMA 3.9. *Suppose that Algorithm* 2.1 *is applied to problem* (1.1) *and that finite termination does not occur. Suppose also that* AS1–AS3, (2.10), *and* (3.11) *hold, that* (2.15) *holds for $k \notin \mathcal{R}$, and that*

$$\Delta_k \leq \min\left[\gamma_0 \delta_\rho, \left(\frac{1}{\kappa_\mu}\right)^{\frac{1}{\mu}}, \left(\frac{\gamma_0^2(1-\gamma_\theta)\kappa_\Delta \kappa_\mu}{\kappa_{usc}\kappa_{ubt}}\right)^{\frac{1}{1-\mu}}\right] \stackrel{\rm def}{=} \delta_{\mathcal{R}}. \tag{3.18}$$

Suppose furthermore that $k > 0$ and that

$$\theta_k \leq \min[\delta_\theta, \delta_n]. \tag{3.19}$$

Then $k \notin \mathcal{R}$.

Proof. Because $\theta_k \leq \delta_n$, we know from (2.10) and Lemma 3.1 that n_k satisfies (2.10) and (3.4). Moreover, since $\theta_k \leq \delta_\theta$, we have that (3.17) also holds. Assume, for the purpose of deriving a contradiction, that $k \in \mathcal{R}$, that is,

$$\|n_k\| > \kappa_\Delta \kappa_\mu \Delta_k^{1+\mu}, \tag{3.20}$$

where we have used (2.12) and the fact that $\kappa_\mu \Delta_k^\mu \leq 1$ because of (3.18). In this case, the mechanism of the algorithm then ensures that $k-1 \notin \mathcal{R}$. Now assume that iteration $k-1$ is unsuccessful. Because of Lemmas 3.6 and 3.8, which hold at iteration $k-1 \notin \mathcal{R}$ because of (3.18), the fact that $\theta_k = \theta_{k-1}$, (2.10), and (3.17), we obtain that

$$\rho_{k-1} \geq \eta_2 \ \text{ and } \ f(x_{k-1} + s_{k-1}) \leq f(x_{k-1}) - \gamma_\theta \theta_{k-1}. \tag{3.21}$$

Hence, given that x_{k-1} is acceptable for the filter at the beginning of iteration $k-1$, if this iteration is unsuccessful, it must be because $x_{k-1} + s_{k-1}$ is not acceptable for the filter and x_{k-1}, which in turn can happen only if

$$\theta(x_{k-1} + s_{k-1}) > (1-\gamma_\theta)\theta_{k-1} = (1-\gamma_\theta)\theta_k$$

because of (3.21) (see the last paragraph of section 2.2). But Lemma 3.4 and the mechanism of the algorithm then imply that

$$(1-\gamma_\theta)\theta_k \leq \kappa_{\rm ubt}\Delta_{k-1}^2 \leq \frac{\kappa_{\rm ubt}}{\gamma_0^2}\Delta_k^2.$$

Combining this last bound with (3.20) and (2.10), we deduce that

$$\kappa_\Delta \kappa_\mu \Delta_k^{1+\mu} < \|n_k\| \leq \kappa_{\rm usc}\theta_k \leq \frac{\kappa_{\rm usc}\kappa_{\rm ubt}}{\gamma_0^2(1-\gamma_\theta)}\Delta_k^2$$

and hence that

$$\Delta_k^{1-\mu} > \frac{\gamma_0^2(1-\gamma_\theta)\kappa_\Delta\kappa_\mu}{\kappa_{\rm usc}\kappa_{\rm ubt}}.$$

Since this last inequality contradicts (3.18), our assumption that iteration $k-1$ is unsuccessful must be false. Thus iteration $k-1$ is successful and $\theta_k = \theta(x_{k-1}+s_{k-1})$. We then obtain from (3.20), (2.10), and (3.9) that

$$\kappa_\Delta \kappa_\mu \Delta_k^{1+\mu} < \|n_k\| \leq \kappa_{\rm usc}\theta_k \leq \kappa_{\rm usc}\kappa_{\rm ubt}\Delta_{k-1}^2 \leq \frac{\kappa_{\rm usc}\kappa_{\rm ubt}}{\gamma_0^2}\Delta_k^2,$$

which is again impossible because of (3.18) and because $(1-\gamma_\theta) < 1$. Hence our initial assumption (3.20) must be false, which yields the desired conclusion. ∎

We now distinguish two mutually exclusive cases. For the first, we consider what happens if there is an infinite subsequence of iterates belonging to the filter.

LEMMA 3.10. *Suppose that Algorithm* 2.1 *is applied to problem* (1.1) *and that finite termination does not occur. Suppose also that* AS1–AS3, (2.10) *hold, and* (2.15) *holds for* $k \notin \mathcal{R}$. *Suppose furthermore that* $|\mathcal{Z}| = \infty$. *Then there exists a subsequence* $\{k_j\} \subseteq \mathcal{Z}$ *such that*

$$\lim_{j\to\infty} \theta_{k_j} = 0 \tag{3.22}$$

and

$$\lim_{j\to\infty} \chi_{k_j} = 0. \tag{3.23}$$

Proof. Let $\{k_i\}$ be any infinite subsequence of $\mathcal{Z}$. We observe that (3.22) follows from Lemma 3.3. Suppose now that

$$\chi_{k_i} \geq \epsilon_2 > 0 \tag{3.24}$$

for all i and some $\epsilon_2 > 0$. Suppose furthermore that there exists $\epsilon_3 > 0$ such that, for all $i \geq i_0$,

$$\Delta_{k_i} \geq \epsilon_3. \tag{3.25}$$

Observe first that (3.22) and (2.10) ensure that

$$\lim_{i\to\infty} \|n_{k_i}\| = 0. \tag{3.26}$$

Thus (3.25) ensures that (2.12) holds for sufficiently large i and thus $k_i \notin \mathcal{R}$ for such i. Now, as we noted in the proof of Lemma 3.5,

$$|m_{k_i}(x_{k_i}) - m_{k_i}(x_{k_i}^{\rm N})| \leq \kappa_{\rm ubg}\|n_{k_i}\| + \frac{1}{2}\kappa_{\rm umh}\|n_{k_i}\|^2,$$

which in turn, with (3.26), yields that

$$\lim_{i\to\infty} [m_{k_i}(x_{k_i}) - m_{k_i}(x_{k_i}^{\rm N})] = 0. \tag{3.27}$$

We also deduce from (2.15) and AS2 that

$$m_{k_i}(x_{k_i}^{\mathrm{N}}) - m_{k_i}(x_{k_i} + s_{k_i}) \geq \kappa_{\mathrm{tmd}}\epsilon_2 \min\left[\frac{\epsilon_2}{\kappa_{\mathrm{umh}}}, \epsilon_3\right] \stackrel{\mathrm{def}}{=} \delta > 0. \tag{3.28}$$

We now decompose the model decrease in its normal and tangential components, that is,

$$m_{k_i}(x_{k_i}) - m_{k_i}(x_{k_i} + s_{k_i}) = m_{k_i}(x_{k_i}) - m_{k_i}(x_{k_i}^{\mathrm{N}}) + m_{k_i}(x_{k_i}^{\mathrm{N}}) - m_{k_i}(x_{k_i} + s_{k_i}).$$

Substituting (3.27) and (3.28) into this decomposition, we find that

$$\liminf_{i\to\infty}[m_{k_i}(x_{k_i}) - m_{k_i}(x_{k_i} + s_{k_i})] \geq \delta > 0. \tag{3.29}$$

We now observe that, because x_{k_i} is added to the filter at iteration k_i, we know from the mechanism of the algorithm that either iteration $k_i \in \mathcal{R}$ or (2.19) must fail. Since we have already shown that $k_i \notin \mathcal{R}$, (2.19) must fail for i sufficiently large, that is,

$$m_{k_i}(x_{k_i}) - m_{k_i}(x_{k_i} + s_{k_i}) < \kappa_\theta \theta_{k_i}^{\psi}. \tag{3.30}$$

Combining this bound with (3.29), we find that θ_{k_i} is bounded away from zero for i sufficiently large, which is impossible in view of (3.22). We therefore deduce that (3.25) cannot hold and obtain that there is a subsequence $\{k_\ell\} \subseteq \{k_i\}$ for which

$$\lim_{\ell\to\infty} \Delta_{k_\ell} = 0.$$

We now restrict our attention to the tail of this subsequence, that is, to the set of indices $k_\ell > 0$ that are large enough to ensure that (3.16), (3.17), and (3.18) hold, which is possible by definition of the subsequence and because of (3.22). For these indices, we may therefore apply Lemma 3.9 and deduce that iteration $k_\ell \notin \mathcal{R}$ for ℓ sufficiently large. Hence, as above, (3.30) must hold for ℓ sufficiently large. However, we may also apply Lemma 3.7, which contradicts (3.30), and therefore (3.24) cannot hold, yielding the desired result. ∎

Thus, if an infinite subsequence of iterates is added to the filter, Lemma 3.2 ensures that there exists a limit point which is a first-order critical point. Our remaining analysis then naturally concentrates on the possibility that there may be no such infinite subsequence. In this case, no further iterates are added to the filter for k sufficiently large. In particular, this means that the number of restoration iterations, $|\mathcal{R}|$, must be finite. In what follows, we assume that $k_0 \geq 0$ is the last iteration for which x_{k_0-1} is added to the filter.

LEMMA 3.11. *Suppose that Algorithm* 2.1 *is applied to problem* (1.1)*, that finite termination does not occur, and that* $|\mathcal{Z}| < \infty$*. Suppose also that* AS1–AS3 *and* (2.10) *hold and that* (2.15) *holds for* $k \notin \mathcal{R}$*. Then we have that*

$$\lim_{k\to\infty} \theta_k = 0. \tag{3.31}$$

Furthermore, n_k *satisfies* (3.4) *for all* $k \geq k_0$ *sufficiently large.*

Proof. Consider any successful iterate with $k \geq k_0$. Since x_k is not added to the filter, it follows from the mechanism of the algorithm that $\rho_k \geq \eta_1$ holds and thus that

$$f(x_k) - f(x_{k+1}) \geq \eta_1[m_k(x_k) - m_k(x_k + s_k)] \geq \eta_1\kappa_\theta\theta_k^{\psi} \geq 0. \tag{3.32}$$

Thus the objective function does not increase for all successful iterations with $k \geq k_0$. But AS1 and AS3 imply (3.3), and therefore we must have, from the first part of this statement, that

$$\lim_{\substack{k \in \mathcal{S} \\ k \to \infty}} f(x_k) - f(x_{k+1}) = 0. \tag{3.33}$$

The limit (3.31) then immediately follows from (3.32) and the fact that $\theta_j = \theta_k$ for all unsuccessful iterations j that immediately follow the successful iteration k, if any. The last conclusion then results from (2.10) and Lemma 3.1. ∎

We now show that the trust-region radius cannot become arbitrarily small if the (asymptotically feasible) iterates stay away from first-order critical points.

LEMMA 3.12. *Suppose that Algorithm* 2.1 *is applied to problem* (1.1)*, that finite termination does not occur, and that* $|\mathcal{Z}| < \infty$*. Suppose also that* AS1–AS3 *hold and* (2.15) *holds for* $k \notin \mathcal{R}$*. Suppose furthermore that* (3.11) *hold for all* $k \geq k_0$*. Then there exists a* $\Delta_{\min} > 0$ *such that*

$$\Delta_k \geq \Delta_{\min}$$

for all k*.*

Proof. Suppose that $k_1 \geq k_0$ is chosen sufficiently large to ensure that (3.19) holds and that n_k satisfies (2.10) for all $k \geq k_1$, which is possible because of Lemma 3.11. Suppose also, for the purpose of obtaining a contradiction, that iteration j is the first iteration following iteration k_1 for which

$$\Delta_j \leq \gamma_0 \min \left[\delta_\rho, \sqrt{\frac{(1-\gamma_\theta)\theta^{\mathrm{F}}}{\kappa_{\mathrm{ubt}}}}, \Delta_{k_1} \right] \stackrel{\mathrm{def}}{=} \gamma_0 \delta_s, \tag{3.34}$$

where

$$\theta^{\mathrm{F}} \stackrel{\mathrm{def}}{=} \min_{i \in \mathcal{Z}} \theta_i$$

is the smallest constraint violation appearing in the filter. Note also that the inequality $\Delta_j \leq \gamma_0 \Delta_{k_1}$, which is implied by (3.34), ensures that $j \geq k_1 + 1$ and hence that $j - 1 \geq k_1$ and thus that $j - 1 \notin \mathcal{R}$. Then the mechanism of the algorithm and (3.34) imply that

$$\Delta_{j-1} \leq \frac{1}{\gamma_0} \Delta_j \leq \delta_s, \tag{3.35}$$

and Lemma 3.6, which is applicable because (3.34) and (3.35) together imply (3.14) with k replaced by $j - 1$, then ensures that

$$\rho_{j-1} \geq \eta_2. \tag{3.36}$$

Furthermore, since n_{j-1} satisfies (2.10), Lemma 3.1 implies that we can apply Lemma 3.4. This, together with (3.34) and (3.35), gives that

$$\theta(x_{j-1} + s_{j-1}) \leq \kappa_{\mathrm{ubt}} \Delta_{j-1}^2 \leq (1 - \gamma_\theta)\theta^{\mathrm{F}}. \tag{3.37}$$

We may also apply Lemma 3.8 because (3.34) and (3.35) ensure that (3.14) holds and because (3.17) also holds for $j - 1 \geq k_1$. Hence we deduce that

$$f(x_{j-1} + s_{j-1}) \leq f(x_{j-1}) - \gamma_\theta \theta_{j-1}.$$

This last relation and (3.37) ensure that $x_{j-1} + s_{j-1}$ is acceptable for the filter and x_{j-1}. Combining this conclusion with (3.36) and the mechanism of the algorithm, we obtain that $\Delta_j \geq \Delta_{j-1}$. As a consequence, and since (2.19) also holds at iteration $j-1$, iteration j cannot be the first iteration following k_1 for which (3.34) holds. This contradiction shows that $\Delta_k \geq \gamma_0 \delta_s$ for all $k > k_1$, and the desired result follows if we define

$$\Delta_{\min} = \min[\Delta_0, \ldots, \Delta_{k_1}, \gamma_0 \delta_s]. \qquad \square$$

We may now analyze the convergence of χ_k itself.

LEMMA 3.13. *Suppose that Algorithm* 2.1 *is applied to problem* (1.1)*, that finite termination does not occur, and that* $|\mathcal{Z}| < \infty$*. Suppose also that* AS1–AS3*,* (2.10) *hold, and* (2.15) *holds for* $k \notin \mathcal{R}$*. Then*

$$\liminf_{k \to \infty} \chi_k = 0. \tag{3.38}$$

Proof. We start by observing that Lemma 3.11 implies that the second conclusion of (2.10) holds for k sufficiently large. Moreover, as in Lemma 3.11, we obtain (3.32) and therefore (3.33) for each $k \in \mathcal{S}$, $k \geq k_0$. Suppose now, for the purpose of obtaining a contradiction, that (3.11) holds, and notice that

$$m_k(x_k) - m_k(x_k + s_k) = m_k(x_k) - m_k(x_k^{\mathrm{N}}) + m_k(x_k^{\mathrm{N}}) - m_k(x_k + s_k). \tag{3.39}$$

Moreover, note, as in Lemma 3.5, that

$$|m_k(x_k) - m_k(x_k^{\mathrm{N}})| \leq \kappa_{\mathrm{ubg}} \|n_k\| + \kappa_{\mathrm{umh}} \|n_k\|^2,$$

which in turn yields that

$$\lim_{k \to \infty} [m_k(x_k) - m_k(x_k^{\mathrm{N}})] = 0$$

because of Lemma 3.11 and the second conclusion of (2.10). This limit, together with (3.32), (3.33), and (3.39), then gives that

$$\lim_{\substack{k \to \infty \\ k \in \mathcal{S}}} [m_k(x_k^{\mathrm{N}}) - m_k(x_k + s_k)] = 0. \tag{3.40}$$

But (2.15), (3.11), AS2, and Lemma 3.12 together imply that for all $k \geq k_0$

$$\begin{aligned} m_k(x_k^{\mathrm{N}}) - m_k(x_k + s_k) &\geq \kappa_{\mathrm{tmd}} \chi_k \min\left[\frac{\chi_k}{\beta_k}, \Delta_k\right] \\ &\geq \kappa_{\mathrm{tmd}} \epsilon \min\left[\frac{\epsilon}{\kappa_{\mathrm{umh}}}, \Delta_{\min}\right], \end{aligned} \tag{3.41}$$

immediately giving a contradiction with (3.40). Hence (3.11) cannot hold and the desired result follows. $\square$

We may summarize all of the above in our main global convergence result.

THEOREM 3.14. *Suppose that Algorithm* 2.1 *is applied to problem* (1.1) *and that finite termination does not occur. Suppose also that* AS1*,* (2.10)*,* AS3*, and* AS2 *hold, and that* (2.15) *holds for* $k \notin \mathcal{R}$*. Let* $\{x_k\}$ *be the sequence of iterates produced by the algorithm. Then either the restoration procedure terminates unsuccessfully by converging to an infeasible first-order critical point of problem* (2.16)*, or there is a subsequence* $\{k_j\}$ *for which*

$$\lim_{j \to \infty} x_{k_j} = x_*$$

and x_* is a first-order critical point for problem (1.1).

Proof. Suppose that the restoration iteration always terminates successfully. From AS3, Lemmas 3.10, 3.11, and 3.13, we obtain that, for some subsequence $\{k_j\}$,

$$\lim_{j\to\infty} \theta_{k_j} = \lim_{j\to\infty} \chi_{k_j} = 0. \tag{3.42}$$

The conclusion then follows from Lemma 3.2. ∎

Can we dispense with AS3 to obtain this result? First, this assumption ensures that the objective remains bounded below and the constraint violation remains bounded above (see (3.3)). This is crucial for the rest of the analysis because the convergence of the iterates to feasibility depends on this fact. Thus, if AS3 does not hold, we have to verify that (3.3) holds for other reasons. The second part of this statement may be ensured quite simply by initializing the filter to $(\theta^{\max}, -\infty)$, for some $\theta^{\max} > \theta_0$, in Step 0 of the algorithm. This has the effect of putting an upper bound on the infeasibility of all iterates, which may be useful in practice. However, this does not prevent the objective function from being unbounded below in

$$\mathcal{C}(\theta^{\max}) = \{x \in \mathbb{R}^n \mid \theta(x) \leq \theta^{\max}\},$$

and we cannot exclude the possibility that a sequence of infeasible iterates might both continue to improve the value of the objective function and satisfy (2.19). If $\mathcal{C}(\theta^{\max})$ is bounded, AS3 is most certainly satisfied. If this is not the case, we could assume that

$$f^{\min} \leq f(x) \text{ and } 0 \leq \theta(x) \leq \theta^{\max} \text{ for } x \in \mathcal{C}(\theta^{\max}) \tag{3.43}$$

for some value of $f^{\min}$ and simply monitor that the values $f(x_k)$ are reasonable—in view of the problem being solved—as the algorithm proceeds. To summarize, we may replace AS1 and AS3 by the following assumption.

AS4. The functions f and c are twice continuously differentiable on an open set containing $\mathcal{C}(\theta^{\max})$, their first and second derivatives are uniformly bounded on $\mathcal{C}(\theta^{\max})$, and (3.43) holds.

The reader should note that AS4 no longer ensures the existence of a limit point, but only that (3.42) holds for some subsequence $\{k_j\}$. Furthermore, the comments following the statement of (2.10) no longer apply if limit points at infinity are allowed.

4. Conclusion and perspectives. We have introduced a trust-region SQP-filter algorithm for general nonlinear programming and have shown this algorithm to be globally convergent to first-order critical points. The proposed algorithm differs from that discussed by Fletcher and Leyffer [18], notably because it uses a decomposition of the step in its normal and tangential components and imposes some restrictions on the length of the former. However, preliminary numerical experiments indicate that its practical performance is similar to that reported in [18]. Since the performance of the latter is excellent, the theory developed in this paper provides the reassurance that filter algorithms also have reasonable convergence properties, which then makes these methods very attractive.

We are aware, however, that the convergence study is not complete, as we have not discussed local convergence properties. As it is possible to exhibit examples where the SQP step increases *both* the objective function *and* the constraint violation,[2] it is

[2]Such an example is provided by Figure 9.3.1 in Fletcher [16], taking the case $\beta = \frac{1}{4}$. Any feasible point close to the origin illustrates the effect.

very likely that such a study will have to introduce second-order corrections (see [16, section 14.4]) to ensure that the Maratos effect does not take place and that a fast (quadratic) rate of convergence can be achieved. Moreover, convergence to second-order critical points also remains, for now, an open question. In this context, the alternative definition of ρ_k presented in (3.15) is also likely to play a role if we choose H_k according to (3.1). In this case, we might choose

$$\Theta_k = \sum_{i \in \mathcal{E} \cup \mathcal{I}} [y_k]_i \langle s_k, \nabla_{xx} c_i(x_k) s_k \rangle$$

in order to ensure that the denominator of the fraction defining ρ_k is a correct model of its numerator not only up to first-order, but also up to second-order. These questions are the subject of ongoing work.

Appendix.

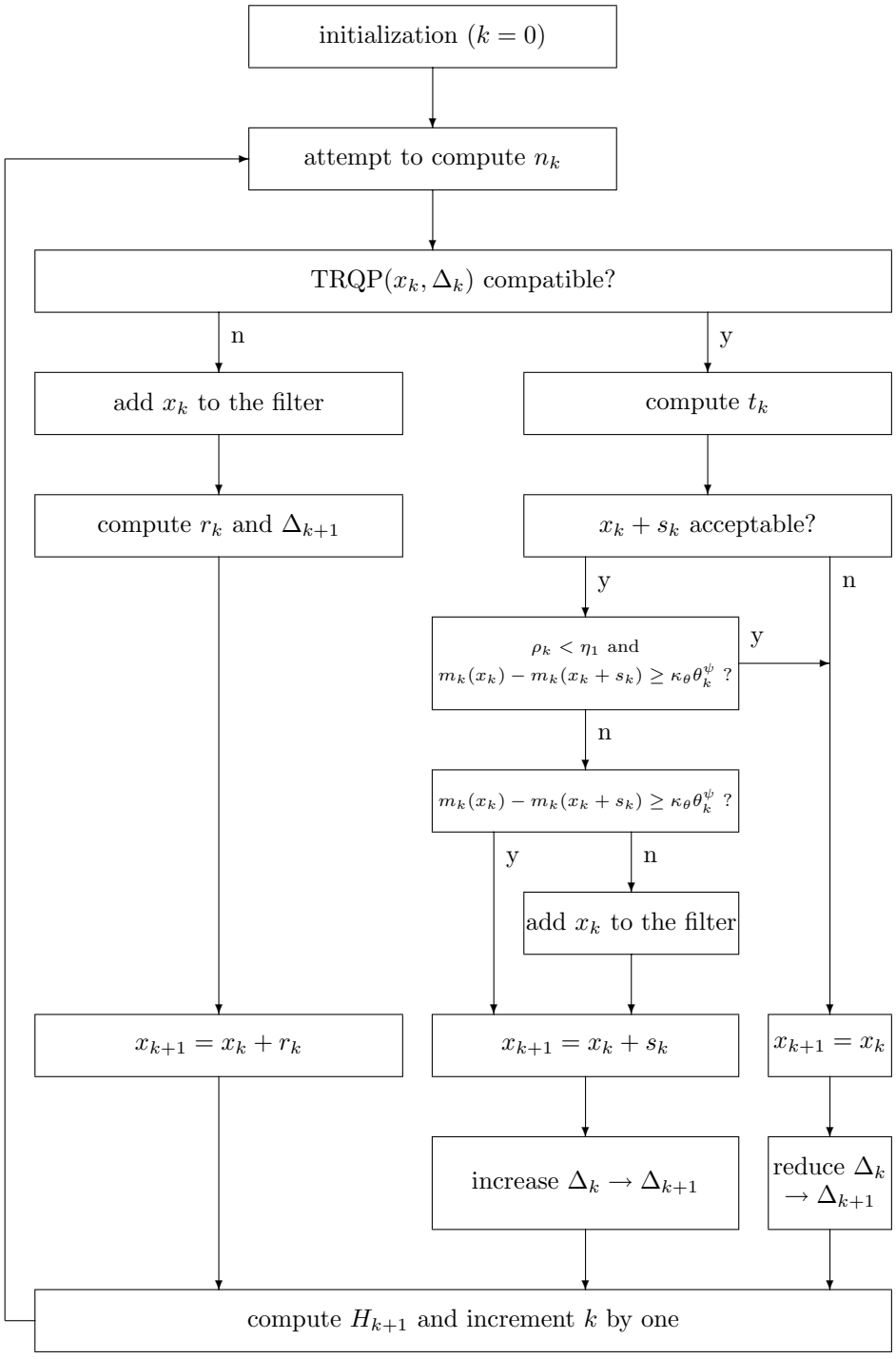


FIG. A.1. *Flowchart of Algorithm* 2.1.

REFERENCES

[1] L. T. Biegler, J. Nocedal, and C. Schmid, *A reduced Hessian method for large-scale constrained optimization*, SIAM J. Optim., 5 (1995), pp. 314–347.

[2] R. H. Bielschowsky and F. A. M. Gomes, *Dynamical Control of Infeasibility in Nonlinearly Constrained Optimization*, presentation at the Optimization '98 Conference, University of Coimbra, Portugal, 1998.

[3] R. H. Byrd, J. C. Gilbert, and J. Nocedal, *A trust region method based on interior point techniques for nonlinear programming*, Math. Program., 89 (2000), pp. 149–186.

[4] R. H. Byrd, M. E. Hribar, and J. Nocedal, *An interior point algorithm for large-scale nonlinear programming*, SIAM J. Optim., 9 (2000), pp. 877–900.

[5] R. H. Byrd, R. B. Schnabel, and G. A. Shultz, *A trust region algorithm for nonlinearly constrained optimization*, SIAM J. Numer. Anal., 24 (1987), pp. 1152–1170.

[6] A. R. Conn, N. Gould, A. Sartenaer, and Ph. L. Toint, *Global convergence of a class of trust region algorithms for optimization using inexact projections on convex constraints*, SIAM J. Optim., 3 (1993), pp. 164–221.

[7] A. R. Conn, N. I. M. Gould, and P. L. Toint, *Trust-Region Methods*, MPS-SIAM Ser. Optim. 1, SIAM, Philadelphia, 2000.

[8] R. S. Dembo and U. Tulowitzki, *On the Minimization of Quadratic Functions Subject to Box Constraints*, Working paper, Series B, 71, School of Organization and Management, Yale University, New Haven, CT, 1983.

[9] J. E. Dennis, Jr., M. El-Alem, and M. C. Maciel, *A global convergence theory for general trust-region based algorithms for equality constrained optimization*, SIAM J. Optim., 7 (1997), pp. 177–207.

[10] J. E. Dennis, Jr., M. El-Alem, and K. Williamson, *A trust-region approach to nonlinear systems of equalities and inequalities*, SIAM J. Optim., 9 (1999), pp. 291–315.

[11] J. E. Dennis and L. N. Vicente, *On the convergence theory of trust-region-based algorithms for equality-constrained optimization*, SIAM J. Optim., 7 (1997), pp. 927–950.

[12] M. El-Alem, *Global convergence without the assumption of linear independence for a trust-region algorithm for constrained optimization*, J. Optim. Theory Appl., 87 (1995), pp. 563–577.

[13] M. El-Alem, *A global convergence theory for Dennis, El-Alem, and Maciel's class of trust-region algorithms for constrained optimization without assuming regularity*, SIAM J. Optim., 9 (1999), pp. 965–990.

[14] M. El-Hallabi and R. A. Tapia, *An Inexact Trust-Region Feasible-Point Algorithm for Nonlinear Systems of Equalities and Inequalities*, Technical Report TR95-09, Department of Computational and Applied Mathematics, Rice University, Houston, TX, 1995.

[15] A. V. Fiacco, *Introduction to Sensitivity and Stability Analysis in Nonlinear Programming*, Academic Press, London, 1983.

[16] R. Fletcher, *Practical Methods of Optimization*, 2nd ed., J. Wiley and Sons, Chichester, England, 1987.

[17] R. Fletcher and S. Leyffer, *User Manual for FilterSQP*, Numerical Analysis Report NA/181, Department of Mathematics, University of Dundee, Dundee, Scotland, 1998.

[18] R. Fletcher and S. Leyffer, *Nonlinear programming without a penalty function*, Math. Program., 91 (2002), pp. 239–269.

[19] R. Fletcher, S. Leyffer, and P. L. Toint, *On the Global Convergence of an SLP-Filter Algorithm*, Technical Report 98/13, Department of Mathematics, University of Namur, Namur, Belgium, 1998.

[20] R. Fletcher, S. Leyffer, and P. L. Toint, *On the Global Convergence of a Filter-SQP Algorithm*, Technical Report 00/15, Department of Mathematics, University of Namur, Namur, Belgium, 2000.

[21] 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.

[22] M. Lalee, J. Nocedal, and T. D. Plantenga, *On the implementation of an algorithm for large-scale equality constrained optimization*, SIAM J. Optim., 8 (1998), pp. 682–706.

[23] X. Liu and Y. Yuan, *A Robust Trust-Region Algorithm for Solving General Nonlinear Programming Problems*, presentation at the International Conference on Nonlinear Programming and Variational Inequalities, Hong Kong, 1998.

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

[25] W. Murray and F. J. Prieto, *A sequential quadratic programming algorithm using an in-*

complete solution of the subproblem, SIAM J. Optim., 5 (1995), pp. 590–640.

[26] K. G. Murty and S. N. Kabadi, *Some NP-complete problems in quadratic and nonlinear programming*, Math. Programming, 39 (1987), pp. 117–129.

[27] E. O. Omojokun, *Trust Region Algorithms for Optimization with Nonlinear Equality and Inequality Constraints*, Ph.D. thesis, University of Colorado, Boulder, CO, 1989.

[28] A. Sartenaer, *Automatic determination of an initial trust region in nonlinear programming*, SIAM J. Sci. Comput., 18 (1997), pp. 1788–1803.

[29] P. L. Toint, *Global convergence of a class of trust region methods for nonconvex minimization in Hilbert space*, IMA J. Numer. Anal., 8 (1988), pp. 231–252.

[30] A. Vardi, *A trust region algorithm for equality constrained minimization: Convergence properties and implementation*, SIAM J. Numer. Anal., 22 (1985), pp. 575–591.

[31] A. Wächter and L. T. Biegler, *Global and Local Convergence of Line Search Filter Methods for Nonlinear Programming*, Technical Report CAPD B-01-09, Department of Chemical Engineering, Carnegie Mellon University, Pittsburgh, PA, 2001.

[32] Y. Yuan, *Trust region algorithms for nonlinear programming*, in Computational Mathematics in China, Contemp. Math. 163, Z. C. Shi, ed., AMS, Providence, RI, 1994, pp. 205–225.

SIAM J. OPTIM.
Vol. 13, No. 3, pp. 660–674

CHEAPER JACOBIANS BY SIMULATED ANNEALING*

UWE NAUMANN†

Abstract. Jacobian matrices can be accumulated by applying the chain rule to vector functions given as computer programs in different orders resulting in varying operations counts while yielding identical results, up to round-off. The minimization of the number of operations performed leads to a computationally hard combinatorial optimization problem based on vertex elimination in computational graphs. This paper discusses simulated annealing as a method for generating nearly optimal Jacobian code.

Key words. automatic differentiation, vertex elimination, simulated annealing

AMS subject classifications. 90C27, 26B10

PII. S1052623400368394

1. Accumulation of Jacobian matrices.

Let

$$F' = F'(\mathbf{x}_0) = \left(\frac{\partial y_i}{\partial x_j}(\mathbf{x}_0)\right)_{j=1,\dots,n}^{i=1,\dots,m}$$

denote the Jacobian matrix (also: Jacobian) of a nonlinear vector function

$$F : \mathbb{R}^n \supseteq D \to \mathbb{R}^m : \quad \mathbf{x} \mapsto \mathbf{y} = F(\mathbf{x})$$

mapping the n components of $\mathbf{x}$ onto the m components of $\mathbf{y}$ and being evaluated at a given argument $\mathbf{x}_0$. The run-time of numerous numerical algorithms is dominated by the time it takes to accumulate F' or to evaluate products of the form $(\mathbb{R}^{m\times l_1}\ni)\ \dot{Y} = F'\dot{X}$ and $(\mathbb{R}^{l_2\times n}\ni)\ \bar{X} = \bar{Y}F'$. This paper presents a method for accumulating F' efficiently. Likely, these ideas will also be useful for computing higher order derivative tensors. Automatic differentiation (AD) [6] will be considered from the point of view of graph theory and combinatorial optimization.

F is assumed to be given as a computer program which decomposes into a sequence of scalar elemental functions $(\mathbb{R}\ni)v_j = \varphi_j(v_i)_{i\prec j}$, where $j = 1,\dots,q$ and $p = q - m$. The direct dependence of v_j on v_i is denoted by $i \prec j$. We write $i \prec^* j$ if there exist $k_1,\dots,k_p$ such that $i \prec k_1 \prec k_2 \prec \cdots \prec k_p \prec j$. So, $\{i|i \prec j\}$ is the index set of the arguments of φ_j and we denote its cardinality by $|\{i|i \prec j\}|$. Within F we distinguish between three types of variables $V = X \cup Z \cup Y$: independent ($X \equiv \{v_{1-n},\dots,v_0\}$), intermediate ($Z \equiv \{v_1,\dots,v_p\}$), and dependent ($Y \equiv \{v_{p+1},\dots,v_q\}$). We set $x_i \equiv v_{i-n}$, $i = 1,\dots,n$, and $y_j \equiv v_{p+j}$, $j = 1,\dots,m$. The numbering $\mathcal{I} : V \to \{(1-n),\dots,q\}$ of the variables of F must induce a topological order with respect to the dependence "$\prec$", i.e., $i \prec^* j \Rightarrow \mathcal{I}(v_i) < \mathcal{I}(v_j)$.

Since the differentiation of F is based on the differentiability of its elemental functions it will be assumed that the φ_j, $j = 1,\dots,q$, have jointly continuous partial derivatives $c_{ji} \equiv \frac{\partial}{\partial v_i}\varphi_j(v_k)_{k\prec j}$, $i \prec j$, on open neighborhoods $\mathcal{D}_j \subset \mathbb{R}^{n_j}$, $n_j \equiv |\{i|i \prec j\}|$, of their domain. The computational graph (or c-graph) $\mathbf{G} = (V, E)$ of F is a directed acyclic graph with $V = \{i|v_i \in F\}$ and $(i,j) \in E$ if $i \prec j$. We

*Received by the editors February 18, 2000; accepted for publication (in revised form) February 7, 2002; published electronically November 6, 2002.
http://www.siam.org/journals/siopt/13-3/36839.html
†Department of Computer Science, University of Hertfordshire, College Lane, Hatfield, AL10 9AB, UK (U.1.Naumann@herts.ac.uk).

assume $\mathbf{G}$ to be linearized in the sense that all partial derivatives of the elemental functions are attached to their corresponding edges, i.e., (i,j) is labeled with c_{ji}.

The *forward mode* of AD [6] computes F' as $\dot{Y} = F'\dot{X}$ by forward propagation of $\dot{X} = I_n$, where $I_n \in \mathbb{R}^{n\times n}$ denotes the identity matrix. A lower bound on the number of multiplications involved in this process is given by $n \cdot q \leq n \cdot |E|$. Analogously, AD's *reverse mode* [6] sets $\bar{Y} = I_m \in \mathbb{R}^{m\times m}$ and accumulates F' as $\bar{X} = \bar{Y}F'$ by reverse propagation of the identity matrix at a cost of at least $m \cdot (p+n) \leq m \cdot |E|$ multiplications.

In the case where F' is sparse, the method of Newsam and Ramsdell [19] may lead to a reduction of these numbers. The application of this method in forward mode yields a minimal cost of $\hat{n}(m+p) \leq \hat{n} \cdot |E|$, where $\hat{n}$ denotes the maximal number of nonzero elements per row in F'. Analogously, Newsam and Ramsdell's version of the reverse mode requires at least $\hat{m}(n+p) \leq \hat{m} \cdot |E|$ multiplications, where $\hat{m}$ is the maximal number of nonzero elements per column in F'.

In contrast to the methods sketched above, the accumulation of F' can be regarded as the process of transforming $\mathbf{G}$ into a subgraph $\mathbf{G}'$ of the complete bi-partite graph $K_{n,m}$ [8]. The n minimal nodes correspond to the independent variables of F, whereas the dependent variables are represented by the m maximal nodes. The number of arithmetic operations actually performed may vary drastically for different application sequences of the chain rule to $\mathbf{G}$.

Prior to the introduction of a particular elimination technique upon which the transformation $\mathbf{G} \to \mathbf{G}' \subseteq K_{n,m}$ can be built, let us look briefly at an example to illustrate the above. Consider a function $F : \mathbb{R}^2 \to \mathbb{R}^2$ given by the following assignments:

$$\begin{aligned} h_1 &= x_1 \cdot x_2, \\ h_2 &= \exp(\sin(h_1)), \\ y_1 &= h_1 \cdot h_2, \\ y_2 &= \cos(h_2). \end{aligned} \tag{1.1}$$

Its c-graph is shown in Figure 1.1, where edges are labeled with local partial derivatives

$$\mathtt{a} = \frac{\partial(v_{-1} \cdot v_0)}{\partial v_{-1}} = v_0 \equiv x_2, \ldots, \qquad \mathtt{c} = \frac{\partial(\sin(v_1))}{\partial v_1} = \cos(v_1), \ldots .$$

From the chain rule it follows that an entry $F'(i,j)$ of the Jacobian can be computed by multiplying the edge labels over all paths connecting the minimal vertex j with the maximal vertex i followed by summing these products [8]. Consequently, the Jacobian of (1.1) is given by

$$F' = \begin{pmatrix} \mathtt{ae} + \mathtt{acdf} & \mathtt{be} + \mathtt{bcdf} \\ \mathtt{acdg} & \mathtt{bcdg} \end{pmatrix}. \tag{1.2}$$

The naive approach to computing (1.2) would take 14 scalar multiplications and 2 additions. The preaccumulation of $r = \mathtt{cd}$, $s = r\mathtt{f}$, and $t = r\mathtt{g}$ would give us

$$F' = \begin{pmatrix} \mathtt{a}(\mathtt{e}+s) & \mathtt{b}(\mathtt{e}+s) \\ \mathtt{a}t & \mathtt{b}t \end{pmatrix} \tag{1.3}$$

at a cost of only 7 multiplications and 2 additions. The algorithms presented in this paper can partially be interpreted as approaches for identifying such reusable expressions.

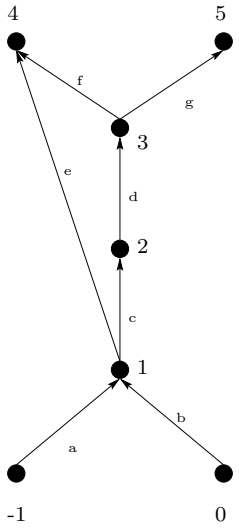


FIG. 1.1.

2. Vertex elimination problem. Elimination of intermediate vertices will be used to transform $\mathbf{G}$ into $\mathbf{G}'$. The following terminology will be assumed: By the chain rule, *updating the existing or generating new edge labels* means that the labels of successive edges (i, j) and (j, k) are multiplied to form the new label of (i, k), whereas labels of parallel edges having both the same source and target are added. Parallel edges will always be merged by performing this addition.

Graphically, the *elimination of an intermediate vertex* j from $\mathbf{G}$ is equivalent to connecting all $i|i \prec j$ with all $k|j \prec k$ followed by updating the existing or generating new edge labels and, finally, the deletion of j. A vertex is deleted together with all edges incident to it. This is illustrated by Figure 2.1.

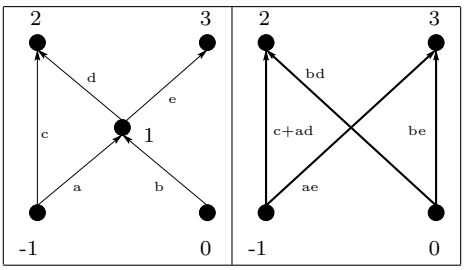


FIG. 2.1. *Vertex elimination in* $\mathbf{G}$.

Relying on the abilities of a wide range of modern floating-point units [9], we assume equal execution times of a *multiply-add-fused* $ab + c$, $a, b, c \in \mathbb{R}$, and a scalar multiplication. Furthermore, we expect memory accesses to take constant time, independent of the actual elimination order (see section 6 for a discussion of this assumption). The number of multiplications involved in the elimination of a vertex j is called the *Markowitz degree* of j and is equal to $\mu_j = |\{k|k \prec j\}| \cdot |\{l|j \prec l\}|$. Again, the local partial derivatives are represented by $\mathtt{a}, \mathtt{b}, \ldots, \mathtt{e}$.

Our objective is to solve the *vertex elimination problem* in c-graphs, i.e., to find

a vertex elimination sequence which minimizes the number of (scalar floating-point) multiplications required for the accumulation of F'. Notice that there may be several different elimination sequences yielding the same computational cost. Here, we are interested in finding only one of them. Motivated by a similar heuristic for the solution of sparse systems of linear equations, Griewank and Reese [8] have proposed a greedy heuristic for the solution of the vertex elimination problem which will be referred to as *lowest Markowitz degree first* (LM). As the name suggests, the idea is to eliminate the *cheapest* vertex first, i.e., the vertex with the lowest Markowitz degree. Obviously, this heuristic will have to be combined with possibly several tie-breakers to ensure a unique choice at every stage. LM is a robust and easy-to-implement method for solving the vertex elimination problem. However, it has its limitations as outlined in [14]. Some of them can be potentially overcome by looking at other methods as done in [15]. *Simulated annealing* represents one of them.

The vertex elimination problem in c-graphs can be represented as a shortest path problem in a (single-source-single-sink, directed acyclic) metagraph with integer vertices (which we will also refer to as the *stages* of the elimination process) as introduced by Bischof and Haghighat in [5]. Its intermediate vertices correspond to all the different c-graphs resulting from vertex elimination applied to the original graph (the source of the metagraph) in order to build the bipartite graph that represents the Jacobian (the sink of the metagraph). Unfortunately, the number of intermediate vertices in the metagraph grows exponentially with p, which limits the applicability of this approach to problems of smaller size.

3. Simulated annealing applied to the vertex elimination problem. The algorithm developed here is motivated by the approach to solving the *traveling salesman problem* presented in [20]. It lead to a software implementation in C++ based on the library of efficient data types and algorithms (LEDA) [11] which will be referred to as SAVE.

As a problem in simulated annealing, the vertex elimination problem is handled as follows.

3.1. Configuration. The intermediate vertices are numbered $i = 1, \dots, p$, and each of them has a Markowitz degree $\mu_i(k)$ depending on the stage k in the metagraph as introduced in section 2. A configuration is a permutation of the indices $1, \dots, p$ interpreted as the order in which the intermediate vertices are eliminated.

3.2. Rearrangements. We use slightly altered versions of the rearrangements suggested in [10] which are based on operations on a certain type of subsequence of a given vertex elimination sequence.

DEFINITION 1. *Let* $(i_1, \dots, i_p)$ *be a vertex elimination sequence.* $(i_j, \dots, i_k)$, $1 \le j \le k \le p$, *is called a* dense subsequence *if for all* j' *with* $j \le j' \le k$ *it holds that* $i_{j'} \in (i_j, \dots, i_k)$.

In order for rearrangements to be suitable for logarithmic simulated annealing as described in [2] they have to exhibit certain properties. One of them is reversibility. It will be guaranteed for the following two types of moves which will be referred to as *reversal* and *transportation.*

1. If $(i_1, \dots, i_p)$ is the current elimination sequence, then we remove a dense subsequence $(i_j, \dots, i_{j+k})$ and replace it with its reverse, making

$$(\dots, i_{j-1}, i_{j+k}, \dots, i_j, i_{j+k+1}, \dots)$$

the next elimination sequence to be regarded, or

2. a dense subsequence $(i_j, \ldots, i_{j+k})$ is removed and then replaced between two indices i_l, i_r on another, randomly chosen, part of the elimination sequence; i.e., $(i_1, \ldots, i_p)$ becomes

$$(\ldots, i_{j-1}, i_{j+k+1}, \ldots, i_l, i_j, \ldots, i_{j+k}, i_r, \ldots)$$

if $l > j + k$. Otherwise, we get

$$(\ldots, i_l, i_j, \ldots, i_{j+k}, i_r, \ldots, i_{j-1}, i_{j+k+1}, \ldots),$$

where $r < j$.

We do not permit other rearrangements apart from these two. Obviously, both rearrangements represent valid neighborhood relationships as they will always transfer a given vertex elimination sequence into a new feasible vertex elimination sequence. A discussion of further rearrangements in the light of logarithmic cooling schedules for inhomogeneous Markov chains can be found in [18]. In this paper we will concentrate on algorithmic issues that arise when applying a well-known simulated annealing algorithm for solving the traveling salesman problem to the vertex elimination problem in linearized computational graphs.

3.3. Objective function. The Markowitz degree of a vertex changes dynamically throughout the elimination process. Our objective function is the sum of the Markowitz degrees of all intermediate vertices at their respective moments of elimination. We will refer to this value as the *overall Markowitz degree*.

Figure 3.1 shows the principle of our simulated annealing algorithm. It starts with an initial elimination sequence (`ES`) which we have chosen as the maximum out of forward (VF) and backward (VB) vertex elimination sequences in terms of the number of multiplications required for the accumulation of the complete Jacobian. At the beginning of the optimization procedure, we do not want the cost `c` (of `ES`) to be close to the minimum. In this case it would become very likely that the algorithm stops after just one iteration without delivering any improvement.

3.4. Annealing schedule. There are two main loops—an outer loop (counter: `olc`) and an inner loop (counter: `ilc`). For both of them we define upper bounds for the number of iterations they will perform (`ol` and `il`). `il*ol` is the maximal number of iterations (the maximal number of elimination sequences that are generated and run) before the algorithm stops, even if it has not converged. It may happen that one and the same elimination sequence is checked twice or even more often, although it is very unlikely. We make sure that we get a result within a reasonable time span by setting `ol` and `il`. Its quality strongly depends on the parameters of the simulated annealing algorithm.

In contrast with the original algorithm, we have chosen the maximal number of iterations performed in the inner loop `il` as the number returned by rounding

$$\mathtt{p} * (2 * \mathtt{atan}(1) - \mathtt{atan}(\mathtt{p}/30 - 1))$$

to the nearest integer. Similarly, the maximal number of outer loop iterations `ol` has been set to the integer nearest to

$$\mathtt{p} * (2 * \mathtt{atan}(1) - \mathtt{atan}(\mathtt{p}/100)) + 50.$$

Figure 3.2 shows the development of `il` and `ol` depending on the number `p` of intermediate variables in $\mathbf{G}$. Notice that for our algorithm we always assume that

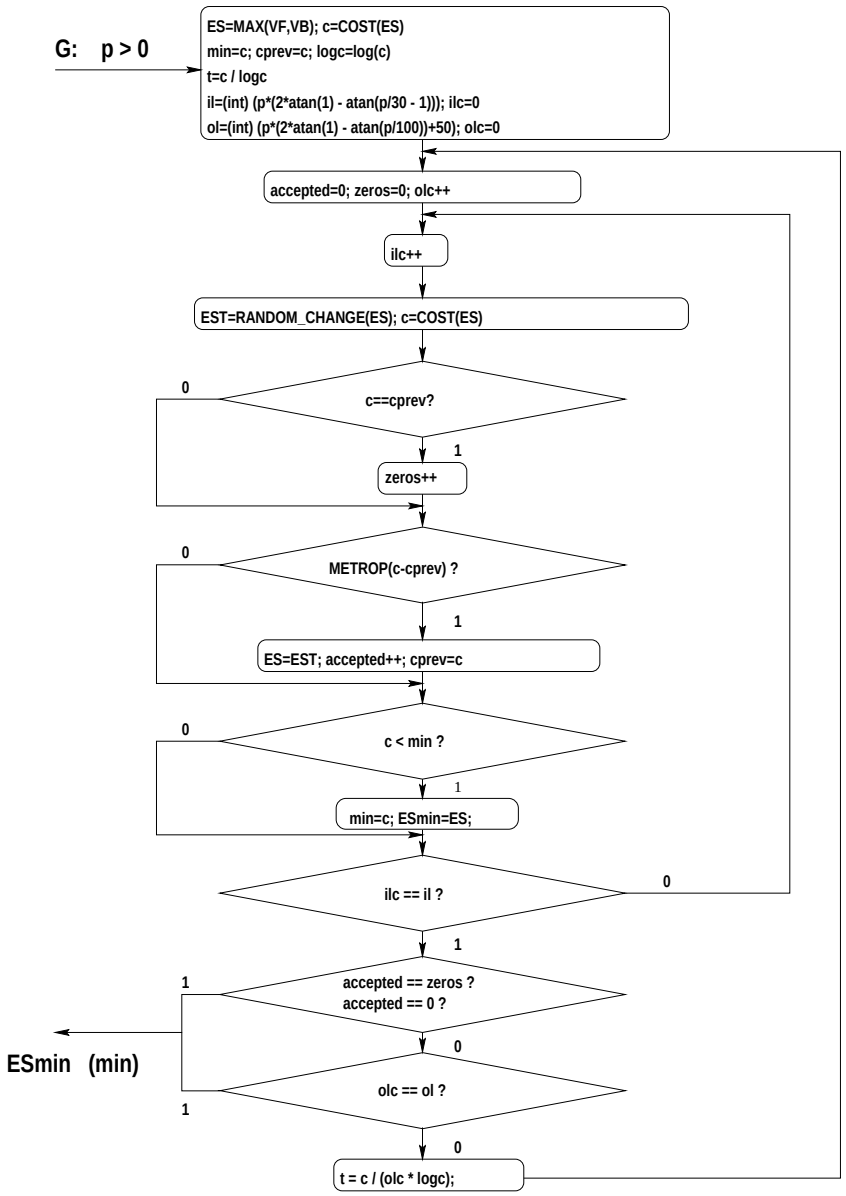


FIG. 3.1. *Simulated annealing algorithm.*

there is at least one vertex to be eliminated from $\mathbf{G}$. The maximal number of cooling steps lies always between 50 and 150. It increases rapidly for small values of `p`, whereas it converges to 150 for larger graphs. Analogously, we observe a steep ascent in the first part (`0<p<50`) of the curve for `il`. Again, it settles for a value around 40 for increasing numbers of intermediate variables. Choosing both values as we did assures that the algorithm will always terminate after a reasonable and in most cases sufficient run-time.

The "temperature" `t` is lowered with every iteration of the outer loop (*inhomogeneous part*) and is kept constant while running through the inner loop (*homogeneous*

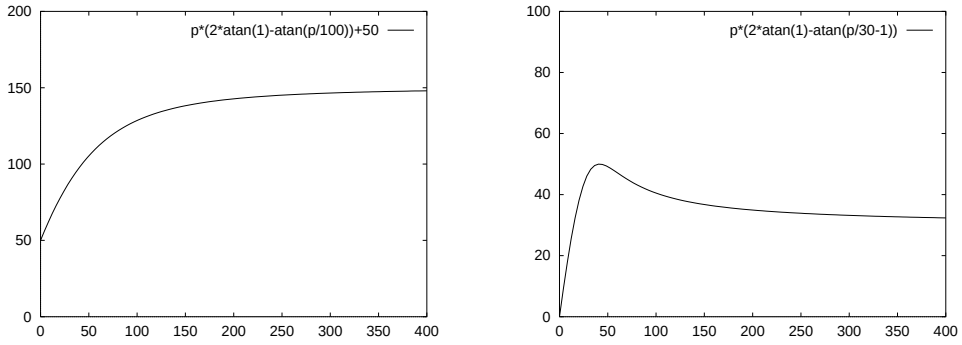


FIG. 3.2. *Bounds on iterations in inner and outer loop.*

part). Within the latter we randomly change the elimination sequence using the two rearrangements suggested above. Whether this new, so far temporary, order (`EST`) is accepted or not depends on the Metropolis criterion [12]. It returns a Boolean variable which issues a verdict on whether to accept a reconfiguration leading to a change `d = c-cprev` in the objective function `COST`. If `d` ≤ 0, then `EST` will always be accepted. If `d` > 0, then the answer of `METROP(d)` will only be positive with probability `exp(-d/t)`. As we have already pointed out, `t` is not changed within the inner loop. However, with every iteration (`olc`) of the outer loop we reinitialize `t` as `t=c/(olc*logc)`. Notice that the new value of `t` depends on the current overall Markowitz degree, i.e., it depends on the cost of the last accepted elimination sequence. With this approach we obtain for the jth outer iteration

$$t_j = \frac{c_{j-1}}{(j-1)\cdot\log(c_0)} \quad \Rightarrow \quad t_{j+1} = \frac{c_j}{j\cdot\log(c_0)}.$$

From the above we can derive the *cooling rate* Δt_j:

$$\Delta t_j = \frac{t_{j+1}}{t_j} = \frac{(j-1)\cdot c_j}{j\cdot c_{j-1}}.$$

Why have we chosen to let `t` develop this way? Suppose we have taken a step upwards (accepted an elimination sequence which results in an increase of the objective function), i.e., $c_j > c_{j-1}$. Then we do not want to continue the cooling process at the same speed as before since this could lead to being unable to leave the current "valley." The temperature is eventually lowered as a result of the increasing outer loop counter `olc`. However, the extent to which the system is cooled depends both on the current temperature and on the change of the overall Markowitz degree during the last iteration. This approach worked well in most cases.

Throughout the entire annealing process we always keep track of the minimal Markowitz degree (`min`) resulting from any of the elimination orders that were checked so far. Thus, we do not depend entirely on the convergence of the algorithm. Even a very good elimination sequence which was found "by luck" in the high temperature phase of the annealing process can be the result of running the algorithm. After each outer loop iteration we check the *exit criteria*. There are three:

1. The maximal number of iterations to be performed is reached (`olc` == `ol`).
2. During the last outer loop iteration, none of the rearrangements has been accepted (`accepted` == 0).

3. There may be rearrangements which do not lead to any change in the cost function. All of them will be accepted. However, if all rearrangements accepted during one outer loop iteration are such, i.e., `accepted == zeros`, then the algorithm will be terminated.

Finally, the simulated annealing algorithm delivers an elimination sequence with the minimal cost, i.e., a vertex elimination order which approximates the minimal number of multiplications required for the accumulation of the whole Jacobian.

Summarizing the above, we have the following list of parameters to experiment with when looking for an optimal annealing schedule: initial elimination sequence, initial temperature, cooling rate, types of rearrangements, number of iterations with constant temperature, number of cooling steps, and acceptance philosophy. Therefore, there is certainly plenty of room for experimentation. The method described above, i.e., the simulated annealing algorithm based on the above configuration and annealing schedule, will be referred to as SAC.

While looking for the "optimal schedule" we have implemented four additional versions of the simulated annealing algorithm. In SAR we use the reversal of dense subsequences as the only rearrangement action. SAT is similar except that only transportation of dense subsequences is allowed. Furthermore, we have experimented with two different annealing schedules. SACS is similar to the described method with

$$t_j = \frac{c_{j-1}}{2j} \quad \Rightarrow \quad \Delta t_j = \frac{t_{j+1}}{t_j} = \frac{j \cdot c_j}{(j+1) \cdot c_{j-1}},$$

i.e., we have slowed the cooling process down. In SACF we increase the temperature in the second step, which can be regarded as the generation of a new random initial elimination sequence, as we accept almost every rearrangement. Then we cool the system down with

$$t_j = \frac{c_{j-1}}{j^2} \quad \Rightarrow \quad \Delta t_j = \frac{t_{j+1}}{t_j} = \frac{j^2 \cdot c_j}{(j+1)^2 \cdot c_{j-1}},$$

which is much faster than in the main method. In the algorithm described by Figure 3.1 we have used the following symbolism:

`ES, EST, ESmin` $\in \mathbb{N}^p$:	elimination sequences,
`c, cprev, min` $\in \mathbb{N} \cup \{0\}$:	cost values,
`il, ilc, ol, olc` $\in \mathbb{N} \cup \{0\}$:	for loops,
`t` $\in \mathbb{R}$:	temperature,
`logc` $\in \mathbb{R}$:	logarithm of initial cost,
`accepted, zeros` $\in \mathbb{N} \cup \{0\}$:	counters.

4. Case studies. The success of simulated annealing depends on extensive testing and variation of parameters. Of course, we would like the algorithm to deliver nearly optimal results in virtually all cases without always having to adapt the parameters to the given problem. Furthermore, it should not run too long even for large-scale problems. We have applied SAVE to many test problems from which we will select a small subset to discuss in detail. As supported by the results achieved, our main version of the simulated annealing algorithm (SAC) performs well on a large number of problems. However, there are certain exceptions indicating that it is probably impossible to come up with a strategy (annealing schedule) which is universally optimal for an a priori fixed stopping time.

4.1. Speelpenning function. The first test problem was used by Speelpenning in his thesis [21] as an illustration of the powers of the reverse mode of AD; it represents a simple chained product of the n independent variables $y = f(\mathbf{x}) = \prod_{i=0}^{n-1} x_i$ (SPF). For $n = 50$ the c-graph contains 48 intermediate vertices. It can be checked that VB is optimal on this problem using 96 multiplications for the computation of the gradient of f. On the other hand, running VF results in an overall Markowitz degree of 1224 as a consequence of the successive increase of the in-degrees of the intermediate vertices. The performance of LM would depend on the choice for the primary tie-break criterion. LM would deliver the optimal solution if the latter were chosen as VB. Setting it to VF would result in an overall Markowitz degree of 142.

We will use SPF as an illustration of the sensitivity of simulated annealing with respect to different annealing schedules. Figure 4.1 shows the results for all variations of our simulated annealing algorithm, which form the five columns of the table below. The three columns in the figure show the development of the temperature over the first 20 outer loop iterations, the changes in the objective function after each inner loop iteration, and the course of the overall Markowitz degree which we intend to minimize, respectively. Apart from differences in the run-time ($t_{\texttt{user}}$), the five methods converged to different final values of the overall Markowitz degree (`min`).

	SAC	SAT	SAR	SACS	SACF
`min`	129	154	189	136	142
$t_{\texttt{user}}$	93 sec	63 sec	105 sec	171 sec	37 sec

Furthermore, we observe the following:

1. Neither a higher nor a lower cooling rate leads to improvements in the value of the objective function. However, the run-time of SACF undercuts that of SAC by a factor of nearly 3. Still, it delivers an acceptable result.

2. Starting with the VF-based elimination sequence, the reversal of substrings could be expected to lead to a good solution in a short time. Obviously, this is not the case. Why? The reason lies in the structure of the Speelpenning function. Whenever we eliminate the vertices of a part of the c-graph backward, this can be regarded as "good" for the value of our objective function. However, running forward is certainly bad. Now, if we allow the reversal of substrings of a given elimination sequence to be the only rearrangement in the annealing process, this can lead to the repeated negation of savings made in the current step by the rearrangements to come.

3. Depending on the accepted rearrangements and the resulting overall Markowitz degree, the temperature is lowered at varying speeds. This becomes especially clear when looking at the graphs for SACS in the fourth row of Figure 4.1.

4. The largest decreases in the objective function value are achieved in the high temperature phase. This is certainly not very surprising.

5. It might often be advantageous to choose a high cooling speed in order to get useful estimates for the savings that can be expected.

It is difficult to state something of general validity when speaking about the practicality of simulated annealing when applied to computationally hard combinatorial optimization problems. However, the approach that we have chosen as our main method (SAC) turns out to deliver the best results in virtually all cases.

4.2. Steady state combustion problem. The steady state combustion problem (SSC) is the variational formulation of an underlying boundary value problem [3]. For the examination of the behavior of simulated annealing we have chosen the case

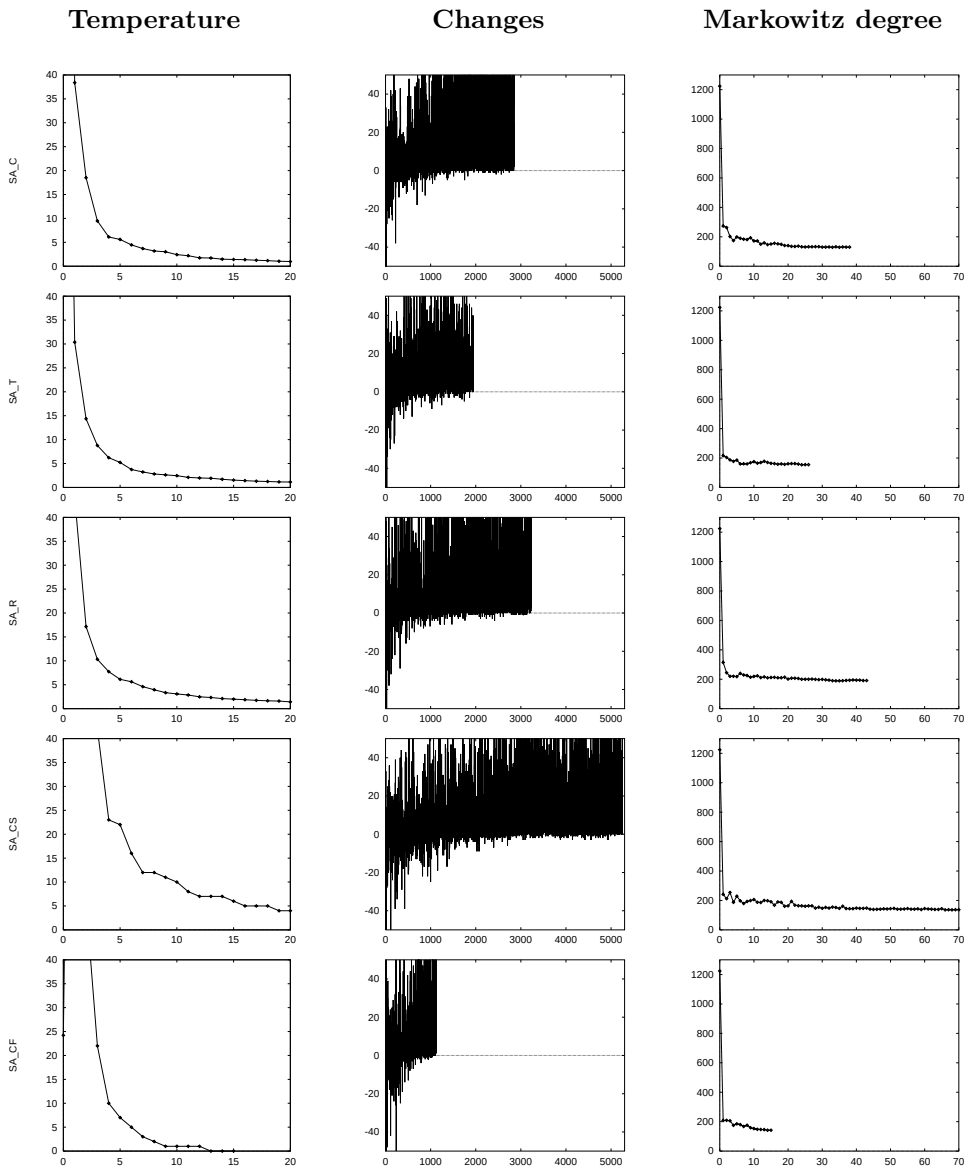


FIG. 4.1. *SAVE on SPF.*

with $n = 4$ independent variables. This leads to a relatively small c-graph containing 103 intermediate vertices which permits a closer look at our algorithm.

SAC converged to an elimination sequence that took 123 multiplications to compute the gradient. The VB-based sequence (142 multiplications) served as the starting point. The best known value of the objective function achieved by applying the *enhanced LM* heuristic proposed in [14] is 122. Figure 4.2 shows the courses taken by the temperature t and by the overall cost c.

With `il` = 47 and `ol` = 130, the algorithm had to terminate after checking 6110 vertex elimination sequences. In fact, it performed only 59 outer loop iterations corresponding to the generation of 2773 (not necessarily different) elimination orders,

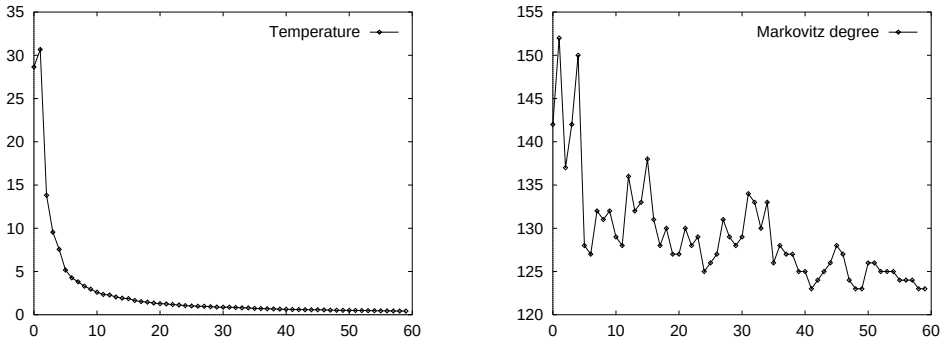


FIG. 4.2. *SSC.*

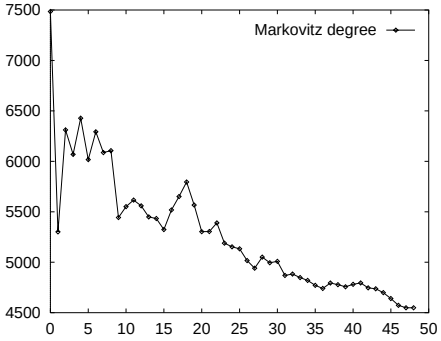


FIG. 4.3. *CQ.*

after which one of the *exit criteria* described in section 3 was met. Exploiting one of the advantages of our simulated annealing algorithm, the overall Markowitz degree is increased repeatedly; thus, it is able to escape from local minima.

Obviously, the savings which can possibly be achieved are not so remarkable that they could justify the effort. It makes sense to apply simulated annealing to the vertex elimination problem in c-graphs if we care about the result only and ignore the time that it took to compute the elimination sequences. In most cases heuristics will deliver sequences nearly as good (or better) at much higher speed.

4.3. Chebyshev quadrature problem. Both the SSC and the Chebyshev quadrature problem (CQ) are taken from the `MINPACK` test problem collection [3]. For the latter we have chosen $n = 15$ and $m = 16$, resulting in a c-graph of reasonable size. Remember that we always take either the VF- or the VB-based elimination sequence as an initialization of the simulated annealing algorithm, depending on which of them delivers the higher cost. Starting with a relatively "bad" initial elimination sequence turned out to be advantageous for the behavior of the algorithm in many cases. Now, CQ represents a counterexample, as the starting sequence is obviously "not good enough" for a problem of the given size and structure. The operations count resulting from the application of the VB mode (8400) is about four times as large as the one of the VF mode (1980). The algorithm "cools the system down" to 4842 which is approximately half the number of multiplications compared to its starting value. However, it does not reach the value which was calculated for VF. Figure 4.3 shows the development of the overall Markowitz degree. After some ups and downs

TABLE 5.1
Numerical results for some MINPACK-2 test problems.

	n	p	m	DM	NR	LM	SAC	NR/SAC
FDC	16	984	16	16000	11000	1338	1234	**8.9**
FCH	32	1209	32	39712	11169	845	851	**13.1**
DIE	20	2499	20	50380	50380	1659	1660	**30.3**
VDI	100	504	100	60400	60400	10401	10337	**5.8**
EXP	96	479	1	575	575	504	504	**1.14**

in the first section of the annealing process, caused by the initially high temperature, the cost decreases continuously, unfortunately, not reaching the minimum. As a way to overcome this problem we could think of slowing the cooling process down in order to allow more iterations to be performed. Representing this approach, SACS results in a cost of 4343. In fact, we observe an improvement compared to SAC. However, it is not very encouraging when taking into account that it took about four times as long to achieve the improvement. With VF as primary tie-break criterion, LM delivered the same solution as VF in this case.

5. Further results. The c-graphs of all test problems were built using the *tape* generated by the AD tool ADOL-C [7]. Our objective was to keep the size of the graphs relatively small. One has to keep in mind that Jacobian accumulation should primarily be regarded as a derivative code optimization technique performed at compile time. Thus, it will be applied to basic blocks or, possibly using profiling information, to unrolled loops performing a small number of iterations. The results presented here show the potential savings that could be achieved following this approach using versions of the corresponding c-graphs generated at run-time.

We will consider the following examples from the MINPACK test problem collection [3]: Flow in a driven cavity problem (FDC), flow in a channel problem (FCH), discrete integral equation function (DIE), variably dimensioned function (VDI), and extended Powell function (EXP).

Vertex elimination in c-graphs fully exploits the structural sparsity of a given problem. Naturally, this applies to both VF and VB. In most cases this alone may lead to significant savings in the overall operations count. The solution of the vertex elimination problem may add another factor which varies from problem to problem.

In Table 5.1 we have compared the values delivered by the SAC method with the best choice out of dense forward and reverse modes (DM) and the corresponding minimum value achieved by one of the two unidirectional methods by Newsam and Ramsdell (NR) as introduced in section 1.

Considering the ratio between the optimal one-sided Newsam–Ramsdell approach and the SAVE, we observe that, in fact, large savings can be achieved. The differences between LM and SAC are not so remarkable, which supports the thesis that local heuristics are still a very good trade-off between efficiency and quality of the result. From the compiler optimization point of view, Jacobian accumulation will certainly be based on heuristics. However, libraries for scientific computing which are extensively used in numerous applications may well apply for optimization using simulated annealing in order to try to improve the corresponding derivative code even further.

Simulated annealing algorithms are well suited for parallelization, which is not the case for greedy heuristics such as LM. The development of parallel simulated annealing algorithms for various elimination techniques in computational graphs [13] is one of the

objectives of an ongoing research project at the University of Hertfordshire, Hatfield, UK [16]. There we also investigate the consequences of restricting simulated annealing to the loop bodies, which is crucial for loops with variable bounds. The simultaneous treatment of two or more iterations may become important if the computational graph of the loop body contains a vertex cut with many fewer elements than the numbers of both minimal and maximal vertices. Also, the combination of several iterations into one local Jacobian may decrease the overall computational effort significantly. However, the question of how to decide when and how many iterations to combine is still open. On a global level, preaccumulation techniques for local Jacobians will have to be used in conjunction with other AD techniques such as sparse vector modes [4], [6] or seed matrix compression techniques [1], [19]. Assuming that the number of derivative values to be computed is large enough to justify the effort required for preaccumulation, one should strive for local Jacobians of large parts of the program. Algorithmic implications thereof are subject to future research.

6. Conclusion and future work. In this paper we have applied simulated annealing to the vertex elimination problem in linearized c-graphs. The intention was to suggest a method for solving this computationally hard combinatorial optimization problem whose robustness can be adjusted by varying certain parameters. It turned out to deliver nearly optimal results in most cases while lacking the implementation simplicity of local heuristics. However, if the run-time of the optimization algorithm is not a crucial parameter, then simulated annealing represents a very flexible alternative. Regarding the generation of efficient Jacobian code as a compile-time procedure, the latter may be the case in many large-scale applications.

Naturally, the minimization of the number of arithmetic operations does not necessarily result in efficient Jacobian code. Memory accesses may dominate the run-time of the algorithm. Their locality can be ensured by regarding local Jacobians of smaller sizes, such as basic blocks or loop bodies, as part of a hierarchical approach as outlined by Bischof and Haghighat in [5]. The aim is certainly not to accumulate the whole Jacobian of some real-world problem given in the form of a computer program. Such programs contain branches and loops with exit criteria which cannot be predicted at compile time, in general. Local Jacobians resulting from, for example, CFD kernels as discussed in [22], are what we are talking about. Profiling information generated by preliminary executions of the program with a typical set of inputs can help identify cost- intensive parts as well as values of loop exit criteria and branches taken. In cases where the values of array indices cannot be determined at compile time, it may still be useful to accumulate local Jacobians without looking at the vertex elimination problem. Although the shape of the c-graph cannot be predicted correctly, the accumulated Jacobian will ensure the exploitation of the graph's (i.e., the Jacobian's) structural sparsity. Depending on whether we choose a VF or VB sequence, this approach would be equivalent to the sparse forward or reverse modes as described in chapter 6 of Griewank [6]. In any case, significant savings can be expected when comparing with DM or Newsam and Ramsdell's method.

What is the main difference between the AD specific optimizations proposed here and the optimizations performed by most modern compilers? First, all other methods for exploiting the structural sparsity of the Jacobian are, in fact, not compile time solutions. They are implemented as special algorithms by the programmer (such as seed matrix compression techniques) or they propagate dependency flags at run-time as implemented in ADIFOR's [4] SparseLinC library. Furthermore, no compiler can currently solve the corresponding combinatorial optimization problem. However, this

does not imply that these algorithms could not become optimizations performed by a differentiation enabled compiler. A collaborative project between the author and NAG Ltd. in Oxford, UK, is currently looking at this issue [17].

Apart from the very practical impact that simulated annealing can make on the run-time of Jacobian code, we expect it to play a central role during the exploration of certain theoretical aspects associated with the *Optimal Jacobian Accumulation* problem. In [13], we have shown that an optimal vertex elimination sequence does, in general, not minimize the number of multiplications involved in the accumulation of Jacobian matrices. *Edge* and *face* elimination techniques were introduced and their superiority was demonstrated. So far, it is unknown how large the *vertex-edge-discrepancy* and the *edge-face-discrepancy* actually are. We conjecture that the best vertex elimination sequence will never involve more than twice the number of multiplications required by the optimal Jacobian code. In order to support this conjecture by numerical results, we are currently implementing simulated annealing algorithms for both edge and face elimination in c-graphs as part of the XCOp project [16] running at the University of Hertfordshire, UK. Anything other than a small discrepancy would mean that the methods used in modern AD tools will not even get close to the possible minimal number of arithmetic operations for selected, probably mostly theoretical, problems.

REFERENCES

[1] M. Powell, A. Curtis, and J. Reid, *On the estimation of sparse Jacobian matrices*, J. Inst. Math. Appl., 13 (1974), pp. 117–119.

[2] A. Albrecht and C. Wong, *On logarithmic simulated annealing*, in Proceedings of IFIP International Conference on Theoretical Computer Science, J. van Leeuwen, O. Watanabe, M. Hagiya, P. Mosses, and I. Ito, eds., Lecture Notes in Comput. Sci. 1872, Springer, Berlin, 2000, pp. 301–314.

[3] B. Averik, R. Carter, and J. More, *The Minpack-2 Test Problem Collection (Preliminary Version)*, Technical Report 150, Mathematical and Computer Science Division, Argonne National Laboratory, Argonne, IL, 1991.

[4] C. Bischof, A. Carle, P. Khademi, and A. Maurer, *The ADIFOR* 2.0 *system for automatic differentiation of Fortran* 77 *programs*, IEEE Comp. Sci. Eng., 3 (1996), pp. 18–32.

[5] C. Bischof and M. Haghighat, *Hierarchical approaches to automatic differentiation*, in Computational Differentiation: Techniques, Applications, and Tools, Proc. Appl. Math. 89, SIAM, Philadelphia, 1996, pp. 83–94.

[6] A. Griewank, *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*, SIAM, Philadelphia, 2000.

[7] A. Griewank, D. Juedes, and J. Utke, *ADOL-C, a package for the Automatic Differentiation of algorithms written in C/C++*, ACM Trans. Math. Software, 22 (1996), pp. 131–167.

[8] A. Griewank and S. Reese, *On the calculation of Jacobian matrices by the Markovitz rule*, in Automatic Differentiation of Algorithms: Theory, Implementation, and Application, G. Corliss and A. Griewank, eds., SIAM, Philadelphia, 1991, pp. 126–135.

[9] R. Jessani and M. Putrino, *Comparison of single- and dual-pass multiply-add fused floating-point units*, IEEE Trans. Comput., 47 (1998), pp. 927–937.

[10] S. Lin, *Computer solutions to the traveling salesman problem*, Bell Systems Tech. J., 44 (1965), pp. 2245–2269.

[11] K. Mehlhorn and S. Näher, *LEDA, a platform for combinatorial and geometric computing*, Commun. ACM, 38 (1996), pp. 96–102.

[12] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, *Equation of state calculations by fast computing machines*, J. Chem. Phys., 21 (1953), pp. 1087–92.

[13] U. Naumann, *Elimination techniques for cheap Jacobians*, in Automatic Differentiation of Algorithms—From Simulation to Optimization, G. Corliss, C. Faure, A. Griewank, L. Hascoet, and U. Naumann, eds., Springer, New York, 2002.

[14] U. Naumann, *An enhanced Markovitz rule for accumulating Jacobians efficiently*, in ALGORITMY 2000—15th Conference on Scientific Computing, A. Handlovicova, M. Ko-

mornikova, K. Mikula, and A. Sevcovic, eds., Slovak University of Technology, Bratislava, Slovakia, 2000, pp. 320–329.

[15] U. Naumann, *Efficient Calculation of Jacobian Matrices by Optimized Application of the Chain Rule to Computational Graphs*, Ph.D. thesis, Technical University Dresden, Dresden, Germany, 1999.

[16] U. Naumann and A. Albrecht, *Combinatorial optimization methods for fast derivative code*, EPSRC grant GR/R38101/01, 2001.

[17] U. Naumann and B. Christianson, *Differentiation enabled compiler technology*, MODDERA/EPSRC grant GR/R55252/01, 2001.

[18] U. Naumann and P. Gottschling, *Prospects for simulated annealing in automatic differentiation*, in Stochastic Algorithms: Foundations and Applications, K. Steinhöfel, ed., Lecture Notes in Comput. Sci. 2264, Springer, New York, 2001.

[19] G. Newsam and J. Ramsdell, *Estimation of sparse Jacobian matrices*, SIAM J. Alg. Dis. Meth., 4 (1983), pp. 404–418.

[20] W. Press, A. Teukolsky, W. Vetterling, and B. Flannery, *Numerical Recipes in C*, Cambridge University Press, Cambridge, UK, 1992.

[21] B. Speelpenning, *Compiling Fast Partial Derivatives of Functions Given by Algorithms*, Ph.D. thesis, University of Illinois, Urbana-Champaign, IL, 1980.

[22] M. Tadjouddine, S. Forth, and J. Pryce, *AD tools and prospects for optimal AD in CFD flux calculations*, in Automatic Differentiation of Algorithms—From Simulation to Optimization, G. Corliss, C. Faure, A. Griewank, L. Hascoet, and U. Naumann, eds., Springer, New York 2002.

SIAM J. OPTIM.
Vol. 13, No. 3, pp. 675–692

NONLINEAR LAGRANGIAN FOR MULTIOBJECTIVE OPTIMIZATION AND APPLICATIONS TO DUALITY AND EXACT PENALIZATION*

X. X. HUANG† AND X. Q. YANG‡

Abstract. Duality and penalty methods are popular in optimization. The study on duality and penalty methods for nonconvex multiobjective optimization problems is very limited. In this paper, we introduce vector-valued nonlinear Lagrangian and penalty functions and formulate nonlinear Lagrangian dual problems and nonlinear penalty problems for multiobjective constrained optimization problems. We establish strong duality and exact penalization results. The strong duality is an inclusion between the set of infimum points of the original multiobjective constrained optimization problem and that of the nonlinear Lagrangian dual problem. Exact penalization is established via a generalized calmness-type condition.

Key words. multiobjective optimization, nonlinear Lagrangian function, duality, exact penalization, stability

AMS subject classifications. 90C29, 90C46

PII. S1052623401384850

1. Introduction and preliminaries.

It is well known that the traditional Lagrange function plays an important role in both theory and methodology for single objective and multiobjective convex optimization problems, such as optimality condition, duality theory, saddle point theory, sensitivity analysis, and solution method [2, 19]. However, it becomes less effective for nonconvex optimization problems. For example, there may be a nonzero duality gap between the single objective nonconvex constrained optimization problem and its Lagrange dual problem. Thus the Lagrange method may fail for nonconvex optimization problems. Moreover, it is worth noting that a zero duality gap can be achieved for a single objective nonconvex optimization problem using an augmented Lagrangian function; see [14]. A more general scheme of the conjugate framework was established for convex and nonconvex cases in [12, 1], respectively. On the other hand, exact penalty functions and their applications in the study of optimality conditions were provided for single objective constrained optimization problems in, e.g., [4, 14, 15] under calmness conditions. See [3] for an excellent review.

Recently, a class of nonlinear Lagrangian functions was introduced and applied to establish a zero duality gap for single objective constrained continuous optimization problems without any convexity requirement [5, 17]. The terminology "nonlinear" refers to the nonlinearity of the objective function of the transformed problems with respect to the objective function of the original constrained optimization problem. The exact penalization result for nonconvex inequality constrained single objective optimization was obtained under a generalized calmness condition in [18]. It is worth

*Received by the editors February 12, 2001; accepted for publication (in revised form) May 3, 2002; published electronically November 6, 2002. This work was partially supported by the Australian Research Council and the Research Grant Council of Hong Kong (grant B-Q359).
http://www.siam.org/journals/siopt/13-3/38485.html

†Department of Mathematics and Computer Science, Chongqing Normal University, Chongqing 400047, People's Republic of China. Current address: Department of Applied Mathematics, The Hong Kong Polytechnic University, Kowloon, Hong Kong (mahuangx@polyu.edu.hk).

‡Department of Applied Mathematics, The Hong Kong Polytechnic University, Kowloon, Hong Kong (mayangxq@polyu.edu.hk).

noting that the early study on the nonlinear Lagrangian can be found in the work [23]. Moreover, a pth power transformation was introduced in [9] to guarantee a zero duality gap for an optimization problem, which is not necessarily convex.

In this paper, we introduce a class of nonlinear Lagrangian functions and nonlinear Lagrangian dual problems for (nonconvex) multiobjective optimization problems. In particular, we obtain a strong duality result between a constrained multiobjective optimization problem and its nonlinear Lagrangian dual problem without any convexity requirement. Several types of exact penalization for nonlinear penalty multiobjective optimization problems are investigated. We study conditions which guarantee

(i) there is a finite penalty parameter vector such that every infimum point of the original constrained multiobjective optimization problem is an infimum point of the nonlinear penalty multiobjective optimization problem (global exact penalization); and

(ii) for each infimum point of the original constrained multiobjective optimization problem, there is a finite penalty parameter vector such that this point is also an infimum point of the nonlinear penalty multiobjective optimization problem (local exact penalization).

The motivation of our study is that there is only limited study on duality and penalty methods for nonconvex multiobjective optimization problems. Yet these approaches are popular solution methods in single objective optimization. For convex multiobjective optimization problems, systematic study of Lagrangian duality and conjugate duality was given in [19, 10] and the references cited therein. To the best of our knowledge, investigation on the conventional penalty function method for constrained multiobjective optimization problems was only given in [16, 20]. We will establish strong duality for multiobjective optimization problems without any convexity requirement. The condition used is the lower semicontinuity of the functions involved, which is much weaker than the continuity assumption in [17]. Moreover, the conditions for exact penalization are a generalization of the ones for single objective optimization in [3, 4, 15, 18]. It is worth noting that nonlinear Lagrangian dual problems studied in this paper provide new models for convex composite optimization problems studied in [6, 7, 21].

Let R^l be an l-dimensional Euclidean space, $C = R^l_+$, and int$C$ be the interior of C. Define the following orderings: for any $z^1, z^2 \in R^l$,

$$\begin{gathered} z^1 \leq_C z^2 \Longleftrightarrow z^2 - z^1 \in C,\ z^1 \not\leq_C z^2 \Longleftrightarrow z^2 - z^1 \notin C, \\ z^1 \leq_{C\setminus\{0\}} z^2 \Longleftrightarrow z^2 - z^1 \in C\setminus\{0\},\ z^1 \not\leq_{C\setminus\{0\}} z^2 \Longleftrightarrow z^2 - z^1 \notin C\setminus\{0\}, \\ z^1 \leq_{\mathrm{int}C} z^2 \Longleftrightarrow z^2 - z^1 \in \mathrm{int}C,\ z^1 \not\leq_{\mathrm{int}C} z^2 \Longleftrightarrow z^2 - z^1 \notin \mathrm{int}C. \end{gathered}$$

Let $e = (1, \ldots, 1) \in \mathrm{int}C$, and $e_i = (0, 0, \ldots, 1, 0, \ldots, 0)$ (the ith component is 1 and the other components are 0's), $i = 1, \ldots, l$.

Consider the following multiobjective constrained optimization problem:

$$\begin{array}{lll} \text{(MOP)} & \inf\limits_{x\in X} & f(x) \\ \text{such that (s.t.)} & & g_j(x) \leq 0, \quad j = 1, \ldots, m, \end{array}$$

where $X \subseteq R^n$ is a nonempty closed set, $f = (f_1, \ldots, f_l) : X \to R^l$ is a vector-valued function such that each of its component function f_i is lower semicontinuous (l.s.c.), and $g_j : X \to R^1$ is l.s.c. for any $j \in \{1, \ldots, m\}$.

By X_0 we denote the set of feasible solutions of (MOP). That is, $X_0 = \{x \in X : g_j(x) \leq 0, j = 1, \ldots, m\}$. It is clear that X_0 is closed.

We say that $x^* \in X_0$ is *an efficient solution* to (MOP) if there exists no $x \in X_0$ such that $f(x) \leq_{C\setminus\{0\}} f(x^*)$. The corresponding function value $f(x^*)$ is called an *efficient point* of (MOP). We denote by $E(0)$ the set of the efficient solutions of (MOP).

The point $x^* \in X_0$ is called *a weakly efficient solution* to (MOP) if there exists no $x \in X_0$ such that $f(x) \leq_{\text{int}C} f(x^*)$. The corresponding point $f(x^*)$ is called a *weakly efficient point* of (MOP). The set of weakly efficient solutions of (MOP) is denoted by $WE(0)$.

The point $x^* \in X_0$ is said to be *a locally weak efficient solution* to (MOP) if there exists $\delta > 0$ such that $f(x) \not\leq_{\text{int}C} f(x^*)$ for any $x \in X_0$ with $\|x - x^*\| \leq \delta$. The set of all locally weak efficient solutions of (MOP) is denoted by $LWE(0)$.

We denote by $V(0)$ the set of *infimum points* of (MOP), i.e., $V(0) = \inf_{x\in X_0} f(x)$. Namely, $z \in V(0)$ if and only if (i) $f(x) \not\leq_{C\setminus\{0\}} z\ \forall x \in X_0$ and (ii) $\exists x_k \in X_0$ such that $f(x_k) \to z$ as $k \to \infty$.

Clearly, if x_0 is an efficient solution to (MOP), then $f(x_0) \in V(0)$.

Without loss of generality, we assume throughout this paper that $\min_{1\leq i\leq l} \inf_{x\in X} f_i(x) \geq 0$. If this assumption does not hold, then consider the following optimization problem:

$$\begin{array}{lll} (\text{MOP}') & \inf\limits_{x\in X} & (\exp(f_1(x)) + 1, \ldots, \exp(f_l(x)) + 1) \\ & \text{s.t.} & g_j(x) \leq 0, \quad j = 1, \ldots, m. \end{array}$$

It is clear that the sets of efficient solutions and weakly efficient solutions of (MOP) are the same as that of (MOP′), respectively.

Throughout this paper, for simplicity, we shall use the notation $\|u\|_\gamma$ to denote the formula $[\sum_{j=1}^m |u_j|^\gamma]^{1/\gamma}$, where $u = (u_1, \ldots, u_m) \in R^m, \gamma \in (0, +\infty)$.

Let $y^1 = (y_1^1, \ldots, y_m^1), y^2 = (y_1^2, \ldots, y_m^2) \in R^m$, define the notation of componentwise product for y^1 and y^2:

$$y^1 * y^2 \equiv (y_1^1 y_1^2, \ldots, y_m^1 y_m^2).$$

Let Z_1 be a subset of a metric space Z, and $z \in Z$. Denote by $d(z, Z_1)$ the distance from the point z to the set Z_1.

The outline of the paper is as follows. In section 2, strong duality for (MOP) and its nonlinear Lagrangian dual problem (DMOP) (see next section) is established. In section 3, conditions are given which are necessary and sufficient for the existence of a global (local) exact penalty parameter. In section 4, we consider saddle points of the nonlinear Lagrangian.

2. Nonlinear Lagrangian functions and duality. Let $A \subseteq R^l \times R^m$. A vector-valued function $p : A \to R^l$ is called *increasing* on the set A if for any $(z^i, y^i) \in A (i = 1, 2)$ with $(z^1, y^1) - (z^2, y^2) \in C \times R_+^m$ we have $p(z^1, y^1) \geq_C p(z^2, y^2)$.

Let p be an increasing vector-valued function defined either on the domain $C \times R^m$ or on the domain $C \times R_+^m$ such that each of its component functions p_i is l.s.c. and p enjoys the following two properties:

(A) There exist positive real numbers $a_j (j = 1, \ldots, m)$ such that for any $z \in C, y = (y_1, \ldots, y_m)$ with (z, y) belonging to the domain of p, $p(z, y) \geq_C z$ and $p(z, y) \geq_C (\max_{1\leq j\leq m}\{a_j y_j\})e$.

(B) $\forall z \in C, p(z, 0, \ldots, 0) = z$.

Remark 2.1. This reduces to the function p of [17] when $l = 1$ and p is continuous.

It is easy to prove the following elementary proposition.

PROPOSITION 2.1. *Let $p(z,y) = p'(p'(z,y),y)$, where p' is an increasing function with properties* (A) *and* (B). *Then p is also an increasing function having properties* (A) *and* (B).

Example 2.1. Let $z = (z_1,\dots,z_l), y = (y_1,\dots,y_m)$, and $(z,y) \in C \times R^m$. Some examples of the increasing function p defined on $C \times R^m$ having properties (A) and (B) are as follows:

$p_\infty(z,y) = \sum_{i=1}^{l} \max\{z_i, y_1,\dots,y_m\}e_i$;

$p_\gamma(z,y) = \sum_{i=1}^{l}(z_i^\gamma + \sum_{j=1}^{m} y_j^{+\gamma})^{1/\gamma}e_i, 0 < \gamma < \infty$, where $y_j^+ = \max\{y_j, 0\}, j = 1,\dots,m$;

$p(z,y) = z + (\sum_{j=1}^{m} b_j y_j^+)e$, where $b_j > 0, j = 1,\dots,m$.

Example 2.2. The restrictions of p_∞, p_γ, p (considered in Example 2.1) to $C \times R_+^m$ are increasing functions defined on $C \times R_+^m$ having properties (A) and (B).

In the rest of this section, p is assumed to be an increasing function defined on $C \times R^m$ with properties (A) and (B), and this section concludes with a remark for the case when p is defined on $C \times R_+^m$.

Let

$$F(x,d) = (f(x), d * g(x)),$$

where $d = (d_1,\dots,d_m) \in R_+^m$ and $g(x) = (g_1(x),\dots,g_m(x))$.

The *nonlinear Lagrangian* function corresponding to p for (MOP) is defined as

$$L(x,d) = p(F(x,d)). \tag{1}$$

DEFINITION 2.2. *The following problem,*

$$\text{(DMOP)} \quad \sup_{d \in R_+^m} q(d),$$

where $q(d) = \inf_{x \in X} L(x,d)\ \forall d \in R_+^m$, is called the nonlinear Lagrangian dual problem to (MOP) *corresponding to p. Here by $z \in \sup_{d \in R_+^m} q(d)$ we mean that*

(i) $(z - q(d)) \bigcap (-C \backslash \{0\}) = \emptyset\ \forall d \in R_+^m$;

(ii) $\exists d^j \in R_+^m$ *and* $z^j \in q(d^j)$ *such that* $z^j \to z$ *as* $j \to +\infty$.

z is called a supremum point of (DMOP).

Remark 2.2. If p is convex, e.g., all the p's except p_γ in the case of $\gamma \in (0,1)$ in Example 2.1, the problem of computing $q(d)$,

$$\inf_{x \in X} p(F(x,d)),$$

is a type of convex composite multiobjective optimization problem studied in [7].

It is elementary to prove the following results.

LEMMA 2.3. *Let p be an increasing function with properties* (A) *and* (B). *Then $p(F(x,d)) = f(x)\ \forall x \in X_0,\ d \in R_+^m$.*

PROPOSITION 2.4 (weak duality). $\forall x \in X_0, d \in R_+^m, (q(d) - f(x)) \bigcap (C \backslash \{0\}) = \emptyset$.

COROLLARY 2.5. *If $x^* \in X_0$ satisfies $f(x^*) \in \sup_{d \in R_+^m} q(d)$, then $x^* \in WE(0)$.*

COROLLARY 2.6. $[\sup_{d \in R_+^m} q(d) - V(0)] \bigcap \text{int } C = \emptyset$.

DEFINITION 2.7 (see [19]). *Let $X \subset R^n$ be a set and $f : X \to R^l$ be a vector-valued function. The set $f(X)$ is said to be externally stable if for any $x \in X$ there exists an efficient solution $x^* \in X$ of f on X such that $f(x^*) \leq_C f(x)$.*

DEFINITION 2.8. *Let $X \subset R^n$ be a set and $f : X \to R^l$ be a vector-valued function. The set $f(X)$ is said to be* inf-*externally stable if for any $x \in X$ there exists an infimum point z^* of $f(X)$ such that $z^* \leq_C f(x)$.*

Remark 2.3. The definition of external stability is given in [19], while the definition of inf-external stability is a weaker concept, which will be used later in this paper.

The following lemma on external stability can be derived from [19, Corollary 3.2.1].

LEMMA 2.9. *Let $X \subset R^n$ be a compact subset. Let $f : X \to R^l$ be a vector-valued function such that each of its component functions is l.s.c. Then $f(X)$ is externally stable.*

It is routine to prove the next lemma.

LEMMA 2.10. *Let $s : C \times R^m \to R^1$ be an increasing l.s.c. function. Let $f : X \to C$ be a vector-valued function such that each component function f_i is l.s.c. Let $g_j : X \to R^1$ $(j = 1, \ldots, m)$ be l.s.c. Then $s(f(x), g(x))$ is l.s.c. on X.*

Let $\xi(z) = \max_{1 \leq i \leq l} \{z_i\}$ $\forall z = (z_1, \ldots, z_l)$.

Clearly, ξ is an increasing, continuous, subadditive, positively homogeneous, and convex function.

DEFINITION 2.11. *Let $X \subset R^n$ be an unbounded set. A vector-valued function $f : X \to R^l$ is said to be coercive on X if*

$$\lim_{\|x\| \to +\infty, x \in X} \xi(f(x)) \to +\infty,$$

where $\|.\|$ is a norm of R^n.

The following result establishes a proper relation between (MOP) and (DMOP).

THEOREM 2.12 (strong duality). *Assume that X is closed, $f(x) \geq_C 0$ $\forall x \in X$, and f is coercive on X if X is unbounded. Then*

$$V(0) \subseteq \sup_{d \in R^m_+} q(d).$$

Proof. Let $z^* \in V(0)$. Then $\exists x^1_k \in X_0$ such that $f(x^1_k) \to z^*$ as $k \to +\infty$.

It follows that $\xi(f(x^1_k)) \to \xi(z^*)$ as $k \to +\infty$. Therefore, $\{x^1_k\}$ is a bounded sequence by the coercivity of f on X. Since X_0 is closed, there exists a subsequence $\{x^1_{k_j}\}$ such that $x^1_{k_j} \to x^*$ for some $x^* \in X_0$. Note that $f_i(x^*) \leq \liminf_{j \to +\infty} f_i(x^1_{k_j}) = (z^*)_i, i = 1, \ldots, l$, where $(z^*)_i$ denotes the ith component of z^*. We have $f(x^*) \leq_C z^*$. This with $z^* \in V(0)$ implies that $f(x^*) = z^*$. Hence $x^* \in E(0)$. Since f is coercive on X, we deduce that $\exists N > 0$ such that

$$\xi(f(x)) \geq \xi(f(x^*)) + 1 \forall x \in X_1 = \{x \in X : \|x\| > N\}. \tag{2}$$

We claim that

$$f(x) \not\leq_{C \setminus \{0\}} f(x^*) \quad \forall x \in X_1. \tag{3}$$

Otherwise, $\xi(f(x)) \leq \xi(f(x^*))$, contradicting (2).

Let $d = ke, k = 1, 2, \ldots$. Since $X_2 = \{x \in X : \|x\| \leq N\}$ is a nonempty compact set and $x^* \in X_2$, by Lemmas 2.9 and 2.10, we obtain a sequence $\{x^2_k\} \subseteq X_2$ such that each x^2_k is an efficient solution to the problem: $\min_{x \in X_2} p(f(x), kg(x))$ and

$$p(f(x^2_k), kg(x^2_k)) \leq_C p(f(x^*), kg(x^*)) = f(x^*). \tag{4}$$

We show that this fact combined with (3) yields that $p(F(x_k^2,d)) \in q(k,\dots,k) = \inf_{x\in X} p(F(x,d))$.

(i) It is obvious that if $x \in X_2$, $p(F(x_k^2,d)) \not\geq_{C\setminus\{0\}} p(F(x,d))$.

(ii) Suppose that $\exists\, \overline{x} \in X_1$ such that

$$p(F(x_k^2,d)) \geq_{C\setminus\{0\}} p(F(\overline{x},d)). \tag{5}$$

Note that

$$p(F(x_k^2,d)) \leq_C f(x^*)$$

and

$$f(x^*) \not\geq_{C\setminus\{0\}} f(\overline{x}).$$

Then

$$p(F(x_k^2,d)) \not\geq_{C\setminus\{0\}} f(\overline{x}). \tag{6}$$

By (5) and (6),

$$p(F(\overline{x},d)) \not\geq_{C\setminus\{0\}} f(\overline{x}),$$

a contradiction with the property (A).

It follows from $\{x_k^2\} \subset X_2$ that there exists a subsequence $\{x_{k_j}^2\}$ such that $x_{k_j}^2 \to x_0 \in X_2$.

Let us show that $x_0 \in X_0$. If not, $d(x_0, X_0) \geq \delta_0$ for some $\delta_0 > 0$. It follows that $d(x_{k_j}^2, X_0) \geq \delta_0/2$ when j is sufficiently large.

Let $X_3 = \{x \in X_2 : d(x, X_0) \geq \delta_0/2\}$ and $\bar{g}(x) = \max_{1\leq j\leq m} g_j(x)$. Since $\bar{g}(x) > 0\ \forall x \in X_3$, X_3 is compact, and $\bar{g}$ is l.s.c, we deduce that $\min_{x\in X_3} \bar{g}(x) = m_0 > 0$.

By property (A) of the function p, there exist positive numbers $a_i (i = 1,\dots,m)$ such that

$$p(f(x_{k_j}^2), k_j g(x_{k_j}^2)) \geq_C \left(m_0 k_j \min_{1\leq i\leq m} a_i\right) e$$

when j is sufficiently large, which contradicts (4). So $x_0 \in X_0$.

Applying property (A) and (4), we have

$$f(x_{k_j}^2) \leq_C p(f(x_{k_j}^2), k_j g(x_{k_j}^2)) \leq_C f(x^*).$$

Thus,

$$f_i(x_{k_j}^2) \leq p_i(f(x_{k_j}^2), k_j g(x_{k_j}^2)) \leq f_i(x^*), \quad i = 1,\dots,l. \tag{7}$$

Applying the lower limit to (7) by letting $j \to \infty$, we conclude that $f_i(x_0) \leq f_i(x^*), i = 1,\dots,l$, which implies that

$$f(x_0) = f(x^*) \tag{8}$$

since $x^* \in E(0)$.

Equation (8) combined with (7) as well as $x_{k_j}^2 \to x_0$ yields that

$$p(f(x_{k_j}^2), k_j g(x_{k_j}^2)) \to f(x^*) \text{ as } j \to +\infty.$$

Finally, it follows directly from Proposition 2.4 that

$$(q(d) - f(x^*)) \bigcap (C \backslash \{0\}) = \emptyset \quad \forall d \in R_+^m.$$

The proof is complete. □

Remark 2.4. 1. When $l = 1$, this theorem improves Theorem 3.1 in [17] by relaxing the assumption of continuity of f and g_j as well as p to lower semicontinuity and dropping the assumption that X_0 is compact.

2. It is evident from the proof of Theorem 2.12 that to solve (MOP) we can solve a series of unconstrained multiobjective programming problems to approach the efficient points of (MOP).

3. The condition that f is coercive on X is important to guarantee the validity of Theorem 2.12. Otherwise, it may fail even if X_0 is compact. Example 2.3 shows this case.

Example 2.3. Let $l = 1, X = [0, +\infty), f(x) = 1/(x+1)\ \forall x \in X$, $g_1(x) = x - 1$ if $0 \le x \le 1; g_1(x) = 1/\sqrt{x} - 1/x$ if $1 < x < +\infty$, $p(y_1, y_2) = \max\{y_1, y_2\}\ \forall y_1, y_2 \in R^1$.

Consider the problem

$$\inf_{x \in X} f(x) \text{ s.t. } g_1(x) \le 0.$$

It is easy to see that $X_0 = [0, 1]$ (which is compact) and $V(0) = \{1/2\}$.

$p(f(x), dg_1(x)) = \max\{f(x), dg_1(x)\} = \max\{1/(x+1), d(1/\sqrt{x} - 1/x)\}\ \forall x \in X \backslash X_0, d \ge 0$.

Clearly, $q(d) = 0\ \forall d \ge 0$. It follows that $\sup_{d \ge 0} q(d) = \{0\}$. Hence $V(0) \subseteq \sup_{d \ge 0} q(d)$ does not hold.

Despite Example 2.3, in actually designing an algorithm based on Theorem 2.12, if X_0 is compact, we can replace $f(x)$ with $f(x) + l(x)e$, where $l : X \to R_+^1$ is an l.s.c. function which satisfies the following condition: there exists a compact set X' such that $X_0 \subseteq X' \subseteq X$ with $l(x) = 0$ if $x \in X'$ and $l(x) \to +\infty$ if $x \in X$ and $\|x\| \to +\infty$. A simple example of such an l is $l(x) = d(x, X_0)\ \forall x \in X$. Thus Theorem 2.12 can be applied to the objective function $f(x) + l(x)e$, which has the same set of (weakly) efficient solutions and the same set of (weakly) efficient points as $f(x)$ on X_0.

Finally, we observe the following two points:

(i) for $z^* \in V(0)$ there may not exist $d^* \in R_+^m$ such that $z^* \in q(d^*)$ even if all the conditions in Theorem 2.12 hold;

(ii) for the conventional Lagrangian, Theorem 2.12 does not, in general, hold.

Counterexamples are given for these two cases in Examples 2.4 and 2.5, respectively.

Example 2.4. Let $l = 1, X = [1/2, +\infty)$, and $f(x) = 1/x$ if $x \in [1/2, 1]$; $f(x) = 2 - x$ if $x \in [1, 2]$; $f(x) = x - 2$ if $x \in (2, +\infty)$. Let $g_1(x) = x - 1$.

Consider the problem

$$\inf_{x \in X} f(x) \text{ s.t. } g_1(x) \le 0.$$

Let $L(x, d) = \max\{f(x), dg_1(x)\},\ d \ge 0,\ x \in X$. Then it is not difficult to derive the following fact: $q(d) = d/(1+d)\ \forall d \ge 0$. Clearly, $q(d) < 1 = \inf_{x \in X_0} f(x)\ \forall d > 0$.

Example 2.5. Let $l = 1$, $X = [0, +\infty)$, $f(x) = x$, $g(x) = x - x^2$. Consider the problem

$$\inf_{x \in X} f(x)$$
$$\text{s.t. } g_1(x) \le 0.$$

It is clear that all the conditions of Theorem 2.12 hold. Let us look at the conventional Lagrangian for this problem: $l(x,\lambda) = f(x)+\lambda g_1(x) = x+\lambda(x-x^2)\ \forall x \in X,\ \lambda \geq 0$. It is easy to check that $\inf_{x\in X} l(x,\lambda) = -\infty\ \forall \lambda > 0$ and $\inf_{x\in X} l(x,0) = 0$. Thus, $\sup_{\lambda\geq 0} \inf_{x\in X} l(x,\lambda) = 0$. However, the optimal value of the original constrained problem is 1.

Based on some conditions on the constraint functions, we also have the following result.

THEOREM 2.13. *Let* $\bar{g}(x) = \max_{1\leq j\leq m} g_j(x)$. *Assume that there exist* $N > 0$ *and* $m_1 > 0$ *such that*

$$\bar{g}(x) \geq m_1 \quad \forall x \in X \text{ with } \|x\| > N. \tag{9}$$

Then $V(0) \subseteq \sup_{d\in R^m_+} q(d)$.

Proof. It follows from (9) that X_0 is a nonempty compact set. For any $z^* = f(x^*) \in V(0)$, by Proposition 2.4 we have that

$$(q(d) - f(x^*)) \bigcap (C\backslash\{0\}) = \emptyset \quad \forall d \in R^m_+.$$

Furthermore, whenever $x \in X$ with $\|x\| > N$,

$$p(f(x), kg(x)) \geq_C \left(km_1 \min_{1\leq i\leq m}\{a_i\}\right) e \geq_{\text{int}C} f(x^*) + e$$

when k is sufficiently large. Consequently, when k is sufficiently large, the set

$$\{x \in X : p(f(x), kg(x)) \leq_C f(x^*)\} (\subseteq \{x \in X : \|x\| \leq N\})$$

is a nonempty compact set. Therefore, when k is sufficiently large, $\exists x_k \in X$ with $\|x_k\| \leq N$ such that x_k is an efficient solution to the problem

$$\min_{x\in X} p(f(x), kg(x))$$

with

$$f(x_k) \leq_C p(f(x_k), kg(x_k)) \leq_C f(x^*). \tag{10}$$

Since $\|x_k\| \leq N$ for k sufficiently large, it follows that there exists a subsequence $\{x_{k_j}\}$ converging to $x' \in X$. We can show as in the proof of Theorem 2.12 that $x' \in X_0$. This fact combined with (10) yields that $f(x') \leq_C f(x^*)$. Therefore, $f(x') = f(x^*)$ since $x^* \in E(0)$. Hence, $p(f(x_{k_j}), k_j g(x_{k_j})) \to f(x^*)$. So $f(x^*) \in \sup_{d\in R^m_+} q(d)$ and the proof is complete. □

The following proposition further clarifies the relation between (MOP) and (DMOP).

PROPOSITION 2.14. *Let* $d^k \in R^m_+\ \forall k$ *and* $d^k \to +\infty$ *as* $k \to \infty$ *(i.e.,* $d^k_i \to +\infty\ \forall i$ *as* $k \to +\infty$*). Suppose that each* x^k *is a weakly efficient solution to* $\inf_{x\in X} L(x, d^k)$. *Then any limiting point of* $\{x^k\}$ *is a weakly efficient solution to* (MOP).

Proof. Without loss of generality, suppose that $x^k \to x^*$. We can show by contradiction that $x^* \in X_0$. In fact, if $d(x^*, X_0) \geq \delta_0$ for some $\delta_0 > 0$, then $d(x^k, X_0) \geq \delta_0/2$ when k is sufficiently large. Since $x^k \to x^*$, we deduce that $\|x^k - x^*\| \leq 1$ when k is sufficiently large.

Let $X_4 = \{x \in X : d(x, X_0) \geq \delta_0/2, \|x - x^*\| \leq 1\}$. Then $x^k \in X_4$ when k is sufficiently large. Let $\bar{g}(x) = \max_{1\leq i\leq m} g_i(x)$. Then $\bar{g}(x^k) \geq \min_{x\in X_3} \bar{g}(x) =$

$m_1 > 0$ when k is sufficiently large. So

$$\begin{aligned} p(f(x^k), d^k * g(x^k)) &\geq_C \bar{g}(x^k) \left(\min_{1\leq i\leq m} a_i \min_{1\leq i\leq m} d_i^k \right) e \\ &\geq_C \left(m_1 \min_{1\leq i\leq m} a_i \min_{1\leq i\leq m} d_i^k \right) e \\ &\geq_{\text{int}C} f(x_0) \end{aligned} \tag{11}$$

for any fixed $x_0 \in X_0$ and k large enough. Moreover, by Lemma 2.3,

$$f(x_0) = p(f(x_0), d^k * g(x_0)). \tag{12}$$

The combination of (11) and (12) contradicts the fact that x^k is a weakly efficient solution to $\min_{x\in X} p(f(x), d^k * g(x))$. Therefore, $x^* \in X_0$.

Now we show that $x^* \in \mathrm{W}(0)$. Otherwise, $\exists x'' \in X_0$ such that $f(x'') \leq_{\text{int}C} f(x^*)$. Therefore,

$$f(x'') \leq_{\text{int}C} f(x^k) \tag{13}$$

when k is sufficiently large since each component function of f is l.s.c.

Note that

$$f(x'') = p(f(x''), d^k * g(x''))$$

and

$$p(f(x_k), d^k * g(x^k)) \geq_C f(x^k);$$

it follows from (13) that

$$p(f(x_0), d^k * g(x_0)) \leq_{\text{int}C} p(f(x^k), d^k * g(x^k))$$

when k is sufficiently large. Namely, x^k is not a weakly efficient solution to

$$\min_{x\in X} \; p(f(x), d^k * g(x))$$

when k is sufficiently large, which cannot be true. The proof is complete. □

Remark 2.5. All the results in this section also hold for the case when p is defined on the domain $C \times R_+^m$, $F_+(x,d) = (f(x), d * g^+(x))$, $g^+(x) = (g_1^+(x), \ldots, g_m^+(x))$, and

$$L(x,d) = p(F_+(x,d)). \tag{14}$$

3. Exact penalization. Consider the following nonlinear penalty function:

$$L_\gamma(x,d) = p_\gamma(f(x), d * g^+(x)) = \sum_{i=1}^{l} \left[f_i^\gamma(x) + \sum_{j=1}^{m} d_j^\gamma g_j^{+\gamma}(x) \right]^{1/\gamma} e_i,$$

where $0 < \gamma < +\infty$.

Let $u = (u_1, \ldots, u_m) \in R^m$. We associate (MOP) with a perturbed problem:

$$(\mathrm{MOP}_u) \quad \begin{array}{ll} \inf\limits_{x\in X} & f(x) \\ \text{s.t.} & g_j(x) \leq u_j, \quad j = 1, \ldots, m, \end{array}$$

where X, f, g_j are defined as in (MOP).

Let

$$X(u) = \{x \in X : g_j(x) \leq u_j, j = 1, \ldots, m\}.$$

We will denote by $E(u), W(u)$, and $V(u)$ the sets of efficient solutions, efficient points, and infimum points of (MOP_u), respectively.

We need the following lemma.

LEMMA 3.1. *For any $x_0 \in X(u)$, there exists $z^* \in V(u)$ such that $z^* \leq_C f(x_0)$.*

Proof. Let $Z = f(X(u))$, $Z_1 = \{z \in \mathrm{cl}(Z) : z \leq_C f(x_0)\}$. Clearly, Z_1 is nonempty and closed and $z \geq_C 0\ \forall z \in Z$. Since $\leq_C$ is a partial order in Z_1, by the well-known Hausdorff maximality principle (see, e.g., [11]), there exists a totally ordered subset Z_2 of Z_1, which is maximal with respect to the set inclusion. Let $z_i^* = \inf\{z_i : (z_1, \ldots, z_i, \ldots, z_l) \in Z_2\}, i = 1, \ldots, l$, and $z^* = (z_1^*, \ldots, z_l^*)$. It is obvious that $0 \leq_C z^* \leq_C f(x_0)$. Furthermore, by the definition of z^* and the fact that Z_2 is totally ordered, we deduce that $z^* \in \mathrm{cl}(Z_2) \subset Z_1$. We assert that $z^* \in Z_2$. Otherwise, as $Z_2 \bigcup \{z^*\}$ is also a totally ordered subset of Z_1 and $Z_2 \subset Z_2 \bigcup \{z^*\}$, this contradicts the maximality of Z_2 with respect to the set inclusion. Finally, we show that $z^* \in V(u)$. We need to prove only that $z \not\leq_{C\setminus\{0\}} z^*\ \forall z \in \mathrm{cl}(Z)$. Let $z \in \mathrm{cl}(Z)$. If $z \not\leq_C f(x_0)$, it can be shown by contradiction that $z \not\leq_{C\setminus\{0\}} z^*$. If $z \leq_C f(x_0)$ and

$$z \leq_{C\setminus\{0\}} z^*, \tag{15}$$

then, by the maximality of Z_2, we have $z \in Z_2$, and thus $z^* \leq_C z$ by the definition of z^*. This contradicts (15). The proof is complete. □

DEFINITION 3.2. *We say that* (MOP) *is γ-rank uniformly weakly stable if there exist $\delta > 0$ and $M > 0$ such that*

$$\left[\frac{V(u) - V(0)}{\|u\|_\gamma^\gamma} + Me\right] \bigcap (-\mathrm{int}C) = \emptyset \tag{16}$$

for any $u \in R_+^m$ with $0 < \|u\|_\gamma \leq \delta$.

Remark 3.1. 1. It is not hard to show that the restriction $u \in R_+^m$ in the definition of the γ-rank uniform weak stability can be replaced by $u \in R^m$. This is also true for the γ-rank weak stability and γ-rank calmness in Definitions 3.4 and 3.7, respectively.

2. If $l = 1$ and $\gamma = 1$, then Definition 3.2 is equivalent to the stability of scalar optimization problems studied by Rosenberg [15]. (Any equality constraint $h(x) = 0$ with h being continuous can be equivalently written as the following inequality constraint: $|h(x)| \leq 0$.) In the definition of γ-rank uniform weak stability of (MOP), the term "uniform" shows the difference from the usual stability in which $V(0)$ in (16) is replaced by a specific point of $V(0)$ and the fact that different points of $V(0)$ may have different M's in (16), and the term "weak" is used in contrast to the stability of (MOP) defined in [19, Definition 6.13, p. 182].

3. Let $0 < \gamma_1 < \gamma_2$. It is not hard to see that if (MOP) is γ_2-rank uniformly weakly stable, then it is also γ_1-rank uniformly weakly stable.

THEOREM 3.3. *If* (MOP) *is γ-rank uniformly weakly stable, then $\exists d^* \in R_+^m$ such that when $d - d^* \in R_+^m$,*

$$V(0) \subseteq q_\gamma(d), \tag{17}$$

where $q_\gamma(d) = \inf_{x\in X} L_\gamma(x, d)$. The converse is also true.

Proof. We begin by proving the first half of this theorem.

If $V(0) = \emptyset$, then the conclusion holds automatically. Now we assume that $V(0) \neq \emptyset$.

Let $\eta(z) = \min_{1\leq i\leq l} z_i \ \forall z = (z_1, \ldots, z_l) \in R^l$. We show by contradiction that $\eta(V(0)) = \{\eta(z) : z \in V(0)\}$ is bounded from above by some $M' > 0$. Otherwise, $\exists z_k \in V(0)$ such that $z_k \to +\infty$. Since $V(0) \neq \emptyset$, it follows that for any $\delta > 0$, $X(u_\delta) \supset X(0) = X_0 \neq \emptyset$, where $u_\delta = (0, 0, \ldots, 0, \delta) \in R^m_+$. Suppose that $x_0 \in X_0 \subset X(u_\delta)$. Then by Lemma 3.1 $\exists z_\delta \in V(u_\delta)$ such that

$$z_\delta \leq_C f(x_0).$$

Hence,

$$(z_\delta - z_k)/\|u_\delta\|^\gamma_\gamma \leq_C (f(x_0) - z_k)/\|u_\delta\|^\gamma_\gamma \to -\infty \text{ as } n \to \infty,$$

which contradicts (16) because $\delta > 0$ can be arbitrarily small.

Suppose that $\exists d_k = (d_{k,1}, \ldots, d_{k,m}) \to +\infty$ and $z_k \in V(0)$ such that $z_k \notin \inf_{x\in X} L_\gamma(x, d_k)$.

By $z_k \in V(0)$, it follows that $\exists x^j_k$ such that $g(x^j_k) \leq 0$ and $f(x^j_k) \to z_k$ as $j \to \infty$.

It follows from $z_k \notin \inf_{x\in X} L_\gamma(x, d_k)$ that $\exists x'_k \in X$ such that

$$L_\gamma(x'_k, d_k) \leq_{C\setminus\{0\}} z_k.$$

That is,

$$\sum_{i=1}^{l} \left[f^\gamma_i(x'_k) + \sum_{j=1}^{m} (d^\gamma_{k,j} g^{+\gamma}_j(x'_k)) \right]^{1/\gamma} e_i \leq_{C\setminus\{0\}} z_k. \tag{18}$$

Using (18), we deduce that $\max_{1\leq j\leq m} g_j(x'_k) > 0$ since $z_k \in V(0)$.

(18) also implies that

$$\sum_{j=1}^{m} d^\gamma_{k,j} g^{+\gamma}_j(x'_k) \leq (z_k)^\gamma_i - f^\gamma_i(x'_k) \leq (z_k)^\gamma_i, \quad i = 1, \ldots, l, \tag{19}$$

where $(z_k)_i$ denotes the ith component of vector z_k.

That is, $[\sum_{j=1}^m d^\gamma_{k,j} g^{+\gamma}_j(x'_k)]^{1/\gamma} \leq \eta(z_k) \leq M'$.

It follows that $g^+_j(x'_k) \to 0$ $(j = 1, \ldots, m)$ as $n \to +\infty$.

Now let $u_{k,j} = g^+_j(x'_k)$ and $u_k = (u_{k,1}, \ldots, u_{k,m})$. Clearly, $\|u_k\|_\gamma > 0$ and $\|u_k\|_\gamma \to 0$. It follows from (19) that $\|u_k\|^\gamma_\gamma \min_{1\leq j\leq m} d^\gamma_{k,j} \leq (z_k)^\gamma_i - f^\gamma_i(x'_k)$. By Lemma 3.1, we deduce that $\exists v_k \in V(u_k)$ such that $v_k \leq_C f(x'_k)$. By the mean-value theorem, we have $(z_k)^\gamma_i - (v_k)^\gamma_i = k(s_k)^{\gamma-1}_i((z_k)_i - (v_k)_i)$, where $(s_k)_i \in ((v_k)_i, (z_k)_i)$. Therefore, it follows from (19) that

$$\|u_k\|^\gamma_\gamma \min_{1\leq j\leq m} d^\gamma_{k,j} \leq k(s_k)^{\gamma-1}_i((z_k)_i - (v_k)_i) \leq \gamma(v_k)^{\gamma-1}_i((z_k)_i - (v_k)_i) \text{ if } \gamma \leq 1; \tag{20}$$

$$\|u_k\|^\gamma_\gamma \min_{1\leq j\leq m} d^\gamma_{k,j} \leq \gamma {M'}^{\gamma-1}((z_k)_i - (v_k)_i) \text{ if } \gamma > 1. \tag{21}$$

Since $\inf_{x\in X} f_i(x) > 0 \ \forall i$, it follows that

$$\min_{1\leq i\leq m} (v_k)_i \geq m_2 > 0. \tag{22}$$

Let $M'' = \max\{M'^{\gamma-1}, m_2^{\gamma-1}\}$. The combination of (20), (21), and (22) yields that

$$\|u_k\|_\gamma^\gamma \min_{1\le j\le m} d_{k,j}^\gamma \le \gamma M''((z_k)_i - (v_k)_i),$$

i.e.,

$$\frac{(v_k)_i - (z_k)_i}{\|u_k\|_\gamma^\gamma} \le -\frac{\min_{1\le j\le m} d_{k,j}^\gamma}{\gamma M''},$$

which contradicts (16). Thus (17) holds.

Now we prove the second half of the theorem by contradiction.

Suppose that $\exists u_k = (u_{k,1}, \dots, u_{k,m}) \in R_+^m$ with $u_k \to 0^+$ and $z_k \in V(u_k), v_k \in V(0)$ such that

$$(z_k - v_k)/\|u_k\|_\gamma^\gamma \to -\infty \text{ as } k \to +\infty,$$

where the virtual element $-\infty$ is such that for any $\alpha \in R_+^1, -\infty \le_{\text{int}C} -\alpha e$. Then $\exists x_k \in X$ with $g_j(x_k) \le u_{k,j} \forall j$ such that

$$(f(x_k) - v_k)/\|u_k\|_\gamma^\gamma \to -\infty \text{ as } k \to +\infty. \tag{23}$$

By the assumption of the theorem, $\exists d^* = (d_1^*, \dots, d_m^*) \in R_+^m$ such that when $d - d^* \in R_+^m$, $v_k \in \inf_{x\in X} L_\gamma(x, d)$. Therefore,

$$L_\gamma(x_k, d^*) \not\le_{C\setminus\{0\}} v_k. \tag{24}$$

We assume that $i^* \in \{1, \dots, l\}$ is such that

$$\left[f_{i^*}^\gamma(x_k) + \sum_{j=1}^m d_j^{*\gamma} g_j^{+\gamma}(x_k)\right]^{1/\gamma} \ge (v_k)_{i^*}.$$

Namely,

$$f_{i^*}^\gamma(x_k) - (v_k)_{i^*}^\gamma \ge -\sum_{j=1}^m d_j^{*\gamma} g_j^{+\gamma}(x_k). \tag{25}$$

It follows from (23) and (24) that $\max_{1\le j\le m} g_j(x_k) > 0$. So from (25) we deduce that

$$f_{i^*}^\gamma(x_k) - (v_k)_{i^*}^\gamma \ge -\max_{1\le j\le m} d_j^{*\gamma} \|u_k\|_\gamma^\gamma.$$

That is,

$$[(v_k)_{i^*}^\gamma - f_{i^*}^k(x_k)]/\|u_k\|_\gamma^\gamma \le \max_{1\le j\le m} d_j^{*\gamma}. \tag{26}$$

Since

$$(v_k)_{i^*}^\gamma - f_{i^*}^\gamma(x_k) = \gamma s_k^{\gamma-1}((v_k)_{i^*} - f_{i^*}(x_k)), \quad s_k \in (f_{i^*}(x_k), (v_k)_i),$$

it follows from the assumption on f that $\exists a > 0$ such that

$$(v_k)_{i^*}^\gamma - f_{i^*}^k(x_k) \ge \gamma a((v_k)_{i^*} - f_{i^*}(x_k)). \tag{27}$$

Equations (26) and (27) yield that

$$[f_{i^*}(x_k) - (v_k)_{i^*}]/\|u_k\|_\gamma^\gamma \geq -\max_{1\leq j\leq m} d_j^{*\gamma}/(ka),$$

which contradicts (23). The proof is complete. □

Remark 3.2. When $l = 1, m = 1$, Theorem 3.3 reduces to Theorem 7.2 in [18].

DEFINITION 3.4. (i) *Let $z^* \in V(0)$. The problem* (MOP) *is said to be γ-rank weakly stable at z^* if there exist positive real numbers δ_{z^*} and M_{z^*} such that*

$$\left[\frac{V(u) - z^*}{\|u\|_\gamma^\gamma} + M_{z^*}e\right] \bigcap (-\text{int}C) = \emptyset$$

for any $u \in R_+^m$ with $0 < \|u\|_\gamma \leq \delta_{z^*}$.

(ii) *The problem* (MOP) *is said to be γ-rank weakly stable if it is γ-rank weakly stable at every $z^* \in V(0)$.*

Remark 3.3. 1. It is clear that if (MOP) is γ-rank uniformly weakly stable, then (MOP) is γ-rank weakly semistable.

2. It is not hard to check that if $f(X(u))$ is externally stable for any $u \in R_+^m$, then the stability of (MOP) defined in [19, Definition 6.1.3, p. 182] implies the 1-rank weak stability of (MOP).

The proof of the next theorem is similar to that of Theorem 3.3 and is thus omitted.

THEOREM 3.5. *Let $z^* \in V(0)$. Then* (MOP) *is γ-rank weakly stable at z^* if and only if there exists a $d^* \in R_+^m$ such that $z^* \in q_\gamma(d)$ whenever $d - d^* \in R_+^m$.*

COROLLARY 3.6. (MOP) *is γ-rank weakly stable if and only if for every z^* there exists a $d^* \in R_+^m$ such that $z^* \in q_\gamma(d)$ whenever $d - d^* \in R_+^m$.*

Remark 3.4. The following simple example shows that (MOP) is 1-rank weakly stable but not 1-rank uniformly weakly stable.

Example 3.1. Let $n = 1$, $l = 2$, $X = R^1$, and $m = 1$. Let $f(x) = (\exp(-x^{1/2}), \exp(-x^{1/2}))$ if $x > 0$; $f(x) = (\exp(x), \exp(-x))$ if $x \leq 0$. Let $g(x) = x\ \forall x \in R^1$. It is easy to check that $V(0) = \{(\exp(x), \exp(-x)) : x \leq 0\}$ and

$$V(u) = \{(\exp(-u^{1/2}),\ \exp(-u^{1/2}))\} \bigcup \{(\exp(x), \exp(-x)) : x < -u^{1/2}\}\ \forall u > 0.$$

It is elementary to prove that (MOP) is 1-rank weakly stable but not 1-rank uniformly weakly stable. By Corollary 3.6, we know that for every $z^* \in V(0)$ there exists $d^* \geq 0$ such that $z^* \in \inf_{x\in R^1}(f(x) + dg^+(x)e)$, where $d \geq d^*$. On the other hand, by Theorem 3.3, we deduce that there exists no $d^* \geq 0$ such that $V(0) \subseteq \inf_{x\in R^1}(f(x) + dg^+(x)e)$, whenever $d \geq d^*$.

DEFINITION 3.7. *Let $x^* \in LWE(0)$. We say that* (MOP) *is γ-rank calm at x^* if there exists $M > 0$ such that for any $u_k = (u_{k,1}, \ldots, u_{k,m}) \in R_+^m$ with $\|u_k\|_\gamma \to 0^+$ (namely, $\|u_k\|_\gamma > 0$ and $\|u_k\|_\gamma \to 0$), for any x_k satisfying $g_j(x_k) \leq u_{k,j}, j = 1, \ldots, m$ and $x_k \to x^*$, there holds*

$$\frac{f(x_k) - f(x^*)}{\|u_k\|_\gamma^\gamma} + Me \notin -C \quad \forall n.$$

Remark 3.5. 1. If $l = 1, \gamma = 1$, then this definition is equivalent to the calmness at a point of a scalar optimization problem (see, e.g., [15, 4]). If $l > 1, \gamma = 1$, then this definition is equivalent to the weak calmness at a point of the multiobjective optimization problem (MOP) defined in [16].

2. If $0 < \gamma_1 < \gamma_2$, then (MOP) is γ_2-rank calm at a point x^*, which implies that it is γ_1-rank calm at x^*.

The following local exact penalization result can also be similarly proved as Theorem 3.3.

THEOREM 3.8. *Let $0 < \gamma < +\infty$. The following statements hold.*

(i) *Assume that x^* is a locally weak efficient solution to* (MOP) *and* (MOP) *is γ-rank calm at x^*. Then there exist $\delta > 0$ and $d^* \in R^m_+$ such that x^* is also a weak efficient solution to the problem* $\min_{x\in X_\delta} L_\gamma(x, d)$, *for any d satisfying $d - d^* \in R^m_+$, where $X_\delta = \{x \in X : \|x - x^*\| \le \delta\}$.*

(ii) *If $x^* \in X_0$ and there exist $d^* \in R^m_+$ and $\delta > 0$ such that x^* is a locally weak efficient solution to the problem* $\min_{x\in X} L_\gamma(x, d^*)$, *then $x^* \in LWE(0)$ and* (MOP) *is γ-rank calm at x^*.*

The next theorem uses a well-known condition in the study of sensitivity of a constrained optimization problem (see, e.g., [12]), i.e., the compactness of the feasible set with a small perturbation. Under this condition, the set of efficient points of (MOP) and that of $L_\gamma(\cdot, d)$ are nonempty. The conclusion follows directly from Theorem 3.3.

THEOREM 3.9. *Assume that there exists $u^0 = (u^0_1, \dots, u^0_m) \in \operatorname{int} R^m_+$ with $\|u^0\| > 0$ sufficiently small such that $X_5 = \{x \in X : g_j(x) \le u^0_j \forall j\}$ is compact. If* (MOP) *is γ-rank uniformly weakly stable, then $\exists d^* \in R^m_+$ such that when $d - d^* \in R^m_+$,*

$$W(0) = f(E(0)) \subseteq \bar{q}_\gamma(d),$$

where $\bar{q}_\gamma(d)$ is the set of efficient points of $L_\gamma(\cdot, d)$ over X. The converse is also true.

The following theorem establishes a further relationship between the solutions of (MOP) and that of the penalty problems based on L_γ.

THEOREM 3.10. *Assume that $X_0 \neq \emptyset$ and $\exists d^* = (d^*_1, d^*_2, \dots, d^*_m) \in R^m_+$ such that for all d satisfying $d - d^* \in R^m_+$, $x^* \in X$ is an efficient solution of the problem* $\min_{x\in X} L_\gamma(x, d)$; *then x^* is an efficient solution of* (MOP).

Proof. Let x^* be an efficient solution of $\min_{x\in X} L_\gamma(x, d)$ for any d satisfying $d - d^* \in R^m_+$. Then we have

$$L_\gamma(x, d) - L_\gamma(x^*, d) \not\le_{C\setminus\{0\}} 0 \quad \forall x \in X, d \text{ satisfying } d - d^* \in R^m_+.$$

For any $x_0 \in X_0$, we have $L_\gamma(x_0, d) = f(x_0)\ \forall d \in R^m_+$ by Lemma 2.3. Thus,

$$f(x_0) - \sum_{i=1}^{l} \left[f^\gamma_i(x^*) + \sum_{j=1}^{m} d^\gamma_j g^{+\gamma}_j(x^*) \right]^{1/\gamma} e_i \not\le_{C\setminus\{0\}} 0\ \forall x_0 \in X_0, d \text{ satisfying } d - d^* \in R^m_+. \tag{28}$$

We claim that $g^+_j(x^*) = 0\ \forall j$ (i.e., $x^* \in X_0$). Otherwise, $\sum_{j=1}^m g^{+\gamma}_j(x^*) > 0$.

It follows from (28) that there exists $i^* \in \{1, \dots, l\}$ such that

$$f^\gamma_{i^*}(x_0) - f^\gamma_{i^*}(x^*) \ge \sum_{j=1}^{m} d^\gamma_j g^{+\gamma}_j(x^*) \ge \left(\min_{1\le j\le m} d^\gamma_j \right) \sum_{j=1}^{m} g^{+\gamma}_j(x^*).$$

Hence,

$$\max_{1\le i\le l} \{f^\gamma_i(x_0) - f^\gamma_i(x^*)\} \ge \sum_{j=1}^{m} d^\gamma_j g^{+\gamma}_j(x^*) \ge \left(\min_{1\le j\le m} d^\gamma_j \right) \sum_{j=1}^{m} g^{+\gamma}_j(x^*),$$

which is impossible if we let $d_j \to +\infty\ \forall j$. Therefore, $x^* \in X_0$. It follows directly from Lemma 2.3 and (28) that $x^* \in E(0)$, and the proof is complete. □

What follows is a characterization of the γ-rank weak stability of (MOP) at a point $z^* \in V(0)$ in terms of the γ-rank stability of a scalar optimization problem (see below).

Let $z^* \in V(0)$. Recall $\xi(z) = \max_{1\le i\le l}\{z_i\}\ \forall z = (z_1, \dots, z_l)$. Consider the following scalar optimization problem:

$$(\mathrm{P}(z^*)) \quad \begin{array}{ll} \inf\limits_{x\in X} & \xi(f(x) - z^*) \\ \text{s.t.} & g_j(x) \le 0, \quad j = 1, \dots, m, \end{array}$$

and its perturbed problem,

$$(\mathrm{P}_u(z^*)) \quad \begin{array}{ll} \inf\limits_{x\in X} & \xi(f(x) - z^*) \\ \text{s.t.} & g_j(x) \le u_j, \quad j = 1, \dots, m \end{array}$$

where $u = (u_1, \dots, u_m) \in R^m_+$ is such that $\|u\|_\gamma > 0$ is sufficiently small.

Clearly, the optimal value of $(\mathrm{P}(z^*))$ is 0. We denote by $\pi(u)$ the optimal value of $(\mathrm{P}_u(z^*))$. $(\mathrm{P}(z^*))$ is said to be *γ-rank stable* if there exist positive numbers δ and M such that

$$\frac{\pi(u)}{\|u\|^\gamma_\gamma} \ge -M$$

for any $u \in R^m_+$ with $0 < \|u\|_\gamma \le \delta$.

Note that this notion of γ-rank stability of $(\mathrm{P}(z^*))$ is equivalent to the *stability* defined in [15] if $\gamma = 1$.

The following conclusion can be straightforwardly proved.

THEOREM 3.11. *Let $z^* \in V(0)$. Then* (MOP) *is γ-rank weakly stable at z^* if and only if $(P(z^*))$ is γ-rank stable.*

COROLLARY 3.12. (MOP) *is γ-rank weakly stable if and only if for any $z^* \in V(0)$, $(P(z^*))$ is γ-rank stable.*

Remark 3.6. As noted in [4, p. 238], for a scalar optimization problem, any constraint qualification (such as the Slater or Mangasarian–Fromowitz condition) which rules out abnormal Lagrangian multipliers at every optimum also guarantees a stronger version of stability of the optimization problem; that is, the optimal value function of $(\mathrm{P}_u(z^*))$ is locally Lipschitz at the origin of R^m.

In the following, we provide some criteria for the γ-rank calmness of (MOP) at a point.

Let $x^* \in LWE(0)$. Let $u \in R^m_+\backslash\{0\}$. We associate (MOP) with the following scalar optimization problem (P') and its perturbed problem (P'_u):

$$(P') \quad \begin{array}{ll} \inf\limits_{x\in X} & \xi(f(x) - f(x^*)) \\ \text{s.t.} & g_j(x) \le 0, \quad j = 1, \dots, m, \end{array}$$

$$(P'_u) \quad \begin{array}{ll} \inf\limits_{x\in X} & \xi(f(x) - f(x^*)) \\ \text{s.t.} & g_j(x) \le u_j, \quad j = 1, \dots, m. \end{array}$$

It is easy to see that x^* is also a local minimum to (P').

(P') is said to be *γ-rank calm* at x^* if there exists $M > 0$ such that for any $u_k = (u_{k,1}, \ldots, u_{k,m}) \in R^m_+$ with $\|u_k\|_\gamma \to 0^+$, for any $x_k \to x^*$ with $g_j(x_k) \le u_{k,j}, \forall j$, we have

$$\xi(f(x_k) - f(x^*))/\|u_k\|^\gamma_\gamma \ge -M.$$

The following proposition establishes the relationship between the γ-rank calmness of (MOP) and that of (P').

PROPOSITION 3.13. *Let x^* be a locally weak efficient solution to* (MOP) *and $0 < \gamma < +\infty$. Then* (MOP) *is γ-rank calm at x^* if and only if (P') is γ-rank calm at x^*.*

A sufficient condition for the calmness of (MOP) *at a point is given in the following proposition.*

PROPOSITION 3.14. *Let $x^* \in X$ and $0 < \gamma < +\infty$. Assume that the following conditions hold:*

(i) *there exists $\lambda \in R^l_+ \backslash \{0\}$ such that x^* is a local minimum to*

$$(P_\lambda) \quad \begin{array}{ll} \inf\limits_{x \in X} & \lambda^T f(x) \\ s.t. & g_j(x) \le 0, \quad j = 1, \ldots, m; \end{array}$$

(ii) *(P_λ) is γ-rank calm at x^*.*

Then (MOP) *is γ-rank calm at x^*.*

The following lemma follows from a statement in [4, p. 239].

LEMMA 3.15. *Let $\gamma \in (0, 1], f_i (i = 1, \ldots, l), g_j (j = 1, \ldots, m)$ be locally Lipschitz functions around a local minimum x^* to (P'). If (P') satisfies either of the following constraint qualifications:*

(i) *Mangasarian–Fromowitz-type constraint qualification: there exists $v \in T^C_X(x^*)$ such that $g^0_j(x^*; v) < 0 \ \forall j \in J(x^*)$, where $J(x^*) = \{j : g_j(x^*) = 0, j = 1, \ldots, m\}$, $g^0_j(x^*; v)$ denotes the Clarke's generalized directional derivative of g_j at x^* in direction v, and $T^C_X(x^*)$ is the Clarke tangent cone of X at x^*,*

(ii) *Slater-type constraint qualification: if X is convex, g_j $(j = 1, \ldots, m)$ is convex around x^* (i.e., $\exists \delta > 0$ such that g_j is convex on the set $X_\delta = \{x \in X : \|x - x^*\| \le \delta\}$), there exists $x_0 \in X_\delta$ such that $g_j(x_0) < 0 \ \forall j \in J(x^*)$,*

then (P') is 1-rank calm at x^*; therefore, it is γ-rank calm at x^*.

It follows from Lemma 3.15 and Proposition 3.13 that we have the following proposition.

PROPOSITION 3.16. *Let $f_i (i = 1, \ldots, l), g_j (j = 1, \ldots, m)$ be locally Lipschitz around a local efficient solution x^* of* (MOP) *and either of* (i) *and* (ii) *in Lemma* 3.15 *hold. Then* (MOP) *is γ-rank calm at x^*.*

Finally, we note that if f is locally Lipschitz and all the constraint functions $g_j, j = 1, \ldots, m$, are affine and X is a polyhedron, then (P') is (1-rank) calm at any of its local minima (see [22], for instance). Thus, by Proposition 3.13, (MOP) is γ-rank calm at any of its local efficient solutions ($\gamma \in (0, 1]$).

4. Saddle points of nonlinear Lagrangian functions. In this section, we consider the saddle point problem of the nonlinear Lagrangian.

Let p be an increasing function defined on $C \times R^m$ (or $C \times R^m_+$) enjoying properties (A) and (B) and let the nonlinear Lagrangian L be defined by (1) (or (14)).

DEFINITION 4.1. *The point $(x^*, d^*) \in X \times R^m_+$ is called a saddle point of the nonlinear Lagrangian L if*

(i) $L(x, d^*) - L(x^*, d^*) \not\leq_{C\setminus\{0\}} 0 \ \forall x \in X$;
(ii) $L(x^*, d) - L(x^*, d^*) \not\geq_{C\setminus\{0\}} 0 \ \forall d \in R^m_+$.

It should be noted that a saddle point may not exist even if all the conditions of Theorem 2.12 *hold (see Example* 2.4 *due to Proposition* 4.2*).*

The following proposition presents the relationship among a saddle point of L, an efficient solution of (MOP), and an efficient solution of (DMOP) in the sense of maximum.

PROPOSITION 4.2. *The point $(x^*, d^*) \in X \times R^m_+$ is a saddle point of L if and only if x^* is an efficient solution of* (MOP)*, $f(x^*) \in q(d^*)$, and d^* is an efficient solution to* (DMOP).

In the following, we compare the Lagrangian function defined analogously as in [19, pp. 185–187] with a special class of nonlinear Lagrangian functions. Then we provide sufficient conditions for the existence of a saddle point of this special class of nonlinear Lagrangian functions.

As in [19], we define a Lagrangian function as follows:

$$L'(x, d) = f(x) + \sum_{j=1}^{m} d_j g_j(x) e,$$

where the dual variable $d = (d_1, \ldots, d_m) \in R^m_+, x \in X$.

Analogous to Definition 4.1, we can define a saddle point of L'.

It is clear that the following inequality holds:

$$\left(\sum_{i=1}^{m} b_i^{\gamma}\right)^{1/\gamma} \geq \sum_{i=1}^{m} b_i \quad \forall b_i \geq 0, \gamma \in (0, 1]. \tag{29}$$

Let $\gamma \in (0, 1]$. Consider the following class of nonlinear Lagrangian functions:

$$L_\gamma(x, d) = \sum_{i=1}^{l} \left[f_i^{\gamma}(x) + \sum_{j=1}^{m} d_j^{\gamma} g_j^{+\gamma}(x) \right]^{1/\gamma} e_i,$$

where $x \in X, d = (d_1, \ldots, d_m) \in R^m_+$. It follows from (29) that

$$L_\gamma(x, d) \geq_C f(x) + \sum_{j=1}^{m} d_j g_j^{+}(x) e \geq_C L'(x, d) \quad \forall x \in X, \ d \in R^m_+. \tag{30}$$

This inequality allows us to establish the following conclusion.

PROPOSITION 4.3. *Assume that $\gamma \in (0, 1]$. Any saddle point of L' is also a saddle point of L_γ.*

The following theorem follows from Theorem 3.5 and Proposition 4.2.

THEOREM 4.4. *Assume that $\gamma \in (0, 1]$ and* (MOP) *is* 1*-rank weakly stable. Then $x^* \in X$ is an efficient solution of* (MOP) *if and only if there exists $d^* \in R^m_+$ such that (x^*, d^*) is a saddle point of L_γ.*

5. Conclusions. In this paper, we introduced nonlinear Lagrangian functions and nonlinear penalty functions for constrained multiobjective optimization problems. We obtained weak and strong duality and saddle point results based on nonlinear Lagrangian functions. We also studied the relationship between the γ-rank weak stability and the exact penalization for inequality constrained multiobjective optimization problems, and the relationship between the γ-rank calmness and the local exact penalization.

Acknowledgment. The authors are grateful to the referees for their detailed comments and criticisms, which have improved the presentation of this paper.

REFERENCES

[1] E. J. Balder, *An extension of duality-stability relations to nonconvex optimization problems*, SIAM J. Control Optim., 15 (1977), pp. 329–343.
[2] M. S. Bazaraa and C. M. Shetty, *Nonlinear Programming: Theory and Algorithms*, John Wiley, New York, 1979.
[3] J. V. Burke, *Calmness and exact penalization*, SIAM J. Control Optim., 29 (1991), pp. 493–497.
[4] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.
[5] C. J. Goh and X. Q. Yang, *A nonlinear Lagrangian theory for nonconvex optimization*, J. Optim. Theory Appl., 109 (2001), pp. 99–121.
[6] 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.
[7] V. Jeyakumar and X. Q. Yang, *Convex composite multi-objective nonsmooth programming*, Math. Programming, 59 (1993), pp. 325–343.
[8] L. S. Lasdon, *Optimization Theory for Large Systems*, Macmillan, New York, 1970.
[9] D. Li, *Zero duality gap for a class of nonconvex optimization problems*, J. Optim. Theory Appl., 85 (1995), pp. 309–324.
[10] D. T. Luc, *Theory of Vector Optimization*, Springer-Verlag, Berlin, 1989.
[11] A. Mukherjea and K. Pothoven, *Real and Functional Analysis*, Plenum Press, New York, 1978.
[12] D. Pallaschke and S. Rolewicz, *Foundations of Mathematical Optimization*, Kluwer Academic Press, Dordrecht, The Netherlands, 1997.
[13] R. T. Rockafellar, *Conjugate Duality and Optimization*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 16, SIAM, Philadelphia, PA, 1974.
[14] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
[15] E. Rosenberg, *Exact penalty functions and stability in locally Lipschitz programming*, Math. Programming, 30 (1984), pp. 340–356.
[16] G. Z. Ruan and X. X. Huang, *Weak calmness and weak stability of multiobjective programming and exact penalty functions*, J. Systems Sci. Math. Sci., 12 (1992), pp. 148–157 (in Chinese).
[17] A. M. Rubinov, B. M. Glover, and X. Q. Yang, *Modified Lagrange and penalty functions in continuous optimization*, Optimization, 46 (1999), pp. 327–351.
[18] A. M. Rubinov, B. M. Glover, and X. Q. Yang, *Decreasing functions with applications to penalization*, SIAM J. Optim., 10 (1999), pp. 289–313.
[19] Y. Sawaragi, H. Nakayama, and T. Tanino, *Theory of Multiobjective Optimization*, Academic Press, New York, London, 1985.
[20] D. J. White, *Multiobjective programming and penalty functions*, J. Optim. Theory Appl., 43 (1984), pp. 583–599.
[21] X. Q. Yang, *Second-order global optimality conditions for convex composite optimization*, Math. Programming, 81 (1998), pp. 327–347.
[22] J. J. Ye, *Optimality conditions for optimization problems with complementarity constraints*, SIAM J. Optim., 9 (1999), pp. 374–387.
[23] Yu. G. Yevtushenko and V. G. Zhadan, *Exact auxiliary functions in optimization problems*, U.S.S.R. Comput. Maths. Math. Phys., 30 (1990), pp. 31–42.

SIAM J. OPTIM.
Vol. 13, No. 3, pp. 693–701

CONVERGENCE PROPERTIES OF THE BFGS ALGORITM*

YU-HONG DAI†

Abstract. The BFGS method is one of the most famous quasi-Newton algorithms for unconstrained optimization. In 1984, Powell presented an example of a function of two variables that shows that the Polak–Ribière–Polyak (PRP) conjugate gradient method and the BFGS quasi-Newton method may cycle around eight nonstationary points if each line search picks a local minimum that provides a reduction in the objective function. In this paper, a new technique of choosing parameters is introduced, and an example with only six cyclic points is provided. It is also noted through the examples that the BFGS method with Wolfe line searches need not converge for nonconvex objective functions.

Key words. unconstrained optimization, conjugate gradient method, quasi-Newton method, Wolfe line search, nonconvex, global convergence

AMS subject classifications. 65K05, 65K10

PII. S1052623401383455

1. The BFGS algorithm.

The BFGS algorithm is one of the most efficient quasi-Newton methods for unconstrained optimization:

$$\min f(x), \quad x \in \mathcal{R}^n. \tag{1.1}$$

The algorithm was proposed by Broyden [2], Fletcher [5], Goldfarb [7], and Shanno [19] individually and can be stated as follows.

ALGORITHM 1.1. THE BFGS ALGORITHM.

Step 0. *Given* $x_1 \in \mathcal{R}^n$; $B_1 \in \mathcal{R}^{n\times n}$ *positive definite;*
Compute $g_1 = \nabla f(x_1)$. *If* $g_1 = 0$, *stop; otherwise, set* $k := 1$.

Step 1. *Set* $d_k = -B_k^{-1} g_k$.

Step 2. *Carry out a line search along* d_k, *getting* $\alpha_k > 0$,
$x_{k+1} = x_k + \alpha_k d_k$, *and* $g_{k+1} = \nabla f(x_{k+1})$;
If $g_{k+1} = 0$, *stop.*

Step 3. *Set*

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{s_k^T y_k}, \tag{1.2}$$

where

$$s_k = \alpha_k d_k, \tag{1.3}$$

$$y_k = g_{k+1} - g_k. \tag{1.4}$$

Step 4. $k := k + 1$; *go to Step* 1.

*Received by the editors January 12, 2001; accepted for publication (in revised form) February 7, 2002; published electronically November 6, 2002. This work was supported by Chinese NSF grants 19801033 and 10171104 and a Youth Innovation Fund of the Chinese Academy of Science.
http://www.siam.org/journals/siopt/13-3/38345.html

†State Key Laboratory of Scientific and Engineering Computing, Institute of Computational Mathematics and Scientific/Engineering Computing, Academy of Mathematics and System Sciences, Chinese Academy of Sciences, P.O. Box 2719, Beijing 100080, People's Republic of China (dyh@lsec.cc.ac.cn).

The line search in Step 2 requires the steplength α_k to meet certain conditions. If exact line search is used, α_k satisfies

$$f(x_k + \alpha_k d_k) = \min_{\alpha>0} f(x_k + \alpha d_k). \tag{1.5}$$

In the implementations of the BFGS algorithm, one normally requires that the steplength α_k satisfies the Wolfe conditions [20]:

$$f(x_k + \alpha_k d_k) - f(x_k) \le \delta_1 \alpha_k d_k^T g_k, \tag{1.6}$$

$$d_k^T \nabla f(x_k + \alpha_k d_k) \ge \delta_2 d_k^T g_k, \tag{1.7}$$

where $\delta_1 \le \delta_2$ are constants in $(0, 1)$. For convenience, we call the line search that satisfies the Wolfe conditions (1.6)–(1.7) the Wolfe line search.

Another famous quasi-Newton method is the DFP method, which was discovered by Davidon [3] and modified by Fletcher and Powell [6]. Broyden [2] proposed a family of quasi-Newton methods:

$$B_{k+1}(\theta) = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{s_k^T y_k} + \theta(s_k^T B_k s_k) v_k v_k^T, \tag{1.8}$$

where $\theta \in \mathcal{R}^1$ is a scalar and $v_k = \frac{y_k}{s_k^T y_k} - \frac{B_k s_k}{s_k^T B_k s_k}$. The choice $\theta = 0$ gives rise to the BFGS update, whereas $\theta = 1$ defines the DFP method.

For uniformly convex functions, Powell [12] showed that the DFP algorithm with exact line searches stops at the unique minimum or generates a sequence that converges to the minimum. Dixon [4] found that all methods in the Broyden family with exact line searches produce the same iterations for general functions. For inexact line searches, Powell [14] first proved the global convergence of the BFGS algorithm with Wolfe line searches for convex functions. His result was extended by Byrd, Nocedal, and Yuan [1] to all methods in the restricted Broyden family with $\theta \in [0, 1)$. However, the following questions have remained open for many years (for example, see Nocedal [9] and Yuan [21]): (i) *does the DFP method with Wolfe line searches converge for convex functions?* and (ii) *does the BFGS method with Wolfe line searches converge for nonconvex functions?*

In this paper, we will consider the $n = 2, m = 8$ example in [15] for the Polak–Ribière–Polyak (PRP) conjugate gradient method [10, 11]. The two-dimensional example shows that the PRP method may cycle around eight nonstationary points if each line search picks a local minimum that provides a reduction in the objective function. By introducing a new technique of choosing parameters, we will present a new example for the PRP method (see section 2). The example has only six cyclic points. Since, in the case that $g_{k+1}^T d_k = 0$ for all k, the BFGS method can produce the same iterations as the PRP method does for two-dimensional functions, it can be shown by the examples that the BFGS method with Wolfe line searches need not converge for nonconvex objective functions (see section 3). Thus a negative answer is given to question (ii). The last section contains some discussions.

2. A counterexample with six cyclic points. The PRP method uses the negative gradient as its initial search direction. For $k \ge 1$, the method defines d_{k+1} as follows:

$$d_{k+1} = -g_{k+1} + \frac{g_{k+1}^T y_k}{\|g_k\|_2^2} d_k. \tag{2.1}$$

Powell [15] constructed a two-dimensional example, showing that the PRP method with the line search (2.2) may cycle around eight nonstationary points:

$$\alpha_k \text{ is a local minimum of } \Phi_k(\alpha) \text{ and such that } \Phi_k(\alpha_k) < \Phi_k(0), \tag{2.2}$$

where $\Phi_k(\alpha)$ is the line search function

$$\Phi_k(\alpha) = f(x_k + \alpha d_k), \quad \text{where } \alpha > 0. \tag{2.3}$$

However, examples with fewer cyclic points do not seem possible from the practice in [15]. In this section, we will introduce a new technique of choosing parameters and provide an example with only six cyclic points.

Assume that $n = 2$. Similar to [15], our example will be constructed so that all the iterations generated by the PRP method converge to the horizontal axis in $\mathcal{R}^2$. For m even, we consider the steps $\{s_k\}$ in the form

$$s_{mj+i} = a_i \begin{pmatrix} 1 \\ b_i\phi^{2j} \end{pmatrix}, \quad s_{mj+\frac{m}{2}+i} = a_i \begin{pmatrix} -1 \\ b_i\phi^{2j+1} \end{pmatrix}, \quad i = 1, \ldots, \frac{m}{2}, \tag{2.4}$$

where ϕ, $\{a_i\}$, $\{b_i\}$ are parameters to be determined, satisfying $\phi \in (0, 1)$ and $a_i > 0\,(i = 1, \ldots, \frac{m}{2})$. To be such that

$$g_{k+1}^T d_k = 0 \quad \text{for all } k, \tag{2.5}$$

we assume that the gradients $\{g_k\}$ have the form

$$\begin{cases} g_{mj+1} = c_1 \begin{pmatrix} b_{\frac{m}{2}}\phi^{2j-1} \\ 1 \end{pmatrix}; \qquad g_{mj+i} = c_i \begin{pmatrix} -b_{i-1}\phi^{2j} \\ 1 \end{pmatrix}, \quad i = 2, \ldots, \dfrac{m}{2}, \\ g_{mj+\frac{m}{2}+1} = c_1 \begin{pmatrix} -b_{\frac{m}{2}}\phi^{2j} \\ 1 \end{pmatrix}; \quad g_{mj+\frac{m}{2}+i} = c_i \begin{pmatrix} b_{i-1}\phi^{2j+1} \\ 1 \end{pmatrix}, \quad i = 2, \ldots, \dfrac{m}{2}, \end{cases} \tag{2.6}$$

where $\{c_i\}$ are also parameters to be determined. In this section, we are interested in the case that $m = 6$.

By relations (2.1) and (2.5), we know that the PRP method satisfies the conjugacy condition

$$d_{k+1}^T y_k = 0 \tag{2.7}$$

and the descent condition

$$d_{k+1}^T g_{k+1} < 0. \tag{2.8}$$

The above conditions require that $g_{6j+i}^T s_{6j+i} = g_{6j+i-1}^T s_{6j+i} < 0$, yielding

$$\begin{cases} c_2(b_2 - b_1) = c_1(b_2 + b_3\phi^{-1}) < 0, \\ c_3(b_3 - b_2) = c_2(b_3 - b_1) < 0, \\ c_1(b_1\phi + b_3) = c_3(b_1\phi + b_2) < 0. \end{cases} \tag{2.9}$$

Denoting $b_0 = -b_3\phi^{-1}$ and $b_4 = -b_1\phi$, we can draw the following conditions on $\{b_i\}$ from (2.9):

$$\begin{cases} (b_2 - b_1)(b_3 - b_2)(b_4 - b_3) = (b_2 - b_0)(b_3 - b_1)(b_4 - b_2), \\ (b_3 - b_4)(b_2 - b_0) > 0, (b_2 - b_1)(b_3 - b_1) > 0, (b_3 - b_2)(b_2 - b_4) > 0. \end{cases} \tag{2.10}$$

Defining $\varphi_i = b_i - b_{i-1}$, the above relations are equivalent to

$$\begin{cases} \varphi_2\varphi_3\varphi_4 = (\varphi_1+\varphi_2)(\varphi_2+\varphi_3)(\varphi_3+\varphi_4), \\ \varphi_4(\varphi_1+\varphi_2) < 0, \quad \varphi_2(\varphi_2+\varphi_3) > 0, \quad \varphi_3(\varphi_3+\varphi_4) < 0. \end{cases} \tag{2.11}$$

Further, letting $t_i = \varphi_{i+1}/\varphi_i$ and noting that $\varphi_4/\varphi_1 = -\phi$, we can obtain

$$\begin{cases} t_1t_2t_3 = (1+t_1)(1+t_2)(1+t_3) = -\phi, \\ t_1 > -1, \quad t_2 > -1, \quad t_3 < -1. \end{cases} \tag{2.12}$$

The first line in (2.12) is equivalent to

$$-t_1t_2t_3 = \frac{t_1t_2(1+t_1)(1+t_2)}{1+t_1+t_2} = \phi. \tag{2.13}$$

Thus for any $\phi \in (0,1)$ and $t_3 < -1$, we may solve t_1 and t_2 from (2.13). If the solved t_1 and t_2 are such that $t_1 > -1$ and $t_2 > -1$, then we can further consider the choices of $\{a_i\}$. In our real construction, we pick $t_3 = -2$. This with (2.13) indicates that

$$t_1t_2 = 1 + t_1 + t_2. \tag{2.14}$$

Further, we find that the following values of $\{t_i\}$ and ϕ satisfy (2.13) and allow suitable $\{a_i; i = 1,2,3\}$:

$$t_1 = -\frac{3}{4}, \quad t_2 = -\frac{1}{7}, \quad t_3 = -2, \quad \phi = \frac{3}{14}. \tag{2.15}$$

Now, by the definitions of φ_i and t_i, we can express $\sum_{i=2}^4 \varphi_i$ in two ways:

$$\begin{aligned} \sum_{i=2}^{4} \varphi_i &\overset{(1)}{=} b_4 - b_1 = -b_1(1+\phi) \\ &\overset{(2)}{=} \varphi_2(1+t_2+t_2t_3) = (b_2-b_1)(1+t_2+t_2t_3). \end{aligned} \tag{2.16}$$

We then get that

$$b_2 = \left[1 - \frac{1+\phi}{1+t_2+t_2t_3}\right] b_1. \tag{2.17}$$

Further, we have

$$b_3 = b_2 + \varphi_3 = b_2 + t_2\varphi_2 = (1+t_2)b_2 - t_2b_1. \tag{2.18}$$

Thus, letting $b_1 = 1$, we have from this, (2.17), and (2.18) that

$$b_1 = 1, \quad b_2 = -\frac{1}{16}, \quad b_3 = \frac{5}{56}. \tag{2.19}$$

Letting $c_2 = 1$, we obtain from (2.9) that

$$c_1 = -3, \quad c_2 = 1, \quad c_3 = -6. \tag{2.20}$$

As will be shown, the parameters chosen above allow the function value to be monotonically decreased. Define f^* to be the limit of $f(x_k)$. Since all the iterations

are required to converge to the horizontal axis and, for each value of the first variable, the dependence of $f(x)$ on the second variable is linear, we have that

$$f(x_k) - f^* = (x_k)_2 (g_k)_2 \quad \text{for all } k \geq 1, \tag{2.21}$$

where $(v)_i$ means the ith component of vector v. Given the limit $\hat{x}_1 = \lim_{j\to\infty} x_{6j+1}$, we can compute $\{x_{6j+i}; i = 1, \ldots, 4\}$ in the following way:

$$\begin{cases} x_{6j+1} = \hat{x}_1 - \sum_{k=j}^{\infty} \sum_{i=1}^{6} s_{6k+i}, \\ x_{6j+i} = x_{6j+i-1} + s_{6j+i-1}, \quad i = 2, 3, 4. \end{cases} \tag{2.22}$$

As a result, the second components of $\{x_{6j+i}; i = 1, \ldots, 4\}$ can be expressed as follows:

$$(x_{6j+i})_2 = -h_i (1-\phi)^{-1} \phi^{2j}, \quad i = 1, \ldots, 4, \tag{2.23}$$

where

$$\begin{cases} h_1 = a_1 b_1 + a_2 b_2 + a_3 b_3, \\ h_2 = a_1 b_1 \phi + a_2 b_2 + a_3 b_3, \\ h_3 = a_1 b_1 \phi + a_2 b_2 \phi + a_3 b_3, \\ h_4 = h_1 \phi. \end{cases} \tag{2.24}$$

Using the relations (2.21) and (2.23) and noting that the structure of this example has some symmetry, we know that the monotonicity of $f(x_k)$ requires $\{a_i\}$ to meet

$$-c_1 h_1 > -c_2 h_2 > -c_3 h_3 > -c_1 h_4. \tag{2.25}$$

This relation can be satisfied if we choose

$$a_1 = 14, \quad a_2 = 160, \quad a_3 = 1. \tag{2.26}$$

In this case, the four terms in (2.25) have the values

$$\frac{687}{56}, \quad \frac{387}{56}, \quad \frac{159}{28}, \quad \text{and} \quad \frac{2061}{784},$$

respectively. So (2.25) is satisfied. Further, if we let $(x_1)_1 = -87.5$, then $\{(x_{6j+i})_1; i = 1, \ldots, 6\}$ have the values -87.5, -73.5, 86.5, 87.5, 73.5, and -86.5, which are all different.

Finally, we discuss how to construct a smooth function $f(x) \in \mathcal{R}^2$ that satisfies the gradient conditions (2.6). At first, for given real numbers p_1, $p_2 (\neq 0)$, p_3, p_4, and any $j \geq 1$, we see that the function

$$\Psi(u_1, u_2) = \left[p_4 + p_2^{-1} p_3 (u_1 - p_1)\right] u_2 \tag{2.27}$$

is such that

$$\nabla \Psi \begin{pmatrix} p_1 \\ p_2 \phi^j \end{pmatrix} = \begin{pmatrix} p_3 \phi^j \\ p_4 \end{pmatrix}. \tag{2.28}$$

Note that $\{x_{6j+i}; i = 1, \ldots, 6\}$ are as follows:

$$\begin{pmatrix} -87.5 \\ -\frac{229}{44}\phi^{2j} \end{pmatrix}, \begin{pmatrix} -73.5 \\ \frac{387}{44}\phi^{2j} \end{pmatrix}, \begin{pmatrix} 86.5 \\ -\frac{53}{44}\phi^{2j} \end{pmatrix}, \begin{pmatrix} 87.5 \\ -\frac{229}{44}\phi^{2j+1} \end{pmatrix}, \begin{pmatrix} 73.5 \\ \frac{387}{44}\phi^{2j+1} \end{pmatrix}, \begin{pmatrix} -86.5 \\ -\frac{53}{44}\phi^{2j+1} \end{pmatrix}.$$

Letting $\mathcal{B}_i = \{u_1; |u_1 - (x_{6j+i})_1| \le 0.1\}$, it is easy to find one-dimensional C^∞ functions ξ and γ such that their values at the intervals $\{\mathcal{B}_i; i = 1, \ldots, 6\}$ are

$$\frac{8251}{458}, \quad -\frac{2847}{387}, \quad -\frac{6981}{212}, \quad \frac{8251}{458}, \quad -\frac{2847}{387}, \quad -\frac{6981}{212}$$

and

$$\frac{55}{229}, \quad -\frac{44}{387}, \quad \frac{33}{106}, \quad -\frac{55}{229}, \quad \frac{44}{387}, \quad -\frac{33}{106},$$

respectively. Then we can test that the function

$$f(u_1, u_2) = [\xi(u_1) + \gamma(u_1)u_1]u_2 \tag{2.29}$$

is a C^∞ function in $\mathcal{R}^2$ and satisfies the gradient conditions (2.6). One deficiency of the function (2.29) is that the point x_{6j+i+1} may not be a local minimum of $\Phi_{6j+i}(\alpha)$ (see (2.3) for the definition of Φ). For example, x_{6j+2}. For this, we can further choose a one-dimensional C^∞ function τ such that for $i = 1, \ldots, 6$ its value at $\mathcal{B}_i$ is equal to $(x_{6j+i})_1$. Then the C^∞ function

$$f(u_1, u_2) = [\xi(u_1) + \gamma(u_1)u_1 + M(u_1 - \tau(u_1))^2]u_2 \tag{2.30}$$

with $M > 0$ sufficiently large can guarantee that each x_{6j+i+1} is a local minimum of $\Phi_{6j+i}(\alpha)$. This completes the construction of our new example.

Thus by introducing the quantities φ_i and t_i, we have obtained a new example. The example shows that the PRP method with the line search (2.2) may cycle around six nonstationary points. One advantage of this example over the one in [15] is that it has only six cyclic points, whereas the latter has eight.

It is easy to see that the above example applies to the BFGS method if the choice of B_1 is such that $B_1 s_1 = -l g_1$, where l is any positive number. If one changes the definition of f in a small neighborhood of x_1 to meet the necessary initial conditions, the example is also efficient for the BFGS method with any positive definite matrix B_1 or the PRP method with $d_1 = -g_1$.

3. Nonconvergence of the BFGS algorithm for nonconvex functions. Generally, the line search (2.2) need not satisfy the Wolfe conditions (1.6)–(1.7). For example, consider the function

$$f(x) = \cos x, \quad x \in \mathcal{R}^1. \tag{3.1}$$

Assume that $x_k = 0$ and $d_k = 1$. For any nonnegative integer i, $\alpha = (2i + 1)\pi$ is a local minimum of $\Phi_k(\alpha)$. Then (1.6) is false if i is large. For the line search in the example of section 2, however, we can directly test that the Wolfe conditions (1.6)–(1.7) hold (see Theorem 3.1). Thus the example in section 2 also shows that the BFGS algorithm with Wolfe line searches need not converge for nonconvex objective functions.

THEOREM 3.1. *Consider the BFGS algorithm with the Wolfe line search* (1.6)–(1.7), *where* $\delta_1 \le \frac{69}{7480}$ *and* $\delta_2 \in (\delta_1, 1)$. *Then for any* $n \ge 2$ *there exists a starting point* x_1 *and a* C^∞ *function* f *in* $\mathcal{R}^n$ *such that the sequence* $\{\|g_k\|_2 : k = 1, 2, \ldots\}$ *generated by the algorithm is bounded away from zero.*

Proof. Consider the example in section 2. For any starting matrix B_1, we may slightly modify the example such that it satisfies the necessary initial conditions. By (2.21), (2.23), and (2.6), we see that

$$f(x_{6j+i}) = f^* - c_i h_i (1 - \phi)^{-1} \phi^{2j}, \quad i = 1, \ldots, 4. \tag{3.2}$$

Still denote $b_0 = -b_3\phi^{-1}$, $b_4 = -b_1\phi$ and let $a_4 = a_1$, $c_4 = c_1$. We have by (2.4) and (2.6) that

$$g_{6j+i}^T s_{6j+i} = a_i c_i (b_i - b_{i-1})\phi^{2j}, \quad i = 1, \ldots, 4. \tag{3.3}$$

Combining (3.2) and (3.3) and noting the symmetry of the example, we know that the first Wolfe condition (1.6) holds with any constant δ_1 satisfying

$$\begin{aligned} \delta_1 &\le \min\left\{ \frac{f(x_{6j+i+1}) - f(x_{6j+i})}{g_{6j+i}^T s_{6j+i}} : i = 1, 2, 3 \right\} \\ &= \frac{1}{1-\phi} \min\left\{ \frac{c_i h_i - c_{i+1} h_{i+1}}{a_i c_i (b_i - b_{i-1})} : i = 1, 2, 3 \right\} \\ &= \frac{69}{7480}. \end{aligned} \tag{3.4}$$

In addition, relations (2.5) and (2.8) imply that the second Wolfe condition (1.7) holds for $\delta_2 \in (\delta_1, 1)$. Thus the example in section 2 shows that the BFGS algorithm with Wolfe line searches need not converge for two-dimensional functions.

In the case when $n \ge 3$, we need only to consider the function

$$\hat{f}(x) = \hat{f}(u_1, u_2, \ldots, u_n) = f(u_1, u_2), \tag{3.5}$$

where f is the function in the example of section 2. This completes our proof. ∎

The parameter δ_1 in the above theorem is required to be no greater than $\frac{69}{7480} \approx 0.0092$. If we consider Powell's example with eight cyclic points, then Theorem 3.1 can be extended to $\delta_1 \le \frac{1}{84} \approx 0.0119$.

4. Some discussions. In this paper, it has been shown by one of Powell's examples in [15] and a new example with six cyclic points that the BFGS algorithm with Wolfe line searches need not converge for nonconvex objective functions. This result also applies to the Hestenes–Stiefel conjugate gradient method [8], the Broyden positive family (1.8) with $\theta \ge 0$, and the limited-memory quasi-Newton methods, since all these methods satisfy both the conjugacy condition (2.7) and the descent condition (2.8) if $g_{k+1}^T d_k = 0$ for all k.

To my knowledge, the parameters δ_1 and δ_2 in (1.6)–(1.7) are often set to 0.01 (or a smaller value) and 0.9, respectively, in the implementations of the BFGS algorithm. According to the remark after it, Theorem 3.1 can be extended to the case where $\delta_1 \le \frac{1}{84}$. Since $\frac{1}{84} > 0.01$, one would be satisfied with this result for the BFGS algorithm. As Professor J. C. Gilbert discussed with me, however, we wonder whether Theorem 3.1 holds for any $\delta_1 < 1$ in theory.

Using the same technique as in section 2, we can show that there do not exist examples of four cyclic points having similar structures. This means that the number of cyclic points, six, cannot be decreased if we assume m to be even. In fact, if $m = 4$, we have by (2.4) and (2.6) that

$$\begin{cases} c_2(b_2 - b_1) = c_1(b_2 + b_2\phi^{-1}) < 0, \\ c_1(b_2 + b_1\phi) = c_2(b_1 + b_1\phi) < 0, \end{cases} \tag{4.1}$$

where $\phi \in (0, 1)$. Denote $b_0 = -b_2\phi^{-1}$, $b_3 = -b_1\phi$, $\varphi_i = b_i - b_{i-1}$ $(i = 1, 2, 3)$, and $t_i = \varphi_{i+1}/\varphi_i (i = 1, 2)$. Similar to (2.10), (2.11), and (2.12), we can obtain

$$\begin{cases} t_1 t_2 = (1 + t_1)(1 + t_2) = -\phi, \\ t_1 > -1, \ t_2 < -1. \end{cases} \tag{4.2}$$

The above imply that $t_2 = -(1 + t_1)$ and $\phi = t_1(1 + t_1)$. Since $\phi \in (0, 1)$, we can then get that $t_1 > 0$. Further, letting $b_1 = 1$, we can, similarly to (2.16), obtain that $b_2 = (1+t_1)^2/t_1$. Since b_1, b_2, and ϕ are all positive, we know by $c_1(b_2+b_1\phi) < 0$ that $c_1 < 0$. Letting $c_1 = -t_1$, we can get by (4.1) that $c_2 = -(1 + t_1)$. In a way similar to (2.21)–(2.25), it is easy to see that the condition $f(x_{4j+1}) > f(x_{4j+2})$ requires

$$-c_1(a_1b_1 + a_2b_2) > -c_2(a_1b_1\phi + a_2b_2). \tag{4.3}$$

Substituting the expressions of ϕ, c_1, and c_2 with t_1, (4.3) is equivalent to

$$-(2 + t_1)t_1^2a_1b_1 - a_2b_2 > 0. \tag{4.4}$$

This is not possible since t_1, a_1, a_2, b_1, and b_2 are all positive. The contradiction shows the nonexistence of examples of four cyclic points.

Under the assumption that $x_k \to \bar{x}$, Powell [13] showed that the BFGS algorithm with exact line searches converges globally for general functions when there are only two variables. This result was extended by Pu and Yu [18] to the case in which $n \geq 2$. Therefore an interesting question may be, If $x_k \to \bar{x}$, is the BFGS algorithm with Wolfe line searches globally convergent for general functions? Another question is, Does there exist an inexact line search that ensures the global convergence of the BFGS method for general functions?

Recently, Powell [16] showed that if the line search always finds the first local minimum of $\Phi_k(\alpha)$ in (2.3), the BFGS method is globally convergent for two-dimensional twice-continuously differentiable functions with bounded level sets. Powell [17] and the author are trying to construct a three-dimensional example showing that the BFGS algorithm with the above line search need not converge.

Acknowledgments. The author is much indebted to Professors Y. Yuan, J. C. Gilbert, and M. J. D. Powell, who discussed with him the idea of this paper and gave many valuable suggestions and comments. Thanks are also due to the two anonymous referees, whose comments and suggestions greatly improved this paper.

REFERENCES

[1] 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.

[2] 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.

[3] W. C. Davidon, *Variable metric methods for minimization,* SIAM J. Optim., 1 (1991), pp. 1–17.

[4] L. C. W. Dixon, *Variable metric algorithms: Necessary and sufficient conditions for identical behavior of nonquadratic functions,* J. Optim. Theory Appl., 10 (1972), pp. 34–40.

[5] R. Fletcher, *A new approach to variable metric algorithms,* Computer J., 13 (1970), pp. 317–322.

[6] R. Fletcher and M. J. D. Powell, *A rapidly convergent descent method for minimization,* Comput. J., 6 (1963), pp. 163–168.

[7] D. Goldfarb, *A family of variable metric methods derived by variational means,* Math. Comp., 24 (1970), pp. 23–26.

[8] M. R. Hestenes and E. Stiefel, *Method of conjugate gradient for solving linear system,* J. Res. Nat. Bur. Standards, 49 (1952), pp. 409–436.

[9] J. Nocedal, *Theory of algorithms for unconstrained optimization,* Acta Numer., 1 (1992), pp. 199–242.

[10] E. Polak and G. Ribière, *Note sur la convergence de methodes de directions conjuguées,* Rev. Française Informat. Recherche Opérationelle, 16 (1969), pp. 35–43.

[11] B. T. Polyak, *Conjugate gradient method in extremal problems,* USSR Comp. Math. and Math. Phys., 9 (1969), pp. 94–112.

[12] M. J. D. Powell, *On the convergence of the variable metric algorithm,* J. Inst. Math. Appl., 7 (1971), pp. 21–36.
[13] M. J. D. Powell, *Quadratic termination properties of minimization algorithm, Part* I *and Part* II, J. Inst. Math. Appl., 10 (1972), pp. 333–357.
[14] M. J. D. Powell, *Some global convergence properties of a variable metric algorithm for minimization without exact line searches,* in Nonlinear Programming, SIAM-AMS Proceedings Vol. IX, R. W. Cottle and C. E. Lemke, eds., SIAM, Philadelphia, PA, 1976, pp. 53–72.
[15] M. J. D. Powell, *Nonconvex minimization calculations and the conjugate gradient method,* in Numerical Analysis, D. F. Griffiths, ed., Lecture Notes in Math. 1066, Springer-Verlag, Berlin, 1984, pp. 122–141.
[16] M. J. D. Powell, *On the convergence of the DFP algorithm for unconstrained optimization when there are only two variables,* Math. Program. Ser. B, 87 (2000), pp. 281–301.
[17] M. J. D. Powell, *private communication*, Department of Applied Mathematics and Theoretical Physics, Cambridge University, Cambridge, UK, 1997.
[18] D. Pu and W. Yu, *On the convergence property of DFP algorithm,* Ann. Oper. Res., 24 (1990), pp. 175–184.
[19] D. F. Shanno, *Conditioning of quasi-Newton methods for function minimization,* Math. Comp., 24 (1970), pp. 647–650.
[20] P. Wolfe, *Convergence conditions for ascent methods,* SIAM Rev., 11 (1969), pp. 226–235.
[21] Y. Yuan, *Numerical Methods for Nonlinear Programming,* Shanghai Scientific and Technical Publishers, Shanghai, 1993 (in Chinese).

SIAM J. OPTIM.
Vol. 13, No. 3, pp. 702–725

ACTIVE SETS, NONSMOOTHNESS, AND SENSITIVITY*

A. S. LEWIS†

Abstract. Nonsmoothness pervades optimization, but the way it typically arises is highly structured. Nonsmooth behavior of an objective function is usually associated, locally, with an *active manifold*: on this manifold the function is smooth, whereas in normal directions it is "vee-shaped." Active set ideas in optimization depend heavily on this structure. Important examples of such functions include the pointwise maximum of some smooth functions and the maximum eigenvalue of a parametrized symmetric matrix. Among possible foundations for practical nonsmooth optimization, this broad class of "partly smooth" functions seems a promising candidate, enjoying a powerful calculus and sensitivity theory. In particular, we show under a natural regularity condition that critical points of partly smooth functions are stable: small perturbations to the function cause small movements of the critical point on the active manifold.

Key words. active set, nonsmooth analysis, subdifferential, generalized gradient, sensitivity, $\mathcal{U}$-Lagrangian, eigenvalue optimization, spectral abscissa, identifiable surface

AMS subject classifications. Primary, 90C31, 49K40; Secondary, 65K10, 15A42

PII. S1052623401387623

1. Introduction.

Optimality conditions throughout the field of optimization are intimately bound up with nonsmoothness. As a simple example, consider the problem of minimizing a sum of Euclidean norms (cf. [1]):

$$\min_{x\in\mathbf{R}^n} h(x) := \sum_{i=1}^{k} \|A^i x - b^i\|$$

for given matrices A^i and vectors b^i. Except at the origin, the Euclidean norm is a *smooth* function, by which we will always mean twice continuously differentiable. Yet its nonsmoothness is crucial to any understanding of this problem. Associated with an optimal solution x_0 is an "active set" $\{i : A^i x_0 = b^i\}$, often nonempty, so the objective function h is nonsmooth at x_0. Furthermore, under reasonable conditions this active set is stable under small perturbations to the problem. (See [6, 20] for active set algorithms.)

This particular problem could be rephrased as a conic quadratic program, amenable to contemporary interior point techniques [1, 3]. Nonetheless, as in linear programming, the active set is an important tool for understanding the problem.

This phenomenon of nonsmoothness inducing a certain "activity" central to optimality conditions repeats many times throughout optimization. Consider the following examples.

(a) *Classical nonlinear programming and minimax.* At an optimal solution of a nonlinear constrained optimization problem, some subset of the inequality constraints is active (that is, those constraints hold with equality): under reasonable conditions (see, for example, [8]), this active set is stable under small perturbations to the problem.

*Received by the editors April 6, 2001; accepted for publication (in revised form) May 6, 2002; published electronically November 14, 2002. This research was supported by the NSERC.
http://www.siam.org/journals/siopt/13-3/38762.html

†Department of Mathematics, Simon Fraser University, Burnaby, BC V5A 1S6, Canada (aslewis@sfu.ca, http://www.math.sfu.ca/∼aslewis).

Somewhat analogously, consider a nonlinear minimax problem

$$\min_{x\in\mathbf{R}^n} h(x) := \max_{i=1,2,\ldots,k} h_i(x,u),$$

where each function h_i is smooth and u denotes a vector of parameters. Under reasonable conditions the active set at an optimal solution x_0,

$$I(x_0,u) := \{i : h_i(x_0,u) = h(x_0)\},$$

is stable under small changes in u.

(b) *Sums of norms.* Rather more generally than our initial example, we could consider the problem

$$\min_{x\in\mathbf{R}^n} h(x) := \sum_{i=1}^{k} \|F_i(x,u)\|,$$

where each function F_i is smooth and u denotes a vector of parameters. Under reasonable conditions the active set at an optimal x_0,

$$I(x_0,u) := \{i : F_i(x_0,u) = 0\},$$

is stable under small changes in u. Any smooth norm could be used in place of the Euclidean norm (cf. [7]).

(c) *Semidefinite programming and eigenvalue optimization.* The primal variable in a semidefinite program is a positive semidefinite matrix (see [16], for example). An optimal solution has a zero eigenvalue with a certain multiplicity: under reasonable conditions, this multiplicity is stable under small perturbations to the problem.

Relatedly, consider the eigenvalue optimization problem (see, for example, [14])

$$\min_{x\in\mathbf{R}^n} h(x) := \lambda_1(F(x,u)),$$

where the smooth function F takes real symmetric matrix values, u denotes a vector of parameters, and the function $\lambda_1(\cdot)$ is the largest eigenvalue. At an optimal solution this largest eigenvalue has a certain multiplicity, which under reasonable conditions is stable under small changes in u.

(d) *Spectral abscissa minimization.* More generally, consider the problem

$$\min_{x\in\mathbf{R}^n} h(x) := \alpha(F(x,u)),$$

where F now takes arbitrary square matrix values and the function $\alpha(\cdot)$ is the *spectral abscissa* (the largest real part of an eigenvalue). An optimal matrix generally has several distinct "active" eigenvalues with real part equal to its spectral abscissa, and each such eigenvalue has an associated algebraic multiplicity (the geometric multiplicity typically being one): under reasonable conditions this pattern of multiplicities is stable under small changes in u (see [5]).

Each of these problems has optimal solutions with a corresponding "activity," which is stable under small perturbations to the problem. In nonlinear minimax or sums of norms, the activity consists of subsets of indices; in eigenvalue optimization or spectral abscissa minimization, it consists of a certain pattern of multiplicities. These "activities" have powerful algorithmic significance: in each case, once the activity of

an optimal solution is known, finding it (at least locally) is a *smooth* minimization problem.

Let us summarize. The problem of minimizing a nonsmooth function is common in practice. But the nonsmoothness of a typical such function is highly structured: it induces a certain "activity" at an optimal solution, which under reasonable conditions is stable under small perturbations to the problem. Once the activity is known, the optimization problem is locally smooth.

The central idea of this current work is that the "activity" corresponds to a *manifold*. Each of the functions h above is what we will call *partly smooth*. Specifically, in a neighborhood of the point of interest x_0 there is a manifold $\mathcal{M}$ (the *active manifold*) containing x_0, with certain properties. Loosely speaking, the function h behaves smoothly as we move on the active manifold $\mathcal{M}$ and "sharply" if we move normal to the manifold; furthermore, in any fixed direction its directional derivative behaves continuously as we move on $\mathcal{M}$ and upper semicontinuously if we allow perturbations off it. (For closed convex functions, for example, this latter semicontinuity property, known as "regularity," is automatic.) We give the precise description in Definition 2.7.

The idea of partial smoothness at first sight appears rather intricate, but we shall find many interesting examples in practice. Each of our four examples is partly smooth under reasonable conditions. Given the parameter vector u, the four active manifolds are defined near x_0 as follows:

(a) $\{x : h_i(x,u) = h_j(x,u) \text{ for all } i,j \in I(x_0,u)\}$;
(b) $\{x : F_i(x,u) = 0 \text{ for all } i \in I(x_0,u)\}$;
(c) $\{x : \lambda_1(F(x,u)) \text{ has same multiplicity as } \lambda_1(F(x_0,u))\}$;
(d) $\{x : F(x,u) \text{ has same active eigenvalue multiplicities as } F(x_0,u)\}$.

Furthermore, we shall see that partly smooth functions have a robust calculus. Thus they form a rich, practical class of nonsmooth functions.

The literature contains many classes of nonsmooth functions more open to analysis than general, potentially pathological nonsmooth functions. A useful example is "amenability" [23, Def. 10.23], a powerful notion for combining smooth and convex techniques, again with a robust calculus. As we shall see, the real function $\sqrt{|\cdot|}$ is partly smooth at the origin relative to the active manifold $\{0\}$, but it is not amenable at the origin (see [23, Ex. 10.25(a)]), and it is not hard to construct similar Lipschitz examples using the fact that amenable functions are locally regular [23, Ex. 10.25(b)]. On the other hand, the convex, piecewise linear-quadratic function $x \mapsto \|x\|_1^2$ is not partly smooth relative to any manifold containing the origin.

The distinctive feature of partial smoothness is the notion of the active manifold: it is this idea that decouples the smooth behavior of the function from its "sharp" behavior. The importance of this general structure was realized for convex functions in [22], although not rigorously developed. The notion of active manifold is also implicit in the approach to polyhedral minimization via "structure functionals" [17]. In the nonconvex case the notion of the active manifold is familiar from active set methods for classical nonlinear programming (see [9], for example). In eigenvalue optimization the role of the active manifold is well known; see [21] and [24], for example. In spectral abscissa minimization, the idea is used heavily in [5].

For *convex* functions, partial smoothness is closely related to the "$\mathcal{U}$-Lagrangian" techniques of [12]: the active manifold is the "gully-shaped valley" of that work, and the normal and tangent spaces to the manifold correspond to the "$\mathcal{U}-\mathcal{V}$ decomposition" originating with the earlier work in [13] and developed for the maximum eigenvalue in [18, 19]. The idea of a "fast track" [15] is also closely related. We link

the $\mathcal{U}$-Lagrangian theory to partial smoothness towards the end of the current work. Notice, however, that many interesting examples of partly smooth functions are not convex: convexity is *not* the real driving force behind this theory.

Another closely related idea is the notion of an "identifiable surface" of a convex set [26], which is a subset of the boundary having a suitable "sharpness" property. In [26] it is shown that, if the solution of an optimization problem posed over such a set lies in an identifiable surface, then various standard constrained optimization algorithms "identify" the surface after a finite number of iterations. Hence the idea of identifiability is a powerful tool for algorithmic analysis.

Remarkably, as we shall see, for convex sets, the ideas of identifiability and partial smoothness coincide, reinforcing the power of this theory. By contrast with identifiability, however, partial smoothness is defined in a more geometric manner, and once again is *not* dependent on convexity.

To demonstrate the power of partly smooth techniques, our culminating result is a sensitivity theorem. In classical nonlinear programming, if a local minimizer has linearly independent active constraints and satisfies strict complementarity and a strong second-order condition, then the minimizer depends smoothly on the parameters of the problem (see [8], for example). An analogous result for eigenvalue optimization appears in [24] and for spectral abscissa minimization, in [5]. Our work here shows how partial smoothness unifies this work. To sketch the idea, suppose the function h is partly smooth at a point x_0 relative to the active manifold $\mathcal{M}$. If x_0 is a strong second-order minimizer of the smooth, restricted function $h|_{\mathcal{M}}$, and is a "sharp" minimizer of the restriction to the normal space $h|_{x_0+N_{\mathcal{M}}(x_0)}$, then the critical point x_0 varies smoothly over $\mathcal{M}$ as the parameters of the problem vary. Back in the context of nonlinear programming, our "sharp minimizer" condition corresponds to the usual strict complementarity condition, and the usual linear independence assumption becomes a transversality condition allowing us to apply a chain rule.

The proof of our sensitivity result amounts to local reduction to a smooth equality-constrained problem. Such a reduction is a standard approach to sensitivity results in nonlinear programming (see [4, Rem. 4.127], for example), and also works in semidefinite programming [4, p. 495]. By comparison, we are able here to consider rather general optimization problems, and without recourse to general nonsmooth second-order theory (such as [23, Chap. 13], for example), but the price of this generality is that we must settle for critical points in our sensitivity analysis, rather than local minimizers (see the example in section 7).

Partial smoothness seems a promising framework for practical nonsmooth optimization. Partly smooth functions form a wide and robust class, with many of the properties sought by previously cited researchers interested in algorithm development, stemming from the decoupling of the smooth and sharp behaviors. We defer algorithmic discussion to a later work.

2. Partial smoothness. We begin with some elementary definitions. We follow the notation and terminology of [23] throughout.

We consider a fixed Euclidean space X (a finite-dimensional real inner product space). We denote the subspace parallel to a nonempty convex set $C \subset X$ by $\operatorname{par} C$. Thus for any point $x \in C$ we have

$$\operatorname{par} C = (\operatorname{aff} C) - x = \mathbf{R}(C - C) = \mathbf{R}_+(C - C),$$

where $\operatorname{aff} C$ is the affine span of C. Easy exercises show $\operatorname{par}(AC) = A\operatorname{par} C$ for any linear map A, and $\operatorname{par}(C_1 \times C_2) = \operatorname{par} C_1 \times \operatorname{par} C_2$ for arbitrary nonempty convex

sets C_1 and C_2. We denote the extended reals by $\overline{\mathbf{R}} = [-\infty, +\infty]$. The *lineality space* of a sublinear function $f : X \to \overline{\mathbf{R}}$ is the subspace

$$\operatorname{lin} f = \{w \in X : f(w) = -f(-w)\}.$$

Let us consider a function $h : X \to \overline{\mathbf{R}}$, finite at a point $x \in X$. We review some definitions from [23]. The *subderivative* $dh(x)(\cdot) : X \to \overline{\mathbf{R}}$ is defined by

$$dh(x)(\bar{w}) = \liminf_{\tau \downarrow 0,\ w \to \bar{w}} \frac{h(x + \tau w) - h(x)}{\tau} \quad (\bar{w} \in X)$$

and the set of *regular subgradients* is (see [23, Ex. 8.4])

$$\hat{\partial} h(x) = \{v \in X : \langle v, w \rangle \leq dh(x)(w) \text{ for all } w \in X\}.$$

The set of *subgradients* is

$$\partial h(x) = \left\{ \lim_r v_r : v_r \in \partial h(x_r),\ x_r \to x,\ h(x_r) \to h(x) \right\},$$

while the set of *horizon subgradients* is

$$\partial^\infty h(x) = \left\{ \lim_r \lambda_r v_r : v_r \in \partial h(x_r),\ x_r \to x,\ h(x_r) \to h(x),\ \lambda_r \downarrow 0 \right\}.$$

Suppose in addition $\partial h(x) \neq \emptyset$. Then h is *(subdifferentially) regular* at x if h is locally lower semicontinuous around x, every subgradient is regular, and furthermore the recession cone (in the sense of convex analysis) $h(x)^\infty$ coincides with $\partial^\infty h(x)$ (see [23, Cor. 8.11]). In this case, the support function of $\partial h(x)$ is the subderivative $dh(x)$ [23, Thm. 8.30]. This is the case in particular for any closed convex function h, and in this case ∂h is the usual subdifferential in the sense of convex analysis.

PROPOSITION 2.1 (lineality space of subderivative). *If the function h is regular at the point $x \in X$, and has a subgradient there, then*

$$\operatorname{lin} dh(x) = (\operatorname{par} \partial h(x))^\perp.$$

Proof. We know $w \notin \operatorname{lin} dh(x)$ if and only if $dh(x)(w) + dh(x)(-w) > 0$, which by [23, Thm. 8.30] is equivalent to the existence of subgradients y and z of h at x satisfying $\langle y - z, w \rangle > 0$, or equivalently $w \notin (\partial h(x) - \partial h(x))^\perp$. The result follows. □

Given a set $\mathcal{M} \subset X$ containing a point x, we call a function $f : \mathcal{M} \to \overline{\mathbf{R}}$ *smooth around* x if x has an open neighborhood V in X such that some smooth function $g : V \to \mathbf{R}$ agrees with f on $\mathcal{M} \cap V$. We call such a function g a *smooth representative* of f around x. Note that in this case f is also smooth around any nearby point in $\mathcal{M}$. We call x a *critical point* of f if

$$f(z) - f(x) = o(\|z - x\|) \text{ for } z \text{ close to } x \text{ in } \mathcal{M}.$$

We call the function f *smooth* if it is smooth around every point in $\mathcal{M}$.

A "manifold" in X, loosely speaking, is a set consisting locally of the solutions of some smooth equations with linearly independent gradients. To be more precise, we say that a set $\mathcal{M} \subset X$ *is a manifold (of codimension m) around* a point $x \in X$ if $x \in \mathcal{M}$ and there is an open set $V \subset X$ containing x such that

$$\mathcal{M} \cap V = \{x \in V : F(x) = 0\},$$

where the smooth function $F : V \to \mathbf{R}^m$ has surjective derivative throughout V. In this case, the *tangent space* to $\mathcal{M}$ at x is given by

$$T_{\mathcal{M}}(x) = \operatorname{Ker}(\nabla F(x))$$

(which is independent of the choice of F), and the *normal space* to $\mathcal{M}$ at x is the orthogonal complement of the tangent space, namely

$$N_{\mathcal{M}}(x) = R(\nabla F(x)^*)$$

(where $R(\cdot)$ denotes range). The set $\mathcal{M}$ is then Clarke regular at x, and its normal cone there is exactly the normal space [23, Ex. 6.8].

We call a set $\mathcal{M}$ a *manifold (of codimension m)* if $\mathcal{M}$ is a manifold of codimension m around every point in $\mathcal{M}$. (More precisely, $\mathcal{M}$ is an "m-codimensional manifold embedded in X"; see [25].) If $\mathcal{M}$ is a manifold around a point x, then $\mathcal{M} \cap U$ is a manifold for some open neighborhood $U \subset X$ of x.

If the function $f : \mathcal{M} \to \overline{\mathbf{R}}$ is smooth around x and $\mathcal{M}$ is a manifold around x, then x is a critical point of f if and only if

$$\nabla g(x) \in N_{\mathcal{M}}(x),$$

where g is any smooth representative of f around x. In particular, this holds if x is a local minimizer of f.

The indicator function $\delta_{\mathcal{M}}$ takes the value 0 on $\mathcal{M}$ and $+\infty$ otherwise.

PROPOSITION 2.2 (subgradients and smoothness). *Suppose the set $\mathcal{M} \subset X$ is a manifold around the point $x \in \mathcal{M}$. For a function $h : X \to \overline{\mathbf{R}}$, if the restriction $h|_{\mathcal{M}}$ is smooth around x, then*

$$\hat{\partial} h(x) \subset \nabla g(x) + N_{\mathcal{M}}(x) \tag{2.3}$$

for any smooth representative g of $h|_{\mathcal{M}}$ around x, and hence

$$\operatorname{par} \hat{\partial} h(x) \subset N_{\mathcal{M}}(x).$$

Proof. For some open neighborhood V of x we have $g + \delta_{\mathcal{M}\cap V} = h + \delta_{\mathcal{M}\cap V}$, so by [23, Cor. 10.9] we deduce

$$\nabla g(x) + N_{\mathcal{M}}(x) = \hat{\partial}(g + \delta_{\mathcal{M}})(x) = \hat{\partial}(h + \delta_{\mathcal{M}})(x) \supset \hat{\partial} h(x),$$

and the result follows. □

Putting this together with the previous result, we arrive at the following proposition.

PROPOSITION 2.4 (smoothness and lineality). *Suppose the set $\mathcal{M} \subset X$ is a manifold around the point x. Suppose also that the function $h : X \to \overline{\mathbf{R}}$ has a subgradient at x and is regular there, and furthermore that the restriction $h|_{\mathcal{M}}$ is smooth around x. Then the subderivative $dh(x)$ is linear on the tangent space, or in other words*

$$\operatorname{lin} dh(x) \supset T_{\mathcal{M}}(x), \tag{2.5}$$

and the horizon subdifferential satisfies

$$\partial^\infty h(x) \subset N_{\mathcal{M}}(x). \tag{2.6}$$

Furthermore, the following properties are equivalent:

(i) *The lineality and tangent spaces coincide:*

$$\operatorname{lin} dh(x) = T_{\mathcal{M}}(x).$$

(ii) *The subdifferential and normal space are parallel:*

$$\operatorname{par} \partial h(x) = N_{\mathcal{M}}(x).$$

(iii) *h is "sharp" in normal directions at x, by which we mean*

$$dh(x)(-w) > -dh(x)(w) \quad \textit{whenever } 0 \neq w \in N_{\mathcal{M}}(x).$$

Finally, if any of the above three properties hold, then $\nabla g(x) \in \operatorname{aff} \partial h(x)$ for any smooth representative g of $h|_{\mathcal{M}}$ and hence the following properties are equivalent:

(a) *x is a critical point of $h|_{\mathcal{M}}$;*
(b) $0 \in \operatorname{aff} \partial h(x)$*;*
(c) $\operatorname{aff} \partial h(x) = N_{\mathcal{M}}(x)$.

Proof. The first inclusion follows from Propositions 2.1 and 2.2, and the second (2.6) follows from the fact that $\partial^\infty h(x)$ is the recession cone of $\partial h(x)$. The equivalence of statements (i) and (ii) is also a consequence of Proposition 2.1. On the other hand, by Proposition 2.2, statement (ii) fails if and only if there exists a nonzero vector w in $N_{\mathcal{M}}(x)$ orthogonal to $\operatorname{par} \partial h(x)$, or in other words satisfying

$$\langle w, u - v \rangle = 0 \quad \text{for all } u, v \in \partial h(x),$$

and since we have

$$dh(x)(w) + dh(x)(-w) = \sup\{\langle w, u - v \rangle : u, v \in \partial h(x)\},$$

this is in turn equivalent to statement (iii) failing.

For the last statement, note that inclusion (2.3), regularity, and property (ii) imply $\nabla g(x) \in \operatorname{aff} \partial h(x)$, and hence $\operatorname{aff} \partial h(x) = \nabla g(x) + N_{\mathcal{M}}(x)$. This shows that properties (a) and (b) are equivalent, and the equivalence of properties (b) and (c) follows from property (ii). □

We are now ready for the key definition.

DEFINITION 2.7. *Suppose that the set $\mathcal{M} \subset X$ contains the point x. The function $h : X \to \overline{\mathbf{R}}$ is* partly smooth at x relative to $\mathcal{M}$ *if $\mathcal{M}$ is a manifold around x and the following four properties hold:*

(i) (restricted smoothness) *the restriction $h|_{\mathcal{M}}$ is smooth around x;*
(ii) (regularity) *at every point close to x in $\mathcal{M}$, the function h is regular and has a subgradient;*
(iii) (normal sharpness) *$dh(x)(-w) > -dh(x)(w)$ for all nonzero directions w in $N_{\mathcal{M}}(x)$;*
(iv) (subgradient continuity) *the subdifferential map ∂h is continuous at x relative to $\mathcal{M}$.*

We say h is partly smooth relative to *a set $\mathcal{M}$ if $\mathcal{M}$ is a manifold and h is partly smooth at each point in $\mathcal{M}$ relative to $\mathcal{M}$.*

DEFINITION 2.8 (partly smooth sets). *A set $S \subset X$ is* partly smooth at *a point x* relative to *a set $\mathcal{M}$ if δ_S is partly smooth at x relative to $\mathcal{M}$. We say S is* partly smooth relative to *a set $\mathcal{M}$ if $\mathcal{M}$ is a manifold and S is partly smooth at each point in $\mathcal{M}$ relative to $\mathcal{M}$.*

NOTE 2.9 (equivalent properties). Some comments may help with this rather lengthy definition.

(a) By Propositions 2.1, 2.2, and 2.4, we could replace property (iii) (normal sharpness) by either of the following properties:
(iii*) (*tangent linearity of subderivative*)

$$\operatorname{lin} dh(x) \subset T_{\mathcal{M}}(x)$$

(or indeed the corresponding equality);
(iii**) (*normals parallel to subdifferential*)

$$N_{\mathcal{M}}(x) \subset \operatorname{par} \partial h(x)$$

(or again the corresponding equality).

(b) Property (i) ensures that h is continuous relative to $\mathcal{M}$, so the subdifferential mapping is always outer semicontinuous relative to $\mathcal{M}$, by [23, Prop. 8.7]. Hence we could replace property (iv) by the following property:
(iv*) (*subgradient inner semicontinuity*) The subdifferential ∂h is *inner semicontinuous* at x relative to $\mathcal{M}$: in other words, for any sequence of points x_r in $\mathcal{M}$ approaching x and any subgradient $y \in \partial h(x)$, there exist subgradients $y_r \in \partial h(x_r)$ approaching y.
Notice that if h is locally Lipschitz (or "strictly continuous" in the terminology of [23]), then the subdifferential $\partial h(x)$ is everywhere nonempty and compact [23, Thm. 9.13], so by [23, Cor. 11.35] we could replace condition (iv) by the following condition:
(iv) (*subderivative continuity*) for all directions $w \in X$, the function $x \in \mathcal{M} \mapsto dh(x)(w)$ is continuous at x_0.
Furthermore, in this case the subderivative reduces to

$$dh(x)(w) = \liminf_{t \downarrow 0} \frac{h(x+tw) - h(x)}{t},$$

and regularity at x amounts to upper semicontinuity of the function $dh(\cdot)(w)$ at x for all directions w [23, Ex. 9.15 and Cor. 8.19]. This justifies the description of partial smoothness we gave in the introduction.

(c) Although the definition of partial smoothness is for a function h defined everywhere on the space X, it extends unchanged to a function defined only close to the point of interest, since partial smoothness depends only on properties of h near that point.

For a partly smooth function, the "normal sharpness" condition (iii), or equivalently, conditions (iii*) (tangent linearity of subderivative) and (iii**) (normals parallel to subdifferential), are all "stable": the fact that they hold at the point x_0 implies that they also hold at all nearby points in the active manifold. That is the content of the following result.

PROPOSITION 2.10 (local normal sharpness). *If the function $h : X \to \overline{\mathbf{R}}$ is partly smooth at the point x_0 relative to the set $\mathcal{M} \subset X$, then all points $x \in \mathcal{M}$ close to x_0 satisfy the condition*

$$dh(x)(-w) > -dh(x)(w) \quad \textit{for all } 0 \neq w \in N_{\mathcal{M}}(x),$$

or equivalently, the condition

$$N_{\mathcal{M}}(x) = \operatorname{par} \partial h(x).$$

Proof. The two properties are equivalent by Note 2.9. By Proposition 2.2 (subgradients and smoothness) we know $N_{\mathcal{M}}(x) \supset \operatorname{par} \partial h(x)$, so if the result fails, then there is a sequence of points $x_r \in \mathcal{M}$ approaching x_0 and a sequence of unit vectors $y_r \in N_{\mathcal{M}}(x_r)$ orthogonal to $\operatorname{par} \partial h(x_r)$. Taking a subsequence, we can suppose that y_r approaches a unit vector $y_0 \in N_{\mathcal{M}}(x_0)$.

Now for arbitrary subgradients $u_0, v_0 \in \partial h(x_0)$, by the continuity of ∂h there exist sequences $u_r \in \partial h(x_r)$ approaching u_0 and $v_r \in \partial h(x_r)$ approaching v_0, and they must satisfy $\langle y_r, u_r - v_r \rangle = 0$. Taking the limit shows $\langle y_0, u_0 - v_0 \rangle = 0$, so since u_0 and v_0 were arbitrary we deduce y_0 is orthogonal to $\operatorname{par} \partial h(x_0) = N_{\mathcal{M}}(x_0)$, which contradicts the fact that y_0 is a unit vector in $N_{\mathcal{M}}(x_0)$. □

We end this section with a simple characterization of partly smooth sets.

PROPOSITION 2.11 (partly smooth sets). *Suppose that the set $\mathcal{M} \subset X$ contains the point x_0. A set $S \subset X$ is partly smooth at x_0 relative to $\mathcal{M}$ if and only if $\mathcal{M}$ is a manifold around x_0 and the following four properties hold:*

(i) *$S \cap \mathcal{M}$ is a neighborhood of x_0 in $\mathcal{M}$;*
(ii) *S is Clarke regular at each point in $\mathcal{M}$ close to x_0;*
(iii) *$N_{\mathcal{M}}(x_0) \subset N_S(x_0) - N_S(x_0)$;*
(iv) *the normal cone map $N_S(\cdot)$ is continuous at x_0 relative to $\mathcal{M}$.*

Proof. This is an easy exercise using the facts that the set S is Clarke regular at the point $x \in S$ if and only if δ_S is regular there, and that $\partial \delta_S(x) = N_S(x)$, and then applying property (iii**) (normals parallel to subdifferential) in Note 2.9. □

The definition of partial smoothness looks a little involved at first sight, but we shall see that there are many important examples.

3. Basic examples. In this section we describe a few basic examples of partly smooth functions. In the next section we describe some calculus rules for building more complex examples.

Example 3.1 (smooth functions). If the open set $\Omega \subset X$ contains the point x and the function $h : \Omega \to \mathbf{R}$ is smooth, then h is partly smooth at x relative to Ω.

Example 3.2 (indicator functions). If $\mathcal{M} \subset X$ is a manifold around the point x, then $\mathcal{M}$ is a partly smooth set at x relative to $\mathcal{M}$. This is an easy consequence of Proposition 2.11 (partly smooth sets).

Example 3.3 (distance functions). If $\mathcal{M} \subset X$ is a manifold around the point x_0, then the distance function $d_{\mathcal{M}} : X \to \mathbf{R}$ defined by

$$d_{\mathcal{M}}(x) = \inf\{\|y - x\| : y \in \mathcal{M}\}$$

is partly smooth at x_0 relative to $\mathcal{M}$. To see this, notice that $\delta_{\mathcal{M}}|_{\mathcal{M}}$ is identically zero, which is smooth. By [23, Ex. 8.53] we know that $d_{\mathcal{M}}$ is regular at each point $x \in \mathcal{M}$ and

$$\partial h(x) = B \cap N_{\mathcal{M}}(x)$$

(where B denotes the closed unit ball in X). Thus the normal space is again parallel to the subdifferential, and this subdifferential varies continuously as x varies in $\mathcal{M}$. In fact, the Euclidean norm could be replaced by any other norm in this example, providing we replace B in the subdifferential formula above with the dual ball.

Notice in particular that the norm $\|\cdot\|$ is partly smooth at the origin relative to the origin.

Example 3.4 (polyhedral functions). Given any function $h : X \to \overline{\mathbf{R}}$ that is *polyhedral* (that is, its epigraph is a polyhedral set) and any point x_0 at which h is

finite, there is a natural manifold about x_0 relative to which h is partly smooth. To see this we express h in the form (see [23, Thm. 2.49])

$$h(x) = \begin{cases} \max\{\langle a^i, x\rangle + b_i : i \in I\} & \text{if } \langle c^j, x\rangle \le d_j \ \text{ for all } j \in J, \\ +\infty & \text{otherwise} \end{cases}$$

for some finite index sets $I \neq \emptyset$ and J and given vectors a^i and c^j in X and reals b_i and d_j (for $i \in I$ and $j \in J$). For any point $x \in X$, define "active" index sets

$$I(x) = \{i \in I : \langle a^i, x\rangle + b_i = h(x)\},$$
$$J(x) = \{j \in J : \langle c^j, x\rangle = d_j\}.$$

Define the set

$$\mathcal{M}_{x_0} = \{x \in X : I(x) = I_0 \text{ and } J(x) = J_0\},$$

where $I_0 = I(x_0)$ and $J_0 = J(x_0)$. It is easy to see that $\mathcal{M}_{x_0}$ is a manifold around x_0. We claim that h is partly smooth at x_0 relative to $\mathcal{M}_{x_0}$.

To see this observe first that for any index $i \in I_0$ we have

$$h(x) = \langle a^i, x\rangle + b_i \ \text{ for all } x \in \mathcal{M}_{x_0},$$

so $h|_{\mathcal{M}_{x_0}}$ is smooth. Second, h is lower semicontinuous and convex, and hence regular whenever it is finite [23, Ex. 7.27]. Now routine calculation (using [23, Thm. 6.46], for example) shows that at any point $x \in \mathcal{M}_{x_0}$ we have

$$N_{\mathcal{M}_{x_0}}(x) = \left\{ \sum_{i\in I_0} \lambda_i a^i + \sum_{j\in J_0} \mu_j c^j : \sum_{i\in I_0} \lambda_i = 0 \right\},$$

$$\partial h(x) = \left\{ \sum_{i\in I_0} \lambda_i a^i + \sum_{j\in J_0} \mu_j c^j : \sum_{i\in I_0} \lambda_i = 1, \quad \lambda_i \ge 0 \ \ (i \in I_0), \right.$$
$$\left. \mu_j \ge 0 \ \ (j \in J_0) \right\}.$$

Thus the normal space is parallel to the subdifferential, which is constant on $\mathcal{M}_{x_0}$.

In particular, the *basic max function* $\mathrm{mx} : \mathbf{R}^n \to \mathbf{R}$ defined by $\mathrm{mx}\, x = \max_i x_i$ is partly smooth at any point $x_0 \in \mathbf{R}^n$ relative to the set

$$\mathcal{M}_{x_0} = \{x \in \mathbf{R}^n : I(x) = I(x_0)\}, \tag{3.5}$$

where

$$I(x) = \left\{ j : x_j = \max_i x_i \right\}.$$

Example 3.6 (largest eigenvalue). The Euclidean space $\mathbf{S}^n$ consists of the n-by-n real symmetric matrices with the inner product $\langle x, y\rangle = \mathrm{trace}(xy)$, for $x, y \in \mathbf{S}^n$. The functions $\lambda_1(x) \ge \lambda_2(x) \ge \cdots \ge \lambda_n(x)$ denote the eigenvalues of x (listed in decreasing

order by multiplicity). Then the largest eigenvalue is partly smooth relative to the manifold

$$\mathcal{M}_m = \{x \in \mathbf{S}^n : \lambda_1(x) \text{ has multiplicity } m\} \quad (1 \leq m \leq n).$$

To see this, note first that the set $\mathcal{M}_m$ above is indeed a manifold (see [18], for example). Furthermore we can write the maximum eigenvalue as

$$\lambda_1(x) = m^{-1} \sum_{j=1}^{m} \lambda_j(x) \text{ for all } x \in \mathcal{M}_m,$$

and the right-hand side is a smooth function of x on $\mathcal{M}_m$ (see [11], for example). Second, λ_1 is convex (see [10], for example) and so is regular everywhere. Now, by [18], as x varies in $\mathcal{M}_m$ there is an n-by-m matrix $Q(x)$, depending continuously on x, whose columns are a basis for the eigenspace of x corresponding to $\lambda_1(x)$, and then we have

$$N_{\mathcal{M}_m}(x) = Q(x)\{w \in \mathbf{S}^n : \operatorname{trace} w = 0\}Q(x)^T,$$
$$\partial\lambda_1(x) = Q(x)\{w \in \mathbf{S}^n_+ : \operatorname{trace} w = 1\}Q(x)^T,$$

where $\mathbf{S}^n_+$ denotes the positive semidefinite matrices [18, Thm. 4.7]. It is easy to see from this that the normal space is parallel to the subdifferential, which varies continuously on $\mathcal{M}_m$.

Example 3.7 (spectral abscissa). The Euclidean space $\mathbf{M}^n$ consists of the n-by-n complex matrices with the (real) inner product $\langle x, y\rangle = \operatorname{Re}\operatorname{trace}(x^*y)$ for $x, y \in \mathbf{M}^n$. The *spectral abscissa* $\alpha(x)$ is the largest of the real parts of the eigenvalues of x.

Given any list $\phi = (n_1, n_2, \ldots, n_r)$ of positive integers with sum no greater than n, let $\mathcal{M}_\phi$ denote the subset of $\mathbf{M}^n$ consisting of matrices x satisfying the following properties:

(i) x has r distinct "active" eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_r$ with real part $\alpha(x)$, and all its other eigenvalues have real part strictly less than $\alpha(x)$;
(ii) each active eigenvalue λ_j has algebraic multiplicity n_j and geometric multiplicity one.

Classic results of Arnold [2] show that $\mathcal{M}_\phi$ is a manifold.

In fact the spectral abscissa α is partly smooth relative to $\mathcal{M}_\phi$ (see [5]).

4. Calculus. In this section we show that partly smooth functions form a robust class by proving a variety of calculus rules. Our fundamental result considers the composition of a partly smooth function with a smooth function, and requires a transversality condition. Consider Euclidean spaces X and Z, an open set $W \subset Z$ containing a point z, a smooth map $\Phi : W \to X$, and a set $\mathcal{M} \subset X$. We say Φ is *transversal to* $\mathcal{M}$ *at* z if $\mathcal{M}$ is a manifold around $\Phi(z)$, and

$$R(\nabla\Phi(z)) + T_{\mathcal{M}}(\Phi(z)) = X$$

or equivalently

$$\operatorname{Ker}(\nabla\Phi(z)^*) \cap N_{\mathcal{M}}(\Phi(z)) = \{0\}. \tag{4.1}$$

THEOREM 4.2 (chain rule). *Given Euclidean spaces X and Z, an open set $W \subset Z$ containing a point z_0, a smooth map $\Phi : W \to X$, and a set $\mathcal{M} \subset X$, suppose Φ is*

transversal to $\mathcal{M}$ at z_0. If the function $h : X \to \overline{\mathbf{R}}$ is partly smooth at $\Phi(z_0)$ relative to $\mathcal{M}$, then the composition $h \circ \Phi$ is partly smooth at z_0 relative to $\Phi^{-1}(\mathcal{M})$.

Proof. An immediate consequence of transversality is that the set $\Phi^{-1}(\mathcal{M})$ is a manifold around any point $z \in \Phi^{-1}(\mathcal{M})$ close to z_0, with normal space

$$N_{\Phi^{-1}(\mathcal{M})}(z) = \nabla\Phi(z)^* N_{\mathcal{M}}(\Phi(z)),$$

and transversality also holds at all such z.

Given a smooth representative g of $h|_{\mathcal{M}}$ around $\Phi(z_0)$, it is easy to see that $g \circ \Phi$ is a smooth representative of $(h \circ \Phi)|_{\Phi^{-1}(\mathcal{M})}$ around z_0, so this latter function is smooth around z_0.

Consider any point $z \in \Phi^{-1}(\mathcal{M})$ close to z_0. By inclusion (2.6) we know

$$\partial^\infty h(\Phi(z)) \subset N_{\mathcal{M}}(\Phi(z)). \tag{4.3}$$

Transversality at z therefore implies

$$\mathrm{Ker}(\nabla\Phi(z)^*) \cap \partial^\infty h(\Phi(z)) = \{0\},$$

so by [23, Thm. 10.6], $h \circ \Phi$ is regular at z, with subdifferential

$$\partial(h \circ \Phi)(z) = \nabla\Phi(z)^* \partial h(\Phi(z)) \neq \emptyset. \tag{4.4}$$

Now, the normal space is parallel to the subdifferential, since

$$\begin{aligned} &\mathrm{par}\,(\partial(h \circ \Phi)(z_0)) = \mathrm{par}\,(\nabla\Phi(z_0)^* \partial h(\Phi(z_0))) \\ &= \nabla\Phi(z_0)^* \mathrm{par}\,(\partial h(\Phi(z_0))) \supset \nabla\Phi(z_0)^* N_{\mathcal{M}}(\Phi(z_0)) = N_{\Phi^{-1}(\mathcal{M})}(z_0), \end{aligned}$$

so it remains only to check the inner semicontinuity property of the subdifferential.

Consider therefore a convergent sequence of points $z_r \to z_0$ in $\Phi^{-1}(\mathcal{M})$, and a subgradient $w \in \partial(h \circ \Phi)(z_0)$. By (4.4) there is a subgradient $y \in \partial h(\Phi(z_0))$ such that $\nabla\Phi(z_0)^* y = w$. Since $\Phi(z_r) \to \Phi(z_0)$ in $\mathcal{M}$ and ∂h is continuous on $\mathcal{M}$ at $\Phi(z_0)$, there must be subgradients $y_r \in \partial h(\Phi(z_r))$ approaching y. But Φ is smooth, so the vectors $\nabla\Phi(z_r)^* y_r \in \partial(h \circ \Phi)(z_r)$ approach w, as required. □

For example, suppose $\Phi(z_0) = 0$ and $\nabla\Phi(z_0)$ is surjective. Then the function $z \mapsto \|\Phi(z)\|$ is partly smooth at z_0 relative to $\Phi^{-1}(0)$.

By applying this result with $h = \delta_S$, we obtain conditions guaranteeing that the set $\Phi^{-1}(S)$ is partly smooth if the set S is smooth.

PROPOSITION 4.5 (separability). *For each $i = 1, 2, \ldots, k$, suppose that X_i is a Euclidean space, that the set $\mathcal{M}_i \subset X_i$ contains the point x_i^0, and that the function $h_i : X_i \to \overline{\mathbf{R}}$ is partly smooth at x_i^0 relative to $\mathcal{M}_i$. Then the function $h : X_1 \times X_2 \times \cdots \times X_k \to \overline{\mathbf{R}}$ defined by*

$$h(x_1, x_2, \ldots, x_k) = \sum_{i=1}^k h_i(x_i) \quad \textit{for } x_i \in X_i, \quad i = 1, 2, \ldots, k,$$

is partly smooth at $(x_1^0, x_2^0, \ldots, x_k^0)$ relative to $\mathcal{M}_1 \times \mathcal{M}_2 \times \cdots \times \mathcal{M}_k$.

Proof. This follows easily from the facts that $\mathcal{M}_1 \times \mathcal{M}_2 \times \cdots \times \mathcal{M}_k$ is a manifold around $(x_1^0, x_2^0, \ldots, x_k^0)$, with normal space

$$N_{\mathcal{M}_1 \times \mathcal{M}_2 \times \cdots \times \mathcal{M}_k}(x_1, x_2, \ldots, x_k) = N_{\mathcal{M}_1}(x_1) \times N_{\mathcal{M}_2}(x_2) \times \cdots \times N_{\mathcal{M}_k}(x_k),$$

and

$$\partial h(x_1, x_2, \ldots, x_k) = \partial h_1(x_1) \times \partial h_2(x_2) \times \cdots \times \partial h_k(x_k),$$

with regularity providing each h_i is regular at x_i [23, Prop. 10.5]. □

For example, the function

$$(x_1, x_2, \ldots, x_k) \mapsto \|x_1\| + \|x_2\| + \cdots + \|x_k\|$$

is partly smooth at the origin relative to the origin.

Applying this result to indicator functions shows that direct products of partly smooth sets are partly smooth.

COROLLARY 4.6 (sum rule). *Consider sets $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_k$ in a Euclidean space Z. Suppose the function $h_i : Z \to \overline{\mathbf{R}}$ is partly smooth at the point z_0 relative to $\mathcal{M}_i$ for each i. Assume furthermore the condition*

$$\sum_{i=1}^{k} y_i = 0 \quad \textit{and} \quad y_i \in N_{\mathcal{M}_i}(z_0) \quad \textit{for each } i \quad \Rightarrow \quad y_i = 0 \quad \textit{for each } i.$$

Then the function $\sum_i h_i$ is partly smooth at z_0 relative to $\cap_i \mathcal{M}_i$.

Proof. We apply the chain rule (Theorem 4.2) and Proposition 4.5 (separability) with

$$\begin{aligned} X &= Z \times Z \times \cdots \times Z \quad (k \text{ copies}), \\ W &= Z, \\ \Phi(z) &= (z, z, \ldots, z) \quad \text{for } z \in Z, \\ \mathcal{M} &= \mathcal{M}_1 \times \mathcal{M}_2 \times \cdots \times \mathcal{M}_k, \\ h(z_1, z_2, \ldots, z_k) &= \sum_i h_i(z_i) \quad \text{for } z_i \in Z,\ i = 1, 2, \ldots, k. \end{aligned}$$

□

Applying this result to indicator functions gives conditions guaranteeing that intersections of partly smooth sets are partly smooth.

COROLLARY 4.7 (smooth perturbation). *If the function $h : X \to \overline{\mathbf{R}}$ is partly smooth at the point x_0 relative to the set $\mathcal{M} \subset X$ and the function $f : X \to \overline{\mathbf{R}}$ is smooth on an open set containing x_0, then the function $h + f$ is partly smooth at x_0 relative to $\mathcal{M}$.*

COROLLARY 4.8 (smooth max function). *Suppose W is an open subset of the Euclidean space Z, and the function $\Phi : W \to \mathbf{R}^n$ is smooth. For any point $z \in W$, define the "active set"*

$$J(z) = \left\{ i : \Phi_i(z) = \max_j \Phi_j(z) \right\}.$$

If the point $z_0 \in W$ satisfies

$$\{\nabla \Phi_i(z_0) : i \in J(z_0)\} \quad \textit{linearly independent}, \tag{4.9}$$

then the function $h : W \to \mathbf{R}$ defined by $h(z) = \max_j \Phi_j(z)$ is partly smooth at z_0 relative to the set

$$\mathcal{M} = \{z \in W : J(z) = J(z_0)\}.$$

Proof. We apply the chain rule (Theorem 4.2) with $X = \mathbf{R}^n$, $\mathcal{M} = \mathcal{M}_{\Phi(z_0)}$ as in (3.5), and $h = \mathrm{mx}$, the basic max function of Example 3.4. The transversality condition follows easily from condition (4.9). □

To apply the idea of partial smoothness to optimization problems with constraints, we need conditions to recognize partly smooth level sets. That is the aim of the last result of this section.

THEOREM 4.10 (level sets). *Consider a point x_0 in a set $\mathcal{M} \subset X$. Suppose that the function $h : X \to \overline{\mathbf{R}}$ is partly smooth at x_0 relative to $\mathcal{M}$, and that x_0 is not a critical point of $h|_{\mathcal{M}}$. Then the level set*

$$L = \{x \in \mathcal{M} : h(x) \leq 0\}$$

is partly smooth at x_0 relative to the set

$$\mathcal{M}_0 = \{x \in \mathcal{M} : h(x) = 0\}.$$

Proof. We can choose an open neighborhood V of x_0, and smooth functions $g : V \to \mathbf{R}$ and $F : V \to \mathbf{R}^m$, such that g agrees with h on the set

$$\mathcal{M} \cap V = \{x \in V : F(x) = 0\}$$

and F has surjective derivative throughout V. If we choose a sufficiently small neighborhood V, then the set

$$\{x \in \mathcal{M} \cap V : h(x) = 0\} \;=\; \{x \in V : F(x) = 0 \text{ and } g(x) = 0\} \tag{4.11}$$

is a manifold around x_0 since

$$\nabla g(x_0) \notin N_{\mathcal{M}}(x_0) = R(\nabla F(x_0)^*).$$

Thus $\mathcal{M}_0$ is indeed a manifold around x_0.

We now need to check the four conditions of Proposition 2.11 (partly smooth sets). Clearly property (i) holds, since $\mathcal{M}_0 \subset \mathcal{M}$.

The assumption that x_0 is not a critical point of $h|_{\mathcal{M}}$ is equivalent to $0 \notin \operatorname{aff} \partial(x_0)$, by Proposition 2.4, so in particular we know $0 \notin \partial h(x_0)$. Since the subdifferential mapping ∂h is continuous relative to $\mathcal{M}$, it follows that $0 \notin \partial h(x)$ for all points $x \in \mathcal{M}$ close to x_0. (In fact this follows just from outer semicontinuity.)

Now consider a point $x \in \mathcal{M}_0$ close to x_0. Notice that h is regular at x_0 and thus locally lower semicontinuous. We can apply [23, Prop. 10.3] to deduce that the level set L is Clarke regular at x (which proves property (ii)), and

$$N_L(x) = (\mathbf{R}_+ \partial h(x)) \cup \partial^\infty h(x).$$

Notice that the right-hand side is closed (since the normal cone is always closed), and it contains $\mathbf{R}_+ \partial h(x)$ and hence also $\operatorname{cl} \mathbf{R}_+ \partial h(x)$. On the other hand, by regularity we have

$$\partial^\infty h(x) = \partial h(x)^\infty \subset \operatorname{cl} \mathbf{R}_+ \partial h(x).$$

Putting these observations together, we deduce the representation

$$N_L(x) = \operatorname{cl} \mathbf{R}_+ \partial h(x). \tag{4.12}$$

By (4.11) we have

$$N_{\mathcal{M}_0}(x_0) = N_{\mathcal{M}}(x_0) + \mathbf{R}\nabla g(x_0),$$

and since h is partly smooth at x_0 relative to $\mathcal{M}$ we also know

$$N_{\mathcal{M}}(x_0) = \mathbf{R}_+(\partial h(x_0) - \partial h(x_0)).$$

Furthermore, Proposition 2.4 implies

$$\nabla g(x_0) \in \operatorname{aff} \partial h(x_0) = \partial h(x_0) + \mathbf{R}_+(\partial h(x_0) - \partial h(x_0)).$$

Hence certainly we have

$$N_{\mathcal{M}_0}(x_0) \subset \mathbf{R}_+\partial h(x_0) - \mathbf{R}_+\partial h(x_0) \subset N_L(x_0) - N_L(x_0),$$

which proves property (iii).

It remains to prove that the normal cone mapping $N_L(\cdot)$ is inner semicontinuous at x_0 relative to $\mathcal{M}_0$, or in other words

$$N_L(x_0) \subset \liminf_{x \to x_0,\ x \in \mathcal{M}_0} N_L(x).$$

Using (4.12) we can rewrite this as

$$\operatorname{cl} \mathbf{R}_+\partial h(x_0) \subset \liminf_{x \to x_0,\ x \in \mathcal{M}_0} \operatorname{cl} \mathbf{R}_+\partial h(x).$$

Since the lim inf is always closed, it suffices to prove

$$\mathbf{R}_+\partial h(x_0) \subset \liminf_{x \to x_0,\ x \in \mathcal{M}_0} \mathbf{R}_+\partial h(x).$$

To this end, suppose that the sequence of points $x_r \in \mathcal{M}_0$ converges to x_0, and consider a vector $y = \mu z$ for some real $\mu \geq 0$ and subgradient $z \in \partial h(x_0)$. Since the subdifferential map ∂h is continuous at x_0 relative to $\mathcal{M}$, there exist subgradients $z_r \in \partial h(x_r)$ approaching z, and then we have vectors $\mu z_r \in \mathbf{R}_+\partial h(x_r)$ approaching y as required. □

COROLLARY 4.13 (smooth constraints). *Suppose W is an open subset of the Euclidean space Z, and the function $\Phi : W \to \mathbf{R}^n$ is smooth. For any point z in the set*

$$L = \{z \in W : \Phi(z) \leq 0\},$$

define the "active set"

$$K(z) = \{k : \Phi_k(z) = 0\}.$$

If the point $z_0 \in L$ satisfies the condition

$$\{\nabla \Phi_k(z_0) : k \in K(z_0)\} \quad \textit{linearly independent},$$

then the set L is partly smooth at z_0 relative to the set

$$\{z \in W : K(z) = K(z_0)\}.$$

Proof. We apply Theorem 4.10 (level sets) to the smooth max function h defined in Corollary 4.8. Notice

$$\partial h(z_0) = \operatorname{conv}\{\nabla \Phi_k(z_0) : k \in K(z_0)\},$$

so $0 \notin \operatorname{aff} \partial h(z_0)$ by the linear independence assumption, and hence z_0 is not a critical point of $h|_{\mathcal{M}}$ for the set $\mathcal{M}$ defined in Corollary 4.8. □

Example 4.14 (semidefinite cone). The convex cone $\mathbf{S}^n_-$ of negative semidefinite matrices is partly smooth relative to the manifold

$$\{x \in \mathbf{S}^n_- : \operatorname{rank} x = k\}$$

for any integer $k = 0, 1, \ldots, n$. To see this, we simply apply Theorem 4.10 to the largest eigenvalue.

Example 4.15 (semistable matrices). A matrix $x \in \mathbf{M}^n$ is *semistable* if all its eigenvalues lie in the closed left half-plane, or in other words, with the notation of Example 3.7 (spectral abscissa), if $\alpha(x) \leq 0$. The (nonconvex) cone of semistable matrices is partly smooth relative to the manifold

$$\{x \in \mathcal{M}_\phi : \alpha(x) = 0\}$$

for any list of multiplicities ϕ. To see this, we apply Theorem 4.10 to the spectral abscissa, using the fact that any subgradient of the spectral abscissa at any point has trace one.

5. Sensitivity. This section considers the stability of critical points of parametric partly smooth functions. Throughout this section we make the following assumption.

Assumption 5.1 (transversal embedding). For Euclidean spaces Y and Z, the set $\mathcal{Q} \subset Y \times Z$ is a manifold containing the point (y_0, z_0) and satisfies the condition

$$(w, 0) \in N_{\mathcal{Q}}(y_0, z_0) \quad \Rightarrow \quad w = 0.$$

Notice that this assumption is "stable": if it holds at the point (y_0, z_0), then it also holds at all nearby points in $\mathcal{Q}$.

For each vector $y \in Y$ we define the set

$$\mathcal{Q}_y = \{z \in Z : (y, z) \in \mathcal{Q}\}.$$

Since the condition in Assumption 5.1 is exactly the transversality condition (4.1) for the map $\Phi : Z \to Y \times Z$ defined by $\Phi(z) = (y_0, z)$, the set $\mathcal{Q}_{y_0}$ is a manifold around z_0. In fact the following result, whose proof is immediate, shows that rather more is true: providing y is close to y_0, the set $\mathcal{Q}_y$ has the structure of a manifold close to z_0.

PROPOSITION 5.2. *If Assumption* 5.1 *holds, then there is an open neighborhood U of z_0 such that for all vectors $y \in Y$ close to y_0 the set $\mathcal{Q}_y \cap U$ is a manifold.*

Throughout this section we consider a function $p : Y \times Z \to \overline{\mathbf{R}}$, and we define a function $p_y : Z \to \overline{\mathbf{R}}$ by

$$p_y(z) = p(y, z) \quad \text{for } y \in Y \text{ and } z \in Z.$$

Clearly if the restriction $p|_{\mathcal{Q}}$ is smooth, then so is the restriction $p_y|_{\mathcal{Q}_y}$. The next result shows an analogous property for partial smoothness.

PROPOSITION 5.3 (partial smoothness with parameters). *Suppose Assumption* 5.1 *holds and the function p is partly smooth relative to the manifold $\mathcal{Q}$. Then there is an open neighborhood U of the point z_0 such that the function p_y is partly smooth relative to $\mathcal{Q}_y \cap U$ for all vectors $y \in Y$ close to y_0.*

Proof. There are open neighborhoods U of z_0 and V of y_0 such that $\mathcal{Q}_y \cap U$ is a manifold for all $y \in V$ and

$$y \in V,\ z \in U,\ (y,z) \in \mathcal{Q},\ (w,0) \in N_{\mathcal{Q}}(y,z) \quad \Rightarrow \quad w = 0.$$

Hence for any points $\hat{y} \in V$ and $\hat{z} \in \mathcal{Q}_{\hat{y}} \cap U$ we can apply the chain rule (Theorem 4.2) at $\hat{z}$ with the map $\Phi : Z \to Y \times Z$ defined by $\Phi(z) = (\hat{y}, z)$ to deduce that the function $p_{\hat{y}} = p \circ \Phi$ is partly smooth at $\hat{z}$ relative to the manifold $\mathcal{Q}_{\hat{y}} \cap U$. □

Our main aim in this work is to study sensitivity of critical points for partly smooth functions. Just as in classical sensitivity analysis for nonlinear programming, we need second-order conditions to make progress.

DEFINITION 5.4. *Given any subset $\mathcal{M}$ of a Euclidean space X, a point x_0 is a* strong local minimizer *of a function $f : \mathcal{M} \to \overline{\mathbf{R}}$ if there exists a real $\delta > 0$ such that $f(x) \geq f(x_0) + \delta\|x - x_0\|^2$ for all $x \in \mathcal{M}$ near x_0.*

We recall some classical sensitivity analysis (see, for example, [8]). Suppose $\mathcal{M} \subset X$ is a manifold around the point $x_0 \in \mathcal{M}$, and the restriction $h|_{\mathcal{M}}$ is smooth around x_0, for some function $h : X \to \overline{\mathbf{R}}$. Let g be any smooth representative of $h|_{\mathcal{M}}$. By definition, x_0 has an open neighborhood $V \subset X$ such that

$$\mathcal{M} \cap V = \{x \in V : F(x) = 0\}$$

for some smooth function $F : V \to \mathbf{R}^m$ with $\nabla F(x_0)$ surjective. The point x_0 is a critical point of $h|_{\mathcal{M}}$ if and only if $\nabla g(x_0) \in N_{\mathcal{M}}(x_0)$, which is equivalent to the existence of a multiplier vector $\mu \in \mathbf{R}^m$ (necessarily unique) such that x_0 is a critical point of the corresponding Lagrangian function $L = g + \mu^T F$. Furthermore, x_0 is a strong local minimizer of $h|_{\mathcal{M}}$ if and only if it is a critical point of $h|_{\mathcal{M}}$ and satisfies the second-order condition

$$y^T \nabla^2 L(x_0) y > 0 \ \text{ whenever } 0 \neq y \in \operatorname{Ker}(\nabla F(x_0)).$$

The following result is also classical.

THEOREM 5.5 (parametric strong minimizers). *Suppose that the function $p|_{\mathcal{Q}}$ is smooth around the point (y_0, z_0), that the point z_0 is a strong local minimizer of the function $p_{y_0}|_{\mathcal{Q}_{y_0}}$, and that Assumption* 5.1 *holds. Then there are open neighborhoods $U \subset Z$ of z_0 and $V \subset Y$ of y_0 and a continuously differentiable function $\Psi : V \to U$ such that $\Psi(y_0) = z_0$ and for all vectors $y \in V$ the function $p_y|_{\mathcal{Q}_y \cap U}$ has a unique critical point $\Psi(y)$, which is furthermore a strong local minimizer.*

To approach a more complete sensitivity theory, we combine the smooth analysis of a partly smooth function on its active manifold with a study of its behavior in normal directions. That is the idea of the following definition.

DEFINITION 5.6 (strong critical point). *For a Euclidean space X, suppose the function $h : X \to \overline{\mathbf{R}}$ is partly smooth at the point x_0 relative to the set $\mathcal{M} \subset X$. We call x_0 a* strong critical point *of h relative to $\mathcal{M}$ if*

(i) *x_0 is a strong local minimizer of $h|_{\mathcal{M}}$, and*

(ii) $0 \in \operatorname{ri} \partial h(x_0)$.

In the next section we see that the condition $0 \in \operatorname{ri} \partial h(x_0)$ could be written equivalently as x_0 being a "sharp" local minimizer of the function $h|_{x_0 + N_{\mathcal{M}}(x_0)}$.

We are now ready for the main result. Comparing it with the classical result Theorem 5.5 above, we see that the extra assumption of strong criticality implies that the parametrized minimizer is also a strong critical point.

THEOREM 5.7 (strong critical points and parameters). *Suppose Assumption* 5.1 *holds and the function p is partly smooth relative to the manifold $\mathcal{Q}$. If the point z_0 is a strong critical point of the function p_{y_0} relative to the set $\mathcal{Q}_{y_0}$, then there are open neighborhoods $U \subset Z$ of z_0 and $V \subset Y$ of y_0 and a continuously differentiable function $\Psi : V \to U$ satisfying $\Psi(y_0) = z_0$ and with the following properties for all vectors $y \in V$:*

(i) *the function $p_y|_{\mathcal{Q}_y \cap U}$ has a unique critical point $\Psi(y)$;*

(ii) *$\Psi(y)$ is a strong critical point of the function p_y relative to the manifold $\mathcal{Q}_y \cap U$.*

Proof. Theorem 5.5 shows the existence of a function Ψ having the required properties, with the exception of property (ii). Proposition 5.3 shows that p_y is partly smooth relative to the manifold $\mathcal{Q}_y \cap U$. Hence to prove property (ii), it suffices to show

$$0 \in \operatorname{ri} \partial p_y(\Psi(y)) \quad \text{for } y \in V \text{ close to } y_0.$$

To this end, as in the proof of Proposition 5.3, we define a map $\Phi_y : Z \to Y \times Z$ by $\Phi_y(z) = (y, z)$ for $z \in Z$, observe that $p_y = p \circ \Phi_y$, and note that Assumption 5.1 allows us to apply the chain rule (Theorem 4.2). By (4.4) we deduce

$$\partial p_y(\Psi(y)) = \operatorname{proj}_Z \partial p(y, \Psi(y)),$$

where $\operatorname{proj}_Z : Y \times Z \to Z$ is the natural projection, whereas a standard calculation shows

$$N_{\mathcal{Q}_y}(\Psi(y)) = \operatorname{proj}_Z N_{\mathcal{Q}}(y, \Psi(y)).$$

We therefore know

$$0 \in \operatorname{ri}(\operatorname{proj}_Z \partial p(y_0, \Psi(y_0))), \tag{5.8}$$

and we want to deduce

$$0 \in \operatorname{ri}(\operatorname{proj}_Z \partial p(y, \Psi(y))) \quad \text{for all } y \text{ close to } y_0.$$

Notice that, by definition, we know $\Psi(y)$ is a critical point of the restriction $p_y|_{\mathcal{Q}_y}$ for y close to y_0, so by partial smoothness, Proposition 2.10 (local normal sharpness) and Proposition 2.4 (smoothness and lineality) we have

$$\operatorname{aff}(\operatorname{proj}_Z \partial p(y, \Psi(y))) = \operatorname{aff} \partial p_y(\Psi(y)) = N_{\mathcal{Q}_y}(\Psi(y)) = \operatorname{proj}_Z N_{\mathcal{Q}}(y, \Psi(y)).$$

If the result fails, then there is a sequence of vectors y_r in Y approaching y_0 such that

$$0 \notin \operatorname{ri}(\operatorname{proj}_Z \partial p(y_r, \Psi(y_r))) \quad \text{for all } r.$$

For all large r we can separate in the subspace $\operatorname{proj}_Z N_{\mathcal{Q}}(y_r, \Psi(y_r))$ to deduce the existence of a unit vector z_r in this subspace, satisfying

$$\inf \langle z_r, \operatorname{proj}_Z \partial p(y_r, \Psi(y_r)) \rangle \geq 0.$$

After taking a subsequence, we can assume z_r approaches a nonzero vector $z \in Z$.

Now, since the point $(y_r, \Psi(y_r))$ converges to the point (y_0, z_0) in the manifold $\mathcal{Q}$, it follows that the subspace $N_{\mathcal{Q}}(y_r, \Psi(y_r))$ converges to the subspace $N_{\mathcal{Q}}(y_0, z_0)$, so Assumption 5.1 implies that the subspace $\operatorname{proj}_Z N_{\mathcal{Q}}(y_r, \Psi(y_r))$ converges to the subspace $\operatorname{proj}_Z N_{\mathcal{Q}}(y_0, z_0)$, by [23, Ex. 4.28]. Hence we deduce

$$z \in \operatorname{proj}_Z N_{\mathcal{Q}}(y_0, z_0).$$

We now claim

$$\inf \langle z, \operatorname{proj}_Z \partial p(y_0, z_0) \rangle \geq 0. \tag{5.9}$$

To see this, consider any vector $u \in \operatorname{proj}_Z \partial p(y_0, z_0)$. Partial smoothness implies that $\partial p(y_r, \Psi(y_r))$ converges to $\partial p(y_0, z_0)$, so again by [23, Ex. 4.28] we deduce that $\operatorname{proj}_Z \partial p(y_r, \Psi(y_r))$ converges to $\operatorname{proj}_Z \partial p(y_0, z_0)$. Thus there is a sequence of vectors $u_r \in \operatorname{proj}_Z \partial p(y_r, \Psi(y_r))$ converging to u. Since $\langle z_r, u_r \rangle \geq 0$ for all r, we deduce $\langle z, u \rangle \geq 0$, as we claimed.

Thus inequality (5.9) holds, so the origin is separated from the convex set $\operatorname{proj}_Z \partial p(y_0, z_0)$ in its affine span (the subspace $\operatorname{proj}_Z N_{\mathcal{Q}}(y_0, z_0)$). But this contradicts relation (5.8), so the proof is complete. □

6. $\mathcal{U} - \mathcal{V}$ decomposition and identifiable surfaces. As we remarked in the introduction, our development is closely related to the $\mathcal{U}$-Lagrangian theory for convex functions of Lemaréchal, Oustry, and Sagastizábal (see, for example, [12]). The key idea of that theory is, for a given *convex* function $h : X \to \overline{\mathbf{R}}$, to decompose X as a sum of two orthogonal subspaces, $\mathcal{U}$ and $\mathcal{V}$: h behaves "sharply" at the point of interest if we perturb in directions in the $\mathcal{V}$ space, whereas it behaves smoothly if we perturb in directions in the $\mathcal{U}$ space.

Our purpose in this section is to draw the connection between this idea and partial smoothness. The development is a nice illustration of various features of the theory of partial smoothness.

We call a local minimizer x of an arbitrary function $h : X \to \overline{\mathbf{R}}$ *sharp* if

$$\liminf_{z \to 0} \frac{h(x+z) - h(x)}{\|z\|} > 0,$$

or equivalently, if $0 \in \operatorname{int} \hat{\partial} h(x)$.

THEOREM 6.1 ($\mathcal{U} - \mathcal{V}$ decomposition). *Suppose the function $h : X \to \overline{\mathbf{R}}$ is partly smooth at the point x relative to the set $\mathcal{M} \subset X$. Define subspaces $\mathcal{U} = T_{\mathcal{M}}(x)$ and $\mathcal{V} = N_{\mathcal{M}}(x)$. Then there exists a function $v : \mathcal{U} \to \mathcal{V}$ with the following three properties:*

(i) *the function v is smooth near the origin;*
(ii) *for small vectors $u \in \mathcal{U}$ and $w \in \mathcal{V}$, $x + u + w \in \mathcal{M} \Leftrightarrow w = v(u)$;*
(iii) *$v(u) = O(\|u\|^2)$ for small $u \in \mathcal{U}$.*

Fix any vector $y \in \operatorname{ri} \partial h(x)$. Then for any small vector $u \in \mathcal{U}$, the function

$$w \in \mathcal{V} \mapsto h(x + u + w) - \langle y, x + u + w \rangle \tag{6.2}$$

has a sharp minimizer at the point $v(u)$.

Furthermore, the point x is a strong critical point of h relative to $\mathcal{M}$ if and only if it is a strong local minimizer of $h|_{\mathcal{M}}$ and a sharp local minimizer of $h|_{x+\mathcal{V}}$.

Proof. By first intersecting with an open set, we can assume $\mathcal{M}$ is a manifold. A standard argument using the implicit function theorem shows the existence of the function v with the properties (i), (ii), and (iii).

Define a map $\Phi : \mathcal{U} \times \mathcal{V} \to X$ by

$$\Phi(u, w) = x + u + w \quad \text{for } u \in \mathcal{U} \text{ and } w \in \mathcal{V}.$$

Clearly Φ is everywhere transversal to $\mathcal{M}$. Hence by the chain rule (Theorem 4.2), the function $h \circ \Phi$ is partly smooth relative to the manifold $\Phi^{-1}(\mathcal{M})$. Consequently, by smooth perturbation (Corollary 4.7) the function $p : \mathcal{U} \times \mathcal{V} \to \overline{\mathbf{R}}$ defined by

$$p(u, w) = h(x + u + w) - \langle y, x + u + w \rangle \quad \text{for } u \in \mathcal{U} \text{ and } w \in \mathcal{V}$$

is partly smooth relative to the manifold

$$\mathcal{Q} = \Phi^{-1}(\mathcal{M}) = \{(u, w) \in \mathcal{U} \times \mathcal{V} : x + u + w \in \mathcal{M}\}.$$

Notice that for a small vector $u \in \mathcal{U}$, the function (6.2) is exactly p_u, and property (ii) shows

$$\mathcal{Q}_u = \{v(u)\}.$$

It is easy to check

$$N_{\mathcal{Q}}(0, 0) = \{0\} \times \mathcal{V},$$

so Assumption 5.1 holds for our function p. Hence we can apply Theorem 5.7 (strong critical points and parameters) to deduce that $v(u)$ is a strong critical point of p_u relative to $\{v(u)\}$. Hence

$$0 \in \operatorname{ri} \partial p_u(v(u)) = \operatorname{int} \partial p_u(v(u)),$$

since $\operatorname{aff} \partial p_u(v(u)) = N_{\mathcal{Q}_u}(v(u)) = \mathcal{V}$.

To see the "only if" direction of the last statement, we simply consider the function (6.2) with $y = 0$ and $u = 0$. In the converse direction, since x is a local minimizer of $h|_{\mathcal{M}}$, we know $\operatorname{aff} \partial h(x) = \mathcal{V}$, by Proposition 2.4 (smoothness and lineality), and since the origin is a sharp local minimizer of the function p_0, we deduce $0 \in \operatorname{int} \partial p_0(0) = \operatorname{int} \operatorname{proj}_{\mathcal{V}} \partial h(x)$, just like the proof of Theorem 5.7. It follows that $0 \in \operatorname{ri} \partial h(x)$. □

The spaces $\mathcal{U}$ and $\mathcal{V}$ in the above result coincide with those in [12] in the convex case.

The idea of partial smoothness is also closely related to the notion of an *identifiable surface* [26] of a convex set. Given a closed convex set $S \subset X$, we call a connected manifold $\mathcal{M} \subset S$ a *(class-C^2) identifiable surface* if either $\mathcal{M}$ is open or for every point $x_0 \in \mathcal{M}$ and every vector $w_0 \in \operatorname{ri} N_S(x_0)$ there exists an open set $V \subset X$ containing x_0 and a smooth function $F : V \to \mathbf{R}^m$ (where m is the codimension of $\mathcal{M}$) such that ∇F is everywhere surjective, $\mathcal{M} \cap V = F^{-1}(0)$, $\nabla F(x)^* \mathbf{R}^m_+ \subset N_S(x)$ for all points $x \in \mathcal{M} \cap V$, and $w_0 \in \nabla F(x_0)^* \mathbf{R}^m_{++}$ (where $\mathbf{R}^m_{++} = \operatorname{int} \mathbf{R}^m_+$).

THEOREM 6.3 (identifiable surfaces). *Consider a closed convex set $S \subset X$ and a connected manifold $\mathcal{M} \subset S$. Then S is partly smooth relative to $\mathcal{M}$ if and only if $\mathcal{M}$ is an identifiable surface.*

Proof. The case when $\mathcal{M}$ is open is immediate, so assume $\mathcal{M}$ has codimension $m > 0$.

Suppose first that $\mathcal{M}$ is an identifiable surface. We need to check the conditions of Proposition 2.11 (partly smooth sets). Condition (i) is immediate, and condition (ii) holds since closed convex sets are everywhere regular. At any point $x_0 \in \mathcal{M}$ we can choose a vector w_0, a neighborhood V, and a function F as in the definition of an identifiable surface, and then we have

$$N_{\mathcal{M}}(x_0) = R(\nabla F(x_0)^*) = \nabla F(x_0)^*(\mathbf{R}^m_+ - \mathbf{R}^m_+) \subset N_S(x_0) - N_S(x_0),$$

so condition (iii) holds.

It remains to show that the normal cone mapping N_S is inner semicontinuous at x_0 relative to $\mathcal{M}$. Since $N_S(x_0)$ is the closure of its relative interior, it suffices to show, for our arbitrary choice $w_0 \in \operatorname{ri} N_S(x_0)$, that for any sequence $\{x_r\} \subset \mathcal{M}$ converging to x_0, there exist vectors $w_r \in N_S(x_r)$ converging to w_0. But since $w_0 \in \nabla F(x_0)^*\mathbf{R}^m_{++}$, there exists a vector $\mu \in \mathbf{R}^m_{++}$ such that $w_0 = \nabla F(x_0)^*\mu$, and then the vector

$$w_r = \nabla F(x_r)^*\mu \in \nabla F(x_r)^*\mathbf{R}^m_+$$

lies in $N_S(x_r)$ for all large r and converges to w_0, as required.

Conversely, suppose that the set S is partly smooth relative to the manifold $\mathcal{M}$, and consider a point $x_0 \in \mathcal{M}$ and a vector $w_0 \in \operatorname{ri} N_S(x_0)$. By Proposition 2.10 (local normal sharpness) we know $N_{\mathcal{M}}(x) = N_S(x) - N_S(x)$ for all points $x \in \mathcal{M}$ close to x_0, and hence the closed convex cone $N_S(x)$ has the same dimension as the subspace $N_{\mathcal{M}}(x)$, namely m. Thus there exist linearly independent vectors $w_1, w_2, \ldots, w_m \in \operatorname{ri} N_S(x_0)$ such that

$$w_0 \in \operatorname{ri}(\operatorname{conv}\{w_1, w_2, \ldots, w_m\}).$$

Since $\mathcal{M}$ is a manifold of codimension m around x_0, there exists an open set $V \subset X$ containing x_0 and a smooth function $G : V \to \mathbf{R}^m$ such that ∇G is everywhere surjective and $\mathcal{M} \cap V = F^{-1}(0)$. Hence for all points $x \in \mathcal{M} \cap V$ we have $N_{\mathcal{M}}(x) = R(\nabla G(x)^*)$. Since $\nabla G(x)^*$ is injective for all points $x \in V$, there exists a basis $\{a^1, a^2, \ldots, a^m\}$ of $\mathbf{R}^m$ satisfying

$$\nabla G(x_0)^* a^j = w_j \quad \text{for } j = 1, 2, \ldots, m.$$

Now the function $F : V \to \mathbf{R}^m$ defined by

$$(F(x))_j = \langle a^j, G(x) \rangle \quad \text{for } x \in V,\ j = 1, 2, \ldots, m,$$

satisfies $F^{-1}(0) = G^{-1}(0) = \mathcal{M} \cap V$ and

$$\nabla F(x)^* e^j = \nabla G(x)^* a^j \quad \text{for } x \in V,\ j = 1, 2, \ldots, m$$

(where $e^j \in \mathbf{R}^m$ denotes the jth unit vector). Thus $\nabla F(x)^*$ is injective, and so $\nabla F(x)$ is surjective for all points $x \in V$. Also,

$$\nabla F(x_0)^* e^j = \nabla G(x_0)^* a^j = w_j \quad \text{for } j = 1, 2, \ldots, m,$$

so $w_0 \in \nabla F(x_0)^*\mathbf{R}^m_{++}$, as required, and furthermore,

$$N_{\mathcal{M}}(x) = R(\nabla F(x)^*) \quad \text{for } x \in \mathcal{M} \cap V.$$

It remains to prove $\nabla F(x)^*\mathbf{R}^m_+ \subset N_S(x)$ for all points $x \in \mathcal{M}$ close to x_0. If this fails, then for some index j there is a sequence $\{x_r\} \subset \mathcal{M}$ approaching x_0 such that

$$\nabla F(x_r)^* e^j \notin N_S(x_r) \quad \text{for all } r.$$

Both the left- and right-hand sides above are contained in the subspace $N_{\mathcal{M}}(x_r)$, so by separating in this subspace, there exists a unit vector $y_r \in N_{\mathcal{M}}(x_r)$ satisfying

$$\langle y_r, \nabla F(x_r)^* e^j \rangle < \langle y_r, v \rangle \quad \text{for all } v \in N_S(x_r),\ r = 1, 2, \ldots.$$

We can assume, after taking a subsequence, that the sequence $\{y_r\}$ converges to some unit vector $y_0 \in N_{\mathcal{M}}(x_0)$, and since $w_j \in \operatorname{ri} N_S(x_0)$, there exists a real $\delta > 0$ such that $w_j - \delta y_0 \in N_S(x_0)$. Now, since the mapping N_S is continuous, there exist vectors $v_r \in N_S(x_r)$ approaching $w_j - \delta y_0$. But we know

$$\langle y_r, \nabla F(x_r)^* e^j \rangle < \langle y_r, v_r \rangle \quad \text{for } r = 1, 2, \ldots,$$

so taking the limit as $r \to \infty$ gives the contradiction

$$\langle y_0, w_j \rangle \leq \langle y_0, w_j - \delta y_0 \rangle. \qquad \square$$

7. Example. The idea of a strong critical point *decouples* behavior in the active manifold from behavior in directions normal to it. Restricting to the active manifold, a strong critical point is a strong local minimizer, whereas, as we saw in the previous section, any point in the active manifold is a sharp local minimizer with respect to perturbations in normal directions.

One might hope that these properties suffice to ensure that strong critical points of reasonable functions are local minimizers. Unfortunately, this is not the case. We present in this section a locally Lipschitz, everywhere regular function $f : \mathbf{R}^2 \to \mathbf{R}$, partly smooth relative to two distinct manifolds containing the origin. Relative to one manifold, the origin is a strong critical point. However, f restricted to the other manifold has a strong local *maximum* at the origin.

We partition $\mathbf{R}^2$ into four disjoint sets

$$\begin{aligned} S_1 &= \{(x, y) : y \leq 0\}, \\ S_2 &= \{(x, y) : 0 < y < 2x^2\}, \\ S_3 &= \{(x, y) : 0 < 2x^2 \leq y \leq 4x^2\}, \\ S_4 &= \{(x, y) : 4x^2 < y\}, \end{aligned}$$

and we define f by

$$f(x, y) = \begin{cases} x^2 - y & \text{on } S_1, \\ \sqrt{x^4 + 2x^2 y - y^2} & \text{on } S_2, \\ 3x^2 - y & \text{on } S_3, \\ y - 5x^2 & \text{on } S_4. \end{cases}$$

It is easy to check that f is everywhere continuous and in fact is continuously differentiable except on the manifolds

$$\begin{aligned} \mathcal{M}_1 &= \{(x, y) : y = 0\}, \\ \mathcal{M}_2 &= \{(x, y) : y = 4x^2\}. \end{aligned}$$

A calculation shows that $\hat{\partial} f(x,y)$ is given by

$$\begin{cases} \{(2x,-1)\} & \text{on } \operatorname{int} S_1, \\ \big[(2x,-1),(2x,1)\big] & \text{on } \mathcal{M}_1, \\ \left\{\left(1+2\left(\frac{y}{x^2}\right)-\left(\frac{y}{x^2}\right)^2\right)^{-1/2}\left(2x\left(1+\left(\frac{y}{x^2}\right)\right),1-\frac{y}{x^2}\right)\right\} & \text{on } S_2, \\ \{(6x,-1)\} & \text{on } S_3 \setminus \mathcal{M}_2, \\ \big[(6x,-1),(-10x,1)\big] & \text{on } \mathcal{M}_2, \\ \{(-10x,1)\} & \text{on } S_4, \end{cases}$$

where $[u,v]$ denotes the line segment between the points $u, v \in \mathbf{R}^2$. The calculation at every point except the origin is routine, since f is either continuously differentiable at such points or can be written locally as the maximum of two continuously differentiable functions. At the origin we use the inequality

$$|3x^2 - y| - 2x^2 \leq f(x,y) \leq |3x^2 - y| \quad \text{for all } x, y.$$

The map

$$\beta \in [0,2] \mapsto (1 + 2\beta - \beta^2)^{-1/2}(1-\beta)$$

has range the interval $[-1,1]$, so for $x \geq 0$,

$$\nabla f(x,y) \in [2x, 6x] \times [-1,1] \quad \text{on } S_2, \tag{7.1}$$

and a similar relation holds if $x \leq 0$. Hence f is everywhere locally Lipschitz, even around the origin.

We next claim $\partial f = \hat{\partial} f$ everywhere, so f is everywhere regular. As above, this is routine everywhere except at the origin, where it follows using (7.1).

Now it is straightforward to check that the function f is partly smooth relative to both the manifolds $\mathcal{M}_1$ and $\mathcal{M}_2$, and that the origin is a strong critical point relative to $\mathcal{M}_1$. But

$$f(x,y) = -x^2 \quad \text{on } \mathcal{M}_2,$$

so the origin is *not* a local minimizer. In summary, although strong criticality is significant for sensitivity analysis, it is *not* a sufficient condition for optimality.

Acknowledgment. Thanks to Henry Wolkowicz for suggesting the connection between partly smooth sets and identifiable surfaces.

REFERENCES

[1] K. D. Anderson, E. Christiansen, A. R. Conn, and M. L. Overton, *An efficient primal-dual interior-point method for minimizing a sum of Euclidean norms*, SIAM J. Sci. Comput., 22 (2000), pp. 243–262.

[2] V. I. Arnold, *On matrices depending on parameters*, Russian Math. Surveys, 26 (1971), pp. 29–43.

[3] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*, SIAM, Philadelphia, 2001.

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

[5] J. V. Burke, A. S. Lewis, and M. L. Overton, *Optimal stability and eigenvalue multiplicity*, Found. Comput. Math., 1 (2001), pp. 205–225.
[6] P. H. Calamai and A. R. Conn, *A stable algorithm for solving the multifacility location problem involving Euclidean distances*, SIAM J. Sci. Stat. Comput., 1 (1980), pp. 512–526.
[7] P. H. Calamai and A. R. Conn, *A projected Newton method for l_p norm location problems*, Math. Programming, 38 (1987), pp. 75–109.
[8] A. V. Fiacco and G. P. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, Wiley, New York, 1968; reprinted, SIAM, Philadelphia, 1990.
[9] R. Fletcher, *Practical Methods of Optimization*, 2nd ed., Wiley, Chichester, UK, 1987.
[10] R. A. Horn and C. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.
[11] T. Kato, *A Short Introduction to Perturbation Theory for Linear Operators*, Springer, New York, 1982.
[12] C. Lemaréchal, F. Oustry, and C. Sagastizábal, *The $\mathcal{U}$-Lagrangian of a convex function*, Trans. Amer. Math. Soc., 352 (2000), pp. 711–729.
[13] C. Lemaréchal and C. Sagastizábal, *Practical aspects of the Moreau–Yosida regularization: Theoretical preliminaries*, SIAM J. Optim., 7 (1997), pp. 367–385.
[14] A. S. Lewis and M. L. Overton, *Eigenvalue optimization*, Acta Numer., 5 (1996), pp. 149–190.
[15] R. Mifflin and C. Sagastizábal, *Proximal Points Are on the Fast Track*, Technical report, Washington State University, Pullman, WA, 2000.
[16] Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM, Philadelphia, 1994.
[17] M. R. Osborne, *Simplicial Algorithms for Minimizing Polyhedral Functions*, Cambridge University Press, Cambridge, UK, 2001.
[18] F. Oustry, *The $\mathcal{U}$-Lagrangian of the maximum eigenvalue function*, SIAM J. Optim., 9 (1999), pp. 526–549.
[19] F. Oustry, *A second-order bundle method to minimize the maximum eigenvalue function*, Math. Program., 89 (2000), pp. 1–33.
[20] M. L. Overton, *A quadratically convergent method for minimizing a sum of Euclidean norms*, Math. Programming, 27 (1983), pp. 34–63.
[21] M. L. Overton and R. S. Womersley, *Second derivatives for optimizing eigenvalues of symmetric matrices*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 697–718.
[22] M. L. Overton and X. Ye, *Towards second-order methods for structured nonsmooth optimization*, in Advances in Optimization and Numerical Analysis, S. Gomez and J.-P. Hennart, eds., Kluwer, Amsterdam, 1994, pp. 97–109.
[23] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer, Berlin, 1998.
[24] A. Shapiro and M. K. H. Fan, *On eigenvalue optimization*, SIAM J. Optim., 5 (1995), pp. 552–569.
[25] I. Vaisman, *A First Course in Differential Geometry*, Marcel Dekker, New York, 1984.
[26] S. J. Wright, *Identifiable surfaces in constrained optimization*, SIAM J. Control Optim., 31 (1993), pp. 1063–1079.

SIAM J. OPTIM.
Vol. 13, No. 3, pp. 726–744

ON BEST APPROXIMATION BY NONCONVEX SETS AND PERTURBATION OF NONCONVEX INEQUALITY SYSTEMS IN HILBERT SPACES*

CHONG LI† AND K. F. NG‡

Abstract. By virtue of convexification techniques, we study best approximations to a closed set C in a Hilbert space as well as perturbation conditions relative to C and a nonlinear inequality system. Some results on equivalence of the best approximation and the basic constraint qualification are established.

Key words. best approximation, nonlinear constraint, nonlinear inequality system, strong CHIP, the basic constraint qualification condition, generalized Mangasarian–Fromowitz constraint qualification and regularity

AMS subject classifications. Primary, 41A65; Secondary, 41A29

PII. S1052623402401373

1. Introduction. Let X, Y be Hilbert spaces over the real field $\mathbb{R}$ (unless specifically stated otherwise), and let C be a closed convex subset of X. Let K consist of all $x \in C$ satisfying the nonconvex inequality system

$$A_i(x) \leqslant 0 \quad \forall i = 1, 2, \dots, m, \tag{NIS}$$

where each A_i is a composite function of the form $H_i \circ F_i$ with $H_i : Y \to \mathbb{R}$, $F_i : X \to Y$ for each i. We assume throughout that, for each i, H_i is continuous convex and F_i is Fréchet differentiable on X with continuous Fréchet derivative denoted by $F_i'(\cdot)$. In general, A_i is nondifferentiable and nonlinear. For each $x \in X$, let $\partial A_i(x)$ denote the subdifferential of A_i at x. Let $x^* \in K$ and $I(x^*)$ denote the set of all active indices i : $I(x^*) = \{i : \ A_i(x^*) = 0\}$. Let P_C and P_K denote the projection operators from X to C and K, respectively. Because it is generally easier to compute P_C than P_K (noting, in particular, that K is not necessarily convex), we stipulate the following definition: x^* is said to have the perturbation property with respect to C and the above (NIS) if for each $x \in X$,

$$x^* = P_K(x) \Longleftrightarrow x^* = P_C\left(x - \sum_1^m \lambda_i h_i\right) \tag{1.1}$$

for some $h_i \in \partial A_i(x^*)$ and $\lambda_i \geqslant 0$, with $\lambda_i = 0$ for all $i \notin I(x^*)$. Here and throughout, $x^* = P_K(x)$ is read as $x^* \in P_K(x)$ if the operator is multivalued. For the special case in which $Y = \mathbb{R}$ and each A_i is affine, this property has been studied by many authors (see, for example, [2, 4, 5, 9, 11, 17, 18]) and has been shown by Deutsch, Li, and Ward (in [10]) to be equivalent to the strong CHIP (strong conical hull intersection

*Received by the editors January 23, 2002; accepted for publication (in revised form) May 14, 2002; published electronically November 14, 2002.

http://www.siam.org/journals/siopt/13-3/40137.html

†Department of Mathematics, Zhejiang University, Hangzhou 310027, People's Republic of China (cli@seu.edu.cn). This author was supported in part by the Natural Science Foundation of China (grant 19971013).

‡Department of Mathematics, Chinese University of Hong Kong, Hong Kong, People's Republic of China (kfng@math.cuhk.edu.hk). This author was supported by a direct grant (CUHK) and Earmarked Grant from the Research Grant Council of Hong Kong.

property) of $\{C, G_1, G_2, \dots, G_m\}$, where each G_i denotes a half-space defined by A_i. Their result has been extended by Li and Jin in [14] to the cases (a) $X = Y$ and each F_i is the identity mapping and (b) $Y = \mathbb{R}$ and each H_i is the identity mapping. In this paper, we consider the case in which each F_i is a general Fréchet differentiable function and each H_i is a general continuous convex function. For each i, let $\tilde{A}_i$ denote the "convexification" of A_i at x^*. (For a definition, see section 2.) Under a regularity condition (which is automatic in the above case (a)), we show in Theorem 3.7 that x^* has the above perturbation property if and only if the convex inequality system

$$\tilde{A}_i(\cdot) \leqslant 0, \quad i = 1, 2, \dots, m,$$

satisfies the basic constraint qualification (BCQ) relative to C at x^*(which is equivalent to the strong CHIP of the family $\{C; \tilde{A}_i^{-1}(\mathbb{R}_-), i = 1, 2, \dots, m\}$ in the case in which each $\tilde{A}_i$ is affine). This generalizes a main result of [10] and that of [14]. Moreover, in the case of $Y = \mathbb{R}^n$, the regularity condition mentioned above is shown to be implied by a constraint qualification of Mangasarian–Fromowitz type (see Theorem 3.13). In section 4, some applications are made to study the inequality system with respect to an abstract convex cone in a (real or complex) Hilbert space.

2. Notations and preparatory results. The notation used in this paper is standard (see [1, 6, 13, 20]). In particular, for a set Z in X (or in Y or $\mathbb{R}^n$), the interior (resp., relative interior, closure, convex hull, convex cone hull, affine space, linear space, negative polar) of Z is defined by int$Z$ (resp., ri$Z$, $\bar{Z}$, conv$Z$, cone$Z$, aff$Z$, span$Z$, Z°), and the normal cone of Z at $\bar{z}$ is denoted by $N_Z(\bar{z})$ and defined by $N_Z(\bar{z}) = (Z - \bar{z})^\circ$. $\mathbb{R}_-$ denotes the subset of $\mathbb{R}$ consisting of all nonpositive real numbers. For a proper extended real-valued function f on X, the subdifferential of f at $x \in X$ is denoted by $\partial f(x)$ and defined by

$$\partial f(x) = \{z \in X : \ f(x) + \langle z, y - x\rangle \leqslant f(y)\ \forall y \in X\}.$$

In particular, $N_Z(\bar{z}) = \partial \mathbf{I}_Z(\bar{z})$. Here and throughout, $\mathbf{I}_Z$ denotes the indicator function of Z: $\mathbf{I}_Z(x) = 0$ if $x \in Z$, and $\mathbf{I}_Z(x) = +\infty$ if $x \in X \setminus Z$.

Let m, C, K, H_i, F_i, and A_i be as in the preceding section. Let $x^* \in K$ and $I(x^*) = \{i : \ A_i(x^*) = 0\}$. For each i, let $\tilde{A}_i$ denote the "convexification" of A_i at x^* defined by

$$\tilde{A}_i(x) = H_i(F_i(x^*) + F_i'(x^*)(x - x^*)) \quad \forall x \in X. \tag{2.1}$$

Note that $\tilde{A}_i$ is continuous and convex (because H_i is, and because $x \mapsto F_i(x^*) + F_i'(x^*)(x - x^*)$ is affine). Note also that

$$\tilde{A}_i(x^*) = A_i(x^*), \quad i = 1, 2, \dots, m. \tag{2.2}$$

Definition 2.1. *An element $d \in X$ is called*

(a) *a convexification feasible direction of* (NIS) *at x^* if*

$$\tilde{A}_i(x^* + d) \leqslant 0, \quad i \in I(x^*),$$

(b) *a sequentially feasible direction of K at x^* if there exist sequences $\{d_k\} \to d$ and a sequence of positive real numbers $\{\delta_k\} \to 0$ such that $\{x^* + \delta_k d_k\} \subseteq K$.*

Let CFD(x^*) (resp., SFD(x^*)) denote the set of all d satisfying (a) (resp., (b)). Note that CFD(x^*) $= [\cap_{i \in I(x^*)} \tilde{A}_i^{-1}(\mathbb{R}_-)] - x^*$ and is a closed convex set containing the

origin (but not necessary a cone), while $\mathrm{SFD}(x^*)$ is a closed cone (but not necessarily convex).

DEFINITION 2.2. *Let $K_S(x^*)$, $K_C(x^*)$, and $K_L(x^*)$ be, respectively, defined by*

$$K_S(x^*) = (x^* + \overline{\mathrm{conv}}(\mathrm{SFD}(x^*))\bigcap C, \tag{2.3}$$

$$K_C(x^*) = (x^* + \mathrm{CFD}(x^*))\bigcap C \tag{2.4}$$

and

$$K_L(x^*) = (x^* + \overline{\mathrm{cone}}(\mathrm{CFD}(x^*)))\bigcap C. \tag{2.5}$$

Note that the three sets are closed convex and that

$$K_C(x^*) = \bigcap_{i\in I(x^*)} \tilde{A}_i^{-1}(\mathbb{R}_-)\bigcap C. \tag{2.6}$$

Note also that

$$K_C(x^*) \subseteq K_L(x^*), \tag{2.7}$$

and that $K_C(x^*) = K_L(x^*)$ when the level set $H_i^{-1}(\mathbb{R}_-)$ is a cone with the vertex $F_i(x^*)$ for all $i \in I(x^*)$. Furthermore, we have the following claim.

PROPOSITION 2.3. *Suppose that the level set $H_i^{-1}(\mathbb{R}_-)$ is a cone with the vertex $F_i(x^*)$ for all $i \in I(x^*)$. Then SFD$(x^*) \subseteq$CFD(x^*), and hence $K_S(x^*) \subseteq K_C(x^*) = K_L(x^*)$.*

Proof. The second assertion follows from the first and the fact that $\mathrm{CFD}(x^*)$ is closed convex. (It is straightforward to verify that $K_C(x^*) = K_L(x^*)$ under the stated assumption.) To prove the first assertion, let $d \in \mathrm{SFD}(x^*)$, and let $\{d_k\}$, $\{\delta_k\}$ be as in Definition 2.1(b). In particular, for each $i \in I(x^*)$, one has $H_i(F_i(x^* + \delta_k d_k)) \leqslant 0$ and hence that $F_i(x^* + \delta_k d_k) \in V_i$, where $V_i := H_i^{-1}(\mathbb{R}_-)$. Therefore

$$\delta_k F_i'(x^*)d_k + o(\|\delta_k d_k\|) \in V_i - F_i(x^*).$$

By the assumption, $V_i - F_i(x^*)$ is a cone. It follows that

$$F_i'(x^*)d_k + o(\|d_k\|) \in V_i - F_i(x^*);$$

passing to the limits, one has that $F_i'(x^*)d \in V_i - F_i(x^*)$. This implies that $d \in \mathrm{CFD}(x^*)$, and the proof is complete. □

PROPOSITION 2.4. *Suppose that* $\mathrm{int}(\overline{\mathrm{cone}}(\mathrm{CFD}(x^*))) \neq \emptyset$ *and that, for each $i \in I(x^*)$, $F_i'(x^*)$ is surjective. Then SFD$(x^*) \subseteq \overline{\mathrm{cone}}(\mathrm{CFD}(x^*))$, and hence $K_S(x^*) \subseteq K_L(x^*)$.*

Proof. We need only to prove the first assertion. As in the proof of Proposition 2.3, let $d \in \mathrm{SFD}(x^*)$, with $\{d_k\}$, $\{\delta_k\}$ as in Definition 2.1(b). Then

$$F_i'(x^*)d_k + o(\|d_k\|) \in V_i - F_i(x^*) \subseteq \overline{\mathrm{cone}}(V_i - F_i(x^*));$$

passing to the limits, one has that $F_i'(x^*)d \in \overline{\mathrm{cone}}(V_i - F_i(x^*))$ for each $i \in I(x^*)$. This shows that

$$\mathrm{SFD}(x^*) \subseteq \bigcap_{i\in I(x^*)} F_i'(x^*)^{-1}(\overline{\mathrm{cone}}(V_i - F_i(x^*))). \tag{2.8}$$

We claim that

$$\operatorname{cone}(\mathrm{CFD}(x^*)) = \bigcap_{i\in I(x^*)} F_i'(x^*)^{-1}(\operatorname{cone}(V_i - F_i(x^*))). \tag{2.9}$$

Indeed, it is clear that the set on the left-hand side of (2.9) is contained in the set on the right-hand side. Conversely, let d belong to the set of the right-hand side in (2.9). Then for each $i \in I(x^*)$ there exists $t_i > 0$ such that $\frac{d}{t_i} \in F_i'(x^*)^{-1}(V_i - F_i(x^*))$; that is,

$$F_i'(x^*)\frac{d}{t_i} \in V_i - F_i(x^*) \quad \forall i \in I(x^*).$$

Set $t := \max_i t_i$. Then, since $V_i - F_i(x^*)$ is a cone,

$$F_i'(x^*)\frac{d}{t} \in V_i - F_i(x^*), \quad i \in I(x^*).$$

This implies that $\frac{d}{t} \in \mathrm{CFD}(x^*)$, and so $d \in \operatorname{cone}(\mathrm{CFD}(x^*))$. Therefore, (2.9) holds. In addition, by (2.9) and the assumption $\operatorname{int}(\overline{\operatorname{cone}}(\mathrm{CFD}(x^*))) \neq \emptyset$,

$$\operatorname{int} \overline{\bigcap_{i\in I(x^*)} (F_i'(x^*)^{-1}(\operatorname{cone}(V_i - F_i(x^*))))} \neq \emptyset.$$

This implies that

$$\begin{aligned}
\overline{\operatorname{cone}}(\mathrm{CFD}(x^*)) &= \overline{\bigcap_{i\in I(x^*)} F_i'(x^*)^{-1}(\operatorname{cone}(V_i - F_i(x^*)))} \\
&= \bigcap_{i\in I(x^*)} \overline{F_i'(x^*)^{-1}(\operatorname{cone}(V_i - F_i(x^*)))} \\
&= \bigcap_{i\in I(x^*)} F_i'(x^*)^{-1}(\overline{\operatorname{cone}}(V_i - F_i(x^*))).
\end{aligned}$$

Here the last equality holds by the open mapping theorem and by the assumption that $F_i'(x^*)$ is surjective. Thus, by (2.8), we have the desired result. □

PROPOSITION 2.5. *Let $\tilde{A}_i$ be defined by* (2.1). *Then it holds that*

$$\partial A_i(x^*) = \partial \tilde{A}_i(x^*) = \partial H_i(F_i(x^*)) \circ F_i'(x^*), \tag{2.10}$$

where, by definition, $z \in \partial H_i(F_i(x^*)) \circ F_i'(x^*)$ if and only if there is $\zeta \in \partial H_i(F_i(x^*))$ such that

$$\langle z, v\rangle = \langle \zeta, F_i'(x^*)v\rangle \quad \forall v \in X.$$

Proof. Recalling that H_i is regular at $F_i(x^*)$ and F_i is strictly differentiable (see [6, Proposition 2.3.6 and section 2.2]), it follows from the chain rule (Theorem 2.3.10 of [6]) that

$$\partial A_i(x^*) = \partial H_i(F_i(x^*)) \circ F_i'(x^*).$$

Similarly, we also have

$$\partial \tilde{A}_i(x^*) = \partial H_i(F_i(x^*)) \circ F_i'(x^*). \qquad \square$$

We shall need the following well-known characterization theorem for the best approximation from a closed convex set G in X; see [3, 9, 10].

PROPOSITION 2.6. *Let G be a closed convex set in X. Then for any $x \in X$, $P_G(x) = g_0$ if and only if $g_0 \in G$, and for any $g \in G$, $\langle x - g_0, g_0 - g\rangle \geqslant 0$, that is, $x - g_0 \in N_G(g_0)$.*

DEFINITION 2.7. (a) *Let $\{C_0, \dots, C_m\}$ be a collection of closed convex sets and $x \in \bigcap_0^m C_j$. The collection is said to have the strong CHIP at x if*

$$N_{\bigcap_0^m C_j}(x) = \sum_{j=0}^{m} N_{C_j}(x).$$

(b) *Let $\{\phi_i : \ i = 1, 2, \dots, m\}$ be a collection of continuous convex functions on X, and let C be a closed convex set in X. The system of convex inequalities*

$$\phi_i(z) \leqslant 0, \quad i = 1, 2, \dots, m, \tag{2.11}$$

is said to satisfy the BCQ relative to C at x if (2.11) *holds for $Z = x$ and*

$$N_{C\cap(\cap_{i=1}^m \phi_i^{-1}(\mathbb{R}_-))}(x) = N_C(x) + \operatorname{cone}(\cup_{i\in I(x)}\{\partial\phi_i(x)\}),$$

where $I(x) = \{i : \ \phi_i(x) = 0\}$.

Remark 2.1. (a) It is known (see [14]) and easy to see that if system (2.11) satisfies the BCQ relative to C at x, then $\{C, \phi_1^{-1}(\mathbb{R}_-), \dots, \phi_m^{-1}(\mathbb{R}_-)\}$ has the strong CHIP. For further discussions relating to the strong CHIP, see also [7, 8, 19].

(b) If ϕ_i is affine, it is well known that

$$\operatorname{cone}(\partial\phi_i(x)) = N_{\phi_i^{-1}(\mathbb{R}_-)}(x), \quad i \in I(x),$$

and hence that

$$\operatorname{cone}\left(\bigcup_{i\in I(x)} \partial\phi_i(x)\right) = \sum_{i\in I(x)} \operatorname{cone}\left(\partial\phi_i(x)\right) = \sum_{i\in I(x)} N_{\phi_i^{-1}(\mathbb{R}_-)}(x) = \sum_{i=1}^{m} N_{\phi_i^{-1}(\mathbb{R}_-)}(x).$$

Thus system (2.11) satisfies the BCQ relative to C at x if and only if $\{C, \phi_1^{-1}(\mathbb{R}_-)$, $\phi_2^{-1}(\mathbb{R}_-), \dots, \phi_m^{-1}(\mathbb{R}_-)\}$ has the strong CHIP at x.

(c) When $C = X$, the definition of the BCQ relative to C at x is the same as the BCQ at x considered in [12, 13]. Note that if $x \in \cap_{j\in I(x)}\phi_j^{-1}(\mathbb{R}_-)$ and $\phi_i(x) = 0$, then $\operatorname{cone}(\partial\phi_i(x)) \subseteq N_{\phi_i^{-1}(\mathbb{R}_-)}(x)$. In addition, some further properties were investigated in [14].

3. Reformulation of the best approximation. We begin with a key lemma that provides a unified tool for the study of best approximation from nonconvex sets.

LEMMA 3.1. *Let $\hat{K}$ be a closed set, C a closed convex set in X, and let $x^* \in X$ be such that $x^* \in \hat{K} \subseteq C$. Let T be a closed convex cone in X. Then the following statements are equivalent:*

(i) $\hat{K} \subseteq (x^* + T) \cap C$;
(i*) $\hat{K} \subseteq x^* + T$;
(ii) $P_{\hat{K}}(x) = x^*$ *whenever* $x \in X$ *with* $P_{(x^*+T)\cap C}(x) = x^*$;
(iii) $P_{\hat{K}}(x) = x^*$ *whenever* $x \in X$ *with* $P_{x^*+T}(x) = x^*$.

Note: By abuse of notations, $P_{\hat{K}}(x) = x^*$ is read as $x^* \in P_K(x)$ when $P_{\hat{K}}(x)$ is multivalued.

Proof. Since $\hat{K} \subseteq C$, (i) $\Leftrightarrow$ (i*). It is trivial that (i) $\Rightarrow$ (ii) $\Rightarrow$ (iii). It remains to show that (iii) $\Rightarrow$ (i*). Suppose that (i*) does not hold; take $\bar{x} \in \hat{K}$ such that $\bar{x} \notin x^* + T$. Let $x^* + e \in P_{x^*+T}(\bar{x})$, where $e \in T$. Write h for $\bar{x} - (x^* + e)$. Then, by Proposition 2.6,

$$\langle \bar{x} - (x^* + e), (x^* + z) - (x^* + e) \rangle \leqslant 0 \quad \forall z \in T;$$

that is, $\langle h, z - e \rangle \leqslant 0$ for each $z \in T$. Letting $z = 2e, e/2$ separately, it follows that $\langle h, e \rangle = 0$, and hence that $\langle h, z \rangle \leqslant 0$ for each $z \in T$. Consequently, $P_{x^*+T}(x_t) = x^*$ for each $t > 0$, where $x_t := x^* + th$; this is because of Proposition 2.6 and

$$\langle \bar{x}_t - x^*, (x^* + z) - x^* \rangle = \langle th, z \rangle \leqslant 0 \quad \forall z \in T.$$

By (iii), it follows that

$$P_{\hat{K}}(x_t) = x^*. \tag{3.1}$$

On the other hand, for $t > 1$ large enough,

$$\begin{aligned} \|x_t - \bar{x}\|^2 &= \|x^* + th - (h + x^* + e)\|^2 \\ &= (t-1)^2 \|h\|^2 + \|e\|^2 \\ &< t^2 \|h\|^2. \end{aligned}$$

Since $\bar{x} \in \hat{K}$, this contradicts (3.1). The proof is complete. □

The following corollary is evident.

COROLLARY 3.2. *Let C be a closed convex set, and let T_1, T_2 be closed convex cones in X; let $x^* \in C$. Then the following statements are equivalent:*

(i) $C \cap (x^* + T_1) = C \cap (x^* + T_2)$;

(ii) *for any $x \in X$, $P_{C\cap(x^*+T_1)}(x) = x^*$ if and only if $P_{C\cap(x^*+T_2)}(x) = x^*$.*

Theorem 3.7 of [14] follows immediately from Lemma 3.1 by applying to $\hat{K} = K$ defined in section 1 and $T = \overline{\text{conv}}(\text{SFD}(x^*))$. Similarly, by letting $T = \overline{\text{cone}}(\text{CFD}(x^*))$ in Lemma 3.1, we have the following result.

PROPOSITION 3.3. *Let $x^* \in K$. Then the following statements are equivalent:*

(i) $K \subseteq K_L(x^*)$;

(ii) *for any $x \in X$, $P_{K_L(x^*)}(x) = x^* \Longrightarrow P_K(x) = x^*$.*

DEFINITION 3.4. *Let $x^* \in K$. Then*

(a) *x^* is called a regular point of K (more precisely, a regular point of K with respect to C and the system (NIS)) if*

$$K \subseteq K_C(x^*) \subseteq K_S(x^*); \tag{3.2}$$

(b) *x^* is called a weakly regular point of K (with respect to C and the system* (NIS)*) if*

$$K \subseteq K_L(x^*) \quad \text{and} \quad K_C(x^*) \subseteq K_S(x^*). \tag{3.3}$$

Remark 3.1. (a) Obviously, a regular point of K must be a weakly regular point of K; the converse is true if the assumption of Proposition 2.3 is satisfied.

(b) If F_i is affine for each $i \in I(x^*)$, then x^* is a regular point of K.

THEOREM 3.5. *Let $x^* \in K$. If x^* is a regular point of K in the above sense, then for any $x \in X$,*

$$P_K(x) = x^* \Longleftrightarrow P_{K_C(x^*)}(x) = x^*. \tag{3.4}$$

Furthermore, if $K \subseteq K_S(x^*)$ and $H_i^{-1}(\mathbb{R}_-)$ is a cone with the vertex $F_i(x^*)$ for each $i \in I(x^*)$, then x^* is regular if and only if (3.4) *holds.*

Proof. Suppose that (3.2) holds. Then

$$P_K(x) = x^* \Longrightarrow P_{K_S(x^*)}(x) = x^* \Longrightarrow P_{K_C(x^*)}(x) = x^* \Longrightarrow P_K(x) = x^*, \tag{3.5}$$

where the first implication holds by [14] (see also Lemma 3.8 below). Hence (3.4) holds. Conversely, suppose that $K \subseteq K_S(x^*)$ and that $H_i^{-1}(\mathbb{R}_-)$ is a cone with the vertex $F_i(x^*)$ (thus $K_C(x^*) = K_L(x^*)$) for each $i \in I(x^*)$. Then it follows from the first implication of (3.5) that, for any $x \in X$, $P_{K_S(x^*)}(x) = x^* \Longleftrightarrow P_K(x) = x^*$. Thus, from (3.4), we have

$$P_{K_S(x^*)}(x) = x^* \Longleftrightarrow P_{K_C(x^*)}(x) = x^*.$$

By Corollary 3.2 and noting that $K_C(x^*) = K_L(x^*)$ and $K_S(x^*)$ are cones with the vertex x^*, this implies that $K_C(x^*) = K_S(x^*)$, and so $K \subseteq K_C(x^*)$, i.e., (3.2) holds. The proof is complete. □

If K is convex, then $K \subseteq K_S(x^*)$. We therefore have the following result.

COROLLARY 3.6. *Let $x^* \in K$. Suppose that K is convex and that $H_i^{-1}(\mathbb{R}_-)$ is a cone with the vertex $F_i(x^*)$ for each $i \in I(x^*)$. Then x^* is regular if and only if* (3.4) *holds.*

We are now ready to present one of our main results. Recall that $\tilde{A}_i$ is defined by (2.1).

THEOREM 3.7. *Let $x^* \in K$. Then the following statements are equivalent:*

(i) *The system of convex inequalities*

$$\tilde{A}_i(z) \leqslant 0, \quad i \in I(x^*), \tag{3.6}$$

satisfies the BCQ relative to C at x^*.

(ii) *The system of convex inequalities*

$$\tilde{A}_i(z) \leqslant 0, \quad i = 1, 2, \ldots, m, \tag{3.7}$$

satisfies the BCQ relative to C at x^*.

(i*) *x^* has the perturbation property with respect to C and the system* (3.6).

(ii*) *x^* has the perturbation property with respect to C and the system* (3.7).

Moreover, if $x^* \in K$ is a regular point of K with respect to C and the system (NIS), *then each of the above statements is also equivalent to the following:*

(iii) *x^* has the perturbation property with respect to C and the system* (NIS).

Proof. The equivalence of (i) $\Longleftrightarrow$(i*) and (ii) $\Longleftrightarrow$(ii*) are given in [14, Theorem 5.1]. By (2.1), $\tilde{A}_i(x^*) = A_i(x^*)$; hence $i \in I(x^*)$ if and only if $\tilde{A}_i(x^*) = 0$. We may assume that $I(x^*)$ is a proper subset of $\{1, 2, \ldots, m\}$ and note that $x^* \in \operatorname{int}(\cap_{i \notin I(x^*)} \tilde{A}_i^{-1}(\mathbb{R}_-))$. Writing C_i for $\tilde{A}_i^{-1}(\mathbb{R}_-)$, it follows from [1, Corollary 2.4, p. 113] that

$$\partial(\mathbf{I}_{C \cap (\cap_{i \in I(x^*)} C_i)}(x^*) + \mathbf{I}_{\cap_{i \notin I(x^*)} C_i}(x^*)) = \partial \mathbf{I}_{C \cap (\cap_{i \in I(x^*)} C_i)}(x^*) + \partial \mathbf{I}_{\cap_{i \notin I(x^*)} C_i}(x^*);$$

that is,

$$N_{C \cap (\cap_{i=1}^m C_i)}(x^*) = N_{C \cap (\cap_{i \in I(x^*)} C_i)}(x^*).$$

Therefore (i) $\Longleftrightarrow$ (ii). Under the additional assumption that x^* is regular (and thus that Theorem 3.5 is applicable), we will show the equivalence of (i*) $\Longleftrightarrow$ (iii). Consider the following statements for $x \in X$:

(1) $P_K(x) = x^*$;
(2) $P_{C\cap(\cap_{i\in I(x^*)}\tilde{A}_i^{-1}(\mathbb{R}_-))}(x) = x^*$;
(3) $P_C(x - \sum_{i\in I(x^*)}\lambda_i h_i) = x^*$ for some $h_i \in \partial A_i(x^*)$ and $\lambda_i \geqslant 0$.

By Theorem 3.5 and the fact that $K_C(x^*) = C \cap (\cap_{i\in I(x^*)}\tilde{A}_i^{-1}(\mathbb{R}_-))$, (1) $\Longleftrightarrow$ (2). Since $\partial A_i(x^*) = \partial\tilde{A}_i(x^*)$ (see Proposition 2.6), (i*) holds if and only if [(2)$\Longleftrightarrow$(3)]. Therefore, (i*) holds if and only if [(1)$\Longleftrightarrow$(3)]; namely, (i*) holds if and only if (iii) holds. □

Remark 3.2. (a) The statement (ii*) simply means (by definition):

(ii′) For any $x \in X$,

$$P_{K_C(x^*)}(x) = x^* \Longleftrightarrow P_C\left(x - \sum_{i\in I(x^*)}\lambda_i h_i\right) = x^* \quad \text{for some } h_i \in \partial\tilde{A}_i(x^*),\ \lambda_i \geqslant 0,$$

where $\partial\tilde{A}_i(x^*)$ can be replaced by $\partial A_i(x^*)$, by Proposition 2.6.

(b) The sufficient part of (ii′) holds in general by Lemma 3.8(ii) below.

(c) The system (3.6) (or (3.7)) may be referred to as a convexification system of (NIS).

LEMMA 3.8. *Let $x^* \in K$ and $x \in X$. The following statements hold:*

(i) *If $P_K(x) = x^*$, then $P_{K_S(x^*)}(x) = x^*$.*

(ii) *If*

$$P_C\left(x - \sum_{i=1}^m \lambda_i h_i\right) = x^* \tag{3.8}$$

for some $h_i \in \partial A_i(x^*)$ and $\lambda_i \geqslant 0$ with $\lambda_i = 0$ for all $i \notin I(x^*)$, then $P_{K_L(x^*)}(x) = x^*$, and so $P_{K_C(x^*)}(x) = x^*$.

Proof. For a proof of (i), see [14]. Next suppose that (3.8) holds. Then, by Proposition 2.6,

$$x - \sum_{i=1}^m \lambda_i h_i - x^* \in N_C(x^*).$$

Hence,

$$\begin{aligned}
x - x^* &\in N_C(x^*) + \sum_{i=1}^m \lambda_i h_i \\
&\subseteq N_C(x^*) + \sum_{i\in I(x^*)} \operatorname{cone}\partial A_i(x^*) \\
&\subseteq N_C(x^*) + \sum_{i\in I(x^*)} N_{\tilde{A}_i^{-1}(\mathbb{R}_-)}(x^*) \\
&\subseteq N_C(x^*) + N_{\cap_{i\in I(x^*)}\tilde{A}_i^{-1}(\mathbb{R}_-)}(x^*) \\
&= N_C(x^*) + \left(\bigcap_{i\in I(x^*)} \tilde{A}_i^{-1}(\mathbb{R}_-) - x^*\right)^\circ \\
&= N_C(x^*) + (\overline{\operatorname{cone}}\mathrm{CFD}(x^*))^\circ \\
&= N_C(x^*) + N_{(x^*+\overline{\operatorname{cone}}\mathrm{CFD}(x^*))}(x^*) \\
&\subseteq N_{C\cap(x^*+\overline{\operatorname{cone}}\mathrm{CFD}(x^*))}(x^*) \\
&= N_{K_L(x^*)}(x^*).
\end{aligned}$$

This implies that $P_{K_L(x^*)}(x) = x^*$ by Proposition 2.6.

The following theorem shows that the regularity condition in Theorem 3.7 can be replaced by weak regularity if a Slater-type condition is satisfied.

THEOREM 3.9. *Let $x^* \in K$ be a weakly regular point of K, and suppose that*

$$\mathrm{ri}(x^* + \mathrm{coneCFD}(x^*)) \bigcap C \neq \emptyset. \tag{3.9}$$

Then the following statements are equivalent:

(i) *System* (3.7) *satisfies the BCQ relative to C at x^*.*

(ii) *x^* has the perturbation property with respect to C and the system* (NIS).

Proof. Suppose that (i) holds. Then, by Theorem 3.7, (ii′) holds. For each $x \in X$, the following implications hold:

$$\begin{array}{ll} P_K(x) = x^* \Longrightarrow P_{K_S(x^*)} = x^* & \text{(Lemma 3.8)} \\ \Longrightarrow P_{K_C(x^*)} = x^* & (K_C(x^*) \subseteq K_S(x^*) \text{ by weak regularity}) \\ \Longrightarrow \text{(3.8) holds} & \text{(Theorem 3.7(ii′))} \\ \Longrightarrow P_{K_L(x^*)} = x^* & \text{(Lemma 3.8(ii))} \\ \Longrightarrow P_K = x^*. & (K \subseteq K_L(x^*) \text{ by weak regularity}) \end{array}$$

This proves that (i)$\Longrightarrow$(ii). ((3.9) is not needed for this implication.)

To prove the opposite implication (ii)$\Longrightarrow$(i), note that, since (3.9) is satisfied,

$$K_L(x^*) = \overline{(x^* + \mathrm{cone}(\mathrm{CFD}(x^*)) \cap C}. \tag{3.10}$$

We will show below that

$$P_{K_C(x^*)}(x) = x^* \Longleftrightarrow P_{K_L(x^*)}(x) = x^*. \tag{3.11}$$

Indeed, since $K_C(x^*) \subseteq K_L(x^*)$, it is sufficient to show

$$P_{K_C(x^*)}(x) = x^* \Longrightarrow P_{K_L(x^*)}(x) = x^*. \tag{3.12}$$

Assume that $P_{K_C(x^*)}(x) = x^*$. By Proposition 2.6, we have

$$\langle x - x^*, x^* - \bar{z}\rangle \geqslant 0 \quad \forall \bar{z} \in K_C(x^*). \tag{3.13}$$

Let $z \in K_L(x^*)$: $z \in C$ and $z = x^* + t(\bar{z} - x^*)$ for some $\bar{z} \in x^* + \mathrm{CFD}(x^*)$ and $t \geqslant 0$. Without loss of generality, assume that $t > 1$. Thus, $\bar{z} = x^* + (1/t)(z - x^*)$, and so $\bar{z} \in C$ since $z \in C$; consequently, $\bar{z} \in K_C(x^*)$. This, with (3.13), implies that

$$\langle x - x^*, x^* - z\rangle = t\langle x - x^*, x^* - \bar{z}\rangle \geqslant 0.$$

Hence, by (3.10), $x - x^* \in N_{K_L(x^*)}(x^*)$. By Proposition 2.6 again, (3.12) holds and so does (3.11). For each $x \in X$, the following implications hold:

$$\begin{array}{ll} P_K(x) = x^* \Longleftrightarrow \text{(3.8) holds} & \text{(by (ii))} \\ \Longrightarrow P_{K_L(x^*)}(x) = x^* & \text{(Lemma 3.8)} \\ \Longrightarrow P_K(x) = x^*. & (K \subseteq K_L(x^*) \text{ by the weak regularity}) \end{array}$$

Combining this with (3.11), (ii′) of Remark 3.2(a) is seen to hold. Thus, by Theorem 3.7, (i) holds. □

Remark 3.3. (a) The implication (i)$\Longrightarrow$(ii) of Theorem 3.9 remains true even if the condition (3.9) is dropped. Example 3.1 below shows that we do require condition (3.9) for the implication (ii)$\Longrightarrow$(i).

(b) In the case in which the condition (3.9) is satisfied, Theorem 3.9 is a genuine extension of Theorem 3.7 (see Example 3.2 below).

Remark 3.4. If x^* is regular, then

$$K \subseteq K_C(x^*) \quad \text{and} \quad K \subseteq K_S(x^*). \tag{3.14}$$

In the following corollaries, we consider (3.14) instead of the regularity.

COROLLARY 3.10. *Suppose that x^* satisfies* (3.14). *Then the following statements are equivalent:*

(i) *The system* (3.7) *satisfies the BCQ relative to C at x^*, and x^* is a regular point of K.*

(ii) *x^* has the perturbation property with respect to C and the system* (NIS).

Proof. By Theorem 3.7, (i)$\Longrightarrow$ (ii). Conversely, suppose that (ii) holds. We claim that, for every $x \in X$,

$$P_{K_S(x^*)}(x) = x^* \Longleftrightarrow P_K(x) = x^* \Longleftrightarrow P_{K_C(x^*)}(x) = x^*. \tag{3.15}$$

Indeed, by (3.14), $P_K(x) = x^*$ if either $P_{K_S(x^*)}(x) = x^*$ or $P_{K_L(x^*)}(x) = x^*$. Conversely, let $x \in X$ and $x^* = P_K(x)$. Then $x^* = P_{K_S(x^*)}(x)$ by Lemma 3.8(i), and it follows from (ii) that $x^* = P_C(x - \sum_1^m \lambda_i h_i)$ for some $h_i \in \partial A_i(x^*)$ and $\lambda_i \geqslant 0$, with $\lambda_i = 0$ for all $i \notin I(x^*)$. By Lemma 3.8(ii), it follows that $x^* = P_{K_C(x^*)}(x)$. Therefore, (3.15) holds. By Lemma 3.1, this implies that $K_C(x^*) \subseteq K_S(x^*)$. Combining this with (3.14), x^* is regular. Now Theorem 3.7 is applicable, and thus (ii)$\Longrightarrow$(i). ∎

COROLLARY 3.11. *Suppose that the system* (3.7) *satisfies the BCQ relative to C at x^* and that*

$$K_C(x^*) = K_S(x^*). \tag{3.16}$$

Then (3.14) *holds if and only if x^* has the perturbation property with respect to C and the system* (NIS).

Proof. In view of the preceding corollary, the necessity part is clear. Conversely, suppose that x^* has the perturbation property with respect to C and the system (NIS). Then we have the following equivalences:

$$\begin{aligned} &P_K(x) = x^* \Longleftrightarrow (3.8) \text{ holds} \\ &\Longleftrightarrow P_{K_C(x^*)}(x) = x^* && ((\text{ii}) \Longleftrightarrow (\text{ii}^*) \text{ of Theorem 3.7}) \\ &\Longleftrightarrow P_{K_S(x^*)}(x) = x^*. && (\text{by } (3.16)) \end{aligned}$$

Thus $K \subseteq K_S(x^*)$ by Lemma 3.1. Combining this with (3.16), we see that (3.14) holds. ∎

A natural question arises from Theorem 3.7: When does the inclusion $K_C(x^*) \subseteq K_S(x^*)$ hold? Apart from the obvious sufficient condition that each F_i, $i \in I(x^*)$, is affine, we give another sufficient condition below in the case when $Y = R^n$. Let $\text{aff}(C)$ denote the linear manifold (i.e., affine subspace) spanned by C. Define

$$E = \{i : \text{int} H_i^{-1}(\mathbb{R}_-) = \emptyset\}, \qquad I_0(x^*) = I(x^*) \setminus E.$$

Write

$$F_i := (F_{i1}, F_{i2}, \dots, F_{in}), \quad i = 1, 2, \dots, m.$$

Note that

$$H_i(x) \geqslant 0 \quad \text{on } X \quad \forall i \in E. \tag{3.17}$$

Let

$$S^* = \{d \in X : H_i(F_i(x^*)+F_i'(x^*)d) = 0,\ i \in E;\ H_i(F_i(x^*)+F_i'(x^*)d) < 0,\ i \in I_0(x^*)\}.$$

Thus $S^* \subseteq \mathrm{CFD}(x^*)$; moreover, by (3.17),

$$(1-t)d_1 + td_2 \in S^* \quad \forall t \in [0,1),\ d_1 \in S^*,\ d_2 \in \mathrm{CFD}(x^*). \tag{3.18}$$

In particular (by letting $d_2 = 0$), one has

$$(1-t)d_1 \in S^* \quad \forall t \in [0,1),\ d_1 \in S^*. \tag{3.19}$$

DEFINITION 3.12. *Let $x^* \in K$, and suppose that $Y = R^n$. We say that* (NIS) *satisfies the generalized MFCQ (Mangasarian–Fromowitz constraint qualification) at x^* if the following conditions are satisfied:*

(a) *The intersection $(x^* + \mathrm{CFD}(x^*)) \cap \mathrm{ri}C$ is nonempty;*

(b) *$\{F_{ij}'(x^*) : i \in E,\ j = 1, 2, \dots, n\}$ are linearly independent on $\mathrm{span}(C - x^*)$;*

(c) *the intersection $S^* \cap \mathrm{span}(C - x^*)$ is nonempty.*

Remark 3.5. In the special case in which $Y = \mathbb{R}$, each H_i is the identity mapping, and C is a subspace of X, the above (a) is automatic, while (b) and (c) are, respectively, equivalent to the following:

(b′) $\{F_{ij}'(x^*) : i \in E,\ j = 1\}$ are linearly independent on C;

(c′) the intersection $S^* \cap C$ is nonempty.

That is, the generalized MFCQ condition coincides with the standard MFCQ on C ([16]; see also [15, 21]).

Our next main result is the following.

THEOREM 3.13. *Let $x^* \in K$ and $Y = R^n$. Suppose that* (NIS) *satisfies the generalized MFCQ at x^* Then*

$$K_C(x^*) \subseteq K_S(x^*). \tag{3.20}$$

If, in addition, for each $i \in I(x^*)$,

$$\tilde{A}_i(z) \leqslant A_i(z) \quad \textit{for each } z \in C, \tag{3.21}$$

then x^* is regular.

Proof. It is easy to verify that $K \subseteq K_C(x^*)$ if (3.21) holds. Thus we need only to prove (3.20). By Definition 3.12(a), it is not difficult to verify that

$$K_C(x^*) = \overline{(x^* + \mathrm{CFD}(x^*)) \cap \mathrm{ri}C}. \tag{3.22}$$

Thus, we need only to prove that

$$(x^* + \mathrm{CFD}(x^*)) \cap \mathrm{ri}C \subseteq K_S(x^*). \tag{3.23}$$

Let $\bar{x}$ belong to the set on the left-hand side of (3.23), and let $d = \bar{x} - x^*$. Then

$$d \in \mathrm{CFD}(x^*) \cap \mathrm{span}(C - x^*). \tag{3.24}$$

By (c), pick $\bar{d}_0 \in S^* \cap \mathrm{span}(C - x^*)$. Define

$$\bar{d}_k = \left(1 - \frac{1}{k}\right) d + \frac{1}{k}\bar{d}_0 \quad \forall k \in \mathbb{N}. \tag{3.25}$$

Then, by (3.18) and (3.24), one has

$$\bar{d}_k \in S^* \cap \operatorname{span}(C - x^*). \tag{3.26}$$

By (b), take a family $\{x_{ij} \in \operatorname{span}(C - x^*) : \ i \in E; j = 1, 2, \dots, n\}$ of vectors in $\operatorname{span}(C - x^*)$, which is dual to $\{F'_{ij}(x^*)\}$ in the sense that

$$F'_{ij}(x^*)x_{hl} = \begin{cases} 1 & \text{if } (i,j) = (h,l), \\ 0 & \text{otherwise.} \end{cases} \tag{3.27}$$

Let Z_k denote the linear subspace of X spanned by $\bar{d}_k$ and the vectors x_{hl}, with $h \in E$ and $l = 1, 2, \dots, n$. We will show that there exist $\bar{\theta}_k \in (0, \frac{1}{k})$ and a continuously differentiable function $\theta \mapsto x_k(\theta)$ defined on $[0, \bar{\theta}_k]$ such that, for each $(i,j) \in E \times \{1, 2, \dots, n\}$ and each $\theta \in [0, \bar{\theta}_k]$,

$$(*) \qquad \begin{cases} x_k(\theta) \in Z_k + x^*, \\ x_k(0) = x^*, \\ x'_k(0) = \bar{d}_k, \\ F_{ij}(x(\theta)) = F_{ij}(x^*) + \theta F'_{ij}(x^*)\bar{d}_k, \quad (i,j) \in E \times \{1, 2, \dots, n\}. \end{cases}$$

Granting this, we show below that $x_k(\theta)$ satisfies (NIS) for sufficiently small $\theta > 0$:

$$H_i(F_i(x_k(\theta)) \leqslant 0, \quad i = 1, 2, \dots, m. \tag{3.28}$$

Since $x_k(0) = x^*$ and by considering smaller θ if necessary, we need only verify the above (3.28) for $i \in I(x^*)$. If $i \in E$, then the last equality in (*) gives

$$H_i(F_i(x(\theta))) = H_i(F_i(x^*) + \theta F'_i(x^*)\bar{d}_k) = 0 \quad \forall \theta \in [0, \bar{\theta}_k],$$

thanks to (3.19) and (3.26). If $i \in I_0(x^*)$, then the Taylor theorem gives

$$F_i(x_k(\theta)) = F_i(x^*) + \theta F'_i(x^*)\bar{d}_k + o(\theta),$$

and thus it follows from the convexity that

$$\begin{aligned} H_i(F_i(x_k(\theta))) &= H_i(F_i(x^*) + \theta(F'_i(x^*)\bar{d}_k + O(\theta))) \\ &\leqslant \theta H_i(F_i(x^*) + F'_i(x^*)\bar{d}_k + O(\theta)) < 0, \end{aligned}$$

provided that $\theta > 0$ is sufficiently small. Here the last inequality holds because

$$H_i(F_i(x^*) + F'_i(x^*)\bar{d}_k) < 0$$

as $\bar{d}_k \in S^*$ and $i \in I_0(x^*)$. Therefore, by taking smaller $\bar{\theta}_k > 0$ if necessary, (3.28) becomes valid for all $\theta \in [0, \bar{\theta}_k]$. By (*), take θ_k with $0 < \theta_k \leqslant \bar{\theta}_k \leqslant 1/k$ such that

$$\left\| \frac{x_k(\theta_k) - x^*}{\theta_k} - \bar{d}_k \right\| < \frac{1}{k}.$$

Then, by (3.25),

$$\left\| \frac{x_k(\theta_k) - x^*}{\theta_k} - d \right\| < \frac{1}{k}(1 + \|d - \bar{d}_0\|).$$

Thus, setting $d_k = \frac{x_k(\theta_k)-x^*}{\theta_k}$, we have $\lim_{k\to\infty} d_k = d$. To verify (3.23), it suffices to show $d \in \mathrm{SFD}(x^*)$. We will establish this by showing that $x_k(\theta_k) \in K$. To do this, note first that, because

$$d_k + x^* = \frac{x_k(\theta_k)}{\theta_k} + \left(1 - \frac{1}{\theta_k}\right) x^* \in \mathrm{aff}(C),$$

it follows from $\bar{x} \in \mathrm{ri}C$ and $\lim_{k\to\infty}(d_k + x^*) = \bar{x}$ that $d_k + x^* \in C$ for k large enough. This implies that $x_k(\theta_k) \in C$ as $x_k(\theta_k) = (1-\theta_k)x^* + \theta_k(d_k + x^*)$. Consequently, it follows from (3.28) that $x_k(\theta_k) \in K$, as required.

To show that there exists x_k with property (*), henceforth we fix k and consider only the special case in which $\bar{d}_k$ is linearly independent from $\{x_{ij}, (i,j) \in E \times \{1,2,\ldots,n\}\}$ (the case in which $\bar{d}_k$ is linearly dependent on $\{x_{ij}\}$ can be dealt with similarly but somewhat more simply); in this case, take a unit vector $x_0 \in Z_k$ such that $\langle x_0, x_{ij}\rangle = 0$ for each $(i,j) \in E \times \{1,2,\ldots,n\}$. Then

$$\bar{d}_k = \langle x_0, \bar{d}_k\rangle x_0 + \sum_{ij} \lambda_{ij} x_{ij} \tag{3.29}$$

for some $\lambda_{ij} \in \mathbb{R}$. We consider the equality system for x in $Z_k + x^*$ near x^*:

$$\begin{cases} F_{ij}(x) = F_{ij}(x^*) + \theta F'_{ij}(x^*)\bar{d}_k, \quad (i,j) \in E \times \{1,2,\ldots,n\}, \\ \langle x_0, x - x^*\rangle = \theta\langle x_0, \bar{d}_k\rangle. \end{cases}$$

For simplicity of notation, we write $\tilde{E}$ for $E \times \{1,2,\ldots,n\}$ and N for the cardinality $|\tilde{E}|$ of $\tilde{E}$. Expressing x in the form

$$x = \alpha_0 x_0 + \sum_{ij} \alpha_{ij} x_{ij} + x^*,$$

the above system can be written as for $(\alpha_0, \alpha_{ij}) \in \mathbb{R}^{1+N}$ near the origin:

$$\begin{cases} F_{ij}(\alpha_0 x_0 + \sum_{ij} \alpha_{ij} x_{ij} + x^*) = F_{ij}(x^*) + \theta F'_{ij}(x^*)\bar{d}_k, \quad (i,j) \in \tilde{E}, \\ \alpha_0 = \theta\langle x_0, \bar{d}_k\rangle. \end{cases} \tag{3.30}$$

The Jacobi matrix J for (3.30) at the origin is nonsingular; in fact, by (3.27),

$$J = \begin{pmatrix} 1 & 0 & \cdots & 0 & F'_{11}(x^*)x_0 \\ 0 & 1 & \cdots & 0 & F'_{12}(x^*)x_0 \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ 0 & 0 & \cdots & 1 & F'_{|E|n}(x^*)x_0 \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix}.$$

By the implicit function theorem, there exist $\theta_k \in (0, \frac{1}{k})$ and continuously differentiable functions, still denoted by α_0, α_{ij}, such that the preceding equality system is satisfied by these functions on $[-\theta_k.\theta_k]$ and such that each of these functions vanishes at $\theta = 0$. Set

$$x_k(\theta) = \alpha_0(\theta)x_0 + \sum \alpha_{ij}(\theta)x_{ij} + x^*, \quad \theta \in [-\theta_k, \theta_k].$$

Then (*) is seen to hold. Indeed, by differentiating each equation in the preceding system at the origin and making use of the dual property (3.27) of $\{x_{ij}\}$ relative to

$\{F'_{ij}(x^*)\}$, one has

$$J \cdot \begin{pmatrix} \alpha'_{11}(0) \\ \vdots \\ \alpha'_{|E|n}(0) \\ \alpha'_0(0) \end{pmatrix} = \begin{pmatrix} F'_{11}(x^*)\bar{d}_k \\ \vdots \\ F'_{|E|n}(x^*)\bar{d}_k \\ \langle x_0, \bar{d}_k \rangle \end{pmatrix}.$$

Computing the last row gives

$$\alpha'_0(0) = \langle x_0, \bar{d}_k \rangle, \tag{3.31}$$

and computing the other rows gives

$$\alpha'_{ij}(0) + F'_{ij}(x^*)x_0 \cdot \alpha'_0(0) = F'_{ij}(x^*)\bar{d}_k \quad \forall (ij) \in \tilde{E}. \tag{3.32}$$

By the dual property of $\{x_{ij}\}$ relative to $\{F'_{ij}(x^*)\}$, it follows from (3.29), (3.31), and (3.32) that $\alpha'_{ij}(0) = \lambda_{ij}$ for each $(i,j) \in \tilde{E}$. Consequently,

$$\begin{aligned} x'_k(0) &= \alpha'_0(0)x_0 + \sum \alpha'_{ij}(0)x_{ij} \\ &= \langle x_0, \bar{d}_k \rangle x_0 + \sum \lambda_{ij}x_{ij} \\ &= \bar{d}_k, \end{aligned}$$

thanks to (3.29). Therefore (*) holds, and the proof is complete. □

COROLLARY 3.14. *Let $Y = R^n$ and $x^* \in K$. Suppose that*

(a) *the intersection $(x^* + \mathrm{CFD}(x^*)) \cap \mathrm{ri}C$ is nonempty;*

(b′) *$\{F'_{ij}(x^*) : i \in I(x^*);\ j = 1, 2, \ldots, n\}$ is linearly independent on $\mathrm{span}(C - x^*)$.*

Then $K_C(x^*) \subseteq K_S(x^*)$.

Proof. It is sufficient to show that the condition (c) of Definition 3.12 is satisfied by virtue of the strengthened condition (b′) (comparing with (b)). If $I_0(x^*) = \emptyset$, then $0 \in S^* \cap \mathrm{span}(C - x^*)$. Hence, we assume that $I_0(x^*) \neq \emptyset$. For any $i \in I_0(x^*)$, let $\alpha_i = (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{in}) \neq 0$ be an element of $\mathbb{R}^n$ satisfying $H_i(F_i(x^*) + \alpha_i) < 0$. From assumption (b′), there exists $\{x_{kl} : (k,l) \in I(x^*) \times \{1, 2, \ldots, n\}\}$ in $\mathrm{span}(C - x^*)$ such that $F'_{ij}(x^*)x_{kl} = \alpha_{ij}$ if $(i,j) = (k,l) \in I_0(x^*) \times \{1, 2, \ldots, n\}$, and $F'_{ij}(x^*)x_{kl} = 0$ otherwise. Then $d_k := \sum_{l=1}^n x_{kl}$ satisfies $F'_i(x^*)d_k = \alpha_i$ if $i = k \in I_0(x^*)$, and $F'_i(x^*)d_k = 0$ otherwise. Let $d := \sum_{k \in I(x^*)} d_k$. Then $d \in S^* \cap \mathrm{span}(C - x^*)$. The proof is complete. □

Example 3.1. Let $X = Y = \mathbb{R}^2$ and $C = [-1/2, 1] \times \{0\}$. Define

$$A(x_1, x_2) = H(F_1(x_1, x_2), F_2(x_1, x_2)),$$

where

$$F_1(x_1, x_2) = x_1(1 + x_2^2) \quad \forall (x_1, x_2) \in \mathbb{R}^2,$$

$$F_2(x_1, x_2) = x_1^2 + x_2 \quad \forall (x_1, x_2) \in \mathbb{R}^2,$$

and

$$H(u, v) = \begin{cases} u^2 + (v-1)^2 - 1, & u \geqslant 0, \\ -u + (v-1)^2 - 1, & u \leqslant 0. \end{cases}$$

Then, if $(x_1, x_2) \in C$,

$$A(x_1, x_2) = \begin{cases} x_1^4 - x_1^2, & x_1 \geqslant 0, \\ x_1(x_1^3 - 2x_1 - 1), & x_1 \leqslant 0. \end{cases}$$

Since $t^3 - 2t - 1 < 0$ on $[-1/2, 0]$, it follows that $K = [0, 1] \times \{0\}$. Let $x^* = (0, 0)$. Then

$$F_1(x^*) = F_2(x^*) = 0,$$

$$F_1'(x^*) = (1, 0), \qquad F_2'(x^*) = (0, 1),$$

and so

$$\tilde{A}(x) = \begin{cases} x_1^2 + (x_2 - 1)^2 - 1, & x_1 \geqslant 0, \\ -x_1 + (x_2 - 1)^2 - 1, & x_1 \leqslant 0. \end{cases}$$

Thus,

(3.33)
$$x^* + \mathrm{CFD}(x^*) = \{(x_1, x_2) \in \mathbb{R}^2 : \ x_2^2 - 2x_2 \leqslant x_1 \leqslant \sqrt{1 - (x_2 - 1)^2}, \ 0 \leqslant x_2 \leqslant 2\},$$

which is the set bounded by a parabola Γ_1 and a semicircle Γ_2 whose tangents at x^* are of slopes $1/2$ and 0, respectively, and hence $x^* + \mathrm{cone}(\mathrm{CFD}(x^*))$ is the polyhedral cone generated by these two tangents. Consequently,

$$K_C(x^*) = \{(0, 0)\}, \qquad K_L(x^*) = [0, 1] \times \{0\},$$

so that

$$K_C(x^*) \subseteq K_S(x^*), \qquad K \subseteq K_L(x^*);$$

that is, x^* is a weakly regular point of S. Furthermore,

$$\partial \tilde{A}(x^*) = \partial A(x^*) = [-1, 0] \times \{-2\}.$$

For any $x = (x_1, x_2) \in X$, $P_K(x) = x^*$ if and only if $x_1 \leqslant 0$. Taking $\lambda = -x_1, \ h = -1$, we get that $P_C(x - \lambda h) = x^*$. This implies that x^* has the perturbation property with respect to C and the system (NIS). However, note that

$$N_{K_C(x^*)}(x^*) = \mathbb{R}^2, \qquad N_C(x^*) = \{0\} \times \mathbb{R},$$

$$\mathrm{cone}(\partial \tilde{A}(x^*)) = \{(x_1, x_2) \in \mathbb{R}^2 : \ x_2 \leqslant 2x_1 \leqslant 0\}.$$

This implies that the system $\tilde{A}(\cdot) \leqslant 0$ does not satisfy the BCQ relative to C. Thus (ii) does not imply (i) in Theorem 3.9 if the condition (3.9) is dropped.

Example 3.2. Let H, F, x^* be defined as in Example 3.1, but let C be defined by

$$C = \{(x_1, x_2) : \ -2x_2 \leqslant x_1 \leqslant 1, \ x_2 \in [0, 1]\}.$$

By Example 3.1, we obtain that $K_L(x^*) = C \supseteq K$. Moreover,

$$\mathrm{ri}(x^* + \mathrm{cone}(\mathrm{CFD}(x^*))) \cap \mathrm{ri} C \neq \emptyset. \tag{3.34}$$

Since $\{F_1'(x^*), F_2'(x^*)\}$ is linearly independent, $K_C(x^*) \subseteq K_S(x^*)$ by Corollary 3.14. It follows that x^* is a weakly regular point of K; hence, by (3.34), the assumptions of Theorem 3.9 are satisfied, and so (i) and (ii) are equivalent. However, $K \nsubseteq K_C(x^*)$ because $(0, 1] \times \{0\} \subset K$, but $(0, 1] \times \{0\}$ is disjoint from $K_C(x^*)$. Hence, Theorem 3.7 cannot be applied. Therefore, in the case in which (3.9) holds, Theorem 3.9 is a genuine extension of Theorem 3.7.

4. Inequality system with respect to cones. In this section, we will apply the results obtained to study an abstract inequality system. Let X, C be as before. Let W be a closed convex cone in $\mathbb{R}^N$. Then W defines a partial order $\succcurlyeq$ on $\mathbb{R}^N$:

$$\bar{a} \succcurlyeq \bar{b} \Longleftrightarrow \bar{a} - \bar{b} \in W. \tag{4.1}$$

Let $G = (g_1, g_2, \dots, g_N)$ be a Fréchet differentiable function from X to $\mathbb{R}^N$, and let $b = (b_1, b_2, \dots, b_N) \in \mathbb{R}^N$. Let $\hat{K}$ consist of all $x \in C$ satisfying the abstract inequality system

$$G(x) \succcurlyeq b, \tag{AIS}$$

namely,

$$\hat{K} = C \cap \{x \in X : \ G(x) \in b + W\}. \tag{4.2}$$

Let $x^* \in \hat{K}$. This system can be rephrased as a system of the form (NIS) by the following device. Define $H : \ \mathbb{R}^N \to \mathbb{R}$ by the Euclidean distance function of W:

$$H(y) = \operatorname{dist}(y, W), \quad y \in \mathbb{R}^N. \tag{4.3}$$

Then $H(\cdot) \geqslant 0$ on $\mathbb{R}^N$, $H(G(x^*) - b) = 0$, $W = \{y \in \mathbb{R}^N : \ H(y) = 0\}$, and, by [13, Example 3.3, p. 259],

$$\partial H(y) = N_W(y) \cap \mathbf{B}(0,1) \quad \forall y \in W, \tag{4.4}$$

where $\mathbf{B}(0,1)$ denotes the unit ball of $\mathbb{R}^N$. Note that x satisfies (AIS) if and only if

$$H(G(x) - b) \leqslant 0. \tag{4.5}$$

Clearly (4.5) is of the type (NIS) with $m = 1$. According to the notation arrangements in sections 1 and 2, let $F, A, \tilde{A}$ be defined by, for each $x \in X$,

$$\left.\begin{array}{l} F(x) = G(x) - b \\ A(x) = H(G(x)) \\ \tilde{A}(x) = H(F(x^*) + F'(x^*)(x - x^*)) \\ \tilde{F}(x) = F(x^*) + F'(x^*)(x - x^*) \end{array}\right\}. \tag{4.6}$$

Let $J(x^*) = \{j : \ g_j(x^*) = b_j\}$.

THEOREM 4.1. *Let $x^* \in \hat{K}$, and suppose that x^* is regular with respect to C and the system* (4.5). *Then the following statements are equivalent:*

(i) $N_{C \cap \tilde{F}^{-1}(W)}(x^*) = N_C(x^*) + N_W(G(x^*) - b) \circ G'(x^*)$;

(ii) *for any $x \in X$, $P_{\hat{K}}(x) = x^* \Longleftrightarrow P_C(x - \sum_{i=1}^N y_i g_i'(x^*)) = x^*$ for some $(y_1, y_2, \dots, y_N) \in W^\circ$ with $\sum_{i \notin J(x^*)} y_i (g_i(x^*) - b_i) = 0$.*

Proof. Clearly, $\tilde{A}(x^*) = A(x^*) = 0$, $G'(\cdot) = F'(\cdot)$, and $\tilde{A}^{-1}(\mathbb{R}_-) = \tilde{F}^{-1}(W)$. By Proposition 2.5 and (4.4), we have

$$\operatorname{cone}\partial \tilde{A}(x^*) = N_W(F(x^*)) \circ F'(x^*). \tag{4.7}$$

Then (i) holds if and only if the convexification system

$$\tilde{A}(x) \leqslant 0 \tag{4.8}$$

corresponding to (4.5) satisfies the BCQ relative to C at x^*.

On the other hand, it is well known and easy to verify that

$$N_W(F(x^*)) = \{y \in W^\circ : \langle y, F(x^*)\rangle = 0\} = \left\{(y_1, y_2, \dots, y_N) \in W^\circ : \sum_{i=1}^N y_i(g_i(x^*) - b_i) = 0\right\} = \left\{(y_1, y_2, \dots, y_N) \in W^\circ : \sum_{i \notin J(x^*)} y_i(g_i(x^*) - b_i) = 0\right\}. \tag{4.9}$$

Combining this with (4.7), one has

$$\text{cone}\partial\tilde{A}(x^*) = \left\{\sum_{i=1}^N y_i g_i'(x^*) : (y_1, y_2, \dots, y_N) \in W^\circ; \sum_{i \notin J(x^*)} y_i(g_i(x^*) - b_i) = 0\right\}.$$

Thus, (ii) is exactly the perturbation property with respect to C and system (4.5). Therefore Theorem 4.1 follows from Theorem 3.7. □

Remark 4.1. Since W is a closed cone, the regularity assumption is equivalent to the weak regularity of x^* (see Proposition 2.3).

An important special case of (AIS) considered above is the following familiar inequality-equality system: $x \in C$ and

$$\begin{cases} g_i(x) = b_i, & i = 1, 2, \dots, m_e, \\ g_i(x) \leqslant b_i, & i = m_e + 1, 2, \dots, m, \end{cases} \tag{4.10}$$

where $m_e \in \{1, 2, \dots .m\}$. Writing N for m and letting

$$W = \{(y_1, y_2, \dots, y_m) : y_i = 0 \ \forall i = 1, \dots, m_e; y_i \leqslant 0 \ \forall i = m_e + 1, \dots, m\}, \tag{4.11}$$

we see that the system (4.10) is of the type considered in (AIS). Let K consist of all $x \in C$ satisfying (4.10), and let $x^* \in K$. Let $I(x^*)$ consist of all i satisfying $g_i(x^*) = b_i$, and let $I_0(x^*) = I(x^*) \setminus \{1, 2, \dots, m_e\}$; thus $I(x^*) := I_0(x^*) \cup \{1, 2, \dots, m_e\}$. We define

$$C_i = \{x \in X : g_i'(x^*)(x - x^*) = 0\}, \quad i \in \{1, 2, \dots, m_e\},$$

$$C_i = \{x \in X : g_i'(x^*)(x - x^*) \leqslant 0\}, \quad i \in I_0(x^*).$$

The following facts are well known (and easy to verify):

$$\left.\begin{array}{l} N_{C_i}(x^*) = \text{span}\{g_i'(x^*)\}, \quad i \in \{1, 2, \dots, m_e\}, \\ N_{C_i}(x^*) = \text{cone}\{g_i'(x^*)\}, \quad i \in I_0(x^*), \\ W^\circ = \{(\lambda_1, \lambda_2, \dots, \lambda_m) : \lambda_i \geqslant 0 \ \forall i = m_e + 1, \dots, m\}. \end{array}\right\} \tag{4.12}$$

Corollary 4.2 (see Theorem 4.1 of [14]). *Let $x^* \in K$ be a regular point with respect to C and the system* (4.10). *Let*

$$D = \{x \in X; \ g_i'(x^*)(x - x^*) = 0 \ \forall i \in \{1, 2, \dots, m_e\}\}.$$

Then the following statements are equivalent:

(i) $\{C, D, C_i : i \in I_0(x^*)\}$ *has the strong CHIP at x^*;*

(ii) *for any* $x \in X$, $P_K(x) = x^* \iff P_C(x - \sum_1^m \lambda_i g_i'(x^*)) = x^*$ *for some* λ_i, $i = 1, \dots, m$, *with*

$$\begin{cases} \lambda_i \geqslant 0 & \forall i \in I_0(x^*), \\ \lambda_i = 0 & \forall i \notin I(x^*). \end{cases}$$

Proof. Let $G : x \mapsto (g_1(x), \dots, g_m(x))$, and let $H, F, A, \tilde{A}, \tilde{F}$ be defined as in (4.3), (4.6). Then,

$$N_{C \cap \tilde{F}^{-1}(W)}(x^*) = N_{C \cap (\cap_{i \in I(x^*)} C_i)}(x^*).$$

Moreover, by (4.12) and (4), one has

$$\begin{aligned} N_W(F(x^*)) &= \left\{ (\lambda_1, \lambda_2, \dots, \lambda_m) \in W^\circ : \sum_{i=1}^m \lambda_i (g_i(x^*) - b_i) = 0 \right\} \\ &= \{(\lambda_1, \lambda_2, \dots, \lambda_m) \in \mathbb{R}^m : \lambda_i \geqslant 0 \ \forall i \geqslant m_e + 1; \lambda_i = 0 \ \forall i \notin I(x^*)\} \end{aligned}$$

and

$$N_W(F(x^*)) \circ F'(x^*) = \left\{ \sum_{i=1}^m \lambda_i g_i'(x^*) : \lambda_i \geqslant 0 \ \forall i \geqslant m_e + 1; \lambda_i = 0 \ \forall i \notin I(x^*) \right\}.$$

Thus, by (4.12),

$$N_W(F(x^*)) \circ F'(x^*) = N_D(x^*) + \sum_{i \in I_0(x^*)} N_{C_i}(x^*).$$

Hence (i) and (ii) are the same as (i) and (ii), respectively, of Theorem 4.1. Therefore Corollary 4.2 follows from Theorem 4.1. □

Finally, we should point out that the results in this paper can be applied to the case when our Hilbert space X is over the complex field $\mathbb{C}$. For the remainder of the paper, let X be a complex Hilbert space and F_j be a Fréchet differentiable complex function defined on X for each $j = 1, 2, \dots, m$. Let $V_1, V_2, \dots, V_m$ be convex closed subsets of the complex plane $\mathbb{C}$. Let C be a closed convex subset of X, and let K consist of all $x \in C$ satisfying the complex system

$$F_j(x) \in V_j, \quad j = 1, 2, \dots, m. \tag{CS}$$

As usual, $\mathbb{C}$ can be metrically viewed as $\mathbb{R}^2$, while X can be regarded as a real Hilbert space with the inner product defined by

$$\langle x, y \rangle_R = \operatorname{Re}\langle x, y \rangle, \quad x, y \in X.$$

Consequently, F_j is a mapping from X into $\mathbb{R}^2$, and V_j is a closed convex subset of $\mathbb{R}^2$. Let $H_j : \mathbb{R}^2 \to \mathbb{R}$ denote the distance function to V_j. Then H_j is a real-valued convex function on $\mathbb{R}^2$ such that

$$V_j = \{y \in \mathbb{C} : H_j(y) \leqslant 0\},$$

and hence K consists of all $x \in C$ satisfying the real system

$$H_j(F_j(x)) \leqslant 0, \quad j = 1, 2, \dots, m. \tag{RS}$$

Thus, Theorems 3.7 and 3.9 can then be applied in a manner similar to what we have done for Theorem 4.1; details need not be repeated here. However, it is worth pointing out that the approach of using $\mathrm{Re}F_j$ and $\mathrm{Im}F_j$ does not work here because, for general closed convex sets V_j, the constraint $F_j(x) \in V_j$ cannot be described by $\mathrm{Re}F_j$ and $\mathrm{Im}F_j$ separately.

REFERENCES

[1] V. Barbu and T. Precupanu, *Convexity and Optimization in Banach Spaces*, Sijthoff & Noordhoff, Groninger, The Netherlands, 1978.

[2] C. De Boor, *On "best" interpolation*, J. Approx. Theory, 16 (1976), pp. 28–48.

[3] B. Brosowski and F. Deutsch, *On some geometric properties of suns*, J. Approx. Theory, 10 (1974), pp. 245–267.

[4] C. Chui, F. Deutsch, and J. Ward, *Constrained best approximation in Hilbert space*, Constr. Approx., 6 (1990), pp. 35–64.

[5] C. Chui, F. Deutsch, and J. Ward, *Constrained best approximation in Hilbert space* II, J. Approx. Theory, 71 (1992), pp. 231–238.

[6] F. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley & Sons, New York, 1983.

[7] 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.

[8] F. Deutsch, W. Li, and J. Swetits, *Fenchel duality and the strong conical intersection property*, J. Optim. Theory Appl., 102 (1997), pp. 681–695.

[9] F. Deutsch, W. Li, and J. Ward, *A dual approach to constrained interpolation from a convex subset of Hilbert space*, J. Approx. Theory, 90 (1997), pp. 385–444.

[10] 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.

[11] F. Deutsch, V. Ubhaya, J. Ward, and Y. Xu, *Constrained best approximation in Hilbert space* III*: Application to n-convex functions*, Constr. Approx., 12 (1996), pp. 361–384.

[12] 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., 86 (1999), pp. 135–160.

[13] J. Hiriart-Urruty and C. Lemarechal, *Convex Analysis and Minimization Algorithms* I, Grundlehren Math. Wiss. 305, Springer, New York, 1993.

[14] C. Li and X. Q. Jin, *Nonlinearly constrained best approximation in Hilbert spaces, the strong conical hull intersection property and the basic constraints qualification condition*, SIAM J. Optim., 13 (2002), pp. 228–239.

[15] O. Mangasarian, *Nonlinear Programming*, McGraw–Hill, New York, 1969.

[16] O. L. Mangasarian and S. Fromowitz, *The Fritz John necessary optimality conditions in the presence of equality constraints*, J. Math. Anal. Appl., 17 (1967), pp. 37–47.

[17] C. Micchelli, P. Smith, J. Swetits, and J. Ward, *Constrained L_p-approximation*, Constr. Approx., 1 (1985), pp. 93–102.

[18] C. A. Micchelli and F. I. Utreras, *Smoothing and interpolation in a convex subset of a Hilbert space*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 728–746.

[19] I. Singer, *Duality for optimization and best approximation over finite intersection*, Numer. Funct. Anal. Optim., 19 (1998), pp. 903–915.

[20] R. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[21] Y. Yuan and W. Sun, *Optimization Theory and Methods*, Science Press, Beijing, 1997 (in Chinese).

SIAM J. OPTIM.
Vol. 13, No. 3, pp. 745–765

COUPLING GENERAL PENALTY SCHEMES FOR CONVEX PROGRAMMING WITH THE STEEPEST DESCENT AND THE PROXIMAL POINT ALGORITHM*

R. COMINETTI† AND M. COURDURIER†

Abstract. We study the asymptotic behavior of the continuous flow generated by coupling the steepest descent method with a general class of penalty schemes for convex programming. We establish convergence of the trajectories towards primal optimal solutions, as well as convergence of some naturally associated dual paths. The results are extended to the sequences generated by an implicit discretization scheme which corresponds to the coupling of an inexact proximal point iteration with the penalty schemes.

Key words. subgradient inclusions, prox method, penalty schemes, convex programming

AMS subject classifications. 34C35, 34D05, 49M10, 49M30, 90C25

PII. S1052623401397242

1. Introduction.

Let $I = \{1, \dots, m\}$ and consider the mathematical program

$$\text{(P)} \qquad v = \min_{x \in \mathbb{R}^n} \{ f_0(x) : f_i(x) \le 0, i \in I \}$$

together with the penalty approximation scheme defined for $r > 0$ by

$$\text{(P}_r\text{)} \qquad v_r = \min_{x \in \mathbb{R}^n} f(x, r)$$

with $f(x,r) = f_0(x) + r \sum_{i \in I} \theta(f_i(x)/r)$. Throughout the paper we assume

$$\text{(H}_0\text{)} \quad \begin{cases} \text{(a) } f_i : \mathbb{R}^n \to \mathbb{R} \text{ is convex for } i = 0, \dots, m; \\ \text{(b) the optimal solution set } S(\text{P}) \text{ is nonempty and bounded;} \\ \text{(c) } \theta : (-\infty, \kappa) \to \mathbb{R} \text{ is smooth and convex with } \kappa \in [0, \infty]; \\ \text{(d) } \theta'(u) > 0 \text{ with } \theta'(u) \to 0 \text{ for } u \to -\infty \text{ and } \theta'(u) \to \infty \text{ for } u \to \kappa; \\ \text{(e) if } \kappa = 0, \text{ there is a Slater point } \bar{x} \text{ with } f_i(\bar{x}) < 0, \forall\, i \in I. \end{cases}$$

When $\kappa < \infty$ we extend θ by setting $\theta(\kappa) = \lim_{u \to \kappa} \theta(u)$ and $\theta(u) = \infty$ for $u > \kappa$. Under these circumstances (P_r) admits an optimal path $x(r)$ and we have $v_r \to v$ and $d(x(r), S(\text{P})) \to 0$ when $r \to 0$, where $d(x, S) = \inf_{y \in S} \|x - y\|$ denotes the distance from the point x to the set S. With further assumptions the path $x(r)$ is uniquely determined and converges to an optimal solution $x^\theta \in S(\text{P})$ (see section 2).

EXAMPLES. *A number of penalty functions have been proposed* [9, 13, 14, 17, 19, 25, 26, 42]*, such as the exponential penalty* $\theta(u) = \exp(u)$ *and the inverse and log-barriers defined for* $u < 0$ *by* $\theta(u) = -1/u$ *and* $\theta(u) = -\ln(-u)$*, respectively. Further examples include the root-barrier* $\theta(u) = -\sqrt{-u}$ *for* $u \le 0$*, the shifted-barriers*

*Received by the editors October 31, 2001; accepted for publication (in revised form) May 9, 2002; published electronically November 14, 2002. This research was partially supported by the FONDAP Program in Applied Mathematics.

http://www.siam.org/journals/siopt/13-3/39724.html

†Universidad de Chile, Departamento de Ingeniería Matemática and Centro de Modelamiento Matemático, Casilla 170/3, Correo 3, Santiago, Chile (rcominet@dim.uchile.cl, mcourdur@dim.uchile.cl).

$u \mapsto \theta(u-1)$ with θ the inverse or log-barrier ($\kappa = 1$), as well as other penalties built by gluing together two or more convex functions.

The standard approach in penalty methods is to closely trace the optimal path $x(r)$ by using some numerical method to estimate $x(r_k)$ for a sequence of penalty parameters $r_k \to 0$. However, approximating $x(r_k)$ with precision may be expensive and unnecessary to generate a convergent algorithm. For instance, one might expect that an economic algorithm performing a single iteration of a descent method for each value r_k could still generate a sequence converging to an optimal solution of (P), regardless of the fact that the iterates approach the optimal path $x(r)$ or not. Such an algorithm may no longer be interpreted as a path-following on the optimal trajectory, and we must rather think of it as a discretization of an underlying continuous flow. If this flow converges to the optimal set, we expect the discrete iterates to be driven towards $S(\mathrm{P})$ by following the stream and not necessarily one specific trajectory. More precisely, since every point lies on an integral curve of the flow, we may think of the discrete iterates as jumping from one streamline to a close but different one, and since all these curves lead to $S(\mathrm{P})$ we may still expect the discrete process to be stable and convergent. Thus, instead of allocating computational resources to restore the proximity to a prespecified trajectory such as the optimal path $x(r)$, we may let the iterates evolve freely along successive integral curves. Proceeding in this way we might expect to attain convergence with a reduced overall computational cost.

In this paper we focus on the flow generated by coupling steepest descent with the penalty scheme, namely

$$\dot{u}(t) \in -\partial f(u(t), r(t)) \quad \text{a.e. } t \geq 0, \tag{SDP}$$

where $r(t)$ is a positive real function tending to 0 as $t \to \infty$, and $\partial f(x, r)$ denotes the subdifferential of $f(\cdot, r)$ at the point x. Under mild assumptions we prove that, no matter how we choose $r(t)$, every solution of (SDP) converges when $t \to \infty$ to a point $u_\infty \in S(\mathrm{P})$ (which may or may not depend on the initial condition $u(0) = u_0$). The freedom in the choice of $r(t)$ provides much flexibility to control the dynamics of (SDP), allowing us even to adapt $r(t)$ on-line depending on the behavior of $u(t)$.

Thus, (SDP) provides a convergent flow whose integral curves could be approximately traced by using a variety of numerical integration schemes such as Adams, Runge–Kutta, extrapolation, Newton-like methods, etc. In section 4 we study a class of numerical methods based on the implicit discretization scheme

$$\frac{u_k - u_{k-1}}{\lambda_k} \in -\partial f(u_k, r_k), \tag{PP}$$

where $r_k \to 0$ and $\lambda_k > 0$ with $\sum \lambda_k = \infty$ (this is the discrete analogue of $t \to \infty$). This iteration may be seen as performing a single proximal point step for the minimization of $f(\cdot, r_k)$ for each value of the penalty parameter r_k. The same assumptions ensuring the convergence of the continuous flow imply that the discrete iterates u_k converge to a point $u_\infty \in S(\mathrm{P})$. To turn (PP) into an implementable algorithm we consider in fact an inexact prox-penalty iteration of the form

$$\frac{u_k - u_{k-1}}{\lambda_k} \in -\partial_{\varepsilon_k} f(u_k, r_k) + \nu_k, \tag{IPP}$$

where we allow for a residual ν_k and we replace the exact by the approximate subdifferential $\partial_{\varepsilon_k} f(u_k, r_k)$. We show that convergence of such an iteration is preserved as long as the errors satisfy $\sum[\varepsilon_k + \|\nu_k\|]\lambda_k < \infty$. As in the continuous case, the

flexibility in the choice of r_k provides a support for algorithms in which the penalty parameter is updated on-line (e.g., to handle numerical instabilities), so that r_k is not known in advance and one cannot impose a priori conditions on it.

We remark that (IPP) is a generic method in the sense that we do not specify how the point u_k is to be computed. For instance, this could be done by using a bundle-type method (with $\varepsilon_k > 0, \nu_k = 0$). Alternatively, noting that the solution u_k of (PP) is the unique minimizer of the strongly convex function $\varphi_k(u) = f(u, r_k) + \frac{1}{2\lambda_k}\|u - u_{k-1}\|^2$, the point u_k may be computed by using an unconstrained minimization method on φ_k. In particular, if φ_k is smooth, we may take $\varepsilon_k = 0$, in which case $\nu_k = \nabla\varphi_k(u_k)$ and then u_k may be computed by using a Newton-like method with a stopping criterion guaranteeing $\sum \|\nabla\varphi_k(u_k)\|\lambda_k < \infty$. Note also that by adjusting the stepsize λ_k one may reduce the distance from u_{k-1} to the minimizer of φ_k, controlling indirectly the number of steps required by the unconstrained method to find the next iterate u_k.

In order to put our results into perspective, we remark that (SDP) falls in the framework of subgradient evolution equations of the form $\dot{u}(t) \in -\partial\varphi_t(u(t))$, where $\{\varphi_t\}_{t\geq 0}$ is a family of closed proper convex functions. In the autonomous case $\varphi_t \equiv \varphi$, a well-known result [15, 16] states that $u(t)$ converges to a minimizer $u^\infty \in \mathrm{Argmin}(\varphi)$, provided the latter set is nonempty. This result was extended to the nonautonomous case in [27], assuming that φ_t converges in an appropriate sense and sufficiently *fast* toward φ. More recently, considering (SDP) for general classes of approximation schemes, convergence of $u(t)$ was established when $r(t)$ tends to 0 sufficiently *slow* [7, 21] or sufficiently *fast* [2, 21]. Unfortunately, for the penalty scheme (P_r) this slow/fast alternative may not cover all possible parameter functions. However, in the recent paper [12] it was shown that in the special case of the exponential penalty for linear programming, any solution of (SDP) tends to a point $u_\infty \in S(\mathrm{P})$ with no restriction on the rate of convergence of $r(t) \to 0$. Our results for (SDP) extend the latter by considering more general penalty schemes as well as nonlinear convex programs.

Concerning (IPP) there exist many previous works such as [1, 10, 11, 20, 31, 32, 34, 35, 39, 40, 41, 43] as well as many other references therein (see [5] for a short survey). A point common to all these works is that, in order to attain convergence, restrictive conditions are imposed on the sequence r_k. In contrast, [5] establishes the discrete analogues of the results in [12] for the exponential penalty in linear programming: for all $r_k \to 0$ any sequence generated by (IPP) converges to a point $u_\infty \in S(\mathrm{P})$ (provided that $\sum \lambda_k = \infty$ and $\sum[\varepsilon_k + \|\nu_k\|]\lambda_k < \infty$). Our results for (IPP) extend the latter to general penalty schemes and nonlinear convex programs.

To conclude this introduction let us mention that, in addition to the convergence of the *primal* trajectories $u(t)$ and u_k generated by the dynamics (SDP) and (IPP), in the case of a linear program (P) and under some extra conditions on the parameters defining the dynamics, we also establish the convergence of some naturally associated *dual* paths $\mu(t)$ and μ_k towards a special dual optimal solution λ^θ. These results generalize the corresponding dual convergence results established for the exponential penalty in [5, 12].

The plan of the paper is as follows. In section 2 we review the convergence theory for the penalty scheme (P_r) and a corresponding dual scheme (D_r). In section 3 we analyze the convergence of the trajectories generated by the steepest-descent-penalty (SDP), as well as the associated dual trajectories. Finally, in section 4 we study the primal and dual convergence of the inexact prox-penalty iteration (IPP).

Throughout the paper we adopt the standard notations in convex analysis [28, 44].

In particular, for φ closed proper and convex we denote φ^* its Fenchel conjugate, $\partial\varphi(x)$ and $\partial_\varepsilon\varphi(x)$ its exact and approximate subdifferentials, and φ^∞ its recession function defined as usual by $\varphi^\infty(d) = \lim_{k\to\infty}[\varphi(x+t_kd_k)-\varphi(x)]/t_k$ for any sequences $t_k \to \infty$, $d_k \to d$, and $x \in \mathrm{dom}(\varphi)$. For $A, B \subset \mathbb{R}^n$ we denote $e(A,B) = \sup_{x\in A} d(x,B)$ the excess of A over B, where $d(x,B) = \inf_{y\in B}\|x-y\|$ is the distance from x to B. Finally, we will denote $\mathrm{aff}(C)$ the affine space spanned by the set $C \subset \mathbb{R}^n$.

2. The penalty scheme. In this section we present a short review of the convergence results for the primal and dual optimal paths associated with (P_r).

2.1. Primal convergence. The following result describes the rationale behind the penalty scheme (P_r). Its proof is rather standard and it can be found in [9].

PROPOSITION 2.1. *Under* (H_0)*, for all* $r > 0$ *the optimal solution set* $S(\mathrm{P}_r)$ *is nonempty and bounded. Moreover, when* $r \downarrow 0$ *we have* $v_r \to v$ *and* $d(x(r), S(\mathrm{P})) \to 0$ *for every optimal path* $x(r) \in S(\mathrm{P}_r)$.

The optimal path $x(r)$ is uniquely determined (see [22]) if we have in addition

(H_1) θ *is strictly convex on* $(-\infty, \kappa)$ and

(H_2) $f_0, \dots, f_m \in \mathcal{Q}$,

where $\mathcal{Q}$ denotes the class of convex functions f which are constant on $\mathrm{aff}(C)$ whenever C is convex and f is constant on C. This class has been considered in [3, 38] and is closely related to the concept of *faithful convexity* [45]. It contains all strictly convex, as well as linear, quadratic, and analytic convex functions. Moreover, $\mathcal{Q}$ is closed under composition: $\sigma\circ f\circ A \in \mathcal{Q}$ whenever σ is increasing and convex, $f \in \mathcal{Q}$, and A is affine.

Proposition 2.1 implies that $x(r)$ is bounded with all its cluster points in $S(\mathrm{P})$. With more structure one can prove that $x(r)$ converges to a particular point $x^\theta \in S(\mathrm{P})$. To prove such a result let $D = \{1, \dots, d\}$ and consider the θ-mean $M_\theta : (-\infty, \kappa)^d \to \mathbb{R}$ and the asymptotic θ-mean $A_\theta : (-\infty, 0)^d \to \mathbb{R}$ defined by

$$M_\theta(v) = \theta^{-1}\left(\tfrac{1}{d}\textstyle\sum_{i\in D}\theta(v_i)\right),$$
$$A_\theta(v) = \lim_{r\to 0} rM_\theta(v/r),$$

where we assume implicitly that

(H_3) *the limit* $A_\theta(v)$ *exists.*

Although for each dimension d we have different means M_θ^d and A_θ^d, we omit the superscript, as d will be clear from the context. Notice that A_θ depends only on the behavior of θ at $-\infty$: if $\theta_1^{-1}\circ\theta_2(u)/u$ converges to some $\beta \in (0,\infty)$ when $u \to -\infty$, then the limit A_{θ_1} exists if and only if A_{θ_2} exists, in which case both asymptotic means coincide. Also, if θ is $\mathcal{C}^2$ with $\theta''(u)/\theta'(u) \to \beta \in (0,\infty)$ for $u \to -\infty$, then (H_3) holds with $A_\theta(v) = \max_{i\in D} v_i$. The proof of these facts can be found in [22].

EXAMPLES. *For the penalty functions mentioned in the introduction,* (H_3) *holds with*

$$\begin{aligned}
A_\theta(v) &= \max_{i\in D} v_i && (\textit{exp-penalty}),\\
A_\theta(v) &= [\tfrac{1}{d}\textstyle\sum_{i\in D} 1/v_i]^{-1} && (\textit{inverse-barrier}),\\
A_\theta(v) &= -[\Pi_{i\in D}(-v_i)]^{1/d} && (\textit{log-barrier}),\\
A_\theta(v) &= -[\tfrac{1}{d}\textstyle\sum_{i\in D}\sqrt{-v_i}]^2 && (\textit{root-barrier}).
\end{aligned}$$

Moreover, since A_θ *depends only on the behavior of* θ *at* $-\infty$*, it follows that for most penalties found in the literature the limit* A_θ *exists and coincides with one of the above.*

In general (see [22]) the function A_θ is convex, continuous, symmetric, positively homogeneous, and componentwise nondecreasing, with

$$\frac{1}{d}\sum_{i\in D} v_i \le A_\theta(v) \le \max_{i\in D} v_i, \tag{2.1}$$

and it can be uniquely extended from $(-\infty,0)^d$ to $\mathbb{R}^d_-$ preserving all these properties. In what follows we keep the notation A_θ to denote this extension, and we suppose in addition that for all $u,v\in\mathbb{R}^d_-$ we have

$$\max_{i\in D} u_i \ne \max_{i\in D} v_i \;\Rightarrow\; \min_{w\in[u,v]} A_\theta(w) < \max\{A_\theta(u), A_\theta(v)\}. \tag{H$_4$}$$

This property is rather weak and it holds, for instance, whenever A_θ is strictly convex or the max function as in the examples above.

DEFINITION 2.2. *For each closed convex nonempty $C\subset S(P)$ we let $E_C=\mathrm{aff}(C)$ and $I_C=\{i\in I: f_i$ is nonconstant on $C\}$. When I_C is nonempty we set $v_C=\min\varphi_C$ and $S_C=\mathrm{argmin}(\varphi_C)$ with $\varphi_C: E_C\to\mathbb{R}\cup\{\infty\}$ given by*

$$\varphi_C(x)=\begin{cases} A_\theta(f_i(x): i\in I_C) & \text{if } x\in C,\\ \infty & \text{otherwise.}\end{cases}$$

The following basic properties will be used throughout the paper.

LEMMA 2.3. *Let $C\subset S(\mathrm{P})$ be closed convex and nonempty.*

(a) *If $I_C\ne\phi$, then there exists $\hat{x}\in C$ with $f_i(\hat{x})<0$ for all $i\in I_C$.*

(b) *Assume* (H$_2$) *and let $d=y-x$ with $x,y\in C$. Then $f_i^\infty(\pm d)=0$ for all $i\notin I_C$.*

(c) *Assume* (H$_2$) *and* (H$_0$)(b). *Then $I_C=\phi$ if and only if C is a singleton.*

(d) *Assume* (H$_3$) *and let $v_j\to v\in\mathbb{R}^d_-$ and $r_j\to 0$. Then $A_\theta(v)\le\liminf r_jM_\theta(v_j/r_j)$, and when $v\in(-\infty,0)^d$ we have $A_\theta(v)=\lim r_jM_\theta(v_j/r_j)$.*

(e) *Assume* (H$_3$), (H$_4$), *and $I_C\ne\phi$. Then S_C is a proper subset of C and there exist $\beta\le 0$ and $j\in I_C$ with $f_j(x)=\max_{i\in I_C} f_i(x)=\beta$ for all $x\in S_C$.*

Proof. (a) For each $i\in I_C$ take $x_i\in C$ with $f_i(x_i)<0$. Then (a) holds with $\hat{x}=\frac{1}{|I_C|}\sum_{i\in I_C} x_i$ (notice that since $C\subset S(\mathrm{P})$ we have $f_i(x)\le 0$ for all $i\in I$ and $x\in C$).

(b) For $i\notin I_C$ the function f_i is constant on $[x,y]$ so that (H$_2$) implies that it is constant on the line passing through x and y. Hence d is a constancy direction for f_i.

(c) Clearly when C is a singleton we have $I_C=\phi$. Conversely, if $I_C=\phi$, then for $x,y\in C$ we have that $d=y-x$ is a constancy direction for all the f_i's, including $i=0$ since $C\subset S(\mathrm{P})$. The boundedness assumption (H$_0$)(b) implies $d=0$ so that $x=y$.

(d) Take $\varepsilon>0$ and let 1_d be a d-dimensional vector with all entries equal to 1. For j large we have $r_jM_\theta([v-\varepsilon 1_d]/r_j)\le r_jM_\theta(v_j/r_j)$ so that letting $j\to\infty$ we get $A_\theta(v-\varepsilon 1_d)\le\liminf r_jM_\theta(v_j/r_j)$. Making $\varepsilon\downarrow 0$ and using the continuity of A_θ, it follows that $A_\theta(v)\le\liminf r_jM_\theta(v_j/r_j)$. A similar argument with $v+\varepsilon 1_d$ shows that when $v\in(-\infty,0)^d$ we have $\limsup r_jM_\theta(v_j/r_j)\le A_\theta(v)$.

(e) We claim that $\max\{f_i(x): i\in I_C\}$ is constant on S_C. Indeed, suppose there exist $x,y\in S_C$ with $\max\{f_i(x): i\in I_C\}\ne\max\{f_i(y): i\in I_C\}$. Denoting $v(z)=(f_i(z): i\in I_C)$ we have $\varphi_C(z)=A_\theta(v(z))$, so that using the convexity of the f_i's, the componentwise monotonicity of A_θ, and (H$_4$), we get the contradiction

$$\min_{z\in[x,y]}\varphi_C(z)\le\min_{w\in[v(x),v(y)]} A_\theta(w)<\max\{A_\theta(v(x)),A_\theta(v(y))\}=v_C.$$

Let $\beta \le 0$ be the constant value of $\max\{f_i(x): i \in I_C\}$ on S_C. Then there is $j \in I_C$ such that $f_j(x) = \beta$ for all $x \in S_C$ (otherwise for each $i \in I_C$ we could find $x^i \in S_C$ with $f_i(x^i) < \beta$ and the point $\bar{x} = \frac{1}{|I_C|}\sum_{i\in I_C} x^i \in S_C$ would satisfy $\max\{f_i(\bar{x}): i \in I_C\} < \beta$, which is impossible). The latter also implies that the inclusion $S_C \subset C$ is strict. ∎

The previous facts allow us to establish the convergence of the optimal path $x(r)$.

THEOREM 2.4. *Assume* (H_0)–(H_4). *Then* (P_r) *has a unique optimal path* $x(r)$ *and we have* $x(r) \to x^\theta$ *as* $r \to 0$, *with* x^θ *the unique solution of the nested hierarchy of minimization problems defined recursively from* (P^0) ≡ (P) *by*

$$(P^{k+1}) \qquad v^{k+1} = \min_{x\in S^k} \varphi_{S^k}(x),$$

where S^k *denotes the optimal solution set of* (P^k).

Proof. The proof is a slight modification of [22, Theorem 5.2]. By Lemma 2.3(e) the set I_{S^k} is strictly decreasing so that $I_{S^k} = \phi$ for all k large and then Lemma 2.3(c) implies that S^k is a singleton $\{x^\theta\}$.

Now, Proposition 2.1 shows that $x(r)$ is bounded with all its cluster points in S^0. Let $x^* = \lim x_j \in S^0$ be such a cluster point, where $x_j = x(r_j)$ for some $r_j \to 0$. We show inductively that $x^* \in S^k$ for all k so that $x^* = x^\theta$ and then $x(r) \to x^\theta$. Suppose $x^* \in S^k$ and set $x^\varepsilon = (1-\varepsilon)x^\theta + \varepsilon\hat{x}$ with $\hat{x} \in S^k$ such that $f_i(\hat{x}) < 0$ for all $i \in I_{S^k}$ (see Lemma 2.3(a)) so that the same strict inequalities hold for x^ε. Setting $x_j^\varepsilon = x_j - x^* + x^\varepsilon$ and using Lemma 2.3(b) we have $f_i(x_j) = f_i(x_j^\varepsilon)$ for $i \notin I_{S^k}$, so that canceling the corresponding terms in the inequality $f(x_j, r_j) \le f(x_j^\varepsilon, r_j)$ a direct computation gives

$$r_j M_\theta(f_i(x_j)/r_j : i \in I_{S^k}) \le r_j M_\theta(f_i(x_j^\varepsilon)/r_j : i \in I_{S^k}).$$

Letting $j \to \infty$ and using Lemma 2.3(d) we get $\varphi_{S^k}(x^*) \le \varphi_{S^k}(x^\varepsilon)$, and then $\varepsilon \downarrow 0$ implies $\varphi_{S^k}(x^*) \le \varphi_{S^k}(x^\theta)$ so that $x^* \in S^{k+1}$, completing the induction step. ∎

We remark that smoothness of θ is not required for this result and also that (H_0)(e) can be replaced by $\theta(0) < \infty$. This theorem first appeared in [22] with a slightly stronger assumption (H_3). Under this form of (H_3) the result was proved in [18] where in addition a wider class of functions f_i is considered. Other previous results include the convergence of the (log-barrier) central path in linear programming [36, 37] and "quasi-analytic" convex programming [38], as well as convergence of the optimal paths for several other penalty functions in [3, 4, 9, 23]. All these results are covered by Theorem 2.4. See also [30] for the case where each constraint "$f_i(x) \le 0$" is penalized with a different penalty function $\theta_i(\cdot)$.

Remark 1. Note that (H_0)(d) implies $\theta^\infty(-1) = 0$ and $\theta^\infty(1) = \infty$, which is the assumption used in [22].

2.2. Dual approximation. Let (D_r) be the dual problem obtained from (P_r) using the perturbation function $\Psi_r(x, y) = f_0(x) + r\sum_{i\in I} \theta([f_i(x) + y_i]/r)$. This dual consists in minimizing the Fenchel conjugate $\Psi_r^*(0, \lambda)$, and a direct computation yields

$$(D_r) \qquad w_r = \min_{\lambda\in\mathbb{R}^m} p(\lambda) + r\sum_{i\in I} \theta^*(\lambda_i),$$

where $p(\lambda) = -\inf_{x\in\mathbb{R}^n}[f_0(x) + \sum_{i\in I} \lambda_i f_i(x)]$. Properties ($H_0$)(c)–(d) imply that θ^* is finite and strictly convex on $(0, \infty)$ with $\theta^*(\mu) = \infty$ for $\mu < 0$, so that (D_r) can be viewed as a *Tikhonov-like* approximation scheme for the dual problem

$$(D) \qquad w = \min_{\lambda\in\mathbb{R}^m} \{p(\lambda) : \lambda_i \ge 0, i \in I\}.$$

Taking $\bar{x}$ feasible for (P) if $\kappa > 0$ or a Slater point if $\kappa = 0$, we have that $\Psi_r(\bar{x}, \cdot)$ is finite and continuous at $y = 0$. Hence duality theory yields $v_r + w_r = 0$ and $S(\mathrm{D}_r) \neq \phi$, and therefore (D_r) has a unique optimal solution $\lambda(r)$. Moreover, if $x(r)$ solves (P_r), then $\lambda_i(r) = \theta'(f_i(x(r))/r)$. Concerning the asymptotic behavior of $\lambda(r)$, we have the following slight extension of [9, Corollary 2.6] and [9, Theorem 3.4].

THEOREM 2.5. *Assume* (H_0), $S(\mathrm{D}) \neq \phi$, *and suppose that either* θ *is bounded from below or* (P) *is a linear program. Then* (D_r) *has a unique optimal path* $\lambda(r)$ *and we have* $\lambda(r) \to \lambda^\theta$ *as* $r \to 0$, *with* λ^θ *the unique solution of*

$$(\mathrm{D}^1) \qquad \min_{\lambda \in S(D)} \sum_{i \in I_0} \theta^*(\lambda_i),$$

where $I_0 = \{i \in I : \lambda_i > 0$ *for some* $\lambda \in S(\mathrm{D})\}$. *Moreover,* $\lambda_i^\theta > 0$ *for all* $i \in I_0$.

The results in [9] assume θ strictly convex with $\theta(u) \to \infty$ as $u \to \kappa$. However, a careful look at the proofs reveals that strict convexity is never used and the divergence condition serves only to ensure $f_i(x(r))/r < \kappa$ for which $(\mathrm{H}_0)(\mathrm{d})$ suffices. Similar results for Tikhonov-like approximations can be found in [6, 23, 29].

3. The steepest-descent-penalty trajectory. In this section we study the coupling of the steepest descent method with penalty schemes as in (SDP). Our analysis is based on two convergence results for nonautonomous subgradient evolution equations developed in section 3.1. The convergence of the primal steepest-descent-penalty trajectories is established in section 3.2, while section 3.3 is concerned with the convergence of some naturally associated dual trajectories. The results in this section are an outgrowth of [24] which is itself an extension of [12].

3.1. Convergence of subgradient inclusions. Let $\varphi_t : E \to \mathbb{R} \cup \{\infty\}$ for $t \geq 0$ be a family of convex functions defined on a finite dimensional Euclidean space E, and let $\alpha : [0, \infty) \to [0, \infty)$ be a measurable function. Suppose that $x : [0, \infty) \to E$ is an absolutely continuous function satisfying

$$(3.1) \qquad \dot{x}(t) \in -\alpha(t)\partial\varphi_t(x(t)) \quad \text{a.e. } t \geq 0.$$

General existence results for nonautonomous subdifferential inclusions of this type can be found in [8, 33]. In the present setting we will take for granted the existence of such a global solution and we concentrate on its asymptotic convergence properties. We prove two convergence results for $x(t)$ depending on whether $\alpha(\cdot)$ is integrable.

The first result generalizes [12, Theorem 2.1].

THEOREM 3.1. *Let* $S \subset E$ *be nonempty and suppose that for* $\varepsilon > 0$ *small enough there exist nonempty sets* B_ε *and* S_ε *in* E *with* S_ε *closed and convex such that*

(a) $e(B_\varepsilon, S_\varepsilon) \to 0$ *and* $e(S_\varepsilon, S) \to 0$ *when* $\varepsilon \to 0$, *and*

(b) $\liminf_{t \to \infty} m_t(\varepsilon) > 0$, *where* $m_t(\varepsilon) = \inf\{\varphi_t(x) - \varphi_t(y) : x \notin B_\varepsilon, y \in S_\varepsilon\}$.

If $\int_{\bar{t}}^{\infty} \alpha(\tau)d\tau = \infty$ *for all* $\bar{t} \geq 0$, *then* $d(x(t), S) \to 0$ *when* $t \to \infty$.

Proof. Let $\varepsilon > 0$ be fixed and consider the function $\psi(t) = \frac{1}{2}d(x(t), S_\varepsilon)^2$. Since $x \mapsto \frac{1}{2}d(x, S_\varepsilon)^2$ is differentiable with gradient equal to $(x - x_\varepsilon)$, where x_ε denotes the projection of x onto S_ε, using (3.1) we obtain a.e. $t \geq 0$

$$(3.2) \qquad \dot{\psi}(t) = \langle -\dot{x}(t), x_\varepsilon(t) - x(t)\rangle \leq \alpha(t)[\varphi_t(x_\varepsilon(t)) - \varphi_t(x(t))].$$

Let $a_\varepsilon = \frac{1}{2}e(B_\varepsilon, S_\varepsilon)^2$ and use (b) to find $\sigma > 0$ and $\bar{t}$ such that $m_t(\varepsilon) \geq \sigma$ for $t \geq \bar{t}$. Then (3.2) implies that a.e. $t \geq \bar{t}$ we have

$$\psi(t) > a_\varepsilon \Rightarrow x(t) \notin B_\varepsilon \Rightarrow \dot{\psi}(t) \leq -\sigma\alpha(t).$$

Hence $\psi(t)$ is decreasing when above a_ε, and since $\int_{\bar t}^{\infty}\alpha(\tau)d\tau=\infty$ we get $\psi(t)\le a_\varepsilon$ for all t large. Thus $\|x(t)-x_\varepsilon(t)\|\le e(B_\varepsilon,S_\varepsilon)$ and then $d(x(t),S)\le e(B_\varepsilon,S_\varepsilon)+e(S_\varepsilon,S)$, so that the conclusion follows from (a). □

For the case when $\alpha(\cdot)$ is integrable we have the following.

THEOREM 3.2. *Suppose that $\varphi_t(x(t))$ is bounded from below and assume there exists a cluster point $\bar x$ of $x(t)$ and $x_\varepsilon\to\bar x$ as $\varepsilon\to 0$ with $\limsup_{t\to\infty}\varphi_t(x_\varepsilon)<\infty$. If $\int_{\bar t}^{\infty}\alpha(\tau)d\tau<\infty$ for some $\bar t$, then $x(t)\to\bar x$.*

Proof. Let $\psi(t)=\frac12\|x(t)-x_\varepsilon\|^2$ and take $M_\varepsilon\ge 0$ and $\hat t\ge\bar t$ such that a.e. $t\ge\hat t$ we have

$$\dot\psi(t)=\langle -\dot x(t),x_\varepsilon-x(t)\rangle\le\alpha(t)[\varphi_t(x_\varepsilon)-\varphi_t(x(t))]\le M_\varepsilon\alpha(t).$$

Then the function $\psi(t)+M_\varepsilon\int_t^\infty\alpha(\tau)d\tau$ is nonnegative and decreasing for $t\ge\hat t$, so it has a limit when $t\to\infty$. Hence $\|x(t)-x_\varepsilon\|$ converges and since $\bar x$ is a cluster point of $x(t)$ the limit is $\|\bar x-x_\varepsilon\|$. The conclusion follows by letting $\varepsilon\to 0$ in the inequality

$$\limsup_{t\to\infty}\|x(t)-\bar x\|\le\lim_{t\to\infty}\|x(t)-x_\varepsilon\|+\|x_\varepsilon-\bar x\|=2\|\bar x-x_\varepsilon\|.\qquad\square$$

3.2. Primal convergence of steepest-descent-penalty. Let us turn to the asymptotic analysis of the subdifferential inclusion (SDP). Concerning the existence of solutions, we may apply the results in [8, 33] to the family of functions $\varphi_t=f(\cdot,r(t))$ in order to find sufficient conditions on $r(\cdot)$ so that for every $u_0\in\mathrm{dom}(\varphi_0)$ there exists a unique $u:[0,\infty)\to\mathbb{R}^n$ absolutely continuous satisfying (SDP) with $u(0)=u_0$. These conditions differ according to the following cases: if $\kappa=\infty$, it suffices to have $r(\cdot)$ absolutely continuous; when $0<\kappa<\infty$ we must have in addition $\frac{dr}{dt}\in L^2_{\mathrm{loc}}(0,\infty)$; while for the case $\kappa=0$ it is enough to have $r(\cdot)$ continuous and nonincreasing. Furthermore, when $\kappa=0$ or $\kappa=\infty$, the domain of φ_t is a constant set and one may allow $r(\cdot)$ to have a finite number of jump discontinuities on each bounded interval.

Taking for granted the existence of a global solution $u(t)$ for (SDP), we will analyze its asymptotic properties with the help of the abstract results in section 3.1. To this end we require the following technical lemma.

LEMMA 3.3. *Assume* (H_0) *and let $S=S(\mathrm{P})$. Then for $\varepsilon\in(0,1)$ there exist nonempty and bounded closed convex sets B_ε and S_ε such that*

(a) *$e(B_\varepsilon,S_\varepsilon)\to 0$ and $e(S_\varepsilon,S)\to 0$ when $\varepsilon\to 0$, and*

(b) *$\liminf_{r\to 0}m_r(\varepsilon)>0$, where $m_r(\varepsilon)=\inf\{f(x,r)-f(y,r):x\notin B_\varepsilon,y\in S_\varepsilon\}$.*

Proof. Let $B_\varepsilon=\{x:f_0(x)\le v+\varepsilon;f_i(x)\le\varepsilon,i\in I\}$ and $S_\varepsilon=(1-\varepsilon)S+\varepsilon\bar x$ with $\bar x$ chosen so that $f_0(\bar x)\le v+1/2$ and $f_i(\bar x)\le 0$ for $i\in I$, with strict inequalities if $\kappa=0$. These sets are clearly convex, nonempty, and compact (since S is compact). Moreover, B_ε decreases to S when $\varepsilon\downarrow 0$ which readily implies (a).

To establish (b) we proceed by contradiction. If this were not the case, we could find sequences $r_k\to 0$, $x_k\notin B_\varepsilon$, $y_k\in S_\varepsilon$, and $\delta_k\to 0$ such that $f(x_k,r_k)\le\delta_k+f(y_k,r_k)$. Since $y_k=(1-\varepsilon)z_k+\varepsilon\bar x$ for some $z_k\in S$, we have $f_0(y_k)\le v+\varepsilon/2$ and $f_i(y_k)\le\varepsilon f_i(\bar x)$ for all $i\in I$, so that $f(y_k,r_k)\le v_k:=v+\varepsilon/2+r_k\sum_{i\in I}\theta[\varepsilon f_i(\bar x)/r_k]$ with $v_k\to v+\varepsilon/2$ (see Remark 1). The inequality $f(x_k,r_k)\le\delta_k+v_k$ gives

$$f_0(x_k)+r_k\sum_{i\in I}\theta(f_i(x_k)/r_k)\le\delta_k+v_k,\tag{3.3}$$

which implies that x_k is bounded. Indeed, otherwise passing to a subsequence we may assume $\|x_k\|\to\infty$ and $x_k/\|x_k\|\to d$ for some $d\ne 0$. Dividing (3.3) by $\|x_k\|$ and letting $k\to\infty$ we get $f_0^\infty(d)+\sum_{i\in I}\theta^\infty(f_i^\infty(d))\le 0$ so that Remark 1 implies

$f_i^\infty(d) \le 0$ for all $i = 0, \dots, m$, contradicting (H_0)(b). Since x_k is bounded, we may assume $x_k \to \hat{x}$ for some $\hat{x}$, and then (3.3) yields $f_0(\hat{x}) + \sum_{i\in I} \theta^\infty(f_i(\hat{x})) \le v + \varepsilon/2$. Again, Remark 1 implies $f_i(\hat{x}) \le 0$ for all $i \in I$ and then $f_0(\hat{x}) \le v + \varepsilon/2$, contradicting the fact that $x_k \notin B_\varepsilon$ for all k. This contradiction proves (b). □

Here is our main result on the coupling of steepest descent and penalty. It extends [12, Theorem 3.1] to nonlinear convex programs and general penalty functions.

THEOREM 3.4. *Let $r : [0,\infty) \to (0,\infty)$ be measurable with $r(t) \to 0$ as $t \to \infty$, and let $u : [0,\infty) \to \mathbb{R}^n$ be an absolutely continuous function satisfying*

$$\dot{u}(t) \in -\partial f(u(t), r(t)) \quad \text{a.e. } t \ge 0. \tag{SDP}$$

Under (H_0) we have $d(u(t), S(P)) \to 0$ as $t \to \infty$. If in addition (H_1)–(H_4) hold and the θ-means M_θ are convex, then $u(t) \to u_\infty$ for some $u_\infty \in S(P)$.

Proof. Assume (H_0). Taking B_ε and S_ε as in Lemma 3.3 we may apply Theorem 3.1 with $\alpha(t) \equiv 1$, $\varphi_t(x) = f(x, r(t))$, and $S = S(P)$ in order to get $d(u(t), S(P)) \to 0$. Since $S(P)$ is bounded, it follows that $u(t)$ is bounded with all its cluster points in $S(P)$. Let $C \subset S(P)$ be the closed convex hull of these cluster points. To establish the convergence of $u(t)$ we must show that C is a singleton or equivalently that $I_C = \phi$ (see Lemma 2.3(c)). To prove the latter we assume (H_1)–(H_4) and M_θ convex.

We proceed by contradiction. Suppose $I_C \ne \phi$ and assume without loss of generality that E_C is a linear subspace. Since $f_i \in \mathcal{Q}$, for $i \notin I_C$ (including $i = 0$) the space E_C is contained in the constancy space of f_i so that $\partial f_i(x) \subset E_C^\perp$. Letting $x(t)$ and $y(t)$ be the projections of $u(t)$ onto E_C and $E_C^\perp$, respectively, we have $y(t) \to 0$, and projecting (SDP) onto E_C we get

$$\dot{x}(t) \in -\sum_{i\in I_C} \theta'(f_i(x(t)+y(t))/r(t)) \partial_C f_i(x(t)+y(t)) \tag{3.4}$$

with $\partial_C f_i(u)$ the projection of $\partial f_i(u)$ on E_C. Choose $\sigma(t) > d(x(t), C)$ with $\sigma(t) \to 0$ as $t \to \infty$, and let $\alpha(t) = |I_C| \theta'(M_\theta(f_i(u(t))/r(t) : i \in I_C))$. Then (3.4) can be written in the form (3.1) by taking $\varphi_t : E_C \to \mathbb{R} \cup \{\infty\}$ as

$$\varphi_t(x) = \begin{cases} r(t) M_\theta(f_i(x+y(t))/r(t) : i \in I_C) & \text{if } d(x,C) \le \sigma(t), \\ \infty & \text{otherwise.} \end{cases}$$

To complete the proof we contradict $I_C \ne \phi$ by analyzing the cases $\alpha(\cdot)$ integrable or not with Theorems 3.2 and 3.1, respectively.

Case 1. $\int_{\bar{t}}^\infty \alpha(\tau) d\tau < \infty$ for some $\bar{t} \ge 0$.

We use Theorem 3.2. Since $u(t)$ is bounded, the same holds for $x(t)$, and (2.1) implies that $\varphi_t(x(t))$ is bounded from below. Let $\bar{x} \in C$ be a cluster point of $x(t)$ and set $x_\varepsilon = (1-\varepsilon)\bar{x} + \varepsilon\hat{x}$ with $\hat{x} \in C$ such that $\max_{i\in I_C} f_i(\hat{x}) < 0$ (see Lemma 2.3(a)). Then the same inequalities hold for x_ε, and Lemma 2.3(d) gives $\varphi_t(x_\varepsilon) \to \varphi_C(x_\varepsilon) < \infty$. Hence Theorem 3.2 applies and we get $x(t) \to \bar{x}$, so that $C = \{\bar{x}\}$, contradicting $I_C \ne \phi$.

Case 2. $\int_{\bar{t}}^\infty \alpha(\tau) d\tau = \infty$ for all $\bar{t} \ge 0$.

We use Theorem 3.1 to show that $d(x(t), S_C) \to 0$ which implies $C \subset S_C$ and contradicts Lemma 2.3(e). To apply the theorem take $B_\varepsilon = \{x \in E_C : d(x, S_C) \le \varepsilon\}$ and let $\omega_\varepsilon = \liminf_{t\to\infty} \inf_{x\notin B_\varepsilon} \varphi_t(x)$. We claim that $\omega_\varepsilon > v_C$. Indeed, if this were not the case, we could find $t_k \to \infty$ and $x_k \notin B_\varepsilon$ with $\lim \varphi_{t_k}(x_k) \le v_C$. In particular for k large we have $\varphi_{t_k}(x_k) < \infty$ so that $d(x_k, C) \le \sigma(t_k) \to 0$, and then x_k is bounded so we may assume that it converges toward some $\bar{x} \in C$. Using Lemma 2.3(d) it

follows that $\varphi_C(\bar{x}) \le \liminf \varphi_{t_k}(x_k) \le v_C$ so that $\bar{x} \in S_C$ contradicting the fact that $d(x_k, S_C) > \varepsilon$ for all k. This proves our claim.

Choose $\hat{x} \in C$ as in Lemma 2.3(a) and set $S_\varepsilon = (1-\alpha_\varepsilon)S_C + \alpha_\varepsilon \hat{x}$ with $\alpha_\varepsilon > 0$ small enough so that $(1-\alpha_\varepsilon)v_C + \alpha_\varepsilon \varphi_C(\hat{x}) < w_\varepsilon$ and $\alpha_\varepsilon \downarrow 0$ when $\varepsilon \to 0$. Then condition (a) of Theorem 3.1 holds trivially, while (b) follows since $\limsup_{t\to\infty} \sup_{y\in S_\varepsilon} \varphi_t(y) < w_\varepsilon$. To prove the latter take $t_k \to \infty$ a sequence attaining this upper limit and $y_k \in S_\varepsilon$ a point where φ_{t_k} is maximal. Passing to a subsequence, we may assume that y_k converges to a certain $\bar{y} \in S_\varepsilon$. Then $f_i(\bar{y}) < 0$ for all $i \in I_C$, and Lemma 2.3(d) gives $\varphi_{t_k}(y_k) \to \varphi_C(\bar{y})$. The definition of S_ε and the convexity of φ_C imply $\varphi_C(\bar{y}) \le (1-\alpha_\varepsilon)v_C + \alpha_\varepsilon \varphi_C(\hat{x}) < w_\varepsilon$ which completes the proof. □

3.3. Dual convergence of steepest-descent-penalty. In this subsection we restrict our attention to the case of linear programming, that is to say, $f_0(x) = c^{\mathrm{T}}x$ and $f_i(x) = a_i^{\mathrm{T}}x - b_i$ for all $i \in I$. Let $u(t)$ be any trajectory satisfying equation (SDP) and consider the associated dual trajectory $\mu(t)$ defined by

$$\mu_i(t) = \theta'([a_i^{\mathrm{T}}u(t) - b_i]/r(t)) \quad \forall\, i \in I.$$

We extend [12, Theorem 3.2] proving that under suitable assumptions $\mu(t)$ converges to the solution λ^θ described in Theorem 2.5.

THEOREM 3.5. *Assume* (H_0) *and let $r(t)$ be smooth and decreasing with $r(t) \to 0$ and $\dot{r}(t)/r(t)$ bounded. Then $\mu(t) \to \lambda^\theta$ and $\dot{u}(t) \to 0$ when $t \to \infty$.*

Proof. Since $\dot{u}(t) = \sum_{i\in I}[\lambda_i^\theta - \mu_i(t)]a_i$ it suffices to prove that $\mu(t) \to \lambda^\theta$. The latter amounts to $\theta'(z_i(t)) \to \lambda_i^\theta$ for all $i \in I$, where $z_i(t) = [a_i^{\mathrm{T}}u(t) - b_i]/r(t)$. Hence, letting $h_i(z) = \theta(z) - \lambda_i^\theta z$, the result boils down to $h_i(z_i(t)) \to \inf_z h_i(z) = -\theta^*(\lambda_i^\theta)$ for $i \in I$.

Since θ may be unbounded from below, we may have $\theta^*(0) = \infty$ so that special care must be taken for dealing with the λ_i^θ's which are zero. Let I_0 be as in Theorem 2.5 and recall that $\lambda_i^\theta > 0$ if and only if $i \in I_0$. Define $\gamma(t) = \sum_{i\notin I_0} \mu_i(t)$ and $\psi(t) = \sum_{i\in I_0} h_i(z_i(t))$. To establish the result we must then prove that $\gamma(t) \to 0$ and $\psi(t) \to \alpha$ with $\alpha = -\sum_{i\in I_0} \theta^*(\lambda_i^\theta)$ (notice that α is finite).

The idea of the proof is the following. For $s = 1, \dots, |I_0^c|$ we let

$$\theta_s(t) = \min_S \left\{ \sum_{i\in S} \theta(z_i(t)) : S \subset I_0^c, |S| = s \right\}, \tag{3.5}$$

$$\gamma_s(t) = \min_S \left\{ \sum_{i\in S} \mu_i(t) \ : S \subset I_0^c, |S| = s \right\}, \tag{3.6}$$

while for $s = 0$ we set $\gamma_0(t) \equiv \theta_0(t) \equiv 0$. Define also

$$\psi_s(t) = \sum_{i\in I} h_i(z_i(t)) - \theta_s(t). \tag{3.7}$$

We will show inductively that $\gamma_s(t) \to 0$ and $\psi_s(t) \to \alpha_s := \alpha - (|I_0^c| - s)\,\theta^*(0)$, from which the result follows since for $s = |I_0^c|$ we have $\gamma_s(t) \equiv \gamma(t)$ and $\psi_s(t) \equiv \psi(t)$. Since the proof is somewhat involved, we split it into a series of lemmas. □

Hereafter we assume (H_0) and $r(t) \to 0$. We remark that $a_i\bar{x} = b_i$ for all $i \in I_0$ and $\bar{x} \in S(\mathrm{P})$ (this follows by complementary slackness since $\lambda_i^\theta > 0$ for $i \in I_0$) and we denote $\bar{u}(t)$ the projection of $u(t)$ onto $S(\mathrm{P})$, $\delta = \mathrm{diam}(S(\mathrm{P}))$, and $\sigma = \max_{i\notin I_0} \|a_i\|$. Notice that since ($\mathrm{H}_0$) holds we have $\|\bar{u}(t) - u(t)\| = d(u(t), S(\mathrm{P})) \to 0$.

Our first lemma shows that when $\|\dot{u}(t)\| \sim 0$ we have $\mu_i(t) \sim 0$ for all $i \notin I_0$.

LEMMA 3.6. *There exist positive constants $\beta > 0$ and $M \geq 0$ such that $\gamma(t) \leq M[\|\dot{u}(t)\| + \theta'(-\beta/r(t))]$ for all t large.*

Proof. If I_0^c is empty, take $\beta = 1$ and $M = 0$. Otherwise let $\beta = \frac{1}{3}\min_{i \notin I_0}(b_i - a_i^{\mathrm{T}}\bar{x})$ with $\bar{x} \in \mathrm{ri}(S(\mathrm{P}))$, so that $\beta > 0$ and then $a_i^{\mathrm{T}}[u(t) - \bar{u}(t)] \leq \beta$ for t large enough. Decompose $I_0^c = I(t) \cup J(t)$, where $I(t)$ contains the i's such that $a_i^{\mathrm{T}}\bar{u}(t) \geq b_i - 2\beta$ and $J(t)$ contains the rest. For $i \in J(t)$ we have $a_i^{\mathrm{T}}u(t) \leq a_i^{\mathrm{T}}\bar{u}(t) + \beta < b_i - \beta$ so that $\mu_i(t) \leq \theta'(-\beta/r(t))$. To obtain a bound on $\mu_i(t)$ for $i \in I(t)$ we observe that $\bar{u}(t)$ and $\bar{x}$ belong to S(P), so that $a_i^{\mathrm{T}}\bar{u}(t) = a_i^{\mathrm{T}}\bar{x}$ for all $i \in I_0$ and then

$$\langle -\dot{u}(t), \bar{u}(t) - \bar{x}\rangle = \sum_{i \notin I_0} \mu_i(t) a_i^{\mathrm{T}}[\bar{u}(t) - \bar{x}].$$

Noting that $a_i^{\mathrm{T}}[\bar{u}(t) - \bar{x}] \geq \beta$ for all $i \in I(t)$, we get

$$\beta \sum_{i \in I(t)} \mu_i(t) \leq \delta \left[\|\dot{u}(t)\| + \sigma \sum_{i \in J(t)} \mu_i(t) \right],$$

and then the result holds with $M = \delta/\beta + |I_0^c|(1 + \sigma\delta/\beta)$. ∎

In the next two lemmas we establish that when $\|\dot{u}(t)\| \sim 0$ we also have $\psi(t) \sim \alpha$. To this end we require some additional notation. Let $E_0 = \mathrm{span}\{a_i : i \in I_0\}$ and for $\varepsilon > 0$ define $g(\varepsilon) = \sup_{w \in E_0}\{\Phi(w) : d(0, \partial\Phi(w)) \leq \varepsilon\}$ with $\Phi : E_0 \to \mathbb{R} \cup \{\infty\}$ given by

$$\Phi(w) = \sum_{i \in I_0} \left[\theta(a_i^{\mathrm{T}}w) - \lambda_i^{\theta} a_i^{\mathrm{T}}w\right].$$

LEMMA 3.7. *Φ is coercive with $\min_{w \in E_0}\Phi(w) = \alpha$ and $g(\varepsilon) \to \alpha$ when $\varepsilon \to 0$.*

Proof. Since $\lambda_i^{\theta} > 0$ for all $i \in I_0$, using Remark 1 it follows easily that Φ is coercive. Also, by definition of the Fenchel conjugate θ^* it is clear that $\Phi(w) \geq \alpha$ for all $w \in E_0$. Take $\bar{x} \in S(\mathrm{P})$ and let $w(r)$ be the projection of $[x(r) - \bar{x}]/r$ onto E_0. For $i \in I_0$ we have $a_i^{\mathrm{T}}\bar{x} = b_i$ so that $\theta'(a_i^{\mathrm{T}}w(r)) = \lambda_i(r)$ and, since Theorem 2.5 implies $\lambda_i(r) \to \lambda_i^{\theta} > 0$, it follows that $a_i^{\mathrm{T}}w(r)$ remains bounded. Hence $w(r)$ is also bounded and then it has a cluster point $\bar{w} \in E_0$ which satisfies $\theta'(a_i^{\mathrm{T}}\bar{w}) = \lambda_i^{\theta}$. The definition of θ^* then gives $\Phi(\bar{w}) = -\sum_{i \in I_0}\theta^*(\lambda_i^{\theta})$ proving $\min_{w \in E_0}\Phi(w) = \alpha$ as claimed. The property $g(\varepsilon) \to \alpha$ is a well-known consequence of coercivity. ∎

LEMMA 3.8. $\alpha \leq \psi(t) \leq g(\|\dot{u}(t)\| + \sigma\gamma(t))$.

Proof. Since $a_i\bar{u}(t) = b_i$ for all $i \in I_0$, taking $w(t)$ as the projection of $[u(t) - \bar{u}(t)]/r(t)$ onto E_0 we get $\psi(t) = \Phi(w(t)) \geq \alpha$. Now, (SDP) and the definition of Φ imply

$$-\dot{u}(t) - \sum_{i \notin I_0} \mu_i(t) a_i = \sum_{i \in I_0} [\mu_i(t) - \lambda_i^{\theta}] a_i \in \partial\Phi(w(t))$$

so that the definition of g yields $\psi(t) = \Phi(w(t)) \leq g(\|\dot{u}(t)\| + \sigma\gamma(t))$. ∎

The previous results lead to the following bound for $\psi_s(t)$.

LEMMA 3.9. *There exists a nondecreasing function $\varepsilon \mapsto H_s(\varepsilon)$ with $H_s(\varepsilon) \to \alpha_s$ when $\varepsilon \downarrow 0$, such that for all t large enough*

$$\alpha_s \leq \psi_s(t) \leq H_s(\|\dot{u}(t)\| + \theta'(-\beta/r(t))). \tag{3.8}$$

Proof. From definition of θ^* it is clear that $\psi_s(t) \geq \alpha_s$. To establish the upper bound we define $b(s) = \sup\{\theta(u) : \theta'(u) \leq s\}$ so that $\theta(z_i(t)) \leq b(\mu_i(t)) \leq b(\gamma(t))$ for all $i \notin I_0$, which combined with Lemma 3.8 gives

$$\psi_s(t) \leq g(\|\dot{u}(t)\| + \sigma\gamma(t)) + (|I_0^c| - s)\, b(\gamma(t)).$$

Thus, using Lemma 3.6, it follows that (3.8) holds for t large enough with $H_s(\varepsilon) = g((1+\sigma M)\varepsilon) + (|I_0^c| - s)\, b(M\varepsilon)$. Clearly $H_s(\cdot)$ is nondecreasing, while $H_s(\varepsilon) \to \alpha_s$ follows from Lemma 3.7 and the obvious fact $b(\varepsilon) \to \inf_u \theta(u) = -\theta^*(0)$. ∎

Our final lemma provides the key for completing the proof of Theorem 3.5.

LEMMA 3.10. *With the hypothesis of Theorem* 3.5, *if* $\gamma_s(t) \to 0$, *then* $\psi_s(t) \to \alpha_s$.

Proof. Using (3.7) we may compute the right derivative of $\psi_s(t)$ as

$$\frac{d\psi_s}{dt^+} = \frac{d}{dt}\left[\sum_{i\in I} h_i(z_i(t))\right] - \frac{d\theta_s}{dt^+},$$

and since $\theta_s(\cdot)$ is a min-function, denoting $\mathcal{S}(t)$ the set of S's which attain the minimum in (3.5), we have $\frac{d\theta_s}{dt^+} = \min\{\frac{d}{dt}\left[\sum_{i\in S}\theta(z_i(t))\right] : S \in \mathcal{S}(t)\}$. Take $S_t \in \mathcal{S}(t)$ attaining the latter minimum. Since $h_i(z_i(t)) = \theta(z_i(t))$ for all $i \in S_t$ we get

$$\frac{d\psi_s}{dt^+} = \sum_{i\notin S_t} \frac{d}{dt}\left[h_i(z_i(t))\right] = \sum_{i\notin S_t} (\mu_i(t) - \lambda_i^\theta)\dot{z}_i(t).$$

Now, setting $\xi(t) = \dot{r}(t)[u(t) - \bar{u}(t)]/r(t)$, a direct computation gives

$$\dot{z}_i(t) = a_i^{\mathrm{T}}[\dot{u}(t) - \xi(t)]/r(t) - \dot{r}(t)[a_i^{\mathrm{T}}\bar{u}(t) - b_i]/r(t)^2,$$

and since $\dot{r}(t) \leq 0$, $[a_i^{\mathrm{T}}\bar{u}(t) - b_i] \leq 0$ and $\lambda_i^\theta[a_i^{\mathrm{T}}\bar{u}(t) - b_i] = 0$, we deduce

$$\frac{d\psi_s}{dt^+} \leq \sum_{i\notin S_t} (\mu_i(t) - \lambda_i^\theta) a_i^{\mathrm{T}}[\dot{u}(t) - \xi(t)]/r(t).$$

Letting $\eta(t) = \sum_{i\in S_t} \mu_i(t) a_i$ we have $\sum_{i\notin S_t}(\mu_i(t) - \lambda_i^\theta)a_i = -[\dot{u}(t) + \eta(t)]$ so that

$$\frac{d\psi_s}{dt^+} \leq -\frac{1}{r(t)}\langle \dot{u}(t) + \eta(t), \dot{u}(t) - \xi(t)\rangle. \tag{3.9}$$

From this inequality it follows easily that $\frac{d\psi_s}{dt^+} \geq 0$ implies $\|\dot{u}(t) - \xi(t)\| \leq \|\xi(t) + \eta(t)\|$ so that, letting $\varepsilon(t) = \|\xi(t)\| + \|\xi(t) + \eta(t)\|$ and $h_s(t) = H_s(\varepsilon(t) + \theta'(-\beta/r(t)))$, we may use Lemma 3.9 to deduce that for t large we have

$$\frac{d\psi_s}{dt^+} \geq 0 \;\Rightarrow\; \|\dot{u}(t)\| \leq \varepsilon(t) \;\Rightarrow\; \alpha_s \leq \psi_s(t) \leq h_s(t). \tag{3.10}$$

We claim that $\varepsilon(t) \to 0$ so that $h_s(t) \to \alpha_s$. Indeed, since $\dot{r}(t)/r(t)$ is bounded and $\|u(t) - \bar{u}(t)\| = d(u(t), S(P)) \to 0$, we get $\xi(t) \to 0$. To see that $\eta(t) \to 0$ we observe that the sets S attaining the minima in (3.6) and (3.5) coincide, and since $S_t \in \mathcal{S}(t)$ we deduce $\|\eta(t)\| \leq \sigma\sum_{i\in S_t}\mu_i(t) = \sigma\gamma_s(t)$ which tends to 0 by assumption.

Using (3.10) it follows that $\psi_s(\cdot)$ is decreasing whenever it is above $h_s(\cdot)$. Since $h_s(t) \to \alpha_s$ we deduce that $\psi_s(t)$ has a limit $\bar{\alpha}_s \geq \alpha_s$ when $t \to \infty$. If $\bar{\alpha}_s > \alpha_s$, then (3.8) implies $\|\dot{u}(t)\| \geq \varepsilon$ for some $\varepsilon > 0$ and all t large, and then by (3.9) the

right derivative of $\psi_s(t)$ tends to $-\infty$ so that $\psi_s(t) \to -\infty$. This contradiction proves $\bar{\alpha}_s = \alpha_s$ and therefore $\psi_s(t) \to \alpha_s$. ∎

We may now conclude the proof of Theorem 3.5. We distinguish two cases.

Proof.

Case 1. $\theta^*(0) < \infty$. Using Lemma 3.10 with $s = 0$ we get $\psi_0(t) \to \alpha_0$. This implies $\psi(t) \to \alpha$, as well as $\theta(z_i(t)) \to -\theta^*(0)$ for all $i \notin I_0$ from which it follows that $\mu_i(t) = \theta'(z_i(t)) \to 0$ and then $\gamma(t) \to 0$.

Case 2. $\theta^*(0) = \infty$. We prove by induction that $\gamma_s(t) \to 0$ for $s = 0, 1, \ldots, |I_0^c|$. For $s = 0$ this holds by definition. Suppose $\gamma_s(t) \to 0$ for some $s < |I_0^c|$ so that Lemma 3.10 implies $\psi_s(t) \to \alpha_s = -\infty$. Since $\psi_s(t) = \psi(t) + \sum_{i \in I_0^c \setminus S_t} \theta(z_i(t))$ with $\psi(t) \geq \alpha$ we may find $i(t) \in I_0^c \setminus S_t$ with $\theta(z_{i(t)}(t)) \to -\infty$, so that $\mu_{i(t)}(t) \to 0$ and then $\gamma_{s+1}(t) \leq \gamma_s(t) + \mu_{i(t)}(t) \to 0$. This achieves the induction step. Taking $s = |I_0^c|$ we deduce $\gamma(t) \equiv \gamma_s(t) \to 0$ and $\psi(t) \equiv \psi_s(t) \to \alpha$ as required. ∎

4. The prox-penalty iteration. In this section we analyze the convergence of the inexact prox-penalty iteration (IPP), proving the analogues of the results obtained for (SDP). Roughly speaking, the analysis is a discretization of the continuous arguments. We begin in section 4.1 with two convergence results for an abstract inexact diagonal prox iteration. The convergence of the primal prox-penalty iterates is studied in section 4.2, while in section 4.3 we study the corresponding dual iterates.

4.1. Convergence of diagonal prox processes. Let $\varphi_k : E \to \mathbb{R} \cup \{\infty\}$ be a sequence of convex functions defined on a finite dimensional Euclidean space E, and let $\alpha_k \geq 0$. Let $x_k \in E$ be any sequence satisfying

$$\frac{x_k - x_{k-1}}{\lambda_k} \in -\partial_{\varepsilon_k}[\alpha_k \varphi_k](x_k) + \nu_k, \tag{4.1}$$

where $\lambda_k > 0$ is a stepsize, $\varepsilon_k \geq 0$ is a tolerance in the computation of approximate subgradients, and $\nu_k \in E$ represents the residual in the inexact resolution of the prox iteration.

For the case $\sum \alpha_k \lambda_k = \infty$ we have the following.

THEOREM 4.1. *Assume that* $\sum \alpha_k \lambda_k = \infty$ *and* $\sum [\varepsilon_k + \|\nu_k\|]\lambda_k < \infty$. *Let* $S \subset E$ *be nonempty and suppose that for* $\varepsilon > 0$ *small enough there exist nonempty sets* B_ε *and* S_ε *in* E, *with* S_ε *closed convex and bounded, such that*

(a) $e(B_\varepsilon, S_\varepsilon) \to 0$ *and* $e(S_\varepsilon, S) \to 0$ *when* $\varepsilon \to 0$, *and*

(b) $\liminf_{k\to\infty} m_k(\varepsilon) > 0$, *where* $m_k(\varepsilon) = \inf\{\varphi_k(x) - \varphi_k(y) : x \notin B_\varepsilon, y \in S_\varepsilon\}$.

Then $d(x_k, S) \to 0$ *when* $k \to \infty$.

Proof. Fix $\varepsilon > 0$ and set $\psi_k = \frac{1}{2}\|x_k - y_k\|^2$ with y_k the projection of x_k onto S_ε. Denoting $\delta_\varepsilon = \operatorname{diam}(S_\varepsilon)$ and $g_k = \nu_k - [x_k - x_{k-1}]/\lambda_k \in \partial_{\varepsilon_k}[\alpha_k \varphi_k](x_k)$ we get

$$\begin{aligned}
\psi_k - \psi_{k-1} &\leq \tfrac{1}{2}\|x_k - y_{k-1}\|^2 - \tfrac{1}{2}\|x_{k-1} - y_{k-1}\|^2 \\
&= \langle x_k - x_{k-1}, \tfrac{x_k + x_{k-1}}{2} - y_{k-1} \rangle \\
&\leq \langle x_k - x_{k-1}, x_k - y_{k-1} \rangle \\
&= \lambda_k \langle g_k, y_{k-1} - x_k \rangle - \lambda_k \langle \nu_k, y_{k-1} - x_k \rangle \\
&\leq \lambda_k \alpha_k [\varphi_k(y_{k-1}) - \varphi_k(x_k)] + \lambda_k \varepsilon_k + \lambda_k \|\nu_k\| [\|x_k - y_k\| + \delta_\varepsilon].
\end{aligned}$$

The inequality $2z \leq 1 + z^2$ implies $\|x_k - y_k\| \leq 1/2 + \psi_k$, so that letting $\rho_k = 1 - \lambda_k \|\nu_k\|$ and $\eta_k = \lambda_k[(\delta_\varepsilon + \frac{1}{2})\|\nu_k\| + \varepsilon_k]$ we deduce

$$\rho_k \psi_k \leq \psi_{k-1} + \eta_k + \lambda_k \alpha_k [\varphi_k(y_{k-1}) - \varphi_k(x_k)].$$

Choose $\bar{k}$ so that $\rho_k > 0$ for $k \geq \bar{k}$. Defining $\rho'_k = [\prod_{\bar{k}}^{k} \rho_i]$ and $\psi'_k = \rho'_k[\psi_k + \sum_{k+1}^{\infty} \eta_i]$, using the fact that $\rho_k \leq 1$ we obtain that for all $k > \bar{k}$ we have

$$\psi'_k \leq \psi'_{k-1} + \lambda_k \alpha_k \rho'_{k-1}[\varphi_k(y_{k-1}) - \varphi_k(x_k)]. \tag{4.2}$$

The assumption $\sum[\varepsilon_k + \|\nu_k\|]\lambda_k < \infty$ implies $\sum_k^{\infty} \eta_i \downarrow 0$ and $\rho'_k \downarrow \bar{\rho}$ for some $\bar{\rho} \in (0,1)$. Then we may find $\hat{k} > \bar{k}$ with $\rho'_k < \bar{\rho}(1+\varepsilon)$ and $\sum_k^{\infty} \eta_i < \varepsilon$ for all $k \geq \hat{k}$, and using assumption (b) we may also suppose that $m_k(\varepsilon) \geq \sigma$ for some $\sigma > 0$ and all $k \geq \hat{k}$.

Let $a_\varepsilon = \bar{\rho}(1+\varepsilon)[\frac{1}{2}e(B_\varepsilon, S_\varepsilon)^2 + \varepsilon]$. If $\psi'_k > a_\varepsilon$, then $\psi_k > \frac{1}{2}e(B_\varepsilon, S_\varepsilon)^2$ so that $x_k \notin B_\varepsilon$, and therefore $[\varphi_k(x_k) - \varphi_k(y_{k-1})] \geq m_k(\varepsilon) \geq \sigma$. Hence (4.2) yields

$$\psi'_k > a_\varepsilon \quad \Rightarrow \quad \psi'_k \leq \psi'_{k-1} - \sigma \alpha_k \lambda_k \rho'_{k-1}.$$

Thus ψ'_k decreases when above a_ε, and since $\sum \rho'_{k-1}\alpha_k\lambda_k = \infty$ it follows that $\psi'_k \leq a_\varepsilon$ for all k large. Since $\bar{\rho}\psi_k \leq \psi'_k$ the latter implies $\|x_k - y_k\| \leq \sqrt{2a_\varepsilon/\bar{\rho}}$ and then $d(x_k, S) \leq \sqrt{2a_\varepsilon/\bar{\rho}} + e(S_\varepsilon, S)$, so that the conclusion follows from (a). ∎

When $\sum \alpha_k \lambda_k < \infty$ we have the following.

THEOREM 4.2. *Assume $\sum \alpha_k\lambda_k < \infty$ and $\sum[\varepsilon_k + \|\nu_k\|]\lambda_k < \infty$. Suppose also that x_k is bounded with $\varphi_k(x_k)$ bounded from below, and let $\bar{x}$ be a cluster point of x_k for which there exists $x_\varepsilon \to \bar{x}$ as $\varepsilon \to 0$ with $\limsup_{k\to\infty} \varphi_k(x_\varepsilon) < \infty$. Then $x_k \to \bar{x}$.*

Proof. Let $\psi_k = \frac{1}{2}\|x_k - x_\varepsilon\|^2$. Proceeding as in the previous proof we get

$$\psi_k - \psi_{k-1} \leq \lambda_k\alpha_k[\varphi_k(x_\varepsilon) - \varphi_k(x_k)] + \lambda_k\varepsilon_k + \lambda_k\langle \nu_k, x_k - x_\varepsilon\rangle.$$

Since x_k is bounded, $\varphi_k(x_k)$ is bounded from below, and $\limsup_k \varphi_k(x_\varepsilon) < \infty$, we may find some constant $M \geq 0$ such that $\psi_k - \psi_{k-1} \leq \lambda_k\tilde{\varepsilon}_k$ with $\tilde{\varepsilon}_k = \varepsilon_k + M[\alpha_k + \|\nu_k\|]$. Then the sequence $\psi_k + \sum_{k+1}^{\infty} \tilde{\varepsilon}_k\lambda_k$ is nonnegative and decreasing, so it has a limit when $k \to \infty$. Hence $\|x_k - x_\varepsilon\|$ converges and since $\bar{x}$ is a cluster point of x_k the limit is $\|\bar{x} - x_\varepsilon\|$. The conclusion follows by letting $\varepsilon \to 0$ in the inequality

$$\limsup_{k\to\infty} \|x_k - \bar{x}\| \leq \lim_{k\to\infty} \|x_k - x_\varepsilon\| + \|x_\varepsilon - \bar{x}\| = 2\|\bar{x} - x_\varepsilon\|. \qquad ∎$$

4.2. Primal convergence of the prox-penalty iteration. Let $r_k > 0$ be a sequence of penalty parameters converging to 0 and consider any sequence $u_k \in \mathbb{R}^n$ satisfying the inexact prox-penalty iteration

$$\frac{u_k - u_{k-1}}{\lambda_k} \in -\partial_{\varepsilon_k} f(u_k, r_k) + \nu_k, \tag{IPP}$$

where as before $f(x, r_k) = f_0(x) + r_k \sum \theta(f_i(x)/r_k)$ denotes the penalty function for the given sequence of penalty parameters, $\lambda_k > 0$ is a stepsize, $\varepsilon_k \geq 0$ is a tolerance in the computation of approximate subgradients, and $\nu_k \in \mathbb{R}^n$ is the residual in the inexact resolution of the prox iteration. In what follows we denote $s_k = -(u_k - u_{k-1})/\lambda_k$ and $g_k = s_k + \nu_k \in \partial_{\varepsilon_k} f(u_k, r_k)$.

We extend [5, Theorem 1.1] to nonlinear programs and general penalty schemes.

THEOREM 4.3. *Suppose* (H_0), $r_k \to 0$, $\sum \lambda_k = \infty$, *and* $\sum[\varepsilon_k + \|\nu_k\|]\lambda_k < \infty$. *Then we have $d(u_k, S(\mathrm{P})) \to 0$. If in addition* (H_1)–(H_4) *hold and the θ-means M_θ are convex, then $u_k \to u_\infty$ for some $u_\infty \in S(\mathrm{P})$.*

Proof. Under (H_0), $\sum \lambda_k = \infty$ and $\sum[\varepsilon_k + \|\nu_k\|]\lambda_k < \infty$, using Lemma 3.3 and Theorem 4.1 with $\alpha_k \equiv 1$, $\varphi_k(x) = f(x, r_k)$, and $S = S(\mathrm{P})$ we get $d(u_k, S(\mathrm{P})) \to 0$. Suppose in addition (H_1)–(H_4) and M_θ convex. To establish the convergence of u_k

we prove that $I_C = \phi$ with C the convex hull of the cluster points of u_k. We proceed by contradiction: suppose $I_C \neq \phi$ and take $S = S_C$. Assume also that E_C is a linear space and decompose $u_k = x_k + y_k$ with $x_k \in E_C$ and $y_k \in E_C^{\perp}$, so that $y_k \to 0$. We show that x_k satisfies an inclusion of the form (4.1). Indeed, for all $w \in \mathbb{R}^n$ we have

$$f(u_k, r_k) + \langle s_k + \nu_k, w - u_k \rangle \leq f(w, r_k) + \varepsilon_k. \tag{4.3}$$

Taking $w = x + y_k$ with $x \in E_C$ and since $f_i(x_k + y_k) = f_i(x + y_k)$ for all $i \notin I_C$, we have

$$q_k(x_k) + \langle s_k + \nu_k, x - x_k \rangle \leq q_k(x) + \varepsilon_k,$$

where $q_k(x) = r_k \sum_{i \in I_C} \theta(f_i(x + y_k)/r_k)$, so that letting $\tilde{s}_k = -(x_k - x_{k-1})/\lambda_k$ and $\tilde{\nu}_k$ the projection of ν_k onto E_C we get $\tilde{s}_k + \tilde{\nu}_k \in \partial_{\varepsilon_k} q_k(x_k)$. Now q_k can be expressed as the composition of $h_k(t) = |I_C| r_k \theta(t/r_k)$ and $A_k(x) = r_k M_\theta(f_i(x + y_k)/r_k : i \in I_C)$, so that [28, Theorem 3.6.1] implies the existence of $\alpha_k \in \partial_{\varepsilon_k} h_k(A_k(x_k))$ such that

$$\tilde{s}_k + \tilde{\nu}_k \in \partial_{\varepsilon_k} [\alpha_k A_k](x_k). \tag{4.4}$$

Take $\sigma_k > d(x_k, C)$ with $\sigma_k \to 0$ and define $\varphi_k : E_C \to \mathbb{R} \cup \{\infty\}$ as $\varphi_k(x) = A_k(x)$ if $d(x, C) \leq \sigma_k$ and $\varphi_k(x) = \infty$ otherwise. Since $h_k(\cdot)$ is increasing we have $\alpha_k \geq 0$, and (4.4) implies that x_k satisfies

$$\frac{x_k - x_{k-1}}{\lambda_k} \in -\partial_{\varepsilon_k} [\alpha_k \varphi_k](x_k) + \tilde{\nu}_k.$$

To complete the proof we contradict $I_C \neq \phi$ by analyzing the cases $\sum \alpha_k \lambda_k$ finite and infinite with Theorems 4.2 and 4.1, respectively.

Case 1. $\sum \alpha_k \lambda_k < \infty$.

We use Theorem 4.2. By assumption we have $\sum \alpha_k \lambda_k < \infty$ and $\sum \varepsilon_k \lambda_k < \infty$, as well as $\sum \|\tilde{\nu}_k\| \lambda_k \leq \sum \|\nu_k\| \lambda_k < \infty$. Since u_k is bounded the same holds for x_k, and (2.1) implies that $\varphi_k(x_k)$ is bounded from below. Let $\bar{x} \in C$ be a cluster point of x_k, and set $x_\varepsilon = (1 - \varepsilon)\bar{x} + \varepsilon \hat{x}$ with $\hat{x} \in C$ such that $\max_{i \in I_C} f_i(\hat{x}) < 0$ (see Lemma 2.3(a)). Then the same inequality holds for x_ε, and Lemma 2.3(d) implies $\varphi_k(x_\varepsilon) \to \varphi_C(x_\varepsilon) < \infty$. Hence Theorem 4.2 applies, and we get $x_k \to \bar{x}$, so that $C = \{\bar{x}\}$, contradicting $I_C \neq \phi$.

Case 2. $\sum \alpha_k \lambda_k = \infty$.

We use Theorem 4.1. By assumption we have $\sum \alpha_k \lambda_k = \infty$ and $\sum \varepsilon_k \lambda_k < \infty$ as well as $\sum \|\tilde{\nu}_k\| \lambda_k \leq \sum \|\nu_k\| \lambda_k < \infty$. Setting $B_\varepsilon = \{x \in E_C : d(x, S_C) \leq \varepsilon\}$ and $S_\varepsilon = (1 - \alpha_\varepsilon) S_C + \alpha_\varepsilon \hat{x}$ as in Case 2 of Theorem 3.4 one can prove that properties (a) and (b) of Theorem 4.1 are satisfied, and then $d(x_k, S_C) \to 0$. This gives $C \subset S_C$ which yields a contradiction with Lemma 2.3(e). □

In order to establish the convergence of the dual variables (in the next section) we do not require u_k to be convergent but only $d(u_k, S(\mathrm{P})) \to 0$. Hence it is worth noting that this also holds when the assumption $\sum \|\nu_k\| \lambda_k < \infty$ is replaced by the (usually weaker) condition $\nu_k \to 0$.

PROPOSITION 4.4. *Assume* (H_0), $r_k \to 0$, $\sum \lambda_k = \infty$, $\sum \varepsilon_k \lambda_k < \infty$, *and* $\nu_k \to 0$. *Then we have* $d(u_k, S(\mathrm{P})) \to 0$.

Proof. This follows from Theorem 4.1 applied to $\varphi_k(x) = f(x, r_k) - \langle \nu_k, x \rangle$. Notice that $\partial_{\varepsilon_k} \varphi_k(u_k) = \partial_{\varepsilon_k} f(u_k, r_k) - \nu_k$, and a slight modification of Lemma 3.3 shows that when $\nu_k \to 0$ conditions (a) and (b) in Theorem 4.1 hold for this φ_k with the same sets B_ε and S_ε, so that the theorem applies. □

4.3. Dual convergence of the prox-penalty iteration. Consider the case of linear programming ($f_0(x)=c^{\mathrm{T}}x$, $f_i(x)=a_i^{\mathrm{T}}x-b_i$ for $i\in I$) and the dual iterates μ^k defined by

$$\mu_i^k=\theta'([a_i^{\mathrm{T}}u_k-b_i]/r_k).$$

We will prove two results ensuring that $\mu^k\to\lambda^\theta$, where λ^θ is the dual optimal solution described in Theorem 2.5. The first one concerns the case where $\varepsilon_k\equiv 0$, while the second allows $\varepsilon_k>0$ but requires the penalty function θ to be bounded from below.

As in the proof of Theorem 3.5 the convergence $\mu^k\to\lambda^\theta$ is equivalent to $\gamma^k\to 0$ and $\psi^k\to\alpha$, where $\gamma^k=\sum_{i\notin I_0}\mu_i^k$, $\psi^k=\sum_{i\in I_0}h_i(z_i^k)$, and $\alpha=-\sum_{i\in I_0}\theta^*(\lambda_i^\theta)$ with $h_i(z)=\theta(z)-\lambda_i^\theta z$ and $z_i^k=[a_i^{\mathrm{T}}u_k-b_i]/r_k$. Since the proofs are somewhat involved, we introduce some additional notation and two technical lemmas.

In what follows we denote $\bar u_k$ the projection of u_k onto $S(\mathrm{P})$. We take E_0, $\Phi(\cdot)$, and $g(\cdot)$ as in section 3.3 and we let $\delta=\mathrm{diam}(S(\mathrm{P}))$ and $\sigma=\max_{i\notin I_0}\|a_i\|$. Moreover, we fix $\bar x\in\mathrm{ri}(S(\mathrm{P}))$ and set $\beta=\frac{1}{3}\min_{i\notin I_0}\beta_i$ with $\beta_i=b_i-a_i^{\mathrm{T}}\bar x>0$ for $i\notin I_0$. Finally, noting that θ' is nondecreasing we define its right inverse as $\theta'^{-1}(\varepsilon)=\sup\{u:\theta'(u)\le\varepsilon\}$.

The first lemma proves that $\|g_k\|\sim 0$ implies $\gamma^k\sim 0$.

LEMMA 4.5. *Assume* (H_0), $r_k\to 0$, $\sum\lambda_k=\infty$, $\sum\varepsilon_k\lambda_k<\infty$, $\varepsilon_k/r_k\to 0$, *and* $\nu_k\to 0$. *Then there exist* $\hat a_k\to 0$ *and an increasing function* $\varepsilon\mapsto\hat\rho(\varepsilon)$ *with* $\lim_{\varepsilon\downarrow 0}\hat\rho(\varepsilon)=0$ *such that* $0\le\gamma^k\le\hat\rho(\|g_k\|+\hat a_k)$ *for all* k *large enough.*

Proof. The result is evident if I_0^c is empty. Otherwise, Proposition 4.4 implies that $\|u_k-\bar u_k\|=d(u_k,S(\mathrm{P}))\to 0$ so that $a_i^{\mathrm{T}}[u_k-\bar u_k]\le\beta$ for k large enough. Decompose $I_0^c=I_k\cup J_k$, where I_k contains the i's such that $a_i^{\mathrm{T}}\bar u_k\ge b_i-2\beta$ and J_k the rest. For $i\in J_k$ we have $z_i^k\le-\beta/r_k$ so that $\mu_i^k\le\theta'(-\beta/r_k)$. Let us bound μ_i^k for $i\in I_k$. Taking $w=u_k+r_k(\bar x-\bar u_k)$ in (4.3) and noting that $c^{\mathrm{T}}(\bar x-\bar u_k)=0$ and $a_i^{\mathrm{T}}(\bar x-\bar u_k)=0$ for all $i\in I_0$, we obtain

$$\sum_{i\notin I_0}\theta(z_i^k)+\langle g_k,\bar x-\bar u_k\rangle\le\varepsilon_k/r_k+\sum_{i\notin I_0}\theta(z_i^k+y_i^k),\tag{4.5}$$

where $y_i^k=a_i^{\mathrm{T}}(\bar x-\bar u_k)$. We have $y_i^k\le-\beta$ for $i\in I_k$ and $y_i^k\le\sigma\delta$ for $i\in J_k$. Then, the monotonicity of θ and the subgradient inequality yield

$$\begin{aligned}
&\theta(z_i^k+y_i^k)\le\ \theta(z_i^k-\beta)\ \le\theta(z_i^k)-\beta\theta'(z_i^k-\beta) && \forall i\in I_k,\\
&\theta(z_i^k+y_i^k)\le\theta(z_i^k+\sigma\delta)\le\theta(z_i^k)+\sigma\delta\theta'(z_i^k+\sigma\delta) && \forall i\in J_k.
\end{aligned}$$

Using these bounds and (4.5) we deduce

$$\begin{aligned}
\beta\textstyle\sum_{i\in I_k}\theta'(z_i^k-\beta)&\le\delta\|g_k\|+\varepsilon_k/r_k+\sigma\delta\textstyle\sum_{i\in J_k}\theta'(z_i^k+\sigma\delta)\\
&\le\delta\|g_k\|+\varepsilon_k/r_k+\sigma\delta|I_0^c|\theta'(\sigma\delta-\beta/r_k).
\end{aligned}$$

Letting $b_k=\varepsilon_k/(r_k\delta)+\sigma|I_0^c|\theta'(\sigma\delta-\beta/r_k)$ it follows that $z_i^k\le\beta+\theta'^{-1}(\delta[\|g_k\|+b_k]/\beta)$ for all $i\in I_k$, and then setting $\nu(\varepsilon)=\theta'(\beta+\theta'^{-1}(\delta\varepsilon/\beta))$ we obtain $\mu_i^k=\theta'(z_i^k)\le\nu(\|g_k\|+b_k)$ for all $i\in I_k$. Therefore

$$\gamma^k=\sum_{i\in J_k}\mu_i^k+\sum_{i\in I_k}\mu_i^k\le|I_0^c|\theta'(-\beta/r_k)+|I_0^c|\nu(\|g_k\|+b_k)$$

and the result holds with $\hat\rho(\varepsilon)=|I_0^c|[\varepsilon+\nu(\varepsilon)]$ and $\hat a_k=\max\{b_k,\theta'(-\beta/r_k)\}$. ∎

The next lemma shows that $\|g_k\|\sim 0$ also implies $\psi^k\sim\alpha$.

LEMMA 4.6. *Assume* (H$_0$), $r_k \to 0$, $\sum \lambda_k = \infty$, $\sum \varepsilon_k \lambda_k < \infty$, $\varepsilon_k / r_k \to 0$, *and* $\nu_k \to 0$. *Then there exist* $\tilde{a}_k \to 0$ *and an increasing function* $\varepsilon \mapsto \tilde{\rho}(\varepsilon)$ *with* $\lim_{\varepsilon \downarrow 0} \tilde{\rho}(\varepsilon) = \alpha$ *such that* $\alpha \le \psi^k \le \tilde{\rho}(\|g_k\| + \tilde{a}_k)$ *for all* k *large enough.*

Proof. Let w_k be the projection of $y_k = [u_k - \bar{x}]/r_k$ on E_0 so that $\psi^k = \Phi(w_k) \ge \alpha$. Set $\varphi_k(y) = c^{\mathrm{T}} y + \sum_{i \in I} \theta(a_i^{\mathrm{T}} y - \beta_i / r_k)$. Taking $w = \bar{x} + r_k y$ in (4.3) it follows

$$\varphi_k(y_k) + \langle g_k, y - y_k \rangle \le \varphi_k(y) + \varepsilon_k / r_k \tag{4.6}$$

so that $g_k \in \partial_{\tilde{\varepsilon}_k} \varphi_k(y_k)$ with $\tilde{\varepsilon}_k = \varepsilon_k / r_k$. Using Brøndsted–Rockafellar's theorem we may find $\tilde{y}_k$ and $\tilde{g}_k \in \partial \varphi_k(\tilde{y}_k)$ such that $\|y_k - \tilde{y}_k\| \le \sqrt{\tilde{\varepsilon}_k}$ and $\|g_k - \tilde{g}_k\| \le \sqrt{\tilde{\varepsilon}_k}$. Let $\tilde{w}_k$ be the projection of $\tilde{y}_k$ onto E_0. Taking $y = \tilde{y}_k$ in (4.6) and noting that we have $\varphi_k(y_k) = \Phi(w_k) + \sum_{i \notin I_0} \theta(a_i^{\mathrm{T}} y_k - \beta_i / r_k)$ and similarly for $\varphi_k(\tilde{y}_k)$, we obtain

$$\psi^k = \Phi(w_k) \le \Phi(\tilde{w}_k) + \tilde{\varepsilon}_k + \sqrt{\tilde{\varepsilon}_k} \|g_k\| + \sum_{i \notin I_0} [\theta(a_i^{\mathrm{T}} \tilde{y}_k - \beta_i / r_k) - \theta(a_i^{\mathrm{T}} y_k - \beta_i / r_k)].$$

The terms in the last sum can be bounded from above by $\theta'(a_i^{\mathrm{T}} \tilde{y}_k - \beta_i / r_k) a_i^{\mathrm{T}} (\tilde{y}_k - y_k)$. Moreover, $a_i^{\mathrm{T}} (\tilde{y}_k - y_k) \le \sigma \sqrt{\tilde{\varepsilon}_k}$ and since $a_i^{\mathrm{T}} \tilde{y}_k - \beta_i / r_k = a_i^{\mathrm{T}} (\tilde{y}_k - y_k) + z_i^k$ and $z_i^k \le \theta'^{-1}(\mu_i^k) \le \theta'^{-1}(\gamma^k)$, for $i \notin I_0$ we get

$$\theta'(a_i^{\mathrm{T}} \tilde{y}_k - \beta_i / r_k) \le \theta'(\sigma \sqrt{\tilde{\varepsilon}_k} + \theta'^{-1}(\gamma^k)) \tag{4.7}$$

leading to the bound

$$\psi^k \le \Phi(\tilde{w}_k) + \tilde{\varepsilon}_k + \sqrt{\tilde{\varepsilon}_k} \|g_k\| + \sigma |I_0^c| \sqrt{\tilde{\varepsilon}_k}\, \theta'(\sigma \sqrt{\tilde{\varepsilon}_k} + \theta'^{-1}(\gamma^k)).$$

If k is large so that $\tilde{\varepsilon}_k \le 1$, letting $\eta_k = \sqrt{\tilde{\varepsilon}_k} + \|g_k\| + \sigma |I_0^c|\, \theta'(\sigma + \theta'^{-1}(\gamma^k))$ we get

$$\psi^k \le \Phi(\tilde{w}_k) + \sqrt{\tilde{\varepsilon}_k}\, \eta_k \le g(\|\nabla \Phi(\tilde{w}_k)\|) + \eta_k. \tag{4.8}$$

Moreover, $\tilde{g}_k = \nabla \Phi(\tilde{w}_k) + \sum_{i \notin I_0} \theta'(a_i^{\mathrm{T}} \tilde{y}_k - \beta_i / r_k) a_i$ so that using (4.7) and the inequality $\|\tilde{g}_k\| \le \sqrt{\tilde{\varepsilon}_k} + \|g_k\|$ we obtain

$$\|\nabla \Phi(\tilde{w}_k)\| \le \|\tilde{g}_k\| + \sigma |I_0^c|\, \theta'(\sigma + \theta'^{-1}(\gamma^k)) \le \eta_k,$$

which combined with (4.8) yields $\psi^k \le g(\eta_k) + \eta_k$. To conclude we observe that Lemma 4.5 gives $\gamma^k \le \hat{\rho}(\|g_k\| + \hat{a}_k)$ so that letting $\chi(\varepsilon) = \varepsilon + \sigma |I_0^c| \theta'(\sigma + \theta'^{-1}(\hat{\rho}(\varepsilon)))$ and $\tilde{a}_k = \max\{\hat{a}_k, \sqrt{\tilde{\varepsilon}_k}\}$ we get $\eta_k \le \chi(\|g_k\| + \tilde{a}_k)$, and then the result holds with $\tilde{\rho}(\varepsilon) = g(\chi(\varepsilon)) + \chi(\varepsilon)$. Notice that $\chi(\varepsilon)$ is increasing and tends to 0 when $\varepsilon \downarrow 0$, so that by Lemma 3.7 we have $\tilde{\rho}(\varepsilon)$ increasing with limit α when $\varepsilon \downarrow 0$. ∎

We may now prove our first dual convergence result.

THEOREM 4.7. *Assume* (H$_0$), $r_k \to 0$, $\sum \lambda_k = \infty$, $\nu_k \to 0$, *and* $\varepsilon_k \equiv 0$. *Suppose also* r_k *nonincreasing with* $q_k = (r_{k-1} - r_k)/(\lambda_k r_{k-1})$ *bounded. Then* $\mu^k \to \lambda^\theta$.

Proof. We prove that $\gamma^k \to 0$ and $\psi^k \to \alpha$. To this end we set $\psi_s^k = \sum_{i \in I} h_i(z_i^k) - \theta_s^k$, where $\gamma_0^k \equiv \theta_0^k \equiv 0$, and for $s = 1, \ldots, |I_0^c|$ we define

$$\theta_s^k = \min_S \left\{ \sum_{i \in S} \theta(z_i^k) : S \subset I_0^c, |S| = s \right\}, \tag{4.9}$$

$$\gamma_s^k = \min_S \left\{ \sum_{i \in S} \mu_i^k \ : S \subset I_0^c, |S| = s \right\}. \tag{4.10}$$

We will show inductively that $\gamma_s^k \to 0$ and $\psi_s^k \to \alpha_s$ with $\alpha_s = \alpha - (|I_0^c| - s)\,\theta^*(0)$. The result then follows since for $s = |I_0^c|$ we have $\gamma_s^k \equiv \gamma^k$ and $\psi_s^k \equiv \psi^k$. For the induction argument we first establish the following.

CLAIM. *If $\gamma_s^k \to 0$, then $\psi_s^k \to \alpha_s$.*

Proof. Let us study the increment $\psi_s^k - \psi_s^{k-1} = (\psi_0^k - \psi_0^{k-1}) + (\theta_s^{k-1} - \theta_s^k)$. Take S_k attaining the minimum in (4.9) so that

$$\begin{aligned}\psi_s^k - \psi_s^{k-1} &\le (\psi_0^k - \psi_0^{k-1}) + \textstyle\sum_{i\in S_k}[\theta(z_i^{k-1}) - \theta(z_i^k)]\\ &= \textstyle\sum_{i\notin S_k}[h_i(z_i^k) - h_i(z_i^{k-1})]\\ &\le \textstyle\sum_{i\notin S_k} h_i'(z_i^k)[z_i^k - z_i^{k-1}]\\ &= \textstyle\sum_{i\notin S_k}(\mu_i^k - \lambda_i^\theta)[z_i^k - z_i^{k-1}].\end{aligned}$$

Setting $\xi_k = \nu_k + q_k(u_{k-1} - \bar u_{k-1})$, a straightforward computation gives

$$[z_i^k - z_i^{k-1}] = (\lambda_k/r_k)[a_i^{\mathrm T}(\xi_k - g_k) + q_k(a_i^{\mathrm T}\bar u_{k-1} - b_i)],$$

and since $(a_i^{\mathrm T}\bar u_{k-1} - b_i) \le 0$ and $\lambda_i^\theta(a_i^{\mathrm T}\bar u_{k-1} - b_i) = 0$ we deduce

$$\psi_s^k - \psi_s^{k-1} \le (\lambda_k/r_k)\sum_{i\notin S_k}(\mu_i^k - \lambda_i^\theta)a_i^{\mathrm T}[\xi_k - g_k].$$

Now $\sum_{i\notin S_k}(\mu_i^k - \lambda_i^\theta)a_i = \nabla f(u_k, r_k) - \eta_k = g_k - \eta_k$ with $\eta_k = \sum_{i\in S_k}\mu_i^k a_i$, so that

$$\psi_s^k - \psi_s^{k-1} \le (\lambda_k/r_k)\langle g_k - \eta_k, \xi_k - g_k\rangle. \tag{4.11}$$

Using (4.11) it follows that when $\psi_s^k > \psi_s^{k-1}$ we have $\|g_k - \xi_k\| \le \|\xi_k - \eta_k\|$ and then $\|g_k\| \le \delta_k$ with $\delta_k = \|\xi_k\| + \|\xi_k - \eta_k\| \to 0$ (note that $\xi_k \to 0$ since q_k is bounded, and $\eta_k \to 0$ since $\sum_{i\in S_k}\mu_i^k = \gamma_s^k \to 0$ by assumption of the claim). Then Lemmas 4.6 and 4.5 imply $\psi^k \le \tilde\rho(\delta_k + \tilde a_k)$ and $\gamma^k \le \hat\rho(\delta_k + \hat a_k)$, so that letting $\chi_s^k = \tilde\rho(\delta_k + \tilde a_k) + (|I_0^c| - s)\theta(\theta'^{-1}(\hat\rho(\delta_k + \hat a_k)))$ and using the definition of ψ_s^k we get

$$\psi_s^k > \psi_s^{k-1} \;\Rightarrow\; \alpha_s \le \psi_s^k \le \chi_s^k.$$

Since $\chi_s^k \to \alpha_s$ it follows that ψ_s^k has a limit $\bar\alpha_s \ge \alpha_s$. If $\bar\alpha_s > \alpha_s$, then Lemmas 4.5 and 4.6 imply $\liminf \|g_k\| > 0$, so we may find a constant $c > 0$ with $\langle g_k - \eta_k, g_k - \xi_k\rangle > c$ for all k large. Then (4.11) gives $c\lambda_k/r_k \le \psi_s^{k-1} - \psi_s^k$ implying $\sum \lambda_k/r_k \le (\psi_s^0 - \bar\alpha_s)/c < \infty$, which is impossible. This contradiction proves $\bar\alpha_s = \alpha_s$ so that $\psi_s^k \to \alpha_s$ establishing the claim. ∎

To complete the proof of Theorem 4.7 we distinguish two cases.

Case 1. $\theta^*(0) < \infty$. Since $\gamma_0^k \equiv 0$ the claim implies $\psi_0^k \to \alpha_0$, from which we get $\psi^k \to \alpha$, as well as $\theta(z_i^k) \to -\theta^*(0)$ for all $i \notin I_0$ so that $\gamma^k \to 0$.

Case 2. $\theta^*(0) = \infty$. We prove inductively that $\gamma_s^k \to 0$ for $s = 0, 1, \dots, |I_0^c|$. This holds by definition for $s = 0$. Suppose $\gamma_s^k \to 0$ for some $s < |I_0^c|$. The claim gives $\psi_s^k \to \alpha_s = -\infty$ and since $\psi_s^k = \psi^k + \sum_{i\in I_0^c\setminus S_k}\theta(z_i^k)$ with $\psi^k \ge \alpha$, there must exist an index $i(k) \in I_0^c \setminus S_k$ such that $\theta(z_{i(k)}^k) \to -\infty$. Then $\mu_{i(k)}^k = \theta'(z_{i(k)}^k) \to 0$ and therefore $\gamma_{s+1}^k \le \gamma_s^k + \mu_{i(k)}^k \to 0$, achieving the induction step. Taking $s = |I_0^c|$ we obtain $\gamma^k \equiv \gamma_s^k \to 0$ and $\psi^k \equiv \psi_s^k \to \alpha$, completing the proof. ∎

Our last result allows $\varepsilon_k > 0$ but requires θ to be bounded from below, extending [5, Theorem 1.2] to more general penalty functions.

THEOREM 4.8. *Assume* (H$_0$), $r_k \to 0$, $\sum \lambda_k = \infty$, $\sum \varepsilon_k \lambda_k < \infty$, $\varepsilon_k / r_k \to 0$, $\nu_k \to 0$, *and either* $\varepsilon_k/\lambda_k \to 0$ *or* $\sum \varepsilon_k / r_k < \infty$. *Suppose also that* θ *is bounded from below and* r_k *nonincreasing with* $q_k = (r_{k-1} - r_k)/(\lambda_k r_{k-1})$ *bounded. Then* $\mu^k \to \lambda^\theta$.

Proof. As in Case 1 above, it suffices to prove $\psi_0^k \to \alpha_0$. This will be established in the following three claims. As before we set $\xi_k = \nu_k + q_k(u_{k-1} - \bar{u}_{k-1}) \to 0$.

CLAIM 1. $\psi_0^k - \psi_0^{k-1} \le [\lambda_k \langle g_k, \xi_k - g_k \rangle + \varepsilon_k]/r_k$.

Proof. We observe that $\psi_0^k = [f(u_k, r_k) - v]/r_k$, where v denotes the optimal value of problem (P). Let $\pi_k = r_{k+1}/r_k$ and define $x_k = (1-\pi_k)\bar{u}_k + \pi_k u_k$. Then $[a_i^{\mathrm{T}} x_{k-1} - b_i] \le r_k [a_i^{\mathrm{T}} u_{k-1} - b_i]/r_{k-1}$ and therefore

$$f(x_{k-1}, r_k) \le (1 - \pi_{k-1}) v + \pi_{k-1} c^{\mathrm{T}} u_{k-1} + r_k \sum_{i \in I} \theta([a_i^{\mathrm{T}} u_{k-1} - b_i]/r_{k-1})$$

so that taking $x = x_{k-1}$ in (4.3) we obtain $\psi_0^k + \langle g_k, x_{k-1} - u_k \rangle / r_k \le \psi_0^{k-1} + \varepsilon_k / r_k$. The conclusion follows since $x_{k-1} - u_k = \lambda_k [g_k - \xi_k]$. ∎

CLAIM 2. ψ_0^k *converges.*

Proof. We distinguish the alternative cases $\varepsilon_k/\lambda_k \to 0$ and $\sum \varepsilon_k / r_k < \infty$.

If $\varepsilon_k/\lambda_k \to 0$, setting $\delta_k = \frac{1}{2}[\|\xi_k\| + \sqrt{\|\xi_k\|^2 + 4\varepsilon_k/\lambda_k}]$ we have $\delta_k \to 0$. If $\psi_0^k > \psi_0^{k-1}$, Claim 1 implies $\|g_k\| \le \delta_k$, and then Lemmas 4.6 and 4.5 give $\psi^k \le \tilde{\rho}(\delta_k + \tilde{a}_k)$ and $\gamma^k \le \hat{\rho}(\delta_k + \hat{a}_k)$. Letting $\chi_0^k = \tilde{\rho}(\delta_k + \tilde{a}_k) + |I_0^c| \theta(\theta'^{-1}(\hat{\rho}(\delta_k + \hat{a}_k)))$ we get

$$\psi_0^k > \psi_0^{k-1} \;\Rightarrow\; \alpha_0 \le \psi_0^k \le \chi_0^k$$

with $\chi_0^k \to \alpha_0$ so that ψ_0^k converges.

The alternative case $\sum \varepsilon_k / r_k < \infty$ is similar. Letting $\tilde{\psi}_0^k = \psi_0^k + \sum_{k+1}^{\infty} \varepsilon_i / r_i$ with ψ_0^k defined as above but with $\delta_k = \|\xi_k\|$, Claim 1 and Lemmas 4.5 and 4.6 imply

$$\tilde{\psi}_0^k > \tilde{\psi}_0^{k-1} \;\Rightarrow\; \|g_k\| \le \delta_k \;\Rightarrow\; \alpha_0 \le \tilde{\psi}_0^k \le \tilde{\chi}_0^k := \chi_0^k + \sum_{k+1}^{\infty} \varepsilon_i / r_i$$

with $\tilde{\chi}_0^k \to \alpha_0$ so that $\tilde{\psi}_0^k$ converges and then ψ_0^k converges as well. ∎

CLAIM 3. $\psi_0^k \to \alpha_0$.

Proof. Using Claim 2 and Lemmas 4.5 and 4.6 it suffices to prove $\liminf \|g_k\| = 0$. If this were not the case, we could find a constant $c > 0$ such that $\langle g_k, g_k - \xi_k \rangle > c$ for all k large, so that Claim 1 implies $[c\lambda_k - \varepsilon_k]/r_k \le \psi_0^{k-1} - \psi_0^k$. Considering the alternative assumption $\varepsilon_k/\lambda_k \to 0$ or $\sum \varepsilon_k / r_k < \infty$, it follows that $\sum \lambda_k / r_k < \infty$, which gives a contradiction. ∎

REFERENCES

[1] P. ALART AND B. LEMAIRE, *Penalization in nonclassical convex programming via variational convergence*, Math. Programming, 51 (1991), pp. 307–331.

[2] O. ALEMANY AND R. COMINETTI, *Steepest descent evolution equations: Asymptotic behavior of solutions and rate of convergence*, Trans. Amer. Math. Soc., 351 (1999), pp. 4847–4860.

[3] F. ALVAREZ, *Métodos Continuos en Optimización Paramétrica: El Método de Newton y Aplicaciones a la Optimización Estructural*, Engineering thesis, Universidad de Chile, Santiago, Chile, 1996.

[4] F. ALVAREZ, *Absolute minimizer in convex programming by exponential penalty*, J. Convex Anal., 7 (2000), pp. 197–202.

[5] F. ALVAREZ AND R. COMINETTI, *Primal and dual convergence of a proximal point exponential penalty method for linear programming*, Math. Program., to appear.

[6] H. ATTOUCH, *Viscosity solutions of minimization problems*, SIAM J. Optim., 6 (1996), pp. 769–806.

[7] H. Attouch and R. Cominetti, *A dynamical approach to convex minimization coupling approximation with the steepest descent method*, J. Differential Equations, 128 (1996), pp. 519–540.

[8] H. Attouch and A. Damlamian, *Strong solutions for parabolic variational inequalities*, Nonlinear Anal., 2 (1978), pp. 329–353.

[9] A. Auslender, R. Cominetti, and M. Haddou, *Asymptotic analysis for penalty methods in convex and linear programming*, Math. Oper. Res., 22 (1997), pp. 43–62.

[10] A. Auslender, J.P. Crouzeix, and P. Fedit, *Penalty-proximal methods in convex programming*, J. Optim. Theory Appl., 55 (1987), pp. 1–21.

[11] M.A. Bahraoui and B. Lemaire, *Convergence of diagonally stationary sequences in convex optimization*, Set-Valued Anal., 2 (1994), pp. 49–61.

[12] B. Baillon and R. Cominetti, *A convergence result for non-autonomous subgradient evolution equations and its application to the steepest descent exponential penalty trajectory in linear programming*, J. Funct. Anal., 187 (2002), pp. 263–273.

[13] A. Ben-Tal and M. Zibulevsky, *Penalty/barrier multiplier methods for convex programming problems*, SIAM J. Optim., 7 (1997), pp. 347–366.

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

[15] H. Brézis, *Opérateurs maximaux monotones et sémigroups de contractions dans les espaces de Hilbert*, Mathematical Studies 5, North-Holland, Amsterdam, 1973.

[16] R.E. Bruck, *Asymptotic convergence of nonlinear contraction semigroups in Hilbert space*, J. Funct. Anal., 18 (1975), pp. 15–26.

[17] C.W. Carroll, *The created response surface technique for optimizing nonlinear restrained systems*, Operations Res., 9 (1961), pp. 169–185.

[18] T. Champion, *Tubularity and asymptotic convergence of penalty trajectories in convex programming*, SIAM J. Optim., 13 (2002), pp. 212–227.

[19] C. Chen and O.L. Mangasarian, *A class of smoothing functions for nonlinear and mixed complementarity problems*, Comput. Optim. Appl., 5 (1996), pp. 97–138.

[20] R. Cominetti, *Coupling the proximal point algorithm with approximation methods*, J. Optim. Theory Appl., 95 (1997), pp. 581–600.

[21] R. Cominetti, *Asymptotic convergence of Euler's method for the exponential penalty in linear programming*, J. Convex Anal., 2 (1995), pp. 145–152.

[22] R. Cominetti, *Nonlinear averages and convergence of penalty trajectories in convex programming*, in Proceedings of the Workshop on Ill-Posed Variational Problems and Regularization Techniques, Trier, 1998, Lecture Notes in Econom. and Math. Systems 477, Springer-Verlag, Berlin, Heidelberg, 1999, pp. 65–78.

[23] R. Cominetti and J. San Martín, *Trajectory analysis for the exponential penalty method in linear programming*, Math. Programming, 67 (1994), pp. 169–187.

[24] M. Courdurier, *Análisis asintótico en penalización convexa*, Memoria de Ingeniería Civil Matemática, Universidad de Chile, Santiago, Chile, 2001.

[25] A. Fiacco and G. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, Wiley, New York, 1968.

[26] K.R. Frisch, *The Logarithmic Potential Method of Convex Programming*, Memorandum, University Institute of Economics, Oslo, 1955.

[27] H. Furuya, K. Miyashiba, and N. Kenmochi, *Asymptotic behavior of solutions to a class of nonlinear evolution equations*, J. Differential Equations, 62 (1986), pp. 73–94.

[28] J.B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms* II, Springer-Verlag, Berlin, 1996.

[29] A.N. Iusem, B.F. Svaiter, and J.X. Da Cruz Neto, *Central paths, generalized proximal point methods, and Cauchy trajectories in Riemannian manifolds*, SIAM J. Control Optim., 37 (1999), pp. 566–588.

[30] A. Iusem and R. Monteiro, *On dual convergence of the generalized proximal point method with Bregman distances*, Math. Oper. Res., 25 (2000), pp. 606–624.

[31] A. Kaplan, *On a convex programming method with internal regularization*, Soviet Math. Dokl., 19 (1978), pp. 795–799.

[32] A. Kaplan and R. Tichatschke, *Proximal methods in view of interior-point strategies*, J. Optim. Theory Appl., 98 (1998), pp. 399–429.

[33] N. Kenmochi, *Nonlinear evolution equations with variable domains in Hilbert spaces*, Proc. Japan Acad. Ser. A Math. Sci., 53 (1977), pp. 163–166.

[34] B. Lemaire, *Coupling optimization methods and variational convergence*, in Trends in Mathematical Optimization, Internat. Ser. Numer. Math., Birkhäuser, Basel, Switzerland, 1988, pp. 163–179.

[35] B. Martinet, *Régularisation d'inéquations variationnelles par approximations successives*, Rev. Française Informat. Recherche Operationnelle, 4 (1970), pp. 154–159.
[36] L. McLinden, *An analogue of Moreau's proximation theorem with applications to the nonlinear complementarity problem*, Pacific J. Math., 88 (1980), pp. 101–161.
[37] N. Megiddo, *Pathways to the optimal set in linear programming*, in Progress in Mathematical Programming: Interior Point Methods, N. Megiddo, ed., Springer-Verlag, New York, 1989, pp. 131–158.
[38] R. Monteiro and F. Zhou, *On the existence and convergence of the central path for convex programming and some duality results*, Comput. Optim. Appl., 10 (1998), pp. 51–77.
[39] K. Mouallif, *Sur la convergence d'une méthode associant pénalisation et réguralisation*, Bull. Soc. Roy. Sci. Liège, 56 (1987), pp. 175–180.
[40] K. Mouallif and P. Tossings, *Une méthode de pénalisation exponentielle associée à une régularisation proximale*, Bull. Soc. Roy. Sci. Liège, 56 (1987), pp. 181–190.
[41] A. Moudafi, *Coupling proximal algorithm and Tikhonov method*, Nonlinear Times Digest, 1 (1994), pp. 203–210.
[42] R. Polyak, *Modified barrier functions: Theory and methods*, Math. Programming, 54 (1992), pp. 177–222.
[43] R.T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.
[44] R.T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[45] R.T. Rockafellar, *Ordinary convex programs without a duality gap*, J. Optim. Theory Appl., 7 (1971), pp. 143–148.

SIAM J. OPTIM.
Vol. 13, No. 3, pp. 766–782

A NEW EFFICIENT LARGE-UPDATE PRIMAL-DUAL INTERIOR-POINT METHOD BASED ON A FINITE BARRIER*

Y. Q. BAI†, M. EL GHAMI‡, AND C. ROOS‡

Abstract. We introduce a new barrier-type function which is not a barrier function in the usual sense: it has finite value at the boundary of the feasible region. Despite this, the iteration bound of a large-update interior-point method based on this function is shown to be $O(\sqrt{n}\,(\log n)\log\frac{n}{\varepsilon})$, which is as good as the currently best known bound for large-update methods. The recently introduced property of *exponential convexity* for the kernel function underlying the barrier function, as well as the strong convexity of the kernel function, are crucial in the analysis.

Key words. linear optimization, interior-point method, primal-dual method, large-update method, polynomial complexity

AMS subject classification. 90C05

PII. S1052623401398132

1. Introduction.

Since the path-breaking paper of Karmarkar [5], linear optimization (LO) revived as an active area of research. Today the resulting interior-point methods (IPMs) are among the most effective methods for solving wide classes of LO problems. Many researchers have proposed and analyzed various IPMs for LO and a large amount of results have been reported. For a survey we refer to recent books on the subject [17, 19, 21]. An interesting fact is that almost all known polynomial-time variants of IPMs use the so-called *central path* [18] as a guideline to the optimal set and some variant of Newton's method to follow the central path approximately. Therefore, analyzing the behavior of Newton's method has been a crucial issue in the theoretical investigation of IPMs. In this paper we consider so-called primal-dual methods. It is generally agreed that these methods are the most efficient methods from a computational point of view (see, e.g., Andersen et al. [1]). These methods use the Newton direction as a search direction; this direction is closely related to the well-known primal-dual logarithmic barrier function.

At present there is still a gap between the practical behavior of the algorithms and the theoretical performance results, in favor of the practical behavior. This is especially true for so-called large-update methods. If n denotes the number of inequalities in the problem, then the theoretical complexity analysis of large-update methods yields an $O(n\log(n/\varepsilon))$ iteration bound, where ε represents the desired accuracy of the solution. In practice, however, large-update methods are much more efficient than the so-called small-update methods for which the theoretical iteration bound is only $O(\sqrt{n}\,\log(n/\varepsilon))$. So the current theoretical bounds differ by a factor $\sqrt{n}$, in favor of the small-update methods. This gap is significant.

Recently, the gap could be narrowed by deviating from the usual approach. Replacing the logarithmic barrier by a so-called *self-regular* barrier function, and modi-

*Received by the editors November 15, 2001; accepted for publication (in revised form) August 7, 2002; published electronically January 3, 2003.

http://www.siam.org/journals/siopt/13-3/39813.html

†Department of Mathematics, Shanghai University, Shanghai, 200436 China (Y.Bai@its.tudelft.nl). This author was supported by the Science Foundation of Shanghai Municipal Commission of Education and Chinese Scholarship Council.

‡Faculty of Information Technology and Systems, Delft University of Technology, P.O. Box 5031, 2600 GA Delft, The Netherlands (M.Elghami@its.tudelft.nl, C.Roos@its.tudelft.nl, http://ssor.twi.tudelft.nl/~roos).

fying the search direction accordingly, a large-update method was obtained for which the theoretical iteration bound is $O(\sqrt{n}\,(\log n)\log(n/\varepsilon))$ [10, 11, 12, 13, 14]. Thus the gap between the theoretical iteration bounds for small- and large-update methods has been narrowed.

In this paper we introduce a new barrier function which is not self-regular. Based on this barrier function we devise a new large-update method which has an $O(\sqrt{n}\ (\log n)\log(n/\varepsilon))$ iteration bound, the currently best bound for large-update methods. The barrier function is different from all known barrier functions in the sense that it is finite at the boundary of the feasible region. Until now all barrier functions used in the analysis of polynomial-time IPMs become unbounded when approaching the boundary of the feasible region. The result of this paper is therefore quite surprising.

The paper is organized as follows. In section 2 we first briefly recall the classical approach and its relation to the well-known (primal-dual) logarithmic barrier function. In section 2.4 we describe the idea underlying the approach of the paper. A crucial observation is that any (univariate) function that is strongly convex on the positive real axis and that attains its minimal value determines in a natural way a primal-dual IPM. The (univariate) function underlying the method is called a kernel function. In section 2.5 we introduce the kernel function considered in this paper. In the analysis of the corresponding algorithm we use the notion of *exponential convexity*, a notion that was also used in [10, 11, 12, 13, 14]. In sections 2.6 and 2.7 we derive some relevant properties that play a crucial role in the analysis of the algorithm; the complexity analysis is performed in section 3. Finally, section 4 contains some concluding remarks and directions for future research.

We use the following notational conventions. Throughout the paper, $\|\cdot\|$ denotes the 2-norm of a vector, whereas $\|\cdot\|_\infty$ denotes the infinity norm. For any $x = (x_1, x_2, \ldots, x_n)^T \in \mathbf{R}^n$, $x_{\min}$ denotes the smallest and $x_{\max}$ the largest value of the components of x. If also $s \in \mathbf{R}^n$, then xs denotes the coordinatewise (or Hadamard) product of the vectors x and s. Furthermore, e denotes the all-one vector of length n. The nonnegative orthant and positive orthant are denoted as $\mathbf{R}^n_+$ and $\mathbf{R}^n_{++}$, respectively. Finally, if $z \in \mathbf{R}^n_+$ and $f : \mathbf{R}_+ \to \mathbf{R}_+$, then $f(z)$ denotes the vector in $\mathbf{R}^n_+$ whose ith component is $f(z_i)$, with $1 \le i \le n$.

2. Preliminaries.

2.1. The central path.

We deal with the LO problem in standard format:

$$(P) \qquad \min\{c^T x : Ax = b, x \ge 0\},$$

where $A \in \mathbf{R}^{m\times n}$, $b \in \mathbf{R}^m$, $c \in \mathbf{R}^n$, and its dual problem

$$(D) \qquad \max\{b^T y : A^T y + s = c, s \ge 0\}.$$

We assume that both (P) and (D) satisfy the interior-point condition (IPC); i.e., there exists (x^0, s^0, y^0) such that

$$Ax^0 = b,\ x^0 > 0, \quad A^T y^0 + s^0 = c,\ s^0 > 0. \tag{1}$$

It is well known that the IPC can be assumed without loss of generality. In fact we may, and will, assume that $x^0 = s^0 = \mathrm{e}$. For this and some other properties mentioned below, see, e.g., [17]. Finding an optimal solution of (P) and (D) is equivalent to

solving the following system:

$$\begin{aligned} Ax &= b, \quad x \geq 0, \\ A^T y + s &= c, \quad s \geq 0, \\ xs &= 0. \end{aligned} \tag{2}$$

The basic idea of primal-dual IPMs is to replace the third equation in (2), the so-called *complementarity condition* for (P) and (D), by the parametrized equation $xs = \mu \mathrm{e}$, with $\mu > 0$. Thus we consider the system

$$\begin{aligned} Ax &= b, \quad x \geq 0, \\ A^T y + s &= c, \quad s \geq 0, \\ xs &= \mu \mathrm{e}. \end{aligned} \tag{3}$$

If $\operatorname{rank}(A) = m$ and the IPC holds, then for each $\mu > 0$ the parameterized system (3) has a unique solution. This solution is denoted as $(x(\mu), y(\mu), s(\mu))$, and we call $x(\mu)$ the μ-center of (P) and $(y(\mu), s(\mu))$ the μ-center of (D). The set of μ-centers (with μ running through all positive real numbers) gives a homotopy path, which is called *the central path* of (P) and (D). The relevance of the central path for LO was first recognized by Sonnevend [18] and Megiddo [6]. If $\mu \to 0$, then the limit of the central path exists and since the limit points satisfy the complementarity condition, the limit yields optimal solutions for (P) and (D).

2.2. Primal-dual path-following methods. IPMs follow the central path approximately. We briefly describe the usual approach. Without loss of generality we assume that $(x(\mu), y(\mu), s(\mu))$ is known for some positive μ. For example, due to the above assumption we may assume this for $\mu = 1$, with $x(1) = s(1) = \mathrm{e}$. We then decrease μ to $\mu := (1-\theta)\mu$ for some fixed $\theta \in (0, 1)$ and we solve the following Newton system:

$$\begin{aligned} A\Delta x &= 0, \\ A^T \Delta y + \Delta s &= 0, \\ s\Delta x + x\Delta s &= \mu \mathrm{e} - xs. \end{aligned} \tag{4}$$

This system uniquely defines a search direction $(\Delta x, \Delta s, \Delta y)$. By taking a step along the search direction, with the step size defined by some line search rules, one constructs a new triple (x, y, s). If necessary, we repeat the procedure until we find iterates that are "close" to $(x(\mu), y(\mu), s(\mu))$. Then μ is again reduced by the factor $1 - \theta$ and we apply Newton's method targeting at the new μ-centers, and so on. This process is repeated until μ is small enough, say until $n\mu \leq \varepsilon$; at this stage we have found an ε-solution of the problems (P) and (D).

Let us mention that in practice many LO solvers use the ε-solution to construct a basic solution and then produce an optimal basic solution by *crossing over* to the Simplex method. An alternative way is to apply a rounding procedure as described by Ye [20] (see also Mehrotra and Ye [7] and Roos, Terlaky, and Vial [17]).

The choice of the so-called barrier update parameter θ plays an important role both in theory and practice of IPMs. Usually, if θ is a constant independent of the dimension n of the problem, for instance $\theta = \frac{1}{2}$, then we call the algorithm a *large-update* (or *long-step*) method. If θ depends on the dimension of the problem, such as $\theta = \frac{1}{\sqrt{n}}$, then the algorithm is named a *small-update* (or *short-step*) method.

Recall that small-update methods have the best iteration bound; they require $O\left(\sqrt{n}\log\frac{n}{\varepsilon}\right)$ iterations to produce an ε-solution. On the other hand, large-update methods, which are in practice much more efficient than small-update methods [1], have a worse iteration bound, namely $O\left(n\log\frac{n}{\varepsilon}\right)$ [17, 19, 21]. This gap between theory and practice has been referred to as the irony of IPM methods [15].

The result of a Newton step with step size α is denoted as

$$x_+ = x + \alpha\Delta x, \quad y_+ = y + \alpha\Delta y, \quad s_+ = s + \alpha\Delta s. \tag{5}$$

The choice of the step size α ($0 < \alpha \leq 1$) is another crucial issue in the analysis of the algorithm. It has to be taken such that the closeness of the iterates to the current μ-center improves by a sufficient amount. In the theoretical analysis the step size α is usually given a value that depends on the closeness of the current iterates to the μ-center.

2.3. Relation with the logarithmic barrier function. Obviously, when analyzing an algorithm as just described, we are in need of a measure for the "closeness" of a primal-dual pair (x, s) to the μ-center $(x(\mu), s(\mu))$. The most popular tool for measuring this closeness is the so-called primal-dual logarithmic barrier function (cf., e.g., [17]), which is a primal-dual variant of the primal logarithmic barrier function introduced by Frisch [3]. Up to the constant $n\log\mu - n$, the primal-dual logarithmic barrier function is given by

$$\Phi_c(x, s; \mu) = \frac{x^T s}{\mu} - \sum_{i=1}^{n} \log\frac{x_i s_i}{\mu} - n. \tag{6}$$

Its usefulness can most easily be understood by introducing the vector

$$v := \sqrt{\frac{xs}{\mu}}. \tag{7}$$

Note that the pair (x, s) coincides with the μ-center $(x(\mu), s(\mu))$ if and only if $v = \mathrm{e}$. Now defining

$$\psi_c(t) := \frac{t^2 - 1}{2} - \log t, \quad t > 0, \tag{8}$$

one may easily verify that $\Phi_c(x, s; \mu)$ can be expressed in terms of the vector v as follows:

$$\Phi_c(x, s; \mu) = 2\Psi_c(v) := 2\sum_{i=1}^{n} \psi_c(v_i). \tag{9}$$

Since $\psi_c(t)$ is strictly convex, and attains its minimal value at $t = 1$, with $\psi_c(1) = 0$, it follows that $\Phi_c(x, s; \mu)$ is nonnegative and vanishes if and only if $v = \mathrm{e}$, i.e., if and only if $xs = \mu\mathrm{e}$. Thus we see that the μ-centers $x(\mu)$ and $s(\mu)$ can be characterized as the minimizers of $\Phi_c(x, s; \mu)$.

Another crucial property of the logarithmic barrier function becomes apparent when applying a widely used scaling scheme. Using the vector v, as defined in (7), we define scaled versions of the displacements Δx and Δs as follows:

$$d_x := \frac{v\Delta x}{x}, \qquad d_s := \frac{v\Delta s}{s}. \tag{10}$$

Now one easily checks that the system (4), which defines the Newton search directions, can be rewritten as

$$\begin{aligned} \bar{A}d_x &= 0, \\ \bar{A}^T\Delta y + d_s &= 0, \\ d_x + d_s &= v^{-1} - v, \end{aligned} \tag{11}$$

where $\bar{A} = AV^{-1}X$, with $V = \operatorname{diag}(v), X = \operatorname{diag}(x)$. The last equation in the above system is called the *scaled centering equation*. Yet we observe that the right-hand side in this equation is nothing else than the negative gradient of $\Psi_c(v)$, as can be easily verified. In other words,

$$d_x + d_s = v^{-1} - v = -\nabla\Psi_c(v). \tag{12}$$

Note that d_x and d_s are orthogonal vectors, since the matrix d_x belongs to the null space and d_s to the row space of the matrix $\bar{A}$. Thus we arrive at the important conclusion that the scaled search directions d_x and d_s form an orthogonal decomposition of the steepest descent direction of the scaled logarithmic barrier function $\Psi_c(v)$.

2.4. Generalization to new barrier functions. Now we are ready to describe the idea underlying the approach in this paper: in the scaled centering equation (12), which defines the search directions, we replace the scaled barrier function $\Psi_c(v)$ by a strictly convex function $\Psi(v)$, $v \in \mathbf{R}^n_{++}$, such that $\Psi(v)$ is minimal at $v = \mathrm{e}$ and $\Psi(\mathrm{e}) = 0$. Thus the new scaled centering equation becomes

$$d_x + d_s = -\nabla\Psi(v). \tag{13}$$

Note that since d_x and d_s are orthogonal, we will have $d_x = 0$ and $d_s = 0$ if and only if $v = \mathrm{e}$, i.e., if and only if $x = x(\mu)$ and $s = s(\mu)$, as it should.

To simplify matters we will restrict ourselves to the case where $\Psi(v)$ is separable with identical coordinate functions. Thus, letting ψ denote the function on the coordinates, we have

$$\Psi(v) = \sum_{i=1}^{n} \psi(v_i), \tag{14}$$

where $\psi(t) : D \to \mathbf{R}_+$, with $\mathbf{R}_{++} \subseteq D$, is strictly convex and minimal at $t = 1$, with $\psi(1) = 0$. We call the univariate function $\psi(t)$ the *kernel function* of the barrier function $\Psi(v)$. Observe that $\psi_c(t)$, as given by (8), is the kernel function of the logarithmic barrier function.

In principle any kernel function gives rise to a primal-dual algorithm. The generic form of this algorithm is given below. The parameters τ, θ and the step size α should be chosen in such a way that the algorithm is "optimized" in the sense that the number of iterations required by the algorithm is as small as possible. Obviously, the resulting iteration bound will depend on the kernel function underlying the algorithm, and our main task becomes finding a kernel function that minimizes the iteration bound. Figure 1 gives some examples of kernel functions that have been analyzed in earlier papers and the complexity results for the corresponding algorithms. In the third and sixth cases, the bound is minimal if $q = \frac{1}{2}\log n$. For this value of q one has

$$qn^{\frac{q+1}{2q}} = q\sqrt{n}\, n^{\frac{1}{2q}} = eq\sqrt{n} = \tfrac{e}{2}\sqrt{n}\,\log n,$$

where we used the identity $n^{\frac{1}{\log n}} = e$. This gives the currently best known iteration bound, namely $O(\sqrt{n}\,(\log n)\log\frac{n}{\varepsilon})$.

GENERIC PRIMAL-DUAL ALGORITHM FOR LO

Input:
A threshold parameter $\tau > 0$;
an accuracy parameter $\varepsilon > 0$;
a fixed barrier update parameter θ, $0 < \theta < 1$;
begin
$x := e$; $s := e$; $\mu := 1$;
while $n\mu \geq \varepsilon$ **do**
begin
$\mu := (1-\theta)\mu$;
while $\Psi(v) > \tau$ **do**
begin
$x := x + \alpha\Delta x$;
$s := s + \alpha\Delta s$;
$y := y + \alpha\Delta y$;
$v := \sqrt{\frac{xs}{\mu}}$;
end
end
end

	kernel function	iteration bound	ref.
1	$\frac{t^2-1}{2} - \log t$	$O(n) \log \frac{n}{\varepsilon}$	e.g., [17]
2	$\frac{1}{2}\left(t - \frac{1}{t}\right)^2$	$O\left(n^{\frac{2}{3}}\right) \log \frac{n}{\varepsilon}$	[8, 9]
3	$\frac{t^2-1}{2} + \frac{t^{1-q}-1}{q-1}$, $q > 1$	$O\left(qn^{\frac{q+1}{2q}}\right) \log \frac{n}{\varepsilon}$	[11, 12]
4	$\frac{t^2-1}{2} + \frac{e^{\frac{1}{t}-e}}{e}$	$O\left(\sqrt{n}\ \log^2 n\right) \log \frac{n}{\varepsilon}$	[16]
5	$\frac{t^2-1}{2} - \int_1^t e^{\frac{1}{\xi}-1} d\xi$	$O\left(\sqrt{n}\ \log^2 n\right) \log \frac{n}{\varepsilon}$	[16]
6	$\frac{t^2-1}{2} + \frac{t^{1-q}-1}{q(q-1)} - \frac{q-1}{q}(t-1)$	$O\left(qn^{\frac{q+1}{2q}}\right) \log \frac{n}{\varepsilon}$	[12]
7	$\frac{t^2-1}{2} + \frac{(e-1)^2}{e}\frac{1}{e^t-1} - \frac{e-1}{e}$	$O\left(n^{\frac{3}{4}}\right) \log \frac{n}{\varepsilon}$	[2]

FIG. 1. *Examples of kernel functions and complexity results.*

2.5. The kernel function considered in this paper. We consider the kernel function[1]

$$\psi(t) = \frac{t^2-1}{2} + \frac{1}{\sigma}\left(e^{\sigma(1-t)} - 1\right) \text{ for some } \sigma \geq 1.$$

[1] In section 4 we will indicate how this function has been found in the search for a kernel function that is likely to yield a good complexity result.

It is worth pointing out that all known kernel functions are coercive, i.e., have the properties $\lim_{t\downarrow 0}\psi(t)=\infty$ and $\lim_{t\to\infty}\psi(t)=\infty$. Our new function has the second property, but it fails to have the first property, because

$$\lim_{t\downarrow 0}\psi(t)=\psi(0)=\frac{e^{\sigma}-1}{\sigma}-\frac{1}{2}<\infty.$$

This means that if either x or s approaches the boundary of the feasible region, then $\Phi(x,s;\mu):=2\Psi(v)$ converges to a finite value, depending on the value of σ. This is a striking feature of the new barrier function. Because of this it is quite surprising that we can show in this paper that if σ is chosen appropriately, then the resulting algorithm has an $O(\sqrt{n}\,(\log n)\log\frac{n}{\varepsilon})$ iteration bound, which is at present the best known iteration bound for large-update methods.

In the analysis of the algorithm based on the present kernel function $\psi(t)$ we need its first three derivatives. For ease of reference we give them here. One has

$$\psi'(t)=t-e^{\sigma(1-t)},\quad \psi''(t)=1+\sigma e^{\sigma(1-t)},\quad \psi'''(t)=-\sigma^2 e^{\sigma(1-t)}. \tag{15}$$

As said before, in the analysis of the algorithm the concepts of strong and exponential convexity are crucial ingredients. We deal with these concepts in the next two sections. To make the paper self-supporting, we include the (elementary) proofs of the results in these sections.

2.6. Consequences of strong convexity. Following the usual terminology (see, e.g., [4]), since $\psi''(t)>1$ we say that $\psi(t)$ is strongly convex. In this section we deal with some important consequences of this property.

LEMMA 2.1. *One has*

$$\frac{1}{2}(t-1)^2\le\psi(t)\le\frac{1}{2}\psi'(t)^2,\quad t>0. \tag{16}$$

Proof. Using that $\psi(1)=\psi'(1)=0$, and $\psi''(t)>1$, we may write

$$\psi(t)=\int_1^t\int_1^{\xi}\psi''(\zeta)\,d\zeta\,d\xi\ge\int_1^t\int_1^{\xi}d\zeta\,d\xi=\frac{1}{2}(t-1)^2,$$

which proves the first inequality. The second inequality is obtained as follows:

$$\begin{aligned}\psi(t)&=\int_1^t\int_1^{\xi}\psi''(\zeta)d\zeta d\xi\le\int_1^t\int_1^{\xi}\psi''(\xi)\psi''(\zeta)d\zeta d\xi\\&=\int_1^t\psi''(\xi)\psi'(\xi)d\xi=\int_1^t\psi'(\xi)d\psi'(\xi)=\frac{1}{2}\psi'(t)^2.\end{aligned}$$

This completes the proof. ∎

In the analysis of the algorithm we also use the *norm-based proximity measure* $\delta(v)$ defined by

$$\delta(v):=\frac{1}{2}\|\nabla\Psi(v)\|=\frac{1}{2}\sqrt{\sum_{i=1}^{n}(\psi'(v_i))^2}. \tag{17}$$

Note that since $\Psi(v)$ is strictly convex and minimal at $v=\mathrm{e}$, whereas the minimal value is zero, we have

$$\Psi(v)=0\Leftrightarrow\delta(v)=0\Leftrightarrow v=\mathrm{e}.$$

COROLLARY 2.2. *One has* $\Psi(v) \leq 2\,\delta(v)^2$.

Proof. Using the second inequality in (16) we may write

$$\Psi(v) = \sum_{i=1}^{n} \psi(v_i) \leq \frac{1}{2} \sum_{i=1}^{n} \psi'(v_i)^2 = \frac{1}{2} \|\nabla \Psi(v)\|^2 = 2\,\delta(v)^2,$$

which is the desired inequality. ∎

COROLLARY 2.3. *One has* $\|v\| \leq \sqrt{n} + \sqrt{2\Psi(v)} \leq \sqrt{n} + 2\delta(v)$.

Proof. Using the first inequality in (16) we obtain

$$2\Psi(v) = 2\sum_{i=1}^{n} \psi(v_i) \geq \sum_{i=1}^{n} (v_i - 1)^2 = \|v\|^2 - 2e^T v + n \geq (\|v\| - \|e\|)^2.$$

This implies $\|v\| \leq \|e\| + \sqrt{2\Psi(v)} = \sqrt{n} + \sqrt{2\Psi(v)}$. ∎

2.7. Exponential convexity. Our first lemma in this section is related to Definition 1 and Lemma 1 in [12].

LEMMA 2.4. *Let* $t_1 \geq \frac{1}{\sigma}$ *and* $t_2 \geq \frac{1}{\sigma}$. *Then*

$$\psi\left(\sqrt{t_1 t_2}\right) \leq \frac{1}{2}\left(\psi(t_1) + \psi(t_2)\right). \tag{18}$$

Proof. One may easily verify that the property in the lemma holds if and only if the function $\psi(e^z)$ is convex for $z \geq -\log\sigma$, and this holds if and only if $\psi'(t) + t\psi''(t) \geq 0$, whenever $t \geq \frac{1}{\sigma}$. Using (15), one obtains

$$\psi'(t) + t\psi''(t) = 2t + (\sigma t - 1)\, e^{\sigma(1-t)}. \tag{19}$$

The last expression is positive if $t \geq \frac{1}{\sigma}$. Hence, the lemma follows. ∎

The above proof makes clear that the property (18) is equivalent to convexity of the composed function $\psi(e^z)$ with respect to z. Following [11], we therefore say that $\psi(t)$ is *exponentially convex*, or shortly, e-convex, whenever $t \geq \frac{1}{\sigma}$.

Contrary to the present kernel function, the kernel functions considered in [10, 11, 12, 13, 14] were all exponentially convex on the whole positive axis. Since we are going to use exponential convexity in our analysis we have to ensure that during the course of the algorithm the coordinates of v stay within the region where $\psi(t)$ is exponentially convex. It is obvious that this will certainly hold if $\Psi(v) \leq \psi(\frac{1}{\sigma})$. Because then we have $\psi(v_i) \leq \psi(\frac{1}{\sigma})$ for each i, and this implies that $v_i \geq \frac{1}{\sigma}$ for each i, since $\sigma \geq 1$. We need a stronger result, however, provided by the next lemma, which makes clear that when v belongs to the level set $\{v : \Psi(v) \leq L\}$, for some given $L \geq 8$, then all coordinates of v are larger than or equal to $\frac{3}{2\sigma}$, provided that the value of σ is large enough.

LEMMA 2.5. *Let* $L \geq 8$ *and* $\Psi(v) \leq L$. *If* σ *satisfies* $\sigma \geq 1 + 2\log(1+L)$, *then* $v_1 := \min(v) > \frac{3}{2\sigma}$.

Proof. First note that $\Psi(v) \leq L$ implies $\psi(v_i) \leq L$ for each $i = 1, \ldots, n$. Hence, putting $t = v_1$, we have

$$\frac{t^2 - 1}{2} + \frac{1}{\sigma}\left(e^{\sigma(1-t)} - 1\right) \leq L.$$

It follows that

$$\frac{1}{\sigma}\left(e^{\sigma(1-t)} - 1\right) \leq L + \frac{1 - t^2}{2} \leq L + \frac{1}{2}. \tag{20}$$

Below we show that if $L \geq 8$ and $\sigma \geq 1 + 2\log(1+L)$, then (20) implies $t > \frac{3}{2\sigma}$, which suffices for the proof of the lemma. If $t \geq 1$, then there is nothing to prove, since $\frac{3}{2\sigma} < 1$. Thus we assume that $t < 1$. Since the left-hand side in (20) is monotonically increasing in σ, without loss of generality we may put σ at its smallest value: $\sigma = 1 + 2\log(1+L)$. Then $\log(1+L)^2 = \sigma - 1$, and hence (20) implies

$$\frac{e^{\sigma(1-t)} - 1}{\sigma} = \frac{e^{\sigma-1}e^{1-\sigma t} - 1}{\sigma} = \frac{(1+L)^2 e^{1-\sigma t} - 1}{1 + 2\log(1+L)} \leq L + \frac{1}{2}.$$

Thus we obtain

$$e^{1-\sigma t} \leq \frac{1 + (1 + 2\log(1+L))\left(L + \frac{1}{2}\right)}{(1+L)^2}.$$

The expression at the right-hand side is monotonically decreasing in L. The value at $L = 8$ is $0.6065 < e^{-\frac{1}{2}}$. Thus we obtain that $e^{1-\sigma t} < e^{-\frac{1}{2}}$, which implies $1 - \sigma t < -\frac{1}{2}$, or $t > \frac{3}{2\sigma}$, proving the lemma. □

3. Analysis of the algorithm. We assume that in the generic algorithm the threshold value τ and the barrier update parameter θ are given. Also, we assume that $\tau = O(n)$. With τ and θ given, we start by finding an upper bound L for the values that are attained by $\Psi(v)$ during the course of the algorithm. After having determined such a bound we use Lemma 2.5 to fix a suitable value for σ. This is done in section 3.1. Then we proceed by estimating the decrease in $\Psi(v)$ during an inner iteration. In doing this we use a fixed value for the step size, which depends on σ and the current value of $\delta(v)$. The analysis is completed by first deriving an upper bound for the number of inner iterations between two subsequent updates of the barrier parameter. By multiplying this upper bound with the number of μ-updates we obtain an upper bound for the total number of iterations required by the algorithm.

3.1. Fixing the values of L and σ. Note that at the start of each outer iteration, just before the update of μ with the factor $1-\theta$, we have $\Psi(v) \leq \tau$. Due to the update of μ the vector v is divided by the factor $\sqrt{1-\theta}$, with $0 < \theta < 1$, which in general leads to an increase in the value of $\Psi(v)$. Then, during the subsequent inner iterations, $\Psi(v)$ decreases until it passes the threshold τ again. Hence, during the course of the algorithm the largest values of $\Psi(v)$ occur just after the updates of μ. That is why in this section we derive an estimate for the effect of a μ-update on the value of $\Psi(v)$. We start with a simple lemma.

LEMMA 3.1. *Let $\beta \geq 1$. Then $\psi(\beta t) \leq \psi(t) + \frac{1}{2}\left(\beta^2 - 1\right)t^2$.*

Proof. Defining $\psi_b(t) = \frac{1}{\sigma}\left(e^{\sigma(1-t)} - 1\right)$, we may write

$$\psi(\beta t) = \frac{\beta^2 t^2 - 1}{2} + \psi_b(\beta t) = \psi(t) + \frac{1}{2}\left(\beta^2 t^2 - t^2\right) + \psi_b(\beta t) - \psi_b(t).$$

Since $\psi_b(t)$ is monotonically decreasing in t, $\psi_b(\beta t) - \psi_b(t) \leq 0$. Hence, the inequality in the lemma follows. □

COROLLARY 3.2. *Let $0 \leq \theta < 1$ and $v_+ = \frac{v}{\sqrt{1-\theta}}$. Then*

$$\Psi(v_+) \leq \Psi(v) + \frac{\theta}{2(1-\theta)}\left(2\Psi(v) + \sqrt{8n\Psi(v)} + n\right).$$

Proof. Using Lemma 3.1, with $\beta = 1/\sqrt{1-\theta}$, we write

$$\Psi(\beta v) \leq \sum_{i=1}^{n} \left(\psi(v_i) + \frac{1}{2} \left(\beta^2 - 1\right) v_i^2 \right) = \Psi\left(v\right) + \frac{\theta \left\|v\right\|^2}{2\left(1-\theta\right)}.$$

Since $\|v\| \leq \sqrt{n} + \sqrt{2\Psi\left(v\right)}$, by Corollary 2.3, we get

$$\Psi(\beta v) \leq \Psi\left(v\right) + \frac{\theta}{2\left(1-\theta\right)} \left(2\Psi\left(v\right) + 2\sqrt{2n\Psi\left(v\right)} + n \right).$$

This proves the lemma. □

Due to Corollary 3.2, after each μ-update we have $\Psi(v) \leq L$, where

$$L = \tau + \frac{\theta}{2\left(1-\theta\right)} \left(2\tau + 2\sqrt{2n\tau} + n \right) = O\left(\frac{\theta n}{1-\theta} \right). \tag{21}$$

The above defined value for L satisfies our desires: L is an absolute upper bound for the values that the barrier function $\Psi(v)$ may take during the course of the algorithm. We define σ as follows:

$$\sigma = 1 + 2\log\left(1 + L\right). \tag{22}$$

Note that since $\tau = O(n)$ and $\theta/(1-\theta) = O(1)$, we have

$$L = O\left(n\right), \quad \sigma = O\left(\log n\right). \tag{23}$$

Without loss of much generality we assume that $L \geq 8$.[2] According to Lemma 2.5 we then have

$$\Psi(v) \leq L \quad \Rightarrow \quad v_i \geq \frac{3}{2\sigma} \text{ for all } i. \tag{24}$$

From now on we assume that L and σ are as given by (21) and (22), respectively.

3.2. Determining a step size. In this section we determine a default step size, namely

$$\bar{\alpha} := \frac{1}{1 + \sigma(1 + 4\delta)}, \tag{25}$$

where $\delta = \delta(v)$. We will show below that this step size not only keeps the iterates feasible but also gives rise to a sufficiently large decrease of the barrier function

$$\Phi(x, s; \mu) := \Psi(v) = \sum_{i=1}^{n} \psi(v_i) \tag{26}$$

in each inner iteration.

Apart from the necessary adaptations to the current context and some simplifications, the analysis below follows the same line of arguments that was first used in [12] and subsequently also in [10, 11, 13, 14].

[2] Assuming $\theta \geq \frac{2}{3}$, this already holds if $\tau \geq 1$ and $n \geq 2$.

In each inner iteration we first compute the search directions Δx, Δy, and Δs from

$$\begin{aligned} A\Delta x &= 0, \\ A^T\Delta y + \Delta s &= 0, \\ s\Delta x + x\Delta s &= -\mu v \nabla\Psi(v). \end{aligned} \tag{27}$$

After a step with size α the new iterates are

$$x_+ = x + \alpha\Delta x, \quad y_+ = y + \alpha\Delta y, \quad s_+ = s + \alpha\Delta s.$$

Recall that during an inner iteration the parameter μ is fixed. Hence, after the step the new v-vector is given by

$$v_+ = \sqrt{\frac{x_+ s_+}{\mu}}. \tag{28}$$

Since

$$x_+ = x\left(e + \alpha\frac{\Delta x}{x}\right) = x\left(e + \alpha\frac{d_x}{v}\right) = \frac{x}{v}\left(v + \alpha d_x\right),$$

and, similarly,

$$s_+ = s\left(e + \alpha\frac{\Delta s}{s}\right) = s\left(e + \alpha\frac{d_s}{v}\right) = \frac{s}{v}\left(v + \alpha d_s\right),$$

also using $xs = \mu v^2$, we obtain

$$v_+ = \sqrt{\left(v + \alpha d_x\right)\left(v + \alpha d_s\right)}. \tag{29}$$

We consider the decrease in Ψ as a function of α. Thus we define

$$f(\alpha) := \Psi\left(v_+\right) - \Psi\left(v\right). \tag{30}$$

Our aim is to find an upper bound for $f(\alpha)$ by using exponential convexity, according to Lemma 2.4. In order to do this we assume for the moment that the step size α is such that the coordinates of the vectors $v + \alpha d_x$ and $v + \alpha d_s$ are not smaller than $\frac{1}{\sigma}$, i.e.,

$$v_i + \alpha d_{xi} \geq \frac{1}{\sigma}, \quad v_i + \alpha d_{si} \geq \frac{1}{\sigma}, \quad 1 \leq i \leq n. \tag{31}$$

Then Lemma 2.4 implies that

$$\Psi\left(v_+\right) \leq \frac{1}{2}\left(\Psi\left(v + \alpha d_x\right) + \Psi\left(v + \alpha d_s\right)\right). \tag{32}$$

As a consequence we have $f(\alpha) \leq f_1(\alpha)$, where

$$f_1(\alpha) = \frac{1}{2}\left(\Psi\left(v + \alpha d_x\right) + \Psi\left(v + \alpha d_s\right)\right) - \Psi\left(v\right). \tag{33}$$

Obviously,

$$f(0) = f_1(0) = 0. \tag{34}$$

Taking the derivative to α, we get

$$f_1'(\alpha) = \frac{1}{2}\sum_{i=1}^{n}\left(\psi'\left(v_i+\alpha d_{xi}\right)d_{xi}+\psi'\left(v_i+\alpha d_{si}\right)d_{si}\right). \tag{35}$$

This gives, also using (13),

$$\begin{aligned} f_1'(0) &= \frac{1}{2}\sum_{i=1}^{n}\left(\psi'\left(v_i\right)d_{xi}+\psi'\left(v_i\right)d_{si}\right) = \frac{1}{2}\nabla\Psi\left(v\right)^T\left(d_x+d_s\right) \\ &= -\frac{1}{2}\nabla\Psi\left(v\right)^T\nabla\Psi\left(v\right) = -2\delta(v)^2. \end{aligned} \tag{36}$$

Differentiating once more in (35), we obtain

$$f_1''(\alpha) = \frac{1}{2}\sum_{i=1}^{n}\left(\psi''\left(v_i+\alpha d_{xi}\right)d_{xi}^{\,2}+\psi''\left(v_i+\alpha d_{si}\right)d_{si}^{\,2}\right). \tag{37}$$

Note that this makes clear that $f_1''(\alpha) > 0$, unless $d_x = d_s = 0$. Thus, since during an inner iteration the iterates x and s are not both at the μ-center, we may conclude that $f_1(\alpha)$ is strictly convex as a function of α.[3]

Recall that $f(\alpha) \le f_1(\alpha)$ and $f(0) = f_1(0) = -2\delta(v)^2 < 0$. It may be worth pointing out at this stage that the best value for α is the one that minimizes $f(\alpha)$. The idea underlying our approach is that the step size that minimizes $f_1(\alpha)$ will be good enough for our purpose. Thus we want to find α^* such that $f_1'(\alpha^*) = 0$. Since $f_1'(\alpha)$ is strictly convex, we have

$$\alpha^* = \max\left\{\alpha \;:\; f_1'(\alpha) \le 0\right\}. \tag{38}$$

The default step size that we are going to use will satisfy $f_1'(\alpha) \le 0$, and as a consequence also $\alpha \le \alpha^*$. This has as an important consequence that our step size will certainly be feasible.

In order to get our default step size, we proceed by deriving an upper bound for the expression at the right in (37). First, letting

$$v_1 := v_{\min}, \qquad \delta := \delta(v), \tag{39}$$

we obtain the inequalities

$$\begin{aligned} v_i+\alpha d_{xi} &\ge v_1-\alpha\left\|d_x\right\| \ge v_1-2\alpha\delta, \\ v_i+\alpha d_{si} &\ge v_1-\alpha\left\|d_s\right\| \ge v_1-2\alpha\delta. \end{aligned} \tag{40}$$

Here we used that $\|(d_x,d_s)\| = 2\delta$, which follows from (13), (17), and the fact that d_x and d_s are orthogonal. Recall from (15) that $\psi'''(t) = -\sigma^2 e^{\sigma(1-t)} < 0$. Hence, $\psi''(t)$ is monotonically decreasing. Therefore, (37) implies that

$$\begin{aligned} f_1''(\alpha) &\le \frac{1}{2}\psi''\left(v_1-2\alpha\delta\right)\sum_{i=1}^{n}\left(d_{xi}^{\,2}+d_{si}^{\,2}\right) \\ &= \frac{1}{2}\psi''\left(v_1-2\alpha\delta\right)\left\|(d_x,d_s)\right\|^2 = 2\delta^2\psi''\left(v_1-2\alpha\delta\right). \end{aligned} \tag{41}$$

[3]Note that $f(\alpha)$ is not necessarily convex as a function of α; that is why the analysis is greatly simplified through the use of the auxiliary function $f_1(\alpha)$, and to justify this we needed exponential convexity!

By integrating we derive from this that

$$f_1'(\alpha) = f_1'(0) + \int_0^\alpha f_1''(\xi)\, d\xi \le -2\delta^2 + \int_0^\alpha 2\delta^2\, \psi''\left(v_1 - 2\xi\delta\right)\, d\xi$$
$$= -2\delta^2 - \delta\left(\psi'\left(v_1 - 2\alpha\delta\right) - \psi'\left(v_1\right)\right).$$

Hence, $f_1'(\alpha) \le 0$ will certainly hold if α satisfies

$$-\psi'\left(v_1 - 2\alpha\delta\right) + \psi'\left(v_1\right) \le 2\delta. \tag{42}$$

Any α satisfying this inequality will also satisfy $\alpha \le \alpha^*$, and hence is a feasible step size. Of course, we want α to be as large as possible. Thus our next task is to find the largest α that satisfies (42). Obviously, this α depends on δ and v_1. Below we use the following strategy: given $\delta > 0$ we find the value of v_1 for which the largest possible step size α is minimal. For that value of v_1 the largest solution of (42) becomes a function of δ alone; this function can be found explicitly as we show below.

Since $\psi''(t)$ is decreasing, the derivative to v_1 of the left-hand side in (42) (i.e., $-\psi''\left(v_1 - 2\alpha\delta\right) + \psi''\left(v_1\right)$) is negative. Hence, with δ fixed, the smaller v_1 is, the smaller the maximal step size α will be. Note that one has

$$\delta = \frac{1}{2}\left\|\nabla\Psi\left(v\right)\right\| \ge \frac{1}{2}\left|\psi'\left(v_1\right)\right| \ge -\frac{1}{2}\psi'\left(v_1\right),$$

and that equality holds throughout if and only if v_1 is the only coordinate in v that differs from 1, and $v_1 \le 1$ (in which case $\psi'\left(v_1\right) \le 0$). Hence, the worst situation for the step size occurs when v_1 satisfies

$$-\frac{1}{2}\psi'\left(v_1\right) = \delta. \tag{43}$$

In that case the largest α satisfying (42) is minimal. For our purpose we need to deal with the worst case, so we will assume (43). Then inequality (42) reduces to

$$-\frac{1}{2}\psi'\left(v_1 - 2\alpha\delta\right) \le 2\delta. \tag{44}$$

The function $-\frac{1}{2}\psi'(t)$ maps the interval $[0,1]$ to the interval $[0, -\frac{1}{2}\psi'(0)]$ and is monotonically decreasing on this interval. Hence, the inverse function $\rho : [0, -\frac{1}{2}\psi'(0)] \to [0,1]$ exists and is monotonically decreasing as well. Therefore, (44) is equivalent to $v_1 - 2\alpha\delta \ge \rho\left(2\delta\right)$. Since, by the definition of ρ and (43), $v_1 = \rho(\delta)$, substitution yields $\rho(\delta) - 2\alpha\delta \ge \rho\left(2\delta\right)$. Thus we obtain that in the worst case the maximal step size that solves (42) is given by

$$\alpha = \frac{\rho\left(\delta\right) - \rho\left(2\delta\right)}{2\delta}. \tag{45}$$

LEMMA 3.3. *With α as given by* (45)*, one has*

$$\alpha \ge \tilde{\alpha} := \frac{1}{\psi''\left(\rho\left(2\delta\right)\right)}. \tag{46}$$

Proof. By the definition of ρ we have $-\psi'(\rho(\delta)) = 2\delta$. Taking derivatives to δ at both sides we get

$$\rho'(\delta) = -\frac{2}{\psi''\left(\rho(\delta)\right)}.$$

Thus we may rewrite (45) as follows:

$$\alpha = \frac{\rho(\delta) - \rho(2\delta)}{2\delta} = \frac{1}{2\delta}\int_{2\delta}^{\delta} \rho'(\xi)\, d\xi = \frac{1}{\delta}\int_{\delta}^{2\delta} \frac{d\xi}{\psi''(\rho(\xi))}.$$

To obtain a lower bound for α, we replace the argument of the last integral by its minimal value. Since ψ'' is monotonically decreasing, $\psi''(\rho(\xi))$ is maximal for $\xi \in [\delta, 2\delta]$ when $\rho(\xi)$ is minimal. Since ρ is monotonically decreasing this occurs when $\xi = 2\delta$. Therefore

$$\alpha = \frac{1}{\delta}\int_{\delta}^{2\delta} \frac{d\xi}{\psi''(\rho(\xi))} \geq \frac{1}{\delta}\frac{\delta}{\psi''(\rho(2\delta))} = \frac{1}{\psi''(\rho(2\delta))},$$

proving the lemma. □

We proceed by deriving a lower bound for the step size $\tilde{\alpha}$, as given by (46). Due to the definition of ρ we may write

$$\tilde{\alpha} = \frac{1}{\psi''(t)}, \quad \text{where } t \text{ is such that} \quad -\psi'(t) = 4\delta.$$

In other words, by (15),

$$\tilde{\alpha} = \frac{1}{1 + \sigma e^{\sigma(1-t)}}, \qquad e^{\sigma(1-t)} - t = 4\delta.$$

From the second expression we derive that

$$e^{\sigma(1-t)} = 4\delta + t \leq 4\delta + 1.$$

Substituting this in the first expression we obtain $\tilde{\alpha} \geq \bar{\alpha}$, where $\bar{\alpha}$ is our default step size as given by (25).

Finally, to validate the above analysis we need to show that $\alpha = \bar{\alpha}$ satisfies (31). Due to (40) it suffices to show that $v_1 - 2\bar{\alpha}\delta \geq \frac{1}{\sigma}$. This is now easy. Using the definition (25) of $\bar{\alpha}$ and $v_1 \geq \frac{3}{2\sigma}$, which is due to (24), we may write

$$v_1 - 2\bar{\alpha}\delta \geq \frac{3}{2\sigma} - \frac{2\delta}{1 + \sigma(1 + 4\delta)} \geq \frac{3}{2\sigma} - \frac{1}{2\sigma} = \frac{1}{\sigma}.$$

3.3. Decrease of the barrier function during an inner iteration. Now that the step size has been determined, the resulting decrease in the barrier function value can be easily established by using the following result; for its (elementary) proof we refer to [12, Lemma 12].

LEMMA 3.4. *Let $h(t)$ be a (univariate) twice differentiable convex function with $h(0) = 0$, $h'(0) < 0$ and let $h(t)$ attain its (global) minimum at $t^* > 0$. If $h''(t)$ is increasing for $t \in [0, t^*]$, then*

$$h(t) \leq \frac{t h'(0)}{2}, \quad 0 \leq t \leq t^*.$$

Using this lemma and (36) and (25) we obtain

$$f(\bar{\alpha}) \leq f_1(\bar{\alpha}) \leq \frac{\bar{\alpha} f_1'(0)}{2} = -\bar{\alpha}\,\delta^2 \leq -\frac{\delta^2}{1 + \sigma(1 + 4\delta)}.$$

This expresses the decrease in one inner iteration in terms of $\delta = \delta(v)$. By Corollary 2.2, we have

$$\delta(v) \geq \sqrt{\frac{\Psi(v)}{2}}.$$

Since the decrease depends monotonically on δ, we may express the decrease in terms of $\Psi = \Psi(v)$ as follows:

$$f(\bar{\alpha}) \leq -\frac{\Psi(v)}{2\left(1+\sigma\left(1+2\sqrt{2\Psi(v)}\right)\right)}. \tag{47}$$

It will be convenient to write instead

$$f(\bar{\alpha}) \leq -\frac{\kappa}{\sigma}\sqrt{\Psi(v)} \tag{48}$$

for some absolute constant $\kappa > 0$ (e.g., $\kappa = \frac{1}{16}$).

3.4. Inner-iteration bound. We need to count how many inner iterations are required to return to the situation where $\Psi(v) \leq \tau$ after a μ-update. We denote the value of $\Psi(v)$ after the μ-update as Ψ_0; the subsequent values in the same outer iteration are denoted as Ψ_k, $k = 1, 2, \ldots$. If K denotes the total number of inner iterations in the outer iteration, we then have

$$\Psi_0 = O\left(\frac{\theta n}{1-\theta}\right), \quad \Psi_{K-1} > \tau, \quad 0 \leq \Psi_K \leq \tau, \tag{49}$$

and, according to (48),

$$\Psi_{k+1} \leq \Psi_k - \frac{\kappa}{\sigma}\left(\Psi_k\right)^{\frac{1}{2}}, \qquad k = 0, 1, \ldots, K-1. \tag{50}$$

At this stage we invoke the following lemma from [12, Lemma 14] without proof.

LEMMA 3.5. *Let $t_0, t_1, \ldots, t_K$ be a sequence of positive numbers such that*

$$t_{k+1} \leq t_k - \lambda t_k^{1-\gamma}, \qquad k = 0, 1, \ldots, K-1, \tag{51}$$

where $\lambda > 0$ and $0 < \gamma \leq 1$. Then $K \leq \left\lfloor \frac{t_0^{\gamma}}{\lambda\gamma} \right\rfloor$.

LEMMA 3.6. *One has*

$$K = O\left(\sigma\sqrt{\frac{\theta n}{1-\theta}}\right). \tag{52}$$

Proof. We apply Lemma 3.5, with $t_k = \Psi_k$, $\lambda = \frac{\kappa}{\sigma}$, and $\gamma = \frac{1}{2}$. This yields

$$K \leq \frac{2\sigma\Psi_0^{\frac{1}{2}}}{\kappa} = O\left(\sigma\sqrt{\frac{\theta n}{1-\theta}}\right),$$

proving the lemma. ∎

3.5. Iteration bound. We just found that the number of inner iterations needed to recenter is given by (52). The number of outer iterations is bounded above by (cf. [17, Lemma II.17])

$$\frac{1}{\theta}\log\frac{n}{\varepsilon}. \tag{53}$$

By multiplying these two numbers we get an upper bound for the total number of iterations, namely

$$O\left(\sigma\sqrt{\frac{n}{\theta\,(1-\theta)}}\right)\log\frac{n}{\varepsilon}.$$

In large-update methods we have $\theta/(1-\theta)=\Theta(1)$. Since $\sigma=O(\log n)$, by (23), the iteration bound then becomes

$$O\left(\sigma\sqrt{n}\,\log\frac{n}{\varepsilon}\right)=O\left(\sqrt{n}\,(\log n)\log\frac{n}{\varepsilon}\right).$$

4. Concluding remarks. Let us start by indicating how the barrier function proposed in this paper has been found. Recall that the default step size $\bar{\alpha}$ used in the paper is obtained from the step size $\tilde{\alpha}$ given in Lemma 3.3:

$$\tilde{\alpha}=\frac{1}{\psi''\left(\rho\left(2\delta\right)\right)}, \tag{54}$$

where ρ is the inverse function of $-\frac{1}{2}\psi'$. Note that this step size essentially depends only on the kernel function $\psi(t)$ and on δ. From the analysis in the paper it may be clear that we may expect an $O\left(\sqrt{n}\log\frac{n}{\varepsilon}\right)$ iteration bound only if $\tilde{\alpha}=\Theta(1/\delta)$. Thus it is natural to ask if there exists a kernel function $\psi(t)$ with this property. The desired property certainly holds if $\psi(t)$ is such that there exist positive a and b such that, for all $\delta\geq 0$,

$$\psi''\left(\rho(2\delta)\right)=a+4b\delta,\qquad -\psi'\left(\rho(\delta)\right)=2\delta. \tag{55}$$

The second equation says that ρ is the inverse function of $-\frac{1}{2}\psi'$, and the first equation says that the default step size (54) is of the order $1/\delta$. Some straightforward calculations lead to the conclusion that the system (55), with the additional requirements $\psi(1)=\psi'(1)=0$, admits only one solution, namely the function

$$\psi(t)=\frac{a}{b}\left(t-1+\frac{1}{b}\left(e^{b(1-t)}-1\right)\right). \tag{56}$$

At first sight the outcome of this analysis is quite disappointing, for at least three reasons: the above function ψ is not a barrier function in the usual sense, it is not strongly convex, and it is not exponentially convex.

Note that the kernel function of this paper arises from (56) by taking $a=b=\sigma$ and by replacing the first linear term $t-1$ by $\frac{1}{2}(t^2-1)$. Although the resulting barrier function is not a barrier function in the usual sense, we have been able to show in this paper that the iteration bound of the corresponding large-update method coincides with the currently best iteration bound for large-update methods. It remains a challenge for future research to analyze a primal-dual method based on the kernel function (56); since this function is not strongly convex, such an analysis will require completely new techniques.

At present no computational results exist for the method presented in this paper. This will be another issue for future research. The extensions to semidefinite optimization and second order cone optimization also deserve to be investigated.

Acknowledgments. The authors kindly acknowledge the assistance of an anonymous referee and the editor in indicating a shortcoming in the derivation of inequality (32) in an earlier version of this paper. The first author acknowledges the hospitality of the Technical University Delft during her visit from March 2001 to March 2002.

REFERENCES

[1] E.D. Andersen, J. Gondzio, Cs. Mészáros, and X. Xu, *Implementation of interior point methods for large scale linear programming*, in Interior Point Methods of Mathematical Programming, T. Terlaky, ed., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996, pp.189–252.

[2] Y.Q. Bai, C. Roos, and M. El ghami, *A primal-dual interior-point algorithm based on an exponential barrier*, Optim. Methods Softw., to appear.

[3] R. Frisch, *The logarithmic potential method for solving linear programming problems*, memorandum, University Institute of Economics, Oslo, Norway, 1955.

[4] J.-P. Hiriart-Urrity and C. Lemaréchal, *Convex Analysis and Minimization Algorithms* I, Springer-Verlag, Berlin, 1996.

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

[6] 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.

[7] S. Mehrotra and Y. Ye, *On finding the optimal facet of linear programs*, Math. Program., 62 (1993), pp. 497–515.

[8] J. Peng, C. Roos, and T. Terlaky, *A new class of polynomial primal-dual methods for linear and semidefinite optimization*, European J. Oper. Res., 143 (2002), pp. 234–256.

[9] J. Peng, C. Roos, and T. Terlaky, *New complexity analysis of the primal-dual Newton method for linear optimization*, Ann. Oper. Res., 99 (2000), pp. 23–39.

[10] J. Peng, C. Roos, and T. Terlaky, *Primal-dual interior-point methods for second-order conic optimization based on self-regular proximities*, SIAM J. Optim., 13 (2002), pp. 179–203.

[11] J. Peng, C. Roos, and T. Terlaky, *A new and efficient large-update interior-point method for linear optimization*, J. Comput. Tech., 6 (2001), pp. 61–80.

[12] J. Peng, C. Roos, and T. Terlaky, *Self-regular functions and new search directions for linear and semidefinite optimization*, Math. Program., 93 (2002), pp. 129–171.

[13] J. Peng, C. Roos, and T. Terlaky, *Self-Regularity: A New Paradigm for Primal-Dual Interior-Point Algorithms*, Princeton University Press, Princeton, NJ, 2002.

[14] J. Peng, C. Roos, T. Terlaky, and A. Yoshise, *Self-regular proximities and new directions for nonlinear $P_*(\kappa)$ complementarity problems*, Math. Oper. Res., submitted.

[15] J. Renegar, *A Mathematical View of Interior-Point Methds in Convex Optimization*, MPS/SIAM Ser. Optim., SIAM, Philadelphia, 2001.

[16] C. Roos, *A Comparative Study of Barrier Functions for Primal-Dual Interior-Point Algorithms in Linear Optimization*, manuscript, 2001.

[17] 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.

[18] G. Sonnevend, *An "analytic center" for polyhedrons and new classes of global algorithms for linear (smooth, convex) programming*, in System Modelling and Optimization, 1985, A. Prékopa, J. Szelezsán, and B. Strazicky, eds., Lecture Notes in Control and Inform. Sci. 84, Springer-Verlag, Berlin, West Germany, 1986, pp. 866–876.

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

[20] Y. Ye, *On the finite convergence of interior-point algorithms for linear programming*, Math. Program., 57 (1992), pp. 325–335.

[21] Y. Ye, *Interior Point Algorithms, Theory and Analysis*, John Wiley and Sons, Chichester, UK, 1997.

SIAM J. OPTIM.
Vol. 13, No. 3, pp. 783–804

INTERIOR-POINT METHODS FOR MASSIVE SUPPORT VECTOR MACHINES*

MICHAEL C. FERRIS† AND TODD S. MUNSON‡

Abstract. We investigate the use of interior-point methods for solving quadratic programming problems with a small number of linear constraints, where the quadratic term consists of a low-rank update to a positive semidefinite matrix. Several formulations of the support vector machine fit into this category. An interesting feature of these particular problems is the volume of data, which can lead to quadratic programs with between 10 and 100 million variables and, if written explicitly, a dense Q matrix. Our code is based on OOQP, an object-oriented interior-point code, with the linear algebra specialized for the support vector machine application. For the targeted massive problems, all of the data is stored out of core and we overlap computation and input/output to reduce overhead. Results are reported for several linear support vector machine formulations demonstrating that the method is reliable and scalable.

Key words. support vector machine, interior-point method, linear algebra

AMS subject classifications. 90C51, 90C20, 62H30

PII. S1052623400374379

1. Introduction.

Interior-point methods [30] are frequently used to solve large convex quadratic and linear programs for two reasons. First, the number of iterations taken is typically either constant or grows very slowly with the problem dimension. Second, the major computation involves solving (one or) two systems of linear equations per iteration, for which many efficient, large-scale algorithms exist. Thus, interior-point methods become more attractive as the size of the problem increases. General-purpose implementations of these methods can be complex, relying upon sophisticated sparse techniques to factor the relevant matrix at each iteration. However, the basic algorithm is straightforward and can be used in a wide variety of problems by simply tailoring the linear algebra to the application.

We are particularly interested in applying an interior-point method to a class of quadratic programs with two properties: each model contains a small number of linear constraints, and the quadratic term consists of a (dense) low-rank update to a positive semidefinite matrix. The key to solving these problems is to exploit structure using block eliminations. One source of massive problems of this type is the data mining community, where several linear support vector machine (SVM) formulations [28, 1, 2, 19] fit into the framework. A related example is the Huber regression problem [17, 21, 31], which can also be posed as a quadratic program of the type considered.

The linear SVM attempts to construct a hyperplane partitioning two sets of observations, where each observation is an element of a low-dimensional space. An

*Received by the editors May 24, 2000; accepted for publication (in revised form) March 21, 2002; published electronically January 3, 2003. This material is based on research supported by National Science Foundation grants CCR-9972372 and CDA-9726385; Air Force Office of Scientific Research grant F49620-01-1-0417; Microsoft Corporation; and the Mathematical, Information, and Computational Sciences Division subprogram of the Office of Advanced Scientific Computing, U.S. Department of Energy, under contract W-31-109-Eng-38.

http://www.siam.org/journals/siopt/13-3/37437.html

†Computer Sciences Department, University of Wisconsin, Madison, WI 53706 (ferris@cs.wisc.edu).

‡Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439 (tmunson@mcs.anl.gov).

interesting characteristic of these models is the volume of data, which can lead to quadratic programs with between 10 and 100 million variables and, if written explicitly, a dense Q matrix. The large number of practical applications of the SVM [6, 26] is indicative of the importance of robust, scalable algorithms to the data mining and machine learning communities.

Sampling techniques [3] can be used to decrease the number of observations needed to construct a good separating surface. However, if we considered a "global" application and randomly sampled only 1% of the current world population, we would generate a problem with around 60 million observations. Recent work [10] has shown that although a random sampling of 20–30% is sufficient for many applications, sampling even as high as 70–80% can produce statistically significant differences in the models. Furthermore, for comparative purposes, a researcher might wish to solve the nonsampled problem to validate the choice of sampling technique.

Solving realistic, large-scale models of this form raises important research issues. In particular, codes targeting massive problems need to handle the required data volume effectively. For example, one dense vector with 50 million double-precision elements requires 400 megabytes of storage. If all data were to be kept in core, we would rapidly exhaust the memory resources of most machines available today. Therefore, we store *all* data out of core and overlap computation and input/output (I/O) to reduce the overhead inherent in such a scheme.

As mentioned above, the crucial implementation details are in the linear algebra calculation. Rather than reimplement a standard predictor-corrector interior-point code [23], we use OOQP [11, 12] as the basis for our work. A key property of OOQP is the object-oriented design, which enables us to tailor the required linear algebra to the application. Our linear algebra implementation exploits problem structure while keeping all of the data out of core. A proximal-point modification [25] to the underlying algorithm is also available to improve robustness on some of the SVM formulations considered.

We begin in section 2 by formally stating the general optimization problem we are interested in solving, and we show specializations of the framework for linear SVMs and Huber regression. In section 3, we describe the interior-point method and linear algebra requirements. The basic proximal-point idea is discussed, and we demonstrate the use of block eliminations to exploit problem structure. The implementation of the linear algebra using out of core computations is presented in section 4, along with some numerical considerations for massive problems. In section 5, we present experimental results for several linear SVM formulations on two large, randomly generated data sets. These results indicate that the method is reliable and scalable to massive problems. In section 6, we summarize our work and briefly outline future efforts.

2. Quadratic programming framework. The general optimization problem we consider has a quadratic term consisting of a low-rank update to a positive semidefinite matrix and a small number of linear constraints. In particular, the problems discussed have m variables, n constraints, and a rank-k update. Let $Q \in \Re^{m\times m}$ be of the form

$$Q = S + RHR^T,$$

where $S \in \Re^{m\times m}$ is symmetric positive semidefinite, $H \in \Re^{k\times k}$ is symmetric positive definite, and $R \in \Re^{m\times k}$. Typically, S is a very large matrix while H is small. We are

concerned with solving the convex problem

$$\begin{array}{ll} \min_x & \frac{1}{2}x^T Q x + c^T x \\ \text{s.t.} & Bx = b, \\ & \ell \le x \le u \end{array} \tag{2.1}$$

for given $B \in \Re^{n\times m}$ with full row rank, $b \in \Re^n$, $c \in \Re^m$, and general bounds, $\ell \in \Re^m \cup \{-\infty\}^m$ and $u \in \Re^m \cup \{+\infty\}^m$ with $\ell < u$. We assume that $k + n \ll m$. That is, the rank of the update and the number of constraints must be small in relation to the overall size of the problem.

To solve instances of this problem, we exploit structure in the matrices generated by an interior-point algorithm using block eliminations. The underlying operations are carried out in the context of the machine learning applications outlined below. As will become evident, in addition to the assumptions made concerning the form of the quadratic program, we also require that the matrices H and $S + T$ can be inverted easily for any positive diagonal matrix T. These assumptions are satisfied in our applications because H and S are diagonal matrices. However, general cases satisfying these criteria clearly exist.

2.1. Linear SVMs. The linear SVM attempts to construct a hyperplane $\{x \mid w^T x = \gamma\}$ correctly separating two point sets with a maximal separation margin. Several quadratic programming formulations exist in the data mining literature [28, 1, 2, 19] for these problems, which are becoming increasingly important because of the large number of practical applications [6, 26]. The common variation among the optimization models is in the choice of the subset of the variables (w and γ) selected to measure the separation margin and the norm used for the misclassification error.

We first introduce some notation chosen to be consistent with that typically used in the data mining literature. We let $A \in \Re^{m\times k}$ be a (typically dense) matrix representing a set of observations drawn from two sample populations, where m is the total number of observations and k the number of features measured for each observation, with $k \ll m$. Typically, the observation matrix A is scaled so that $\|A\|_\infty \approx k$. Let $D \in \Re^{m\times m}$ be a diagonal matrix defined as

$$D_{i,i} := \begin{cases} +1 & \text{if } i \in P_+, \\ -1 & \text{if } i \in P_-, \end{cases}$$

where P_+ and P_- are the indices of the elements in the two populations. We use the notation e to represent a vector of all ones of the appropriate dimension.

The standard SVM [28, 6] is the following optimization problem:

$$\begin{array}{ll} \min_{w,\gamma,y} & \frac{1}{2}\|w\|_2^2 + \nu e^T y \\ \text{subject to} & D(Aw - e\gamma) + y \ge e, \\ & y \ge 0. \end{array} \tag{2.2}$$

The essential idea is to minimize a weighted sum of the one-norm of the misclassification error, $e^T y$, and the two-norm of w, the normal to the hyperplane being derived. The relationship between minimizing $\|w\|_2$ and maximizing the margin of separation is described, for example, in [22]. Here, ν is a parameter weighting the two compet-

ing goals related to misclassification error and margin of separation. The inequality constraints implement the misclassification error.

Various modifications of (2.2) are developed in the literature. The motivation for many of them is typically to improve the tractability of the problem and to allow novel reformulations in the solution phase. For example, one formulation incorporates γ into the objective function:

$$\begin{array}{ll} \min_{w,\gamma,y} & \frac{1}{2}\|w,\gamma\|_2^2 + \nu e^T y \\ \text{subject to} & D(Aw - e\gamma) + y \geq e, \\ & y \geq 0. \end{array} \tag{2.3}$$

This formulation is described in [20] to allow successive overrelaxation to be applied to the (dual) problem.

A different permutation replaces the one-norm of y in (2.3) with the two-norm, such that the nonnegativity constraint on y becomes redundant. The resulting problem, first introduced in [22], is then

$$\begin{array}{ll} \min_{w,\gamma,y} & \frac{1}{2}\|w,\gamma\|_2^2 + \frac{\nu}{2}\|y\|_2^2 \\ \text{subject to} & D(Aw - e\gamma) + y \geq e. \end{array} \tag{2.4}$$

An active set method on the Wolfe dual of (2.4) is proposed in [22] to calculate a solution. Concurrent with work described here, Mangasarian and Musicant advocated the use of the Sherman–Morrison–Woodbury update formula in their active set algorithm.

Another variant considered [5] is a slight modification of (2.4):

$$\begin{array}{ll} \min_{w,\gamma,y} & \frac{1}{2}\|w\|_2^2 + \frac{\nu}{2}\|y\|_2^2 \\ \text{subject to} & D(Aw - e\gamma) + y \geq e. \end{array} \tag{2.5}$$

We can also use a one-sided Huber M-estimator [16] for the misclassification error within the linear SVM. This function is a convex quadratic for small values of its argument and is linear for large values. The resulting quadratic program is a combination of (2.2) and (2.5),

$$\begin{array}{ll} \min_{w,\gamma,y,t} & \frac{1}{2}\|w\|_2^2 + \frac{\nu_1}{2}\|t\|_2^2 + \nu_2 e^T y \\ \text{subject to} & D(Aw - e\gamma) + t + y \geq e, \\ & y \geq 0, \end{array} \tag{2.6}$$

where ν_1 and ν_2 are two parameters. We note that $\frac{\nu_2}{\nu_1}$ is the switching point between the quadratic and linear error terms. When $\nu_1 \to \infty$ or $\nu_2 \to \infty$, we recover (2.2) or (2.5), respectively. A similar unification of (2.3) and (2.4) can be made by incorporating γ into the objective function of (2.6). For completeness, this problem is

$$\begin{array}{ll} \min_{w,\gamma,y,t} & \frac{1}{2}\|w,\gamma\|_2^2 + \frac{\nu_1}{2}\|t\|_2^2 + \nu_2 e^T y \\ \text{subject to} & D(Aw - e\gamma) + t + y \geq e, \\ & y \geq 0. \end{array} \tag{2.7}$$

As stated, these problems are not in a form matching (2.1). However, the Wolfe duals [18] of (2.2)–(2.7) are, respectively,

$$\begin{array}{ll} \min_x & \frac{1}{2}x^T DAA^T Dx - e^T x \\ \text{subject to} & e^T Dx = 0, \\ & 0 \le x \le \nu e, \end{array} \tag{2.8}$$

$$\begin{array}{ll} \min_x & \frac{1}{2}x^T DAA^T Dx + \frac{1}{2}x^T Dee^T Dx - e^T x \\ \text{subject to} & 0 \le x \le \nu e, \end{array} \tag{2.9}$$

$$\begin{array}{ll} \min_x & \frac{1}{2\nu}x^T x + \frac{1}{2}x^T DAA^T Dx + \frac{1}{2}x^T Dee^T Dx - e^T x \\ \text{subject to} & x \ge 0, \end{array} \tag{2.10}$$

$$\begin{array}{ll} \min_x & \frac{1}{2\nu}x^T x + \frac{1}{2}x^T DAA^T Dx - e^T x \\ \text{subject to} & e^T Dx = 0, \\ & x \ge 0, \end{array} \tag{2.11}$$

$$\begin{array}{ll} \min_x & \frac{1}{2\nu_1}x^T x + \frac{1}{2}x^T DAA^T Dx - e^T x \\ \text{subject to} & e^T Dx = 0, \\ & 0 \le x \le \nu_2 e, \end{array} \tag{2.12}$$

$$\begin{array}{ll} \min_x & \frac{1}{2\nu_1}x^T x + \frac{1}{2}x^T DAA^T Dx + \frac{1}{2}x^T Dee^T Dx - e^T x \\ \text{subject to} & 0 \le x \le \nu_2 e, \end{array} \tag{2.13}$$

which are of the desired form. In addition to the papers cited above, several specialized codes have been applied to solve (2.8); for example, see [24]. Once the dual problems above are solved, the hyperplane in the primal problems can be recovered as follows:

- $w = A^T Dx$, and γ is the multiplier on $e^T Dx = 0$ for (2.2), (2.5), and (2.6).
- $w = A^T Dx$, and $\gamma = -e^T Dx$ for (2.3), (2.4), and (2.7).

Clearly, (2.8)–(2.13) are in the class of problems considered. Rather than become embroiled in a debate over the various formulations, we show that our method can be successfully applied to any of them, and we leave the relative merits of each to be discussed by application experts in the machine learning field.

2.2. Huber regression. A problem related to the SVM is to determine a Huber M-estimator, as discussed in [17, 21, 28, 31]. For an inconsistent system of equations, $Aw = b$, an error residual is typically minimized, namely, $\sum_{i=1}^m \rho((Aw-b)_i)$. In order to deemphasize outliers and avoid nondifferentiability when $\rho(\cdot) = |\cdot|$, the Huber M-estimator [16] has been used.

The corresponding optimization problem is a convex quadratic program,

$$\begin{array}{ll} \min_{w,y,t} & \frac{1}{2}\|t\|_2^2 + \nu e^T y \\ \text{subject to} & -y \le Aw - b - t \le y, \end{array}$$

whose dual has the form

$$\begin{array}{ll} \min_x & \frac{\nu}{2}\|x\|_2^2 + b^T x \\ \text{subject to} & A^T x = 0, \\ & -e \le x \le e. \end{array}$$

The dual has the structure considered whenever the number of observations m is enormous and the number of features k is small. The aforementioned references indicate how to recover a primal solution from the dual.

3. Interior-point method. Since (2.1) is a convex quadratic program, the Karush–Kuhn–Tucker first-order optimality conditions [18] are both necessary and sufficient. These optimality conditions can be written as the mixed complementarity problem

$$\begin{bmatrix} S+RHR^T & -B^T & -I & I \\ B & 0 & 0 & 0 \\ I & 0 & 0 & 0 \\ -I & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ \lambda \\ w \\ v \end{bmatrix} + \begin{bmatrix} c \\ -b \\ -\ell \\ u \end{bmatrix} \perp \begin{array}{l} x \text{ free}, \\ \lambda \text{ free}, \\ w \geq 0, \\ v \geq 0, \end{array} \tag{3.1}$$

where we have augmented the system with slack variables to handle general lower and upper bounds. The $\perp$ notation is defined componentwise by using

- $a \perp b \geq 0$ if and only if $a \geq 0$, $b \geq 0$, and $ab = 0$,
- $a \perp b$ free if and only if $a = 0$.

If the lower or upper bounds are infinite, then the corresponding w and v variables are removed from the problem. The reason for adding slack variables is to eliminate the bounds on x and make the initial starting-point calculation easy.

The basic idea of an interior-point method for (3.1) is to solve the equivalent nonlinear system of equations

$$\begin{aligned} (S+RHR^T)x - B^T\lambda - w + v &= -c, \\ Bx &= b, \\ x - \ell &= y, \\ u - x &= z, \\ Wy &= 0, \\ Vz &= 0, \end{aligned} \tag{3.2}$$

with $w \geq 0$, $v \geq 0$, $y \geq 0$, and $z \geq 0$, where W and V are the diagonal matrices formed from w and v. Furthermore, y and z represent the variables complementary to w and v. Convergence results for these methods can be found in [30] and are not discussed here. Specializations of the interior-point method to the SVM case can be found in [27].

The Mehrotra predictor-corrector method [23] is a specific type of interior-point method. The iterates for the algorithm are guaranteed to remain interior to the simple bounds; that is, $w_i > 0$, $v_i > 0$, $y_i > 0$, and $z_i > 0$ for each iteration i. During the predictor phase, we calculate the Newton direction for (3.2), while the corrector moves the iterate closer to the central path. The direction $(\Delta x, \Delta\lambda, \Delta w, \Delta v, \Delta y, \Delta z)$ is calculated by solving the linearization

$$\begin{bmatrix} S+RHR^T & -B^T & -I & I & 0 & 0 \\ B & 0 & 0 & 0 & 0 & 0 \\ I & 0 & 0 & 0 & -I & 0 \\ -I & 0 & 0 & 0 & 0 & -I \\ 0 & 0 & Y_i & 0 & W_i & 0 \\ 0 & 0 & 0 & Z_i & 0 & V_i \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta\lambda \\ \Delta w \\ \Delta v \\ \Delta y \\ \Delta z \end{bmatrix}$$
$$= \begin{bmatrix} -c - (S+RHR^T)x_i + B^T\lambda_i + w_i - v_i \\ b - Bx_i \\ \ell - x_i + y_i \\ -u + x_i + z_i \\ -W_iy_i + \sigma\frac{(w_i)^Ty_i + (v_i)^Tz_i}{2m}e \\ -V_iz_i + \sigma\frac{(w_i)^Ty_i + (v_i)^Tz_i}{2m}e \end{bmatrix},$$

where $\sigma \in [0,1]$ is a chosen parameter. Different choices for σ give rise to the predictor and the corrector steps, respectively. Since these two systems just differ in the right-hand sides, in the following we simplify our presentation by omitting the details of these vectors and replacing them with generic terms r.

The particular structure of the above linearization allows us to eliminate the variables Δw, Δv, Δy, and Δz from the system. Thus, at each iteration of the algorithm, we solve two systems of linear equations of the form

$$\begin{bmatrix} C + RHR^T & -B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta \lambda \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \end{bmatrix}, \tag{3.3}$$

where

$$C := S + Y_i^{-1} W_i + Z_i^{-1} V_i \tag{3.4}$$

is iteration dependent and r_1 and r_2 are the appropriate right-hand sides. The right-hand sides are the only values that change between the predictor and corrector step. See [11] for further information on the calculation of the right-hand sides and diagonal modification.

For the remainder of this section, we look at the linear algebra necessary to calculate the direction at each iteration. We initially develop the case where S is positive definite and we have only simple bounds. We then discuss the modification made for arbitrary linear constraints. We finish with the most general case, where S is not assumed to be positive definite.

3.1. Simple bound-constrained case. We first describe the method in the simplest context, that of the SVM formulation in (2.10). In this case, $S = \frac{1}{\nu} I$ is positive definite, $R = D \begin{bmatrix} A & -e \end{bmatrix}$, $H = I$, and B is not present. The linear system (3.3) reduces to

$$(C + RHR^T)\Delta x = r_1. \tag{3.5}$$

While the matrices R and H are constant over iterations, the matrix C in (3.4) and r_1 are iteration dependent.

The matrix in (3.5) is a rank-k update to an easily invertible matrix. Therefore, we can use the Sherman–Morrison–Woodbury [13] formula

$$(C + RHR^T)^{-1} = C^{-1} - C^{-1}R(H^{-1} + R^T C^{-1} R)^{-1} R^T C^{-1}$$

to solve for Δx. It is trivial to form C^{-1} and H^{-1} because they are both positive definite diagonal matrices. The matrix $H^{-1} + R^T C^{-1} R$ is a (small) symmetric $k \times k$ matrix that, once formed, can be handled by standard dense linear algebra subroutines. Since this matrix is independent of r_1, we have to form and factor this small dense matrix only once per iteration. That is, we can use the same factors in both the predictor and the corrector steps. To summarize, to solve (3.5), we carry out the following steps.

Algorithm SMW.

1. Calculate $t_1 = R^T C^{-1} r_1$.
2. Solve $(H^{-1} + R^T C^{-1} R) t_2 = t_1$.
3. Determine $\Delta x = C^{-1}(r_1 - R t_2)$.

Note that t_1 and t_2 are small k-vectors. Furthermore, the calculation in step 1 can be carried out at the same time the matrix required in step 2 is being formed. Thus, a complete solve requires two passes through the data stored as R, namely, one for steps 1 and 2 and one for step 3. This feature is important for the out of core implementation discussed in section 4.

3.2. Constrained case. We now turn to the case where the quadratic program under consideration still has a positive definite Q matrix but the problem has a small number of linear constraints. For example, problem (2.11) falls into this class, where $S = \frac{1}{\nu}I$ is positive definite, $R = DA$, $H = I$, and $B = e^T D$. Note that B is a nonzero, $1 \times m$ matrix with full row rank.

The predictor-corrector method requires the solution of (3.3) at each iteration. We have already shown how to apply $(C + RHR^T)^{-1}$ using Algorithm SMW. We use this observation to eliminate $\Delta x = (C + RHR^T)^{-1}(r_1 + B^T\Delta\lambda)$ from (3.3) and generate the following system in $\Delta\lambda$:

$$B(C + RHR^T)^{-1}B^T\Delta\lambda = r_2 - B(C + RHR^T)^{-1}r_1. \tag{3.6}$$

Since B has full row rank and $C + RHR^T$ is symmetric positive definite, we conclude that $B(C+RHR^T)^{-1}B^T$ is symmetric and positive definite. Hence, it is nonsingular, and the linear system (3.6) is solvable for any r_1 and r_2.

To use (3.6), we must solve the system

$$(C + RHR^T)\begin{bmatrix} T_1 & t_2 \end{bmatrix} = \begin{bmatrix} B^T & r_1 \end{bmatrix}$$

with multiple right-hand sides corresponding to the columns of B^T and r_1. However, we never need to form or factor $(C + RHR^T)$ explicitly, since we can solve for all the right-hand sides simultaneously using Algorithm SMW, incurring the cost only of storing T_1, an $m \times n$ matrix, and t_2. Note that in our SVM examples, $n = 1$.

Let us review the steps needed to solve (3.3).

1. Form $T_1 = (C + RHR^T)^{-1}B^T$ and $t_2 = (C + RHR^T)^{-1}r_1$ using a simultaneous application of Algorithm SMW.
2. Calculate $t_3 = r_2 - Bt_2$ using the solution from step 1.
3. Form the $n \times n$ matrix $T_2 = BT_1$.
4. Solve $T_2\Delta\lambda = t_3$, for the solution of (3.6).
5. Calculate $\Delta x = t_2 + T_1\Delta\lambda$.

Steps 2 and 3 can be done concurrently with step 1. Specifically, we can accumulate T_2 and t_3 as the elements in T_1 and t_2 become available from step 3 of Algorithm SMW. Per iteration, this scheme requires only two passes through the data in R, all in step 1, and one pass through T_1 in step 5.

Furthermore, since the predictor-corrector method requires two solves of the form (3.3) per iteration with different r_1 and r_2, the extra storage used for T_1 means that we need to calculate T_1 only once per iteration. For efficiency, we reuse the factors of $C + RHR^T$ in step 2 of Algorithm SMW and T_2 in step 4 of the above algorithm in both the predictor and corrector steps of the interior-point algorithm.

3.3. General case. Unfortunately, this is not the end of the story because formulations (2.8) and (2.9) do not have a positive definite matrix S but instead use $S = 0$. In fact, these problems also have lower and upper bounds. In this setting, while the matrix $C = Y^{-1}W + Z^{-1}V$ (for appropriately defined W, V, Y, and Z) is positive definite on the interior of the box defined by the bound constraints, the

interior-point method typically runs into numerical difficulties when the solution approaches the boundary of the box constraints.

Algorithmically, we would like the optimization problem to have a positive definite S matrix. When S is already positive definite, no modifications are needed in (2.1). For example, (2.10) and (2.11) have positive definite Q matrices and are strongly convex quadratic programs.

However, when S is only positive semidefinite (or zero), we can use a proximal-point modification [25]. Proximal-point algorithms augment the objective function with a strongly convex quadratic term and repeatedly solve the resulting quadratic program until convergence is achieved. That is, given x_i, they solve the quadratic program

$$\begin{array}{ll} \min_x & \frac{1}{2}x^TQx + c^Tx + \frac{\eta}{2}\|x - x_i\|_2^2 \\ \text{subject to} & Bx = b, \\ & x \geq 0 \end{array} \tag{3.7}$$

for some $\eta > 0$, possibly iteration dependent, to find a new x^{i+1}. The algorithm repeatedly solves subproblems of the form (3.7) until convergence occurs. Properties of such algorithms are developed in [25, 7], where it is shown that if the original problem has a solution, then the proximal-point algorithm converges to a particular element in the solution set of the original problem. Furthermore, each of the quadratic subproblems is strongly convex.

This approach may be used to solve (2.8) and (2.9), for example. However, rather than solving each subproblem (3.7) exactly, we instead solve the subproblems inexactly by applying just one step of the interior-point method before updating the subproblem. Thus, in effect, we are solving at each iteration the system of equations

$$\begin{bmatrix} C + RHR^T + \eta I & -B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta\lambda \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \end{bmatrix}.$$

Therefore, when using the proximal-point perturbation algorithm, we use the same interior-point implementation and simply modify the C matrix.

Since a proximal-point perturbation can cause the algorithm to take many iterations, our code initiates the proximal-point perturbation algorithm only when numerical difficulties are encountered. We identify numerical difficulties with an increase in the error for satisfying the equations. This technique switches to the proximal-point algorithm when necessary.

The linear algebra issues are now the same as the issues already covered above except for the particular values present in S. The remaining challenge is to solve massive problems. The implementation is discussed in the next section, where we use an out of core computation to reduce memory requirements.

4. Implementation. An interesting feature of the SVM problem is the volume of data, which can lead to quadratic programs with between 10 and 100 million variables and a Q matrix that would be dense if formed explicitly. Quadratic programming codes explicitly using the Q matrix will not work well for these problems. We need a method for which we can utilize specialized linear algebra. Therefore, we use the Mehrotra predictor-corrector algorithm [23] as implemented in OOQP [11] as the basis for our interior-point method. The OOQP code is written in such a way that we can easily tailor the linear algebra to the application. This feature can be exploited to enable the solution of large data mining problems.

The linear algebra outlined in section 3 is used in our implementation. As mentioned in section 3, we simultaneously solve systems of equations involving different right-hand side vectors and also reuse appropriate vectors and matrices for the predictor and corrector steps. However, because of the target size, we must effectively deal with the volume of data. Potentially, round-off or accumulation errors could become significant, so we want to minimize these as much as possible. Finally, we want to use a termination condition independent of the problem size. These topics are discussed in the following subsections along with further information on selecting a starting point. Clearly, the fact that interior-point algorithms typically require only a small number of iterations is crucial for performance.

The scaling of the problem can affect the behavior of the numerical linear algebra used to calculate the solution. For example, we experimented with the substitution $\tilde{x} = \frac{x}{\nu}$ in (2.8) when $\nu > 1$. This helped to some extent when the standard OOQP starting point was used, but was not necessary with the special starting point described next.

4.1. Starting point. The starting point chosen for the method can significantly impact both the theoretical and practical performance of the algorithm. To achieve flexibility in the starting-point choice, we use the augmented system in (3.2) that has removed the bounds on x.

We know that the majority of the variables are zero at a solution to the SVM problem because the zero variables correspond to those observations correctly classified. Therefore, the starting point uses $x^0 = 0$ and $\lambda^0 = 0$. We choose w^0 and v^0 so that $w^0 - v^0 = c$. That is, the residual in the first equation of (3.2) at the starting point is zero. Since $c = -e$ for the SVM, we set $w^0 = (\nu + 1)e$ and $v^0 = (\nu + 2)e$. We are then left with a choice for the slack variables, y^0 and z^0, added to the augmented system for w and v. To retain parity with our choice for w^0 and v^0, we set $y^0 = (\nu + 2)e$ and $z^0 = (\nu + 1)e$. We use the same starting point for the formulations without upper bounds but note that v and z are removed from the problem.

Better numerical performance might be achieved by the algorithm with an alternative starting point. For example, we expect that at a solution, most elements of y will be zero and most elements of z will be ν. This fact is not reflected in the current choice of y^0. We did not perform any further investigation of this topic.

4.2. Data issues. Consider a model with 50 million observations and suppose there are 35 features, each represented by a 1-byte quantity. Then, the observation matrix R is $50,000,000 \times 35$ and consumes 1.75 gigabytes of storage. If the features are measured as double-precision values, the storage requirement balloons to 14 gigabytes. Furthermore, the quadratic program has 50 million variables. Therefore, each double-precision vector requires 400 megabytes of space. If we assume 10 vectors are used, an additional 4 gigabytes of storage is necessary. Thus, the total space requirement for the algorithm on a problem of this magnitude is between 5.75 and 18 gigabytes. Clearly, an in core solution is not possible on today's machines.

We must attempt to perform most, if not all, of the operations using data kept out of core, while still achieving adequate performance. All of the linear algebra discussed in section 3 accesses the data sequentially. Therefore, while working on one buffer (block) of data, we can be reading the next from disk. The main computational component is constructing the matrix $M = H^{-1} + R^T C^{-1} R$ (see step 2 of Algorithm SMW). We begin by splitting R and C^{-1} into p buffers of data and calculate

$$M = H^{-1} + \sum_{j=1}^{p} R_j^T (C^{-1})_j R_j.$$

Note that C is a diagonal matrix in the examples considered but that more general matrices can be handled with more sophisticated splitting techniques.

To summarize, we perform the following steps to calculate M.

1. Request R_1 and $(C^{-1})_1$ from disk, and set $M = H^{-1}$.
2. For $j = 1$ to $p - 1$ do
   (a) Wait for R_j and C_j^{-1} to finish loading.
   (b) Request R_{j+1} and C_{j+1}^{-1} from disk.
   (c) Accumulate $M = M + R_j^T (C^{-1})_j R_j$.
3. Wait for R_p and C_p^{-1} to finish loading.
4. Accumulate $M = M + R_p^T (C^{-1})_p R_p$.

The code uses asynchronous I/O constructs to provide the request and wait functionality. The remainder of the linear algebra in section 3 can be calculated similarly. The code performs as many of the required steps as possible concurrently with the reading of the R_j buffers from disk.

The amount of data kept in core is significantly reduced with such a scheme. The tradeoff is that the code is not as fast as an in core implementation. In section 5, we quantify the impact of the out of core calculation.

4.3. Numerical considerations. Because of the number of variables in the problems solved, we can run into significant round-off errors while performing the linear algebra, particularly when accumulating the matrices. A naive implementation of Algorithm SMW that does nothing to address these problems results in divergence of the interior-point method for a moderately sized problem with one million observations. In an attempt to limit the effect of these numerical errors, we use a combination of aggregation to identify a block and bucketing within the block for the computations.

4.3.1. Aggregation. Consider the construction of the matrix $H^{-1} + R^T C^{-1} R$ using the above technique. The aggregation technique accumulates the $R_j^T (C^{-1})_j R_j$ components in temporary matrices, M_l for $l = 1, \ldots, L$, and then merges these as $M = \sum_{l=1}^{L} M_l$. Specifically, the initialization and accumulation steps are updated from the algorithm above into the following final form.

1. Request R_1 and $(C^{-1})_1$ from disk, and set $M_1 = H^{-1}$ and $M_l = 0$ for $l = 2, \ldots, L$.
2. For $j = 1$ to $p - 1$ do
   (a) Wait for R_j and C_j^{-1} to finish loading.
   (b) Request R_{j+1} and C_{j+1}^{-1} from disk.
   (c) Accumulate $M_{(j \mod L)+1} = M_{(j \mod L)+1} + R_j^T (C^{-1})_j R_j$.
3. Wait for R_p and C_p^{-1} to finish loading.
4. Accumulate $M_{(p \mod L)+1} = M_{(p \mod L)+1} + R_p^T (C^{-1})_p R_p$.
5. Merge $M = \sum_{l=1}^{L} M_l$.

Our merge is implemented by repeatedly adding the $\frac{L}{2}$ neighbors as depicted in Figure 4.1 (termed pairwise summation in [15]). A similar procedure is used for the vector computations. The code uses $L = 8$ for the calculations. We note that the above algorithm is dependent on the buffer size read from disk. This dependency is removed in the code by further partitioning R_j and C_j^{-1} into smaller buffers with 50,000 elements. This is a heuristic to limit the size of the intermediate summation values without having to perform an expensive sorting operation.

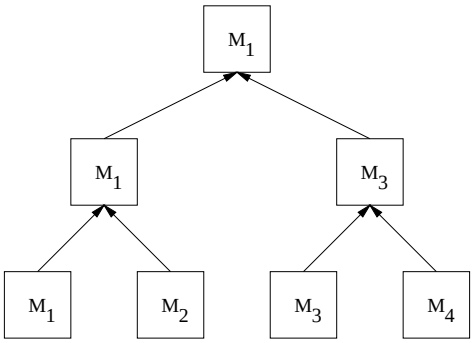


FIG. 4.1. *Accumulation diagram.*

4.3.2. Bucketing. While aggregation accumulates small batches of results, the bucketing strategy accumulates results of the same order of magnitude. The code uses 11 buckets with the ranges listed in Table 4.1. Whenever a result needs to be accumulated, it is assigned to the appropriate bucket. At the end of the computation, the buckets are merged. We decided to add first the positive and negative buckets of the same magnitude, and then accumulate the buckets starting with the smallest in magnitude. Again, this is a heuristic that does not require a sort of the data being accumulated. For summations involving numbers of the same sign, the accumulation from smallest to largest is as recommended in [29]. The addition of the positive and negative buckets of the same magnitude is designed to alleviate cancellation effects.

An example is given in [15] to test the effects of ordering on summations. The example has a large value M such that in floating-point arithmetic $1 + M \equiv M$, with the requirement that the values $1, 2, 3, 4, M, -M$ should be summed. Our bucketing and summation scheme results in the following summation:

$$(((1 + 2) + 3) + 4) + (M - M).$$

Furthermore, the correct result is calculated independent of the initial ordering of the values, and no sort is required. Further details on other orderings and examples can be found in [15].

Since some of our calculations have mixed signs and some involve just positive numbers, the combination of heuristics, aggregation to identify a block and bucketing within each block, was found to be very effective.

TABLE 4.1
Bucket ranges.

	Range	
Bucket	Lower bound	Upper bound
1	$-\infty$	-10^8
2	-10^8	-10^4
3	-10^4	-1
4	-1	-10^{-4}
5	-10^{-4}	-10^{-8}
6	-10^{-8}	10^{-8}
7	10^{-8}	10^{-4}
8	10^{-4}	1
9	1	10^4
10	10^4	10^8
11	10^8	∞

4.4. Termination criteria. The termination criterion is based on the inf-norm of the Fischer–Burmeister function [9] for the complementarity problem (3.1), with an appropriate modification for the presence of equations [8]. If we denote all variables in (3.1) by x and the affine function on the left of (3.1) by $F(x)$, then each component of the Fischer–Burmeister function is defined by

$$\phi(x_i, F_i(x)) := \sqrt{x_i^2 + F_i(x)^2} - x_i - F_i(x)$$

for those variables with lower bounds of zero and by $\phi(x_i, F_i(x)) = -F_i(x)$ for variables without bounds. We can see from this definition that $\phi(x_i, F_i(x)) = 0$ if and only if the complementarity relationship is satisfied between x_i and $F_i(x)$. The inf-norm is independent of the number of variables in the problem and can be stably calculated given evaluations of the linear functions in (3.1). We further note that the function F can be evaluated during the calculation of the right-hand side in the predictor step. Therefore, the function calculation does not cost an additional pass through the data. We use a termination criterion of 10^{-6} for the Fischer–Burmeister function within the code, which is much more stringent than the default criterion for OOQP. In Figure 4.2 we plot the (log) residual as a function of the iteration for problem (2.8) with 10 million observations.

We terminate unsuccessfully whenever the iteration limit is reached or we fail to achieve a decrease in the residual for satisfying the equations in the interior-point method for six consecutive iterations and the complementarity residual $(w^T y + v^T z)$ is less than $\frac{10^{-15}}{2m}$.

The machine learning community sometimes terminates an algorithm based upon conditions other than optimality, such as tuning set accuracy [20]. Similar criteria could be used within our code, but we prefer to terminate at an optimal solution to the quadratic program.

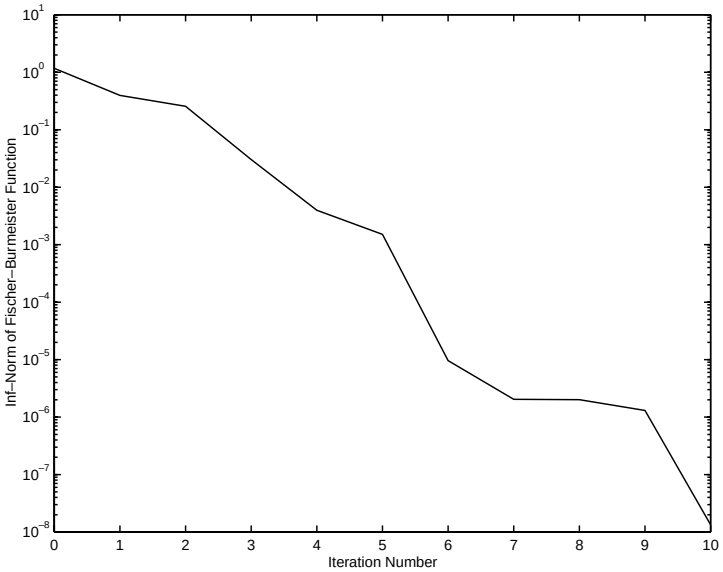


FIG. 4.2. *Log residual as a function of iterations for problem* (2.8) *with* 10 *million observations.*

5. Computational results. All of the tests were run on a 296 MHz Sun Ultrasparc with two processors and 768 megabytes of RAM. We stored all data on a locally mounted disk with 18 gigabytes of storage space available. This disk is not backed up. This setup prevents all overhead due to network communication and disk contention with nightly backups. Since the disk is not dedicated, our results reflect some effects due to contention with other users.

The asynchronous I/O routines are implemented using threads. Thus, both of the processors can be used for the tests. However, the workstation is shared by many individuals. During our tests the second processor was typically running a different user's jobs. Further results on a uniprocessor machine indicate that the impact of the second processor is minimal.

5.1. Data sets. For experimentation, we generated a separable, random data set with 34 features. We did this by constructing a separating hyperplane and then creating data points and classifying them with the hyperplane. The data generated contains 60 million observations of 34 features, where each feature has an integer value between 1 and 10. Multiplication by D was performed while the data was generated, with De being encoded as an additional column to the observation set. Each of the feature measurements is a 1-byte quantity. A nonseparable dataset was constructed by randomly changing the classification of the observations with a 1% probability. The nonseparable dataset has exactly 600,108 misclassified observations.

We limited the size to 60 million observations to avoid problems with the 2-gigabyte file size restriction imposed by various operating systems. To increase the size further without changing operating system, we could store the original data in multiple files.

5.2. Out of core impact. The impact on performance of using an out of core implementation was tested by using the formulation in (2.10) with $\nu = 1$ on the separable dataset. Since S is positive definite in this case, no proximal-point modification was added.

The first property investigated was the effect of out of core computations on performance using asynchronous I/O. To test the performance, we ran problems for sizes varying between 200,000 and 1 million observations. A data buffer size of 100,000 observations (elements) for each matrix (vector) was used for the out of core computations. We ran each of the tests five times and used the minimum values in the figures. The average time per iteration is reported in Figure 5.1 for in core, asynchronous I/O, and synchronous I/O implementations. While the asynchronous I/O is not as fast as keeping everything in core, we note only an 8.2–9.9% increase in time over the in core implementation for the chosen buffer and problem sizes. Synchronous I/O results in a 9.4–13.1% increase. For both of these tests the maximum percentage increase in time occurred with a problem size of 800,000 elements. We conclude that an out of core implementation of the algorithm uses limited memory but results in increased time. We believe that the enormous decrease in the amount of RAM used for a less than 10% increase in time is a reasonable tradeoff to make. A case can also be made for using the easier to implement synchronous I/O.

The next set of experiments was designed to determine the impact of modifying the file buffer size. For these tests, we fixed the problem size to 1 million observations and varied the file buffer size from 50,000 to 500,000 elements. The average time per iteration is plotted in Figure 5.2. The results indicate that a file buffer size of around 250,000 elements is optimal with a 9.0% increase in time over the in core solution. The total amount of data buffered in main memory is between 110 and 152 megabytes

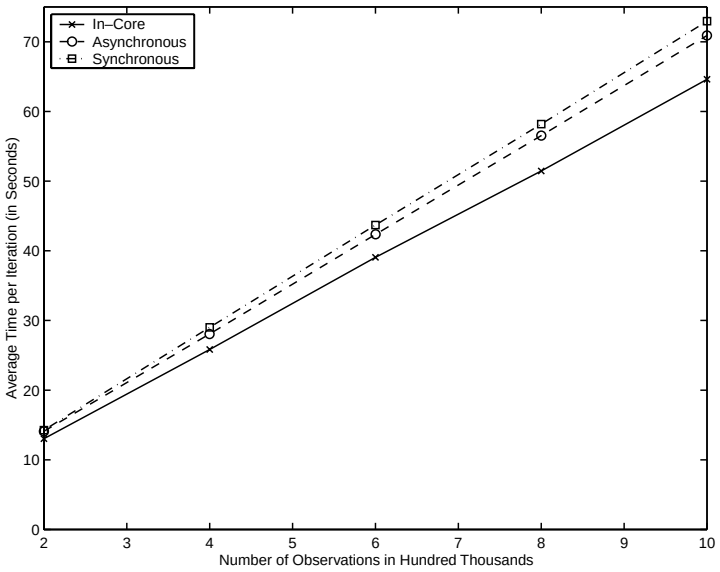


FIG. 5.1. *Average time per iteration for various problem sizes with a fixed file buffer size of* 100,000 *elements.*

depending on the problem formulation used. Based on these results, we decided to use asynchronous I/O and a buffer size of 250,000 elements for the remainder of the numerical experiments.

5.3. Baseline comparison. We next investigated the performance of our interior-point algorithm compared with other methods from the machine learning community.

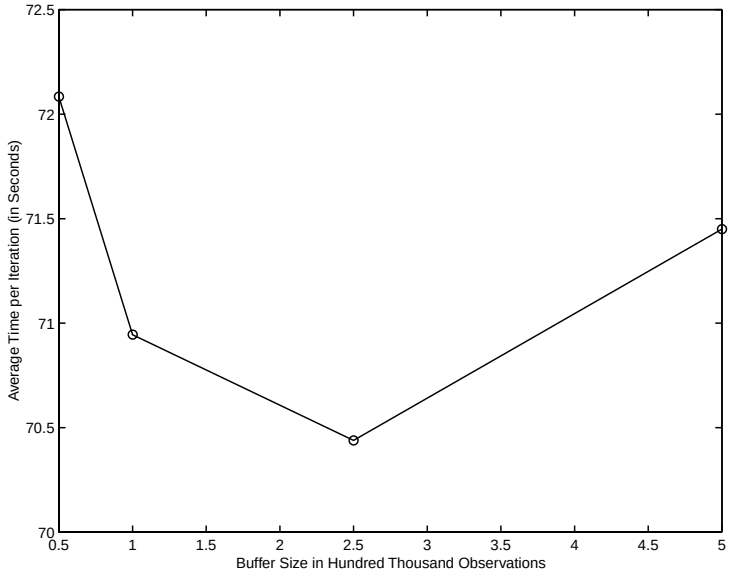


FIG. 5.2. *Average time per iteration for various file buffer sizes with a fixed problem size of* 1,000,000 *observations.*

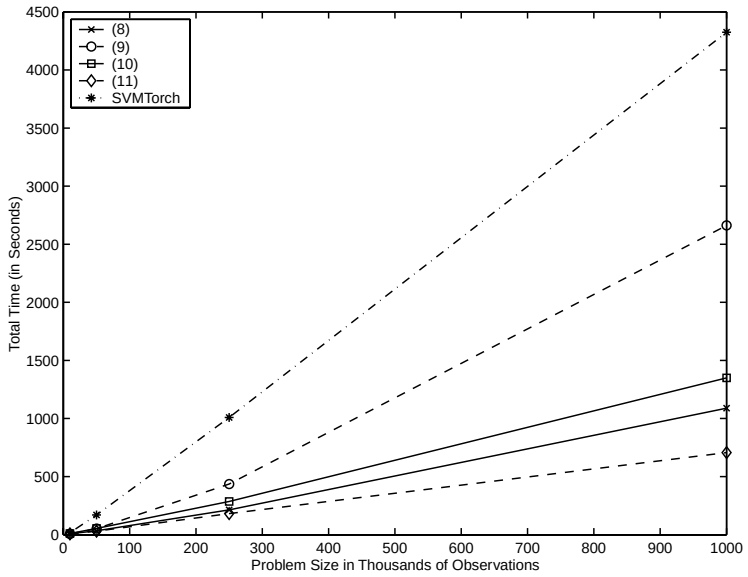


FIG. 5.3. *Total time comparison of the different formulations and SVMTorch with varying problem sizes on the separable dataset.*

We used SVMTorch [4] for this series of tests because the code is freely available and, according to the documentation, is specifically tailored for large-scale problems. We compiled both codes with the same compiler and options and converted the datasets into the binary format requested by SVMTorch. We ran SVMTorch using a linear kernel with $\nu = 1$. All other options, including the termination tolerance, were set to their default values.

Figure 5.3 reports the total running time for SVMTorch and each of (2.8)–(2.11) on the separable dataset with various numbers of observations. From these results, SVMTorch is 1.6–6.1 times slower than our codes on the separable dataset depending on the size and formulation chosen.

Results on the nonseparable dataset are more dramatic. SVMTorch took 1156.3 seconds to find a solution with 10,000 observations. Our interior-point codes took 5.8, 5.8, 8.6, and 9.8 seconds with formulations (2.8)–(2.11), respectively. These numbers indicate that SVMTorch is between 116 and 196 times slower on the nonseparable dataset. The magnitude becomes even larger when the number of observations is increased. With 50,000 observations, we let SVMTorch spend over 15 hours of CPU time in 380,000 iterations before terminating the SVMTorch code with a "current error" of 2.15. These numbers indicate that the SVMTorch code is at best more than 1,060 times slower than our interior-point code on this particular dataset. We did not perform any further tests with this code.

5.4. Sensitivity to ν. The next set of experiments was to determine the sensitivity of the method to increases in ν. In all of these tests, a proximal-perturbation of $\eta = 10^{-5}$ was added for the models in (2.8) and (2.9) when the error in solving the equations increased. We report in Figures 5.4 and 5.5 the number of iterations taken by the interior-point methods for various values of ν between 1 and 10,000 on the separable and nonseparable datasets with 1 million observations, respectively. On the separable dataset we notice increases in the number of iterations taken to find a

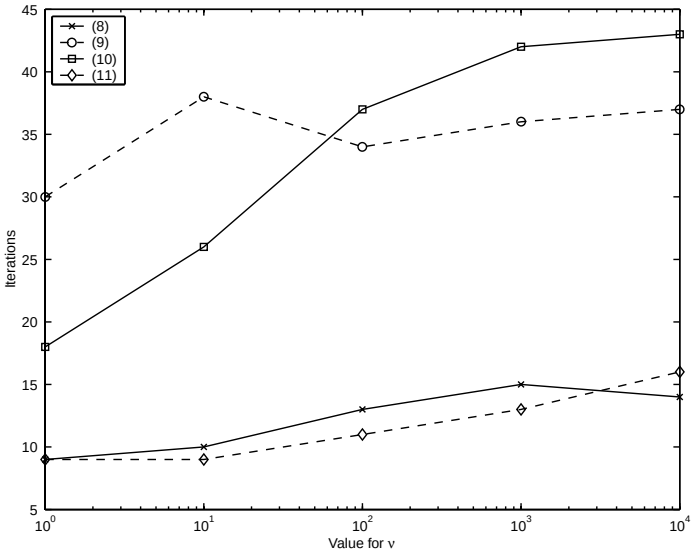


FIG. 5.4. *Iteration comparison of the different formulations with varying ν on the separable dataset.*

solutions. The iteration counts taken on the nonseparable dataset also increase. For (2.8) and (2.9), a small percentage of the variables at the solution were at the upper bound for all values of ν in both the separable and nonseparable tests. We note that on the nonseparable dataset for $\nu = 1,000$ and $\nu = 10,000$, all of the formulations failed to achieve termination tolerances; they stopped with final and best residuals reported in Table 5.1. Further improvements to the linear algebra implementation

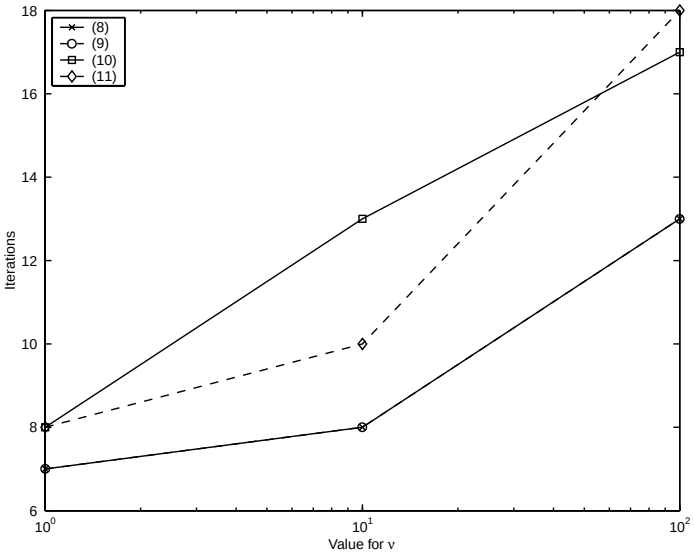


FIG. 5.5. *Iteration comparison of the different formulations with varying ν on the nonseparable dataset.*

TABLE 5.1
Final and best residuals reported for the different formulations with $\nu = 1,000$ and $\nu = 10,000$ on the nonseparable dataset.

	$\nu = 1,000$		$\nu = 10,000$	
Formulation	Final	Best	Final	Best
(2.8)	2.12e-6	1.38e-6	2.78e-5	1.71e-5
(2.9)	2.65e-6	1.36e-6	1.84e-5	1.04e-5
(2.10)	5.73e-6	3.76e-6	6.87e-5	3.50e-5
(2.11)	8.57e-6	3.93e-6	3.73e-5	2.60e-5

would need to be investigated in order to produce reasonable results for larger values of ν.

5.5. Massive problems. The final set of experiments was designed to determine the reliability of the algorithm on the various formulations and the scalability of the implementation to massive problems. Specifically, we varied the problem size between 1 and 60 million observations. In all of these tests $\nu = 1$ was used, and for the models in (2.8) and (2.9) a proximal-perturbation of $\eta = 10^{-5}$ was added when the error in solving (3.3) increased.

Each model was run one time with problem sizes of 1, 5, 10, 20, and 60 million observations. We plot average time per iteration in Figure 5.6 and number of iterations as functions of problem size in Figures 5.7 and 5.8, respectively. The similarity in the average time per iteration between formulations (2.10) and (2.11) (and also between (2.8) and (2.9)) is indistinguishable. To avoid clutter, we plot the results only for (2.8) and (2.11) in Figure 5.6. The total times are reported in Figures 5.9 and 5.10.

The average time per iteration appears to grow almost linearly with the problem size. This result is to be expected, as the majority of the time taken per iteration is in constructing $H^{-1} + R^T C^{-1} R$. The number of floating-point operations necessary

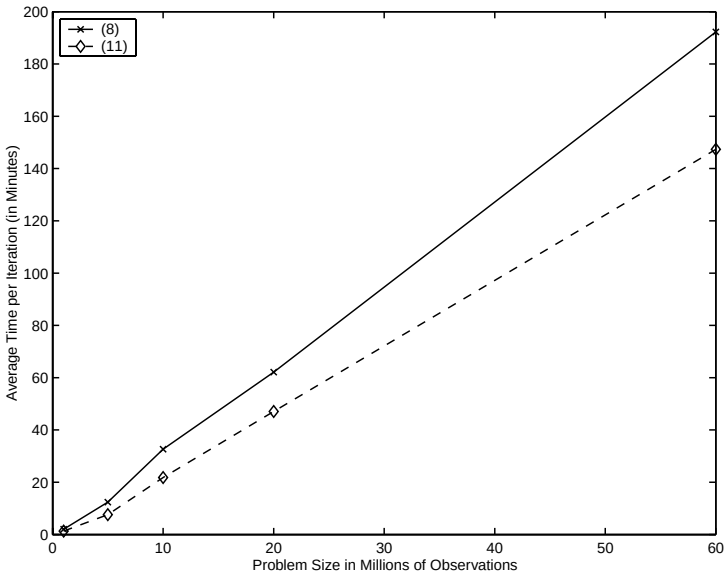


FIG. 5.6. *Average time per iteration comparison of the different formulations with varying problem size.*

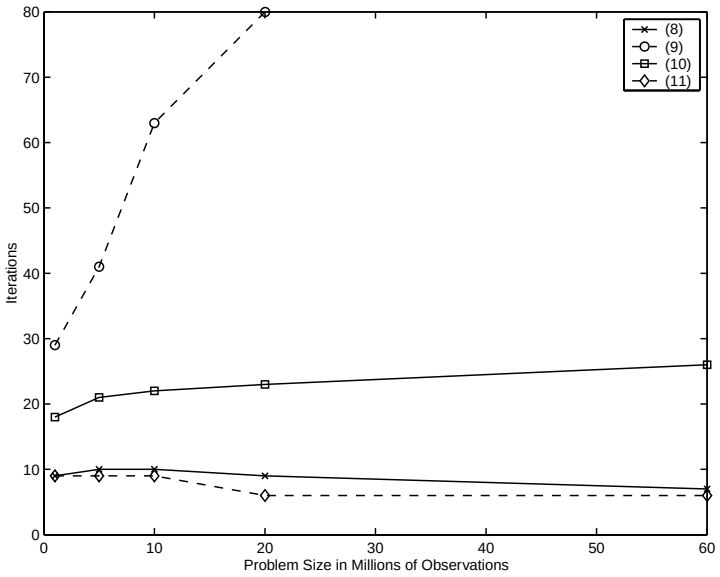


FIG. 5.7. *Iteration comparison of the different formulations with varying problem size on the separable dataset.*

to calculate this quantity grows linearly with problem size m (but quadratically with the number of features k). The extra time needed for (2.8) is due to the treatment of upper bounds.

A surprising result for the constrained formulations, (2.8) and (2.11), on the separable dataset is that the number of iterations remains fairly flat as the problem

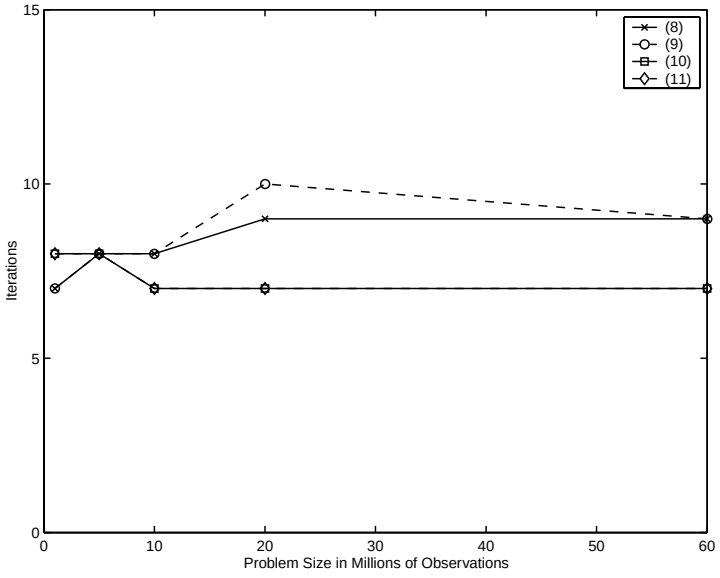


FIG. 5.8. *Iteration comparison of the different formulations with varying problem size on the nonseparable dataset.*

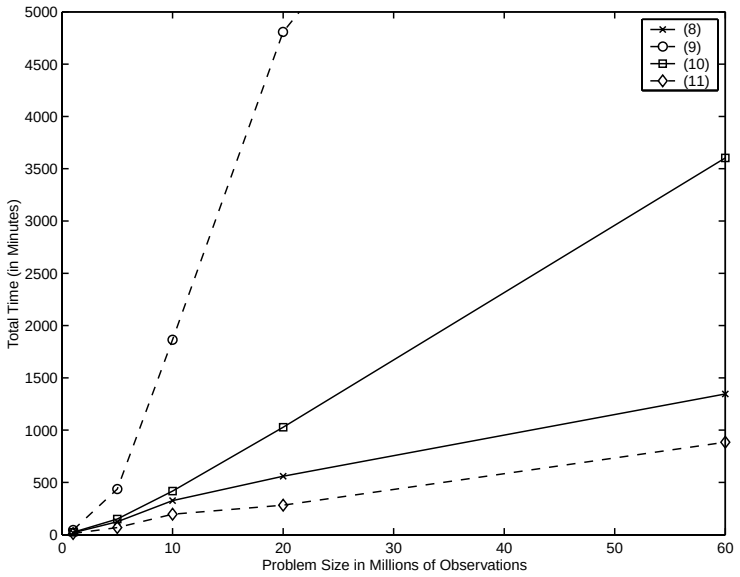


FIG. 5.9. *Total time comparison of the different formulations with varying problem size on the separable dataset.*

size increases and even decreases for some of the larger problems. As expected, the number of iterations taken for (2.9) and (2.10) increases with the dimension of the problem. We note a large difference in the number of iterations taken to converge for formulations (2.8) and (2.9) even though the problems are similar. The main reason for this difference is that many small steps are taken when solving (2.9).

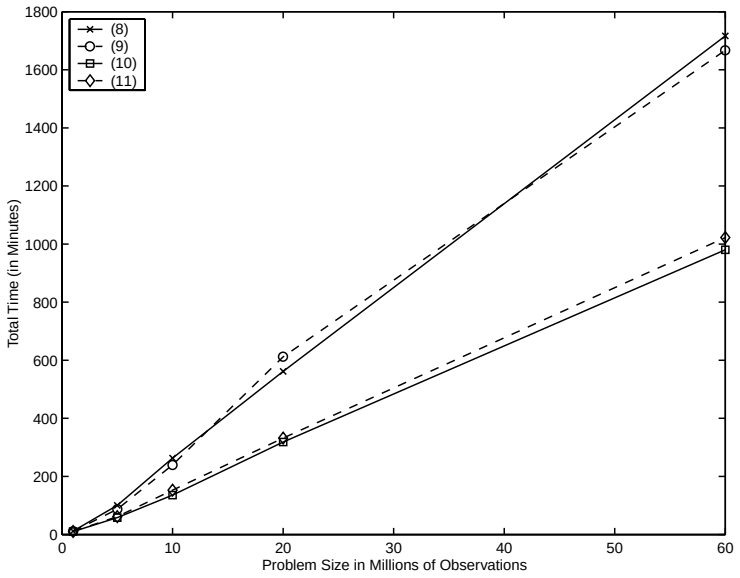


FIG. 5.10. *Total time comparison of the different formulations with varying problem size on the nonseparable dataset.*

All of the formulations performed extremely well on the nonseparable dataset, with very little variation in the number of iterations. These facts are counterintuitive and are probably related to the random nature of the model. However, more tests on "real" datasets need to be performed before drawing any firm conclusions.

The constrained formulations (2.8) and (2.11) appear to be the most tractable for interior-point methods. Both of these formulations solved the 60-million observation problem in 15–22.5 hours on a standard workstation. Formulations (2.9) and (2.10) also work for the 60-million observation problem but take longer times.

We believe the strength of this approach is its scalability and reliability. While it may be possible to adjust the parameters of the interior-point method or the parameters of the proximal-point iteration for improved performance, we have elected to use the same defaults on all problems and have not encountered any numerical difficulties beyond those documented in section 5.4.

6. Conclusions. We have developed an interior-point code for solving several quadratic programming formulations of the linear SVM. We are able to solve large problems reasonably by exploiting the linear algebra and using out of core computations. Scalability of the approach has been demonstrated.

The linear algebra can be parallelized easily, and further speedups can be realized through storage of the data across multiple disks. More sophisticated corrector implementations [14] of the interior-point code can be used to further reduce the iteration count. These are topics for future work, along with extensions to nonlinear SVM, and techniques to further reduce the number of data scans.

Acknowledgments. We thank Olvi Mangasarian for bringing this problem to our attention; Yuh-Jye Lee and David Musicant for numerous insightful discussions on machine learning problems; Mike Gertz, Jeff Linderoth, and Stephen Wright for making a preliminary version of OOQP available to us; and Michael Saunders and two anonymous referees for encouraging us to further explore the numerical properties of the code.

REFERENCES

[1] P. S. Bradley and O. L. Mangasarian, *Massive data discrimination via linear support vector machines*, Optim. Methods Softw., 13 (2000), pp. 1–10.
[2] C. J. C. Burges, *A tutorial on support vector machines for pattern recognition*, Data Mining and Knowledge Discovery, 2 (1998), pp. 121–167.
[3] W. G. Cochran, *Sampling Techniques*, 3rd ed., John Wiley and Sons, New York, 1977.
[4] R. Collobert and S. Bengio, *SVMTorch: Support vector machines for large-scale regression problems*, J. Mach. Learn. Res., 1 (2001), pp. 143–160.
[5] C. Cortes and V. Vapnik, *Support vector networks*, Machine Learning, 20 (1995), pp. 273–297.
[6] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, Cambridge, UK, 2000.
[7] M. C. Ferris, *Finite termination of the proximal point algorithm*, Math. Program., 50 (1991), pp. 359–366.
[8] M. C. Ferris, C. Kanzow, and T. S. Munson, *Feasible descent algorithms for mixed complementarity problems*, Math. Program., 86 (1999), pp. 475–497.
[9] A. Fischer, *A special Newton-type optimization method*, Optimization, 24 (1992), pp. 269–284.
[10] V. Ganti, J. Gehrke, and R. Ramakrishnan, *A framework for measuring changes in data characteristics*, in Proceedings of the Eighteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, May 31–June 2, 1999, Philadelphia, Pennsylvania, ACM Press, New York, 1999, pp. 126–137.
[11] M. Gertz and S. Wright, *Object-Oriented Software for Quadratic Programming*, Preprint ANL/MCS-P891-1000, Argonne National Laboratory, Argonne, IL, 2001.

[12] M. Gertz and S. Wright, *OOQP User Guide*, Technical Memorandum ANL/MCS-TM-252, Argonne National Laboratory, Argonne, IL, 2001.
[13] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, MD, 1996.
[14] J. Gondzio, *Multiple centrality corrections in a primal-dual method for linear programming*, Comput. Optim. Appl., 6 (1996), pp. 137–156.
[15] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 1996.
[16] P. J. Huber, *Robust Statistics*, John Wiley and Sons, New York, 1981.
[17] W. Li and J. J. Swetits, *The linear ℓ_1 estimator and the Huber M-estimator*, SIAM J. Optim., 8 (1998), pp. 457–475.
[18] O. L. Mangasarian, *Nonlinear Programming*, McGraw–Hill, New York, 1969; Classics Appl. Math. 10, SIAM, Philadelphia, 1994.
[19] O. L. Mangasarian, *Generalized support vector machines*, in Advances in Large Margin Classifiers, A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, eds., MIT Press, Cambridge, MA, 2000, pp. 135–146.
[20] O. L. Mangasarian and D. R. Musicant, *Successive overrelaxation for support vector machines*, IEEE Transactions on Neural Networks, 10 (1999), pp. 1032–1037.
[21] O. L. Mangasarian and D. R. Musicant, *Robust linear and support vector regression*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22 (2000), pp. 950–955.
[22] O. L. Mangasarian and D. R. Musicant, *Active set support vector machine classification*, in Advances in Neural Information Processing Systems 13, T. K. Leen, T. G. Dietterich, and V. Tresp, eds., MIT Press, Cambridge, MA, 2001, pp. 577–583.
[23] S. Mehrotra, *On the implementation of a primal-dual interior point method*, SIAM J. Optim., 2 (1992), pp. 575–601.
[24] J. Platt, *Sequential minimal optimization: A fast algorithm for training support vector machines*, in Advances in Kernel Methods—Support Vector Learning, B. Schölkopf, C. J. C. Burges, and A. J. Smola, eds., MIT Press, Cambridge, MA, 1999, pp. 185–208.
[25] R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.
[26] B. Schölkopf, C. Burges, and A. Smola, eds., *Advances in Kernel Methods: Support Vector Machines*, MIT Press, Cambridge, MA, 1998.
[27] A. J. Smola, *Learning with Kernels*, Ph.D. thesis, GMD Research Series 25, Technische Universität Berlin, Germany, 1998.
[28] V. N. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed., Springer-Verlag, New York, 2000.
[29] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, London, 1965.
[30] S. J. Wright, *Primal–Dual Interior–Point Methods*, SIAM, Philadelphia, 1997.
[31] S. J. Wright, *On reduced convex QP formulations of monotone LCP problems*, Math. Program., 90 (2001), pp. 459–474.

SIAM J. OPTIM.
Vol. 13, No. 3, pp. 805–841

SEMISMOOTH NEWTON METHODS FOR OPERATOR EQUATIONS IN FUNCTION SPACES*

MICHAEL ULBRICH†

Abstract. We develop a semismoothness concept for nonsmooth superposition operators in function spaces. The considered class of operators includes nonlinear complementarity problem (NCP)-function-based reformulations of infinite-dimensional nonlinear complementarity problems and thus covers a very comprehensive class of applications. Our results generalize semismoothness and α-order semismoothness from finite-dimensional spaces to a Banach space setting. For this purpose, a new infinite-dimensional generalized differential is used that is motivated by Qi's finite-dimensional C-subdifferential [*Research Report AMR*96/5, School of Mathematics, University of New South Wales, Australia, 1996]. We apply these semismoothness results to develop a Newton-like method for nonsmooth operator equations and prove its local q-superlinear convergence to regular solutions. If the underlying operator is α-order semismooth, convergence of q-order $1+\alpha$ is proved. We also establish the semismoothness of composite operators and develop corresponding chain rules. The developed theory is accompanied by illustrative examples and by applications to NCPs and a constrained optimal control problem.

Key words. Newton-like methods, semismoothness, superposition operators, generalized differentials, nonlinear complementarity problems, superlinear convergence, optimal control problems

AMS subject classifications. 49M15, 65K05, 90C33, 49J52, 47J25, 47H30

PII. S1052623400371569

1. Introduction.

In this paper, we develop a semismoothness concept for nonsmooth operators in function spaces and establish q-superlinear convergence of a Newton-like method for semismooth operator equations. Results on convergence with rate >1 are also presented. The class of operators we consider includes those obtained by nonlinear complementarity problem (NCP)-function-based reformulations of NCPs in function spaces. These problems arise frequently in practice, e.g., in the form of first-order optimality conditions of constrained elliptic [62, 63], parabolic [64], and flow control problems [62, 60]. As an illustrative example of the application to optimal control, we will discuss the elliptic control problem (1.6) in detail. The numerical results in [63, 62, 60] show that the semismooth Newton method developed in this paper solves constrained control problems very efficiently.

The notion of semismoothness was introduced by Mifflin [43] for real-valued functions defined on finite-dimensional spaces. Qi [50] and Qi and Sun [52] extended semismoothness to mappings between finite-dimensional spaces and showed that, although the underlying mapping is in general nonsmooth, Newton's method can be generalized to semismooth equations and converges locally with q-superlinear rate to a regular solution [49, 50, 52]. For related early approaches to nonsmooth Newton methods, we refer to [40, 41, 47]. In particular, Kummer [40, 41] has established q-superlinear convergence for a general, abstract class of nonsmooth Newton methods under conditions that include (1.1).

Written in a form most convenient for our purposes, a mapping $f:\mathbb{R}^k\to\mathbb{R}^l$ is called semismooth at x if f is Lipschitz near x and directionally differentiable at x,

*Received by the editors May 2, 2000; accepted for publication (in revised form) June 14, 2002; published electronically January 3, 2003.
http://www.siam.org/journals/siopt/13-3/37156.html
†Fachbereich Mathematik, Universität Hamburg, Bundesstr. 55, D-20146 Hamburg, Germany (ulbrich@math.uni-hamburg.de). The author was supported by Deutsche Forschungsgemeinschaft (DFG) grant Ul157/3-1 and by CRPC grant CCR–9120008.

and if

$$\max_{M\in\partial f(x+h)} \|f(x+h)-f(x)-Mh\| = o\big(\|h\|\big) \quad \text{as } h\to 0, \tag{1.1}$$

where ∂f denotes Clarke's generalized Jacobian [13]. See section 2 for details. Further, if f is α-order semismooth, $0<\alpha\le 1$, then the small-order term in (1.1) can be improved to $O(\|h\|^{1+\alpha})$. An important source of semismooth equations are reformulations of the NCP

$$y_i\ge 0, \quad Z_i(y)\ge 0, \quad y_iZ_i(y)=0, \quad i=1,\dots,k, \tag{1.2}$$

with a continuously differentiable function $Z:\mathbb{R}^k\to\mathbb{R}^k$. In this approach, which can also be applied to more general problems (mixed complementarity problem, MCP; variational inequality problem, VIP), an NCP-function [57], i.e., a function $\phi:\mathbb{R}^2\to\mathbb{R}$ with the property

$$\phi(x)=0 \quad\Longleftrightarrow\quad x_1\ge 0, \quad x_2\ge 0, \quad x_1x_2=0,$$

is applied componentwise to the NCP to rewrite it equivalently in the form

$$\Phi(y)=0, \quad \text{where} \quad \Phi(y)=\big(\phi(y_1,Z_1(y)),\dots,\phi(y_k,Z_k(y))\big)^T. \tag{1.3}$$

Frequently used NCP-functions are $\phi(x)=\min\{x_1,x_2\}$ as well as the Fischer–Burmeister function (see [22])

$$\phi_{FB}(x)=\sqrt{x_1^2+x_2^2}-x_1-x_2. \tag{1.4}$$

Both are semismooth (of order 1), and thus the function Φ in (1.3) is also semismooth. Therefore, semismooth Newton methods can be applied to solve (1.3). The strong theoretical properties of this approach, its numerical potential, and extensions to more general problems (MCP, VIP) have been extensively studied in recent years (see, e.g., [17, 19, 20, 36, 37, 61]) and have led to very efficient Newton-like methods (see, e.g., [45]). Although smooth NCP-functions can be constructed [42], they suffer from the fact that $\nabla\phi(0)=0$ necessarily must hold since the curve $\{x\in\mathbb{R}^2:\phi(x)=0\}$ has a kink at $x=0$. As a consequence, the use of smooth NCP-functions requires a strict complementarity condition, whereas this can be avoided by working with nondifferentiable NCP-functions. Since the introduction of the semismooth Fischer–Burmeister function, many researchers agree that semismooth NCP-functions are a very powerful tool for developing efficient algorithms with strong theoretical properties.

The objective of this paper is to extend the notions of semismoothness and α-order semismoothness, respectively, to nonlinear superposition operators in function spaces, and to develop a corresponding superlinearly convergent Newton-like method. Hereby, we are motivated by applications arising in mathematical modeling and optimal control, which often (see below) can be cast as pointwise bound-constrained VIP posed in function spaces. As our main example we consider the following NCP: Find $y\in L^p(\Omega)$ such that

$$y\ge 0, \quad Z(y)\ge 0, \quad yZ(y)=0 \tag{1.5}$$

holds pointwise a.e. on Ω, where $\Omega\subset\mathbb{R}^n$ is Lebesgue measurable with positive and finite measure, $L^p(\Omega)$ is the Lebesgue space of p-integrable functions, and the operator $Z:L^p(\Omega)\to L^r(\Omega)$, $1\le r<p\le\infty$, is continuously Fréchet differentiable. For the purpose of illustration, we now show how a particular optimal control problem can

be converted to an NCP of the form (1.5). The problem we describe will serve as a model problem (chosen to be simple for convenience) to which our theory and the developed Newton method are readily applicable. Consider the following distributed optimal control problem of an elliptic partial differential equation with upper bounds on the control:

$$\begin{aligned} &\underset{w\in L^2(\Omega)}{\text{minimize}} \quad J(w) \stackrel{\text{def}}{=} \frac{1}{2}\,\|u(w)-u_d\|^2_{L^2(\Omega)} + \frac{\lambda}{2}\,\|w-w_d\|^2_{L^2(\Omega)} \\ &\text{subject to} \quad w \le b \quad \text{on } \Omega, \end{aligned} \tag{1.6a}$$

where $u = u(w) \in H^1_0(\Omega)$ (the usual Sobolev space) is the weak solution of the uniformly elliptic state equation

$$-\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial u}{\partial x_j}\right) = w \quad \text{on } \Omega. \tag{1.6b}$$

We assume $\lambda > 0$, $a_{ij} \in L^\infty(\Omega)$, $u_d \in L^2(\Omega)$, and $w_d, b \in L^\infty(\Omega)$. Denoting by $\nabla J(w) \in L^2(\Omega)$ the L^2-Riesz representation of the gradient of J, it will be shown in Example 5.6 that $\bar w$ solves the control problem if and only if $\bar y = b - \bar w$ solves the NCP (1.5) with $Z(y) = -\nabla J(b-y)$. We will further discuss this problem in Example 5.6 and in section 6.2. We stress that this problem is meant for the purpose of illustration, and thus we decided to consider this particularly simple linear-quadratic control problem, which we hope is easily accessible to most readers. For more advanced applications to the optimal control of nonlinear partial differential equations, we refer the interested reader to [62, 60].

In order to reformulate (1.5) as a nonsmooth operator equation, we use an NCP-function to rewrite the pointwise complementarity conditions in (1.5) as equations. Doing this, (1.5) can be cast equivalently in form of the operator equation

$$\Phi(y) = 0, \quad \text{where} \quad \Phi(y)(\omega) \stackrel{\text{def}}{=} \phi(y(\omega), Z(y)(\omega)), \quad \omega \in \Omega. \tag{1.7}$$

In this paper, we consider superposition operators of the more general form

$$\Psi : Y \to L^r(\Omega), \quad \Psi(y)(\omega) = \psi(F(y)(\omega)), \tag{1.8}$$

with mappings $\psi : \mathbb{R}^m \to \mathbb{R}$ and $F : Y \to \prod_{i=1}^m L^{r_i}(\Omega)$, where $1 \le r \le r_i < \infty$, Y is a real Banach space, and $\Omega \subset \mathbb{R}^n$ is a bounded open domain. Obviously, choosing $Y = L^p(\Omega)$, $r_1 = r_2 = r$, $m = 2$, $\psi = \phi$, and $F : y \in Y \mapsto (y, Z(y))$, we have $\Psi = \Phi$ with Φ and Ψ as in (1.7) and (1.8), respectively, so that reformulated NCPs are included as special cases in our analysis. Essentially, our working assumptions are that ψ is Lipschitz continuous and semismooth and that F is continuously Fréchet differentiable. The detailed assumptions are given below. The main result of this paper is a semismoothness-like estimate of the form

$$\sup_{M\in\partial^\circ\Psi(y+s)} \|\Psi(y+s) - \Psi(y) - Ms\|_{L^r} = o(\|s\|_Y) \quad \text{as } s \to 0 \text{ in } Y. \tag{1.9}$$

We also give conditions under which the remainder term in (1.9) is of the order $O(\|s\|_Y^{1+\alpha})$, $0 < \alpha \le 1$. In this case we call Ψ α-order semismooth. The multifunction (i.e., set-valued mapping) $\partial^\circ\Psi : Y \rightrightarrows \mathcal{L}(Y, L^r)$ denotes an appropriate vector-valued generalized differential of Ψ, which is related to, and motivated by, Qi's finite-dimensional C-subdifferential [51]. The estimate (1.9) generalizes (1.1) to the function space setting. We will not require that Ψ be directionally differentiable, because this is not needed in the analysis of Newton's method. We remark that several

authors [24, 40, 44, 67] have studied conditions of the form (1.9) in finite dimensions independently of the papers [50] and [52].

Based on (1.9), we develop a locally q-superlinearly convergent Newton method for the nonsmooth operator equation

$$\Psi(y) = 0. \tag{1.10}$$

Moreover, in the case in which Ψ is α-order semismooth, we prove convergence with q-rate $1 + \alpha$. In analogy to BD (Bouligand differential)-regularity assumptions for finite-dimensional semismooth Newton methods, we impose a regularity condition on the elements of the generalized differential. Further, as was observed earlier in the context of related local convergence analyses in function space [38, 64], we have to incorporate a smoothing step to overcome the nonequivalence of norms. We also will provide an example showing that this smoothing step can be indispensable.

Recently, a different semismoothness concept for operator equations was proposed by Chen, Nashed, and Qi [12]. Our approach differs significantly from the one in [12]. There, the notion of a slanting function is introduced, and a generalized derivative, the slant derivative, is obtained as the collection of all limits of the slanting function as $y_k \to y$. Semismoothness is then defined by imposing appropriate conditions on the approximation properties of the slanting function and the slant derivative.

Although the differentiability properties of superposition operators with smooth ψ are well investigated (see, e.g., the expositions [6] and [7]), this is not the case for nonsmooth functions ψ. Further, even if ψ is smooth, for operator equations of the form (1.10) the availability of local convergence results for Newton-like methods appears to be very limited.

As an important application and to illustrate our results, we discuss reformulations (1.7) of the NCP (1.5). Furthermore, we show how the constrained elliptic control problem (1.6) can be converted to an equivalent NCP that meets all our assumptions.

There are also close connections between the NCP-function approach and noninterior path-following methods for NCPs [11], which recently were introduced and analyzed in finite dimensions. Hereby, the NCP-function ϕ is embedded in a class of smooth perturbations ϕ_σ, where $\sigma \geq 0$ is a parameter. For $\sigma > 0$ the function ϕ_σ is smooth, whereas $\phi_0 = \phi$. For the Fischer–Burmeister function ϕ_{FB}, e.g., the functions ϕ_σ can be obtained by adding the term σ under the square root. The main idea of these methods, transcribed to our setting, consists of following the trajectory of solutions to the corresponding perturbed operator equations $\Phi_\sigma(y) = 0$ as $\sigma \to 0$. Usually corrector steps are computed by Newton's method. In the asymptotic phase $\sigma \to 0$, the behavior of Newton's method on the unperturbed equation plays a key role in achieving fast local convergence. We therefore believe that the results presented in this paper will also be helpful to investigate path-following methods in a function space setting.

We emphasize that the number of applications fitting into our framework is huge, in particular those involving complementarity; see [15, 18, 21, 26, 28, 39, 46, 48]. Many of these applications arise from infinite-dimensional variational inequalities that model systems continuous in time and/or space [15, 18, 26, 39, 46], and are therefore posed in function spaces. Hence, the development and analysis of efficient abstract algorithms for the solution of the infinite-dimensional problem (1.5) is very desirable in order to derive robust, efficient, and mesh-independent methods for the solution of the discretized problem. The nonsmooth Newton method developed in this paper is directly applicable to NCP-function-based reformulations of the NCP (1.5) and can therefore be seen as a generalization of semismooth Newton methods for finite-

dimensional NCPs.

For the development of a semismoothness concept we have to choose an appropriate vector-valued generalized differential for the operator Ψ. Although the available literature on generalized differentials and subdifferentials is mainly focused on real-valued functions (see, e.g., [10, 13, 14, 55] and the references therein), several authors have proposed and analyzed generalized differentials for nonlinear operators between infinite-dimensional spaces [16, 25, 32, 53, 58]. In our approach, we work with a generalized differential that exploits the structure of Ψ. Roughly speaking, our general guidance hereby is to transcribe, at least formally, componentwise operations in $\mathbb{R}^k$ to pointwise operations in function spaces. To sketch the idea, note that the finite-dimensional analogue of the operator Ψ is the mapping

$$\Psi^{\mathsf{f}} : \mathbb{R}^k \to \mathbb{R}^l, \quad \Psi^{\mathsf{f}}_j(x) = \psi(F^j(x)), \quad j = 1, \dots, l,$$

with ψ as above and C^1-mappings $F^j : \mathbb{R}^k \to \mathbb{R}^m$. We have the correspondences $\omega \in \Omega \leftrightarrow j \in \{1, \dots, l\}$, $y \in Y \leftrightarrow x \in \mathbb{R}^k$, and $F(y)(\omega) \leftrightarrow F^j(x)$. Componentwise application of the chain rule for Clarke's generalized gradient [13] shows that the C-subdifferential of Ψ^{f} consists of matrices $M \in \mathbb{R}^{l \times k}$ having rows of the form

$$M_j = \sum_{i=1}^{m} d_i^j (F_i^j)'(x), \quad \text{with} \quad d^j \in \partial\psi(F^j(x)).$$

Note that the collection of all these matrices M can be an overestimate of the C-subdifferential, since the chain rule asserts only that $\partial[\psi(F^j(x))] \subset \partial\psi(F^j(x))(F^j)'(x)$. Carrying out the same construction for Ψ in a purely formal manner suggests that we choose a generalized differential for Ψ consisting of operators of the form

$$v \in Y \mapsto \sum_{i=1}^{m} d_i \cdot (F_i'(x)v), \quad \text{with} \quad (d_1, \dots, d_m)(\omega) \in \partial\psi(F(y)(\omega)) \text{ a.e. on } \Omega,$$

where the inclusion on the right is meant in the sense of measurable selections. One advantage of this approach, which motivates our choice of the generalized differential $\partial^\circ\Psi$, is that it consists of relatively "concrete" objects as compared to those investigated in, e.g., [16, 25, 32, 53, 58], which necessarily are more abstract since they are not restricted to a particular structure of the underlying operator. It is not the objective of this paper to investigate the connections between the generalized differential $\partial^\circ\Psi$ and other generalized differentials. There are close relationships, but we leave it as a topic for future research. Here, we concentrate on the development of a semismoothness concept based on $\partial^\circ\Psi$, a related nonsmooth Newton's method, and the relations to the respective finite-dimensional analogues.

As already mentioned, the literature on Newton-like methods for the solution of NCPs or, closely related, bound-constrained optimization problems posed in function spaces is very limited. Here, we call an iteration Newton-like if each iteration essentially requires the solution of a linear operator equation. We point out that in this sense sequential quadratic programming (SQP) methods for problems involving inequality constraints [1, 2, 3, 4, 5, 29, 59] are not Newton-like, since each iteration requires the solution of a quadratic programming problem (or, put differently, a linearized generalized equation), which is in general significantly more expensive than solving a linear operator equation. Therefore, the methods considered in this paper, rather than being applied directly to the nonlinear problem, could also be of interest as subproblem solvers for SQP methods.

Probably the prior investigations most closely related to ours are the analysis of Bertsekas' projected Newton method by Kelley and Sachs [38] and the investigation

of affine-scaling interior-point Newton methods by Ulbrich and Ulbrich [64]. Both papers deal with bound-constrained minimization problems in function spaces and establish the local q-superlinear convergence of their respective Newton-like methods. In both approaches the convergence results are obtained by directly estimating the remainder terms appearing in the analysis of the Newton iteration. In that way, specific properties of the solution are exploited, and a strict complementarity condition is assumed in both papers. During the revision of our paper, a very similar investigation by Hintermüller, Ito, and Kunisch [30] came to our attention.[1] Therein, linear complementarity problems motivated by constrained optimal control problems are considered. It is shown that the primal dual active set strategy [8, 9] can be interpreted as a semismooth Newton method of the form investigated in the present paper. In [30], fast local convergence is proved by a direct analysis, and, in its recent revision [31], a second proof is given by applying the semismoothness results that we develop in the current work. The general framework of our presentation requires, except in special situations, that we augment the nonsmooth Newton iteration by a smoothing step. The algorithm in [30] does not require a smoothing step, since, in our terminology, a special NCP-function is used and the underlying operator has a smoothing property. The idea here is very similar to the construction of smoothing steps described in Example 6.7 and is discussed in Remark 6.8; see also [31]. We refer the interested reader to [62, Chapter 4], where we develop a general class of smoothing-step-free Newton methods for VIPs.

In the present paper, we develop our results for the general problem class (1.10) and derive the applicability to NCPs as a simple, but important special case. In the context of NCPs and optimization, we do not have to assume any strict complementarity condition. Further, we organize our analysis of Newton's methods by decomposing it in two steps: First, we develop a semismoothness result that replaces differentiability in ordinary Newton methods. Second, an invertibility condition on the members of the generalized differential is introduced. This regularity condition can be verified conveniently by using the sufficient conditions that we recently developed in [63, 62]¿

In section 2 we review some concepts of finite-dimensional nonsmooth analysis that are important in our context, in particular, generalized differentials and semismoothness. Our working assumptions are stated in section 3. In section 4 we introduce the generalized differential $\partial^\circ \Psi$ and investigate some of its properties. In section 5 a semismoothness and α-order semismoothness concept for the operator Ψ is proposed and studied in detail. The results are illustrated by applications to NCPs. In particular, we demonstrate the necessity of our assumptions by several (counter-) examples. In section 6 we propose a Newton-like method for the solution of the nonsmooth operator equation (1.10) and use our semismoothness results to establish its q-superlinear convergence. In the case of an α-order semismooth operator Ψ, we prove convergence of q-order $1+\alpha$. Applications to NCPs are provided as illustrative examples, and the computation of smoothing steps is discussed. We also show how to avoid smoothing steps in certain situations. Furthermore, we consider the application of the semismooth Newton method to the elliptic control problem (1.6) and address its discretization. In section 7 we show that under appropriate assumptions the composition of semismooth operators is again semismooth, and we develop two chain rules. Finally, in section 8, we establish some further properties of our generalized differential.

Notation. Given a Banach space Y, we denote by $\|\cdot\|_Y$ its norm, by B_Y its open unit ball, and by $\bar{B}_Y$ its closed unit ball; in the special case $Y = (\mathbb{R}^n, \|\cdot\|_p)$, we prefer

[1] The author is thankful to Michael Hintermüller for sending him the paper [30] and the revised manuscript [31].

to write B_p^n and $\bar{B}_p^n$, respectively. On a product space $\prod_i Y_i$, we choose $\|y\|_{\Pi_i Y_i} = \sum_i \|y\|_{Y_i}$ as a norm. $\mathcal{L}(Y,Z)$ denotes the Banach space of bounded linear operators from the Banach space Y to the Banach space Z, equipped with the operator norm $\|\cdot\|_{Y,Z}$. By $\langle v, w\rangle_\Omega$ we denote the dual pairing between $v \in L^p(\Omega)$ and $w \in L^{p'}(\Omega)$, $1/p + 1/p' = 1$. The indicator function of a measurable set $Q \subset \Omega$, taking the value one on Q and zero on its complement $Q^c = \Omega \setminus Q$, is denoted by $\mathbf{1}_Q$. We write μ for the Lebesgue measure on $\mathbb{R}^n$. Given a function $w \in L^\infty(\Omega)$ and an operator $A \in \mathcal{L}(Y, L^p(\Omega))$, we define the operator $w \cdot A \in \mathcal{L}(Y, L^p(\Omega))$ that takes $y \in Y$ to the function $\omega \in \Omega \mapsto w(\omega)(Ay)(\omega)$. The Fréchet derivative of an operator H is denoted by H'. For convenience, we will write $\sum_i$ and $\prod_i$ instead of $\sum_{i=1}^m$ and $\prod_{i=1}^m$.

2. Generalized differentials and semismoothness in finite dimensions. We begin with an overview of the semismoothness concept in finite dimensions. Let the vector-valued function $f : \mathbb{R}^k \to \mathbb{R}^l$ be given. We first collect some notions from nonsmooth analysis. Assume that f is locally Lipschitz continuous. According to Rademacher's theorem, the set $U_f \subset \mathbb{R}^k$ of all points x at which f fails to be differentiable is a Lebesgue null set. Here, the fact that f is a mapping between finite-dimensional spaces is crucial. Using this, generalized Jacobians can be constructed as follows.

DEFINITION 2.1. *Let f be locally Lipschitz. We define the following generalized Jacobians of f at x:*

(a) *the Bouligand (B-) subdifferential:*

$$\partial_B f(x) \stackrel{\text{def}}{=} \left\{ M \in \mathbb{R}^{l\times k} : \exists (x_j) \subset \mathbb{R}^k \setminus U_f :\ x_j \to x,\ f'(x_j) \to M \right\},$$

where f' denotes the Jacobian of f,

(b) *Clarke's generalized Jacobian, the convex hull of $\partial_B f(x)$:*

$$\partial f(x) \stackrel{\text{def}}{=} co\, \partial_B f(x),$$

(c) *Qi's C-subdifferential:* $\partial_C f(x) \stackrel{\text{def}}{=} \partial f_1(x) \times \cdots \times \partial f_l(x)$.

These generalized differentials induce multifunctions $\partial_B f, \partial f, \partial_C f : \mathbb{R}^k \rightrightarrows \mathbb{R}^{l\times k}$. They have the following properties:

(1) $\partial_B f$, ∂f, and $\partial_C f$ are nonempty and compact-valued. Moreover, ∂f and $\partial_C f$ are convex-valued.

(2) The multifunctions $\partial_B f$, ∂f, and $\partial_C f$ are upper semicontinuous (see Definition A.2 or [13, p. 29]).

(3) $\partial_B f(x) \subset \partial f(x) \subset \partial_C f(x)$ for all x.

Based on Clarke's generalized Jacobian, Qi [50] and Qi and Sun [52] introduced the following notion of semismoothness.

DEFINITION 2.2. *f is* semismooth *at $x \in \mathbb{R}^k$ if it is locally Lipschitz and, for all $h \in \mathbb{R}^k$, the limit*

$$\lim_{\substack{M\in\partial f(x+th')\\ h'\to h, t\to 0^+}} Mh'$$

exists and is finite.

The following characterization, however, is more appropriate for our purposes.

PROPOSITION 2.3. *Let f be locally Lipschitz. Then f is semismooth at x if and only if f is directionally differentiable at x and*

$$\max_{M\in\partial f(x+h)} \|f(x+h) - f(x) - Mh\|_2 = o(\|h\|_2) \quad \text{as } h \to 0. \tag{2.1}$$

Proof. In [52, Theorem 2.3] it is shown that the locally Lipschitz continuous function f is semismooth at x if and only if f is directionally differentiable at x and

$$\max_{M\in\partial f(x+h)} \|Mh - f'(x,h)\|_2 = o(\|h\|_2) \quad \text{as } h \to 0. \tag{2.2}$$

Furthermore, since f is locally Lipschitz continuous on the finite-dimensional space $\mathbb{R}^k$, directional differentiability implies B-differentiability (see [56]):

$$\|f(x+h) - f(x) - f'(x,h)\|_2 = o(\|h\|_2) \quad \text{as } h \to 0. \tag{2.3}$$

It is now straightforward to see that, under (2.3), the conditions (2.1) and (2.2) are equivalent. ∎

DEFINITION 2.4. *f is* α-order semismooth*, $0 < \alpha \le 1$, at $x \in \mathbb{R}^k$ if it is locally Lipschitz and directionally differentiable at x, and if*

$$\max_{M\in\partial f(x+h)} \|Mh - f'(x,h)\|_2 = O\big(\|h\|_2^{1+\alpha}\big) \quad \textit{as } h \to 0.$$

The following consequence of α-order semismoothness will be important.

PROPOSITION 2.5 (see [23, Lemmas 2 and 17]). *Let f be α-order semismooth at x, $0 < \alpha \le 1$. Then*

$$\max_{M\in\partial f(x+h)} \|f(x+h) - f(x) - Mh\|_2 = O(\|h\|_2^{1+\alpha}) \qquad \textit{as } h \to 0, \tag{2.4}$$

$$\|f(x+h) - f(x) - f'(x,h)\|_2 = O\big(\|h\|_2^{1+\alpha}\big) \qquad \textit{as } h \to 0. \tag{2.5}$$

It is obvious that useful semismoothness concepts can also be obtained by replacing ∂f by other suitable generalized derivatives. This was investigated in a general framework by Jeyakumar [33, 34] and by Xu [66, 67]. Here, we sketch only Jeyakumar's approach, in which he introduced the concept of $\partial^* f$-semismoothness, where $\partial^* f$ is an approximate Jacobian [35]. For the definition of approximate Jacobians, we refer to [35]; in what follows, it is sufficient to know that an approximate Jacobian of $f : \mathbb{R}^k \mapsto \mathbb{R}^l$ is a closed-valued multifunction $\partial^* f : \mathbb{R}^k \rightrightarrows \mathbb{R}^{l\times k}$ with nonempty values and that $\partial_B f$, ∂f, and $\partial_C f$ are approximate Jacobians.

DEFINITION 2.6. *Let $f : \mathbb{R}^k \mapsto \mathbb{R}^l$ be continuous, and let an approximate Jacobian $\partial^* f$ of f be given.*

(a) *The function f is called* weakly $\partial^* f$-semismooth *at x if*

$$\sup_{M\in\overline{co}\,\partial^* f(x+h)} \|f(x+h) - f(x) - Mh\|_2 = o(\|h\|_2) \quad \textit{as } h \to 0. \tag{2.6}$$

(b) *The function f is* $\partial^* f$-semismooth *at x if*

(i) *f is B-differentiable at x (e.g., locally Lipschitz near x and directionally differentiable at x; see* [56]*) and*

(ii) *f is weakly $\partial^* f$-semismooth at x.*

Note that ∂f-semismoothness coincides with semismoothness. Obviously, we can define weak $\partial^* f$-semismoothness of order α by requiring the order $O(\|h\|_2^{1+\alpha})$ in (2.6).

Finally, we consider a Newton-like method for the solution of the nonsmooth equation

$$f(x) = 0, \tag{2.7}$$

where $f : \mathbb{R}^k \to \mathbb{R}^k$ is weakly $\partial^* f$-semismooth or weakly $\partial^* f$-semismooth of the order α, respectively, at the solution $\bar{x}$. For this system of equations, Newton-like methods

were developed that converge locally q-superlinearly [33, 49, 50, 52]; see also [40, 41]. A representative result is the following.

PROPOSITION 2.7. *Denote by $\bar{x} \in \mathbb{R}^k$ a solution of (2.7), and let the initial point $x_0 \in \mathbb{R}^k$ be given. Consider the following Newton-like iteration:*

For $j = 0, 1, 2, \ldots$ as long as $f(x_j) \neq 0$:
Choose $M_j \in \partial^* f(x_j)$ and compute $x_{j+1} = x_j + s_j$, where

$$M_j s_j = -f(x_j).$$

Assume that

(a) *f is weakly $\partial^* f$-semismooth (or weakly $\partial^* f$-semismooth of the order α) at $\bar{x}$.*
(b) *There exist $\eta > 0$ and $C > 0$ such that, for all $x \in \bar{x} + \eta B_2^k$, every $M \in \partial^* f(x)$ is nonsingular with $\|M^{-1}\|_2 \leq C$ (regularity assumption).*

Then there exists $\delta > 0$ such that for all $x_0 \in \bar{x} + \delta B_2^k$ the above iteration either terminates with $x_j = \bar{x}$ or generates a sequence (x_j) that converges q-superlinearly (or with q-order $1 + \alpha$) to $\bar{x}$.

Proof. As long as $x_j \in \bar{x} + \eta B_2^k$, the iteration is well defined by (b). Setting $e_j = x_j - \bar{x}$ and using $f(\bar{x}) = 0$, we have

$$M_j e_{j+1} = M_j s_j + M_j e_j = -f(x_j) + M_j e_j = f(\bar{x}) - f(\bar{x} + e_j) + M_j e_j.$$

This, (a), and (b) yield

$$\|e_{j+1}\|_2 \leq \|M_j^{-1}\|_2 \, \|f(\bar{x} + e_j) - f(\bar{x}) - M_j e_j\|_2 = o(\|e_j\|_2) \quad \text{as } x_j \to \bar{x}. \tag{2.8}$$

By (a) we can choose $\delta \in (0, \eta]$ so small that

$$\|f(\bar{x} + h) - f(\bar{x}) - Mh\|_2 \leq \frac{\|h\|_2}{2C} \quad \text{for all } M \in \partial^* f(\bar{x} + h) \text{ and all } h \in \delta B_2^k. \tag{2.9}$$

Note that this holds trivially for $h = 0$. Hence, for all $x_j \in \bar{x} + \delta B_2^k$ with $x_j \neq \bar{x}$, we have $\|e_{j+1}\|_2 \leq \|e_j\|_2 / 2$ by (2.8), and thus $x_{j+1} \in \bar{x} + (\|e_j\|_2 / 2) B_2^k \subset \bar{x} + (\delta/2) B_2^k$. Inductively, we conclude that for all $x_0 \in \bar{x} + \delta B_2^k$ the algorithm is well defined and either terminates finitely or generates a sequence (x_j) converging to $\bar{x}$. In the case of finite termination, we have $f(x_j) = 0$ and, by (2.9) and the choice of δ, we see that, for any $M \in \overline{\text{co}} \partial^* f(\bar{x} + e_j)$,

$$\frac{\|e_j\|_2}{2} \geq C \, \|f(x_j) - f(\bar{x}) - M e_j\|_2 \geq \|M^{-1}\|_2 \, \|M e_j\|_2 \geq \|e_j\|_2 ;$$

hence $x_j = \bar{x}$. On the other hand, if the algorithm generates an infinite sequence $x_j \to \bar{x}$, then we see from (2.8) that the rate of convergence is q-superlinear. If f is weakly $\partial^* f$-semismooth of order α at $\bar{x}$, then we can improve the order in (2.8) to $O(\|e_j\|_2^{1+\alpha})$ and obtain convergence with q-rate $1 + \alpha$. ∎

Remark 2.8. In many cases, the approximate Jacobian is upper semicontinuous and compact-valued, particularly if $\partial_B f$, ∂f, or $\partial_C f$ are used. Then it is easy to show that the regularity condition of Proposition 2.7(b) is already satisfied if all $M \in \partial^* f(\bar{x})$ are nonsingular.

3. Assumptions. In the rest of the paper, we will impose the following assumptions on F and ψ.

Assumption 3.1. There are $1 \leq r \leq r_i < q_i \leq \infty$, $1 \leq i \leq m$, such that
(a) the operator $F : Y \to \prod_i L^{r_i}(\Omega)$ is continuously Fréchet differentiable;

(b) the mapping $y \in Y \mapsto F(y) \in \prod_i L^{q_i}(\Omega)$ is locally Lipschitz continuous—i.e., for all $y \in Y$ there exists an open neighborhood $U = U(y)$ and a constant $L_F = L_F(U)$ such that

$$\sum\nolimits_i \|F_i(y_1) - F_i(y_2)\|_{L^{q_i}} \le L_F \|y_1 - y_2\|_Y \quad \text{for all } y_1, y_2 \in U;$$

(c) the function $\psi : \mathbb{R}^m \to \mathbb{R}$ is Lipschitz continuous of rank $L_\psi > 0$—i.e.,

$$|\psi(x_1) - \psi(x_2)| \le L_\psi \|x_1 - x_2\|_1 \quad \text{for all } x_1, x_2 \in \mathbb{R}^m;$$

(d) ψ is semismooth.

Remark 3.2. Since by assumption the domain Ω is bounded, we have the continuous embedding $L^q(\Omega) \subset L^p(\Omega)$ whenever $1 \le p \le q \le \infty$.

Remark 3.3. Note that in Assumption 3.1 the only difference between the operators in (a) and (b) is the topology of the range space. As mentioned in Remark 3.2, the L^{q_i}-norms are stronger than the corresponding L^{r_i}-norms.

For semismoothness of order > 0 we will strengthen Assumption 3.1 to obtain the following.

Assumption 3.4. As Assumption 3.1, but with (a) and (d) replaced as follows: There exists $\alpha \in (0, 1]$ such that

(a′) the operator $F : Y \to \prod_i L^{r_i}(\Omega)$ is α-order Hölder continuously Fréchet differentiable;

(d′) ψ is α-order semismooth.

Note that for the special case $Y = \prod_i L^{q_i}(\Omega)$ and $F = \mathrm{id}_Y$ we have

$$\Psi : y \in Y \mapsto \psi(y),$$

and it is easily seen that Assumptions 3.1 or 3.4, respectively, reduce to (c) and (d) or (c) and (d′), respectively.

Under Assumption 3.1, the operator Ψ defined in (1.8) is well defined and locally Lipschitz continuous.

PROPOSITION 3.5. *Let Assumption* 3.1 *hold. Then for all* $1 \le q \le q_i$, $1 \le i \le m$, *and thus in particular for* $q = r$, *the operator* Ψ *defined in* (1.8) *maps* Y *locally Lipschitz continuously into* $L^q(\Omega)$.

Proof. Using Lemma A.1, we first prove $\Psi(Y) \subset L^q(\Omega)$, which follows from

$$\begin{aligned}\|\Psi(y)\|_{L^q} = \|\psi(F(y))\|_{L^q} &\le \|\psi(0)\|_{L^q} + \|\psi(F(y)) - \psi(0)\|_{L^q} \\ &\le c_{q,\infty}(\Omega)|\psi(0)| + L_\psi \sum\nolimits_i \|F_i(y)\|_{L^q} \\ &\le c_{q,\infty}(\Omega)|\psi(0)| + L_\psi \sum\nolimits_i c_{q,q_i}(\Omega) \|F_i(y)\|_{L^{q_i}}.\end{aligned}$$

To establish the local Lipschitz continuity, denote by L_F the local Lipschitz constant in Assumption 3.1(b) on the set U, and let $y_1, y_2 \in U$ be arbitrary. Then, again by Lemma A.1,

$$\begin{aligned}\|\Psi(y_1) - \Psi(y_2)\|_{L^q} &\le L_\psi \sum\nolimits_i \|F_i(y_1) - F_i(y_2)\|_{L^q} \\ &\le L_\psi \sum\nolimits_i c_{q,q_i}(\Omega) \|F_i(y_1) - F_i(y_2)\|_{L^{q_i}} \\ &\le L_\psi L_F \left(\max_{1 \le i \le m} c_{q,q_i}(\Omega) \right) \|y_1 - y_2\|_Y. \qquad \square\end{aligned}$$

4. An infinite-dimensional generalized differential. For the development of a semismoothness concept for the operator Ψ defined in (1.8), we have to choose

an appropriate generalized differential. As we already mentioned in the introduction, our aim is to work with a differential that is as closely connected to finite-dimensional generalized Jacobians as possible. Hence, we will propose a generalized differential $\partial^\circ\Psi$ in such a way that its natural finite-dimensional discretization contains Qi's C-subdifferential; see section 6.2.

Our construction is motivated by a formal pointwise application of the chain rule. In fact, suppose for the moment that the operator $y \in Y \mapsto F(y) \in C(\bar\Omega)^m$ is strictly differentiable, where $C(\bar\Omega)$ denotes the space of continuous functions equipped with the max-norm. Then for fixed $\omega \in \Omega$ the function $f : y \mapsto F(y)(\omega)$ is strictly differentiable with derivative $f'(y) \in \mathcal{L}(Y, \mathbb{R}^m)$,

$$f'(y) : v \mapsto (F'(y)v)(\omega).$$

The chain rule for generalized gradients [13, Theorem 2.3.10], applied to the real-valued mapping $y \mapsto \Psi(y)(\omega) = \psi(f(y))$, yields

(4.1)
$$\partial\big(\Psi(y)(\omega)\big) \subset \partial\psi\big(f(y)\big) \circ f'(y) = \left\{ g \in Y^* \;\middle|\; \begin{array}{l} \langle g, v\rangle = \sum_i d_i(\omega)\big(F_i'(y)v\big)(\omega), \\ d(\omega) \in \partial\psi\big(F(y)(\omega)\big) \end{array} \right\}.$$

Furthermore, we can replace "$\subset$" by "$=$" if ψ or $-\psi$ is regular (e.g., if ψ is convex or concave) or if the linear operator $f'(y)$ is onto; see [13, Theorem 2.3.10]. Following the above motivation and returning to the general setting of Assumption 3.1, we define the generalized differential $\partial^\circ\Psi(y)$ in such a way that for all $M \in \partial^\circ\Psi(y)$ the linear form $v \mapsto (Mv)(\omega)$ is an element of the right-hand side in (4.1), as follows.

Definition 4.1 (generalized differential $\partial^\circ\Psi$). *Let Assumption* 3.1 *hold. For* Ψ *as defined in* (1.8) *we define the generalized differential* $\partial^\circ\Psi : Y \rightrightarrows \mathcal{L}(Y, L^r)$,

$$\partial^\circ\Psi(y) \stackrel{\text{def}}{=} \left\{ M \in \mathcal{L}(Y, L^r) \;\middle|\; \begin{array}{l} M : v \mapsto \sum_i d_i \cdot \big(F_i'(y)v\big), \\ d \text{ measurable selection of } \partial\psi\big(F(y)\big) \end{array} \right\}. \tag{4.2}$$

Remark 4.2. The superscript "$\circ$" is chosen to indicate that this generalized differential is designed for superposition operators.

The generalized differential $\partial^\circ\Psi(y)$ is nonempty. To show this, we first prove the next claim.

Lemma 4.3. *Let the Assumption* 3.1(a) *hold, and let* $d \in L^\infty(\Omega)^m$ *be arbitrary. Then the operator*

$$M : v \in Y \mapsto \sum\nolimits_i d_i \cdot \big(F_i'(y)v\big)$$

is an element of $\mathcal{L}(Y, L^r)$, *and*

$$\|M\|_{Y,L^r} \le \sum\nolimits_i c_{r,r_i}(\Omega)\, \|d_i\|_{L^\infty}\, \|F_i'(y)\|_{Y,L^{r_i}}. \tag{4.3}$$

Proof. By Assumption 3.1(a) and Lemma A.1

$$\begin{aligned} \|Mv\|_{L^r} = \Big\|\sum\nolimits_i d_i \cdot \big(F_i'(y)v\big)\Big\|_{L^r} &\le \sum\nolimits_i \|d_i\|_{L^\infty}\, \|F_i'(y)v\|_{L^r} \\ &\le \Big(\sum\nolimits_i c_{r,r_i}(\Omega)\, \|d_i\|_{L^\infty}\, \|F_i'(y)\|_{Y,L^{r_i}}\Big)\, \|v\|_Y \qquad \text{for all } v \in Y, \end{aligned}$$

which shows that (4.3) holds and $M \in \mathcal{L}(Y, L^r)$. ∎

In the next step, we show that the multifunction

$$\partial\psi(F(y)) : \omega \in \Omega \mapsto \partial\psi(F(y)(\omega)) \subset \mathbb{R}^m$$

is measurable (see Definition A.3 or [54, p. 160]).

LEMMA 4.4. *Any closed-valued, upper semicontinuous multifunction* $\Gamma : \mathbb{R}^k \rightrightarrows \mathbb{R}^l$ *is Borel measurable.*

Proof. Let $C \subset \mathbb{R}^l$ be compact. We show that $\Gamma^{-1}(C)$ is closed. To this end, let $x_k \in \Gamma^{-1}(C)$ be arbitrary with $x_k \to x^*$. Then there exist $z_k \in \Gamma(x_k) \cap C$, and, due to the compactness of C, we achieve by transition to a subsequence that $z_k \to z^* \in C$. Since $x_k \to x^*$, upper semicontinuity yields that there exist $\hat{z}_k \in \Gamma(x^*)$ with $(z_k - \hat{z}_k) \to 0$ and thus $\hat{z}_k \to z^*$. Therefore, since $\Gamma(x^*)$ is closed, we obtain $z^* \in \Gamma(x^*) \cap C$. Hence, $x^* \in \Gamma^{-1}(C)$, which proves that $\Gamma^{-1}(C)$ is closed and therefore a Borel set. ∎

COROLLARY 4.5. *The multifunction* $\partial\psi(F(y)) : \Omega \rightrightarrows \mathbb{R}$ *is measurable.*

Proof. By Lemma 4.4, the compact-valued and upper semicontinuous multifunction $\partial\psi$ is Borel measurable. Now, for all closed sets $C \subset \mathbb{R}^m$, we have, setting $u = F(y) \in \prod_i L^{r_i}(\Omega)$,

$$\partial\psi(F(y))^{-1}(C) = u^{-1}(\partial\psi^{-1}(C)).$$

This set is measurable since $\partial\psi^{-1}(C)$ is a Borel set and u is a (class of equivalent) measurable function(s). ∎

The next result is a direct consequence of Lipschitz continuity; see [13, 2.1.2].

LEMMA 4.6. *Under Assumption* 3.1(c) *there holds* $\partial\psi(x) \subset [-L_\psi, L_\psi]^m$ *for all* $x \in \mathbb{R}^m$.

Combining this with Corollary 4.5 yields the following.

LEMMA 4.7. *Let Assumption* 3.1 *hold. Then for all* $y \in Y$, *the set*

$$K(y) = \{d : \Omega \to \mathbb{R}^m : d \text{ measurable selection of } \partial\psi(F(y))\} \tag{4.4}$$

is a nonempty subset of $L_\psi \bar{B}^m_{L^\infty} \subset L^\infty(\Omega)^m$.

Proof. By the theorem on measurable selections [54, Corollary 1C] and Corollary 4.5, $\partial\psi(F(y))$ admits at least one measurable selection $d : \Omega \to \mathbb{R}^m$, i.e.,

$$d(\omega) \in \partial\psi(F(y)(\omega)) \quad \text{a.e. on } \Omega.$$

From Lemma 4.6 it follows that $d \in L_\psi \bar{B}^m_{L^\infty}$. ∎

We now can prove the next result.

PROPOSITION 4.8. *Under Assumption* 3.1, *for all* $y \in Y$ *the generalized differential* $\partial^\circ\Psi(y)$ *is nonempty and bounded in* $\mathcal{L}(Y, L^r)$.

Proof. Lemma 4.7 ensures that there exist measurable selections d of $\partial\psi(F(y))$ and that all these d are contained in $L_\psi \bar{B}^m_{L^\infty}$. Hence, Lemma 4.3 shows that

$$M : v \mapsto \sum\nolimits_i d_i \cdot \big(F_i'(y)v\big)$$

is in $\mathcal{L}(Y, L^r)$. The boundedness of $\partial^\circ\Psi(y)$ follows from (4.3). ∎

We now have everything at hand for introducing a semismoothness concept that is based on the generalized differential $\partial^\circ\Psi$. We postpone the investigation of further properties of $\partial^\circ\Psi$ to sections 7 and 8. There, we will establish chain rules, the convex-valuedness, weak compact-valuedness, and the weak graph closedness of $\partial^\circ\Psi$.

5. Semismoothness in function spaces. In this section, we develop a semismoothness concept for the operator Ψ defined in (1.8). Our notion of semismoothness is similar to Jeyakumar's weak semismoothness in Definition 2.6(a). In place of the finite-dimensional approximate Jacobian, we work with the generalized differential $\partial^\circ\Psi$. Since we will show in Theorem 8.1 that $\partial^\circ\Psi$ is convex and closed (even compact) in the weak operator topology, there is no need to take the closed convex hull of $\partial^\circ\Psi$ as is done in (2.6).

DEFINITION 5.1. *The operator* Ψ *is* semismooth *at* $y \in Y$ *if*

$$\sup_{M \in \partial^\circ \Psi(y+s)} \|\Psi(y+s) - \Psi(y) - Ms\|_{L^r} = o(\|s\|_Y) \quad \text{as } s \to 0 \text{ in } Y. \tag{5.1}$$

Ψ *is* α-order semismooth, $0 < \alpha \le 1$, *at* $y \in Y$ *if*

$$\sup_{M \in \partial^\circ \Psi(y+s)} \|\Psi(y+s) - \Psi(y) - Ms\|_{L^r} = O\big(\|s\|_Y^{1+\alpha}\big) \quad \text{as } s \to 0 \text{ in } Y. \tag{5.2}$$

This definition is easily extended to general operators between Banach spaces. Of course, an appropriate generalized differential must be available. In this paper, we only deal with the superposition operator Ψ, and thus we dispense with a more general definition of semismoothness.

In the following main theorem we establish the semismoothness and the β-order semismoothness, respectively, of the operator Ψ.

THEOREM 5.2.

(a) *Under Assumption* 3.1*, the operator* Ψ *is semismooth.*

(b) *Let Assumption* 3.4 *hold. Assume that there exists a* $\gamma > 0$ *such that the set*

$$\Omega_\varepsilon = \left\{ \omega : \max_{\|h\|_1 \le \varepsilon} \Big(\rho\big(F(y)(\omega), h\big) - \varepsilon^{-\alpha} \|h\|_1^{1+\alpha} \Big) > 0 \right\}, \quad \varepsilon > 0,$$

with the residual function $\rho : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ *given by*

$$\rho(x,h) = \max_{z^T \in \partial\psi(x+h)} |\psi(x+h) - \psi(x) - z^T h|,$$

has the following decrease property:

$$\mu(\Omega_\varepsilon) = O(\varepsilon^\gamma) \quad \text{as } \varepsilon \to 0^+. \tag{5.3}$$

Then the operator Ψ *is* β*-order semismooth at* y *with*

$$\begin{aligned} &\beta = \min\left\{ \frac{\gamma\nu}{1+\gamma/q_0}, \frac{\alpha\gamma\nu}{\alpha+\gamma\nu} \right\}, \quad \text{where} \\ &q_0 = \min_{1 \le i \le m} q_i, \qquad \nu = \frac{q_0 - r}{q_0 r} \quad \text{if } q_0 < \infty, \qquad \nu = \frac{1}{r} \quad \text{if } q_0 = \infty. \end{aligned} \tag{5.4}$$

The proof of this theorem will be presented in section 5.1.

Remark 5.3. Condition (5.3) requires the measurability of the set Ω_ε, which will be verified in the proof. We also remark that the α-order semismoothness of ψ implies $\mu(\Omega_\varepsilon) \to 0$ as $\varepsilon \to 0$; see the discussion after Remark 5.4.

Remark 5.4. As we will see in Lemma 5.8, it would be sufficient to require only the β-order Hölder continuity of F' in Assumption 3.4(a′), with $\beta \le \alpha$ as defined in (5.4).

It might be helpful to give an explanation of the abstract condition (5.3) here. For convenient notation, let $x = F(y)(\omega)$. Due to the α-order semismoothness of ψ provided by Assumption 3.4, we have $\rho\big(x,h\big) = O(\|h\|_1^{1+\alpha})$ as $h \to 0$; see Proposition 2.5. In essence, Ω_ε is the set of all $\omega \in \Omega$ where there exists $h \in \varepsilon \bar{B}_1^m$ for which this asymptotic behavior is not yet observed, because the remainder term $\rho\big(x,h\big)$ exceeds $\|h\|_1^{1+\alpha}$ by a factor of at least $\varepsilon^{-\alpha}$, which grows infinitely as $\varepsilon \to 0$. From the continuity of the Lebesgue measure it is clear that $\mu(\Omega_\varepsilon) \to 0$ as $\varepsilon \to 0$. The decrease condition (5.3) essentially states that the measure of the set Ω_ε where $F(y)$ takes "bad values," i.e., values at which the radius of small residual is very small, decreases with the rate ε^γ.

The following Example 5.5 demonstrates the applicability of Theorem 5.2 to NCPs. It also provides a very concrete interpretation of condition (5.3).

Example 5.5 (application to NCPs). The reformulation of NCPs (1.5) in the form (1.7) leads to an important special case of the operator equations (1.10) under consideration. Let the operator $Z : L^p(\Omega) \to L^r(\Omega)$, $1 \le r < p \le \infty$, be given, and consider the NCP (1.5), which we restate for convenience:

$$y \ge 0, \quad Z(y) \ge 0, \quad yZ(y) = 0. \tag{5.5}$$

In section 1 we showed that we can apply an NCP-function ϕ to transform (5.5) into the equivalent operator equation

$$\Phi(y) = 0, \quad \text{where} \quad \Phi(y)(\omega) = \phi(y(\omega), Z(y)(\omega)), \quad \omega \in \Omega. \tag{5.6}$$

We now view the operator Φ as a special case of the more general class of operators Ψ defined in (1.8) and interpret Assumptions 3.1 and 3.4 in this context. To this end, we choose $Y = L^p(\Omega)$, set $r_1 = r_2 = r$, and define

$$F : y \in Y \mapsto (y, Z(y)) \in L^{r_1}(\Omega) \times L^{r_2}(\Omega).$$

Then (5.6) is equivalent to (1.10) with $\psi = \phi$. Assume that

(1) the operator $Z : L^p(\Omega) \to L^r(\Omega)$ is continuously Fréchet differentiable;
(2) there is $q \in (r, \infty]$ such that $Z : L^p(\Omega) \to L^q(\Omega)$ is locally Lipschitz continuous;
(3) ϕ is Lipschitz continuous;
(4) ϕ is semismooth.

Then Assumption 3.1 is satisfied with $q_1 = p$ and $q_2 = q$. In fact, (1) and the continuous embedding $L^p(\Omega) \subset L^r(\Omega)$ imply Assumption 3.1(a). Further, (2) and the Lipschitz continuity of the identity $u \in L^p(\Omega) \mapsto u \in L^p(\Omega)$ yield Assumption 3.1(b). Finally, (3) and (4) imply Assumption 3.1(c)–(d). Therefore, we can apply Theorem 5.2 and obtain that Φ is semismooth:

$$\sup_{M \in \partial^\circ \Phi(y+s)} \|\Phi(y+s) - \Phi(y) - Ms\|_{L^r} = o(\|s\|_{L^p}) \quad \text{as } s \to 0 \text{ in } L^p(\Omega). \tag{5.7}$$

Further, we have for all $M \in \partial^\circ \Phi(u)$ and $v \in Y$

$$Mv = d_1 v + d_2 \cdot (Z'(y)v), \tag{5.8}$$

where $d \in L^\infty(\Omega)^2$ is a measurable selection of $\partial\phi(y, Z(y))$.

In Example 5.6 we will show that the optimal control problem (1.6) can be converted to an equivalent NCP for which the above assumptions (1), (2) are satisfied.

In the rest of this example we focus on semismoothness of order $\beta > 0$. As above, we see that Assumption 3.4 holds if instead of (1) and (4) we require the following:

(1′) The operator $Z : L^p(\Omega) \to L^r(\Omega)$ is α-Hölder continuously Fréchet differentiable.
(4′) ϕ is α-order semismooth.

If condition (5.3) is also satisfied, we can apply Theorem 5.2 to derive the β-order semismoothness of Φ.

Once we have chosen a particular NCP-function, condition (5.3) can be made very concrete. We discuss this for the Fischer–Burmeister function $\phi = \phi_{FB}$, which is Lipschitz continuous and 1-order semismooth and thus satisfies Assumptions 3.4(c) and (d′) with $\alpha = 1$. Further, this function is C^∞ on $\mathbb{R}^2 \setminus \{0\}$ with derivatives

$$\nabla\phi(x) = \frac{x}{\|x\|_2} - \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad \nabla^2\phi(x) = \frac{1}{\|x\|_2^3} \begin{pmatrix} x_2^2 & -x_1 x_2 \\ -x_1 x_2 & x_1^2 \end{pmatrix}.$$

The eigenvalues of $\nabla^2\phi(x)$ are 0 and $\|x\|_2^{-1}$. In particular, we see that $\|\nabla^2\phi(x)\|_2 = \|x\|_2^{-1}$ explodes as $x \to 0$. If $0 \notin [x, x+h]$, then Taylor expansion of $\phi(x)$ about $x+h$ yields, with appropriate $\tau \in [0,1]$,

$$\rho(x,h) = |\phi(x+h) - \phi(x) - \nabla\phi(x+h)^T h| = \frac{1}{2}|h^T\nabla^2\phi(x+\tau h)h| \le \frac{\|h\|_2^2}{2\,\|x+\tau h\|_2}.$$

Furthermore, $\rho(0,h) = 0$, $\rho(x,0) = 0$. Our aim is to show that (5.3) is equivalent to the condition

$$\mu\left(\{0 < \|F(y)\|_1 < \varepsilon\}\right) = O(\varepsilon^\gamma) \quad \text{as } \varepsilon \to 0. \tag{5.9}$$

Obviously, this follows easily when we have established the following relation:

$$\{0 < \|F(y)\|_1 < \varepsilon\} \subset \Omega_\varepsilon \subset \{0 < \|F(y)\|_1 < (1 + 2^{-1/2})\varepsilon\}. \tag{5.10}$$

To show the first inclusion in (5.10), let ω be such that $x = F(y)(\omega)$ satisfies $0 < \|x\|_1 < \varepsilon$, and choose $h = -tx$, where $t \in (1, \sqrt{2})$ is such that $\|h\|_1 \le \varepsilon$. Then a straightforward calculation yields

$$\rho(x,h) = 2\,\|x\|_2 \ge \sqrt{2}\,\|x\|_1 = \frac{\sqrt{2}}{t}\,\|h\|_1 > \|h\|_1 \ge \varepsilon^{-1}\,\|h\|_1^2.$$

This implies that $\omega \in \Omega_\varepsilon$ and thus proves the first inclusion.

To show the second inclusion in (5.10), let $u = F(y)$. If $u(\omega) = 0$, then certainly $\omega \notin \Omega_\varepsilon$, since then $\rho(u(\omega), \cdot) \equiv 0$. If on the other hand $\|u(\omega)\|_1 \ge (1 + 2^{-1/2})\varepsilon$, then we have for all $h \in \varepsilon \bar{B}_1^2$

$$\rho\big(u(\omega), h\big) \le \frac{\|h\|_2^2}{2\,\|u(\omega)+\tau h\|_2} \le \frac{\|h\|_1^2}{\sqrt{2}\,\|u(\omega)+\tau h\|_1} \le \varepsilon^{-1}\,\|h\|_1^2,$$

and thus $\omega \notin \Omega_\varepsilon$.

Having established the equivalence of (5.3) and (5.9), the meaning of (5.3) becomes apparent: The set $\{0 < \|F(y)\|_1 < \varepsilon\}$ on which the decrease rate in measure is assumed is the set of all ω where strict complementarity holds, but is less than ε, i.e., $0 < |y(\omega)| + |Z(y)(\omega)| < \varepsilon$. In a neighborhood of these points the curvature of ϕ is very large since $\|\nabla^2\phi\|$ is big. This requires that $|F(y+s)(\omega) - F(y)(\omega)|$ must be very small in order to have a sufficiently small residual $\rho(F(y)(\omega), F(y+s)(\omega) - F(y)(\omega))$.

We stress that a violation of strict complementarity, i.e., $y(\omega) = Z(y)(\omega) = 0$, does not cause any problems, since then $\rho(F(y)(\omega), \cdot) = \rho(0, \cdot) \equiv 0$.

In the next example, we return to the control problem (1.6).

Example 5.6 (application to a control problem). We consider the constrained elliptic control problem (1.6) and show that it is equivalent to an NCP satisfying the conditions (a) and (b) derived in the previous Example 5.5. Further, we establish additional results that will be useful in section 6.2 where we describe how the developed semismooth Newton method can be applied to solve the control problem.

Denote by $A \in \mathcal{L}(H_0^1, H^{-1})$ the linear operator on the left-hand side of (1.6b). Due to the uniform ellipticity assumption, it is well known that A is a homeomorphism, so that, using the continuous embedding $L^2(\Omega) \subset H^{-1}(\Omega) = H_0^1(\Omega)^*$, the control-to-state mapping $w \in L^2(\Omega) \mapsto u(w) = A^{-1}w \in H_0^1(\Omega)$ is continuous linear and thus smooth with Fréchet derivative $u'(w) : v \in L^2(\Omega) \mapsto A^{-1}v \in H_0^1(\Omega)$. Therefore, denoting by $\nabla J(w) \in L^2(\Omega)$ the L^2-Riesz representation of the gradient of J, we have

$$\nabla J(w) = (A^{-1})^*(A^{-1}w - u_d) + \lambda(w - w_d).$$

The first-order necessary (and here also sufficient) optimality conditions for (1.6a) result in the pointwise complementarity system

$$w \le b, \quad \nabla J(w) \le 0, \quad (w-b)\nabla J(w) = 0 \quad \text{on } \Omega. \tag{5.11}$$

Introducing the new unknown $y = b - w \in L^2(\Omega)$ and the operator $Z : L^2(\Omega) \to L^2(\Omega)$, $Z(y) = -\nabla J(b-y)$, the optimality system (5.11) is equivalent to the NCP (5.5); their solutions are related via the identity $w = b - y$. Now choose p such that

(5.12)

$$p \in (2,\infty] \text{ if } n = 1, \quad p \in (2,\infty) \text{ if } n = 2, \quad \text{and} \quad p \in \left(2, \frac{2n}{(n-2)}\right) \text{ if } n \ge 3.$$

Then the continuous embedding $H_0^1(\Omega) \subset L^p(\Omega)$ holds. We have

$$Z(y) = G(y) + \lambda y, \quad \text{where} \quad G(y) = (A^{-1})^*(A^{-1}(y-b) + u_d) + \lambda(w_d - b).$$

Since $(A^{-1})^*(A^{-1}(y-b)+u_d) \in H_0^1(\Omega) \subset L^p(\Omega)$ for $y \in L^2(\Omega)$ and $\lambda(w_d - b) \in L^\infty(\Omega)$, G maps $L^2(\Omega)$ continuously affine linearly into $L^p(\Omega)$.

Next, consider a solution y of the NCP. If $y(x) = 0$, then $0 \le Z(y)(x) = G(y)(x) + \lambda y(x) = G(y)(x)$. If $y(x) \ne 0$, then $y(x) > 0$ and $Z(y)(x) = 0$, which implies $y(x) = -\lambda^{-1}G(y)(x) > 0$. This shows $y = \max\{-\lambda^{-1}G(y), 0\} \in L^p(\Omega)$.

Therefore, with p as in (5.12), the NCP corresponding to the control problem has the following properties:

(1) Any solution of the NCP lies in $L^p(\Omega)$, with $p > 2$ as in (5.12).
(2) $Z : L^2(\Omega) \to L^2(\Omega)$ is continuous affine linear.
(3) $Z(y) = G(y) + \lambda y$, where $G : L^2(\Omega) \to L^p(\Omega)$, $p > 2$ as in (1), is continuous affine linear. In particular, Z maps $L^p(\Omega)$ continuously affine linearly to $L^p(\Omega)$.

From these results we can immediately derive the assumptions (1), (2), and (1′) in Example 5.5. In fact, from (1) here we see that we can pose the problem in $L^p(\Omega)$ instead of $L^2(\Omega)$. Now let $q = p$ and $r = 2$. Then (2) shows that Z maps $L^p(\Omega)$ continuously affine linearly to $L^r(\Omega)$, and thus condition (1) of Example 5.5, and even condition (1′) with $\alpha = 1$, hold. From (3) we conclude that Z maps $L^p(\Omega)$ continuously affine linearly to $L^q(\Omega)$ with $q = p$. This establishes condition (2) of Example 5.5.

The control problem of the previous example is further considered in section 6.2.

Remark 5.7. In Example 5.6 we saw that NCPs arising in practice sometimes satisfy stronger assumptions than those stated in Example 5.5. A typical situation is the following: The NCP is posed in the Hilbert space $L^2(\Omega)$, and $Z : L^2(\Omega) \to L^2(\Omega)$ is continuously Fréchet differentiable. Further, one can find $p, q > 2$ such that Z maps $L^p(\Omega)$ locally Lipschitz continuously to $L^q(\Omega)$. Finally, any solution of the NCP can be shown to lie in $L^p(\Omega)$. This is the situation we had in Example 5.6.

5.1. Proof of Theorem 5.2. We can simplify the analysis by exploiting the following fact.

LEMMA 5.8. *Let Assumption* 3.1 *hold and suppose that the operator*

$$\Lambda : u \in \prod\nolimits_i L^{q_i}(\Omega) \mapsto \psi(u) \in L^r(\Omega)$$

is semismooth at $u = F(y)$. Then the operator $\Psi : Y \to L^r(\Omega)$ defined in (1.8) *is semismooth at y. Further, if Assumption* 3.4 *holds and Λ is α-order semismooth at $u = F(y)$, then Ψ is α-order semismooth at y.*

Proof. We first observe that, given any $M \in \partial^\circ \Psi(y+s)$, there is $M_\Lambda \in \partial^\circ \Lambda(F(y+s))$ such that $M = M_\Lambda F'(y+s)$. In fact, there exists a measurable selection $d \in$

$L^\infty(\Omega)^m$ of $\partial\psi(\omega)$ such that $M = \sum_i d_i \cdot F_i'(y+s)$, and obviously $M_\Lambda : v \mapsto \sum_i d_i v_i$ yields an element of $\partial^\circ\Lambda(F(y+s))$ with the desired property. A more general chain rule will be established in Theorem 7.2.

Setting $u = F(y)$, $v = F(y+s) - F(y)$, and $w = F(y+s)$, we have

$$\begin{aligned}
&\sup_{M\in\partial^\circ\Psi(y+s)} \|\Psi(y+s) - \Psi(y) - Ms\|_{L^r} \\
&\quad\le \sup_{M_\Lambda\in\partial^\circ\Lambda(w)} \|\Lambda(w) - \Lambda(u) - M_\Lambda F'(y+s)s\|_{L^r} \\
&\quad\le \sup_{M_\Lambda\in\partial^\circ\Lambda(w)} \|\Lambda(w) - \Lambda(u) - M_\Lambda v\|_{L^r} \\
&\qquad + \sup_{M_\Lambda\in\partial^\circ\Lambda(w)} \big\|M_\Lambda\big(F(y+s) - F(y) - F'(y+s)s\big)\big\|_{L^r} \stackrel{\text{def}}{=} \rho_\Lambda + \rho_{MF}.
\end{aligned}$$

By the local Lipschitz continuity of F and the semismoothness of Λ, we obtain

$$\rho_\Lambda = o(\|v\|_{\Pi_i L^{q_i}}) = o(\|s\|_Y) \quad \text{as } s \to 0 \text{ in } Y.$$

Further, since $d \in L_\psi \bar{B}^m_{L^\infty}$ by Lemma 4.7, we have by Assumption 3.1(a)

$$\begin{aligned}
\|\rho_{MF}\|_{L^r} &\le L_\psi \sum\nolimits_i \|F_i(y+s) - F_i(y) - F_i'(y+s)s\|_{L^r} \\
&\le L_\psi \sum\nolimits_i c_{r,r_i}(\Omega)\, \|F_i(y+s) - F_i(y) - F_i'(y+s)s\|_{L^{r_i}} \\
&= o(\|s\|_Y) \quad \text{as } s \to 0 \text{ in } Y.
\end{aligned}$$

This proves the first result.

Now let Assumption 3.4 hold and Λ be α-order semismooth at $u = F(y)$. Then ρ_Λ and ρ_{MF} are both of the order $O(\|s\|_Y^{1+\alpha})$, which implies the second assertion. ∎

For the proof of Theorem 5.2 we need, as a technical intermediate result, the Borel measurability of the function

$$\rho : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}, \qquad \rho(x,h) = \max_{z^T\in\partial\psi(x+h)} |\psi(x+h) - \psi(x) - z^T h|. \tag{5.13}$$

We prove this by showing that ρ is upper semicontinuous. Readers familiar with this type of result might want to skip the proof of Lemma 5.9.

Recall that a function $f : \mathbb{R}^l \to \mathbb{R}$ is upper semicontinuous at x if

$$\limsup_{x'\to x} f(x') \le f(x).$$

Equivalently, f is upper semicontinuous if and only if $\{x : f(x) \ge a\}$ is closed for all $a \in \mathbb{R}$.

LEMMA 5.9. *Let $f : (x,z) \in \mathbb{R}^l \times \mathbb{R}^m \mapsto \mathbb{R}$ be upper semicontinuous. Moreover, let the multifunction $\Gamma : \mathbb{R}^l \rightrightarrows \mathbb{R}^m$ be upper semicontinuous and compact-valued. Then the function*

$$g : \mathbb{R}^l \to \mathbb{R}, \quad g(x) = \max_{z\in\Gamma(x)} f(x,z),$$

is well defined and upper semicontinuous.

Proof. For $x \in \mathbb{R}^l$, let $(z_k) \subset \Gamma(x)$ be such that

$$\lim_{k\to\infty} f(x,z_k) = \sup_{z\in\Gamma(x)} f(x,z).$$

Since $\Gamma(x)$ is compact, we may assume that $z_k \to z^*(x) \in \Gamma(x)$. Now, by the upper semicontinuity of f,

$$f\big(x, z^*(x)\big) \geq \limsup_{k\to\infty} f(x, z_k) = \sup_{z\in\Gamma(x)} f(x, z) \geq f\big(x, z^*(x)\big).$$

Thus, g is well defined, and there exists $z^* : \mathbb{R}^l \to \mathbb{R}^m$ with $g(x) = f(x, z^*(x))$.

We now prove the upper semicontinuity of g at x. Let $(x_k) \subset \mathbb{R}^l$ tend to x in such a way that

$$\lim_{k\to\infty} g(x_k) = \limsup_{x'\to x} g(x'),$$

and set $z_k = z^*(x_k) \in \Gamma(x_k)$. By the upper semicontinuity of Γ, there exists $(\hat{z}_k) \subset \Gamma(x)$ with $(\hat{z}_k - z_k) \to 0$ as $k \to \infty$.

Since $\Gamma(x)$ is compact, a subsequence can be selected such that the sequence $(\hat{z}_k)$, and thus (z_k), converges to some $\hat{z} \in \Gamma(x)$. Now, using that f is upper semicontinuous and $\hat{z} \in \Gamma(x)$,

$$\limsup_{x'\to x} g(x') = \lim_{k\to\infty} g(x_k) = \lim_{k\to\infty} f(x_k, z_k) = \limsup_{k\to\infty} f(x_k, z_k) \leq f(x, \hat{z}) \leq g(x).$$

Therefore, g is upper semicontinuous at x. ∎

LEMMA 5.10. *Let $\psi : \mathbb{R}^m \to \mathbb{R}$ be locally Lipschitz continuous. Then the function ρ defined in* (5.13) *is well defined and upper semicontinuous.*

Proof. Since $\partial\psi$ is upper semicontinuous and compact-valued, the multifunction

$$(x, h) \in \mathbb{R}^m \times \mathbb{R}^m \mapsto \partial\psi(x + h)$$

is upper semicontinuous and compact-valued as well. Further, the mapping

$$(x, h, z) \mapsto |\psi(x + h) - \psi(x) - z^T h|$$

is continuous, and we may apply Lemma 5.9, which yields the assertion. ∎

Proof of Theorem 5.2. By Lemma 5.8, it suffices to prove the semismoothness (of order β) of the operator

$$\Lambda : u \in \prod\nolimits_i L^{q_i}(\Omega) \mapsto \psi(u) \in L^r(\Omega).$$

(a) Semismoothness. In Lemma 5.10 we showed that the function

$$\rho : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}, \quad \rho(x, h) = \max_{z^T\in\partial\psi(x+h)} |\psi(x + h) - \psi(x) - z^T h|,$$

is upper semicontinuous and thus Borel measurable. Hence, for $u, v \in \prod_i L^{r_i}(\Omega)$, the function $\rho(u, v)$ is measurable. We define the measurable function

$$a = \frac{\rho(u, v)}{\|v\|_1 + \mathbf{1}_{\{v=0\}}}.$$

Since $\rho(u(\omega), v(\omega)) = 0$ whenever $v(\omega) = 0$, we obtain

$$\rho(u, v) = a\,\|v\|_1.$$

Furthermore,

$$a(\omega) = \frac{\rho\big(u(\omega), v(\omega)\big)}{\|v(\omega)\|_1 + \mathbf{1}_{\{v=0\}}(\omega)} = \frac{o\big(\|v(\omega)\|_1\big)}{\|v(\omega)\|_1 + \mathbf{1}_{\{v=0\}}(\omega)} \to 0 \qquad \text{as } v(\omega) \to 0. \tag{5.14}$$

Due to the Lipschitz continuity of ψ, we have

$$\rho(x,h) \le 2L_\psi \|h\|_1, \tag{5.15}$$

which implies $a \in 2L_\psi \bar{B}_{L^\infty}$.

Now let (v_k) tend to zero in the space $\prod_i L^{q_i}(\Omega)$, and set $a_k = a|_{v=v_k}$. Then every subsequence of (v_k) itself contains a subsequence $(v_{k'})$ such that $v_{k'} \to 0$ a.e. on Ω. By (5.14), this implies $a_{k'} \to 0$ a.e. on Ω. Since $(a_{k'})$ is bounded in $L^\infty(\Omega)$, we conclude

$$\lim_{k'\to\infty} \|a_{k'}\|_{L^t} = 0 \quad \text{for all } t \in [1,\infty).$$

Hence, in $L^t(\Omega)$, $1 \le t < \infty$, zero is an accumulation point of every subsequence of (a_k). This proves $a_k \to 0$ in all spaces $L^t(\Omega)$, $1 \le t < \infty$.

Since the sequence (v_k), $v_k \to 0$, was arbitrary, we thus have proven that, for all $1 \le t < \infty$,

$$\|a\|_{L^t} \to 0 \quad \text{as } \|v\|_{\Pi_i L^{q_i}} \to 0.$$

Now we can use Hölder's inequality to obtain

$$\begin{aligned} \|\rho(u,v)\|_{L^r(\Omega)} &\le \sum\nolimits_i \|a v_i\|_{L^r} \le \sum\nolimits_i \|a\|_{L^{p_i}} \|v_i\|_{L^{q_i}} \\ &\le \left(\max_{1\le i\le m} \|a\|_{L^{p_i}}\right) \|v\|_{\Pi_i L^{q_i}} = o\big(\|v\|_{\Pi_i L^{q_i}}\big) \quad \text{as } \|v\|_{\Pi_i L^{q_i}} \to 0, \end{aligned} \tag{5.16}$$

where $p_i = \frac{q_i r}{q_i - r}$ if $q_i < \infty$, and $p_i = r$ if $q_i = \infty$. Note that here we exploited the fact that $r < q_i$. This proves the semismoothness of Λ.

(b) Semismoothness of order β. We now suppose that Assumption 3.4 and, in addition, (5.3) hold. First, note that for fixed $\varepsilon > 0$ the function

$$(x,h) \in \mathbb{R}^m \times \mathbb{R}^m \mapsto \rho(x,h) - \varepsilon^{-\alpha} \|h\|_1^{1+\alpha}$$

is upper semicontinuous and that the multifunction

$$x \in \mathbb{R}^m \mapsto \varepsilon \bar{B}_1^m$$

is compact-valued and upper semicontinuous. Hence, by Lemma 5.9, the function

$$x \in \mathbb{R}^m \mapsto \max_{\|h\|_1 \le \varepsilon} \left(\rho(x,h) - \varepsilon^{-\alpha} \|h\|_1^{1+\alpha}\right)$$

is upper semicontinuous and therefore Borel measurable. This proves the measurability of the set Ω_ε appearing in (5.3). For $\varepsilon > 0$ and $0 < \beta \le \alpha$ we define the set

$$\Omega_{\beta\varepsilon} = \left\{\omega : \rho\big(u(\omega), v(\omega)\big) > \varepsilon^{-\beta} \|v(\omega)\|_1^{1+\beta}\right\}$$

and observe that

$$\Omega_{\beta\varepsilon} \subset \Omega_\varepsilon \cup \{\|v\|_1 > \varepsilon\} \stackrel{\text{def}}{=} \Omega_\varepsilon \cup \Omega'_\varepsilon.$$

In fact, let $\omega \in \Omega_{\beta\varepsilon}$ be arbitrary. The nontrivial case is $\|v(\omega)\|_1 \le \varepsilon$. We then obtain for $h = v(\omega)$

$$\rho\big(u(\omega), h\big) > \varepsilon^{-\beta} \|h\|_1^{1+\beta} = \varepsilon^{-\alpha}\varepsilon^{\alpha-\beta} \|h\|_1^{1+\beta} \ge \varepsilon^{-\alpha} \|h\|_1^{\alpha-\beta} \|h\|_1^{1+\beta} = \varepsilon^{-\alpha} \|h\|_1^{1+\alpha},$$

and thus, since $\|h\|_1 \le \varepsilon$,

$$\max_{\|h\|_1 \le \varepsilon} \left(\rho\big(u(\omega), h\big) - \varepsilon^{-\alpha} \|h\|_1^{1+\alpha} \right) > 0,$$

showing that $\omega \in \Omega_\varepsilon$.

In the case $q_0 = \min_{1\le i\le m} q_i < \infty$ we derive the estimate

$$\begin{aligned}\mu(\Omega'_\varepsilon) = \mu\left(\{\|v\|_1 > \varepsilon\}\right) &\le \left\|\varepsilon^{-1} \|v\|_1\right\|^{q_0}_{L^{q_0}(\Omega'_\varepsilon)} \\ &\le \varepsilon^{-q_0} \left(\max_i c_{q_0,q_i}(\Omega'_\varepsilon) \right)^{q_0} \|v\|^{q_0}_{\Pi_i L^{q_i}} = \varepsilon^{-q_0} O\left(\|v\|^{q_0}_{\Pi_i L^{q_i}} \right).\end{aligned}$$

If we choose $\varepsilon = \|v\|^\lambda_{\Pi_i L^{q_i}}$, $0 < \lambda < 1$, then

$$\mu(\Omega_{\beta\varepsilon}) \le \mu(\Omega_\varepsilon) + \mu(\Omega'_\varepsilon) = O\big(\|v\|^{\gamma\lambda}_{\Pi_i L^{q_i}}\big) + O\big(\|v\|^{(1-\lambda)q_0}_{\Pi_i L^{q_i}}\big).$$

This estimate is also true in the case $q_0 = \infty$, since then $\mu(\Omega'_\varepsilon) = 0$ as soon as $\|v\|_{\Pi_i L^{q_i}} < 1$. This can be seen by noting that then for a.a. $\omega \in \Omega$ the following holds:

$$\|v(\omega)\|_1 \le \| \|v\|_1 \|_{L^\infty} \le \|v\|_{\Pi_i L^{q_i}} \le \|v\|^\lambda_{\Pi_i L^{q_i}} = \varepsilon.$$

Introducing $\nu = \frac{q_0 - r}{q_0 r}$ if $q_0 < \infty$, and $\nu = 1/r$ otherwise, for all $0 < \beta \le \alpha$, we obtain, using (5.15) and Lemma A.1,

$$\begin{aligned}\|\rho(u,v)\|_{L^r(\Omega_{\beta\varepsilon})} &\le \left\|2L_\psi \|v\|_1 \right\|_{L^r(\Omega_{\beta\varepsilon})} \le 2L_\psi c_{r,q_0}(\Omega_{\beta\varepsilon}) \|v\|_{L^{q_0}(\Omega_{\beta\varepsilon})^m} \\ &\le 2L_\psi \mu(\Omega_{\beta\varepsilon})^\nu \|v\|_{L^{q_0}(\Omega_{\beta\varepsilon})^m} \\ &= O\big(\|v\|^{1+\gamma\lambda\nu}_{\Pi_i L^{q_i}}\big) + O\big(\|v\|^{1+(1-\lambda)\nu q_0}_{\Pi_i L^{q_i}}\big).\end{aligned} \tag{5.17}$$

Again, we have used here the fact that $r < q_0 \le q_i$, which allowed us to take advantage of the smallness of the set $\Omega_{\beta\varepsilon}$.

Finally, on $\Omega^c_{\beta\varepsilon}$, $(1+\beta)r \le q_0$, $0 < \beta \le \alpha$, holds with our choice $\varepsilon = \|v\|^\lambda_{\Pi_i L^{q_i}}$

$$\begin{aligned}\|\rho(u,v)\|_{L^r(\Omega^c_{\beta\varepsilon})} &\le \left\|\varepsilon^{-\beta} \|v\|_1^{1+\beta}\right\|_{L^r(\Omega^c_{\beta\varepsilon})} \le c_{r,\frac{q_0}{1+\beta}}(\Omega^c_{\beta\varepsilon}) \|v\|^{-\beta\lambda}_{\Pi_i L^{q_i}} \|v\|^{1+\beta}_{L^{q_0}(\Omega^c_{\beta\varepsilon})^m} \\ &= O\big(\|v\|^{1+\beta(1-\lambda)}_{\Pi_i L^{q_i}}\big).\end{aligned}$$

Therefore,

$$\|\rho(u,v)\|_{L^r} = O\big(\|v\|^{1+\gamma\lambda\nu}_{\Pi_i L^{q_i}}\big) + O\big(\|v\|^{1+(1-\lambda)\nu q_0}_{\Pi_i L^{q_i}}\big) + O\big(\|v\|^{1+\beta(1-\lambda)}_{\Pi_i L^{q_i}}\big).$$

We now choose $0 < \lambda < 1$ and $\beta > 0$ with $\beta \le \alpha$, $(1+\beta)r \le q_0$, in such a way that the order of the right-hand side is maximized. In the case $(1+\alpha)r \ge q_0$ the minimum of all three exponents is maximized for the choice $\beta = \frac{q_0 - r}{r} = \nu q_0$ and $\lambda = \frac{q_0}{\gamma + q_0}$. Then all three exponents are equal to $1 + \frac{\gamma\nu q_0}{\gamma + q_0}$, and thus

$$\|\rho(u,v)\|_{L^r} = O\left(\|v\|^{1+\frac{\gamma\nu q_0}{\gamma+q_0}}_{\Pi_i L^{q_i}} \right). \tag{5.18}$$

If, on the other hand, $(1+\alpha)r < q_0$, then the third exponent is smaller than the second one for all $0 < \lambda < 1$ and $0 < \beta \le \alpha$. Further, it is not difficult to see that under these constraints the first and third exponent become maximal for $\beta = \alpha$ and $\lambda = \frac{\alpha}{\alpha + \gamma\nu}$ and attain the value $1 + \frac{\alpha\gamma\nu}{\alpha+\gamma\nu}$. Hence,

$$\|\rho(u,v)\|_{L^r} = O\left(\|v\|^{1+\frac{\alpha\gamma\nu}{\alpha+\gamma\nu}}_{\Pi_i L^{q_i}} \right). \tag{5.19}$$

Combining (5.18) and (5.19) proves the β-order semismoothness of Λ with β as in (5.4). ∎

5.2. Illustrations. In this section we give two examples to illustrate the above analysis by pointing out the necessity of the main assumptions and by showing that the derived results cannot be improved in several respects.

In order to prevent our examples from being too academic, we will not work with the simplest choices possible. Rather, we will throughout use reformulations of NCPs based on the Fischer–Burmeister function.

The examples address the following items:

- Example 5.11 shows the necessity of the norm gap between L^{q_i}- and L^r-norms.
- Example 5.12 discusses the sharpness of our order of semismoothness β in Theorem 5.2 for varying values of γ.

At the indicated places (5.16) and (5.17) in the above proof, we needed the gap between the L^r- and L^{q_i}-norms in order to apply Hölder's inequality. The following example illustrates that Theorem 5.2 does not hold in general if we drop the condition $r_i < q_i$ in Assumption 3.1.

Example 5.11 (Necessity of the norm gap $r < q_i$). We return to the setting of NCPs as described in Example 5.5. Under the assumptions stated there, we obtain from Theorem 5.2 that the estimate (5.7) holds, where $1 \le r < q \le \infty$. Our aim here is to show that the requirement $r < q$ is indispensable in the sense that (5.7) is violated in general for $r \ge q$.

As we will see in section 6, the estimate (5.7) at a solution y of the NCP is the main tool for proving fast local convergence of Newton's method. Hence, we will construct a simple NCP with a unique solution for which (5.7) fails to hold whenever $r \ge q$. Hereby, we use the Fischer–Burmeister NCP-function ϕ_{FB} defined in (1.4) for the reformulation (5.6) of the NCP.

Let $1 < p \le \infty$ be arbitrary, choose $\Omega = (0,1)$, and set

$$Z(y)(\omega) = y(\omega) + \omega.$$

Obviously, $\bar{y} \equiv 0$ is the unique solution of the NCP. Choosing $q = p$, $\phi = \phi_{FB}$, and $\alpha = 1$, the assumption in Example 5.5, and hence also Assumption 3.4, is satisfied for all $r \in [1, p)$. To show that the requirement $r < p$ is really necessary to obtain the semismoothness of Φ, we will investigate the residual

$$R(s) \stackrel{\text{def}}{=} \Phi(\bar{y} + s) - \Phi(\bar{y}) - Ms, \quad M \in \partial^\circ \Phi(\bar{y} + s), \tag{5.20}$$

at $\bar{y} \equiv 0$ with $s \in L^\infty(\Omega)$, $s \ge 0$, $s \ne 0$. Our aim is to show that for all $r \in [1, \infty]$

$$\|R(s)\|_{L^r} = o(\|s\|_{L^p}) \quad \text{as } s \to 0 \text{ in } L^\infty \quad \Longrightarrow \quad r < p \tag{5.21}$$

holds. To this end, let $s \in L^p(\Omega)$, $s \ge 0$, $y, v \in L^p(\Omega)$ and define $F(y) = (y, Z(y))$. Then

$$\begin{aligned} \Phi(y)(\omega) &= \phi(y, Z(y))(\omega) = \phi(y(\omega), y(\omega) + \omega), \\ (F'(y)v)(\omega) &= (v, Z'(y)v)(\omega) = (v(\omega), v(\omega)) \end{aligned}$$

for almost all ω. Also, since $\bar{y} \equiv 0$, any $M \in \partial^\circ \Phi(\bar{y} + s) = \partial^\circ \Phi(s)$ satisfies, for almost all ω,

$$(Mv)(\omega) \in \partial\phi(F(s)(\omega))(F'(s)v)(\omega) = \partial\phi(\sigma, \sigma + \omega)(v(\omega), v(\omega))$$

with $\sigma = s(\omega)$. Thus,

$$(Mv)(\omega) = \phi'(\sigma, \sigma + \omega)(v(\omega), v(\omega)),$$

since ϕ is smooth except at the origin and $(s(\omega), s(\omega)+\omega) \neq (0,0)$ for almost all ω. Using this, a straightforward calculation gives

$$|R(s)(\omega)| = |\phi(\sigma, \sigma+\omega) - \phi(0,\omega) - \phi'(\sigma,\sigma+\omega)(\sigma,\sigma)| = \omega - \frac{\omega(\sigma+\omega)}{\sqrt{2\sigma^2+2\sigma\omega+\omega^2}}$$

a.e. on $\Omega = (0,1)$. Now let $0 < \varepsilon < 1$. For the special choice $s_\varepsilon \overset{\text{def}}{=} \varepsilon \mathbf{1}_{(0,\varepsilon)}$, i.e., $s_\varepsilon(\omega) = \varepsilon$ for $\omega \in (0,\varepsilon)$, and $s_\varepsilon(\omega) = 0$ otherwise, we obtain

$$\|s_\varepsilon\|_{L^p} = \varepsilon^{\frac{p+1}{p}} \quad (1 < p < \infty), \quad \|s_\varepsilon\|_{L^\infty} = \varepsilon.$$

In particular, $s_\varepsilon \to 0$ in L^∞ as $\varepsilon \to 0$. For almost all $0 < \omega < \varepsilon$ there holds

$$|R(s_\varepsilon)(\omega)| \geq \omega \left(1 - \sup_{0<t<1} \frac{1+t}{\sqrt{2+2t+t^2}}\right) = \frac{5-2\sqrt{5}}{5}\omega \geq \frac{\omega}{10}.$$

Hence, $\|R(s_\varepsilon)\|_{L^\infty} \geq \frac{\varepsilon}{10} \geq \frac{\|s_\varepsilon\|_{L^p}}{10}$, and for all $r \in [p,\infty)$

$$\|R(s_\varepsilon)\|_{L^r} \geq \frac{1}{10}\left(\int_0^\varepsilon \omega^r d\omega\right)^{\frac{1}{r}} = \frac{\varepsilon^{\frac{r+1}{r}}}{10(r+1)^{\frac{1}{r}}} \geq \frac{\|s_\varepsilon\|_{L^p}}{10(r+1)^{\frac{1}{r}}}.$$

Therefore, (5.21) is proved. This shows that in (5.7) the norm on the left-hand side must be stronger than on the right-hand side. ∎

Next, we show that, at least in the case $q_0 \leq (1+\alpha)r$, the order of our semismoothness result is sharp. By showing this for varying values of γ, we also observe that decreasing values of γ reduce the maximum order of semismoothness exactly as stated in Theorem 5.2. Hence, our result does not overestimate the role of γ.

Example 5.12 (order of semismoothness and its dependence on γ). We consider the following NCP, which generalizes the one in Example 5.11: Let $1 < p \leq \infty$ be arbitrary, set $\Omega = (0,1)$, and choose

$$Z(y)(\omega) = y(\omega) + \omega^\theta, \quad \theta > 0.$$

Obviously, $\bar{y} \equiv 0$ is the unique solution of the NCP. Choosing $q = p$, $\phi = \phi_{FB}$, and $\alpha = 1$, the assumptions in Example 1.5, and hence also Assumption 3.4, is satisfied for all $r \in [1,p)$.

From $Z(\bar{y})(\omega) = (0,\omega^\theta)$ it follows that $\gamma = 1/\theta$ is the maximum value for which condition (5.9), and thus the equivalent condition (5.3), is satisfied.

With the residual $R(s)$ as defined in (5.20), we obtain

$$|R(s)(\omega)| = \omega^\theta - \frac{\omega^\theta(s(\omega)+\omega^\theta)}{\sqrt{2s(\omega)^2+2s(\omega)\omega^\theta+\omega^{2\theta}}}.$$

For $\varepsilon \in (0,1)$ and $s_\varepsilon \overset{\text{def}}{=} \varepsilon^\theta \mathbf{1}_{(0,\varepsilon)}$ we have

$$\|s_\varepsilon\|_{L^p} = \varepsilon^{\frac{p\theta+1}{p}} \quad (1 < p < \infty), \quad \|s_\varepsilon\|_{L^\infty} = \varepsilon^\theta.$$

Further, for $0 < \omega < \varepsilon$ we have

$$|R(s_\varepsilon)(\omega)| \geq \omega^\theta \left(1 - \sup_{0<t<1} \frac{1+t}{\sqrt{2+2t+t^2}}\right) = \frac{5-2\sqrt{5}}{5}\omega^\theta \geq \frac{\omega^\theta}{10}.$$

Hence, for all $r \in [1,p)$

$$\|R(s_\varepsilon)\|_{L^r} \geq \frac{1}{10}\left(\int_0^\varepsilon \omega^{r\theta} d\omega\right)^{\frac{1}{r}} = \frac{\varepsilon^{\frac{r\theta+1}{r}}}{10(r\theta+1)^{\frac{1}{r}}} \geq \frac{\|s_\varepsilon\|_{L^p}^{\frac{pr\theta+p}{pr\theta+r}}}{10(r\theta+1)^{\frac{1}{r}}} = \frac{\|s_\varepsilon\|_{L^p}^{1+\frac{\gamma\nu}{1+\gamma/q_0}}}{10(r\theta+1)^{\frac{1}{r}}},$$

with $q_0 = p = q$, $\gamma = 1/\theta$, and ν as in (5.4). This shows that the value of β given in Theorem 5.2 is sharp for all values of θ (and thus γ) at least as long as $q_0 \leq (1+\alpha)r$, which in the current setting can be written as $p \leq (1+\alpha)r$.

We think that in the case $q_0 > (1+\alpha)r$ our value of β could still be slightly improved by splitting Ω into more than the two parts $\Omega_{\beta\varepsilon}$ and $\Omega^c_{\beta\varepsilon}$ by choosing different values ε_k for ε that correspond to different powers of $\|v\|_{\Pi_i L^{q_i}}$. In order to keep the analysis as clear as possible, we will not pursue this idea any further in the current paper.

6. Semismooth Newton method. We now apply the developed semismoothness results to derive a superlinearly convergent Newton-type method for the solution of the nonsmooth operator equation

$$\Psi(y) = 0, \tag{6.1}$$

with Ψ as defined in (1.8). Throughout this chapter, let $\bar{y} \in Y$ denote a solution to (6.1). We impose the following regularity condition on $\partial^\circ \Psi$.

Assumption 6.1. There exist a Banach space $Y_0 \supset Y$ (Y continuously embedded) and positive constants η, $C_{M^{-1}}$ such that, for all $y \in \bar{y} + \eta B_Y$, every $M \in \partial^\circ \Psi(y)$ can be extended to an invertible operator $M \in \mathcal{L}(Y_0, L^r)$ with $\|M^{-1}\|_{L^r, Y_0} \leq C_{M^{-1}}$.

Example 6.2 (application to NCP). In the following, we want to discuss why the introduction of the additional space Y_0 is of importance. To this end, we consider the reformulation of the NCP (1.5) in the form (5.6), as described in Example 5.5. Recall that the operators $M \in \partial^\circ \Phi(y)$ assume the form (5.8). Now define $\Omega_1 = \{\omega \in \Omega : d_2(\omega) = 0\}$. Then for all $\omega \in \Omega_1$ we have

$$(Mv)(\omega) = d_1(\omega) v(\omega).$$

This shows that (i) M can only be expected to be invertible (between appropriate spaces) if $d_1 \neq 0$ on Ω_1 and (ii) Mv is in general not more regular (in the L^p-sense) than v and vice versa. Therefore, it is not appropriate to assume that $M \in \mathcal{L}(Y, L^r)$ is continuously invertible, as the norm on $Y = L^p$ is stronger than on L^r. However, it is reasonable to assume that M is an L^r-automorphism. This leads to the regularity Assumption 6.1 with $Y_0 = L^r(\Omega)$, which can be verified to hold for many NCPs arising in practice; see [63, 62]. In [63] and [62] sufficient conditions for regularity are established that are widely applicable and easy to apply.

Being aware of the potential gap between the Y_0- and Y-norms, we propose the following Newton method for the solution of (6.1). The algorithm includes a smoothing step to overcome the discrepancy of norms, which will be discussed in section 6.1.

Algorithm 6.3 (semismooth Newton method).

0. Choose an initial point $y_0 \in Y$ sufficiently close to a solution $\bar{y} \in Y$ of (6.1). Set $k = 0$.
1. If $\Psi(y_k) = 0$, then stop with solution y_k.
2. Compute $M_k \in \partial^\circ \Psi(y_k)$, determine $s_k \in Y_0$ by solving

$$M_k s_k = -\Psi(y_k),$$

   and set $y^n_{k+1} = y_k + s_k$.
3. Perform a smoothing step:

$$y^n_{k+1} \in Y^0 \mapsto y_{k+1} \in Y.$$

4. Increment k by one and go to Step 1.

For the smoothing step we require the following.

Assumption 6.4. There exists $C_S > 0$ such that, for all k, the following holds:

$$\|y_{k+1} - \bar{y}\|_Y \le C_S \|y_{k+1}^n - \bar{y}\|_{Y_0}.$$

The local convergence proof for Algorithm 6.3 will clarify the role of the smoothing step.

THEOREM 6.5. *Let Assumptions* 3.1, 6.1, *and* 6.4 *hold. Then there exists* $\delta > 0$ *such that for all* $y_0 \in \bar{y} + \delta B_Y$ *Algorithm* 6.3 *is well defined and either terminates with a solution* y_k *of* (6.1) *or generates a sequence* $(y_k) \subset Y$ *that converges q-superlinearly to* $\bar{y}$.

Under the stronger Assumption 3.4 *and* (5.3), *the rate of convergence is of q-order* $1 + \beta$, *with* $\beta > 0$ *given in* (5.4).

Proof. Let $y_k \in \bar{y} + \delta B_Y$ with $\delta \in (0, \eta]$ sufficiently small. Then, by Assumption 6.1, the step s_k is well defined. Furthermore, using Assumption 6.1, $\Psi(\bar{y}) = 0$, and Theorem 5.2 gives, as $\delta \to 0$,

$$\begin{aligned} \|y_{k+1}^n - \bar{y}\|_{Y_0} &= \|y_k - M_k^{-1}\Psi(y_k) - \bar{y}\|_{Y_0} = \|M_k^{-1}(M_k(y_k - \bar{y}) - \Psi(y_k))\|_{Y_0} \\ &\le \|M_k^{-1}\|_{L^r, Y_0} \, \|0 - \Psi(y_k) - M_k(\bar{y} - y_k)\|_{L^r} \\ &\le C_{M^{-1}} \, \|\Psi(\bar{y}) - \Psi(y_k) - M_k(\bar{y} - y_k)\|_{L^r} = o(\|y_k - \bar{y}\|_Y), \end{aligned} \tag{6.2}$$

and thus, due to the properties of the smoothing step (see Assumption 6.4),

$$\|y_{k+1} - \bar{y}\|_Y \le C_S \|y_{k+1}^n - \bar{y}\|_{Y_0} = o(\|y_k - \bar{y}\|_Y).$$

We conclude that, if δ is sufficiently small and $y_0 \in \bar{y} + \delta B_Y$, then inductively, as long as $\Psi(y_k) \neq 0$, the new point y_{k+1} is well defined and $y_{k+1} \in \bar{y} + \delta B_Y$. Furthermore,

$$\|y_{k+1} - \bar{y}\|_Y = o(\|y_k - \bar{y}\|_Y).$$

This establishes the q-superlinear convergence.

Under Assumption 3.4 and (5.3), we can strengthen (6.2) to

$$\|y_{k+1}^n - \bar{y}\|_{Y_0} = O\big(\|y_k - \bar{y}\|_Y^{1+\beta}\big) \quad \text{as } k \to \infty,$$

where β is given by (5.4). Hence, using the properties of the smoothing step,

$$\|y_{k+1} - \bar{y}\|_Y = O\big(\|y_k - \bar{y}\|_Y^{1+\beta}\big) \quad \text{as } k \to \infty,$$

which proves convergence with q-order $1 + \beta$. ∎

6.1. Remarks on smoothing steps. Examples 5.11 and 6.2 demonstrate that the incorporation of a smoothing step into the Newton method in general cannot be avoided. However, the smoothing step is needed only in pathological cases, and it turns out to be quite common in practice that such bad situations do not occur very often. Since the design of smoothing steps is by no means trivial and its computation usually requires at least an additional evaluation of F, it would be valuable to have criteria at hand that indicate whether a smoothing step is needed. The underlying idea would be to run the algorithm without a smoothing step unless the indicator tells us that a smoothing is required. In the following, we discuss several aspects of this issue.

1. If the norms on Y_0 and Y are equivalent, then no smoothing step is needed; i.e., $y_{k+1} = y_{k+1}^n$ can be chosen for all k.

2. If in the kth iteration there holds

$$\|y_{k+1}^n - \bar{y}\|_Y \le C_S \|y_{k+1}^n - \bar{y}\|_{Y_0}, \tag{6.3}$$

then the smoothing step can be skipped, i.e., $y_{k+1} = y_{k+1}^n$ can be chosen. However, since $\bar{y}$ is not available, this condition cannot be checked at runtime.

3. We now derive a condition that necessarily holds if a smoothing step may be skipped. To this end, assume that y_{k+1}^n satisfies (6.3) and that y_k satisfies the smoothness condition

$$\|y_k - \bar{y}\|_Y \le C_S \|y_k - \bar{y}\|_{Y_0}. \tag{6.4}$$

Then, as shown in the proof of Theorem 6.5, for any $\kappa > 0$ there is a $\delta > 0$ such that for all $y_k \in \bar{y} + \delta B_Y$

$$\|y_{k+1}^n - \bar{y}\|_{Y_0} \le \kappa \|y_k - \bar{y}\|_Y \le \kappa C_S \|y_k - \bar{y}\|_{Y_0}$$

holds, and thus

$$\|y_{k+1}^n - \bar{y}\|_Y \le C_S \|y_{k+1}^n - \bar{y}\|_{Y_0} \le \kappa C_S \|y_k - \bar{y}\|_Y \le \kappa C_S^2 \|y_k - \bar{y}\|_{Y_0}.$$

Therefore,

$$\begin{aligned} \|s_k\|_{Y_0} &\ge \|y_k - \bar{y}\|_{Y_0} - \|y_{k+1}^n - \bar{y}\|_{Y_0} \ge (1 - \kappa C_S) \|y_k - \bar{y}\|_{Y_0}, \\ \|s_k\|_Y &\le \|y_k - \bar{y}\|_Y + \|y_{k+1}^n - \bar{y}\|_Y \le (1 + \kappa C_S) C_S \|y_k - \bar{y}\|_{Y_0}, \end{aligned}$$

and for $\kappa < 1/C_S$ we conclude that

$$\|s_k\|_Y \le \frac{1 + \kappa C_S}{1 - \kappa C_S} C_S \|s_k\|_{Y_0} \to C_S \|s_k\|_{Y_0} \quad \text{as } \kappa \to 0.$$

We obtain the following result.

LEMMA 6.6. *If, for fixed $\hat{C}_S > C_S$, y_k is sufficiently close to $\bar{y}$ in Y and*

$$\|s_k\|_Y > \hat{C}_S \|s_k\|_{Y_0}, \tag{6.5}$$

then at least one of the conditions (6.3), (6.4) *is violated.*

Therefore, if (6.5) occurs and we have good reasons to believe that (6.4) is satisfied (e.g., good residual reduction $\|\Psi(y_k)\|_{L^q} \ll \|\Psi(y_{k-1})\|_{L^q}$ with $q = \max_i q_i$ and smoothness of s_{k-1} in the sense that, e.g., $\|s_{k-1}\|_Y \le \hat{C}_S/2 \|s_{k-1}\|_Y$), we will perform a smoothing step to obtain y_{k+1} from y_{k+1}^n. If, on the other hand, it is doubtful that y_k satisfies (6.4), we have to return to iteration k and recompute y_k from y_k^n by a smoothing step.

Numerical tests showed that the following simpler rule without backtracking works well in practice: Perform a smoothing step $y_{k+1}^n \mapsto y_{k+1}$ if (6.5) holds, and choose $y_{k+1} = y_{k+1}^n$, otherwise.

So far, we have not described how smoothing steps can be obtained. We do this now for the case of NCP reformulations.

Example 6.7 (smoothing steps for NCPs). We consider operators arising from nonsmooth reformulations of NCPs as described in Example 5.5 and further investigated in the Examples 5.11 and 6.2. The following construction of a smoothing step follows an idea in [38]; see also [64]. In addition to the assumptions stated in Example 5.5, let us assume that the operator $Z : L^p(\Omega) \to L^r(\Omega)$ assumes the form $Z(y) = G(y) + \lambda y$, where $\lambda \in L^\infty(\Omega)$ is positive and bounded away from zero, and $G : L^r(\Omega) \mapsto L^p(\Omega)$ is Lipschitz continuous. Note that $G(y)$ is smoother than its

preimage y, since $L^p(\Omega) \subset L^r(\Omega)$ with nonequivalent norms. This form of Z arises, e.g., in the first-order necessary optimality conditions of a large class of optimal control problems with bounds on the control and L^2-regularization [38, 60, 62, 63, 64]. In particular, in Example 5.6(2) we already observed this structure of Z when we considered the elliptic control problem (1.6). This is further discussed in section 6.2.

It is well known and easy to verify that $\bar{y} \in L^p(\Omega)$ solves the NCP if and only if

$$S(\bar{y}) \stackrel{\text{def}}{=} (\bar{y} - \lambda^{-1} Z(\bar{y}))_+ = \bar{y},$$

where $u_+(\omega) \stackrel{\text{def}}{=} \max\{u(\omega), 0\}$. Further, for all $y \in L^r(\Omega)$ there holds $S(y) = \lambda^{-1} G(y)_-$ with $u_- \stackrel{\text{def}}{=} (-u)_+$. Hence, using $|u_- - v_-| \le |u - v|$, we obtain for all $y \in L^r(\Omega)$

$$|S(y) - \bar{y}| = |S(y) - S(\bar{y})| = \lambda^{-1} \left| G(y)_- - G(\bar{y})_- \right| \le \lambda^{-1} \left| G(y) - G(\bar{y}) \right|,$$

and therefore

$$\|S(y) - \bar{y}\|_{L^p} \le \|\lambda^{-1}\|_{L^\infty} \|G(y) - G(\bar{y})\|_{L^p} \le L_G \|\lambda^{-1}\|_{L^\infty} \|y - \bar{y}\|_{L^r},$$

where L_G is the Lipschitz constant of G. This shows that the mapping $y_k^n \mapsto y_k \stackrel{\text{def}}{=} S(y_k^n)$ is a smoothing step with $C_S = L_G \|\lambda^{-1}\|_{L^\infty}$ for $Y = L^p(\Omega)$ and $Y_0 = L^r(\Omega)$.

Remark 6.8. The idea of constructing smoothing steps can actually be used for a reformulation so that no smoothing step is required in the Newton method because the resulting operator is semismooth from $L^r(\Omega)$ to $L^r(\Omega)$. The following approach can be found in more generality in [62] and is motivated by [30]. In fact, let $\lambda > 0$, and consider the NCP with $Z(y) = G(y) + \lambda y$, where we assume that, for suitable $1 \le r < p \le \infty$, the operator $G : L^r(\Omega) \to L^r(\Omega)$ is continuously Fréchet-differentiable and that $G : L^r(\Omega) \to L^p(\Omega)$ is locally Lipschitz continuous. According to Example 6.7, the NCP is equivalent to

$$y - S(y) = 0.$$

The operator $S(y) = (y - \lambda^{-1} Z(y))_+ = \max\{-\lambda^{-1} G(y), 0\}$ is a superposition operator $S(y) = \Psi(y) = \psi(F(y))$ with

$$\psi(t) = \max\{-\lambda^{-1} t, 0\}$$

and $F(y) = G(y)$. Application of Theorem 5.2 with $m = 1$, $Y = L^r(\Omega)$, $r_1 = r$, and $q_1 = p$ now yields the semismoothness of the operator $\Psi : L^r(\Omega) \to L^r(\Omega)$. Choosing the special NCP-function $\phi(x) = x_1 - \max\{x_1 - \lambda^{-1} x_2, 0\} = \min\{x_1, \lambda^{-1} x_2\}$, we see that

$$\Phi(y) = \phi(y, Z(y)) = y - \max\{y - \lambda^{-1} Z(y), 0\} = y - S(y)$$

holds. It is straightforward to verify the identity $\partial^\circ \Phi(y) = I - \partial^\circ \Psi(y)$. Due to the semismoothness of Ψ, we obtain that $\Phi : L^r(\Omega) \to L^r(\Omega)$ is semismooth:

$$\begin{aligned} &\sup_{M \in \partial^\circ \Phi(y+s)} \|\Phi(y+s) - \Phi(y) - Ms\|_{L^r} \\ = &\sup_{M \in \partial^\circ \Psi(y+s)} \|\Psi(y+s) - \Psi(y) - Ms\|_{L^r} = o(\|s\|_{L^r}) \quad \text{as } \|s\|_{L^r} \to 0. \end{aligned}$$

Application of the semismooth Newton iteration to $\Phi(y) = 0$ then essentially results in the method described and analyzed in [30], which is equivalent to the primal dual active set strategy [9].

6.2. Application to a control problem. In this section we show how Algorithm 6.3 can be used to solve the constrained elliptic control problem (1.6). In Example 5.6 we converted (1.6) to an equivalent NCP and analyzed its properties. We recall that our choices were $Y = L^p(\Omega)$, $r_1 = r_2 = r = 2$, and $q_1 = q_2 = q = p$ with $p > 2$ as in (5.12). We observed that the operator Z meets all of the assumptions of Example 5.5. Therefore, the superposition operator Φ resulting from a reformulation as operator equation (5.6) is semismooth if the NCP-function ϕ is Lipschitz continuous and semismooth. The operator Φ is β-order semismooth with β as in Theorem 5.2 if, in addition, ϕ is α-order semismooth and (5.3) holds. Let us choose $Y_0 = L^2(\Omega)$.

For the application of Algorithm 6.3, several operations have to be performed. For convenience, we drop the index k and denote by $y \in L^p(\Omega)$ the current iterate.

We first describe the computation of

$$Z(y) = G(y) + \lambda y, \quad G(y) = (A^{-1})^*(A^{-1}(y-b) + u_d) + \lambda(w_d - b).$$

This requires the solution of two elliptic equations, the state equation (1.6b) with right-hand side $w = b - y$ and the *adjoint equation*

$$-\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ji}\frac{\partial v}{\partial x_j}\right) = u_d - u \quad \text{on } \Omega; \tag{6.6}$$

then $G(y) = v + \lambda(w_d - b)$ and $Z(y) = G(y) + \lambda y$. Now $\Phi(y)$ is easily obtained by applying the NCP-function ϕ pointwise to the pair of functions $(y, Z(y))$.

Next, we need to know what an element M of the generalized differential $\partial^\circ \Phi(y)$ looks like. We have

$$\begin{aligned} M = d_1 \cdot I + d_2 \cdot Z'(y) &= d_1 \cdot I + d_2 \cdot ((A^{-1})^* A^{-1} + \lambda I) \\ &= (d_1 + \lambda d_2) \cdot I + d_2 \cdot (A^{-1})^* A^{-1}, \end{aligned}$$

with $d_i \in L^\infty(\Omega)$, $(d_1, d_2) \in \partial\phi(y, z)$ a.e. on Ω, where $z = Z(y)$. For $\phi = \phi_{FB}$ we obtain $(d_1, d_2) = \phi'(y, z)$ pointwise a.e. on $\{x : \big(y(x), z(x)\big) \neq (0,0)\}$ and $(d_1, d_2) \in \partial\phi(0,0) = \{(\tau_1 - 1, \tau_2 - 1) : \tau_1^2 + \tau_2^2 \leq 1\}$ pointwise a.e. on $\{x : \big(y(x), z(x)\big) = (0,0)\}$. It is easy to see that $d_1, d_2 \leq 0$ and $2 - \sqrt{2} \leq |d_1 + d_2| \leq 2 + \sqrt{2}$ a.e. on Ω. Other NCP-functions, e.g., $\phi(s,t) = \min\{s,t\}$, have similar properties. Therefore, the Newton system in step 2 of Algorithm 6.3 assumes the form

$$ds + d_2 \cdot (A^{-1})^* A^{-1} s = -\Phi(y) \quad \text{with} \quad d = d_1 + \lambda d_2. \tag{6.7}$$

Note that d and d^{-1} are bounded in L^∞ and that $d_1 d_2 \geq 0$.

We briefly sketch the way in which (6.7) can be solved efficiently by multigrid methods. With $s_1 = A^{-1}s$ and $s_2 = (A^{-1})^* s_1$ we have $s = -d^{-1}(\Phi(y) + d_2 s_2)$, and $s_1, s_2 \in H_0^1(\Omega)$ solve the weakly coupled elliptic system

$$As_1 = -d^{-1}\Phi(y) - d^{-1}d_2 s_2, \qquad A^* s_2 = s_1. \tag{6.8}$$

This system can be solved very efficiently by multigrid methods; see, e.g., [27, section 11]. Alternatively, we can eliminate s_2 from (6.8) and obtain the compact fixed point problem

$$s_1 = -A^{-1}\big(d^{-1}\Phi(y) + d^{-1}d_2 \cdot (A^{-1})^* s_1\big),$$

to which a multigrid method of the second kind [27, section 16] can be applied. Within each iteration, premultiplication by A^{-1} and $(A^{-1})^*$ has to be performed, which again can be done by invoking fast solvers.

Finally, a smoothing step is required. Note that the operator $Z(y) = G(y)+\lambda y$ has exactly the structure we need to construct smoothing steps as described in Example 6.7, since G maps continuous affine linearly (and thus Lipschitz continuously) to $L^p(\Omega)$, with $p > 2$ as in (5.12). Computation of a smoothing step is not cheap, since, as shown above, evaluation of G requires us to solve two elliptic equations, the state equation (1.6b) and the adjoint equation (6.6). Therefore, it is advantageous to avoid smoothing steps if possible, which can be done by using the heuristics that we developed in section 6.1. Alternatively, the smoothing step can be avoided by using the special reformulation described in Remark 6.8.

The regularity condition in Assumption 6.1 can be verified by using either of the sufficient conditions derived in [63] and [62]; see those papers for details. The convergence results of Theorem 6.5 are thus applicable.

We end this section by addressing discretization. For simplicity, we consider a finite difference approximation on a regular computational grid covering Ω and consisting of N interior grid points. Corresponding to the functions u, w, u_d, w_d, b, we obtain the grid functions $\mathbf{u}, \mathbf{w}, \mathbf{u_d}, \mathbf{w_d}, \mathbf{b} \in \mathbb{R}^N$, which represent the node values. Furthermore, using an appropriate finite-difference stencil, we obtain the discrete state equation

$$\mathbf{A}\mathbf{u} = \mathbf{w}, \tag{6.9}$$

where $\mathbf{A} \in \mathbb{R}^{N\times N}$ approximates the differential operator A. Let the diagonal matrix $\mathbf{L} \in \mathbb{R}^{N\times N}$ represent the discrete L^2-inner product, e.g., $\mathbf{L}_{ii} = h^n$ if the grid is equidistant with step size h. The discrete objective function is

$$\mathbf{J}(\mathbf{w}) = \frac{1}{2}(\mathbf{A}^{-1}\mathbf{w} - \mathbf{u_d})^T\mathbf{L}(\mathbf{A}^{-1}\mathbf{w} - \mathbf{u_d}) + \frac{\lambda}{2}(\mathbf{w} - \mathbf{w_d})^T\mathbf{L}(\mathbf{w} - \mathbf{w_d}).$$

The pointwise control constraint $w \le b$ is discretized by $\mathbf{w} \le \mathbf{b}$ (componentwise). The *Euclidean* gradient of $\mathbf{J}$ is

$$\nabla\mathbf{J}(\mathbf{w}) = (\mathbf{A}^{-1})^T\mathbf{L}(\mathbf{A}^{-1}\mathbf{w} - \mathbf{u_d}) + \lambda\mathbf{L}(\mathbf{w} - \mathbf{w_d}).$$

For proper scaling, we have to transform $\nabla\mathbf{J}(\mathbf{w})$ to the discrete L^2 inner product represented by $\mathbf{L}$. The resulting gradient, which is the discrete counterpart of $\nabla J(w)$, is given by

$$\mathbf{J}'(\mathbf{w}) \stackrel{\text{def}}{=} \mathbf{L}^{-1}\nabla\mathbf{J}(\mathbf{w}) = \mathbf{L}^{-1}(\mathbf{A}^{-1})^T\mathbf{L}(\mathbf{A}^{-1}\mathbf{w} - \mathbf{u_d}) + \lambda(\mathbf{w} - \mathbf{w_d}).$$

The discrete optimality system reads (noting that $\mathbf{L}$ is positive diagonal)

$$\mathbf{w}_i \le \mathbf{b}_i, \quad \mathbf{J}'_i(\mathbf{w}) \le 0, \quad (\mathbf{w} - \mathbf{b})_i\mathbf{J}'_i(\mathbf{w}) = 0, \quad i = 1, \dots, N, \tag{6.10}$$

and corresponds to (5.11). As in the continuous case, we introduce $\mathbf{y} = \mathbf{b} - \mathbf{w}$ and

$$\mathbf{Z} : \mathbb{R}^N \to \mathbb{R}^N, \quad \mathbf{Z}(\mathbf{y}) = -\mathbf{J}'(\mathbf{b} - \mathbf{y}).$$

Then (6.10) is equivalent to the finite-dimensional NCP (1.2) with $k = N$, $y = \mathbf{y}$, and $Z = \mathbf{Z}$. We apply an NCP-function ϕ to write the NCP equivalently in the form

$$\Phi(\mathbf{y}) = 0, \quad \text{where} \quad \Phi(\mathbf{y}) = \big(\phi(\mathbf{y}_1, \mathbf{Z}_1(\mathbf{y})), \dots, \phi(\mathbf{y}_N, \mathbf{Z}_N(\mathbf{y}))\big)^T.$$

As a discretization of $\partial^\circ\Phi$ we choose $\partial^\circ\Phi(\mathbf{y})$, the set of all matrices $\mathbf{M} \in \mathbb{R}^{N\times N}$,

$$\mathbf{M} = \mathbf{D}_1 + \mathbf{D}_2\mathbf{Z}'(\mathbf{y}) = \mathbf{D}_1 + \mathbf{D}_2\big(\mathbf{L}^{-1}(\mathbf{A}^{-1})^T\mathbf{L}\mathbf{A}^{-1} + \lambda\mathbf{I}\big),$$
$$\mathbf{D}_1, \mathbf{D}_2 \in \mathbb{R}^{N\times N} \ \text{ diagonal}, \quad (\mathbf{D}_1, \mathbf{D}_2)_{ii} \in \partial\phi(\mathbf{y}_i, \mathbf{Z}_i(\mathbf{y})).$$

As discussed earlier, we have for the ith row of $\partial^\circ\Phi(\mathbf{y})$

$$[\partial^\circ\Phi(\mathbf{y})]_i = \partial\phi(\mathbf{y}_i, \mathbf{Z}_i(\mathbf{y}))\frac{d}{d\mathbf{y}}\begin{pmatrix}\mathbf{y}_i\\ \mathbf{Z}_i(\mathbf{y})\end{pmatrix} \supset \partial\Phi_i(\mathbf{y})$$

by the chain rule for generalized gradients, with equality if, e.g., ϕ or $-\phi$ is regular. Therefore, $\partial_C\Phi(\mathbf{y}) \subset \partial^\circ\Phi(\mathbf{y})$, and hence we can choose $\mathbf{M}$ as in the ordinary finite-dimensional semismooth Newton method. Since computing elements of $\partial^\circ\Phi(\mathbf{y})$ can be easier than computing those of $\partial_C\Phi(\mathbf{y})$, we point out that the estimate

$$\sup_{\mathbf{M}\in\partial^\circ\Phi(\mathbf{y}+\mathbf{s})} \|\Phi(\mathbf{y}+\mathbf{s}) - \Phi(\mathbf{y}) - \mathbf{M}\mathbf{s}\| = o(\|\mathbf{s}\|) \quad \text{as } \mathbf{s}\to 0$$

is easy to prove if ϕ is semismooth and $\mathbf{Z}$ is continuously differentiable (which is the case here). The estimate holds with "$o(\|\mathbf{s}\|)$" replaced by "$O(\|\mathbf{s}\|^{1+\alpha})$" if ϕ is α-order semismooth, $0<\alpha\le 1$, and $\mathbf{Z}'$ is α-order Hölder continuous (which is the case here). Therefore, the discrete equivalent of the infinite-dimensional semismooth Newton method converges q-superlinearly to regular solutions (with order $1+\alpha$ in the case of α-order $\partial^\circ\Phi$-semismoothness); see Proposition 2.7. For numerical results, we refer to [63, 62].

7. Semismooth composite operators and chain rules. In this section we show that our class of semismoothness operators is closed under composition, which is helpful, e.g., for proving semismoothness of a particular operator by breaking it up into simpler pieces. Furthermore, we establish chain rules for composite operators. We consider the scenario in which $F = G\circ H$ is a composition of the operators

$$G : X \mapsto \prod\nolimits_i L^{r_i}(\Omega), \qquad H : Y \mapsto X,$$

with X a Banach space, and in which $\psi = \psi_1\circ\psi_2$ is a composition of the functions

$$\psi_1 : \mathbb{R}^l \to \mathbb{R}, \qquad \psi_2 : \mathbb{R}^m \to \mathbb{R}^l.$$

We impose assumptions on ψ_1, ψ_2, G, and H to ensure that F and ψ satisfy Assumption 3.1. The following is one way to do this.

Assumption 7.1. There are $1\le r\le r_i < q_i \le\infty$, $1\le i\le m$, such that

(a) the operators $G : X\to\prod_i L^{r_i}(\Omega)$ and $H : Y\to X$ are continuously Fréchet differentiable;
(b) the operator G maps X locally Lipschitz continuously into $L^{q_i}(\Omega)$;
(c) the functions ψ_1 and ψ_2 are Lipschitz continuous;
(d) ψ_1 and ψ_2 are semismooth.

It is straightforward to strengthen these assumptions such that they imply Assumption 3.4. For brevity, we will not discuss the extension of the next theorem to semismoothness of order β, which is easily established by slight modifications of the assumptions and the proofs.

THEOREM 7.2. *Let Assumption* 7.1 *hold, and let* $F = G\circ H$ *and* $\psi = \psi_1\circ\psi_2$. *Then the following hold:*

(a) F *and* ψ *satisfy Assumption* 3.1.
(b) Ψ *as defined in* (1.8) *is semismooth.*
(c) *The operator* $\Psi_G : z\in X\mapsto\psi(G(z))\in L^r(\Omega)$ *is semismooth and the following* chain rule *holds:*

$$\partial^\circ\Psi(y) = \partial^\circ\Psi_G(H(y))H'(y) = \{M_G H'(y) : M_G\in\partial^\circ\Psi_G(H(y))\}.$$

(d) *If $l = 1$ and ψ_1 is strictly differentiable* (*see* [13, p. 30]), *then the operator $\Psi_2 : y \in Y \mapsto \psi_2(F(y)) \in L^r(\Omega)$ is semismooth and the following* chain rule *holds:*

$$\partial^\circ \Psi(y) = \psi_1'(\Psi_2(y))\partial^\circ \Psi_2(y) = \{\psi_1'(\Psi_2(y)) \cdot M_2 : M_2 \in \partial^\circ \Psi_2(y)\}.$$

Proof. (a) Assumption 7.1(a) implies Assumption 3.1(a); Assumption 3.1(b) follows from Assumption 7.1(a),(b); Assumption 7.1(c) implies Assumption 3.1(c); and Assumption 3.1(d) holds by Assumption 7.1(d), since the composition of semismooth functions is semismooth.

(b) By (a), we can apply Theorem 5.2.

(c) Assumption 7.1 implies Assumption 3.1 with G and X instead of F and Y. Hence, Ψ_G is semismooth by Theorem 5.2.

For the proof of the "$\subset$" part of the chain rule, let $M \in \partial^\circ \Psi(y)$ be arbitrary. By definition, there exists a measurable selection d of $\partial\psi(F(y))$ such that

$$M = \sum\nolimits_i d_i \cdot F_i'(y).$$

Now, since $F_i'(y) = G_i'(H(y))H'(y)$,

$$M = \sum\nolimits_i d_i \cdot G_i'\big(H(y)\big)H'(y) = M_G H'(y), \quad \text{where}$$

$$M_G = \sum\nolimits_i d_i \cdot G_i'\big(H(y)\big). \tag{7.1}$$

Obviously, we have $M_G \in \partial^\circ \Psi_G(H(y))$.

To prove the reverse inclusion, note that any $M_G \in \partial^\circ \Psi_G(H(y))$ assumes the form (7.1) with appropriate measurable selection $d \in \partial\psi(F(y))$. Then

$$M_G H'(y) = \sum\nolimits_i d_i \cdot \big(G_i'(H(y))H'(y)\big) = \sum\nolimits_i d_i \cdot F_i'(y),$$

which shows $M_G H'(y) \in \partial^\circ \Psi\big(y\big)$.

(d) Certainly, F and ψ_2 satisfy Assumption 3.1 (with ψ_2 replaced by ψ). Hence, Theorem 5.2 yields the semismoothness of Ψ_2. We proceed by noting that a.e. on Ω

$$\psi_1'(\Psi_2(y)(\omega))\partial\psi_2(F(y)(\omega)) = \partial\psi(F(y)(\omega)) \tag{7.2}$$

holds, where we have applied the chain rule for generalized gradients [13, Theorem 2.3.9] and the identity $\partial\psi_1 = \{\psi_1'\}$; see [13, Proposition 2.2.4].

We first prove the "$\supset$" direction of the chain rule. Let $M_2 \in \partial^\circ \Psi_2$ be arbitrary. It assumes the form

$$M_2 = \sum\nolimits_i \hat{d}_i \cdot F_i'(y),$$

where $\hat{d} \in L^\infty(\Omega)^m$ is a measurable selection of $\partial\psi_2(F(y))$. Now for any operator M contained in the right-hand side of the assertion we have with $d \overset{\text{def}}{=} \psi_1'(\Psi_2(y))\hat{d}$

$$M = \psi_1'\big(\Psi_2(y)\big) \cdot M_2 = \sum\nolimits_i d_i \cdot F_i'(y).$$

Obviously, $d \in L^\infty(\Omega)^m$ and, by (7.2), d is a measurable selection of $\partial\psi(F(y))$. Hence, $M \in \partial^\circ \Psi(y)$.

Conversely, to prove "$\subset$", let $M \in \partial^\circ \Psi(y)$ be arbitrary, and denote by $d \in L^\infty(\Omega)^m$ the corresponding measurable selection of $\partial\psi(F(y))$. Now let $\tilde{d} \in L^\infty(\Omega)^m$ be a measurable selection of $\partial\psi_2(F(y))$, and define $\hat{d} \in L^\infty(\Omega)^m$ by

$$\hat{d}(\omega) = \tilde{d}(\omega) \ \text{ on } \Omega_0 = \{\omega : \psi_1'(\Psi_2(y)(\omega)) = 0\}, \quad \hat{d}(\omega) = \frac{d(\omega)}{\psi_1'(\Psi_2(y)(\omega))} \ \text{ on } \Omega \setminus \Omega_0.$$

Then $\hat{d}$ is measurable and $d = \psi_1'(\Psi_2(y))\hat{d}$. Further, $\hat{d}(\omega) = \tilde{d}(\omega) \in \partial\psi_2(F(y))$ on Ω_0 and, using (7.2),

$$\hat{d}(\omega) = \frac{d(\omega)}{\psi_1'(\Psi_2(y)(\omega))} \in \frac{\psi_1'(\Psi_2(y)(\omega))\partial\psi_2(F(y))}{\psi_1'(\Psi_2(y)(\omega))} = \partial\psi_2(F(y)) \quad \text{on } \Omega \setminus \Omega_0.$$

Thus, $\hat{d}$ is a measurable selection of $\partial\psi_2(F(y))$, and consequently also $\hat{d} \in L^\infty(\Omega)^m$ due to the Lipschitz continuity of ψ_2. Therefore,

$$M_2 = \sum\nolimits_i \hat{d}_i \cdot F_i'(y) \in \partial^\circ\Psi_2(y),$$

and thus $M \in \psi_1'(\Psi_2(y)) \cdot \partial^\circ\Psi_2(y)$, as asserted. ∎

8. Further properties of the generalized differential. We now establish that our generalized differential is convex-valued, weak compact-valued, and weakly graph closed. These properties can provide a basis for future research on the connections between $\partial^\circ\Psi$ and other generalized differentials, particularly the Thibault generalized differential [58] and the Ioffe–Ralph generalized differential [32, 53]. As weak topology on $\mathcal{L}(Y, L^r)$ we use the weak operator topology, which is defined by the seminorms $M \mapsto |\langle w, Mv\rangle_\Omega|$, $v \in Y$, $w \in L^{r'}(\Omega)$, the dual space of $L^r(\Omega)$.

The following result will be of importance.

LEMMA 8.1. *Under Assumption* 3.1*, the set* $K(y)$ *defined in* (4.4) *is convex and weak* sequentially compact in* $L^\infty(\Omega)^m$ *for all* $y \in Y$*.*

Proof. From Lemma 4.7 we know that $K(y) \subset L_\psi \bar{B}^m_{L^\infty}$ is nonempty and bounded. Further, the convexity of $\partial\psi(x)$ implies the convexity of $K(y)$. Now let $s_k \in K(y)$ tend to s in $L^2(\Omega)^m$. Then for a subsequence $s_{k'}(\omega) \to s(\omega)$ holds for almost all $\omega \in \Omega$. Since $\partial\psi(u(\omega))$ is compact, this implies that for almost all $\omega \in \Omega$, $s(\omega) \in \partial\psi(u(\omega))$ holds and thus $s \in K(y)$. Hence, $K(y)$ is a bounded, closed, and convex subset of $L^2(\Omega)^m$ and therefore weak sequentially compact in $L^2(\Omega)^m$. Therefore, $K(y)$ is also weak* sequentially closed in $L^\infty(\Omega)^m$, for, if $(s_k) \subset K(y)$ converges weakly* to s in $L^\infty(\Omega)^m$, then $\langle w, s_k - s\rangle_\Omega \to 0$ for all $w \in L^1(\Omega)^m \supset L^2(\Omega)^m$, showing that $s_k \to s$ weakly in $L^2(\Omega)^m$. Thus, $K(y)$ is weak* sequentially closed and bounded in $L^\infty(\Omega)^m$. Since $L^1(\Omega)^m$ is separable, this yields that $K(y)$ is weak* sequentially compact. ∎

8.1. Convexity and weak compactness. As further useful properties of $\partial^\circ\Psi$ we establish the convexity and weak compactness of its images as follows.

THEOREM 8.2. *Under Assumption* 3.1*, the generalized differential* $\partial^\circ\Psi(y)$ *is nonempty, convex, and weakly sequentially compact for all* $y \in Y$*. If* Y *is separable, then* $\partial^\circ\Psi(y)$ *is also weakly compact for all* $y \in Y$*.*

Proof. The nonemptiness was already stated in Theorem 4.8. Convexity follows immediately from the convexity of the set $K(y)$ derived in Lemma 4.7. We now prove weak sequential compactness. Let $(M_k) \subset \partial^\circ\Psi(y)$ be any sequence. Then

$$M_k = \sum\nolimits_i d_{ki} \cdot F_i'(y),$$

with $d_k \in K(y)$; see (4.4). Lemma 8.1 yields that $K(y)$ is weak* sequentially compact in $L^\infty(\Omega)^m$. Hence, we can select a subsequence such that (d_k) converges weak* to $d^* \in K(y)$ in $L^\infty(\Omega)^m$. Define $M^* = \sum_i d_i^* \cdot F_i(y)$ and observe that $M^* \in \partial^\circ\Psi(y)$ since $d^* \in K(y)$. It remains to prove that $M_k \to M^*$ weakly. Let $w \in L^{r'}(\Omega) = L^r(\Omega)'$ and $v \in Y$ be arbitrary. We set $z_i = w \cdot F_i'(y)v$ and note that $z_i \in L^1(\Omega)$. Hence,

$$\begin{aligned} |\langle w, (M_k - M^*)v\rangle_\Omega| &\le \sum\nolimits_i |\langle w, (d_k - d^*)_i \cdot F_i'(y)v\rangle_\Omega| \\ &= \sum\nolimits_i |\langle z_i, (d_k - d^*)_i\rangle_\Omega| \longrightarrow 0 \quad \text{as } k \to \infty. \end{aligned} \tag{8.1}$$

Therefore, the weak sequential compactness is shown.

By Lemma 4.3, $\partial^\circ\Psi(y)$ is contained in a closed ball in $\mathcal{L}(Y, L^r)$, on which the weak topology is metrizable if Y is separable (note that $1 \le r < \infty$ implies that $L^r(\Omega)$ is separable). Hence, in this case the weak compactness follows from the weak sequential compactness. ∎

8.2. Weak graph closedness of the generalized differential. Finally, we prove that the multifunction $\partial^\circ\Psi$ is weakly graph closed as follows.

THEOREM 8.3. *Let Assumption* 3.1 *be satisfied, and let* $(y_k) \subset Y$ *and* $(M_k) \subset \mathcal{L}(Y, L^r(\Omega))$ *be sequences such that* $M_k \in \partial^\circ\Psi(y_k)$ *for all* k, $y_k \to y^*$ *in* Y, *and* $M_k \to M^*$ *weakly in* $\mathcal{L}(Y, L^r(\Omega))$. *Then* $M^* \in \partial^\circ\Psi(y^*)$ *holds. If, in addition,* Y *is separable, then the above assertion also holds if we replace the sequences* (y_k) *and* (M_k) *by nets.*

Proof. Let $y_k \to y^*$ in Y and $\partial^\circ\Psi(y_k) \ni M_k \to M^*$ weakly. We have the representations $M_k = \sum_i d_{ki} \cdot F_i'(y_k)$ with measurable selections d_k of $\partial\psi(u_k)$, where $u_k = F(y_k)$. We also introduce $u^* = F(y^*)$. The multifunction $\omega \in \Omega \mapsto \partial\psi(u^*(\omega))$ is closed-valued (even compact-valued) and measurable. Furthermore, the function $(\omega, h) \mapsto \|d_k(\omega) - h\|_2$ is a normal integrand on $\Omega \times \mathbb{R}^m$ [54, Corollary 2P]. Hence, by [54, Theorem 2K], the multifunctions $S_k : \Omega \to \mathbb{R}^m$,

$$S_k(\omega) = \operatorname*{arg\,min}_{h \in \partial\psi(u^*(\omega))} \|d_k(\omega) - h\|_2 ,$$

are closed-valued (even compact-valued) and measurable. We choose measurable selections s_k of S_k. The sequence (s_k) is contained in the (by Lemma 8.1) sequentially weak* compact set $K(y^*) \subset L^\infty(\Omega)^m$. Further, by Lemma 4.7, we have $d_k \in L_\psi \bar{B}^m_{L^\infty}$.

Hence, by transition to subsequences, we achieve $s_k \to \bar{s} \in K(y^*)$ weak* in $L^\infty(\Omega)^m$, and $d_k \to \bar{d} \in L_\psi \bar{B}^m_{L^\infty}$ weak* in $L^\infty(\Omega)^m$. Therefore, $(d_k - s_k) \to (\bar{d} - \bar{s})$ weak* in $L^\infty(\Omega)^m$ and thus also weakly in $L^2(\Omega)^m$. Since $u_k \to u^*$ in $\prod_i L^{q_i}(\Omega)$, we achieve by transition to a further subsequence that $u_k \to u^*$ a.e. on Ω. Hence, since $d_k(\omega) \in \partial\psi(u_k(\omega))$ for almost all $\omega \in \Omega$ and $\partial\psi$ is upper semicontinuous, we obtain from the construction of s_k that $(d_k - s_k) \to 0$ a.e. on Ω. The sequence $(d_k - s_k)$ is bounded in $L^\infty(\Omega)^m$ and thus the Lebesgue convergence theorem yields $(d_k - s_k) \to 0$ in $L^2(\Omega)^m$. From $(d_k - s_k) \to 0$ and $(d_k - s_k) \to (\bar{d} - \bar{s})$ weakly in $L^2(\Omega)^m$, we see $\bar{d} = \bar{s}$. We thus have

$$d_k \to \bar{d} = \bar{s} \in K(y^*) \quad \text{weak}^* \text{ in } L^\infty(\Omega)^m.$$

This shows that $\bar{M} \overset{\text{def}}{=} \sum_i \bar{d}_i \cdot F_i'(y^*) \in \partial^\circ\Psi(y^*)$. It remains to prove that $M_k \to \bar{M}$ weakly. To show this, let $w \in L^{r'}(\Omega) = L^r(\Omega)'$ and $v \in Y$ be arbitrary. Then with $z_{ki} = w \cdot F_i'(y_k)v$ and $z_i = w \cdot F_i'(y^*)v$, $z_{ki}, z_i \in L^1(\Omega)$ holds and

$$\|z_{ki} - z_i\|_{L^1} \le \|w\|_{L^{r'}} \|F_i'(y_k)v - F_i'(y^*)v\|_{L^r} \to 0 \quad \text{as } k \to \infty.$$

Hence we obtain, similarly as in (8.1),

$$\begin{aligned} |\langle w, (M_k - \bar{M})v\rangle_\Omega| &\le \sum\nolimits_i \left|\langle w, d_{ki} \cdot F_i'(y_k)v - \bar{d}_i \cdot F_i'(y^*)v\rangle_\Omega\right| \\ &= \sum\nolimits_i \left|\langle d_{ki}, z_{ki}\rangle_\Omega - \langle \bar{d}_i, z_i\rangle_\Omega\right| \\ &\le \sum\nolimits_i \left(|\langle \bar{d}_i - d_{ki}, z_i\rangle_\Omega| + \|d_{ki}\|_{L^\infty} \|z_i - z_{ki}\|_{L^1}\right) \to 0 \quad \text{as } k \to \infty. \end{aligned}$$

This implies that $M^* = \bar{M} \in \partial^\circ\Psi(y^*)$ and completes the proof of the first assertion.

Now let $(y_\kappa) \subset Y$ and $(M_\kappa) \subset \mathcal{L}(Y, L^r(\Omega))$ be nets such that $M_\kappa \in \partial^\circ\Psi(y_\kappa)$ for all κ, $y_\kappa \to y^*$ in Y, and $M_\kappa \to M$ weakly in $\mathcal{L}(Y, L^r(\Omega))$. Since (y_κ) finally stays in

any neighborhood of y^* and since F' is continuous, we see from (4.3) that without loss of generality we may assume that (M_κ) is contained in a bounded ball $\mathcal{B} \subset \mathcal{L}(Y, L^r)$. Since, due to the assumed separability of Y, $\mathcal{B}$ is metrizable with respect to the weak topology, we see that we can work with sequences instead of nets. ∎

9. Concluding remarks and future work. In this work, a new semismoothness theory for superposition operators in function spaces was developed. Our semismoothness concept uses a new generalized differential that generalizes Qi's finite-dimensional C-subdifferential. The developed results were shown to be applicable to NCP-function-based reformulations of NCPs posed in function spaces. Using this semismoothness theory, a Newton-like method for nonsmooth operator equations was developed, which, depending on the order of semismoothness of the operator, converges q-superlinearly or with q-order $1 + \alpha$ to a regular solution. For illustration, the application of the algorithm to the control-constrained optimal control of an elliptic partial differential equation was discussed in detail. We also established the semismoothness of composite operators and developed corresponding chain rules. Furthermore, the multifunction $\partial^\circ \Psi$ was shown to have several useful properties, in particular weak graph closedness, which can be helpful, e.g., in the development of relationships between $\partial^\circ \Psi$ and other vector-valued generalized differentials.

In the author's Habilitation thesis [62], the presented results are further developed in various directions. In particular, it is shown how our semismooth Newton method can be extended to handle mixed problems of the form

$$\Psi(y) = 0, \quad G(y) = 0,$$

where $G : Y \to Z$ is a smooth operator. This problem class includes reformulations of Karush–Kuhn–Tucker conditions for many optimal control and variational inequality problems. The main challenge hereby is the choice of a suitable regularity condition on the operators $(M, G'(y))$, $M \in \partial^\circ \Psi(y)$, and the development of sufficient conditions for regularity that extend the ones given in [61]. It is also possible to establish superlinear convergence of an inexact semismooth Newton method under a Dennis–Moré-type condition. A further interesting question is how our locally convergent Newton method can be made globally convergent in an efficient way. Here, one can use that the merit function $y \in Y \mapsto \|\Psi(y)\|_{L^2}^2 /2$ is continuously differentiable under reasonable assumptions, which are satisfied, e.g., for $\psi = \phi_{FB}$ and $q_i \geq 2$. Therefore, a convergence theory similar to the one developed in [65] for affine-scaling trust-region methods for bound-constrained nonlinear optimization in function spaces is transferable to our setting. For the finite-dimensional analogue of the presented algorithm, globalization techniques were developed in, e.g., [17, 19, 36, 61]. A particular trust-region globalization for our semismooth Newton method can be found in [62]. The proposed class of Newton methods was successfully applied to the elliptic control problem (1.6) (see [63]), nonlinear elliptic control problems [62], obstacle problems [62], and flow control problems [62, 60]. In all cases, the method was very efficient and achieved a superlinear rate of convergence. We plan further numerical tests and will report on the results in forthcoming papers.

We plan further investigations in the future. In particular, it would be interesting to establish the mesh-independence of the proposed semismooth Newton method. Also, the efficient implementation of the algorithm presents further challenges. In particular, the possibility of obtaining approximations of M_k by replacing $F_i'(y_k)$ with quasi-Newton matrices is a question that should be addressed. Furthermore, depending on the particular problem, multigrid methods can provide a powerful tool for the computation of Newton steps. We have sketched this approach briefly in section 6.2. Our preliminary numerical tests with multilevel semismooth Newton methods,

reported in [62], are very promising, and we are currently starting to investigate this multilevel approach in more detail.

Appendix.

A consequence of Hölder's inequality. The following estimate is frequently used in our analysis. It follows immediately from Hölder's inequality.

LEMMA A.1. *Let Ω be bounded, $1 \le p \le q \le \infty$, and*

$$c_{p,q}(\Omega) \stackrel{\text{def}}{=} \mu(\Omega)^{\frac{q-p}{pq}} \quad \text{if } p < q < \infty, \qquad c_{p,\infty}(\Omega) \stackrel{\text{def}}{=} \mu(\Omega)^{1/p} \quad \text{if } p < \infty,$$
$$c_{p,q}(\Omega) \stackrel{\text{def}}{=} 1 \quad \text{if } p = q.$$

Then for all $v \in L^q(\Omega)$ there holds

$$\|v\|_{L^p} \le c_{p,q}(\Omega)\, \|v\|_{L^q}.$$

Upper semicontinuity and measurability of multifunctions. For convenience, we also provide the definition of upper semicontinuity and measurability of multifunctions (see [13, 54]).

DEFINITION A.2. *A multifunction $\Gamma : U \rightrightarrows \mathbb{R}^l$ defined on $U \subset \mathbb{R}^k$ is* upper semicontinuous *at $x \in U$ if for all $\varepsilon > 0$ there exists $\delta > 0$ such that*

$$\Gamma(x') \subset \{z + h : z \in \Gamma(x),\ \|h\| < \varepsilon\} \quad \text{for all } x' \in U,\ \|x' - x\| < \delta.$$

DEFINITION A.3. *A multifunction $\Gamma : U \rightrightarrows \mathbb{R}^l$ defined on the measurable set $U \subset \mathbb{R}^k$ is called* measurable [54, p. 160] *if it is closed-valued and if for all closed (or open, or compact; see* [54, Proposition 1A]*) sets $C \subset \mathbb{R}^l$ the preimage*

$$\Gamma^{-1}(C) = \{x \in U : \Gamma(x) \cap C \neq \emptyset\}$$

is measurable.

Acknowledgments. The author would like to thank the referees for their helpful comments. This work was done while the author was visiting the Department of Computational and Applied Mathematics and the Center for Research on Parallel Computation at Rice University, which provided an excellent research environment. In particular, the author would like to thank John Dennis and Matthias Heinkenschloss for their hospitality and support.

REFERENCES

[1] W. ALT, *The Lagrange-Newton method for infinite-dimensional optimization problems*, Numer. Funct. Anal. Optim., 11 (1990), pp. 201–224.
[2] W. ALT, *Parametric optimization with applications to optimal control and sequential quadratic programming*, Bayreuth. Math. Schr., 35 (1991), pp. 1–37.
[3] W. ALT, *Sequential quadratic programming in Banach spaces*, in Advances in Optimization (Proceedings of the 6th French-German Colloquium on Optimization, Lambrecht, Germany, 1991), W. Oettli and D. Pallaschke, eds., Springer, Berlin, 1992, pp. 281–301.
[4] W. ALT AND K. MALANOWSKI, *The Lagrange-Newton method for nonlinear optimal control problems*, Comput. Optim. Appl., 2 (1993), pp. 77–100.
[5] W. ALT, R. SONTAG, AND F. TRÖLTZSCH, *An SQP method for optimal control of weakly singular Hammerstein integral equations*, Appl. Math. Optim., 33 (1996), pp. 227–252.
[6] J. APPELL, *The superposition operator in function spaces—A survey*, Expo. Math., 6 (1988), pp. 209–270.
[7] J. APPELL AND P. P. ZABREJKO, *Nonlinear Superposition Operators*, Cambridge University Press, Cambridge, 1990.
[8] M. BERGOUNIOUX, M. HADDOU, M. HINTERMÜLLER, AND K. KUNISCH, *A comparison of a Moreau–Yosida-based active set strategy and interior point methods for constrained optimal control problems*, SIAM J. Optim., 11 (2000), pp. 495–521.

[9] M. Bergounioux, K. Ito, and K. Kunisch, *Primal-dual strategy for constrained optimal control problems*, SIAM J. Control Optim., 37 (1999), pp. 1176–1194.
[10] J. M. Borwein and Q. J. Zhu, *A survey of subdifferential calculus with applications*, Nonlinear Anal., 38 (1999), pp. 687–773.
[11] B. Chen and N. Xiu, *A global linear and local quadratic noninterior continuation method for nonlinear complementarity problems based on Chen–Mangasarian smoothing functions*, SIAM J. Optim., 9 (1999), pp. 605–623.
[12] X. Chen, Z. Nashed, and L. Qi, *Smoothing methods and semismooth methods for nondifferentiable operator equations*, SIAM J. Numer. Anal., 38 (2000), pp. 1200–1216.
[13] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley and Sons, New York, 1983.
[14] F. H. Clarke, Y. S. Ledyaev, R. J. Stern, and P. R. Wolenski, *Nonsmooth Analysis and Control Theory*, Springer-Verlag, New York, 1998.
[15] T. F. Coleman, Y. Li, and A. Verma, *A Newton method for American option pricing*, J. Comput. Finance, 5 (2002).
[16] B. D. Craven and B. M. Glover, *An approach to vector subdifferentials*, Optimization, 38 (1996), pp. 237–251.
[17] T. De Luca, F. Facchinei, and C. Kanzow, *A semismooth equation approach to the solution of nonlinear complementarity problems*, Math. Programming, 75 (1996), pp. 407–439.
[18] G. Duvaut and J.-L. Lions, *Inequalities in Mechanics and Physics*, Springer-Verlag, Berlin, 1976; also Grundlehren Math. Wiss., 219.
[19] F. Facchinei and C. Kanzow, *A nonsmooth inexact Newton method for the solution of large-scale nonlinear complementarity problems*, Math. Programming, 76 (1997), pp. 493–512.
[20] F. Facchinei and J. Soares, *A new merit function for nonlinear complementarity problems and a related algorithm*, SIAM J. Optim., 7 (1997), pp. 225–247.
[21] M. C. Ferris and J. S. Pang, *Engineering and economic applications of complementarity problems*, SIAM Rev., 39 (1997), pp. 669–713.
[22] A. Fischer, *A special Newton-type optimization method*, Optimization, 24 (1992), pp. 269–284.
[23] A. Fischer, *Solution of monotone complementarity problems with locally Lipschitzian functions*, Math. Programming, 76 (1997), pp. 513–532.
[24] S. A. Gabriel and J.-S. Pang, *A trust region method for constrained nonsmooth equations*, in Large Scale Optimization (Proceedings of the Large Scale Optimization Conference, Gainesville, FL, 1993), W. W. Hager, D. W. Hearn, and P. M. Pardalos, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1994, pp. 155–181.
[25] B. M. Glover and D. Ralph, *First order approximations to nonsmooth mappings with application to metric regularity*, Numer. Funct. Anal. Optim., 15 (1994), pp. 599–620.
[26] R. Glowinski, J.-L. Lions, and R. Trémolières, *Numerical Analysis of Variational Inequalities*, North–Holland, Amsterdam, 1981.
[27] W. Hackbusch, *Multigrid Methods and Applications*, Springer-Verlag, Berlin, 1985.
[28] P. T. Harker and J.-S. Pang, *Finite-dimensional variational inequality and nonlinear complementarity problems: A survey of theory, algorithms and applications*, Math. Programming, 48 (1990), pp. 161–220.
[29] M. Heinkenschloss and F. Tröltzsch, *Analysis of the Lagrange-SQP-Newton method for the control of a phase field equation*, Control Cybernet., 28 (1999), pp. 177–211.
[30] M. Hintermüller, K. Ito, and K. Kunisch, *The Primal-Dual Active Set Strategy as Semismooth Newton Method*, Technical report 214, Spezialforschungsbereich F 003, Optimierung und Kontrolle, Karl-Franzens-Universität Graz, Graz, Austria, 2001.
[31] M. Hintermüller, K. Ito, and K. Kunisch, *The Primal-Dual Active Set Strategy as Semismooth Newton Method*, Technical report 214, Spezialforschungsbereich F 003, Optimierung und Kontrolle, Karl-Franzens-Universität Graz, Graz, Austria, Revised version, 2002.
[32] A. D. Ioffe, *Nonsmooth analysis: Differential calculus of nondifferentiable mappings*, Trans. Amer. Math. Soc., 266 (1981), pp. 1–56.
[33] V. Jeyakumar, *Simple Characterizations of Superlinear Convergence for Semismooth Equations via Approximate Jacobians*, Applied Mathematics Research Report AMR98/28, School of Mathematics, University of New South Wales, Sydney, New South Wales, Australia, 1998.
[34] V. Jeyakumar, *Solving B-Differentiable Equations*, Applied Mathematics Research Report AMR98/27, School of Mathematics, University of New South Wales, Sydney, New South Wales, Australia, 1998.
[35] V. Jeyakumar and D. T. Luc, *Approximate Jacobian matrices for nonsmooth continuous maps and C^1-optimization*, SIAM J. Control Optim., 36 (1998), pp. 1815–1832.
[36] H. Jiang, M. Fukushima, L. Qi, and D. Sun, *A trust region method for solving generalized complementarity problems*, SIAM J. Optim., 8 (1998), pp. 140–157.
[37] H. Jiang and L. Qi, *A new nonsmooth equations approach to nonlinear complementarity problems*, SIAM J. Control Optim., 35 (1997), pp. 178–193.
[38] C. T. Kelley and E. W. Sachs, *Multilevel algorithms for constrained compact fixed point*

problems, SIAM J. Sci. Comput., 15 (1994), pp. 645–667.
[39] D. Kinderlehrer and G. Stampacchia, *An Introduction to Variational Inequalities and Their Applications*, Academic Press/Harcourt Brace Jovanovich Publishers, New York, 1980.
[40] B. Kummer, *Newton's method for nondifferentiable functions*, in Advances in Mathematical Optimization, J. Guddat, B. Bank, H. Hollatz, P. Kall, D. Klatte, B. Kummer, K. Lommatzsch, K. Tammer, M. Vlach, and K. Zimmerman, eds., Akademie-Verlag, Berlin, 1988, pp. 114–125.
[41] B. Kummer, *Newton's method based on generalized derivatives for nonsmooth functions: Convergence analysis*, in Advances in Optimization (Proceedings of the 6th French-German Colloquium on Optimization, Lambrecht, Germany, 1991), W. Oettli and D. Pallaschke, eds., Springer, Berlin, 1992, pp. 171–194.
[42] O. L. Mangasarian, *Equivalence of the complementarity problem to a system of nonlinear equations*, SIAM J. Appl. Math., 31 (1976), pp. 89–92.
[43] R. Mifflin, *Semismooth and semiconvex functions in constrained optimization*, SIAM J. Control Optim., 15 (1977), pp. 959–972.
[44] V. S. Mikhalevich, A. M. Gupal, and V. I. Norkin, *Methods of Nonconvex Optimization*, Nauka, Moskow, 1987.
[45] T. S. Munson, F. Facchinei, M. C. Ferris, A. Fischer, and C. Kanzow, *The semismooth algorithm for large scale complementarity problems*, INFORMS J. Comput, 13 (2001), pp. 294–311.
[46] P. D. Panagiotopoulos, *Inequality Problems in Mechanics and Applications. Convex and Nonconvex Energy Functions*, Birkhäuser Boston, Boston, MA, 1985.
[47] J.-S. Pang, *A B-differentiable equation-based, globally and locally quadratically convergent algorithm for nonlinear programs, complementarity and variational inequality problems*, Math. Programming, 51 (1991), pp. 101–131.
[48] J.-S. Pang, *Complementarity problems*, in Handbook of Global Optimization, R. Horst and P. M. Pardalos, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1995, pp. 271–338.
[49] J.-S. Pang and L. Qi, *Nonsmooth equations: Motivation and algorithms*, SIAM J. Optim., 3 (1993), pp. 443–465.
[50] L. Qi, *Convergence analysis of some algorithms for solving nonsmooth equations*, Math. Oper. Res., 18 (1993), pp. 227–244.
[51] L. Qi, *C-differential operators, C-differentiability and generalized Newton methods*, Research Report AMR96/5, School of Mathematics, University of New South Wales, Sydney, New South Wales, Australia, 1996.
[52] L. Qi and J. Sun, *A nonsmooth version of Newton's method*, Math. Programming, 58 (1993), pp. 353–367.
[53] D. Ralph, *Rank*-1 *support functionals and the rank*-1 *generalized Jacobian, piecewise linear homeomorphisms*, Ph.D. thesis, Computer Sciences Department, University of Wisconsin, Madison, WI, 1990.
[54] R. T. Rockafellar, *Integral functionals, normal integrands and measurable selections*, in Nonlinear Operators and the Calculus of Variations, Lecture Notes in Math. 543, J. P. Gossez, E. J. Lami Dozo, J. Mawhin, and L. Waelbroeck, eds., Springer, Berlin, 1976, pp. 157–207.
[55] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
[56] A. Shapiro, *On concepts of directional differentiability*, J. Optim. Theory Appl., 66 (1990), pp. 477–487.
[57] D. Sun and L. Qi, *On NCP-functions*, *Computational optimization—A tribute to Olvi Mangasarian, Part* II, Comput. Optim. Appl., 13 (1999), pp. 201–220.
[58] L. Thibault, *On generalized differentials and subdifferentials of Lipschitz vector-valued functions*, Nonlinear Anal., 6 (1982), pp. 1037–1053.
[59] F. Tröltzsch, *An SQP method for the optimal control of a nonlinear heat equation*, Control Cybernet., 23 (1994), pp. 267–288.
[60] M. Ulbrich, *Constrained optimal control of Navier–Stokes flow by semismooth Newton methods*, Systems Control Lett., to appear.
[61] M. Ulbrich, *Nonmonotone trust-region methods for bound-constrained semismooth equations with applications to nonlinear mixed complementarity problems*, SIAM J. Optim., 11 (2001), pp. 889–917.
[62] M. Ulbrich, *Nonsmooth Newton-like Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*, Habilitationsschrift, Zentrum Mathematik, Technische Universität München, Munich, Germany, 2001.
[63] M. Ulbrich, *On a nonsmooth Newton method for nonlinear complementarity problems in function space with applications to optimal control*, in Complementarity: Applications, Algorithms and Extensions (Proceedings of the International Conference on Complementarity, Madison, WI, 1999), M. C. Ferris, O. L. Mangasarian, and J.-S. Pang, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2001, pp. 341–360.

[64] M. Ulbrich and S. Ulbrich, *Superlinear convergence of affine-scaling interior-point Newton methods for infinite-dimensional nonlinear problems with pointwise bounds*, SIAM J. Control Optim., 38 (2000), pp. 1938–1984.

[65] M. Ulbrich, S. Ulbrich, and M. Heinkenschloss, *Global convergence of trust-region interior-point algorithms for infinite-dimensional nonconvex minimization subject to pointwise bounds*, SIAM J. Control Optim., 37 (1999), pp. 731–764.

[66] H. Xu, *Point-Based Set-Valued Approximations, C-Differential Operators and Applications*, Report 6/98, School of Information Technology and Mathematical Science, University of Ballarat, Australia, 1998.

[67] H. Xu, *Set-valued approximations and Newton's methods*, Math. Program., 84 (1999), pp. 401–420.

SIAM J. OPTIM.
Vol. 13, No. 3, pp. 842–864

REOPTIMIZATION WITH THE PRIMAL-DUAL INTERIOR POINT METHOD*

JACEK GONDZIO† AND ANDREAS GROTHEY†

Abstract. Reoptimization techniques for an interior point method applied to solving a sequence of linear programming problems are discussed. Conditions are given for problem perturbations that can be absorbed in merely one Newton step. The analysis is performed for both short-step and long-step feasible path-following methods. A practical procedure is then derived for an infeasible path-following method. It is applied in the context of *crash* start for several large-scale structured linear programs. Numerical results with OOPS, a new object-oriented parallel solver, demonstrate the efficiency of the approach. For large structured linear programs, crash start leads to about 40% reduction in the number of iterations and translates into a 25% reduction of the solution time. The crash procedure parallelizes well, and speed-ups between 3.1–3.8 on four processors are achieved.

Key words. interior point methods, warm-start, crash start

AMS subject classifications. 90C51, 90C06

PII. S1052623401393141

1. Introduction.

A number of optimization algorithms require solving a sequence of linear programs. This is a common situation, for example, in the cutting plane methods [13], decomposition [6, 4], branch-and-bound and branch-and-cut approaches for mixed integer optimization [24], and many others. The problems in such a sequence are often similar, i.e., the later instance is only a minor perturbation of the earlier one. Hence the optimal solution of the earlier problem (or, more generally, a close-to-optimality solution of it with some desirable properties) should be a good starting point for the subsequent problem.

Interior point methods are reputed to have difficulties when they are (naively) applied in this context. Indeed, the efficiency of a practical interior point algorithm critically depends on the ability of the algorithm to stay close to the central path. It is no surprise that, after the problem has been perturbed, the optimal solution of the earlier problem (that necessarily must have been very close to the boundary of the feasible region) is a very bad starting point for a new problem with a different optimal partition. The difficulty of reoptimization may be decreased by the choice of a suitable nonoptimal point lying in the neighborhood of the central path [9] and reoptimization from it. Such an approach has an intuitive justification: a suitable nonoptimal point is not yet too close to the boundary of the feasible region, and hence it is able to absorb larger perturbations to the problem.

This approach has already been used in two different classes of reoptimization problems. In the first one, the size of the problem increases: a set of new constraints or a set of new variables is added to the previous linear program. This is, for example, the case in a cutting plane method (or a column generation method) widely used in combinatorial optimization or branch-and-cut approaches to integer programming. The advantages of the application of interior point methods in this context have been

*Received by the editors July 31, 2001; accepted for publication (in revised form) August 1, 2002; published electronically February 12, 2003. Supported by the Engineering and Physical Sciences Research Council of UK, EPSRC grant GR/M68169.

http://www.siam.org/journals/siopt/13-3/39314.html

†School of Mathematics, University of Edinburgh, King's Buildings, Edinburgh EH9 3JZ, Scotland (gondzio@maths.ed.ac.uk, agr@maths.ed.ac.uk).

recognized by Mitchell and Todd [17, 19], who were the first to try practical reoptimization procedures in the primal projective algorithm. An implementation of the warm-starting technique for the infeasible primal-dual method applied to solving restricted master problems in the cutting plane scheme was described in [9]. The reader interested in the use of interior point methods to solve combinatorial optimization problems should consult [18] and the references therein.

In the second important class of problems that need reoptimization, the size of the problem does not change, but the parameters such as the coefficient matrix, the right-hand side, the objective function, and/or the variable bounds do change. This is the case, for example, when subproblems in Dantzig–Wolfe decomposition [6] are solved (the objective of the linear program changes), when subproblems in Benders decomposition [4] are solved (the right-hand-side vector of the linear program changes), or when a variable has its bound tightened in the branch-and-bound technique (this again results in the perturbation of the right-hand-side vector). Preliminary results from applying an interior point-based reoptimization procedure for solving subproblems to the decomposition of large-scale structured linear programs have been reported in [11], where a straightforward extension of the method of [9] was used.

A reoptimization strategy similar to the one proposed here has been analyzed recently by Yıldırım and Wright [27]. In their approach, all intermediate iterates (approximate μ-centers) are stored. Once the perturbation to the problem is known, one of the earlier stored μ-centers is chosen. This point is supposed to be sufficiently far away from the optimal solution to absorb the perturbation of the problem and to restore feasibility in only one Newton step. Yıldırım and Wright derive bounds on the size of absorbable perturbations. Their bounds depend on the problem size and the problem condition number. Two different condition numbers are used, the one of Nunez and Freund [20] and the one that follows from Dikin [7]. Unfortunately, none of these numbers can be easily computed. Therefore it is not obvious how to use these results in practice.

The approach proposed in this paper is different. We introduce a new relative measure of perturbations. The perturbations in the primal and dual spaces are compared with the primal and dual slack variables, respectively, corresponding to a given starting point. We derive bounds on the largest possible perturbations that can be absorbed by a given approximate μ-center without significantly affecting its proximity measures hence allowing easy continuation of the path-following algorithm. We discuss the two cases of short-step and long-step methods, including a recent result on $\mathcal{O}(\sqrt{n}\log n\log(1/\epsilon))$ complexity of the latter due to Peng, Roos, and Terlaky [21].

We are aware of the gap between theory and practice. The former provides complexity estimates and necessarily relies on worst-case analysis. The approach proposed in this paper can be used in practice. Our relative measure of perturbations needs very little effort to be computed; it can thus be determined for a list of candidate starting points and used to choose the most suitable one. We also discuss how the theoretical developments that we have made for this feasible path-following method can be applied in the infeasible method known to be the most efficient interior point method in practice [14, 3].

One of the difficulties in the implementation of interior point methods is the choice of the starting point; cf. [3] and the references therein. Most implementations of interior point methods use some variation of Mehrotra's starting point [16]. Although for the self-dual embedding, which is believed to be less sensitive, any starting point is acceptable, there is still an issue of finding a good initial point [2]. In this paper

we will consider only the standard primal-dual interior point method.

Unlike the simplex method, which can take advantage of an advanced starting basis [5, 12, 15], the interior point method is known to be incapable of doing so. It is common to consider warm-starting interior point methods from an approximate μ-center. However, reoptimization techniques are also of interest if an advanced starting point is known not from a previous solve of a similar problem but, for example, by a crash procedure. The issues involved are the same, but the situation is more challenging since the advanced starting points are generally not μ-centers (or even close to one).

In the second part of this paper we study this problem in the context of interior point methods applied to solving very large structured linear programs, and we provide evidence that advanced *crash* starting points can be constructed for them. Our crashing procedure relies on decomposition but in general constructs infeasible and not necessarily well-centered starting points. We use the reoptimization procedure presented in this paper to start the primal-dual algorithm from such points. Numerical results obtained with OOPS, the object-oriented parallel solver [10], confirm that our crash routine can save up to 30-40% of iterations compared with use of the standard starting point.

The paper is organized as follows. In section 2 we briefly state the problem, recall some known facts about the worst-case complexity of the path-following methods, and introduce the notation used throughout the paper. In section 3 we derive bounds on the largest perturbations (primal and dual infeasibilities) that can be absorbed by a well-centered point in one Newton step. We show that the proximity measure of the updated point is worsened only slightly so that the path-following method may continue from this point without affecting the worst-case complexity result. In section 4 we translate our findings into computational practice. In particular, we discuss how to deal with large perturbations of the problem by gradually taking them into account in subsequent iterations. We also relax the constraint of maintaining close proximity to the central path; instead, we rely on the use of the multiple centrality corrections technique [8]. In section 5 we formulate desired properties of good candidates for the starting point in an interior point method and discuss how such points can be obtained through the use of decomposition techniques for large structured linear programs. In section 6 we illustrate our findings with computational results for a number of structured linear programs. One class of problems originates from network optimization [10] and the other is a well-studied multistage stochastic programming formulation of the asset liability management problem [28]. In section 7 we give our conclusions.

2. Preliminaries. We consider a primal-dual path-following method for linear programming. The theory for this class of methods has been discussed in detail in the excellent book of Wright [25]; in our developments we shall refer to several results that can be found in this book. In this section we shall deal with the *feasible* method.

Consider a pair of linear programs, the primal

$$\begin{aligned} &\text{minimize} && c_0^T x \\ &\text{subject to} && A_0 x = b_0, \\ & && x \geq 0, \end{aligned} \tag{2.1}$$

where $c_0, x \in \mathcal{R}^n, b_0 \in \mathcal{R}^m$, and $A_0 \in \mathcal{R}^{m \times n}$ has full row rank, and its dual

$$\text{maximize} \qquad b_0^T y$$

$$\begin{aligned} &\text{subject to } A_0^T y + s = c_0, \\ &\qquad\qquad s \geq 0, \end{aligned} \tag{2.2}$$

where $y \in \mathcal{R}^m$ and $s \in \mathcal{R}^n$. We assume that the feasible sets of the primal and dual problems (2.1) and (2.2) have nonempty interiors

$$\mathcal{F}^0 = \{(x, y, s) | A_0 x = b_0,\ A_0^T y + s = c_0,\ (x, s) > 0\} \neq \emptyset.$$

Hence for any $\mu > 0$ there exists a uniquely defined point $(x(\mu), y(\mu), s(\mu))$, $x(\mu) > 0, s(\mu) > 0$, that satisfies the following first-order optimality conditions for the associated barrier problem:

$$\begin{aligned} A_0 x &= b, \\ A_0^T y + s &= c, \\ XSe &= \mu e, \end{aligned} \tag{2.3}$$

where X and S are diagonal matrices with the elements x_j and s_j, respectively, $e \in \mathcal{R}^n$ is the n-vector of all ones, and $\mu > 0$ is the barrier parameter. Such a point is called a μ-center.

We assume that a feasible path-following algorithm is used so that all of its iterates are primal and dual feasible. However, they are not necessarily perfectly centered. We shall consider two neighborhoods of the central path appropriate for the short-step and the long-step algorithms, respectively. The short-step algorithm keeps all its iterates in

$$N_2(\theta) = \{(x, y, s) \in \mathcal{F}^0 \,|\, \|XSe - \mu e\|_2 \leq \theta\mu\}, \tag{2.4}$$

where $0 \leq \theta < 1$. For the long-step algorithm we shall use the following neighborhood:

$$N_\infty(\gamma_l, \gamma_u) = \{(x, y, s) \in \mathcal{F}^0 \,|\, \gamma_l\, \mu \leq x_j s_j \leq \gamma_u\, \mu \ \ \forall j\}, \tag{2.5}$$

where $0 < \gamma_l \leq 1 \leq \gamma_u$. The reader should notice that our definition of the $N_\infty(\gamma_l, \gamma_u)$ neighborhood is a slight modification of the usual $N_{-\infty}(\gamma)$ neighborhood of [25].

2.1. Complexity bounds for the path-following algorithms. Below we remind the reader of the current best complexity results for linear optimization with the path-following algorithm. We recall them in a form that explicitly uses the parameter κ associated with the quality of the initial solution. Assume that we seek an ϵ-optimal solution of a linear program, and that an initial well-centered feasible point is given such that $\mu^0 = (1/\epsilon)^\kappa$. The short-step method finds the ϵ-optimal solution in at most $\mathcal{O}((\kappa + 1)\sqrt{n}\log(1/\epsilon))$ iterations. The classical long-step method finds the ϵ-optimal solution in at most $\mathcal{O}((\kappa+1)n\log(1/\epsilon))$ iterations. Peng, Roos, and Terlaky [21] have recently given a new result for a large-update method. Their method performs only $\mathcal{O}((\kappa + 1)\log(1/\epsilon))$ updates of the barrier parameter. However, it requires many so-called inner iterations ($\mathcal{O}(\sqrt{n})$) to restore centrality after the barrier update, giving an overall complexity bound of $\mathcal{O}((\kappa + 1)\sqrt{n}\log n\, \log(1/\epsilon))$. It is not obvious that this method should be viewed as a long-step algorithm, but we link our results to it as well.

2.2. Reoptimization problem. Assume that an approximate μ-center has been found for the primal-dual pair (2.1)–(2.2) and that the linear optimization problem has changed. Namely, all its data A_0, b_0, and c_0 has been replaced with the new values

A, b, and c (where again A is assumed to have full row rank). Unlike in the method of [9], we assume that the size of the linear problem has not changed. We can thus use the approximate μ-center as an iterate for the new problem. In the new first-order conditions all three equations may possibly be violated. Let us define the residuals

$$\begin{bmatrix} \xi_b \\ \xi_c \\ \xi_\mu \end{bmatrix} = \begin{bmatrix} b - Ax \\ c - A^T y - s \\ \mu e - XSe \end{bmatrix}. \tag{2.6}$$

All entries of ξ_μ are of order $\mathcal{O}(\mu)$ because the point (x, y, s) is an approximate μ-center of (2.1)–(2.2). However, primal and dual infeasibilities ξ_b and ξ_c can be arbitrarily large.

We shall express these infeasibilities in the scaling related to the current primal-dual point. In the following section we shall prove sufficient conditions that such scaled perturbations have to satisfy to make reoptimization possible. For a current approximate μ-center (x, y, s), primal perturbation ξ_b, and dual perturbation ξ_c we define the relative residual vectors

$$\tilde{\xi}_b = X^{-1}A^T(AA^T)^{-1}\xi_b \quad \text{and} \quad \tilde{\xi}_c = S^{-1}\xi_c, \tag{2.7}$$

where X^{-1} and S^{-1} is the usual notation for diagonal $n \times n$ matrices built of elements x_j^{-1} and s_j^{-1}, respectively.

Before we go any further, the reader should be warned that one cannot expect to eliminate terms that depend on the problem dimension from the complexity bounds. Instead, following [27], we will concentrate on terms that depend on the quality of the initial point, namely, on the parameter κ present in every bound. The aim of our warm-starting procedure is to find a new point $(\bar{x}, \bar{y}, \bar{s})$ that is primal and dual feasible for the perturbed problem and corresponds to a new barrier parameter $\bar{\mu}$ with the value close to μ.

3. Absorbing primal and dual infeasibilities. Assume that an approximate μ-center has been found for (2.1)–(2.2) and that it is used to compute the Newton direction for the new linear program in which both primal and dual feasibility is violated. Consider the following Newton equation system:

$$\begin{bmatrix} A & 0 & 0 \\ 0 & A^T & I \\ S & 0 & X \end{bmatrix} \cdot \begin{bmatrix} \Delta x \\ \Delta y \\ \Delta s \end{bmatrix} = \begin{bmatrix} \xi_b \\ \xi_c \\ 0 \end{bmatrix}. \tag{3.1}$$

We draw the reader's attention to the fact that the Newton direction attempts to correct only primal and dual infeasibilities, but the complementarity products $x_j s_j$ are not recentered (although they will obviously change, possibly worsen, once the step is made). A few manipulations give the following Newton direction for dual variables,

$$\Delta y = (AXS^{-1}A^T)^{-1}(AXS^{-1}\xi_c + \xi_b), \tag{3.2}$$

and the following Newton direction for primal variables and dual slacks,

$$\Delta x = (XS^{-1}A^T(AXS^{-1}A^T)^{-1}AXS^{-1} - XS^{-1})\,\xi_c + XS^{-1}A^T(AXS^{-1}A^T)^{-1}\xi_b, \tag{3.3}$$

$$\Delta s = (I - A^T(AXS^{-1}A^T)^{-1}AXS^{-1})\,\xi_c - A^T(AXS^{-1}A^T)^{-1}\xi_b. \tag{3.4}$$

In our analysis we shall rely on bounds of norms of the following matrix:

$$Q = I - S^{-1}A^T(AXS^{-1}A^T)^{-1}AX.$$

LEMMA 3.1. *If* $(x, y, s) \in N_2(\theta)$, *then* $\|Q\|_2 \leq (\frac{1+\theta}{1-\theta})^{1/2}$.

Proof. Since $(x, y, s) \in N_2(\theta)$, we have $(1-\theta)\mu \leq x_j s_j \leq (1+\theta)\mu$ for all $j = 1, 2, \ldots, n$. Further, we have $\frac{1}{(x_j s_j)^{1/2}} \leq \frac{1}{(1-\theta)^{1/2}\mu^{1/2}}$ and $(x_j s_j)^{1/2} \leq (1+\theta)^{1/2}\mu^{1/2}$. Since

$$Q = X^{-1/2}S^{-1/2}(I - X^{1/2}S^{-1/2}A^T(AXS^{-1}A^T)^{-1}AX^{1/2}S^{-1/2})X^{1/2}S^{1/2}$$

and the matrix in outer parenthesis is an orthogonal projection on the null space of $AX^{1/2}S^{-1/2}$ (and hence its 2-norm is equal to 1), we can write

$$\|Q\|_2 \leq \|X^{-1/2}S^{-1/2}\|_2 \cdot 1 \cdot \|X^{1/2}S^{1/2}\|_2 \leq \left(\frac{1+\theta}{1-\theta}\right)^{1/2},$$

which completes the proof. ∎

LEMMA 3.2. *If* $(x, y, s) \in N_\infty(\gamma_l, \gamma_u)$, *then* $\|Q\|_2 \leq (\frac{\gamma_u}{\gamma_l})^{1/2}$.

The proof is omitted because it is very similar to the proof of Lemma 3.1.

Let us observe that, for both neighborhoods, the bound on the norm of Q depends only on the constants that define the proximity of the point to the central path. Below we shall also use the ∞-norm of Q. We shall rely on the simple relation $\|Q\|_\infty \leq \sqrt{n}\,\|Q\|_2$ that holds for any square $n \times n$ matrix.

Our reoptimization procedure is divided into two steps. In the first step the total infeasibility is absorbed by making a full step in the Newton direction. In the next step the good quality of the proximity to the central path has to be restored. The second step is needed by the short-step algorithm and by Peng, Roos, and Terlaky's [21] large-update algorithm, but it may be omitted in the large-step path-following algorithm.

We analyze independently the cases of perturbations in the primal and dual spaces and use techniques that are similar to those applied by Yıldırım and Todd [26] to analyze the sensitivity of interior point solutions subject to perturbations in vectors b and c (cf. Propositions 1 and 2 in [26]). One of the key features of the approach presented in this paper is that the primal perturbation ξ_b is used only in the primal direction Δx, while the dual perturbation ξ_c is used only in the dual direction $(\Delta y, \Delta s)$. More precisely, our feasibility restoration directions are obtained by substituting $\xi_c = 0$ into (3.3) and $\xi_b = 0$ into (3.2) and (3.4).

Decoupling the primal and dual directions gives the algorithm added flexibility to recover from infeasibilities, particularly in situations where primal and dual infeasibilities differ substantially, as is the case when subproblems in Benders or Dantzig–Wolfe decomposition are solved [11] or in our application to crash start (section 5). Note that decoupling the steps results in only a slight increase in computational cost, since both steps can be obtained using the same Cholesky factors.

3.1. Restoring dual feasibility. After setting $\xi_b = 0$ (i.e., ignoring primal infeasibility), from (3.4) and (2.7) we obtain

$$S^{-1}\Delta s = (I - S^{-1}A^T(AXS^{-1}A^T)^{-1}AX)\,S^{-1}\xi_c = Q\,\tilde{\xi}_c. \tag{3.5}$$

We shall now analyze two cases: the short-step and the long-step path-following algorithms.

LEMMA 3.3. *Let $(x, y, s) \in N_2(\theta)$ and $\beta < \sqrt{n}$. If $\|\tilde{\xi}_c\|_2 \leq \frac{\beta}{\sqrt{n}} \cdot (\frac{1-\theta}{1+\theta})^{1/2}$, then the full Newton step in the dual space is feasible and it absorbs the total infeasibility ξ_c. For $\theta = 0.25, \beta = 0.1$, and $\sqrt{n} \geq 100$ the new point $(\bar{x}, \bar{y}, \bar{s}) = (x, y + \Delta y, s + \Delta s) \in N_2(0.5)$.*

Proof. From Lemma 3.1, (3.5), and the assumption of this lemma, we find that $\|S^{-1}\Delta s\|_2 \leq \beta/\sqrt{n} < 1$. Therefore $|\Delta s_j/s_j| \leq \beta/\sqrt{n} < 1$, and hence the full Newton step in the dual space is feasible. After this step, the dual feasibility is restored.

It remains to prove the result about the proximity of the new point to the new $\bar{\mu}$-center. We prove first that the new barrier parameter after a full step in the Newton direction in the dual space does not change significantly compared with the previous one. This new barrier parameter is defined as follows:

$$n\bar{\mu} = \sum_j \left(x_j s_j + x_j s_j \frac{\Delta s_j}{s_j} \right).$$

The inequality $\|S^{-1}\Delta s\|_2 \leq \beta/\sqrt{n}$ implies that $-\beta/\sqrt{n} \leq \frac{\Delta s_j}{s_j} \leq \beta/\sqrt{n}$ and

$$\frac{-\beta}{\sqrt{n}} x_j s_j \leq x_j s_j \frac{\Delta s_j}{s_j} \leq \frac{\beta}{\sqrt{n}} x_j s_j;$$

hence

$$\frac{-\beta}{\sqrt{n}} n\mu \leq \sum_j x_j s_j \frac{\Delta s_j}{s_j} \leq \frac{\beta}{\sqrt{n}} n\mu,$$

and so the new barrier parameter $\bar{\mu}$ satisfies

$$n\mu(1 - \beta/\sqrt{n}) \leq n\bar{\mu} \leq n\mu(1 + \beta/\sqrt{n}).$$

We need to evaluate the proximity of the new primal-dual pair to the new $\bar{\mu}$-center. First observe that

$$\|\bar{X}\bar{S}e - \bar{\mu}e\|_2 \leq \|\bar{X}\bar{S}e - XSe\|_2 + \|XSe - \mu e\|_2 + \|\mu e - \bar{\mu}e\|_2.$$

Each of these three terms can be bounded from above:

$$\|\bar{X}\bar{S}e - XSe\|_2 = \left(\sum_j (x_j s_j \frac{\Delta s_j}{s_j})^2 \right)^{1/2} \leq \left(n\,(1+\theta)^2\,\mu^2\,\frac{\beta^2}{n} \right)^{1/2} = (1+\theta)\,\beta\,\mu,$$

$$\|XSe - \mu e\|_2 \leq \theta\mu,$$

$$\|\mu e - \bar{\mu}e\|_2 \leq \left(n\,\frac{\beta^2}{n}\,\mu^2 \right)^{1/2} = \beta\mu.$$

Therefore

$$\|\bar{X}\bar{S}e - \bar{\mu}e\|_2 \leq ((1+\theta)\beta + \theta + \beta)\mu.$$

For $\theta = 0.25$ and $\beta = 0.1$ we have $(1+\theta)\beta + \theta + \beta = 0.475$. Also since $\mu(1 - \beta/\sqrt{n}) \leq \bar{\mu} \leq \mu(1 + \beta/\sqrt{n})$ for $\sqrt{n} \geq 100$, we have $0.99\mu \leq \bar{\mu} \leq 1.01\mu$ and

$$\|\bar{X}\bar{S}e - \bar{\mu}e\|_2 \leq 0.475\mu \leq \frac{0.475}{0.99}\bar{\mu} \leq 0.5\bar{\mu},$$

which completes the proof. ∎

The choice of constants in Lemma 3.3 has ensured that the new point $(\bar{x}, \bar{y}, \bar{s})$ belongs to a slightly larger neighborhood of the central path, $N_2(0.5)$. It suffices to make only one pure centering step to get back to the original smaller neighborhood $N_2(0.25)$. This completes the analysis for the short-step method.

The analysis in the case of the long-step algorithm differs slightly because we assume that the current μ-center belongs to a large neighborhood $N_\infty(0.5, 2.0)$, and that, after absorbing infeasibility, the new point belongs to a larger neighborhood $N_\infty(0.25, 2.5)$. We do not need any recentering step in this case (we just accept a slightly larger neighborhood; cf. [25]).

LEMMA 3.4. *Let $(x, y, s) \in N_\infty(\gamma_l, \gamma_u)$ and $\beta < 1$. If $\|\tilde{\xi}_c\|_\infty \le \beta/\|Q\|_\infty$, then the full Newton step is feasible and it absorbs the total infeasibility ξ_c. For $\gamma_l = 0.5, \gamma_u = 2.0$, and $\beta = 0.1$ the new point $(\bar{x}, \bar{y}, \bar{s}) = (x, y + \Delta y, s + \Delta s) \in N_\infty(0.25, 2.5)$.*

Proof. From (3.5) and the assumption of this lemma, we find that $\|S^{-1}\Delta s\|_\infty \le \beta < 1$ and from $S\Delta x + X\Delta s = 0$ also that $\|X^{-1}\Delta x\|_\infty = \|S^{-1}\Delta s\|_\infty < 1$. Hence the full Newton step is feasible. After this step, feasibility is restored.

Similarly to the proof of Lemma 3.3, we bound the new barrier parameter $\bar{\mu}$ and analyze the proximity of the new iterate to the central path. From the inequality $\|S^{-1}\Delta s\|_\infty \le \beta$ we get

$$-\beta\, n\mu \le \sum_j x_j s_j \frac{\Delta s_j}{s_j} \le \beta\, n\mu$$

and thus the new barrier parameter $\bar{\mu}$ satisfies

$$n\mu(1-\beta) \le n\bar{\mu} \le n\mu(1+\beta).$$

We still need to prove that componentwise the centrality has not been worsened too much. To prove the claim, we need to show that

$$0.25\bar{\mu} \le \bar{x}_j \bar{s}_j \le 2.5\bar{\mu}.$$

Indeed, we find that

$$\bar{x}_j\bar{s}_j = x_j s_j + x_j s_j \frac{\Delta s_j}{s_j} \ge (\gamma_l - \gamma_u\beta)\mu \ge \frac{\gamma_l - \gamma_u\beta}{1+\beta}\,\bar{\mu},$$
$$\bar{x}_j\bar{s}_j = x_j s_j + x_j s_j \frac{\Delta s_j}{s_j} \le (\gamma_u + \gamma_u\beta)\mu \le \frac{\gamma_u + \gamma_u\beta}{1-\beta}\,\bar{\mu}.$$

For $\gamma_l = 0.5$, $\gamma_u = 2.0$, and $\beta = 0.1$ we obtain the required result. ∎

Note that the first part of Lemma 3.4 also follows from Proposition 2 in [26].

The case of the large-update method [21] requires one more comment. This algorithm uses a different proximity measure

$$\Psi(v) = \sum_j \psi(v_j), \quad \text{where} \quad \psi(v_j) = \frac{v_j^2 - 1}{2} + \frac{v_j^{1-q} - 1}{q-1},$$

in which $v_j = \sqrt{\frac{x_j s_j}{\mu}}$, $j = 1, 2, \ldots, n$, and $q \ge 1$ is an additional parameter. From Lemma 5.3 in [21] we have

$$\Psi(v) \le \frac{1}{2}\,\|v^{-q} - v\|_2^2.$$

We leave it to the reader to prove the following bridge result.

LEMMA 3.5. *Let two constants γ_l and γ_u such that $0 < \gamma_l \le 1 \le \gamma_u$ be given. If $(x, y, s) \in N_\infty(\gamma_l, \gamma_u)$, then $\Psi(v) \le C\, n$, where C is a constant independent of n.*

As a consequence of Lemma 3.5, we can apply Lemma 2.7 from [21] to conclude that after at most $\mathcal{O}(n^{\frac{q+1}{2q}})$ recentering iterations the proximity measure of the new point, $\Psi(v')$, will be reduced from $\mathcal{O}(n)$ to $\mathcal{O}(1)$, thus allowing the continuation of the large-update method.

Summing up, for three different variants of the path-following algorithm we have given conditions on the perturbation in the dual space that can be easily absorbed in one Newton step without affecting the key properties of a given algorithm.

3.2. Restoring primal feasibility. The case of restoring primal feasibility can be dealt with in a very similar way to that of section 3.1.

After setting $\xi_c = 0$ (i.e., ignoring dual infeasibility), from (3.3) and (2.7) we obtain

$$\begin{aligned} X^{-1}\Delta x &= S^{-1}A^T(AXS^{-1}A^T)^{-1}\xi_b \\ &= S^{-1}A^T(AXS^{-1}A^T)^{-1}AXX^{-1}A^T(AA^T)^{-1}\xi_b \\ &= (I-Q)\,\tilde{\xi}_b. \end{aligned} \tag{3.6}$$

It is worth noting that we use a relative primal residual $\tilde{\xi}_b = X^{-1}A^T(AA^T)^{-1}\xi_b \in \mathcal{R}^n$ instead of the real perturbation vector $\xi_b \in \mathcal{R}^m$. This has a common-sense justification that any primal infeasibility has to be absorbed through changes of the primal variables.

Let us observe that computing the direction Δx could also involve, through (3.2) and (3.4), the computation of directions in the dual space $(\Delta y, \Delta s)$, but the latter are skipped without being used. Although we do apply the primal-dual framework to compute feasibility restoration directions, we use the primal perturbation only in the primal feasibility restoration direction (3.6), and the dual perturbation only in the dual feasibility restoration direction (3.5).

We omit the detailed analysis of the case of primal perturbations. We formulate two lemmas analogous to Lemmas 3.3 and 3.4 but skip their proofs because they are almost identical to those of Lemmas 3.3 and 3.4. For these omitted proofs the reader should observe that the bound on the relative step in the primal space (3.6) involves the matrix $I - Q$, and by using Lemma 3.1 we have

$$\|I-Q\|_2 \le \|Q\|_2 + 1 \le \left(\frac{1+\theta}{1-\theta}\right)^{1/2} + 1.$$

LEMMA 3.6. *Let $(x, y, s) \in N_2(\theta)$ and $\beta < \sqrt{n}$. If $\|\tilde{\xi}_b\|_2 \le \frac{\beta}{\sqrt{n}}/((\frac{1+\theta}{1-\theta})^{1/2}+1)$, then the full Newton step in the primal space is feasible and it absorbs the total infeasibility ξ_b. For $\theta = 0.25, \beta = 0.1$, and $\sqrt{n} \ge 100$ the new point $(\bar{x}, \bar{y}, \bar{s}) = (x + \Delta x, y, s) \in N_2(0.5)$.*

LEMMA 3.7. *Let $(x, y, s) \in N_\infty(\gamma_l, \gamma_u)$ and $\beta < 1$. If $\|\tilde{\xi}_b\|_\infty \le \beta/(\|Q\|_\infty + 1)$, then the full Newton step is feasible and it absorbs the total infeasibility ξ_b. For $\gamma_l = 0.5$, $\gamma_u = 2.0$, and $\beta = 0.1$ the new point $(\bar{x}, \bar{y}, \bar{s}) = (x + \Delta x, y, s) \in N_\infty(0.25, 2.5)$.*

The case of large update method [21] is covered by the use of Lemmas 3.7 and 3.5.

For ease of the presentation, we have split our analysis into two independent cases for the primal and the dual spaces. It is possible to perform the analysis in the presence of perturbations in both spaces at the same time. Then, however, to evaluate the new

complementarity products $\bar{x}_j\bar{s}_j = (x_j + \Delta x_j)(s_j + \Delta s_j)$, we would have to consider the products $\Delta x_j \Delta s_j$. Having bounds on $\|X^{-1}\Delta x\|_\infty$ and $\|S^{-1}\Delta s\|_\infty$ allows us to bound these products with terms proportional to $x_j s_j$. Hence the componentwise changes to the complementarity products can also be bounded by terms proportional to $x_j s_j$, which themselves are bounded by terms proportional to μ. Therefore very similar analysis holds when both infeasibilities have to be absorbed at the same time, but the proofs become longer.

We have left the discussion of the simultaneous treatment of both primal and dual perturbations to the following section, in which also some other issues of implementation of our reoptimization technique are addressed.

4. From theory to practice. There still exists an important gap between the theory and the computational practice of interior point methods. The theory for feasible path-following algorithms is more elegant and provides better complexity bounds than that for the infeasible algorithms (cf. [25], Chapters 5 and 6). On the other hand, the implementations of interior point methods use infeasible algorithms. In these approaches the primal and dual feasibility is expected to be attained together with optimality, although, in practice, the infeasibilities are often reduced much earlier than the duality gap gets decreased below the optimality tolerance.

The infeasible primal-dual algorithm follows the central path; i.e., in subsequent iterations it makes (one) damped step in the Newton direction towards the solution of the first-order optimality conditions for the barrier problem and reduces the barrier parameter. The iterates of this algorithm stay in a large neighborhood of the central path (cf. [25], page 109). In our implementation of the infeasible primal-dual method (applied to (2.1)–(2.2)), this neighborhood is defined by the following inequalities:

$$\begin{aligned} \|A_0 x - b_0\| &\le \epsilon_p(\mu)(||r_b^0|| + 1), \\ \|A_0^T y + s - c_0\| &\le \epsilon_d(\mu)(||r_c^0|| + 1), \\ \gamma_l \mu \le x_j s_j &\le \gamma_u \mu, \qquad j = 1, 2, \ldots, n. \end{aligned} \tag{4.1}$$

The two vectors $r_b^0 = A_0 x^0 - b_0$ and $r_c^0 = A_0^T y^0 + s^0 - c_0$ are the violations of the primal and dual constraints, respectively, at the initial point (x^0, y^0, s^0). The relative feasibility tolerances $\epsilon_p(\mu)$ and $\epsilon_d(\mu)$ decrease with μ and reach zero at $\mu = 0$. The parameters γ_l and γ_u control the discrepancy between the largest and the smallest complementarity products.

If we expect that reoptimizations will be done, then we impose stronger requirements on the reduction of the primal and dual infeasibilities by a fast reduction of the feasibility tolerances. This occasionally requires an additional recentering step in which we preserve the duality gap but reduce the infeasibilities. Such a step uses multiple centrality correctors [8] and is expected to reduce the discrepancy between the largest and the smallest complementarity products.

Following [9], if we solve a given linear optimization problem (2.1)–(2.2) and expect that reoptimizations will be done later, then we modify the infeasible primal-dual algorithm and ask for a nearly optimal μ-center to be saved for future reoptimization. That is, we find a point (x, y, s) that satisfies

$$\begin{aligned} \|A_0 x - b_0\| &\approx 0, \\ \|A_0^T y + s - c_0\| &\approx 0, \\ \gamma_l \mu \le x_j s_j &\le \gamma_u \mu, \qquad j = 1, 2, \ldots, n, \end{aligned} \tag{4.2}$$

for some small barrier parameter μ, which guarantees near-optimality.

Let us observe that an exact μ-center satisfies both primal and dual feasibility constraints and the parameter μ controls its distance to optimality since the duality gap at this point is

$$c_0^T x - b_0^T y = c_0^T x - x^T A_0^T y = x^T (c_0 - A_0^T y) = x^T s = n\mu. \tag{4.3}$$

In the usual applications, the path-following algorithm terminates when the duality gap drops below a predetermined relative optimality tolerance ϵ, i.e., when

$$|c_0^T x - b_0^T y| \leq \epsilon(|c_0^T x| + 1), \tag{4.4}$$

with ϵ usually taking the values 10^{-6} or 10^{-8}. However, instead of using (4.4) as the stopping criterion, we can use (4.2). Here the algorithm stops at an approximate μ-center corresponding to the predetermined barrier parameter μ (thereby controlling the required distance to optimality). Once we obtain a rough estimate of the optimal objective value $\tilde{z} = c_0^T x$, e.g., when (4.4) is already satisfied for $\epsilon_0 = 10^{-1}$, we limit the decrease of the barrier to

$$\mu = \hat{\epsilon}\frac{|\tilde{z}|}{n}. \tag{4.5}$$

From (4.2), (4.3), and (4.5) we see that an approximate μ-center corresponding to such a μ is a nearly optimal solution with the relative precision $\hat{\epsilon}$.

The choice of the tolerance $\hat{\epsilon}$ (and, in consequence, the parameter μ) depends on how significant the expected changes to the problem might be. If we expect violent modifications of the problem, then a larger value of $\hat{\epsilon}$, say 10^{-1} or 10^{-2}, is suggested. For expected small perturbations to the problem, we suggest a closer-to-optimality point with $\hat{\epsilon}$ equal to 10^{-3} or 10^{-4}. It is also possible to store several candidate points and to delay the decision on which of them should be used in reoptimization to the time when primal and dual perturbations ξ_b and ξ_c have become known.

From now on we assume that an approximate μ-center (x, y, s) that satisfies (4.2) is stored and that the data in the linear program changes from A_0, b_0, and c_0 to A, b, and c. Naturally, we cannot expect that the feasible point for an earlier problem will be feasible for the new one. We accept the possible violation of both primal and dual feasibility constraints in the new problem,

$$\xi_b = b - Ax \neq 0 \quad \text{and} \quad \xi_c = c - A^T y - s \neq 0,$$

and compute the relative perturbation vectors $\tilde{\xi}_b$ and $\tilde{\xi}_c$ from (2.7). Following the theoretical developments of section 3, since we work with the long-step (infeasible) path-following algorithm, we could use the ∞-norms of the relative perturbation vectors and verify whether they satisfy the assumptions of Lemmas 3.4 and 3.7. If we knew $\|Q\|_\infty$, we could check whether $\|\tilde{\xi}_c\|_\infty \leq \beta/\|Q\|_\infty$ and $\|\tilde{\xi}_b\|_\infty \leq \beta/(\|Q\|_\infty + 1)$. Since we do not know $\|Q\|_\infty$, we could replace it with an upper bound $\sqrt{n}\,(\frac{\gamma_u}{\gamma_l})^{1/2}$ or with what we expect to be its reasonable estimate $|Q|$. Regardless of whether the real norm $\|Q\|_\infty$ or only its estimate $|Q|$ are used, this would still allow us to predict how successful the feasibility restoration direction could be. In particular, if the conditions of Lemmas 3.4 and 3.7 were not satisfied, we could use for warm-starting another μ-center which is further from optimality. Such a point could possibly absorb larger perturbations in the primal and dual spaces. If we had a whole history of iterates

stored as suggested in [27], we could backtrack to an approximate μ-center corresponding to a barrier parameter that is sufficiently large to absorb the perturbations ξ_b and ξ_c.

In the practical algorithm, we compute $S^{-1}\Delta s$ from (3.5) and $X^{-1}\Delta x$ from (3.6) and then perform the ratio tests for the stepsizes in the primal and dual spaces

$$\hat{\alpha}_P := \max\{\alpha > 0 : \ x + \alpha \Delta x \geq 0\}, \tag{4.6}$$

$$\hat{\alpha}_D := \max\{\alpha > 0 : \ s + \alpha \Delta s \geq 0\}. \tag{4.7}$$

Obviously, $\hat{\alpha}_P \geq 1/\|X^{-1}\Delta x\|_\infty > 0$ and $\hat{\alpha}_D \geq 1/\|S^{-1}\Delta s\|_\infty > 0$. If the stepsizes $\hat{\alpha}_P$ and $\hat{\alpha}_D$ are small, say they fall below a prescribed tolerance $\hat{\alpha}_P \leq \alpha_{min}$ or $\hat{\alpha}_D \leq \alpha_{min}$, then we spread the absorption of primal and dual perturbations across a few subsequent iterations. To achieve this, we scale infeasibilities and use

$$\xi_b^{'} = \delta_P \, \xi_b \quad \text{and} \quad \xi_c^{'} = \delta_D \, \xi_c, \tag{4.8}$$

where $\delta_P, \delta_D \in (0,1)$ specify the fraction of infeasibilities that we expect could be absorbed in a single Newton step. We could obviously set $\delta_P = \hat{\alpha}_P$ and $\delta_D = \hat{\alpha}_D$; however, we have found that this is sometimes too pessimistic. Therefore we choose $\delta > 1$ and define

$$\delta_P = \delta \, \hat{\alpha}_P \quad \text{and} \quad \delta_D = \delta \, \hat{\alpha}_D. \tag{4.9}$$

Although the analysis deals independently with the infeasibilities in two spaces, our practical reoptimization procedure takes them into account at the same time. We use the primal-dual framework (i.e., Newton equation system (3.1)) and ignore primal infeasibility in the dual step and dual infeasibility in the primal step. We thus have to solve two systems of equations like (3.1) with $\xi_c = 0$ and with $\xi_b = 0$, respectively. (Both these systems use the same factorization, of course.) For the first few iterations of the reoptimization algorithm the step in the primal space results from ξ_b, the step in the dual space results from ξ_c, and only the recentering steps use both directions at the same time.

It is important to mention that we combine the step in which feasibility perturbations are absorbed with the use of multiple centrality correctors [8]. Hence after the step has been made, the proximity of the new point to the central path is not necessarily worsened compared with that of the previous iterate. This is an important feature of our approach because we expect that not all the perturbations can be absorbed in this single interior point iteration; instead, few more iterations may be needed to restore feasibility in the perturbed problem. Therefore we need the intermediate points be as well centered as possible to be good candidates for absorbing the remaining feasibility perturbations.

Summing up, if large perturbations have to be dealt with, our practical reoptimization procedure absorbs them gradually, making slow progress towards optimality in a new problem at the same time. One could interpret this as passing through a family of problems with data changing gradually from A_0, b_0, c_0 to A, b, c.

Let us summarize our findings in the reoptimization algorithm below. We assume that (x, y, s) is an approximate μ-center for (2.1) and (2.2), and that in the new linear program this point produces infeasibilities ξ_b and ξ_c given by (2.6). The procedure uses the following parameters: $\gamma_l = 0.5$ and $\gamma_u = 2.0$ define the neighborhood of the central path (2.5), $\delta = 2.0$ determines how much of the perturbations we expect to absorb in one primal-dual iteration (cf. (4.9)), and $\alpha_{min} = 0.1$ is a threshold for

acceptable stepsizes in the primal and dual spaces. We draw the reader's attention to the fact that our procedure does not choose how much backtracking is needed because we use only one μ-center saved for future warm-starting. The procedure could obviously be enhanced by a simple test based on Lemmas 3.4 and 3.7 to choose a suitable starting point from the list of candidates if such a list were available.

REOPTIMIZATION WITH THE PRIMAL-DUAL METHOD

Input

(x, y, s): approximate μ-center (4.2);

Parameters

γ_l, γ_u: relative threshold values for outlier complementarity products;
δ: parameter in (4.8) and (4.9);
α_{min}: the minimum acceptable stepsize;

Initialize

Δx: primal feasibility restoring direction (3.6);
$\Delta y, \Delta s$: dual feasibility restoring direction (3.2) (with $\xi_b = 0$) and (3.5);
$\hat{\alpha}_P, \hat{\alpha}_D$: stepsizes (4.6) and (4.7) in the primal and dual spaces;

Absorb infeasibility

while $(\hat{\alpha}_P \leq \alpha_{min}$ or $\hat{\alpha}_D \leq \alpha_{min})$, **do**
  if $(\hat{\alpha}_P \leq \alpha_{min})$, **then**
    scale primal direction:
    $\Delta x := \delta \hat{\alpha}_P \Delta x$;
  endif
  if $(\hat{\alpha}_D \leq \alpha_{min})$, **then**
    scale dual direction:
    $\Delta y := \delta \hat{\alpha}_D \Delta y$,
    $\Delta s := \delta \hat{\alpha}_D \Delta s$;
  endif
  define the predictor direction $\Delta_p = (\Delta x, \Delta y, \Delta s)$;
  $\Delta =$ `Recenter`(Δ_p);
  `MakeStep` (Δ);
  at the new point, recompute:
    $(\Delta x, \Delta y, \Delta s)$ from (3.6), (3.2) (with $\xi_b = 0$), and (3.5);
    $\hat{\alpha}_P, \hat{\alpha}_D$ from (4.6) and (4.7);
end-while

In our implementation $\gamma_l = 0.5$, $\gamma_u = 2.0$, $\delta = 2.0$, and $\alpha_{min} = 0.1$.

Two procedures in this algorithm need further comments. In the `Recenter`(Δ_p) procedure, the direction $\Delta_p = (\Delta x, \Delta y, \Delta s)$ is used as a predictor direction for the multiple centrality correctors technique [8]. Centrality correctors usually alter Δ_p and replace it with a new direction Δ. Centrality correctors aim at two goals: firstly to increase the stepsizes from $1/\delta = 0.5$ to larger values $\alpha_P, \alpha_D \in (0.5, 1)$ and secondly to improve the centrality of the new iterate, i.e., to decrease the spread between the largest and the smallest complementarity products in it. In the procedure `MakeStep`, the maximum feasible stepsizes in the primal and dual spaces are determined along direction Δ, the variables are updated from (x, y, s) to $(\bar{x}, \bar{y}, \bar{s})$, and infeasibilities ξ_b and ξ_c are recomputed.

The reoptimization procedure terminates when a significant portion of the initial perturbations in the primal and dual spaces have already been absorbed, and for the

remaining infeasibilities ξ_b and ξ_c the stepsizes in both spaces exceed the threshold α_{min}. We have found that at this point there is no more need for special feasibility restoration steps. The usual infeasible primal-dual method can be used to terminate the optimization.

5. Crash start of an interior point method. In this paper, we report results on the application of the reoptimization technique to the problem of finding a good advanced starting point. In the next section we will present a decomposition-based technique which aims to find an approximate μ-center quickly without using an interior point method for the whole system. This point is then used as the starting point for our reoptimization method. Since this advanced solution is not the result of solving a similar linear program, we cannot assume that we have a well-centered point to begin with. In other words, in addition to possible large perturbations of the primal and dual feasibility, ξ_μ in (2.6) may also be large.

Guided by the theoretical results of section 3, we realize that if some of the primal or dual slack variables are very close to zero, then the relative primal and dual perturbations (2.7) may become huge and this would inevitably lead to very small stepsizes $\hat{\alpha}_P$ and $\hat{\alpha}_D$ in the reoptimization procedure (cf. (4.6) and (4.7)). Therefore, when constructing a candidate for a starting point, we shall bound these variables away from zero. Additionally, since the theory indicates through Lemma 3.2 that the ratio between the largest and the smallest complementarity products (bounded by γ_u/γ_l) contributes to the increase of $\|Q\|$, we shall pay particular attention to limiting the spread of complementarity products even at the expense of increasing primal and dual infeasibilities ξ_b and ξ_c.

Summing up, we shall look for a candidate initial point (x^*, y^*, s^*) that satisfies the following requirements:

1. $\exists \mu^* : \gamma_l \mu^* \leq x_j^* s_j^* \leq \gamma_u \mu^*$ for some $0 < \gamma_l \leq 1 \leq \gamma_u$ with γ_u/γ_l small;
2. μ^* is small;
3. primal and dual infeasibilities $\xi_b = b_0 - A_0 x^*$ and $\xi_c = c_0 - A_0^T y^* - s^*$ are small.

One could design different heuristics or more rigorous algorithms that would generate a good candidate point (x^*, y^*, s^*) along these lines. We want to apply our *crash* procedure to the solution of large structured linear programs. Hence we expect that the key to its success lies in exploiting the structure of the problem. Our approach uses one iteration of a decomposition method to guess the initial point. We describe our heuristic in detail in the following sections.

Many real-life linear programs display some particular block structures. The structure usually results from system dynamics, uncertainty, spatial distribution, or other factors that lead to the creation of huge problems made up of small, nearly identical parts which have to be coordinated through time, uncertainty, space, or other dimensions. Moreover, modeling very complicated real-life optimization problems often requires nested embedding of structures. The presence of special structure in the problem should be exploited by an interior point method. It can simplify and/or accelerate the execution of linear algebra operations [10].

The problem structure may also be used to find an advanced starting point. We shall illustrate this idea on two well-known classes of specially structured problems: the primal block-angular and the dual block-angular ones. Although for ease of presentation we shall restrict our discussion to those two classes, we shall apply a similar approach also to more complicated linear programs that display nested block structures which combine those two.

Let us recall that the constraint matrix of the linear program with the primal block-angular structure has the following form:

$$A = \begin{pmatrix} A_1 & & & & \\ & A_2 & & & \\ & & \ddots & & \\ & & & A_n & \\ B_1 & B_2 & \cdots & B_n & B_0 \end{pmatrix}, \tag{5.1}$$

where $A_i \in \mathcal{R}^{m_i \times n_i}, i = 1, \ldots, n$, and $B_i \in \mathcal{R}^{m_0 \times n_i}, i = 0, \ldots, n$. Matrix A then has $M = m_0 + \sum_{i=1}^{n} m_i$ rows and $N = \sum_{i=0}^{n} n_i$ columns. The constraint matrix of the linear program with the dual block-angular structure has the following form:

$$A = \begin{pmatrix} A_1 & & & C_1 \\ & A_2 & & C_2 \\ & & \ddots & \vdots \\ & & A_n & C_n \end{pmatrix}, \tag{5.2}$$

where $A_i \in \mathcal{R}^{m_i \times n_i}, i = 1, \ldots, n$, and $C_i \in \mathcal{R}^{m_i \times k}, i = 1, \ldots, n$. Matrix A then has $M = \sum_{i=1}^{n} m_i$ rows and $N = k + \sum_{i=1}^{n} n_i$ columns. In the following sections we will use the partitioning of objective and variable vectors $c = [c_1, \ldots, c_n, c_0], x = [x_1, \ldots, x_n, x_0]$ and likewise the vectors of right-hand sides $b = [b_1, \ldots, b_n, b_0]$ in (5.1) and $b = [b_1, \ldots, b_n]$ in (5.2). Dual variables y for constraints and s for the nonnegativity constraints on x are partitioned as b and x, respectively.

5.1. Decomposition-based starting point. A problem whose constraint matrix is of the forms (5.1), (5.2) lends itself obviously to a decomposition approach. In our experience such a strategy is efficient at finding a near-optimal point quickly; however, it might take a long time to converge to within a specified tolerance. The idea is therefore to construct a starting point for the interior point method from information obtained after one iteration of a decomposition scheme applied to the original problem. One major aim will be to construct a point which is as close to primal and dual feasibility as possible. Assume we apply the Dantzig–Wolfe decomposition [6] to (5.1). After dualizing the coupling constraint, the problem decomposes, and each of the subproblems could be solved independently. We could then combine the subproblem solutions to obtain an advanced starting point for the interior point method. The difficulty with this approach is that while the resulting point is dual feasible in the complete problem, there would be a considerable violation of primal feasibility in the (earlier ignored) coupling constraint. Similarly, applying Benders decomposition [4] to (5.2) and combining subproblem solutions would yield a point which is primal feasible while violating dual feasibility (corresponding to ignored linking variables). Our idea therefore is to combine the two decomposition approaches. Problems (5.1) and (5.2) are extended into forms that allow the application of both Dantzig–Wolfe and Benders decomposition. We will first apply Dantzig–Wolfe decomposition; from its solution, values of complicating variables to use in Benders decomposition can be derived. The Benders subproblem is then solved, and from the solutions of both sets of subproblems we will construct a point which is close to both primal and dual feasibility.

A scheme to solve optimization problems by iterating between Dantzig–Wolfe and Benders subproblems has been suggested as cross-decomposition by van Roy [22] and

Vlahos [23]. We will revise the cross-decomposition algorithm applied to problems of form (5.1) and (5.2). Note, however, that cross-decomposition requires strong assumptions about boundedness and feasibility of the resulting subproblems, which are not satisfied in our case. We will therefore suggest modifications to the algorithm which make it suited for our application.

5.2. Cross-decomposition for primal block-angular structure. Assume in the next two sections that the system matrix A is of form (5.1). The problem could be solved by Dantzig–Wolfe decomposition. However, introducing extra variables $h_i, i = 0, \dots, n$, and constraints $B_i x_i - h_i = 0, i = 0, \dots, n$, leads to the augmented problem

$$\min_{\substack{x_0,\dots,x_n \geq 0 \\ h_0,\dots,h_n}} \sum_{i=0}^{n} c_i^T x_i \quad \text{s.t.} \quad \begin{aligned} A_i x_i &= b_i, \quad i = 1, \dots, n, \\ B_i x_i - h_i &= 0, \quad i = 0, \dots, n, \\ \sum_{i=0}^{n} h_i &= b_0 \end{aligned} \tag{5.3}$$

and enables us to apply Benders decomposition using $h = (h_0, \dots, h_n)$ as complicating variables. The cross-decomposition scheme applied to (5.3) would proceed as follows: a guess of the multiplier $\hat{y}_0$ on the $\sum h_i = b_0$ constraint is obtained. With this the Dantzig–Wolfe subproblem

$$v_D(\hat{y}_0) = \min_{x_0,\dots,x_n \geq 0} \sum_{i=0}^{n} (c_i - B_i^T \hat{y}_0)^T x_i \quad \text{s.t.} \quad A_i x_i = b_i, \quad i = 1, \dots, n, \tag{5.4}$$

is solved. From its solution $(x_0^*, \dots, x_n^*)$, values $\hat{h}_i = B_i x_i^*$ of the Benders complicating variables are obtained, and with these the Benders subproblem

$$v_P(\hat{h}) = \min_{x_0,\dots,x_n \geq 0} \sum_{i=0}^{n} c_i^T x_i \quad \text{s.t.} \quad \begin{aligned} A_i x_i &= b_i, \quad i = 1, \dots, n, \\ B_i x_i &= \hat{h}_i, \quad i = 0, \dots, n, \end{aligned} \tag{5.5}$$

is solved. Multipliers $y_{0,i}^*$ on the $B_i x_i = \hat{h}_i$ constraints are obtained and averaged $\hat{y}_0 = (\sum_i y_{0,i}^*)/(n+1)$, which is then used again in (5.4). Note that problems (5.4) and (5.5) separate into n and $n+1$ smaller problems, respectively. However, in (5.4) the subproblems might be unbounded, and in (5.5) the subproblems might be infeasible. We will now show how this procedure can be used to construct an advanced starting point.

5.3. Crash start for a primal block-angular problem. We are aiming for a point that is both near to primal and dual feasibility and close to the central path. Using the particular system matrix (5.1), we therefore aim to satisfy (compare (2.3))

$$c_i - A_i^T y_i - B_i^T y_0 - s_i = 0, \tag{5.6}$$

$$c_0 - B_0^T y_0 - s_0 = 0, \tag{5.7}$$

$$A_i x_i = b_i, \tag{5.8}$$

$$\sum_{i=0}^{n} B_i x_i = b_0, \tag{5.9}$$

$$S_i X_i e = \mu e, \tag{5.10}$$

$$S_0 X_0 e = \mu e, \tag{5.11}$$

$$s_i, x_i, s_0, x_0 \geq 0. \tag{5.12}$$

Let us now assume that an estimate $\hat{y}_0$ of the complicating constraint multiplier is available (we have used $\hat{y}_0 = e$ in the tests). With this the Dantzig–Wolfe subproblem (5.4) is solved. At the solution (denoted by superscripts $^{(1)}$) the following KKT conditions hold:

$$c_i - A_i^T y_i^{(1)} - B_i^T \hat{y}_0 - s_i^{(1)} = 0, \quad i = 1, \ldots, n, \tag{5.13}$$

$$A_i x_i^{(1)} = b_i, \quad i = 1, \ldots, n. \tag{5.14}$$

Note that as long as $c_i - B_i^T \hat{y}_0 \geq 0$, problem (5.4) is bounded, a condition which will be satisfied in our test problems. From this, estimates $h_i^{(1)} = B_i x_i^{(1)}$ of the Benders complicating variables are obtained. Projecting them on the $\sum h_i = b_0$ constraint by $(\hat{h}_i)_j = (h_i^{(1)})_j (b_0)_j / (\sum h_i^{(1)})_j$, the Benders subproblem (5.5) is solved with $\hat{h}_i$ as complicating variables. Note that this subproblem is not necessarily feasible, and thus (5.5) is replaced by a penalized version

$$v_P(\hat{h}) = \min_{\substack{x_0,\ldots,x_n \geq 0 \\ p_i^+, p_i^- \geq 0}} \sum_{i=0}^{n} c_i^T x_i + \gamma e^T (p_i^+ + p_i^-) \quad \text{s.t.} \quad \begin{aligned} A_i x_i &= b_i, & i &= 1, \ldots, n, \\ B_i x_i - \hat{h}_i &= p_i^+ - p_i^-, & i &= 0, \ldots, n, \end{aligned} \tag{5.15}$$

whose solution (denoted by superscripts $^{(2)}$) satisfies the KKT conditions

$$c_i - A_i^T y_i^{(2)} - B_i^T y_{0,i}^{(2)} - s_i^{(2)} = 0, \tag{5.16}$$

$$A_i x_i^{(2)} = b_i, \tag{5.17}$$

$$B_i x_i^{(2)} - \hat{h}_i = p_i^+ - p_i^-. \tag{5.18}$$

After solving (5.4) and (5.15), we can accumulate an estimated solution (x^*, y^*, s^*) to the linear programming problem of form (5.1) as follows:

$$\begin{aligned} x_i^* &= x_i^{(2)}, \\ y_i^* &= y_i^{(1)}, \\ s_i^* &= s_i^{(1)}, \\ x_0^* &= \max\left\{ B_0^{-1} \left(b_0 - \sum_{i=1}^{n} B_i x_i^{(2)} \right), 0 \right\}, \\ y_0^* &= \hat{y}_0, \\ s_0^* &= c_0 - B_0^T \hat{y}_0, \end{aligned}$$

where it is assumed that B_0^{-1} is easy to compute (such as when B_0 is diagonal). With these choices, dual feasibility is ensured by (5.13) and the definition of s_0^*. $A_i x_i^* = b_i$ holds by (5.17), and $\sum_{i=1}^{0} B_i x_i^* - b_0 = \sum_{i=1}^{n} (p_i^+ - p_i^-)$, which should be small due to our choice of objective in (5.15). Thus the guess is close to primal feasibility. A good spread of complementarity products is achieved as follows.

All subproblems are solved by an interior point method. Rather than ensuring convergence of the subproblem, we are aiming for a point on the central path for a relatively small μ. To achieve this, we choose a target value $\hat{\mu}$ ($\hat{\mu} = 0.01$ has been used in the tests) and obtain an estimate $\tilde{z}_i \approx c_i^T x_i$ of the optimal objective value for each subproblem. The subproblems are solved using

$$\epsilon_i = \frac{\hat{\mu} n_i}{|\tilde{z}_i|} \tag{5.19}$$

in (4.4) as a stopping criterion (see the discussion leading to (4.5)), and two additional recentering steps using multiple centrality correctors are performed. Further, we set $(x_0^*)_j = \max\{1, (x_0^*)_j\}, (s_0^*)_j = \max\{\hat{\mu}, (s_0^*)_j\}$. This ensures that

$$X_i^* S_i^* e \approx \hat{\mu} e, \qquad X_0^* S_0^* e \geq \hat{\mu} e, \tag{5.20}$$

so that the point (x^*, y^*, s^*) should be reasonably well centered for the application of the interior point method.

5.4. Cross-decomposition for dual block-angular structure. For the case in which the system matrix A is of form (5.2) we proceed similarly. We will start by stating the cross-decomposition for this case. In order to apply the Dantzig–Wolfe scheme, we need to introduce additional variables $x_{0,i}, i = 1, \ldots, n$, and constraints $x_{0,i} - x_0 = 0, i = 1, \ldots, n$, to arrive at the augmented problem

$$\min_{\substack{x_0, x_i, x_{0,i} \geq 0 \\ i=1,\ldots,n}} \sum_{i=0}^{n} c_i^T x_i \quad \text{s.t.} \quad \begin{aligned} A_i x_i + C_i x_{0,i} &= b_i, \qquad i = 1, \ldots, n, \\ x_{0,i} - x_0 &= 0, \qquad i = 1, \ldots, n. \end{aligned} \tag{5.21}$$

The cross-decomposition starts by relaxing the $x_{0,i} - x_0 = 0$ constraints. Initial multipliers $\hat{\lambda} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_n)$ are guessed and the subproblem

$$v_D(\hat{\lambda}) = \min_{\substack{x_0, x_i, x_{0,i} \geq 0 \\ i=1,\ldots,n}} \sum_{i=0}^{n} c_i^T x_i + \sum_{i=1}^{n} \hat{\lambda}_i^T (x_{0,i} - x_0) \quad \text{s.t.} \quad A_i x_i + C_i x_{0,i} = b_i, \quad i = 1, \ldots, n, \tag{5.22}$$

is solved. The optimal value of x_0 is obtained and used as complicating variable $\hat{x}_0$ in the Benders subproblem

$$v_P(\hat{x}_0) = \min_{\substack{x_i, x_{0,i} \geq 0 \\ i=1,\ldots,n}} \sum_{i=1}^{n} c_i^T x_i + c_0^T \hat{x}_0 \quad \text{s.t.} \quad \begin{aligned} A_i x_i + C_i x_{0,i} &= b_i, \\ x_{0,i} &= \hat{x}_0. \end{aligned} \tag{5.23}$$

Multipliers λ_i^* on the $x_{0,i} = \hat{x}_0$ constraints are obtained and used as new $\hat{\lambda}_i$ in the next iteration of (5.22). Note that again (5.23) and (5.22) separate into smaller subproblems. Further, (5.22) might be unbounded, just as (5.23) might be infeasible. We will now derive the way in which we modify this scheme to construct an advanced starting point which is close to primal and dual feasibility.

5.5. Crash start for the dual block-angular problem. From the KKT conditions of the augmented system (5.21), $\sum_{i=1}^{n} \lambda_i = c_0$ needs to be satisfied at the solution. We therefore restrict ourselves to such choices of $\hat{\lambda}$ and use, for instance, $\hat{\lambda}_i = c_0/n$ as a starting guess. With this choice, subproblem (5.22) simplifies to

$$v_D(\hat{\lambda}) = \min_{\substack{x_i, x_{0,i} \geq 0 \\ i=1,\ldots,n}} \sum_{i=1}^{n} (c_i^T x_i + \hat{\lambda}_i^T x_{0,i}) \quad \text{s.t} \quad A_i x_i + C_i x_{0,i} = b_i, \quad i = 1, \ldots, n, \tag{5.24}$$

and separates into n smaller subproblems. The Benders complicating variables $\hat{x}_0$ could be obtained by $\hat{x}_0 = \sum x_{0,i}/n$. In our test problems, however, the C_i are the negatives of projection matrices, so that a large value of $\hat{x}_0$ is likely to lead to a feasible Benders subproblem. Therefore we have used $(\hat{x}_0)_j = \max_j\{(x_{0,i})_j\}$. Problem (5.23), however, might still be infeasible, so again it is replaced by a penalized version

$$v_P(\hat{x}_0) = \min_{\substack{x_1, \ldots, x_n \geq 0 \\ p_i^+, p_i^- \geq 0}} \sum_{i=1}^{n} (c_i^T x_i + \gamma e^T (p_i^+ + p_i^-)) \quad \text{s.t.} \quad A_i x_i + p_i^+ - p_i^- = b_i - C_i \hat{x}_0, \tag{5.25}$$

where the $x_{0,i}$ have also been substituted out. If new estimates $\hat{\lambda}_i$ are needed, they can be obtained as $\hat{\lambda}_i = -C_i^T y_i^* - s_i^*$, as can be motivated by the KKT conditions for the augmented system.

The estimate for an advanced solution is again obtained by combining the solutions of (5.24) and (5.25). We are aiming to find a point close to the central path for an LP with system matrix (5.2), so we aim to satisfy

$$c_i - A_i^T y_i - s_i = 0, \tag{5.26}$$

$$c_0 - \sum_{i=1}^{n} C_i^T y_i - s_0 = 0, \tag{5.27}$$

$$A_i x_i + C_i x_0 = b_i, \tag{5.28}$$

$$S_i X_i e = \mu e, \tag{5.29}$$

$$S_0 X_0 e = \mu e, \tag{5.30}$$

$$s_i, x_i, s_0, x_0 \geq 0. \tag{5.31}$$

The solutions to the Lagrangian subproblem (5.24) (denoted by superscripts $^{(1)}$) and the Benders subproblem (5.25) (superscripts $^{(2)}$) satisfy, respectively,

$$A_i^T y_i^{(1)} + s_i^{(1)} = c_i, \tag{5.32}$$

$$C_i^T y_i^{(1)} + s_{0,i}^{(1)} = \hat{\lambda}_i, \tag{5.33}$$

$$A_i x_i^{(1)} + C_i x_{0,i}^{(1)} = b_i, \tag{5.34}$$

$$A_i^T y_i^{(2)} + s_i^{(2)} = c_i, \tag{5.35}$$

$$A_i x_i^{(2)} + C_i \hat{x}_0 + p_i^+ - p_i^- = b_i. \tag{5.36}$$

We combine these solutions by choosing

$$\begin{aligned} x_i^* &= x_i^{(2)}, \\ y_i^* &= y_i^{(1)}, \\ s_i^* &= s_i^{(1)}, \\ x_0^* &= \hat{x}_0, \\ s_0^* &= \sum s_{0,i}^{(1)}. \end{aligned}$$

With these choices, (5.26) is satisfied by (5.32); (5.27) follows from (5.33), together with the definition of s_0^* and the fact that $c_0 = \sum \hat{\lambda}_i$. Together the advanced solution is dual feasible. Further, we have that the residual of (5.28) is $p_i^+ - p_i^-$, which should be small.

To obtain a fairly well centered point, we apply the same heuristic as in the last section. We solve subproblems only to a relative accuracy of ϵ_i given by (5.19) and set $(x_0^*)_j = \max\{1, (x_0^*)_j\}, (s_0^*)_j = \max\{\hat{\mu}, (s_0^*)_j\}$.

6. Numerical results. We have tested our approach in the context of OOPS, the object-oriented parallel solver [10]. We have implemented the reoptimization procedure and our crash procedure for constructing an advanced starting point as described in section 5. We have applied the *crash* starting in the solution of several different classes of structured linear programs.

TABLE 6.1
Problem statistics.

Problem	Rows	Columns	Blks
MCNF RN	14232	72996	119
MCNF R2	3621	16600	60
MCNF R4	6981	20850	70
MCNF R6	8715	51300	86
MCNF R13	88613	460000	288
MCNF R14	159602	637600	398
MSND PB1	22213	72514	81
MSND PB2	59021	207901	102
MSND PB3	54657	188266	109
MSND PB4	83561	294735	123
MSND PB5	242570	886178	179
JOSBP T1	1021	2400	26
JOSBP T2	3414	7266	43
JOSBP T3	13053	26860	80
JOSBP P1	3241	6970	42
JOSBP P2	6492	13978	59
JOSBP P3	14221	32760	92
ALM P1	66666	166665	101
ALM P2	666666	1666665	101
ALM P3	1222221	3333330	101

Our computational results demonstrate the performance of the crash procedure and give an insight into the practical behavior of the reoptimization strategy. All classes of problems solved in this paper are well documented in the literature, so we restrict their description to an explanation of the associated block structures. These problems originate from network optimization [1] and multistage stochastic programming applied to asset liability management [28].

Multicommodity network flow problems (MCNF) are of primal block-angular structure; all other problems display a nested dual block-angular structure: multicommodity survivable network design (MSND) and joint optimal synthesis of base and spare capacity (JOSBP) problems have primal block-angular subblocks. Detailed formulations of all these problems can be found in [10]. The asset liability management (ALM) problems have dual block-angular subblocks.

We have used the algorithm to generate an advanced starting point as described in section 5 with a few minor variations. First note that not all the slack variables p_i^+, p_i^- in (5.15) and (5.25) are necessary since some slacks might be easily picked up by the problem variables. We have removed these slacks from the subproblems as far as possible. Further, for the ALM problems the complicating variable costs c_0 are zero and the default choice of $\hat{\lambda}$ ($\hat{\lambda}_i = c_0/n = 0$) would lead to unbounded Lagrangian relaxation subproblems. We have therefore used a different choice of $\hat{\lambda}$, still satisfying $\sum_i \hat{\lambda}_i = c_0 = 0$, that guarantees bounded subproblems.

In Table 6.1 we report problem statistics. The problems are grouped by category. For each problem we give its size in numbers of rows and columns and the number of diagonal blocks. In Table 6.2 we report the results of our method. The column following the problem name contains the number of iterations to reach optimal solution from the default starting point (which is based on Mehrotra [16]). The final block states the results for our algorithm. Its first column reports the number of iterations of the interior point method starting from the crash point and using the reoptimization. The following columns give some useful numbers: $\|\tilde{\xi}_b\|_\infty$ and $\|\tilde{\xi}_c\|_\infty$ at the start of the reoptimization, γ_u/γ_l as a measure of initial centrality of the generated point, and

TABLE 6.2
Solution statistics.

Problem	Default Iters	Advanced starting point Iters	$\|\tilde{\xi_b}\|_\infty$	$\|\tilde{\xi_c}\|_\infty$	$(\gamma_u/\gamma_l)^{\frac{1}{2}}$	itf	α_P	α_D
MCNF RN	31	22	7.1e+3	6.9e-1	2236	22	3.0e-2	7.9e-2
MCNF R2	20	12	4.0e+3	4.8e-4	875	12	1.8e-2	7.2e-2
MCNF R4	18	10	2.7e+4	1.1e-3	504	8	1.6e-1	2.8e-1
MCNF R6	20	12	7.4e+4	3.4e-3	1240	11	1.3e-2	7.5e-2
MCNF R13	37	26	3.9e+5	5.0e-3	3435	26	8.5e-3	3.3e-2
MCNF R14	49	38	3.3e+5	3.0e-3	3332	37	1.1e-2	1.4e-2
MSND PB1	25	20	2.81	0.97	269	10	8.3e-2	5.8e-2
MSND PB2	33	20	1.94	0.96	369	14	3.1e-2	6.6e-2
MSND PB3	29	17	1.40	0.97	359	12	1.0e-1	1.6e-2
MSND PB4	36	22	1.66	0.97	462	18	1.2e-1	1.8e-3
MSND PB5	51	29	1.59	0.97	589	20	2.6e-2	4.6e-3
JOSBP T1	15	12	14.3	0.99	74	8	3.6e-2	3.0e-1
JOSBP T2	22	19	13.6	0.97	108	18	9.3e-3	1.3e-2
JOSBP T3	28	22	1.23	1.00	32	13	1.3e-2	1.2e-2
JOSBP P1	25	19	1.17	1.00	42	13	1.6e-2	4.6e-2
JOSBP P2	27	22	0.96	1.00	41	13	1.1e-2	4.2e-2
JOSBP P3	40	37	0.70	1.00	34	7	1.0e-2	3.5e-2
ALM P1	21	12	2.7e+3	0.96	741	4	3.5e-1	3.9e-1
ALM P2	44	22	4.6e+3	0.97	2280	5	3.6e-3	2.7e-1
ALM P3	66	30	8.4e+4	0.98	2626	6	2.6e-2	2.5e-1

TABLE 6.3
Speed-ups for parallel implementation.

Prob	Default start 1 proc Time	2 procs Time	2 procs s-up	4 procs Time	4 procs s-up	Advanced start 1 proc Time	2 procs Time	2 procs s-up	4 procs Time	4 procs s-up
R13	1855	952	1.95	468	3.96	1427	734	1.94	377	3.79
R14	3560	1782	1.99	892	3.99	2730	1387	1.97	715	3.82
PB4	644	373	1.73	185	3.38	625	361	1.73	180	3.47
PB5	3255	1859	1.75	1003	3.25	2662	1491	1.78	850	3.13
P2	6070	3124	1.94	1674	3.62	3593	1813	1.98	1067	3.36
P3	29752	15143	1.96	7804	3.81	16039	8350	1.92	5118	3.13

itf, the number of steps needed to reach $\max\{\|\xi_c\|_\infty, \|\xi_b\|_\infty\} \leq 0.01$, together with the initial stepsizes in primal and dual spaces as an indication of how fast primal and dual feasibility is regained.

It can be seen that in all cases the interior point method needed fewer iterations to converge to a solution from our advanced starting point than from the default starting point. On average about 33% of iterations could be saved. For the six largest of our test problems, an average of 40% of iterations was saved, and this translated into a 25% decrease in CPU time, as can be seen in Table 6.3.

For all problem classes the advanced starting point is dual feasible. However, for the dual block-angular problems s_0^* might have to be bounded away from zero, resulting in a scaled dual infeasibility of ≈ 1. For classes MSND and JOSPB, our choice for x_0^* is expected to lead to a solution of (5.25) with zero slacks, resulting in a primal feasible advanced starting point; for the other classes, however, primal feasibility is more difficult to achieve. These observations are reflected in the results.

It should be noted that for most problems the measure of centrality γ_u/γ_l is fairly large. This results as expected from the earlier discussion in small initial stepsizes α_P, α_D and hence a large number of iterations to regain feasibility. However, en-

couragingly, even in these adverse circumstances, the reoptimization strategy obtains good results.

Our implementation, OOPS, of the primal-dual method is a parallel solver, and the crash start has been implemented in parallel in the obvious way by distributing the Benders and Dantzig–Wolfe subproblems amongst processors. The program has been run on a Sun Enterprise 6500 parallel computer with four processors. Table 6.3 states the CPU times (first values) and obtained speed-ups (second values) on two and four processors for the largest of our test problems. Speed-ups between 3.1–3.8 were obtained. These are consistent with the speed-ups obtained for OOPS without crash start.

7. Conclusions. We have discussed a strategy that exploits a well-centered solution of one linear program to generate a good starting point for a perturbed problem. We have given bounds on the size of perturbations of the primal and dual feasibility that can be absorbed in merely one Newton step. Our bounds critically depend on the size of perturbations measured in the scaled space of primal-dual solutions and can easily be computed in practical situations that require warm-starting. They can thus be used to facilitate a choice of one well-centered point from the list of candidates if such a list is available.

We have performed the analysis for the *feasible* path-following methods. We have shown that the measure of proximity to the central path appropriate for each of these methods will not be corrupted by the feasibility restoration step (if the perturbations are small, of course). We have then translated our findings into a computational practice of *infeasible* path-following method. The practical reoptimization strategy spreads the process of restoring feasibility into few subsequent iterations. If the perturbations are large, then only a fraction of them is absorbed in a single iteration.

Finally, we have applied the reoptimization technique to allow the start of an interior point method from almost an arbitrary point. We have provided numerical results which confirm that this strategy works well for large structured linear programs and that it can be efficiently implemented in parallel.

Acknowledgment. We are grateful to the anonymous referee for his comments.

REFERENCES

[1] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin, *Network Flows*, Prentice–Hall, New York, 1993.

[2] 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, 2000, pp. 197–232.

[3] E. D. Andersen, J. Gondzio, C. Mészáros, and X. Xu, *Implementation of interior point methods for large scale linear programming*, in Interior Point Methods in Mathematical Programming, T. Terlaky, ed., Kluwer Academic Publishers, Norwell, MA, 1996, pp. 189–252.

[4] J. F. Benders, *Partitioning procedures for solving mixed-variables programming problems*, Numer. Math., 4 (1962), pp. 238–252.

[5] R. E. Bixby, *Implementing the simplex method: The initial basis*, ORSA J. Computing, 4 (1992), pp. 267–284.

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

[7] I. I. Dikin, *On the speed of an iterative process*, Upravlaemye Sistemy, 12 (1974), pp. 54–60.

[8] J. Gondzio, *Multiple centrality corrections in a primal-dual method for linear programming*, Comput. Optim. Appl., 6 (1996), pp. 137–156.

[9] J. Gondzio, *Warm start of the primal-dual method applied in the cutting plane scheme*, Math. Program., 83 (1998), pp. 125–143.

[10] J. Gondzio and R. Sarkissian, *Parallel Interior Point Solver for Structured Linear Programs*, Technical report MS-00-025, Department of Mathematics and Statistics, University of Edinburgh, Edinburgh, UK, 2000; revised 2002.
[11] J. Gondzio and J.-P. Vial, *Warm start and ε-subgradients in cutting plane scheme for block-angular linear programs*, Comput. Optim. Appl., 14 (1999), pp. 17–36.
[12] N. I. M. Gould and J. K. Reid, *New crash procedures for large-scale systems of linear constraints*, Math. Programming, 45 (1989), pp. 475–501.
[13] J. E. Kelley, Jr., *The cutting-plane method for solving convex programs*, J. SIAM, 8 (1960), pp. 703–712.
[14] I. J. Lustig, R. E. Marsten, and D. F. Shanno, *Interior point methods for linear programming: Computational state of the art*, ORSA J. Computing, 6 (1994), pp. 1–14.
[15] I. Maros and G. Mitra, *Startegies for creating advanced bases for large-scale linear programming problems*, INFORMS J. Computing, 10 (1998), pp. 248–260.
[16] S. Mehrotra, *On the implementation of a primal-dual interior point method*, SIAM J. Optim., 2 (1992), pp. 575–601.
[17] J. E. Mitchell, *Karmakar's Algorithm and Combinatorial Optimization Problems*, Ph.D. thesis, Cornell University, Ithaca, NY, 1988.
[18] J. E. Mitchell, *Computational experience with an interior point cutting plane algorithm*, SIAM J. Optim., 10 (2000), pp. 1212–1227.
[19] J. E. Mitchell and M. J. Todd, *Solving combinatorial optimization problems using Karmarkar's algorithm*, Math. Programming, 56 (1992), pp. 245–284.
[20] M. A. Nunez and R. M. Freund, *Condition measures and properties of the central trajectory*, Math. Program., 83 (1998), pp. 1–28.
[21] J. Peng, C. Roos, and T. Terlaky, *A new and efficient large-update interior-point method for linear optimization*, J. Comput. Technol., 6 (2001), pp. 61–80.
[22] T. J. Van Roy, *Cross decomposition for mixed integer programming*, Math. Programming, 25 (1983), pp. 46–63.
[23] K. Vlahos, *Generalised Cross Decomposition Application to Electricity Capacity Planning*, Technical report, London Business School, London, 1991.
[24] L. A. Wolsey, *Integer Programming*, John Wiley and Sons, New York, 1998.
[25] S. J. Wright, *Primal-Dual Interior-Point Methods*, SIAM, Philadelphia, 1997.
[26] E. A. Yildirim and M. J. Todd, *Sensitivity analysis in linear programming and semidefinite programming using interior-point methods*, Math. Program., 90 (2001), pp. 229–261.
[27] E. A. Yildirim and S. J. Wright, *Warm-start strategies in interior-point methods for linear programming*, SIAM J. Optim., 12 (2002), pp. 782–810.
[28] W. T. Ziemba and J. M. Mulvey, *Worldwide Asset and Liability Modeling*, Publications of the Newton Institute, Cambridge University Press, Cambridge, UK, 1998.

SIAM J. OPTIM.
Vol. 13, No. 3, pp. 865–888

THE PRIMAL-DUAL ACTIVE SET STRATEGY AS A SEMISMOOTH NEWTON METHOD*

M. HINTERMÜLLER†, K. ITO‡, AND K. KUNISCH†

Abstract. This paper addresses complementarity problems motivated by constrained optimal control problems. It is shown that the primal-dual active set strategy, which is known to be extremely efficient for this class of problems, and a specific semismooth Newton method lead to identical algorithms. The notion of slant differentiability is recalled and it is argued that the max-function is slantly differentiable in L^p-spaces when appropriately combined with a two-norm concept. This leads to new local convergence results of the primal-dual active set strategy. Global unconditional convergence results are obtained by means of appropriate merit functions.

Key words. complementarity problems, function spaces, semismooth Newton method

AMS subject classifications. 90C33, 65K10, 65J99

PII. S1052623401383558

1. Introduction. This paper is motivated by linearly constrained quadratic problems of the type

$$\text{(P)} \qquad \begin{cases} \min J(y) = \frac{1}{2}(y, Ay) - (f, y) \\ \text{subject to } y \le \psi, \end{cases}$$

where A is positive definite and f, ψ are given. In previous contributions [IK1, IK2, BIK, BHHK] we proposed a primal-dual active set strategy as an extremely efficient method to solve (P). We shall show in the present work that the primal-dual active set method can be interpreted as a semismooth Newton method. This opens up a new interpretation and perspective of analyzing the primal-dual active set method. Both the finite dimensional case with $y \in \mathbb{R}^n$ and the infinite dimensional case with $y \in L^2(\Omega)$ will be considered. While our results are quite generally applicable the main motivation arises from infinite dimensional constrained variational problems and their discretization. Frequently such problems have a special structure which can be exploited. For example, in the case of discretized obstacle problems A can be an M-matrix, and for constrained optimal control problems A is a smooth additive perturbation of the identity operator.

The analysis of semismooth problems and the Newton algorithm to solve such problems has a long history for finite dimensional problems. We refer to selected papers [Q1, Q2, QS] and the references therein. Typically, under appropriate semismoothness and regularity assumptions locally superlinear convergence rates of semismooth Newton methods are obtained. Since many definitions used in the above papers depend on Rademacher's theorem, which has no analogue in infinite dimensions, very recently, e.g., in [CNQ, U] new concepts for generalized derivatives and semismoothness in infinite dimensional spaces were introduced. In our work we primarily use the notion of slant differentiability from [CNQ] which we recall for the

*Received by the editors January 15, 2001; accepted for publication (in revised form) August 7, 2002; published electronically February 12, 2003.
http://www.siam.org/journals/siopt/13-3/38355.html
†Institute of Mathematics, University of Graz, A-8010 Graz, Austria (michael.hintermueller@uni-graz.at, karl.kunisch@uni-graz.at).
‡North Carolina State University, Raleigh, NC 27695 (kito@unity.ncsu.edu).

reader's convenience at the end of this section. For the problem under consideration it coincides with the differentiability concept in [U]. This will be explained in section 4.

Let us briefly outline the structure of the paper. In section 2 the relationship between the primal-dual active set method and semismooth Newton methods is explained. Local as well as global convergence for finite dimensional problems, which is unconditional with respect to initialization in certain cases, is addressed in section 3. The global convergence results depend on properties of the matrix A. For instance, the M-matrix property required in Theorem 3.2 is typically obtained when discretizing obstacle problems (see, e.g., [H, KNT]) by finite differences or finite elements. Theorem 3.3 can be connected to discretizations of control constrained optimal control problems. Some relevant numerical aspects of the conditions of Theorem 3.3 are discussed at the end of section 4. An instance of the perturbation result of Theorem 3.4 is given by discretized optimal control problems with sufficiently small cost parameter. Perturbations of M-matrices resulting from discretized obstacle problems and state constrained optimal control problems (see, e.g., [Ca]) fit into the framework of Theorem 3.4. In section 4 slant differentiability properties of the max-function between function spaces are analyzed. Superlinear convergence of semismooth Newton methods for optimal control problems with pointwise control constraints is proved. Several alternative methods were analyzed to solve optimal control problems with pointwise constraints on the controls. Among them are the projected Newton method, analyzed, e.g., in [HKT, KS] and affine scaling interior point Newton methods [UU]. We plan to address nonlinear problems in a future work. Let us stress, however, that nonlinear iterative methods frequently rely on solving auxiliary problems of the type (P), and solving them efficiently is important.

To briefly describe some of the previous work in the primal-dual active set method, we recall that this method arose as a special case of generalized Moreau–Yosida approximations to nondifferentiable convex functions [IK1]. Global convergence proofs based on a modified augmented Lagrangian merit function are contained in [BIK]. In [BHHK] comparisons between the primal-dual active set method and interior point methods are carried out. In [IK2] the primal-dual active set method was used to solve optimal control of variational inequalities problems. For this class of problems, convergence proofs are not yet available.

We now turn to the notion of differentiability which will be used in this paper. Let X and Z be Banach spaces and consider the nonlinear equation

$$F(x) = 0\,, \tag{1.1}$$

where $F\colon D \subset X \to Z$, and D is an open subset of X.

DEFINITION 1. *The mapping $F\colon D \subset X \to Z$ is called slantly differentiable in the open subset $U \subset D$ if there exists a family of mappings $G\colon U \to \mathcal{L}(X, Z)$ such that*

$$\lim_{h\to 0} \frac{1}{\|h\|} \|F(x+h) - F(x) - G(x+h)h\| = 0 \tag{A}$$

for every $x \in U$.

We refer to G as a slanting function for F in U. Note that G is not required to be unique to be a slanting function for F in U. The definition of slant differentiability in an open set is a slight adaptation of the terminology introduced in [CNQ], where in addition it is required that $\{G(x) : x \in U\}$ is bounded in $\mathcal{L}(X, Z)$. In [CNQ] also

the term slant differentiability at a point is introduced. In applications to Newton's method this presupposes knowledge of the solution, whereas slant differentiability of F in U requires knowledge of a set which contains the solution. Under the assumption of slant differentiability in an open set, Newton's method converges superlinearly for appropriate choices of the initialization. While this discussion is perhaps more philosophical than mathematical, we mention it because the assumption of slant differentiability in an open set parallels the hypothesis of knowledge of the domain within which a second order sufficient optimality condition is satisfied for smooth problems.

Kummer [K2] introduced a notion similar to slant differentiability at a point and coined the name Newton map. He also pointed out the discrepancy between the requirements needed for numerical realization and for the proof of superlinear convergence of the semismooth Newton method.

The following convergence result is already known [CNQ].

THEOREM 1.1. *Suppose that x^* is a solution to (1.1) and that F is slantly differentiable in an open neighborhood U containing x^* with slanting function $G(x)$. If $G(x)$ is nonsingular for all $x \in U$ and $\{\|G(x)^{-1}\| : x \in U\}$ is bounded, then the Newton iteration*

$$x^{k+1} = x^k - G(x^k)^{-1}F(x^k)$$

converges superlinearly to x^*, provided that $\|x^0 - x^*\|$ is sufficiently small.

We provide the short proof since it will be used to illustrate the subsequent discussion.

Proof. Note that the Newton iterates satisfy

$$\|x^{k+1} - x^*\| \le \|G(x^k)^{-1}\| \, \|F(x^k) - F(x^*) - G(x^k)(x^k - x^*)\|, \tag{1.2}$$

provided that $x^k \in U$. Let $B(x^*, r)$ denote a ball of radius r centered at x^* contained in U and let M be such that $\|G(x)^{-1}\| \le M$ for all $x \in B(x^*, r)$. We apply (A) with $x = x^*$. Let $\eta \in (0, 1]$ be arbitrary. Then there exists $\rho \in (0, r)$ such that

$$\|F(x^* + h) - F(x^*) - G(x^* + h)h\| < \frac{\eta}{M} \, \|h\| \le \frac{1}{M} \, \|h\| \tag{1.3}$$

for all $\|h\| < \rho$. Consequently, if we choose x^0 such that $\|x^0 - x^*\| < \rho$, then by induction from (1.2), (1.3) with $h = x^k - x^*$ we have $\|x^{k+1} - x^*\| < \rho$ and in particular $x^{k+1} \in B(x^*, \rho)$. It follows that the iterates are well-defined. Moreover, since $\eta \in (0, 1]$ is chosen arbitrarily $x^k \to x^*$ converges superlinearly. ∎

Note that replacing property (A) by a condition of the type

$$\lim_{h \to 0} \frac{1}{\|h\|} \, \|F(x) - F(x - h) - G(x)h\| = 0$$

would require a uniformity assumption with respect to $x \in U$ for Theorem 1.1 to remain valid in the case where X is infinite dimensional.

Let us put the concept of slant differentiability into perspective with the notion of semismoothness as introduced in [Mi] for real-valued functions and extended in [QS] to finite dimensional vector-valued functions. Semismoothness of $F : U \subset \mathbb{R}^n \to \mathbb{R}^m$ in the sense of Qi and Sun [QS] implies

$$\|F(x + h) - F(x) - Vh\| = o(\|h\|) \tag{1.4}$$

for $x \in U$, where V is an arbitrary element of the generalized Jacobian $\partial F(x+h)$ in the sense of Clarke [C, Prop. 2.6.2]. Thus, slant differentiability introduced in Definition 1 is a more general concept. In fact, the slanting functions according to Definition 1 are not required to be elements of $\partial F(x+h)$. On the other hand, if (1.4) holds for $x \in U \subset \mathbb{R}^n$, then a single-valued selection $V(x) \in \partial F(x)$, $x \in U$, serves as a slanting function in the sense of Definition 1.

We shall require the notion of a P-matrix which we recall next.

DEFINITION 2. *An $n \times n$-matrix is called a* P*-matrix if all its principal minors are positive.*

It is well known [BP] that A is a P-matrix if and only if all real eigenvalues of A and of its principal submatrices are positive. Here B is called a principal submatrix of A if it arises from A by deletion of rows and columns from the same index set $\mathcal{J} \subset \{1, \dots, n\}$.

2. The primal-dual active set strategy as semismooth Newton method. In this section we consider complementarity problems of the form

$$\begin{cases} Ay + \lambda = f, \\ y \le \psi,\ \lambda \ge 0,\ (\lambda, y - \psi) = 0\,, \end{cases} \tag{2.1}$$

where $(\cdot,\cdot)$ denotes the inner product in $\mathbb{R}^n$, A is an $n \times n$-valued P-matrix, and f, $\psi \in \mathbb{R}^n$. The assumption that A is a P-matrix guarantees the existence of a unique solution $(y^*, \lambda^*) \in \mathbb{R}^n \times \mathbb{R}^n$ of (2.1) [BP]. In the case where A is symmetric positive definite (2.1) is the optimality system for

$$\begin{cases} \min J(y) = \dfrac{1}{2}(y, Ay) - (f, y) \\ \text{subject to } y \le \psi. \end{cases} \tag{P}$$

Note that the complementarity system given by the second line in (2.1) can equivalently be expressed as

$$\mathcal{C}(y, \lambda) = 0, \text{ where } \mathcal{C}(y, \lambda) = \lambda - \max(0, \lambda + c(y - \psi)) \tag{2.2}$$

for each $c > 0$. Here the max-operation is understood componentwise.

Consequently, (2.1) is equivalent to

$$\begin{cases} Ay + \lambda = f, \\ \mathcal{C}(y, \lambda) = 0. \end{cases} \tag{2.3}$$

The primal-dual active set method is based on using (2.2) as a prediction strategy; i.e., given a current primal-dual pair (y, λ), the choice for the next active and inactive sets is given by

$$\mathcal{I} = \{i \colon \lambda_i + c(y - \psi)_i \le 0\} \text{ and } \mathcal{A} = \{i \colon \lambda_i + c(y - \psi)_i > 0\}.$$

This leads to the following algorithm.

PRIMAL-DUAL ACTIVE SET ALGORITHM.

(i) Initialize y^0, λ^0. Set $k = 0$.

(ii) Set $\mathcal{I}_k = \{i \colon \lambda_i^k + c(y^k - \psi)_i \le 0\}$, $\mathcal{A}_k = \{i \colon \lambda_i^k + c(y^k - \psi)_i > 0\}$.

(iii) Solve

$$Ay^{k+1} + \lambda^{k+1} = f,$$

$$y^{k+1} = \psi \text{ on } \mathcal{A}_k, \lambda^{k+1} = 0 \text{ on } \mathcal{I}_k.$$

(iv) Stop, or set $k = k + 1$ and return to (ii).

Above we utilize $y^{k+1} = \psi$ on $\mathcal{A}_k$ to stand for $y_i^{k+1} = \psi_i$ for $i \in \mathcal{A}_k$. Let us now argue that the above algorithm can be interpreted as a semismooth Newton method. For this purpose it will be convenient to arrange the coordinates in such a way that the active and inactive ones occur in consecutive order. This leads to the block matrix representation of A as

$$A = \begin{pmatrix} A_{\mathcal{I}_k} & A_{\mathcal{I}_k\mathcal{A}_k} \\ A_{\mathcal{A}_k\mathcal{I}_k} & A_{\mathcal{A}_k} \end{pmatrix},$$

where $A_{\mathcal{I}_k} = A_{\mathcal{I}_k\mathcal{I}_k}$ and analogously for $A_{\mathcal{A}_k}$. Analogously the vector y is partitioned according to $y = (y_{\mathcal{I}_k}, y_{\mathcal{A}_k})$ and similarly for f and ψ. In section 3 we shall argue that $v \to \max(0, v)$ from $\mathbb{R}^n \to \mathbb{R}^n$ is slantly differentiable with a slanting function given by the diagonal matrix $G_m(v)$ with diagonal elements

$$G_m(v)_{ii} = \begin{cases} 1 & \text{if} \quad v_i > 0, \\ 0 & \text{if} \quad v_i \le 0. \end{cases}$$

Here we use the subscript m to indicate particular choices for the slanting function of the max-function. Note that G_m is also an element of the generalized Jacobian (see [C, Definition 2.6.1]) of the max-function. Semismooth Newton methods for generalized Jacobians in Clarke's sense were considered, e.g., in [Q1, QS].

The choice G_m suggests a semismooth Newton step of the form

$$\begin{pmatrix} A_{\mathcal{I}_k} & A_{\mathcal{I}_k\mathcal{A}_k} & I_{\mathcal{I}_k} & 0 \\ A_{\mathcal{A}_k\mathcal{I}_k} & A_{\mathcal{A}_k} & 0 & I_{\mathcal{A}_k} \\ 0 & 0 & I_{\mathcal{I}_k} & 0 \\ 0 & -cI_{\mathcal{A}_k} & 0 & 0 \end{pmatrix} \begin{pmatrix} \delta y_{\mathcal{I}_k} \\ \delta y_{\mathcal{A}_k} \\ \delta\lambda_{\mathcal{I}_k} \\ \delta\lambda_{\mathcal{A}_k} \end{pmatrix} = - \begin{pmatrix} (Ay^k + \lambda^k - f)_{\mathcal{I}_k} \\ (Ay^k + \lambda^k - f)_{\mathcal{A}_k} \\ \lambda^k_{\mathcal{I}_k} \\ -c(y^k - \psi)_{\mathcal{A}_k} \end{pmatrix}, \tag{2.4}$$

where $I_{\mathcal{I}_k}$ and $I_{\mathcal{A}_k}$ are identity matrices of dimensions $\operatorname{card}(\mathcal{I}_k)$ and $\operatorname{card}(\mathcal{A}_k)$. The third equation in (2.4) implies that

$$\lambda^{k+1}_{\mathcal{I}_k} = \lambda^k_{\mathcal{I}_k} + \delta\lambda_{\mathcal{I}_k} = 0 \tag{2.5}$$

and the last one yields

$$y^{k+1}_{\mathcal{A}_k} = \psi_{\mathcal{A}_k}. \tag{2.6}$$

Equations (2.5) and (2.6) coincide with the conditions in the second line of step (iii) in the primal-dual active set algorithm. The first two equations in (2.4) are equivalent to $Ay^{k+1} + \lambda^{k+1} = f$, which is the first equation in step (iii).

Combining these observations we can conclude that the semismooth Newton update based on (2.4) is equivalent to the primal-dual active set strategy.

We also note that the system (2.4) is solvable since the first equation in (2.4) together with (2.5) gives

$$(A\ \delta y)_{\mathcal{I}_k} + (A\ y^k)_{\mathcal{I}_k} = f_{\mathcal{I}_k},$$

and consequently by (2.6)

$$A_{\mathcal{I}_k}\ y^{k+1}_{\mathcal{I}_k} = f_{\mathcal{I}_k} - A_{\mathcal{I}_k\mathcal{A}_k}\ \psi_{\mathcal{A}_k}. \tag{2.7}$$

Since A is a P-matrix, $A_{\mathcal{I}_k}$ is regular and (2.7) determines $y^{k+1}_{\mathcal{I}_k}$. The second equation in (2.4) is equivalent to

$$\lambda^{k+1}_{\mathcal{A}_k} = f_{\mathcal{A}_k} - (Ay^{k+1})_{\mathcal{A}_k}. \tag{2.8}$$

In section 4 we shall consider (P) in the space $L^2(\Omega)$. Again one can show that the semismooth Newton update and the primal-dual active set strategy coincide.

3. Convergence analysis: The finite dimensional case. This section is devoted to local as well as global convergence analysis of the primal-dual active set algorithm to solve

$$\begin{cases} Ay + \lambda = f, \\ \lambda - \max(0, \lambda + c(y - \psi)) = 0, \end{cases} \tag{3.1}$$

where $f \in \mathbb{R}^n$, $\psi \in \mathbb{R}^n$, $A \in \mathbb{R}^{n\times n}$ is a P-matrix, and the max-operation is understood componentwise. To discuss slant differentiability of the max-function we define for an arbitrarily fixed $\delta \in \mathbb{R}^n$ the matrix-valued function $G_m \colon \mathbb{R}^n \to \mathbb{R}^{n\times n}$ by

$$G_m(y) = \operatorname{diag}\,(g_1(y_1), \dots, g_n(y_n)), \tag{3.2}$$

where $g_i \colon \mathbb{R} \to \mathbb{R}$ is given by

$$g_i(z) = \begin{cases} 0 & \text{if} \quad z < 0\,, \\ 1 & \text{if} \quad z > 0\,, \\ \delta_i & \text{if} \quad z = 0\,. \end{cases}$$

LEMMA 3.1. *The mapping $y \to \max(0, y)$ from $\mathbb{R}^n$ to $\mathbb{R}^n$ is slantly differentiable on $\mathbb{R}^n$, and G_m defined in* (3.2) *is a slanting function for every $\delta \in \mathbb{R}^n$.*

Proof. Clearly, $G_m \in \mathcal{L}(\mathbb{R}^n)$ and $\{\|G_m(y)\| \colon y \in \mathbb{R}^n\}$ is bounded. We introduce $D \colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ by

$$D(y, h) = \|\max(0, y + h) - \max(0, y) - G_m(y + h)h\|.$$

It is simple to check that

$$D(y, h) = 0 \ \text{ if } \ \|h\|_\infty < \min\{|y_i| \colon y_i \neq 0\} =: \beta.$$

Consequently, the max-function is slantly differentiable. ∎

Remark 3.1. Note that the value of the generalized derivative G_m of the max-function can be assigned an arbitrary value at the coordinates satisfying $y_i = 0$. The numerator D in Definition 1 satisfies $D(y, h) = 0$ if $\|h\|_\infty < \beta$. Moreover, for every $\gamma > \beta$ there exists h satisfying

$$D(y, h) \geq \beta \ \text{ and } \ \|h\|_\infty = \gamma.$$

Here we assume that $\beta := 0$ whenever $\{i | y_i \neq 0\} = \emptyset$. Consequently, for $\beta > 0$ the mapping

$$\gamma \mapsto \sup\{\|\max(0, y + h) - \max(0, y) - G_m(y + h)h\|_\infty \colon \|h\|_\infty = \gamma\}$$

is discontinuous at $\gamma = \beta$ and equals zero for $\gamma \in (0, \beta)$. ∎

Let us now turn to the convergence analysis of the primal-dual active set method or, equivalently, the semismooth Newton method for (3.1). Note that the choice G_m for the slanting function in section 2 corresponds to a slanting function with $\delta = 0$. In view of (2.5)–(2.8) for $k \geq 1$ the Newton update (2.4) is equivalent to

$$\begin{pmatrix} A_{\mathcal{I}_k} & 0 \\ A_{\mathcal{A}_k\mathcal{I}_k} & I_{\mathcal{A}_k} \end{pmatrix} \begin{pmatrix} \delta y_{\mathcal{I}_k} \\ \delta\lambda_{\mathcal{A}_k} \end{pmatrix} = \ - \begin{pmatrix} A_{\mathcal{I}_k\mathcal{A}_k}\delta y_{\mathcal{A}_k} + \delta\lambda_{\mathcal{I}_k} \\ A_{\mathcal{A}_k}\delta y_{\mathcal{A}_k} \end{pmatrix} \tag{3.3}$$

and

$$\delta\lambda_i = -\lambda_i^k, \ i \in \mathcal{I}_k, \ \text{ and } \ \delta y_i = \psi_i - y_i^k, \ i \in \mathcal{A}_k. \tag{3.4}$$

Let us introduce $F\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^n$ by

$$F(y,\lambda) = \begin{pmatrix} Ay + \lambda - f \\ \lambda - \max(0, \lambda + c(y - \psi)) \end{pmatrix},$$

and note that (3.1) is equivalent to $F(y,\lambda) = 0$. As a consequence of Lemma 3.1 the mapping F is slantly differentiable and the system matrix of (2.4) is a slanting function for F with the particular choice G_m for the slanting function of the max-function. We henceforth denote the slanting function of F by G_F.

Let (y^*, λ^*) denote the unique solution to (3.1) and $x^0 = (y^0, \lambda^0)$ the initial values of the iteration. From Theorem 1.1 we deduce the following fact.

THEOREM 3.1. *The primal-dual active set method or, equivalently, the semismooth Newton method converge superlinearly to $x^* = (y^*, \lambda^*)$, provided that $\|x^0 - x^*\|$ is sufficiently small.*

The boundedness requirement of $(G_F)^{-1}$ according to Theorem 1.1 can be derived analogously to the infinite dimensional case; see the proof of Theorem 4.1.

In our finite dimensional setting this result can be obtained alternatively by observing that G_m corresponds to a generalized Jacobian in Clarke's sense combined with the convergence results for semismooth Newton methods in [Q1, QS]. In fact, from (2.4) we infer that $G_F(x^*)$ is a nonsingular generalized Jacobian, and Lemma 3.1 proves the semismoothness of F at x^*. Hence, Theorem 3.2 of [QS] yields the locally superlinear convergence property. For a discussion of the semismoothness concept in finite dimensions we refer the reader to [Q1, QS].

Furthermore, since (3.1) is strongly semismooth, by utilizing Theorem 3.2 of [QS] the convergence rate can even be improved. Indeed, the primal-dual active set strategy converges locally with a q-quadratic rate. For the definition of strong semismoothness we refer the reader to [FFKP].

We also observe that if the iterates $x^k = (y^k, \lambda^k)$ converge to $x^* = (y^*, \lambda^*)$, then they converge in finitely many steps. In fact, there are only finitely many choices of active/inactive sets and if the algorithm would determine the same sets twice, then this contradicts convergence of x^k to x^*. We refer to [FK] for a similar observation for a nonsmooth Newton method of the types discussed in [Q1, QS, K1], for example.

Let us address global convergence next. In the following two results sufficient conditions for convergence for arbitrary initial data $x^0 = (y^0, \lambda^0)$ are given. We recall that A is referred to as an M-matrix, if it is nonsingular, $(m_{ij}) \le 0$, for $i \ne j$, and $M^{-1} \ge 0$. Our notion of an M-matrix coincides with that of nonsingular M-matrices as defined in [BP].

THEOREM 3.2. *Assume that A is an* M*-matrix. Then $x^k \to x^*$ for arbitrary initial data. Moreover, $y^* \le y^{k+1} \le y^k$ for all $k \ge 1$ and $y^k \le \psi$ for all $k \ge 2$.*

For a proof of Theorem 3.2 we can utilize the proof of Theorem 1 in [H], where a (primal) active set algorithm is proposed and analyzed. However, we provide a proof in Appendix A since, in contrast to the algorithm in [H], the primal-dual active set strategy makes use of the dual variable λ and includes arbitrarily fixed $c > 0$. From the proof in Appendix A it can be seen that for unilaterally constrained problems c drops out after the first iteration. We point out that, provided the active and inactive sets coincide, the linear systems that have to be solved in every iteration of both algorithms coincide. In practice, however, λ and c play a significant role and make a distinct difference between the performance of the algorithm in [H] and the primal-dual active set strategy. In fact, the primal-dual active set strategy fixes $\lambda_i^{k+1} = 0$ for $i \in \mathcal{I}_k$. The decision whether an inactive index $i \in \mathcal{I}_k$ becomes an active one,

i.e., whether $i \in \mathcal{A}_{k+1}$, is based on

$$\lambda_i^{k+1} + c(y_i^{k+1} - \psi_i) > 0\,.$$

In contrast, the (primal) active set algorithm in [H] uses the criterion

$$f_i - (Ay^{k+1})_i + (y_i^{k+1} - \psi_i) > 0$$

instead. Clearly, if the linear system of both algorithms are solved approximately (e.g., by some iterative procedure) then the numerical behavior may differ.

Remark 3.2. Concerning the applicability of Theorem 3.2 we recall that many discretizations of second order differential operators give rise to M-matrices. ∎

For a rectangular matrix $B \in \mathbb{R}^{n\times m}$ we denote by $\|\cdot\|_1$ the subordinate matrix norm when both $\mathbb{R}^n$ and $\mathbb{R}^m$ are endowed with the one-norms. Moreover, B_+ denotes the $n \times m$-matrix containing the positive parts of the elements of B. The following result can be applied to discretizations of constrained optimal control problems. We refer to the end of section 4 for a discussion of the conditions of Theorem 3.3 in the case of control constrained optimal control problems.

THEOREM 3.3. *If A is a* P*-matrix and for every partitioning of the index set into disjoint subsets $\mathcal{I}$ and $\mathcal{A}$ we have $\|(A_{\mathcal{I}}^{-1}A_{\mathcal{I}\mathcal{A}})_+\|_1 < 1$ and $\sum_{i\in\mathcal{I}}(A_{\mathcal{I}}^{-1}y_{\mathcal{I}})_i \geq 0$ for $y_{\mathcal{I}} \geq 0$, then $\lim_{k\to\infty} x^k = x^*$.*

Proof. From (3.3) we have

$$(y^{k+1} - \psi)_{\mathcal{I}_k} = (y^k - \psi)_{\mathcal{I}_k} + A_{\mathcal{I}_k}^{-1}A_{\mathcal{I}_k\mathcal{A}_k}(y^k - \psi)_{\mathcal{A}_k} + A_{\mathcal{I}_k}^{-1}\lambda_{\mathcal{I}_k}^k$$

and upon summation over the inactive indices

$$\begin{aligned}\sum_{\mathcal{I}_k}(y_i^{k+1} - \psi_i) = \sum_{\mathcal{I}_k}(y_i^k - \psi_i) + \sum_{\mathcal{I}_k}\left(A_{\mathcal{I}_k}^{-1}A_{\mathcal{I}_k\mathcal{A}_k}(y^k - \psi)_{\mathcal{A}_k}\right)_i \\ + \sum_{\mathcal{I}_k}(A_{\mathcal{I}_k}^{-1}\lambda_{\mathcal{I}_k}^k)_i.\end{aligned} \tag{3.5}$$

Adding the obvious equality

$$\sum_{\mathcal{A}_k}(y_i^{k+1} - \psi_i) - \sum_{\mathcal{A}_k}(y_i^k - \psi_i) = -\sum_{\mathcal{A}_k}(y_i^k - \psi_i)$$

to (3.5) implies

$$\sum_{i=1}^n(y_i^{k+1} - y_i^k) \leq -\sum_{\mathcal{A}_k}(y_i^k - \psi_i) + \sum_{\mathcal{I}_k}(A_{\mathcal{I}_k}^{-1}A_{\mathcal{I}_k\mathcal{A}_k}(y^k - \psi)_{\mathcal{A}_k})_i\,. \tag{3.6}$$

Here we used the fact $\lambda_{\mathcal{I}_k}^k = -\delta\lambda_{\mathcal{I}_k} \leq 0$, established in the proof of Theorem 3.2. There it was also argued that $y_{\mathcal{A}_k}^k \geq \psi_{\mathcal{A}_k}$. Hence, it follows that

$$\sum_{i=1}^n(y_i^{k+1} - y_i^k) \leq -\|y^k - \psi\|_{1,\mathcal{A}_k} + \|(A_{\mathcal{I}_k}^{-1}A_{\mathcal{I}_k\mathcal{A}_k})_+\|_1\,\|y^k - \psi\|_{1,\mathcal{A}_k} < 0\,, \tag{3.7}$$

unless $y^{k+1} = y^k$. Consequently,

$$y^k \to \mathcal{M}(y^k) = \sum_{i=1}^n y_i^k$$

acts as a merit function for the algorithm. Since there are only finitely many possible choices for active/inactive sets there exists an iteration index $\bar{k}$ such that $\mathcal{I}_{\bar{k}} = \mathcal{I}_{\bar{k}+1}$. Moreover, $(y^{\bar{k}+1}, \lambda^{\bar{k}+1})$ is a solution to (3.1). In fact, in view of (iii) of the algorithm it suffices to show that $y^{\bar{k}+1}$ and $\lambda^{\bar{k}+1}$ are feasible. This follows from the fact that due to $\mathcal{I}_{\bar{k}} = \mathcal{I}_{\bar{k}+1}$ we have $c(y_i^{\bar{k}+1} - \psi_i) = \lambda_i^{\bar{k}+1} + c(y_i^{\bar{k}+1} - \psi_i) \leq 0$ for $i \in \mathcal{I}_{\bar{k}}$ and $\lambda_i^{\bar{k}+1} + c(y_i^{\bar{k}+1} - \psi_i) > 0$ for $i \in \mathcal{A}_{\bar{k}}$. Thus the algorithm converges in finitely many steps. ∎

Remark 3.3. Let us note as a corollary to the proof of Theorem 3.3 that in the case where A is an M-matrix then $\mathcal{M}(y^k) = \sum_{i=1}^n y_i^k$ is always a merit function. In fact, in this case the conditions of Theorem 3.3 are obviously satisfied. ∎

A perturbation result. We now discuss the primal-dual active set strategy for the case where the matrix A can be expressed as an additive perturbation of an M-matrix.

THEOREM 3.4. *Assume that $A = M + K$ with M an* M*-matrix and with K an $n \times n$-matrix. Then, if $\|K\|_1$ is sufficiently small,* (3.1) *admits a unique solution $x^* = (y^*, \lambda^*)$, the primal-dual active set algorithm is well-defined, and $\lim_{k\to\infty} x^k = x^*$.*

Proof. Recall that as a consequence of the assumption that M is an M-matrix all principal submatrices of M are nonsingular M-matrices as well [BP]. Let $\mathcal{S}$ denote the set of all subsets of $\{1, \ldots, n\}$, and define

$$\rho = \sup_{\mathcal{I}\in\mathcal{S}} \|M_{\mathcal{I}}^{-1} K_{\mathcal{I}}\|_1 .$$

Let K be chosen such that $\rho < \frac{1}{2}$. For every subset $\mathcal{I} \in \mathcal{S}$ the inverse of $A_{\mathcal{I}}$ exists and can be expressed as

$$A_{\mathcal{I}}^{-1} = \left(I_{\mathcal{I}} + \sum_{i=1}^{\infty} \left(-M_{\mathcal{I}}^{-1} K_{\mathcal{I}}\right)^i \right) M_{\mathcal{I}}^{-1} .$$

As a consequence the algorithm is well-defined. Proceeding as in the proof of Theorem 3.3 we arrive at

$$\begin{aligned} \sum_{i=1}^{n} (y_i^{k+1} - y_i^k) = &- \sum_{i\in\mathcal{A}} (y_i^k - \psi_i) + \sum_{i\in\mathcal{I}} \left(A_{\mathcal{I}}^{-1} A_{\mathcal{I}\mathcal{A}} (y^k - \psi)_{\mathcal{A}}\right)_i \\ &+ \sum_{i\in\mathcal{I}} (A_{\mathcal{I}}^{-1} \lambda_{\mathcal{I}}^k)_i , \end{aligned} \tag{3.8}$$

where $\lambda_i^k \leq 0$ for $i \in \mathcal{I}$ and $y_i^k \geq \psi_i$ for $i \in \mathcal{A}$. Here and below we drop the index k with $\mathcal{I}_k$ and $\mathcal{A}_k$. Setting $g = -A_{\mathcal{I}}^{-1} \lambda_{\mathcal{I}}^k \in \mathbb{R}^{|\mathcal{I}|}$ and since $\rho < \frac{1}{2}$ we find

$$\begin{aligned} \sum_{i\in\mathcal{I}} g_i &\geq \|M_{\mathcal{I}}^{-1} \lambda_{\mathcal{I}}^k\|_1 - \sum_{i=1}^{\infty} \|M_{\mathcal{I}}^{-1} K_{\mathcal{I}}\|_1^i \|M_{\mathcal{I}}^{-1} \lambda_{\mathcal{I}}^k\|_1 \\ &\geq \frac{1 - 2\rho}{1 - \rho} \|M^{-1} \lambda_{\mathcal{I}}^k\|_1 \geq 0 , \end{aligned}$$

and consequently by (3.8)

$$\sum_{i=1}^{n} (y_i^{k+1} - y_i^k) \leq - \sum_{i\in\mathcal{A}} (y_i^k - \psi_i) + \sum_{i\in\mathcal{I}} (A_{\mathcal{I}}^{-1} A_{\mathcal{I}\mathcal{A}} (y^k - \psi)_{\mathcal{A}})_i .$$

Note that $A_{\mathcal{I}}^{-1} A_{\mathcal{I}\mathcal{A}} \leq M_{\mathcal{I}}^{-1} K_{\mathcal{I}\mathcal{A}} - M_{\mathcal{I}}^{-1} K_{\mathcal{I}} (M + K)_{\mathcal{I}}^{-1} A_{\mathcal{I}\mathcal{A}}$. Here we have used $(M + K)_{\mathcal{I}}^{-1} - M_{\mathcal{I}}^{-1} = -M_{\mathcal{I}}^{-1} K_{\mathcal{I}} (M + K)_{\mathcal{I}}^{-1}$ and $M_{\mathcal{I}}^{-1} M_{\mathcal{I}\mathcal{A}} \leq 0$. Since $y^k \geq \psi$ on $\mathcal{A}$,

it follows that $\|K\|_1$ can be chosen sufficiently small such that $\sum_{i=1}^n (y_i^{k+1} - y_i^k) < 0$ unless $y^{k+1} = y^k$, and hence

$$y^k \mapsto \mathcal{M}(y^k) = \sum_{i=1}^n y_i^k$$

is a merit function for the algorithm. The proof is now completed in the same manner as that of Theorem 3.3. ∎

The assumptions of Theorem 3.4 do not require A to be a P-matrix. From its conclusions existence of a solution to (3.1) for arbitrary f follows. This is equivalent to the fact that A is a P-matrix [BP, Thm. 10.2.15]. Hence, it follows that Theorem 3.4 represents a sufficient condition for A to be a P-matrix.

Observe further that the M-matrix property is not stable under arbitrarily small perturbations since off-diagonal elements may become positive. This implies certain limitations of the applicability of Theorem 3.2. Theorem 3.4 guarantees that convergence of the primal-dual active set strategy for arbitrary initial data is preserved for sufficiently small perturbations K of an M-matrix. Therefore, Theorem 3.4 is also of interest in connection with numerical implementations of the primal-dual active set algorithm.

Remark 3.4. The primal-dual active set strategy can be interpreted as a prediction strategy which, on the basis of (y^k, λ^k), predicts the *true* active and inactive sets, i.e.,

$$\mathcal{A}^* = \{i : \lambda_i^* + c(y_i^* - \psi_i) > 0\} \quad \text{and} \quad \mathcal{I}^* = \{1, \dots, n\} \setminus \mathcal{A}^* .$$

To further pursue this point we define the following partitioning of the index set at iteration level k:

$$\mathcal{I}_G = \mathcal{I}_k \cap \mathcal{I}^*, \quad \mathcal{I}_B = \mathcal{I}_k \cap \mathcal{A}^*, \quad \mathcal{A}_G = \mathcal{A}_k \cap \mathcal{A}^*, \quad \mathcal{A}_B = \mathcal{A}_k \cap \mathcal{I}^* .$$

The sets $\mathcal{I}_G$, $\mathcal{A}_G$ give a *good* prediction, the sets $\mathcal{I}_B$ and $\mathcal{A}_B$ a *bad* prediction. Let us denote by $G_F(x^k)$ the system matrix of (2.4) and let $\Delta y = y^{k+1} - y^*$, $\Delta\lambda = \lambda^{k+1} - \lambda^*$. If the primal-dual active set method is interpreted as a semismooth Newton method, then the convergence analysis is based on the identity

$$G_F(x^k) \begin{pmatrix} \Delta y_{\mathcal{I}_k} \\ \Delta y_{\mathcal{A}_k} \\ \Delta\lambda_{\mathcal{I}_k} \\ \Delta\lambda_{\mathcal{A}_k} \end{pmatrix} = -\big(F(x^k) - F(x^*) - G_F(x^k)(x^k - x^*)\big) =: \Psi(x^k) . \tag{3.9}$$

Without loss of generality we can assume that the components of the equation $\lambda - \max\{0, \lambda + c(y - \psi)\} = 0$ are ordered as $(\mathcal{I}_G, \mathcal{I}_B, \mathcal{A}_G, \mathcal{A}_B)$. Then the right-hand side of (3.9) has the form

$$\Psi(x^k) = -\mathrm{col}\left(0_{\mathcal{I}_k}, 0_{\mathcal{A}_k}, 0_{\mathcal{I}_G}, \lambda^*_{\mathcal{I}_B}, 0_{\mathcal{A}_G}, c(\psi - y^*)_{\mathcal{A}_B}\right) , \tag{3.10}$$

where $0_{\mathcal{I}_k}$ denotes a vector of zeros of length $|\mathcal{I}_k|$, $\lambda^*_{\mathcal{I}_B}$ denotes a vector of λ^* coordinates with index set $\mathcal{I}_B$, and analogously for the remaining terms. Since $y^k \geq \psi$ on $\mathcal{A}_k$ and $\lambda^k \leq 0$ on $\mathcal{I}_k$ we have

$$\|\psi - y^*\|_{\mathcal{A}_B} \leq \|y^k - y^*\|_{\mathcal{A}_B} \quad \text{and} \quad \|\lambda^*\|_{\mathcal{I}_B} \leq \|\lambda^k - \lambda^*\|_{\mathcal{I}_B} . \tag{3.11}$$

Exploiting the structure of $G_F(x^k)$ and (3.10) we find

$$\Delta y_{\mathcal{A}_G} = 0, \quad \Delta y_{\mathcal{A}_B} = (\psi - y^*)_{\mathcal{A}_B}, \quad \Delta\lambda_{\mathcal{I}_G} = 0, \quad \Delta\lambda_{\mathcal{I}_B} = -\lambda^*_{\mathcal{I}_B}\,. \tag{3.12}$$

On the basis of (3.9)–(3.12) we can draw the following conclusions:

(i) If $x^k \to x^*$, then there exists an index $\bar{k}$ such that $\mathcal{I}_B = \mathcal{A}_B = \emptyset$ for all $k \geq \bar{k}$. Consequently, $\Psi(x^{\bar{k}}) = 0$ and, as we noted before, if $x^k \to x^*$, then convergence occurs in finitely many steps.

(ii) By (3.9)–(3.11) there exists a constant $\kappa \geq 1$ independent of k such that

$$\|\Delta y\| + \|\Delta\lambda\| \leq \kappa \left(\|(y^k - y^*)_{\mathcal{A}_B}\| + \|(\lambda^k - \lambda^*)_{\mathcal{I}_B}\|\right)\,.$$

Thus if the incorrectly predicted sets are small in the sense that

$$\|(y^k - y^*)_{\mathcal{A}_B}\| + \|(\lambda^k - \lambda^*)_{\mathcal{I}_B}\| \leq \tfrac{1}{2\kappa - 1}\Big(\|(y^k - y^*)_{\mathcal{A}_{B,c}}\| + \|(\lambda^k - \lambda^*)_{\mathcal{I}_{B,c}}\|\Big)\,,$$

where $\mathcal{A}_{B,c}$ ($\mathcal{I}_{B,c}$) denotes the complement of the indices $\mathcal{A}_B$ ($\mathcal{I}_B$), then

$$\|y^{k+1} - y^*\| + \|\lambda^{k+1} - \lambda^*\| \leq \tfrac{1}{2}\left(\|y^k - y^*\| + \|\lambda^k - \lambda^*\|\right)\,,$$

and convergence follows.

(iii) If $y^* < \psi$ and $\lambda^0 + c(y^0 - \psi) \leq 0$ (e.g., $y^0 = \psi$, $\lambda^0 = 0$), then the algorithm converges in one step. In fact, in this case $\mathcal{A}_B = \mathcal{I}_B = \emptyset$ and $\Psi(x^0) = 0$. ∎

Finally, we shall point out that Theorems 3.2–3.4 establish global convergence of the primal-dual active set strategy or, equivalently, semismooth Newton method without the necessity of a line search. The rate of convergence is locally superlinear. Moreover, it can be observed from (2.4) that if $\mathcal{I}_k = \mathcal{I}_{k'}$ for $k \neq k'$, then $y^k = y^{k'}$ and $\lambda^k = \lambda^{k'}$. Hence, in case of convergence no cycling of the algorithm is possible, and termination at the solution of (2.1) occurs after finitely many steps.

4. The infinite dimensional case. In this section we first analyze the notion of slant differentiability of the max-operation between various function spaces. Then we turn to the investigation of convergence of semismooth Newton methods applied to (P). We close the section with a numerical example for superlinear convergence.

Let X denote a space of functions defined over a bounded domain or manifold $\Omega \subset \mathbb{R}^n$ with Lipschitzian boundary $\partial\Omega$, and let $\max(0, y)$ stand for the pointwise maximum operation between 0 and $y \in X$. Let $\delta \in \mathbb{R}$ be fixed arbitrarily. We introduce candidates for slanting functions G_m of the form

$$G_m(y)(x) = \begin{cases} 1 & \text{if} \quad y(x) > 0\,, \\ 0 & \text{if} \quad y(x) < 0\,, \\ \delta & \text{if} \quad y(x) = 0\,, \end{cases} \tag{4.1}$$

where $y \in X$.

PROPOSITION 4.1.

(i) *G_m can in general not serve as a slanting function for* $\max(0,\cdot)\colon\ L^p(\Omega) \to L^p(\Omega)$ *for* $1 \leq p \leq \infty$.

(ii) *The mapping* $\max(0,\cdot)\colon L^q(\Omega) \to L^p(\Omega)$ *with* $1 \leq p < q \leq \infty$ *is slantly differentiable on* $L^q(\Omega)$ *and G_m is a slanting function.*

The proof is deferred to Appendix A.

We refer to [U] for a related investigation of the *two-norm problem* involved in Proposition 4.1 in the case of superposition operators. An example in [U] proves the necessity of the norm gap for the case in which the complementarity condition is expressed by means of the Fischer–Burmeister functional.

We now turn to (P) posed in $L^2(\Omega)$. For convenience we repeat the problem formulation

$$\text{(P)} \qquad \begin{cases} \min J(y) = \dfrac{1}{2}(y, Ay) - (f, y) \\ \text{subject to } y \le \psi, \end{cases}$$

where $(\cdot,\cdot)$ now denotes the inner product in $L^2(\Omega)$, f, and $\psi \in L^2(\Omega)$, $A \in \mathcal{L}(L^2(\Omega))$ is self-adjoint, and

$$\text{(H1)} \qquad (Ay, y) \ge \gamma \|y\|^2$$

for some $\gamma > 0$ independent of $y \in L^2(\Omega)$. There exists a unique solution y^* to (P) and a Lagrange multiplier $\lambda^* \in L^2(\Omega)$ such that (y^*, λ^*) is the unique solution to

$$\text{(4.2)} \qquad \begin{cases} Ay^* + \lambda^* = f, \\ \mathcal{C}(y^*, \lambda^*) = 0, \end{cases}$$

where $\mathcal{C}(y, \lambda) = \lambda - \max(0, \lambda + c(y - \psi))$, with the max-operation defined pointwise a.e. and $c > 0$ fixed. The primal-dual active set strategy is analogous to the finite dimensional case. We repeat it for convenient reference.

PRIMAL-DUAL ACTIVE SET ALGORITHM IN $L^2(\Omega)$.

(i) Choose y^0, λ^0 in $L^2(\Omega)$. Set $k = 0$.

(ii) Set $\mathcal{A}_k = \{x\colon \lambda^k(x) + c(y^k(x) - \psi(x)) > 0\}$ and $\mathcal{I}_k = \Omega \backslash \mathcal{A}_k$.

(iii) Solve

$$\begin{aligned} &Ay^{k+1} + \lambda^{k+1} = f, \\ &y^{k+1} = \psi \ \text{ on } \ \mathcal{A}_k, \lambda^{k+1} = 0 \ \text{ on } \ \mathcal{I}_k. \end{aligned}$$

(iv) Stop, or set $k = k + 1$ and return to (ii).

Under our assumptions on A, f, and ψ it is simple to argue the solvability of the system in step (iii) of the above algorithm.

For the semismooth Newton step as well we can refer back to section 2. At iteration level k with $(y^k, \lambda^k) \in L^2(\Omega) \times L^2(\Omega)$ given, it is of the form (2.4) where now $\delta y_{\mathcal{I}_k}$ denotes the restriction of δy (defined on Ω) to $\mathcal{I}_k$ and analogously for the remaining terms. Moreover, $A_{\mathcal{I}_k\mathcal{A}_k} = E^*_{\mathcal{I}_k} A\, E_{\mathcal{A}_k}$, where $E_{\mathcal{A}_k}$ denotes the extension-by-zero operator for $L^2(\mathcal{A}_k)$ to $L^2(\Omega)$-functions, and its adjoint $E^*_{\mathcal{A}_k}$ is the restriction of $L^2(\Omega)$-functions to $L^2(\mathcal{A}_k)$, and similarly for $E_{\mathcal{I}_k}$ and $E^*_{\mathcal{I}_k}$. Moreover, $A_{\mathcal{A}_k\mathcal{I}_k} = E^*_{\mathcal{A}_k} A\, E_{\mathcal{I}_k}$, $A_{\mathcal{I}_k} = E^*_{\mathcal{I}_k} A\, E_{\mathcal{I}_k}$, and $A_{\mathcal{A}_k} = E^*_{\mathcal{A}_k} A\, E_{\mathcal{A}_k}$. It can be argued precisely as in section 2 that the primal-dual active set strategy and the semismooth Newton updates coincide, provided that the slanting function of the max-function is taken according to

$$\text{(4.3)} \qquad G_m(u)(x) = \begin{cases} 1 & \text{if} \quad u(x) > 0, \\ 0 & \text{if} \quad u(x) \le 0, \end{cases}$$

which we henceforth assume.

Proposition 4.1 together with Theorem 1.1 suggest that the semismooth Newton algorithm applied to (4.2) may not converge in general. We therefore restrict our attention to operators A of the form

$$\text{(H2)} \qquad A = C + \beta I, \quad \text{with } C \in \mathcal{L}(L^2(\Omega), L^q(\Omega)), \quad \text{where } \beta > 0,\ q > 2.$$

We show next that a large class of optimal control problems with control constraints can be expressed in the form (P) with (H2) satisfied.

Example 1. We consider the optimal control problem

$$\text{(4.4)} \qquad \begin{cases} \text{minimize} & \frac{1}{2}\|y - z\|^2_{L^2} + \frac{\beta}{2}\|u\|^2_{L^2} \\ \text{subject to} & -\Delta y = u \text{ in } \Omega,\ y = 0 \text{ on } \partial\Omega\,, \\ & u \le \psi,\ u \in L^2(\Omega)\,, \end{cases}$$

where $z \in L^2(\Omega)$, $\psi \in L^q(\Omega)$, and $\beta > 0$. Let $B \in \mathcal{L}(H^1_o(\Omega), H^{-1}(\Omega))$ denote the operator $-\Delta$ with homogeneous Dirichlet boundary conditions. Then (4.4) can equivalently be expressed as

$$\text{(4.5)} \qquad \begin{cases} \text{minimize} & \frac{1}{2}\|B^{-1}u - z\|^2_{L^2} + \frac{\beta}{2}\|u\|^2_{L^2} \\ \text{subject to} & u \le \psi,\ u \in L^2(\Omega)\,. \end{cases}$$

In this case $A \in \mathcal{L}(L^2(\Omega))$ turns out to be $Au = B^{-1}\mathcal{J}B^{-1}u + \beta u$, where $\mathcal{J}$ is the embedding of $H^1_o(\Omega)$ into $H^{-1}(\Omega)$, and $f = B^{-1}z$. Condition (H2) is obviously satisfied.

In (4.4) we considered the distributed control case. A related boundary control problem is given by

$$\text{(4.6)} \qquad \begin{cases} \text{minimize} & \frac{1}{2}\|y - z\|^2_{L^2(\Omega)} + \frac{\beta}{2}\|u\|^2_{L^2(\partial\Omega)} \\ \text{subject to} & -\Delta y + y = 0 \text{ in } \Omega,\ \frac{\partial y}{\partial n} = u \text{ on } \partial\Omega\,, \\ & u \le \psi,\ u \in L^2(\partial\Omega)\,, \end{cases}$$

where n denotes the unit outer normal to Ω along $\partial\Omega$. This problem is again a special case of (P) with $A \in \mathcal{L}(L^2(\partial\Omega))$ given by $Au = B^{-*}\mathcal{J}B^{-1}u + \beta u$, where $B^{-1} \in \mathcal{L}(H^{-1/2}(\Omega), H^1(\Omega))$ denotes the solution operator to

$$-\Delta y + y = 0 \text{ in } \Omega,\ \frac{\partial y}{\partial n} = u \text{ on } \partial\Omega\,,$$

and $f = B^{-*}z$. Moreover, $C = B^{-*}\mathcal{J}B^{-1}_{|L^2(\Omega)} \in \mathcal{L}(L^2(\partial\Omega), H^{1/2}(\partial\Omega))$ with $\mathcal{J}$ the embedding of $H^{1/2}(\Omega)$ into $H^{-1/2}(\partial\Omega)$, and hence (H2) is satisfied as a consequence of the Sobolev embedding theorem.

For the sake of illustration it is also worthwhile to specify (2.5)–(2.8), which were found to be equivalent to the Newton update (2.4) for the case of optimal control problems. We restrict ourselves to the case of the distributed control problem (4.4). Then (2.5)–(2.8) can be expressed as

$$\text{(4.7)} \qquad \begin{cases} \lambda^{k+1}_{\mathcal{I}_k} = 0, \quad u^{k+1}_{\mathcal{A}_k} = \psi_{\mathcal{A}_k}, \\ E^*_{\mathcal{I}_k}\left[(B^{-2} + \beta I)E_{\mathcal{I}_k}u^{k+1}_{\mathcal{I}_k} - B^{-1}z + (B^{-2} + \beta I)E_{\mathcal{A}_k}\psi_{\mathcal{A}_k}\right] = 0\,, \\ E^*_{\mathcal{A}_k}\left[\lambda^{k+1} + B^{-2}u^{k+1} + \beta u^{k+1} - B^{-1}z\right] = 0\,, \end{cases}$$

where we set $B^{-2} = B^{-1}\mathcal{J}B^{-1}$. Setting $p^{k+1} = B^{-1}z - B^{-2}u^{k+1}$, a short computation shows that (4.7) is equivalent to

$$
\begin{cases} -\Delta y^{k+1} = u^{k+1} \quad \text{in } \Omega\,, \quad y^{k+1} = 0 \quad \text{on } \partial\Omega\,, \\ -\Delta p^{k+1} = z - y^{k+1} \quad \text{in } \Omega\,, \quad p^{k+1} = 0 \quad \text{on } \partial\Omega\,, \\ \quad p^{k+1} = \beta u^{k+1} + \lambda^{k+1} \quad \text{in } \Omega\,, \\ \quad u^{k+1} = \psi \quad \text{in } \mathcal{A}_k\,, \lambda^{k+1} = 0 \quad \text{in } \mathcal{I}_k\,. \end{cases} \tag{4.8}
$$

This is the system in the primal variables (y, u) and adjoint variables (p, λ), previously implemented in [BHHK, BIK] for testing the algorithm. □

At this point we remark that the primal-dual active set strategy has no straightforward infinite dimensional analogue for state constrained optimal control problems and obstacle problems [H]. For state constrained optimal control problems the Lagrange multiplier is only a measure in general, and hence the core steps (ii) and (iii) of our algorithm are no longer meaningful. For details on the regularity issue we refer the reader to [Ca]. Theorem 3.2 proves global convergence of the primal-dual active set strategy or, equivalently, semismooth Newton method for discretized obstacle problems. However, no comparable result can be expected in infinite dimensions. The main reason comes from the fact that the systems that would have to be solved in step (iii) are the first order conditions related to the problems

$$
\min \quad \tfrac{1}{2}(Ay, y)_{L^2(\Omega)} - (f, y)_{L^2(\Omega)} \quad \text{such that} \quad y = \psi \quad \text{a.e. on } \mathcal{A}_k\,.
$$

Again the multiplier associated with the equality constraint is only a measure in general.

Our main intention is to consider control constrained problems as in Example 1. To prove convergence under assumptions (H1), (H2) we utilize a reduced algorithm which we explain next.

The operators $E_{\mathcal{I}}$ and $E_{\mathcal{A}}$ denote the extension by zero, and their adjoints are restrictions to $\mathcal{I}$ and $\mathcal{A}$, respectively. The optimality system (4.2) does not depend on the choice of $c > 0$. Moreover, from the discussion in section 2 the primal-dual active set strategy is independent of $c > 0$ after the initialization phase. For the specific choice $c = \beta$ system (4.2) can equivalently be expressed as

$$
\beta y^* - \beta\psi + \max(0, Cy^* - f + \beta\psi) = 0\,, \tag{4.9}
$$

$$
\lambda^* = f - Cy^* - \beta y^*\,. \tag{4.10}
$$

We shall argue in the proof of Theorem 4.1 that the primal-dual active set method in $L^2(\Omega)$ for (y, λ) is equivalent to the following algorithm for the reduced system (4.9)–(4.10), which will be shown to converge superlinearly.

REDUCED ALGORITHM.

(i) Choose $y^0 \in L^2(\Omega)$ and set $k = 0$.

(ii) Set $\mathcal{A}_k = \{x : (f - Cy_k - \beta\psi)(x) > 0\}$, $\mathcal{I}_k = \Omega \setminus \mathcal{A}_k$.

(iii) Solve

$$
\beta y_{\mathcal{I}_k} + (C(E_{\mathcal{I}_k} y_{\mathcal{I}_k} + E_{\mathcal{A}_k}\psi_{\mathcal{A}_k}))_{\mathcal{I}_k} = f_{\mathcal{I}_k}
$$

and set $y^{k+1} = E_{\mathcal{I}_k} y_{\mathcal{I}_k} + E_{\mathcal{A}_k}\psi_{\mathcal{A}_k}$.

(iv) Stop, or set $k = k + 1$ and return to (ii).

THEOREM 4.1. *Assume that* (H1), (H2) *hold and that* ψ *and* f *are in* $L^q(\Omega)$. *Then the primal-dual active set strategy or, equivalently, the semismooth Newton method converge superlinearly if* $\|y^0 - y^*\|$ *is sufficiently small and* $\lambda^0 = \beta(y^0 - \psi)$.

The proof is given in Appendix A. It consists essentially of two steps. In the first equivalence between the reduced algorithm and the original one is established, and in the second one slant differentiability of the mapping $\hat{F} : L^2(\Omega) \to L^2(\Omega)$ given by $\hat{F}(y) = \max(0, Cy - f + \beta\psi)$ is shown. With respect to the latter we can alternatively utilize the theory of semismoothness of composite mappings as developed in [U]. For this purpose we first recall the notion of semismoothness as introduced in [U]. Suppose we are given the superposition operator

$$\tilde{\Psi} : Y \to L^r(\Omega), \quad \tilde{\Psi}(y)(x) = \tilde{\psi}(H(y)(x)),$$

where $\tilde{\psi} : \mathbb{R}^m \to \mathbb{R}$ and $H : Y \to \prod_{i=1}^m L^{r_i}(\Omega)$, with $1 \le r \le r_i < \infty$, and Y is a Banach space. Then $\tilde{\Psi}$ is called semismooth at $y \in Y$ if

$$\sup_{G \in \partial_s \tilde{\Psi}(y+h)} \|\tilde{\Psi}(y+h) - \tilde{\Psi}(y) - Gh\|_{L^r} = o(\|h\|_Y) \quad \text{as } h \to 0 \text{ in } Y. \tag{4.11}$$

Here $\partial_s \tilde{\Psi}$ denotes the generalized differential

$$\partial_s \tilde{\Psi}(y) = \left\{ G \in \mathcal{L}(Y, L^r) \,\middle|\, \begin{array}{l} G : v \mapsto \sum_i d_i(y)(H_i'(y)v), \text{ where } d(y) \\ \text{is a measurable selection of } \partial\tilde{\psi}(H(y)) \end{array} \right\}, \tag{4.12}$$

where $\partial\tilde{\psi}$ is Clarke's generalized Jacobian [C], and prime denotes the Fréchet derivative. In our context $Y = L^r(\Omega) = L^2(\Omega)$, $m = 2$, $r_i = 2$, $H(y) = (0, Cy - f + \beta\psi)$, and $\tilde{\psi}(a, b) = \max(a, b)$. Clearly, H is affine with respect to the second component. By (H2), and since $\psi \in L^q(\Omega)$, $f \in L^q(\Omega)$, it follows that H is Lipschitz from $L^2(\Omega)$ to $(L^q(\Omega))^2$, with $q > 2$. Moreover, $\tilde{\psi}$ is semismooth in the sense of [QS]. Consequently, $\tilde{\Psi}$ is semismooth in the sense of (4.11) by [U, Thm. 5.2].

In general, a slanting function G according to Definition 1 need not satisfy $G(y) \in \partial_s \tilde{\Psi}(y)$. However, the particular slanting function

$$\hat{G}(y)v = G_m(Cy - f + \beta\psi)Cv$$

with

$$G_m(u)(x) = \begin{cases} 1 & \text{if } u(x) \ge 0, \\ 0 & \text{if } u(x) < 0 \end{cases}$$

satisfies $\hat{G}(y) \in \partial_s \tilde{\Psi}(y)$. In fact, $d(y) = (d_1(y), d_2(y)) = (0, G_m(Cy - f + \beta\psi))$ is a measurable selection of $\partial \max(0, Cy - f + \beta\psi)$. Thus, (4.12) yields

$$\partial_s \tilde{\Psi}(y)v \ni G(y)v = \sum_i d_i(y)(H_i'(y)v) = G_m(Cy - f + \beta\psi)Cv = \hat{G}(y)v.$$

Consequently, from the proof of Theorem 6.4 in [U] we infer that the reduced algorithm converges locally superlinearly.

Let us point out that the semismooth Newton method in [U] requires a smoothing step while our primal-dual active set strategy does not. To explain the difference of the two approaches, we note that with respect to (P) the following NCP problem is considered in [U]: Find $y \in Y$ such that

$$y - \psi \le 0, \quad Z(y) := Ay - f \ge 0, \quad (y - \psi)Z(y) = 0. \tag{4.13}$$

Then (4.13) is reformulated by utilizing an NCP function. In our context, this yields

$$\max(y-\psi, f-Ay)=0. \tag{4.14}$$

Following [U] one chooses $Y=L^p(\Omega)$, $p>2$, and considers $y\mapsto \max(y-\psi, f-Ay)$ from $L^p(\Omega)$ to $L^2(\Omega)$ in order to introduce the norm gap which is required for semismoothness according to (4.11). In Algorithm 6.3 of [U] the Newton step first produces an update in $L^2(\Omega)$, which requires smoothing to obtain the new iterate in $L^p(\Omega)$ which is utilized in (4.14). In our formulation, (4.13) is reformulated as (4.9) rather than (4.14). Here we can take advantage of the fact that (4.9) allows us to directly exploit the smoothing property of the operator C. Consequently, we obtain a superlinearly convergent Newton method without the necessity of a smoothing step.

If an appropriate growth condition is satisfied, then the superlinear convergence result of Theorem 4.1 can be improved to superlinear convergence with a specific rate. Let us suppose that there exists $\alpha>0$ such that

$$\lim_{h\to 0}\frac{1}{\|h\|^{1+\alpha}}\|F(x^*+h)-F(x^*)-G(x^*+h)h\|=0\,. \tag{A'}$$

Then an inspection of the proof of Theorem 1.1 shows that the rate of convergence of x^k to x^* is of q-order $1+\alpha$; i.e., we have $\|x^{k+1}-x^*\|=\mathcal{O}(\|x^k-x^*\|^{1+\alpha})$ as $k\to\infty$. To investigate (A') for the specific F appearing in the proof of Theorem 4.1 one can apply the general theory in [U]. We prefer to give an independent proof adapted to our problem formulation. Let the assumptions of Theorem 4.1 hold and recall that $F: L^2(\Omega)\to L^2(\Omega)$ is given by $F(y)=\beta y-\beta\psi+\max(0, Cy-f+\beta\psi)$. First we consider the case $2<q<+\infty$. The relevant difference quotient for the nonlinear term which must be analyzed for (A') to hold is given by

$$\begin{aligned}&\frac{1}{\|h\|_{L^2}^{1+\alpha}}\|\max(0,C(y^*+h)-f+\beta\psi)-\max(0,Cy^*-f+\beta\psi)\\ &\qquad -G_m(Cy^*+Ch-f+\beta\psi)(Ch)\|_{L^2}\\ =&\frac{1}{\|Ch\|_{L^q}^{1+\alpha}}\|\max(0,w+Ch)-\max(0,w)-G_m(w+Ch)(Ch)\|_{L^2}\frac{\|Ch\|_{L^q}^{1+\alpha}}{\|h\|_{L^2}^{1+\alpha}}\,,\end{aligned}$$

where we set $w=Cy^*-f+\beta\psi$. Utilizing the fact that $C\in\mathcal{L}(L^2(\Omega),L^q(\Omega))$ it suffices to consider

$$\frac{1}{\|h\|_{L^q}^{1+\alpha}}\|D_{w,h}\|_{L^2}=\frac{1}{\|h\|_{L^q}^{1+\alpha}}\|\max(0,w+h)-\max(0,w)-G_m(w+h)h\|_{L^2}\,.$$

Here and below we use the notation introduced in the proof of Proposition 4.1(ii). Proceeding as in the proof of Proposition 4.1(ii) we find for $\frac{1}{\sigma}+\frac{1}{\tau}=1$, $\sigma\in(1,\infty)$,

$$\begin{aligned}\frac{1}{\|h\|_{L^q}^{1+\alpha}}\|D_{w,h}\|_{L^2}\le\frac{1+|\delta|}{\|h\|_{L^q}^{1+\alpha}}&\left[|\Omega_\epsilon(h)|^{1/2\tau}\left(\int_{\Omega_\epsilon(h)}|w(x)|^{2\sigma}dx\right)^{1/2\sigma}\right.\\ &\left.+|\Omega_\epsilon(w)|^{1/2\tau}\left(\int_{\Omega_0(h)\setminus\Omega_\epsilon(h)}|w(x)|^{2\sigma}dx\right)^{1/2\sigma}\right]\\ =\frac{1+|\delta|}{\|h\|_{L^q}^{1+\alpha}}&(|\Omega_\epsilon(h)|^{1/2\tau}+|\Omega_\epsilon(w)|^{1/2\tau})\left(\int_{\Omega_0(h)}|w(x)|^{2\sigma}dx\right)^{1/2\sigma}.\end{aligned} \tag{4.15}$$

Let us set $r = \frac{q}{1+\alpha}$. We have

$$\left(\int_{\Omega_0(h)} |w(x)|^{2\sigma} dx\right)^{1/2\sigma} \leq \left(\int_{\Omega_0(h)} |w(x)|^{\frac{2\sigma q}{r}} |w(x)|^{\frac{2\sigma(r-q)}{r}} dx\right)^{1/2\sigma}$$
$$\leq \left(\int_{\Omega_0(h)} |w(x)|^{\frac{2\sigma q}{r}\frac{r}{2\sigma}}\right)^{1/r} \left(\int_{\Omega_0(h)} |w(x)|^{\frac{2\sigma(r-q)}{r}\frac{r}{r-2\sigma}}\right)^{(r-2\sigma)/2r\sigma}$$
$$= \left(\int_{\Omega_0(h)} |w(x)|^q dx\right)^{1/r} \left(\int_{\Omega_0(h)} \frac{1}{|w(x)|^{\frac{2\sigma(q-r)}{r-2\sigma}}} dx\right)^{(r-2\sigma)/2r\sigma},$$

where it is assumed that $r = \frac{q}{1+\alpha} > 2\sigma > 2$. Since $|w(x)| \leq |h(x)|$ for $x \in \Omega_0(h)$ we find

$$\frac{1}{\|h\|_{L^q}^{1+\alpha}} \|D_{w,h}\|_{L^2}$$
$$\leq (1+|\delta|)(|\Omega_\epsilon(h)|^{1/2\tau} + |\Omega_\epsilon(w)|^{1/2\tau}) \left(\int_{\Omega_0(h)} \frac{1}{|w(x)|^{\frac{2\sigma(q-r)}{r-2\sigma}}} dx\right)^{(r-2\sigma)/2r\sigma}.$$

Suppose that

$$\int_{\{x:|w(x)|\neq 0\}} \frac{1}{|w(x)|^{\frac{2\sigma(q-r)}{r-2\sigma}}} dx < +\infty. \tag{4.16}$$

Then, following the argument in the proof of Proposition 4.1(ii), we have

$$\lim_{\|h\|_{L^q}\to 0} \frac{1}{\|h\|_{L^q}^{1+\alpha}} \|D_{w,h}\|_{L^2} = 0,$$

and hence (A') holds. Let us interpret the conditions on α and q. As already pointed out we must have $q > 2(1+\alpha)$ which for $\alpha = 0$ is consistent with the requirement that there must be a norm gap. The exponent in (4.16) can equivalently be expressed as $Q(\alpha, q) = \frac{2\sigma\alpha q}{q-2\sigma(1+\alpha)}$. Hence, for fixed q, the quotient $Q(\alpha, q)$ is increasing with α and (4.16) is more likely to be satisfied for small rather than for large α. Similarly, for fixed α, $Q(\alpha, q)$ is decreasing with respect to q ($> 2\sigma(1+\alpha)$), and hence (4.16) has a higher chance to be satisfied for large rather than small q.

Convergence of q-order larger than 2 is possible if $q > 2$ and (4.16) holds for the associated values of q and α. If w is Lipschitzian, then it must be of at most linear growth across the boundary of the set $\{x : w(x) \neq 0\}$. For this reason it is of interest to consider the range of α-values satisfying $\frac{2\alpha q}{q-2(1+\alpha)} < 1$. This necessitates $\alpha < \frac{1}{2}$.

In the case $q = +\infty$ we have for every $\sigma > 1$

$$\left(\int_{\Omega_0(h)} |w(x)|^{2\sigma} dx\right)^{1/2\sigma} = \left(\int_{\Omega_0(h)} |w(x)|^{2\sigma(1+\alpha)} |w(x)|^{-2\sigma\alpha} dx\right)^{1/2\sigma}$$
$$\leq \|h\|_{L^\infty}^{1+\alpha} \left(\int_{\Omega_0(h)} |w(x)|^{-2\sigma\alpha} dx\right)^{1/2\sigma}.$$

This estimate and (4.15) for $q = +\infty$ yield

$$\frac{1}{\|h\|_{L^\infty}^{1+\alpha}}\|D_{w,h}\|_{L^2} \leq (1+|\delta|)(|\Omega_\epsilon(h)|^{1/2\tau} + |\Omega_\epsilon(w)|^{1/2\tau}) \cdot \left(\int_{\Omega_0(h)} \frac{1}{|w(x)|^{2\sigma\alpha}} dx\right)^{1/2\sigma}.$$

Now suppose that for some $\sigma > 1$,

$$\int_{\{x:|w(x)|\neq 0\}} \frac{1}{|w(x)|^{2\sigma\alpha}} dx < +\infty\,. \tag{4.17}$$

Then, again following the arguments in the proof of Proposition 4.1, we obtain

$$\lim_{\|h\|_{L^\infty}\to 0} \frac{1}{\|h\|_{L^\infty}^{1+\alpha}}\|D_{w,h}\|_{L^2} = 0\,,$$

which shows that (A') is satisfied.

Example 1 (continued). As already observed Theorem 4.1 is directly applicable to problems (4.4) and (4.6) and confirms local superlinear convergence of the semismooth Newton algorithm.

Convergence for (4.4) was already analyzed in [BIK] where it was proved that a modified augmented Lagrangian acts as a merit function, provided that

$$\beta + \gamma \leq c \leq \beta - \frac{\beta^2}{\gamma} + \frac{\beta^2}{\|\Delta^{-1}\|^2} \tag{4.18}$$

for some $\gamma > 0$. Here $\|\Delta^{-1}\|$ denotes the operator norm of Δ^{-1} in $\mathcal{L}(L^2(\Omega))$. This previous convergence result is unconditional with respect to the initial condition, but it restricts the range of β. Theorem 4.1 is a local result with respect to initialization but does not restrict the range of $\beta > 0$. Further, the discussion following Theorem 4.1 provides rate of convergence results.

Let us also comment on the discretized version of (4.4). To be specific we consider a two dimensional domain Ω endowed with a uniform rectangular grid, with Δ_h denoting the five-point-star discretization of Δ, and functions z, ψ, y, u discretized by means of grid functions at the nodal points. Numerical results for this case were reported in [BIK] and [BHHK], and convergence can be argued provided the discretized form of (4.18) holds. Let us consider to which extent Theorems 3.2–3.4 provide new insight on confirming convergence, which was observed numerically in practically all examples. Theorem 3.2 is not applicable since $A_h = \beta I + \Delta_h^{-2}$ is not an M-matrix. Theorem 3.4 is applicable with $M = \beta I$ and $K = \Delta_h^{-2}$, and asserts convergence if β is sufficiently large. We also tested numerically the applicability of Theorem 3.3 and found that for $\Omega = (0,1)^2$ the norm condition was satisfied in all cases we tested with grid-size $h \in [10^{-2}, 10^{-1}]$ and $\beta \geq 10^{-4}$, whereas the cone condition $\sum_{i\in\mathcal{I}}(A_{\mathcal{I}}^{-1}y_{\mathcal{I}})_i \geq 0$ for $y_{\mathcal{I}} \geq 0$ was satisfied only for $\beta \geq 10^{-2}$, for the same range of grid-sizes. Still the function $y^k \to \mathcal{M}(y^k)$ utilized in the proof of Theorem 3.4 behaved as a merit function for the wider range of $\beta \geq 10^{-3}$. Note that the norm and cone condition of Theorem 3.4 involve only the system matrix A, whereas $\mathcal{M}(y^k)$ also depends on the specific choice of f and ψ. □

Remark 4.1. Throughout the paper we used the function $\mathcal{C}$ defined in (2.2) as a complementarity function. Another popular choice of complementarity function is

given by the Fischer–Burmeister function

$$\mathcal{C}_{FB}(y,\lambda) = \sqrt{y^2+\lambda^2} - (y+\lambda)\,.$$

Note that $\mathcal{C}_{FB}(0,\lambda) = \sqrt{\lambda^2} - \lambda = 2\max(0,-\lambda)$, and hence by Proposition 4.1 the natural choices for slanting functions do not satisfy property (A). □

Remark 4.2. Condition (H2) can be considered as yet another incidence, where a *two-norm concept* for the analysis of optimal control problems is essential. It utilizes the fact that the control-to-solution mapping of the differential equation is a smoothing operation. Two-norm concepts were used for second order sufficient optimality conditions and the analysis of SQP-methods in [M, I, IK3], for example, and also for semismooth Newton methods in [U]. □

In view of the fact that (P) consists of a quadratic cost functional with affine constraints the question arises whether superlinear convergence coincides with one step convergence after the active/inactive sets are identified by the algorithm. The following example illustrates the fact that this is not the case.

Example 2. We consider Example 1 with the specific choices

$$z(x_1,x_2) = \sin(5x_1) + \cos(4x_2), \quad \psi \equiv 0, \quad \beta = 10^{-5}, \text{ and } \Omega = (0,1)^2.$$

A finite difference based discretization of (4.4) with a uniform grid of mesh size $h = \frac{1}{100}$ and the standard five-point-star discretization of the Laplace operator was used. The primal-dual active set strategy with initialization given by solving the unconstrained problem and setting $\lambda_h^0 = 0$, was used. The exact discretized solution $(u_h^*, \lambda_h^*, y_h^*)$ was attained in eight iterations. In Table 1 we present the values for

$$q_u^k = \frac{|u_h^k - u_h^*|}{|u_h^{k-1} - u_h^*|}, \quad q_\lambda^k = \frac{|\lambda_h^k - \lambda_h^*|}{|\lambda_h^{k-1} - \lambda_h^*|},$$

where the norms are discrete L^2-norms. Clearly these quantities indicate superlinear convergence of u_h^k and λ_h^k.

TABLE 1

k	1	2	3	4	5	6	7
q_u^k	1.0288	0.8354	0.6837	0.4772	0.2451	0.0795	0.0043
q_λ^k	0.6130	0.5997	0.4611	0.3015	0.1363	0.0399	0.0026

We also tested whether the quantities appearing in the rate of convergence discussion are reflected in the numerical results. For this purpose note that for the problem under consideration w appearing in (4.16) and (4.17) is given by $w = \Delta^{-2}u^* + \Delta^{-1}z + \beta\psi$. Roughly, (4.16) and (4.17) have a higher chance to be satisfied with larger value for α if w is not smooth across the boundary of the set $\{x : w(x) = 0\}$. In a numerical test we kept all problem data identical to those specified above except for changing ψ to $\psi(x_1,x_2) = x_1x_2 - 1$. Note that this new ψ increases the chance that (4.16) and (4.17) are satisfied. Moreover, increasing β (for the same ψ) results in an increase of the influence of ψ to w. Thus we expect an improved convergence as β is increased. For the new ψ and small β the algorithm finds the solution in one less iteration. Increasing β results in a further reduction of three iterations; see Tables 1 and 2.

TABLE 2

	q_u^k					
k	1	2	3	4	5	6
$\beta = 10^{-5}$	1.0443	0.8359	0.6780	0.4679	0.2342	0.0614
$\beta = 10^{-3}$	0.1410	0.0455	0.0041	–	–	–

Appendix A.

Proof of Theorem 3.2. The assumption that A is an M-matrix implies that for every index partition $\mathcal{I}$ and $\mathcal{A}$ we have $A_{\mathcal{I}}^{-1} \geq 0$ and $A_{\mathcal{I}}^{-1} A_{\mathcal{I}\mathcal{A}} \leq 0$; see [BP, p. 134]. Let us first show the monotonicity property of the y-component. Observe that for every $k \geq 1$ the complementarity property

$$\lambda_i^k = 0 \quad \text{or} \quad y_i^k = \psi_i\,, \quad \text{for all } i \text{ and } k \geq 1\,, \tag{A.1}$$

holds. For $i \in \mathcal{A}_k$ we have $\lambda_i^k + c(y_i^k - \psi_i) > 0$, and hence by (A.1) either $\lambda_i^k = 0$, which implies $y_i^k > \psi_i$, or $\lambda_i^k > 0$, which implies $y_i^k = \psi_i$. Consequently, $y^k \geq \psi = y^{k+1}$ on $\mathcal{A}_k$ and $\delta y_{\mathcal{A}_k} = \psi_{\mathcal{A}_k} - y_{\mathcal{A}_k}^k \leq 0$. For $i \in \mathcal{I}_k$ we have $\lambda_i^k + c(y_i^k - \psi_i) \leq 0$ which implies $\delta\lambda_{\mathcal{I}_k} \geq 0$ by (2.4) and (A.1). Since $\delta y_{\mathcal{I}_k} = -A_{\mathcal{I}_k}^{-1} A_{\mathcal{I}_k\mathcal{A}_k} \delta y_{\mathcal{A}_k} - A_{\mathcal{I}_k}^{-1} \delta\lambda_{\mathcal{I}_k}$ by (3.3) it follows that $\delta y_{\mathcal{I}_k} \leq 0$. Therefore $y^{k+1} \leq y^k$ for every $k \geq 1$.

Next we show that y^k is feasible for all $k \geq 2$. Due to the monotonicity of y^k it suffices to show that $y^2 \leq \psi$. Let $V = \{i : y_i^1 > \psi_i\}$. For $i \in V$ we have $\lambda_i^1 = 0$ by (A.1), and hence $\lambda_i^1 + c(y_i^1 - \psi_i) > 0$ and $i \in \mathcal{A}_1$. Since $y^2 = \psi$ on $\mathcal{A}_1$ and $y^2 \leq y^1$ it follows that $y^2 \leq \psi$.

To verify that $y^* \leq y^k$ for all $k \geq 1$ note that

$$\begin{aligned} f_{\mathcal{I}_{k-1}} &= \lambda^*_{\mathcal{I}_{k-1}} + A_{\mathcal{I}_{k-1}} y^*_{\mathcal{I}_{k-1}} + A_{\mathcal{I}_{k-1}\mathcal{A}_{k-1}} y^*_{\mathcal{A}_{k-1}} \\ &= A_{\mathcal{I}_{k-1}} y^k_{\mathcal{I}_{k-1}} + A_{\mathcal{I}_{k-1}\mathcal{A}_{k-1}} \psi_{\mathcal{A}_{k-1}}\,. \end{aligned}$$

It follows that

$$A_{\mathcal{I}_{k-1}} \left(y^k_{\mathcal{I}_{k-1}} - y^*_{\mathcal{I}_{k-1}} \right) = \lambda^*_{\mathcal{I}_{k-1}} + A_{\mathcal{I}_{k-1}\mathcal{A}_{k-1}} \left(y^*_{\mathcal{A}_{k-1}} - \psi_{\mathcal{A}_{k-1}} \right)\,.$$

Since $\lambda^*_{\mathcal{I}_{k-1}} \geq 0$ and $y^*_{\mathcal{A}_{k-1}} \leq \psi_{\mathcal{A}_{k-1}}$ the M-matrix properties of A imply that $y^k_{\mathcal{I}_{k-1}} \geq y^*_{\mathcal{I}_{k-1}}$ for all $k \geq 1$.

Turning to the feasibility of λ^k assume that for a pair of indices $(\bar{k}, i)$, $\bar{k} \geq 1$, we have $\lambda_i^{\bar{k}} < 0$. Then necessarily $i \in \mathcal{A}_{\bar{k}-1}$, $y_i^{\bar{k}} = \psi_i$, and $\lambda_i^{\bar{k}} + c(y_i^{\bar{k}} - \psi_i) < 0$. It follows that $i \in \mathcal{I}_{\bar{k}}$, $\lambda_i^{\bar{k}+1} = 0$, and $\lambda_i^{\bar{k}+1} + c(y_i^{\bar{k}+1} - \psi_i) \leq 0$, since $y_i^{k+1} \leq \psi_i$, $k \geq 1$. Consequently, $i \in \mathcal{I}_{\bar{k}+1}$ and by induction $i \in \mathcal{I}_k$ for all $k \geq \bar{k} + 1$. Thus, whenever a coordinate of λ^k becomes negative at iteration $\bar{k}$, it is zero from iteration $\bar{k} + 1$ onwards, and the corresponding primal coordinate is feasible. Due to finite dimensionality of $\mathbb{R}^n$ it follows that there exists k_o such that $\lambda^k \geq 0$ for all $k \geq k_o$.

Monotonicity of y^k and $y^* \leq y^k \leq \psi$ for $k \geq 2$ imply the existence of $\bar{y}$ such that $\lim y^k = \bar{y} \leq \psi$. Since $\lambda^k = Ay^k + f \geq 0$ for all $k \geq k_o$, there exists $\bar{\lambda}$ such that $\lim \lambda^k = \bar{\lambda} \geq 0$. Together with (A.1) it follows that $(\bar{y}, \bar{\lambda}) = (y^*, \lambda^*)$. ∎

Remark A.1. From the proof it follows that if $\lambda_i^{\bar{k}} < 0$ for some coordinate i at iteration $\bar{k}$, then $\lambda_i^k = 0$ and $y_i^k \leq \psi_i$ for all $k \geq \bar{k} + 1$.

Proof of Proposition 4.1. (i) It suffices to consider the one dimensional case $\Omega = (-1, 1) \subset \mathbb{R}$. We show that property (A) does not hold at $y(x) = -|x|$. Let us

define $h_n(x) = \frac{1}{n}$ on $(-\frac{1}{n}, \frac{1}{n})$ and $h_n(x) = 0$ otherwise. Then

$$\int_{-1}^{1} |\max(0, y + h_n)(x) - \max(0, y)(x) - (G_m(y + h_n)(h_n))(x)|^p \, dx$$
$$= \int_{\{x : y(x) + h_n(x) > 0\}} |y(x)|^p dx = \int_{-\frac{1}{n}}^{\frac{1}{n}} |y(x)|^p dx = \frac{2}{p+1} \left(\frac{1}{n}\right)^{p+1},$$

and $\|h_n\|_{L^p} = \sqrt[p]{2/n^{p+1}}$. Consequently,

$$\lim_{n\to\infty} \frac{1}{\|h_n\|_{L^p}} \|\max(0, y + h_n) - \max(0, y) - G_m(y + h_n)h_n\|_{L^p} = \sqrt[p]{\frac{1}{p+1}} \neq 0,$$

and hence (A) is not satisfied at y for any $p \in [1, \infty)$.

To consider the case $p = \infty$ we choose $\Omega = (0, 1)$ and show that (A) is not satisfied at $y(x) = x$. For this purpose define for $n = 2, \ldots$

$$h_n(x) = \begin{cases} -(1 + \frac{1}{n})x & \text{on } (0, \frac{1}{n}], \\ (1 + \frac{1}{n})x - \frac{2}{n}(1 + \frac{1}{n}) & \text{on } (\frac{1}{n}, \frac{2}{n}], \\ 0 & \text{on } (\frac{2}{n}, 1]. \end{cases}$$

Observe that $E_n = \{x : y(x) + h_n(x) < 0\} \supset (0, \frac{1}{n}]$. Therefore

$$\lim_{n\to\infty} \frac{1}{\|h_n\|_{L^\infty([0,1])}} \|\max(0, y + h_n) - \max(0, y) - G_m(y + h_n)h_n\|_{L^\infty([0,1])}$$
$$= \lim_{n\to\infty} \frac{n^2}{n+1} \|y\|_{L^\infty(E_n)} \geq \lim_{n\to\infty} \frac{n}{n+1} = 1,$$

and hence (A) cannot be satisfied.

(ii) Let $\delta \in \mathbb{R}$ be fixed arbitrarily and y, $h \in L^q(\Omega)$, and set

$$D_{y,h}(x) = \max(0, y(x) + h(x)) - \max(0, y(x)) - G_m(y + h)(x)h(x).$$

A short computation shows that

$$|D_{y,h}(x)| \begin{cases} \leq |y(x)| & \text{if } (y(x) + h(x))y(x) < 0, \\ \leq (1 + |\delta|)\,|y(x)| & \text{if } y(x) + h(x) = 0, \\ = 0 & \text{otherwise.} \end{cases} \tag{A.2}$$

For later use we note that from Hölder's inequality we obtain for $1 \leq p < q \leq \infty$

$$\|w\|_{L^p} \leq |\Omega|^r \|w\|_{L^q}, \quad \text{with } r = \begin{cases} \frac{q-p}{pq} & \text{if } q < \infty, \\ \frac{1}{p} & \text{if } q = \infty. \end{cases}$$

From (A.2) it follows that only

$$\Omega_0(h) = \{x \in \Omega : y(x) \neq 0,\ y(x)(y(x) + h(x)) \leq 0\}$$

requires further investigation. For $\epsilon > 0$ we define subsets of $\Omega_0(h)$ by

$$\Omega_\epsilon(h) = \{x \in \Omega : |y(x)| \geq \epsilon,\ y(x)(y(x) + h(x)) \leq 0\}.$$

Note that $|y(x)| \geq \epsilon$ a.e. on $\Omega_\epsilon(h)$ and therefore

$$\|h\|_{L^q(\Omega)} \geq \epsilon |\Omega_\epsilon(h)|^{1/q} \text{ for } q < \infty.$$

It follows that

$$\lim_{\|h\|_{L^q(\Omega)}\to 0} |\Omega_\epsilon(h)| = 0 \quad \text{for every fixed } \epsilon > 0\,. \tag{A.3}$$

For $\epsilon > 0$ we further define sets

$$\Omega^\epsilon(y) = \{x \in \Omega : 0 < |y(x)| \le \epsilon\} \subset \{x : y(x) \neq 0\}\,.$$

Note that $\Omega^\epsilon(y) \subset \Omega^{\epsilon'}(y)$ whenever $0 < \epsilon \le \epsilon'$ and $\bigcap_{\epsilon>0} \Omega^\epsilon(y) = \emptyset$. As a consequence

$$\lim_{\epsilon\to 0^+} |\Omega^\epsilon(y)| = 0\,. \tag{A.4}$$

From (A.2) we find

$$\begin{aligned}
\frac{1}{\|h\|_{L^q}}\|D_{y,h}\|_{L^p} &\le \frac{1+|\delta|}{\|h\|_{L^q}} \left(\int_{\Omega_0(h)} |y(x)|^p dx\right)^{1/p} \\
&\le \frac{1+|\delta|}{\|h\|_{L^q}} \left[\left(\int_{\Omega_\epsilon(h)} |y(x)|^p dx\right)^{1/p} + \left(\int_{\Omega_0(h)\setminus\Omega_\epsilon(h)} |y(x)|^p dx\right)^{1/p}\right] \\
&\le \frac{1+|\delta|}{\|h\|_{L^q}} \left[|\Omega_\epsilon(h)|^{(q-p)/(qp)} \left(\int_{\Omega_\epsilon(h)} |y(x)|^q dx\right)^{1/q} \right. \\
&\qquad \left. + |\Omega^\epsilon(y)|^{(q-p)/(qp)} \left(\int_{\Omega_0(h)\setminus\Omega_\epsilon(h)} |y(x)|^q dx\right)^{1/q}\right] \\
&\le (1+|\delta|)\left(|\Omega_\epsilon(h)|^{(q-p)/(qp)} + |\Omega^\epsilon(y)^{(q-p)/(qp)}|\right)\,.
\end{aligned}$$

Choose $\eta > 0$ arbitrarily and note that by (A.4) there exists $\bar\epsilon > 0$ such that $(1 + |\delta|)|\Omega^{\bar\epsilon}(y)|^{(q-p)/(qp)} < \eta$. Consequently,

$$\frac{1}{\|h\|_{L^q}}\|D_{y,h}\|_{L^p} \le (1+|\delta|)|\Omega_{\bar\epsilon}(h)|^{(q-p)/(qp)} + \eta$$

and by (A.3)

$$\lim_{\|h\|_{L^q}\to 0} \frac{1}{\|h\|_{L^q}}\|D_{y,h}\|_{L^p} \le \eta\,.$$

Since $\eta > 0$ is arbitrary the claim holds for $1 \le p < q < \infty$.

The case $q = \infty$ follows from the result for $1 \le p < q < \infty$. ∎

Proof of Theorem 4.1. Let y^k, $k \ge 1$, denote the iterates of the reduced algorithm and define

$$\lambda^{k+1} = \begin{cases} 0 & \text{on } \mathcal{I}_k\,, \\ (f - Cy^{k+1} - \beta\psi)_{\mathcal{A}_k} & \text{on } \mathcal{A}_k \end{cases} \qquad \text{for } k = 0, 1, \ldots.$$

We obtain $\lambda^k + \beta(y^k - \psi) = f - Cy^k - \beta\psi$ for $k = 1, 2, \ldots$, and hence the active sets $\mathcal{A}_k$, the iterates y^{k+1} produced by the reduced algorithm and by the algorithm in the two variables (y^{k+1}, λ^{k+1}), coincide for $k = 1, 2, \ldots$, provided the initialization strategies coincide. This, however, is the case since due to our choice of λ^0 and $\beta = c$

we have $\lambda^0 + \beta(y^0 - \psi) = f - Cy^0 - \beta\psi$, and hence the active sets coincide for $k = 0$ as well.

To prove convergence of the reduced algorithm we utilize Theorem 1.1 with $F : L^2(\Omega) \to L^2(\Omega)$ given by $F(y) = \beta y - \beta\psi + \max(0, Cy - f + \beta\psi)$. From Proposition 4.1(ii) it follows that F is slantly differentiable. In fact, the relevant difference quotient for the nonlinear term in F is

$$\frac{1}{\|Ch\|_{L^q}} \big\| \max(0, Cy - f + \beta\psi + Ch) - \max(0, Cy - f + \beta\psi) - G_m(Cy - f + \beta\psi + Ch)(Ch) \big\|_{L^2} \frac{\|Ch\|_{L^q}}{\|h\|_{L^2}},$$

which converges to 0 for $\|h\|_{L^2} \to 0$. Here

$$G_m(Cy - f + \beta\psi + Ch)(x) = \begin{cases} 1 & \text{if } (C(y+h) - f + \beta\psi)(x) \geq 0, \\ 0 & \text{if } (C(y+h) - f + \beta\psi)(x) < 0, \end{cases}$$

so that in particular δ of (4.1) was set equal to 1 which corresponds to the "$\leq$" sign in the definition of $\mathcal{I}_k$. A slanting function G_F of F at y in direction h is therefore given by

$$G_F(y + h) = \beta I + G_m(Cy - f + \beta\psi + Ch)C.$$

It remains to argue that $G_F(z) \in \mathcal{L}(L^2(\Omega))$ has a bounded inverse. Since for arbitrary $z \in L^2(\Omega)$, $h \in L^2(\Omega)$

$$G_F(z)h = \begin{pmatrix} \beta I_{\mathcal{I}} + C_{\mathcal{I}} & C_{\mathcal{I}\mathcal{A}} \\ 0 & \beta I_{\mathcal{A}} \end{pmatrix} \begin{pmatrix} h_{\mathcal{I}} \\ h_{\mathcal{A}} \end{pmatrix},$$

where $\mathcal{I} = \{x : (Cz - f + \beta\psi)(x) \geq 0\}$ and $\mathcal{A} = \{x : (Cz - f + \beta\psi)(x) < 0\}$, it follows from (H1) that $G_F(z)^{-1} \in \mathcal{L}(L^2(\Omega))$. Above we denoted $C_{\mathcal{I}} = E_{\mathcal{I}}^* C E_{\mathcal{I}}$ and $C_{\mathcal{I}\mathcal{A}} = E_{\mathcal{I}}^* C E_{\mathcal{A}}$. □

Acknowledgment. We appreciate many helpful comments by the referees.

REFERENCES

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

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

[BP] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Computer Science and Scientific Computing Series, Academic Press, New York, 1979.

[Ca] E. Casas, *Control of an elliptic problem with pointwise state constraints*, SIAM J. Control Optim., 24 (1986), pp. 1309–1318.

[CNQ] X. Chen, Z. Nashed, and L. Qi, *Smoothing methods and semismooth methods for nondifferentiable operator equations*, SIAM J. Numer. Anal., 38 (2000), pp. 1200–1216.

[C] F.H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, 1983.

[FFKP] F. Facchinei, A. Fischer, C. Kanzow, and J.-M. Peng, *A simply constrained optimization reformulation of KKT systems arising from variational inequalities*, Appl. Math. Optim., 40 (1999), pp. 19–37.

[FK] A. Fischer and C. Kanzow, *On finite termination of an iterative method for linear complementarity*, Math. Programming, 74 (1996), pp. 279–292.

[HKT] M. Heinkenschloss, C.T. Kelley, and H.T. Tran, *Fast algorithms for nonsmooth compact fixed-point problems*, SIAM J. Numer. Anal., 29 (1992), pp. 1769–1792.

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

[I] 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.

[IK1] K. Ito and K. Kunisch, *Augmented Lagrangian methods for nonsmooth convex optimization in Hilbert spaces*, Nonlinear Anal., 41 (2000), pp. 573–589.

[IK2] K. Ito and K. Kunisch, *Optimal control of elliptic variational inequalities*, Appl. Math. Optim., 41 (2000), pp. 343–364.

[IK3] K. Ito and K. Kunisch, *Newton's method for a class of weakly singular optimal control problems*, SIAM J. Optim., 10 (2000), pp. 896–916.

[KS] C.T. Kelley and E.W. Sachs, *Multilevel algorithms for constrained compact fixed point problems*, SIAM J. Sci. Comput., 15 (1994), pp. 645–667.

[K1] B. Kummer, *Newton's method for nondifferentiable functions*, in Advances in Optimization Math. Res. 45, J. Guddat et al., eds., Akademie-Verlag, Berlin, 1988, pp. 114–125.

[K2] B. Kummer, *Generalized Newton and NCP methods: Convergence, regularity, actions*, Discuss. Math. Differ. Incl. Control Optim., 20 (2000), pp. 209–244.

[KNT] Y.A. Kuznetsov, P. Neittaanmäki, and P. Tarvainen, *Overlapping block methods for obstacle problems with convection-diffusion operators*, in Complementarity and Variational Problems, M.C. Ferris and J.-S. Pang, eds., SIAM, Philadelphia, 1997.

[Mi] R. Mifflin, *Semismooth and semiconvex functions in constrained optimization*, SIAM J. Control Optim., 15 (1977), pp. 959–972.

[M] H. Maurer, *First and second order sufficient optimality conditions in mathematical programming and optimal control*, Math. Prog. Study, 14 (1981), pp. 43–62.

[Q1] L. Qi, *Convergence analysis of some algorithms for solving nonsmooth equations*, Math. Oper. Res., 18 (1993), pp. 227–244.

[Q2] H.D. Qi, *A regularized smoothing Newton method for box constrained variational inequality problems with P_0-functions*, SIAM J. Optim., 10 (1999), pp. 315–330.

[QS] L. Qi and J. Sun, *A nonsmooth version of Newton's method*, Math. Programming, 58 (1993), pp. 353–367.

[UU] M. Ulbrich and S. Ulbrich, *Superlinear convergence of affine-scaling interior-point Newton methods for infinite-dimensional nonlinear problems with pointwise bounds*, SIAM J. Control Optim., 38 (2000), pp. 1938–1984.

[U] M. Ulbrich, *Semismooth Newton methods for operator equations in function spaces*, SIAM J. Optim., 13 (2003), pp. 805–842.

SIAM J. OPTIM.
Vol. 13, No. 3, pp. 889–903

ANALYSIS OF GENERALIZED PATTERN SEARCHES*

CHARLES AUDET† AND J. E. DENNIS, JR.‡

Abstract. This paper contains a new convergence analysis for the Lewis and Torczon generalized pattern search (GPS) class of methods for unconstrained and linearly constrained optimization. This analysis is motivated by a desire to understand the successful behavior of the algorithm under hypotheses that are satisfied by many practical problems. Specifically, even if the objective function is discontinuous or extended-valued, the methods find a limit point with some minimizing properties. Simple examples show that the strength of the optimality conditions at a limit point depends not only on the algorithm, but also on the directions it uses and on the smoothness of the objective at the limit point in question. The contribution of this paper is to provide a simple convergence analysis that supplies detail about the relation of optimality conditions to objective smoothness properties and to the defining directions for the algorithm, and it gives previous results as corollaries.

Key words. pattern search algorithm, linearly constrained optimization, surrogate-based optimization, nonsmooth optimization, derivative-free convergence analysis

AMS subject classifications. 90C30, 90C56, 65K05

PII. S1052623400378742

1. Introduction.

Generalized pattern search (GPS) algorithms were defined and analyzed by Torczon [29] for derivative-free unconstrained optimization on continuously differentiable functions using positive spanning directions. Lewis and Torczon [24] introduced the idea of using positive spanning directions with GPS. In [23], they showed that if the objective is continuously differentiable and if the set of directions that define the local search is chosen properly with respect to the boundary of the feasible region, then the GPS framework and convergence theory extend to bound-constrained optimization. In [25], they showed the same results for problems with a finite number of linear constraints. Both these extensions use the appealing "barrier" strategy of declaring any infeasible point to be unacceptable as a next iterate. Our purpose here is to provide a new unified analysis for the methods in [29, 23, 25] and to help elucidate the relationship between the algorithm, the search directions, and the local smoothness properties of the objective at certain specified limit points of the algorithm.

The optimization problem considered in this paper is

$$\min_{x\in\Omega} f(x), \qquad \text{where } f : \Re^n \to \Re \cup \{\infty\}. \tag{1.1}$$

We assume as in [25] that $\Omega = \{x \in \Re^n : \ell \le Ax \le u\}$, where $A \in \mathcal{Q}^{m\times n}$ is a rational matrix, $\ell, u \in \{\Re \cup \{\pm\infty\}\}^m$, and $\ell \le u$.

*Received by the editors September 26, 2000; accepted for publication (in revised form) July 6, 2002; published electronically February 12, 2003. This work was supported by DOE DE-FG03-95ER25257, AFOSR F49620-01-1-0013, The Boeing Company, Sandia LG-4253, ExxonMobil, and the LANL Computer Science Institute (LACSI) contract 03891-99-23.

http://www.siam.org/journals/siopt/13-3/37874.html

†Département de Mathématiques et de Génie Industriel, École Polytechnique de Montréal, C.P. 6079, Succ. Centre-ville, Montréal, QC, H3C 3A7 Canada (Charles.Audet@gerad.ca, http://www.gerad.ca/Charles.Audet). This author was supported by NSERC (Natural Sciences and Engineering Research Council) fellowship PDF-207432-1998 during a postdoctoral stay at Rice University.

‡Computational and Applied Mathematics Department, Rice University, MS 134, 6100 Main Street, Houston, TX 77005-1892 (dennis@caam.rice.edu, http://www.caam.rice.edu/∼dennis).

GPS methods are extremely effective for some engineering design problems with expensive function evaluations when used with less expensive surrogates [5, 6]. For these and many other applied problems, a call to the subroutine that evaluates $f(x)$ may result unexpectedly in no value being returned even for a feasible x, which we model as $f(x) = \infty$. Reasons for this behavior are discussed in [5], where GPS with surrogates is shown to be effective on a helicopter rotor design example, for which no value is returned roughly 66% of the time. The issue is discussed in a different algorithmic and application context in [7, 8]. In such instances, we cannot assume global smoothness, not even continuity. We are not the first to observe that GPS can work well on nonsmooth problems, but previous convergence theorems do not apply to such problems.

We view the barrier approach as applying the algorithm not to f but to the barrier function $f_\Omega = f + \psi_\Omega$, where ψ_Ω is the indicator function for Ω. It is zero on Ω, and ∞ elsewhere. Clearly then, we do not evaluate $f(x)$ if x is infeasible, because we know that its value is immaterial since the algorithm works with f_Ω, and the value of f_Ω is $+\infty$ on all points that are either infeasible or at which f is declared to be $+\infty$:

$$f_\Omega(x) = \begin{cases} f(x) & \text{if } x \in \Omega, \\ \infty & \text{otherwise.} \end{cases}$$

The reason that we treat together all the methods in [29, 23, 25] that use the barrier approach is that, by viewing them as the same algorithm applied to f_Ω, we can treat them by corollaries of a single result, Theorem 3.7, that allows for extended values and other nonsmooth behavior. Our approach is first to identify a class of promising limit points produced by GPS applied to extended-valued discontinuous functions like f_Ω. To make statements about optimality conditions at these limit points, we work not with f_Ω but with f. If f is lower semicontinuous at such a limit point, we can make a weak optimality statement. Then we apply the Clarke calculus [9] locally to f at such a point to relate progressively stronger optimality conditions to progressively stronger local smoothness assumptions at the limit point.

Thus, the structure of our results will be that, at some limit point whose existence is asserted independent of certain assumptions, we make those additional assumptions to draw stronger conclusions. This is standard for Newton or quasi-Newton methods (e.g., [27, Theorem 8.6, p. 216] or virtually all of [22]), but it has not been the norm for direct search methods.

Specifically, without assuming any smoothness, we observe that there is a convergent subsequence of the sequence $\{x_k\}$ of iterates produced by the algorithm. Since $\{f(x_k)\}$ generated by the algorithm is nonincreasing, it is convergent to a finite limit if it is bounded below. Thus, if f is lower semicontinuous at any limit point $\bar{x}$ of the sequence of iterates, then $f(\bar{x}) \leq \liminf_k f(x_k) = \lim_k f(x_k)$. Our analysis is of interest for the heat intercept design problem given in [21], where f is not continuous at one of the limit points generated, but a plot suggests that it is lower semicontinuous.

Again without any smoothness assumptions, we show that there is a limit point $\hat{x}$ of a subsequence of $\{x_k\}$ consisting of iterates on progressively finer meshes. (A formal definition of the mesh is given in section 2.) These specific iterates of interest are *mesh local optimizers* in that they minimize the function on a positive spanning set of neighboring mesh points. This will be made precise in section 2.

The directional tests that led GPS to refine the mesh at mesh local optimizers are exactly that difference quotients be nonnegative for the Clarke generalized directional derivative at $\hat{x}$. If the Clarke derivatives exist at $\hat{x}$, as they will if f is locally

Lipschitz at $\hat{x}$, then these nonnegative difference quotients pass through the limit to be nonnegative Clarke derivatives in the directions used.

Nonnegative directional derivatives on a set of positive spanning directions for $\Re^n$ are a necessary condition for optimality, but that is not the usual first order condition. To get the usual condition that the gradient is zero, we assume in addition that the generalized gradient of f is a singleton. This extra smoothness causes the above directional optimality conditions to hold for all directions in $\Re^n$. We give examples that supplement those in [1] and show that our results are sharp in that they predict the behavior of the algorithm.

The remainder of the paper is organized as follows: in the next section, we will give a brief description of the GPS algorithm class. We adhere to a slightly different, but equivalent, version of the Lewis and Torczon algorithm. In section 3, we present the assumptions together with a discussion of our local smoothness conditions, give the key result and some immediate corollaries for unconstrained problems together with a discussion of these results, and then go on to the results for the linear constraints. Section 4 is devoted to some concluding remarks.

2. GPS algorithms. GPS algorithms for unconstrained or linearly constrained minimization generate a sequence of iterates $\{x_k\}$ in $\Re^n$ with nonincreasing objective function values. Each iteration is divided into two phases: an optional SEARCH and a local POLL, defined next.

In the SEARCH step, the barrier objective function f_Ω is evaluated at a finite number of points on a mesh (a discrete subset of $\Re^n$ defined below, whose fineness is parameterized by the *mesh size parameter* $\Delta_k > 0$) to try to find one that yields a lower objective function value than the incumbent. Any strategy may be used to select the mesh points that are candidates to replace the incumbent, as long as only finitely many points (including none) are selected.

This is a key point. The SEARCH step accommodates whatever heuristics the user was already using to attack his or her problem using surrogates. One might do some random search on the mesh using the surrogate, or, as in the Boeing Design Explorer software [4], one might apply SQP to the surrogate problem and then move the solution to a nearby mesh point to choose the candidates at which to evaluate the expensive objective function in hopes of obtaining a better next iterate. Coope and Price [11] offer a possibility for a related framework that does not require pushing a surrogate solution to the mesh for it to become an acceptable trial point. In [13], they apply the Clarke analysis given here with their related methods.

On the other hand, the freedom of the SEARCH step is definitely a theoretical liability. In [1] and here, there are examples of nonempty searches that spoil chances for the algorithm to find KKT points, and of empty searches that mire the algorithm at a poor point when a naive random selection from the current mesh in the SEARCH would generally lead to success. Regardless, this freedom must be retained. Indeed, for the Boeing example [5, 6], the algorithm with surrogates is much more efficient than Serafini's implementation [28] of the Dennis–Torczon MDS/PDS algorithm [14]. This is not to disparage the MDS algorithm, which is very robust on that example.

Below, we will offer terminology consistent with that of Coope and Price to replace the usual "successful/unsuccessful" terminology in the GPS literature. The original terminology was adequate until it was recognized that the "unsuccessful" iterations were the important ones because they produce *mesh local optimizers*, while successful iterations produce only *improved mesh points*, which we define now.

When the incumbent is replaced, i.e., when $f_\Omega(x_{k+1}) < f_\Omega(x_k)$, or equivalently,

when $f(x_{k+1}) < f(x_k)$, then x_{k+1} is said to be an *improved mesh point*. When the SEARCH step fails to provide an improved mesh point, the POLL step is invoked. This second step consists of evaluating the barrier objective function at the neighboring mesh points to see whether a lower function value can be found there.

When the POLL step fails to provide an improved mesh point, then the current incumbent solution is said to be a *mesh local optimizer* (i.e., its objective function value is less than or equal to that of neighboring mesh points). The algorithm then refines the mesh by setting the mesh size parameter

$$\Delta_{k+1} = \tau^{w_k}\Delta_k \tag{2.1}$$

for $0 < \tau^{w_k} < 1$, where $\tau > 1$ is a rational number that remains constant over all iterations and $w_k \leq -1$ is an integer bounded below by the constant $w^- \leq -1$.

A feature first noted in Torczon [29] and also supported in the analysis given here is that if either the SEARCH or POLL step produces an improved mesh point, the current iteration can stop, and the new point $x_{k+1} \neq x_k$ has a strictly lower objective function value, the mesh size parameter is kept the same or is increased to carry out the next SEARCH step, and the process is reiterated. The coarsening of the mesh follows the rule

$$\Delta_{k+1} = \tau^{w_k}\Delta_k, \tag{2.2}$$

where $\tau > 1$ is defined above and $w_k \geq 0$ is an integer bounded above by $w^+ \geq 0$. Our experience with surrogate-based SEARCH steps [5, 6] is that a great deal of progress can be made with few function values, and at least $n + 1$ function evaluations are needed to show only local mesh optimality, which indicates that the mesh needs to be refined (see [24] for defining a minimal number of polling directions).

By modifying the mesh size parameters as above, it follows that for any $k \geq 0$ there exists an integer $r_k \in \mathcal{Z}$ such that

$$\Delta_k = \tau^{r_k}\Delta_0. \tag{2.3}$$

The basic ingredient in the definition of the mesh is a set of positive spanning directions D in $\Re^n$ (more precisely, nonnegative linear combinations of the elements of the set D span $\Re^n$). There is great freedom in choosing these directions; only the following additional rule needs to be respected: each direction $d_j \in D$ (for $j = 1, 2, \ldots, |D|$) is the product $G\bar{z}_j$ of the nonsingular generating matrix $G \in \Re^{n\times n}$ by an integer vector $\bar{z}_j \in \mathcal{Z}^n$. Note that the same generating matrix is used for all directions. For convenience, the set D is also viewed as a real $n \times |D|$ matrix. Similarly, we denote the matrix whose columns are $\bar{z}_j$, for $j = 1, 2, \ldots, |D|$, by $\bar{Z}$; we can therefore write $D = G\bar{Z}$. At iteration k, the mesh is centered around the current iterate $x_k \in \Re^n$, and its fineness is parameterized through the mesh size parameter Δ_k as follows:

$$M_k = \{x_k + \Delta_k Dz : z \in \mathcal{Z}_+^{|D|}\}, \tag{2.4}$$

where $\mathcal{Z}_+$ is the set of nonnegative integers. This way of describing the mesh differs from [29, 23, 25].

At each iteration, some positive spanning matrix D_k composed of columns of D is used to construct the POLL set. We write $D_k \subseteq D$ to signify that the matrix D_k is composed of columns of D. The poll set is composed of mesh points neighboring the current iterate x_k in the directions of the columns of D_k:

$$\text{POLL set:} \quad \{x_k + \Delta_k d : d \in D_k\}. \tag{2.5}$$

Rules for selecting D_k may depend on the user's dynamic intervention during the current run, or, for example, on the iteration number or the current iterate, i.e., $D_k = D(k, x_k) \subseteq D$.

The algorithm is stated formally as follows.

A BASIC GPS ALGORITHM.

- *Initialization*:
  Let x_0 be such that $f_\Omega(x_0)$ is finite. Let D be a positive spanning set, and let M_0 be the mesh on $\Re^n$ defined by $\Delta_0 > 0$ and D_0 (see (2.4)). Set the iteration counter k to 0.
- SEARCH *and* POLL *steps*:
  Perform the SEARCH and possibly the POLL steps (or only part of them) until an improved mesh point x_{k+1} with the lowest f_Ω value so far is found on the mesh M_k defined by (2.4).
  - Optional SEARCH: Evaluate f_Ω on a finite subset of trial points on the mesh M_k defined by (2.4). (The strategy that gives the set of points is usually provided by the user; it must be finite and the set can be empty.)
  - Local POLL: Evaluate f_Ω on the poll set defined in (2.5).
- *Parameter update*:
  If the SEARCH or the POLL step produced an improved mesh point, i.e., a feasible iterate $x_{k+1} \in M_k \cap \Omega$ for which $f_\Omega(x_{k+1}) < f_\Omega(x_k)$, then update $\Delta_{k+1} \geq \Delta_k$ according to rule (2.2).
  Otherwise, $f_\Omega(x_k) \leq f_\Omega(x_k + \Delta_k d)$ for all $d \in D_k$, and so x_k is a mesh local optimizer. Set $x_{k+1} = x_k$ and update $\Delta_{k+1} < \Delta_k$ according to rule (2.1).
  Increase $k \leftarrow k + 1$ and go back to the SEARCH and POLL step.

The SEARCH strategy is the key to the algorithm's effectiveness, as we discussed above. The convergence analysis is independent of the SEARCH step, provided that it is finite and returns a point (or points) on the mesh. The POLL step applied to f_Ω, as we will see, guarantees that the limit point provided by the algorithm satisfies optimality conditions whose strength depends on the local smoothness of f at the limit point.

3. Convergence analysis. Theorem 3.7 is our main result. It and Theorem 3.1 make no special assumptions about the crucial relationship between the directions D and the feasible region Ω. This means that they apply to quite general uses of GPS (see also the remark following Theorem 3.14); but, without a connection between Ω and D, the resulting constrained optimality conditions are weak even when f is smooth. Theorem 3.9 is the strongest result we expect for stationarity in the unconstrained case (see [1] for supporting examples).

Since one of the objectives of the paper is to simplify the convergence analysis of GPS, we include the proofs of all the results leading to our main theorem, even if some of them can essentially be found in previous work modulo the slightly different way of defining the mesh (we indicate the appropriate references).

3.1. Assumptions and smoothness requirements. We make the standard assumption that all iterates produced by GPS lie in a compact set (see [2, 3, 10, 11, 12, 16, 17, 18]). A sufficient condition for this to hold is that the level set $L(x_0) = \{x \in \Omega : f(x) \leq f(x_0)\}$ be compact. We cannot assume that $L(x_0)$ is compact because we allow discontinuities and even $f(x) = \infty$, and so we do not know that $L(x_0)$ is closed. However, we can assume that $L(x_0)$ is bounded so that its closure is compact.

Whatever we assume to ensure that the iterates are in a compact set, this already implies that there are convergent subsequences of the iteration sequence. This is enough to say that if f is lower semicontinuous at such a limit point $\bar{x}$, then $f(\bar{x}) \leq \lim_k f(x_k)$ for the entire iteration sequence. Of course, arbitrarily near a point at which it is lower semicontinuous, f can be infinite, which means that there can be points of the sort mentioned above at which f fails to evaluate arbitrarily near $\bar{x}$, but it also means that we can say nothing about any derivatives at such an $\bar{x}$. For that, we will consider an interesting set of subsequences identified by the algorithm. Specifically, we will be concerned here, as in [2, 11, 12], with the iterates x_k that are mesh local optimizers for meshes that get infinitely fine. We will use $\bar{x}$ to denote generic limit points of the sequence of iterates, and $\hat{x}$ for limit points of mesh local optimizers for meshes that get infinitely fine. It is only at mesh local optimizers that Δ_k is reduced. The analysis would be simpler if we assumed that the mesh size was never coarsened, since obviously then the meshes would become infinitely fine for every sequence of mesh local optimizers. However, we will not use this assumption, since mesh coarsening can lead more rapidly to a deeper basin than might be found without it.

To summarize, the convergence analysis provided below relies only on the following assumptions.

A1: A function $f_\Omega = f + \psi_\Omega : \Re \to \Re \cup \{+\infty\}$ and initial point $x_0 \in \Re^n$ (with $f_\Omega(x_0) < \infty$) are available.
A2: The constraint matrix A is rational.
A3: All iterates $\{x_k\}$ produced by the GPS algorithm lie in a compact set.

We now prove the following result with an immediate, but rather strange, implication—stationary points are the least interesting locally smooth limit points that GPS produces, in the sense that all limit points have the same function value but there are descent directions leading from any locally smooth nonstationary points. Of course, if all the limit points are stationary points, then all are equally interesting.

THEOREM 3.1. *Under assumptions* A1 *and* A3*, there exists at least one limit point of the iteration sequence* $\{x_k\}$*. If* f *is lower semicontinuous at such a limit point* $\bar{x}$*, then* $\lim_k f(x_k)$ *exists and is greater than or equal to* $f(\bar{x})$*. If* f *is continuous at every limit point of* $\{x_k\}$*, then every limit point has the same function value.*

Proof. Since f is lower semicontinuous at $\bar{x}$, we know that for any subsequence $\{x_k\}_{k\in K}$ of the iteration sequence that converges to $\bar{x}$, $\liminf_{k\in K} f(x_k) \geq f(\bar{x})$, which is finite. But since the subsequence of function values is a subsequence of a nonincreasing sequence, they have the same limit inferior. Thus, the entire sequence is also bounded below by $f(\bar{x})$, and thus it converges. ∎

To prove more, we will need to assume more. In addition to A1–A3, previous work on pattern search algorithms assumes continuous differentiability of the function f on a neighborhood of the level set $L(x_0) = \{x \in \Omega : f(x) \leq f(x_0)\}$ (see [2, 23, 25, 29, 11, 12]). In the unconstrained case, Torczon [29] shows that for GPS there exists a limit point $\bar{x}$ satisfying $\nabla f(\bar{x}) = 0$, and our [2] shows the same result for every limit point $\hat{x}$ of any sequence of mesh local optimizers for which $\lim_k \Delta_k = 0$. Note that, since every limit point of the GPS sequence is a point of continuity in this case, nonstationary limit points, whose possible existence is shown in [1], are very interesting because with the right SEARCH step, or the right choice of directions, one can proceed to a point with a better value of f. Our analysis below uses the weaker assumption of strict differentiability (defined in the first paragraph of section 3.4) at such a limit point instead of continuous differentiability on $L(x_0)$).

First we easily show (under no smoothness assumptions) the existence of at least one limit point of a subsequence of mesh local optimizers on meshes that get infinitely fine. Then, for those limit points at which f is strictly differentiable, we show that the gradient is zero. To avoid confusion about the relative strength of assuming in the context of GPS that f is locally Lipschitz, strictly differentiable at a point, or continuously differentiable, we will provide examples following Theorems 3.7 and 3.9 for which those results apply and earlier results do not. The proofs of the mesh refinement results were first given in [29] with a different description of the meshes.

We now proceed with some results on the behavior of the mesh and mesh size parameter. These results do not depend at all on the smoothness of f_Ω; they use just the definition of the algorithm and integrality of the matrix $\overline{Z}$ used to construct the set of directions D. For a different framework, Coope and Price [11] relax the conditions on the mesh, but they assume that the meshes become infinitely fine. This is an interesting tradeoff that puts the burden for ensuring that the meshes become infinitely fine into the implementation but allows for search points off the mesh and more freedom in the definition of the meshes.

3.2. Mesh refinement. The main result of this section is that there is a subsequence of mesh local optimizers for which the mesh size parameter goes to zero. The first lemma shows that for each mesh M_k defined by (2.4), the minimal distance over all pairs of distinct mesh points is bounded below by the mesh size parameter Δ_k times a scalar. In the Euclidean norm, the proof involves the smallest singular value of G (see [29]).

LEMMA 3.2. *For any integer $k \geq 0$, any norm for which any nonzero integer vector has norm at least* 1*, and M_k defined by* (2.4)*,*

$$\min_{u \neq v \in M_k} \|u - v\| \;\geq\; \frac{\Delta_k}{\|G^{-1}\|}.$$

Proof. Using (2.4), we let $u = x_k + \Delta_k D z_u$ and $v = x_k + \Delta_k D z_v$ be two distinct points on M_k, with both z_u and z_v in $\mathcal{Z}_+^{|D|}$. Then

$$\|u - v\| \;=\; \Delta_k \|D(z_u - z_v)\| \;=\; \Delta_k \|G\bar{Z}(z_u - z_v)\| \;\geq\; \Delta_k \frac{\|\bar{Z}(z_u - z_v)\|}{\|G^{-1}\|} \;\geq\; \frac{\Delta_k}{\|G^{-1}\|}.$$

The last part of the inequality is due to the fact that $\bar{Z}(z_u - z_v)$ is a nonzero integer vector; thus its norm is greater than or equal to one. ∎

The separation between mesh points shown by Lemma 3.2 depends on the directions in D being integer linear combinations of the columns of a fixed nonsingular $n \times n$ generating matrix. For example, in $\Re^1$, positive integer combinations of the columns of $D = [-1, +\pi]$ are a dense subset of the real line. This is not a counterexample to Lemma 3.2, because the matrix $[-1, +\pi]$ cannot be written as a scalar multiple of a 1×2 integer matrix.

The next lemma shows that the mesh size parameters generated by the algorithm are bounded above. (It is similar to a result in [2] for categorical variables.)

LEMMA 3.3. *Under assumptions* A1 *and* A3*, there exists a positive integer r^+ such that $\Delta_k \leq \Delta_0 \tau^{r^+}$ for any integer $k \geq 0$.*

Proof. Using assumption A3, we let $\mathcal{X}$ be a compact set in $\Re^n$ that contains all iterates, and denote its diameter by γ (i.e., the maximal distance between two of its points). If $\Delta_k > \gamma \cdot \|G^{-1}\|$, then Lemma 3.2 with $(v = x_k)$ ensures that any trial point $u \in M_k$ different from x_k would have been outside of $\mathcal{X}$. But since no iterate is outside

$\mathcal{X}$, it follows that at any iteration whose mesh size parameter exceeds $\gamma \cdot \|G^{-1}\|$, the iterate x_k is a mesh local optimizer. Thus Δ_k is bounded above by $\gamma \cdot \|G^{-1}\| \tau^{w^+}$, and the result follows by setting r^+ large enough so that $\Delta_0 \tau^{r^+} \geq \gamma \cdot \|G^{-1}\| \tau^{w^+}$. ∎

The proof of the next result is identical in spirit to that of the same result in Torczon [29] and that adapted in [2] for categorical variables.

PROPOSITION 3.4. *Under assumptions* A1 *and* A3*, the mesh size parameters satisfy* $\liminf_{k \to +\infty} \Delta_k = 0$.

Proof. Suppose, by way of contradiction, that there exists a negative integer ρ such that $0 < \Delta_0 \tau^\rho \leq \Delta_k$ for all $k \geq 0$. Combining (2.3) with Lemma 3.3 implies that for any $k \geq 0$, r_k takes its value among the integers of the finite set $\{\rho, \rho+1, \ldots, r^+\}$.

Since $x_{k+1} \in M_k$, (2.4) assures that $x_{k+1} = x_k + \Delta_k D z_k$ for some $z_k \in \mathcal{Z}_+^{|D|}$. Using (2.3) by substituting $\Delta_k = \Delta_0 \tau^{r_k}$, it follows that for any integer $N \geq 1$

$$x_N = x_0 + \sum_{k=0}^{N-1} \Delta_k D z_k \;=\; x_0 + \Delta_0 D \sum_{k=0}^{N-1} \tau^{r_k} z_k \;=\; x_0 + \frac{p^\rho}{q^{r^+}} \Delta_0 D \sum_{k=0}^{N-1} p^{r_k - \rho} q^{r^+ - r_k} z_k,$$

where p and q are relatively prime integers satisfying $\tau = p/q$. Since for any k the term $p^{r_k - \rho} q^{r^+ - r_k} z_k$ appearing in this last sum is an integer, it follows that all iterates lie on the translated integer lattice generated by x_0 and the columns of $p^\rho / q^{r^+} \Delta_0 D$.

Therefore, since all iterates belong to a compact set, it follows that there are only finitely many different iterates, and thus one of them must be visited infinitely many times. Therefore the rule presented in (2.2) is applied only finitely many times, and the one in (2.1) is applied infinitely many times. This contradicts the hypothesis that $\Delta_0 \tau^\rho$ is a lower bound for the mesh size parameter. ∎

3.3. Main convergence result. Since the mesh size parameter shrinks only when a mesh local optimizer is detected, Proposition 3.4 guarantees that there are infinitely many mesh local optimizers. The following definition specifies the subsequences we use.

DEFINITION 3.5. *A subsequence of the GPS iterates consisting of mesh local optimizers,* $\{x_k\}_{k \in K}$ *(for some subset of indices* K*), is said to be a* refining subsequence *if* $\{\Delta_k\}_{k \in K}$ *converges to zero.*

The following shows the existence of convergent refining subsequences. Notice that if coarsening of the mesh were not allowed (i.e., w^+ were set at 0 in (2.2)), then every subsequence of mesh local optimizers would be a refining subsequence, and so the next result would be trivial.

THEOREM 3.6. *Under assumptions* A1 *and* A3*, there exists at least one convergent refining subsequence.*

Proof. Let K'' be the set of indices of iterates that are mesh local optimizers. Since the mesh is refined only at iterations when a local mesh optimizer is detected, Proposition 3.4 guarantees that there exists a subset of indices $K' \subset K''$ for which $\{\Delta_k\}_{k \in K'} \downarrow 0$. Assumption A3 ensures that there exists a subset of indices $K \subset K'$ for which the subsequence of iterates $\{x_k\}_{k \in K}$ converges. ∎

We show below that the limit of any refining subsequence satisfies directional first order optimality conditions appropriate to the local smoothness of f. It is shown in [1] that, even for a continuously differentiable f, the entire iteration sequence might not converge. There may even be infinitely many limit points, and not all of these limit points are stationary points.

Next is our basic, but key, result in which we apply Clarke's [9] generalized directional derivatives in a very straightforward way to the pattern search analysis. The

results that follow specialize this result. Clarke's derivative at $\hat{x}$ in the direction d is defined for locally Lipschitz functions. Loosely speaking, it is defined to be the limit superior of the directional derivatives (in the direction d) of sequences converging to $\hat{x}$. The precise definition is given in the proof (see (3.1)).

THEOREM 3.7. *Under assumptions* A1–A3*, if $\hat{x}$ is any limit of a refining subsequence, if d is any direction in D for which f at a* POLL *step was evaluated for infinitely many iterates in the subsequence, and if f is Lipschitz near $\hat{x}$, then the generalized directional derivative of f at $\hat{x}$ in the direction d is nonnegative, i.e., $f^\circ(\hat{x}; d) \geq 0$.*

Proof. Let $\{x_k\}_{k \in K}$ be a refining subsequence and $\hat{x}$ its limit point obtained as in the statement of the Theorem. Since f is locally Lipschitz near $\hat{x}$, we have from Clarke [9] by definition that

$$f^\circ(\hat{x}; d) \equiv \limsup_{y \to \hat{x},\ t \downarrow 0} \frac{f(y + td) - f(y)}{t} \ \geq \ \limsup_{k \in K} \frac{f(x_k + \Delta_k d) - f(x_k)}{\Delta_k} \ . \tag{3.1}$$

We need to know that the difference quotients are defined. First note that since f is Lipschitz near $\hat{x}$, it must be finite near $\hat{x}$. Note also that since a main point of the paper is to allow for extended-valued functions and to justify the expedient of dealing with constraints by declining to evaluate the function f at infeasible points, we made the hypothesis that f was actually evaluated infinitely many times in the direction d. Therefore, for k sufficiently large all the POLL steps in the direction d, $x_k + \Delta_k d$ are feasible. If they had not been, then f_Ω would have been infinite there, and so f would not have been evaluated. (Recall that if $x \notin \Omega$, then $f_\Omega(x)$ is set at $+\infty$ and $f(x)$ is not evaluated.)

Thus, we have that infinitely many of the right-hand quotients of (3.1) are defined, and in fact they are the same as for f_Ω. But since they are defined, all of them must be nonnegative or else the corresponding POLL step would have been successful in identifying an improved mesh point. (Recall that refining subsequences are constructed from mesh local optimizers.) ∎

In the unconstrained case, there will always be a positive spanning set of directions that satisfy the hypotheses of the previous theorem. In the constrained case, there may be no such d if D is defined in a way incompatible with the geometry of the constraints. (See the example in [23].) Thus in the next section, we will appeal to the construction in [25] to ensure that a sufficiently rich set of directions is used for bound or linear constraints. Again, we emphasize that GPS is a directional method, and the choice of directions is crucial.

The following example illustrates Theorem 3.7 on a Lipschitz function. This function looks like a convex function (quadratic, in fact) that has been contaminated by local noise that decreases in amplitude near the minimizer. This behavior is common enough in practice to be the target class for implicit filtering algorithms [19].

Example 3.8. Consider the function $f : \Re \to \Re$ defined as $f(x) = x^2(2 + \sin(\frac{\pi}{x}))$. This function possesses infinitely many local optima near 0. One can show that f is Lipschitz near 0, but it is not strictly differentiable there, and so certainly it is not continuously differentiable. In fact, the generalized gradient satisfies $\partial f(0) = [-\pi, \pi]$.

If the GPS algorithm with empty SEARCH steps, $x_0 = \frac{1}{3}$, $\Delta_0 = 1$, $D = \{-1, 1\}$, $\Delta_{k+1} = \Delta_k$ when an improved mesh point is found, and $\Delta_{k+1} = \frac{1}{2}\Delta_k$ when a mesh local optimizer is detected, is applied to this problem, then the sequence of iterates $\{x_k\}$ converges to 0, where $f^\circ(0; \pm 1) = \pi \geq 0$, as Theorem 3.7 guarantees. The proof of this claim can be seen from Table 3.1.

Theorem 3.7 is the key to our analysis. Its proof follows so directly from Clarke's

TABLE 3.1

In four consecutive iterations, the iterates go from $x_k = 1/\alpha, \Delta_k = 3/\alpha$, where α is a positive integer, to $x_{k+4} = x_k/4, \Delta_{k+4} = \Delta_k/4$.

k	x_k	$f(x_k)$	Δ_k	$f(x_k - \Delta_k)$	$f(x_k + \Delta_k)$	Iteration status
$4i$	$\frac{1}{\alpha}$	$\frac{2}{\alpha^2}$	$\frac{3}{\alpha}$	$f(\frac{1-3}{\alpha}) \geq \frac{4}{\alpha^2}$	$f(\frac{1+3}{\alpha}) \geq \frac{16}{\alpha^2}$	mesh local optimizer
$4i+1$	$\frac{1}{\alpha}$	$\frac{2}{\alpha^2}$	$\frac{3}{2\alpha}$	$f(\frac{2-3}{2\alpha}) = \frac{1}{2\alpha^2}$	$f(\frac{2+3}{2\alpha}) \geq \frac{25}{4\alpha^2}$	improved mesh point
$4i+2$	$\frac{-1}{2\alpha}$	$\frac{1}{2\alpha^2}$	$\frac{3}{2\alpha}$	$f(\frac{-1-3}{2\alpha}) \geq \frac{4}{\alpha^2}$	$f(\frac{-1+3}{2\alpha}) = \frac{2}{\alpha^2}$	mesh local optimizer
$4i+3$	$\frac{-1}{2\alpha}$	$\frac{1}{2\alpha^2}$	$\frac{3}{4\alpha}$	$f(\frac{-2-3}{4\alpha}) \geq \frac{25}{16\alpha^2}$	$f(\frac{-2+3}{4\alpha}) = \frac{1}{8\alpha^2}$	improved mesh point
$4(i+1)$	$\frac{1}{4\alpha}$	$\frac{1}{8\alpha^2}$	$\frac{3}{4\alpha}$			

definition of the generalized directional derivative because unsuccessful polling at mesh local optimizers belonging to convergent refining sequences provides exactly the nonnegative difference quotients that Clarke's derivatives need since $x_k \to \hat{x}$ and $\Delta_k \downarrow 0$. We believe that this illustrates an intimate relationship between Clarke's generalized directional derivatives and the directional algorithm GPS.

3.4. Corollaries for unconstrained optimization. Before we add the complication of choosing directions for linear constraints, we give some corollaries of Theorem 3.7 for the unconstrained case. In addition to the assumption that f is Lipschitz near $\hat{x}$, we assume that the generalized gradient of f at $\hat{x}$ is a singleton. This is equivalent to assuming that f is strictly differentiable at $\hat{x}$, i.e., that there exists a $D_s f(\hat{x}) \in \Re^n$ such that $\lim_{y \to \hat{x}, t \downarrow 0} \frac{f(y+tw)-f(y)}{t} = D_s f(\hat{x})^T w$ for all $w \in \Re^n$ (see [9, Proposition 2.2.1 or Proposition 2.2.4]). Since the generalized gradient is a singleton $\partial f(\hat{x}) = \{D_s f(\hat{x})\}$, we use the standard notation for the gradient $\nabla f(\hat{x}) = D_s f(\hat{x})$.

THEOREM 3.9. *Under assumptions* A1 *and* A3*, let $\Omega = \Re^n$ and let $\hat{x}$ be any limit of a refining subsequence. If f is strictly differentiable at $\hat{x}$, then $\nabla f(\hat{x}) = 0$.*

Proof. Again from [9], if f is strictly differentiable at $\hat{x}$, then for any direction $w \neq 0$, $f^\circ(\hat{x}; w) = \nabla f(\hat{x})^T w$. Now let $\hat{D}$ be any positive spanning set that is used infinitely many times in the refining subsequence; there must be at least one since D is finite. Then by Theorem 3.7, for each $d \in \hat{D}$, $0 \leq \nabla f(\hat{x})^T d$. Thus, if we write w as a nonnegative linear combination of the elements of $\hat{D}$, then we see immediately that $\nabla f(\hat{x})^T w \geq 0$. However, the same construction for $-w$ shows that $-\nabla f(\hat{x})^T w \geq 0$ and so $\nabla f(\hat{x}) = 0$. ∎

The following example, based on a function taken from [20], illustrates the applicability of Theorem 3.9 by showing that any realization of GPS converges to the global minimizer for this convex function, which is strictly differentiable at its minimizer but not continuously differentiable. Previous GPS analysis techniques that use global continuous differentiability do not apply to this example.

Example 3.10. Consider the convex function $f : \Re \to \Re$ defined as $f(x) = \int_0^x \varphi(u) du$, where

$$\varphi(u) = \begin{cases} u & \text{if } u \leq 0, \\ \frac{1}{1+\kappa} & \text{if } \kappa + 1 > \frac{1}{u} \geq \kappa \in \mathcal{Z}_+. \end{cases}$$

The function f is Lipschitz near $\hat{x} = 0$. It is shown in [20] that f has kinks at $\frac{1}{\kappa}$ with $\partial f(\frac{1}{\kappa}) = [\frac{1}{\kappa+1}, \frac{1}{\kappa}]$ for $\kappa = 1, 2, \ldots$. The corollary of Proposition 2.2.4 in [9] guarantees that f is not continuously differentiable near $\hat{x}$. Furthermore, $\partial f(0)$ reduces to the singleton $\{0\}$, and the same Proposition ensures that f is strictly differentiable at $\hat{x}$.

Applying Theorem 3.9 guarantees that any instance of any pattern search algorithm with any set of initial parameters generates a subsequence of iterates that converges to the global minimizer $\hat{x} = 0$, where $\nabla f(\hat{x}) = 0$, since the function is Lipschitz everywhere, and 0 is the only point at which Clarke's generalized derivatives are nonnegative in all directions of a positive spanning set.

We certainly are not claiming that the weaker smoothness conditions that we use imply that GPS methods *always* find a minimizer. This has been known to be false since the inception of GPS methods. Simple convex counterexamples come from starting at just the wrong point and choosing just the right ill-suited directions.

This can be seen by considering $f(x) = |x_1| + |x_2|$ on $\Re^2$ and starting with $x_0 = (1, 0)^T$ with $D = \{(1, 0)^T, (-1, 1)^T, (-1, -1)^T\}$. The initial point x_0 is a mesh local optimizer for every $\Delta > 0$, and so the iteration never moves from x_0 with an empty SEARCH step. Our theorem applies to this simple example and describes exactly what happens; f is regular at $\hat{x}$, and the directional derivatives along the members of D are nonnegative.

The following two corollaries assume continuous differentiability. We have discussed how, for our applications, this assumption is unlikely to be satisfied, except perhaps locally. We include these results only to tie our results here to earlier results that use global continuous differentiability. The first corollary strengthens our result in [2]. It shows that the limit of the gradient for any refining subsequence converges to zero, even if the subsequence itself does not converge.

COROLLARY 3.11. *Let* A1 *and* A3 *hold for* $\Omega = \Re^n$ *and* f *continuously differentiable on a neighborhood of a compact set containing all the iterates* $\{x_k\}$. *Then for any refining subsequence* $\{x_k\}_{k \in K}$, $0 = \lim_{k \in K} \nabla f(x_k)$.

Proof. If $\hat{x}$ is any limit point of a refining subsequence, then continuous differentiability implies strict differentiability at $\hat{x}$, and so $\nabla f(\hat{x}) = 0$ from Theorem 3.9. Since the continuous image of a compact set is compact, the entire sequence of gradients of any refining subsequence is in a compact set. Thus, there must be a subsequence $\{x_k\}_{k \in K'}$ of the refining subsequence for which $\lim_{k \in K'} \nabla f(x_k) = \limsup_k \nabla f(x_k)$. But then $\{x_k\}_{k \in K'}$ has a convergent subsequence, and its limit point has a zero gradient because it is a limit point of a refining subsequence, and so $0 = \limsup_k \nabla f(x_k)$. ∎

A consequence of the previous result is that, under the assumption that f is continuously differentiable, any limit point of a refining sequence has a zero gradient.

The fact that under the assumption of continuous differentiability the limit of the gradients of any refining subsequence is zero was pointed out in [15]. Earlier, under strong restrictions on the algorithm, it was shown in [29] that $0 = \lim_k \nabla f(x_k)$. One of those restrictions is that $\lim \Delta_k = 0$, which we proved above already is enough to say that the limit of the gradients at the mesh local optimizers is zero since then they are a refining subsequence. Thus, we will not discuss the restrictions needed for the stronger result, since they are too constraining for our class of problems.

The next corollary is Torczon's result from [29], strengthened by the same result from [15].

COROLLARY 3.12. *Let* A1 *and* A3 *hold for* $\Omega = \Re^n$, *and let* f *be continuously differentiable on a neighborhood of a compact set containing all the iterates* $\{x_k\}$; *then some limit point* $\hat{x}$ *of* $\{x_k\}$ *satisfies* $\nabla f(\hat{x}) = 0$. *The limit of the gradients for any refining subsequence is zero.*

Proof. Every refining subsequence is a subsequence of $\{x_k\}$. ∎

In summary, if assumptions A1 and A3 are satisfied, then the algorithm guaran-

tees the following hierarchy of convergence behavior:

(i) If f is lower semicontinuous at any limit point $\bar{x}$ of the GPS iteration sequence, then Theorem 3.1 says that $f(\bar{x}) \leq \lim_k f(x_k)$.

(ii) Every limit point of the iteration sequence at which f is continuous has the same function value $\lim_k f(x_k)$, whether or not it is a stationary point. Thus, although there is always at least one limit point that is a stationary point, if GPS produces a nonstationary limit point [1], then it is more promising than any stationary limit point because they have the same function value, but there is a descent direction from the nonstationary limit point. The conclusion is that the directions used were poorly suited to the problem.

(iii) There is at least one $\hat{x}$ that is a limit point of a refining subsequence; i.e., $\hat{x}$ is a limit point of a sequence of local optimizers on meshes that get infinitely fine. If the function f is lower semicontinuous but not even Lipschitz near $\hat{x}$, then nothing additional to the above is claimed about optimality conditions satisfied by $\hat{x}$.

(iv) If f is Lipschitz near $\hat{x}$, then Theorem 3.7 holds and Clarke's generalized derivatives satisfy $f^\circ(\hat{x}; d) \geq 0$ for some directions $d \in D$ that form a positive spanning set. In addition, $f(\hat{x}) = \lim_k f(x_k)$ since f is continuous at $\hat{x}$.

(v) If f is regular[1] at $\hat{x}$, then the directional derivatives satisfy $f'(\hat{x}; d) \geq 0$ for some directions $d \in D$, a positive spanning set, and $f(\hat{x}) = \lim_k f(x_k)$.

(vi) If f is strictly differentiable at $\hat{x}$, then Theorem 3.9 holds and $\nabla f(\hat{x}) = 0$, but its function value $\lim_k f(x_k)$ is the same as at any other limit point of the entire GPS iteration sequence at which f is continuous (by (ii)).

(vii) If f is globally continuously differentiable (as assumed in earlier analyses), this means that every limit point of a refining subsequence is a stationary point as in item (vi) and that the gradients of a refining subsequence converge to zero, whether or not the subsequence converges. However, as was shown in [1], there still can be limit points of the entire GPS iteration sequence that are not stationary points. Though such points have the same function value as the stationary points, there is a descent direction from such points that leads to lower function values.

3.5. Linearly constrained convergence results. In this section, we will consider only the case in which Ω is defined through a finite set of linear constraints. In order to prove the relevant optimality results, we will have to assume that D, even though finite, is rich enough to generate POLL sets that conform to the geometry of the boundary of Ω. Furthermore, to apply our proof technique, we must ensure that the spanning sets that reflect this geometry get used infinitely many times as we converge to a point on the boundary. Lewis and Torczon [25] show how to use standard linear algebra tools to generate the requisite positive spanning matrices $D_k \subseteq D$. The convergence analysis relies on assumption A2, the rationality of the constraint matrix A.

We pause to remind the reader that, for $x \in \Omega$, the tangent cone to Ω at x is $T_\Omega(x) = \text{cl}\{\mu(w - x) : \mu \geq 0, w \in \Omega\}$. The normal cone to Ω at x is $N_\Omega(x)$ and can be written as the polar of the tangent cone: $N_\Omega(x) = \{v \in \Re^n : \forall w \in T_\Omega(x),\ v^T w \leq 0\}$. It is the nonnegative span of all the outwardly pointing constraint normals at x.

It would add unnecessary length to this paper to rewrite the construction given

[1] The function f is said to be *regular* at x if, for all v, the one-sided directional derivative exists and coincides with $f^\circ(x; v)$ (see Clarke [9]).

by Lewis and Torczon [25] for D and the choice rule for D_k from D at each iteration (their notation for D_k is Γ_k). The construction is presented there quite succinctly in section 8 of [25] where they consider implementation issues, including difficulties inherent to degenerate constraints. We will use the following abstracted version of their direction choice.

DEFINITION 3.13. *A rule for selecting the positive spanning sets $D_k = D(k, x_k) \subseteq D$* conforms to Ω *for some $\epsilon > 0$ if, at each iteration k and for each y in the boundary of Ω for which $\|y - x_k\| < \epsilon$, $T_\Omega(y)$ is generated by nonnegative linear combinations of the columns of a subset D_k^y of D_k.*

With this definition, we are ready for our next convergence result. Note that if $x_k \in \Omega$ is not near the boundary, then D_k need only provide a positive spanning set for $\Re^n$, which is completely sensible. However, in our experience, it is best not to take ϵ too small so that when the iterates approach the boundary with small values of the mesh size parameter, the rule for selecting the mesh size parameter conforms to the boundary of Ω. This is mitigated somewhat by allowing variable coarsening of the mesh as in (2.2).

THEOREM 3.14. *Under assumptions* A1–A3*, if f is strictly differentiable at a limit point $\hat{x}$ of a refining subsequence and if the rule for selecting the positive spanning sets $D_k = D(k, x_k) \subseteq D$ conforms to Ω for an $\epsilon > 0$, then $\nabla f(\hat{x})^T w \geq 0$ for all $w \in T_\Omega(\hat{x})$ and $-\nabla f(\hat{x}) \in N_\Omega(\hat{x})$. Thus, $\hat{x}$ is a KKT point.*

Proof. If $\hat{x}$ is interior to Ω, then the result is just Theorem 3.9, and thus we can proceed directly to the case in which $\hat{x}$ is on the boundary of Ω.

Suppose that the rule for selecting $D_k \subseteq D$ conforms to Ω for some fixed $\epsilon > 0$ and that there are finitely many linear constraints; then $D_k^{\hat{x}}$ generates $T_\Omega(\hat{x})$ for large $k \in K$. It follows that there can be only finitely many different such sets $D_k^{\hat{x}}$ for $k \in K$. Let $D^{\hat{x}} \subseteq D$ be one of them that occurs infinitely many times.

Theorem 3.7 implies that $\nabla f(\hat{x})^T d \geq 0$ for every column d of $D^{\hat{x}}$. But since every $w \in T_\Omega(\hat{x})$ is a nonnegative linear combination of the columns of $D^{\hat{x}}$, then $\nabla f(\hat{x})^T w \geq 0$. To complete the proof, we multiply both sides by -1 and conclude that $-\nabla f(\hat{x})$ is in $N_\Omega(\hat{x})$. ∎

Remark 3.15. If f were only assumed to be Lipschitz near $\hat{x}$, then we could still conclude, as in Theorem 3.7, that $f^\circ(\hat{x}; d) \geq 0$ for every column d of $D^{\hat{x}}$.

The following corollary is Lewis and Torczon's result from [25], which relies on a stronger differentiability assumption.

COROLLARY 3.16. *If* A1–A3 *hold and f is continuously differentiable on a neighborhood of a compact set containing all the iterates $\{x_k\}$, and if the rule for selecting the positive spanning sets $D_k = D(k, x_k) \subseteq D$ conforms to Ω for an $\epsilon > 0$, then there exists a limit point $\hat{x}$ of $\{x_k\}$ such that $\nabla f(\hat{x})^T w \geq 0$ for all $w \in T_\Omega(\hat{x})$ and $-\nabla f(\hat{x}) \in N_\Omega(\hat{x})$. Thus, $\hat{x}$ is a KKT point.*

Proof. The proof follows from Theorem 3.14, since every refining subsequence is a subsequence of $\{x_k\}$ and continuous differentiability implies strict differentiability. ∎

4. Concluding remarks. This paper puts together ways to choose the directions and results on properties of the mesh by Lewis and Torczon, some observations of ours about what is needed to obtain convergence of those algorithms (such as refining subsequences), and elements of nonsmooth analysis set forth by Clarke. Clarke's analysis is perfectly suited to exposing the first order optimality conditions at limit points of certain subsequences of the GPS iterates under weakened assumptions that correspond to some real problems for which GPS is quite effective.

We believe that our analysis helps confirm an observation of [25] that GPS methods for general constraints will not be based on the appealingly simple barrier strategy of placing a high function value on infeasible trial points. This is because, to prove the efficacy of the barrier strategy, the positive spanning set D, from which all the GPS directions are chosen, is finite, and thus it cannot be certain to generate the tangent cone at every boundary point of a nonpolygonal feasible region that the iteration approaches.

In [3], we suggest and analyze a GPS algorithm for general constraints, based not on a single objective but on the new filter approach of Fletcher and collaborators [16, 17, 18]. In [26], Lewis and Torczon give a successive augmented Lagrangian pattern search approach together with its convergence analysis. Ongoing work by Coope and Price along the lines of [12] and [13] promises alternatives for general constraints yet to be realized.

Acknowledgments. We wish to acknowledge a helpful referee and Major Mark Abramson, USAF, for many insightful comments that improved the presentation of this work.

REFERENCES

[1] C. Audet, *Convergence Results for Pattern Search Algorithms Are Tight*, Technical report TR98-24, Department of Computational and Applied Mathematics, Rice University, Houston, TX, 1998.

[2] C. Audet and J.E. Dennis, Jr., *Pattern search algorithms for mixed variable programming*, SIAM J. Optim., 11 (2000), pp. 573–594.

[3] C. Audet and J.E. Dennis, Jr., *A Pattern Search Filter Method for Nonlinear Programming without Derivatives*, Technical report TR00-09, Department of Computational and Applied Mathematics, Rice University, Houston, TX, 2000.

[4] C. Audet, A.J. Booker, J.E. Dennis, Jr., P.D. Frank, and D. Moore, *A Surrogate-Model-Based Method For Constrained Optimization*, AIAA 2000-4891, in Proceedings of the 8th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization, Long Beach, CA, 2000.

[5] A.J. Booker, J.E. Dennis, Jr., P.D. Frank, D.B. Serafini, V. Torczon, and M.W. Trosset, *A rigorous framework for optimization of expensive functions by surrogates*, Structural Optim., 17 (1999), pp. 1–13.

[6] A.J. Booker, J.E. Dennis, Jr., P.D. Frank, D.W. Moore, and D.B. Serafini (1999), *Managing Surrogate Objectives to Optimize a Helicopter Rotor Design—Further Experiments*, AIAA Paper 98-4717, in Proceedings of the 7th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization, St. Louis, MO, 1998.

[7] T.D. Choi, O.J. Eslinger, C.T. Kelley, J.W. David, and M. Etheridge, *Optimization of automotive valve train components with implicit filtering*, Optim. Engrg., 1 (2000), pp. 9–27.

[8] T.D. Choi and C.T. Kelley, *Superlinear convergence and implicit filtering*, SIAM J. Optim., 10 (1999), pp. 1149–1162.

[9] F.H. Clarke, *Optimization and Nonsmooth Analysis*, SIAM Classics in Appl. Math. 5, SIAM, Philadelphia, 1990.

[10] 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.

[11] I.D. Coope and C.J. Price, *On the convergence of grid-based methods for unconstrained optimization*, SIAM J. Optim., 11 (2001), pp. 859–869.

[12] I.D. Coope and C.J. Price, *Positive Bases in Optimization*, Report UCDMS2000/12, Department of Mathematics and Statistics, University of Canterbury, Canterbury, UK, 2000.

[13] I.D. Coope and C.J. Price, *Frames and Grids in Unconstrained and Linearly Constrained Optimization: A Non-Smooth Approach*, Report UCDMS2002/1, Department of Mathematics and Statistics, University of Canterbury, Canterbury, UK, 2002.

[14] J.E. Dennis and V. Torczon, *Direct search methods on parallel machines*, SIAM J. Optim., 1 (1991), pp. 448–474.

[15] E. Dolan, M. Lewis, and V. Torczon, *On the local convergence of pattern search*, ICASE Technical report 2000-36, NASA Langley Research Center, Hampton, VA, 2000.
[16] R. Fletcher and S. Leyffer, *Nonlinear programming without a penalty function*, Math. Program., 91 (2002), pp. 239–269.
[17] R. Fletcher, S. Leyffer, and Ph.L. Toint, *On the global convergence of a filter–SQP algorithm*, SIAM J. Optim., 13 (2002), pp. 44–59.
[18] R. Fletcher, N.I.M. Gould, S. Leyffer, Ph.L. Toint, and A. Wachter, *Global convergence of a trust-region SQP-filter algorithm for general nonlinear programming*, SIAM J. Optim., 13 (2002), pp. 635–659.
[19] P. Gilmore and C.T. Kelley, *An implicit filtering algorithm for optimization of functions with many local minima*, SIAM J. Optim., 5 (1995), pp. 269–285.
[20] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms*, Springer-Verlag, Berlin, New York, 1993.
[21] M. Kokkolaras, C. Audet, and J.E. Dennis, Jr., *Mixed variable optimization of the number and composition of heat intercepts in a thermal insulation system*, Optim. Engrg., 2 (2001), pp. 5–29.
[22] C-J. Lin and J.J. Moré, *Newton's method for large bound-constrained optimization problems*, SIAM J. Optim., 9 (1999), pp. 1100–1127.
[23] R.M. Lewis and V. Torczon, *Pattern search algorithms for bound constrained minimization*, SIAM J. Optim., 9 (1999), pp. 1082–1099.
[24] R.M. Lewis and V. Torczon, *Rank Ordering and Positive Basis in Pattern Search Algorithms*, ICASE Technical report TR 96-71, NASA Langley Research Center, Hampton, VA, 1996.
[25] R.M. Lewis and V. Torczon, *Pattern search methods for linearly constrained minimization*, SIAM J. Optim., 10 (2000), pp. 917–941.
[26] 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.
[27] J. Nocedal and S.J. Wright, *Numerical Optimization*, Springer Ser. Oper. Res., Springer, NY, 1999.
[28] D. Serafini, *A Framework for Managing Models in Nonlinear Optimization of Computationally Expensive Functions*, Ph.D. Thesis, Rice University, Houston, TX, 1999.
[29] V. Torczon, *On the convergence of pattern search algorithms*, SIAM J. Optim., 7 (1997), pp. 1–25.

SIAM J. OPTIM.
Vol. 13, No. 3, pp. 904–920

SHARP VARIATIONAL CONDITIONS FOR CONVEX COMPOSITE NONSMOOTH FUNCTIONS*

V. JEYAKUMAR† AND D. T. LUC‡

Abstract. In this paper, we present first- and second-order variational conditions for a convex composite function $g \circ F$, where g is a nonsmooth convex function and F is a vector-valued map. The first-order results, which apply to (not necessarily locally Lipschitz) continuous maps F, not only recapture the results of the special cases where F is locally Lipschitz or Gâteaux differentiable but also yield sharp necessary variational conditions in these cases. The results are achieved by applying a new strengthened notion of approximate Jacobian, called a Gâteaux (G-) approximate Jacobian, without the use of the upper semicontinuity of the approximate Jacobian. These variational results are generally derived by using a chain rule formula or by constructing upper convex approximations to the composite function. These approaches often need the upper semicontinuity requirement of a generalized Jacobian map. Such a requirement not only limits the derivation of sharp optimality conditions, as the "small" approximate Jacobians (or generalized subdifferentials) lack an upper semi-continuity property, but also restricts the treatment of Gâteaux differentiable maps F. This situation is overcome by the use of G-approximate Jacobians. The second-order variational conditions are shown to hold, in particular, in the case where F is continuously Gâteaux differentiable.

Key words. Gâteaux approximate Jacobians, sharp necessary conditions, convex composite functions, second-order variational conditions

AMS subject classifications. 49A52, 90C30, 26A24

PII. S1052623401396509

1. Introduction.

Consider the convex composite function $g \circ F$, where $F : \mathbb{R}^n \to \mathbb{R}^m$ is a vector-valued continuous map and $g : \mathbb{R}^m \to \mathbb{R}$ is a nonsmooth convex function. The convex composite model functions $g \circ F$, which arise in a variety of practical optimization problems, have been extensively studied in the literature (see [7, 11, 12, 18, 19, 23, 27, 28]). The new class of convex composite continuous functions where F is continuous appear in the form of the norm, $||F(x)||$, when solving nonlinear equations $F_i(x) = 0, \quad i = 1, 2, \ldots, m$, of continuous functions. They also arise in the form of the max function, $\max(F_i, 0)$, when finding a feasible point of a system of continuous nonlinear inequalities $F_i(x) \leq 0, \quad i = 1, 2, \ldots, m$. Recently, it has been demonstrated that convex composite continuous functions play an important role in the study of spectral functions such as the spectral abscissa and spectral radius [2], which are convex composite continuous functions but are not locally Lipschitzian. Variational analysis of such composite functions are of great interest in control theory and related areas. A variant of the nonsmooth composite model function $g \circ F$ where g is differentiable and F is continuous also comes to light in the optimization reformulation of complementarity problems (see [6]).

The first-order variational conditions for the function $g \circ F$ are well known in the cases where F is either continuously differentiable [1] or F is locally Lipschitzian [4]. In these cases, the conditions have been obtained by applying a chain rule for-

*Received by the editors October 17, 2001; accepted for publication (in revised form) August 22, 2002; published electronically February 27, 2003. This research was partially supported by a grant from the Australian Research Council.

http://www.siam.org/journals/siopt/13-3/39650.html

†Department of Applied Mathematics, University of New South Wales, Sydney 2052, Australia (jeya@maths.unsw.edu.au).

‡Departement de Mathématiques, Université d'Avignon, 33 Rue Louis Pasteur, 8400 Avignon, France (dtluc@univ-avignon.fr).

mula for the Clarke generalized Jacobian of F. These situations, where F is either locally Lipschitzian and Gâteaux differentiable [12] or F is (not necessarily Lipschitz) Gâteaux differentiable [25], were separately treated by constructing suitable upper convex approximations to the composite functions $g \circ F$, as the Clarke generalized Jacobian-based chain rules are not flexible enough for handling Gâteaux differentiability. The approximate Jacobians (see [13, 16, 14]) are flexible tools for studying (not necessarily locally Lipschitz) continuous functions, and they yield sharp optimality conditions for locally Lipschitz functions [26]. However, the development of a chain rule formula for composite functions requires the upper semicontinuity of an approximate Jacobian map (see [6, 15]). Such a requirement not only limits the derivation of sharp optimality conditions (see [26]), as the "small" generalized subdifferentials or approximate Jacobians lack an upper semicontinuity property, but also restricts the treatment of Gâteaux differentiable maps F. To overcome this situation and to obtain variational conditions for the convex composite function $g \circ F$, where F is neither locally Lipschitzian nor Gâteaux differentiable, we introduce a slightly stronger version of the approximate Jacobian, called a Gâteaux (G-) approximate Jacobian. Both the locally Lipschitz maps and the Gâteaux differentiable maps admit such a G-approximate Jacobian, and so the composite function in these cases enjoy sharp first-order variational conditions.

On the other hand, second-order variational conditions have so far been given in the cases where F is either twice continuously differentiable [1, 3] or continuously differentiable with locally Lipschitz derivatives (i.e., $C^{1,1}$) [18, 19]. We present second-order variational conditions in terms of G-approximate Jacobians for the composite function $g \circ F$, in particular, for the case where F is continuously Gâteaux differentiable, but the derivative ∇F is not necessarily locally Lipschitz. These conditions are obtained by extending the approach of Burke [1], Burke and Poliquin [3], and Ioffe [10] and using approximate Hessians.

The outline of the paper is as follows. In section 2, we present definitions and summarize the calculus of approximate Jacobians and introduce Gâteaux approximate Jacobians and their connections to approximate Jacobians. In section 3, we present constructions of approximate Jacobians for the composite function $g \circ F$ without the upper semicontinuity of the approximate Jacobians and subsequently derive first-order variational conditions. Finally, in section 4, we present general second-order conditions for the composite function under much reduced smoothness condition on F.

2. G-approximate Jacobians. We begin this section with some background material on approximate Jacobians for nonsmooth maps and then present a new notion, called G-approximate Jacobians, which are admitted, in particular, by both locally Lipschitz maps and Gâteaux differentiable maps. Let $F : \mathbb{R}^n \to \mathbb{R}^m$. For each $v \in \mathbb{R}^m$ the composite function, $(vF) : \mathbb{R}^n \to \mathbb{R}$, is defined by

$$(vF)(x) = \langle v,\ F(x)\rangle = v^T F(x),$$

where $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product. The upper Dini directional derivative of (vF) at x in the direction $u \in \mathbb{R}^n$ is defined by

$$(vF)^+(x,u) := \limsup_{t \downarrow 0} \frac{(vF)(x+tu) - (vF)(x)}{t}.$$

We denote by $L(\mathbb{R}^n, \mathbb{R}^m)$ the space of all $(m \times n)$ matrices. The *convex hull* and the *closed convex hull* of a set A are denoted by $\operatorname{co} A$ and $\overline{\operatorname{co}}\, A$, respectively. Let us now

define the notion of an approximate Jacobian.

DEFINITION 2.1 (approximate Jacobian). *The map $F : \mathbb{R}^n \to \mathbb{R}^m$ admits an* approximate Jacobian*, $\partial F(x)$, at $x \in \mathbb{R}^n$ if $\partial F(x) \subseteq L(\mathbb{R}^n, \mathbb{R}^m)$ is closed and for each $v \in \mathbb{R}^m$*

$$(vF)^+(x,u) \leq \sup_{M \in \partial F(x)} \langle v, Mu \rangle \quad \forall u \in \mathbb{R}^n.$$

An element M of $\partial F(x)$ is called an *approximate Jacobian matrix* of F at x. The set-valued map $\partial F : \mathbb{R}^n \rightrightarrows L(\mathbb{R}^n, \mathbb{R}^m)$ is an approximate Jacobian of F if, for each $x \in \mathbb{R}^n$, $\partial F(x)$ is an approximate Jacobian of F at x. We allow infinite values on both sides of the above inequality.

Let us now introduce the notion of approximate Hessian for a Gâteaux differentiable function $f : \mathbb{R}^n \to \mathbb{R}$. Note that the Gâteaux derivative of f which is denoted by ∇f is a map from $\mathbb{R}^n$ to $\mathbb{R}^n$.

DEFINITION 2.2 (approximate Hessian). *The Gâteaux differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ admits an* approximate Hessian $\partial^2 f(x)$ *at x if this set is an approximate Jacobian to the derivative ∇f at x.*

Note that $\partial^2 f(x) = \partial \nabla f(x)$ and the matrix $M \in \partial^2 f(x)$ is an approximate Hessian matrix of f at x. Clearly, if f is twice differentiable at x, then $\nabla^2 f(x)$ is a symmetric approximate Hessian matrix of f at x. Let us define $\partial_B^2 f(x)$ by

$$\partial_B^2 f(x) = \left\{ M \ : \ M = \lim_{n \to \infty} \nabla^2 f(x_n),\ x_n \in \Delta,\ x_n \to x \right\},$$

where Δ is the set of points in $\mathbb{R}^n$ where f is twice differentiable.

Recall that if $f : \mathbb{R}^n \to \mathbb{R}$ is $C^{1,1}$, where f is continuously Gâteaux differentiable with the locally Lipschitz derivative ∇f, then the generalized Hessian in the sense of Hiriart-Urruty, Strodiot, and Hien Nguyen [8] is given by $\partial_H^2 f(x) = \mathrm{co}\partial_B^2 f(x)$. Clearly, $\partial_H^2 f(x)$ is a nonempty convex compact set of symmetric matrices. The second-order directional derivative of f at x in the directions $(u,v) \in \mathbb{R}^n \times \mathbb{R}^n$ is defined by

$$f^{\circ\circ}(x;u,v) = \limsup_{\substack{y \to x \\ s \to 0}} \frac{\langle \nabla f(y+su), v \rangle - \langle \nabla f(y), v \rangle}{s}.$$

Since for each $(u,v) \in \mathbb{R}^n \times \mathbb{R}^n$,

$$(v\nabla f)^+(x,u) \leq f^{\circ\circ}(x;u,v) = \max_{M \in \partial_B^2 f(x)} \langle Mu, v \rangle = \max_{M \in \partial_B^2 f(x)} \langle Mv, u \rangle,$$

$\partial_B^2 f(x)$ is an approximate Hessian of f at x.

Recall that a set-valued map $T : \mathbb{R}^n \rightrightarrows \mathbb{R}^k$ is said to be *upper semicontinuous at $x \in \mathbb{R}^n$* if for every $\epsilon > 0$ there exists $\delta > 0$ such that $T(x + \delta B_n) \subseteq T(x) + \epsilon B_k$, where B_n and B_k are the closed unit balls in $\mathbb{R}^n$ and $\mathbb{R}^k$, respectively.

We list here some elementary and useful calculus rules for approximate Jacobians without proofs for the sake of convenience. We refer the interested reader to [13, 16, 17, 14] for details, examples, and applications of approximate Jacobians and Hessians.

(i) *Nonuniqueness.* If $\partial F(x) \subseteq L(\mathbb{R}^n, \mathbb{R}^m)$ is an approximate Jacobian of F at x, then every closed subset of $L(\mathbb{R}^n, \mathbb{R}^m)$ which contains $\partial F(x)$ is an approximate Jacobian of F at x.

(ii) *Gâteaux differentiability.* If $F : \mathbb{R}^n \to \mathbb{R}^m$ is Gâteaux differentiable at x with the derivative $\nabla F(x)$, then $\{\nabla F(x)\}$ is an approximate Jacobian of F at x and

any closed set containing $\nabla F(x)$ is also an approximate Jacobian of F at x. Moreover, F is Gâteaux differentiable at x if and only if it admits a singleton approximate Jacobian at this point. On the other hand, if F is continuously differentiable at x and if $\partial F(x)$ is any approximate Jacobian of F at x, then $\nabla F(x) \in \overline{\text{co}}\, \partial F(x)$.

(iii) *Locally Lipschitzian.* Suppose that F is locally Lipschitz at $x \in \mathbb{R}^n$. Then the Clarke generalized Jacobian [5, 24]

$$\partial_C F(x) = \text{co}\left\{ \lim_{n\to\infty} \nabla F(x_n) : x_n \to x,\ \{x_n\} \subset \Omega \right\}$$

is an approximate Jacobian of F at x [13]. The set Ω is a dense set of points in $\mathbb{R}^n$ on which F is differentiable. Moreover, if the locally Lipschitz map F admits an approximate Jacobian map ∂F which is upper semicontinuous at x, then $\partial_C F(x) \subset \overline{\text{co}}\, \partial F(x)$.

(iv) *Addition.* If $F, G : \mathbb{R}^n \to \mathbb{R}^m$ and if $\partial F(x)$ and $\partial G(x)$ are approximate Jacobian of F and G at x, respectively, then the closure of the set $\partial F(x) + \partial G(x)$ is an approximate Jacobian of $F + G$ at x.

(v) *Cartesian products.* Let $F : \mathbb{R}^n \to \mathbb{R}^m$ and $G : \mathbb{R}^n \to \mathbb{R}^l$ be continuous. If $\partial F(x) \subseteq L(\mathbb{R}^n, \mathbb{R}^m)$ and $\partial G(x) \subseteq L(\mathbb{R}^n, \mathbb{R}^l)$ are approximate Jacobians of F and G at x, respectively, then $\partial F(x) \times \partial G(x)$ is an approximate Jacobian of $F \times G$ at x. In particular, if $F = (f_1, \dots, f_m)$ and $\partial f_1(x), \dots, \partial f_m(x)$ are generalized subdifferentials of the scalar component functions $f_1, \dots, f_m$ at x, respectively, then $\partial f_1(x) \times \cdots \times \partial f_m(x)$ is an approximate Jacobian of f at x.

(vi) *Extremality.* Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuous. If $\partial f(x)$ is an approximate Jacobian of f at x and if x is a local minimizer (or a maximizer) of f, then $0 \in \overline{\text{co}}\partial f(x)$.

(vii) *Generalized mean value theorem.* Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be continuous, and let ∂F be an approximate Jacobian of F. Then for each pair of points $a, b \in R^n$, one has

$$F(b) - F(a) \in\ \overline{\text{co}}(\partial F[a, b](b - a)).$$

If $f : \mathbb{R}^n \to \mathbb{R}$ is continuous and if ∂f is an approximate Jacobian of f, then there exists some $c \in (a, b)$ such that

$$f(b) - f(a) \in\ \overline{\text{co}}(\partial f(c)(b - a)).$$

(viii) *Generalized Taylor's expansion.* Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously Gâteaux differentiable on $\mathbb{R}^n$; let x, $y \in \mathbb{R}^n$. Suppose that for each $z \in [x, y]$, $\partial^2 f(z)$ is an approximate Hessian of f at z. Then there exists $\zeta \in (x, y)$ such that

$$f(y) \in f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2}\overline{\text{co}}\, \langle \partial^2 f(\zeta)(y - x), (y - x)\rangle.$$

DEFINITION 2.3 (G-approximate Jacobian [20]). *The map $F : \mathbb{R}^n \to \mathbb{R}^m$ admits a* G-approximate Jacobian $\partial F(x)$ *at $x \in \mathbb{R}^n$ if $\partial F(x) \subseteq L(\mathbb{R}^n, \mathbb{R}^m)$ is closed and if*

$$(\forall u \in R^n)\ \ (\forall t > 0)\ \ (\exists\ M_t \in \partial F(x)),\quad F(x + tu) - F(x) = M_t(tu) + \text{o}(t),$$

where $\frac{\text{o}(t)}{t} \to 0$, as $t \to 0$.

Clearly, both M_t and $\text{o}(t)$ depend on u. However, this dependence is suppressed in the above definition for notational convenience. It is immediate from the definition

that if F is Gâteaux differentiable at x with the derivative $\nabla F(x)$, then $\{\nabla F(x)\}$ is a G-approximate Jacobian of F at x. Now we see in the next proposition that every G-approximate Jacobian at a point is an approximate Jacobian at that point.

PROPOSITION 2.1. *Let $F : \mathbb{R}^n \to \mathbb{R}^m$. If $\partial F(x)$ is a G-approximate Jacobian F at $x \in \mathbb{R}^n$, then it is also an approximate Jacobian of F at x.*

Proof. Let $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^m$. Let $\{t_i\}$ be a sequence of positive numbers converging to 0 such that

$$(vF)^+(x,u) = \lim_{i\to\infty} \frac{(vF)(x+t_iu) - (vF)(x)}{t_i}.$$

Since $\partial F(x)$ is a G-approximate Jacobian of F at x, for each i there exists $M_{t_i} \in \partial F(x)$ such that

$$\frac{\langle v,\ F(x+t_iu)\rangle - \langle v,\ F(x)\rangle}{t_i} = \langle v,\ M_{t_i}u\rangle + \left\langle v,\ \frac{\mathrm{o}(t_i)}{t_i}\right\rangle.$$

Passing to the limit, we get that $\lim_{i\to\infty} \frac{\langle v,\ \mathrm{o}(t_i)\rangle}{t_i} = 0$ and

$$\begin{aligned}(vF)^+(x,u) = \lim_{i\to\infty} \frac{(vF)(x+t_iu) - (vF)(x)}{t_i} &\le \sup_{N\in\partial F(x)} \left(\langle v, Nu\rangle + \frac{\langle v,\ \mathrm{o}(t_i)\rangle}{t_i}\right)\\ &= \sup_{N\in\partial F(x)} \langle v, Nu\rangle,\end{aligned}$$

which shows that $\partial F(x)$is an approximate Jacobian of F at x. □

The following example shows that an approximate Jacobian of a map at a point is not necessarily a G-approximate Jacobian of the map at the point.

EXAMPLE 2.1. *Consider the function $F : \mathbb{R}^2 \to \mathbb{R}^2$ defined by*

$$F(x,y) = (|x| - |y|,\quad |y| - |x|)^T.$$

Then F is locally Lipschitz. It is simple to verify that the set

$$\partial F(0) = \left\{ \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix},\ \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \right\}$$

is an approximate Jacobian of F at 0. However, it is not a G-approximate Jacobian of F at 0. However, the following set $\partial^* F(0)$ is both an approximate Jacobian of F at 0 and a G-approximate Jacobian of F at 0:

$$\partial^* F(0) = \left\{ \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix},\ \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix},\ \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix},\ \begin{pmatrix} -1 & -1 \\ 1 & 1 \end{pmatrix} \right\}.$$

Note that the Clarke Jacobian of F at 0 is given by

$$\partial_C F(0) = \mathrm{co}\left\{ \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix},\ \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix},\ \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix},\ \begin{pmatrix} -1 & -1 \\ 1 & 1 \end{pmatrix} \right\}.$$

The next example shows that a locally Lipschitz function may admit a bounded G-approximate Jacobian at a point, and the convex hull of the G-approximate Jacobian is strictly contained in the Clarke generalized Jacobian at that point.

EXAMPLE 2.2. *Let* $f : \mathbb{R}^2 \to \mathbb{R}$ *be defined by*

$$f(x,y) = \begin{cases} x^2 \sin(\frac{1}{x}) + |y| & \text{if} \quad x \neq 0, \\ |y| & \text{if} \quad x = 0. \end{cases}$$

Then f *is locally Lipschitz at* 0 *and the set*

$$\partial f(0) = \{(0,\beta) : \beta \in [-1,1]\}$$

is a G-approximate Jacobian of f *at* 0*, which is strictly contained in the Clarke Jacobian, given by*

$$\partial_C f(0) = \{(\alpha,\beta) : \alpha,\beta \in [-1,1]\}.$$

PROPOSITION 2.2. *Let the map* $F : \mathbb{R}^n \to \mathbb{R}^m$ *be continuous. If* ∂F *is an approximate Jacobian map of* F *which is upper semicontinuous at* x*, then* $\overline{\text{co}}\partial F(x)$ *is a G-approximate Jacobian of* F *at* x*.*

Proof. Let $u \in \mathbb{R}^n$ and $u \neq 0$, and let $t > 0$. Then it follows from the generalized mean value theorem that

$$F(x+tu) - F(x) \in \overline{\text{co}}(\partial F[x, x+tu]tu) \subset \text{co}(\partial F[x,\ x+tu]tu) + t^2||u||B_m.$$

Now, using Caratheodory's theorem, we find $nm+1$ matrices $N_{t,1}, \ldots, N_{t,nm+1}$ of $\partial F[x,\ x+tu]$, $b_t \in B_m$, and $\lambda_{t,1}, \ldots, \lambda_{t,nm+1} \in [0,1]$ such that $\sum_{i=1}^{nm+1} \lambda_{t,i} = 1$ and

$$F(x+tu) - F(x) = \sum_{i=1}^{nm+1} \lambda_{t,i} N_{t,i}(tu) + t^2||u||b_t.$$

Since ∂F is upper semicontinuous at x, we can find $M_{t,i} \in \partial F(x)$ such that $||M_{t,i} - N_{t,i}|| \to 0$ as $t \to 0$. Now define

$$M_t := \sum_{i=1}^{nm+1} \lambda_{t,i} M_{t,i} \in \text{co}\partial F(x)$$

and

$$\text{o}(t) := \left(\sum_{i=1}^{nm+1} \lambda_{t,i}(N_{t,i} - M_{t,i})(tu) + t^2||u||b_t \right).$$

Then $F(x+tu) - F(x) = M_t tu + \text{o}(t)$ and

$$\lim_{t\to 0} \frac{\text{o}(t)}{t} = \lim_{t\to 0} \left(\sum_{i=1}^{nm+1} \lambda_{t,i}(N_{t,i} - M_{t,i})u + t||u||b_t \right) = 0,$$

and hence $\overline{\text{co}}\partial F(x)$ is a G-approximate Jacobian of F at x. □

It is worth noting from the previous proposition that for a locally Lipschitz map F the Clarke generalized Jacobian $\partial_C F(x)$ at x is a bounded G-approximate Jacobian of F at x.

Let us look at several more numerical examples to illustrate the nature of G-approximate Jacobians. The first example illustrates that a continuous map which is not locally Lipschitz may admit a bounded G-approximate Jacobian at a point.

EXAMPLE 2.3. *Let $f : \mathbb{R} \to \mathbb{R}$ be defined by*

$$f(x) = \begin{cases} x^2 \sin(\frac{1}{x^2}) & \text{if} \quad x \neq 0, \\ 0 & \text{if} \quad x = 0. \end{cases}$$

Then f is continuous but is not locally Lipschitz at 0. The set $\partial f(0) = \{0\}$ is a bounded G-approximate Jacobian of f at 0.

The following example shows that for a continuous map a G-approximate Jacobian at a point may be an unbounded set.

EXAMPLE 2.4. *Let $f : \mathbb{R} \to \mathbb{R}$ be defined by*

$$f(x) = \begin{cases} \sqrt{x} & \text{if} \quad x > 0, \\ -x & \text{if} \quad x \leq 0. \end{cases}$$

Then f is continuous but is not locally Lipschitz at 0. It is easy to see that for each $u \in \mathbb{R}$ and $t > 0$ the condition

$$f(0 + tu) - f(0) = M_t(u)(tu) + \mathrm{o}(t)$$

is satisfied by $\mathrm{o}(t) \equiv 0$, and M_t is defined by

$$M_t(u) = \begin{cases} \frac{1}{\sqrt{tu}} & \text{if} \quad u > 0, \\ -1 & \text{if} \quad u \leq 0. \end{cases}$$

Hence, the set $\partial f(0) = [-1, \infty)$ is a G-approximate Jacobian of f at 0.

We say that a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ admits a G-*approximate Hessian* $\partial^2 f(x)$ at x if this set is a G-approximate Jacobian to the derivative ∇f at x.

3. Chain rules and first-order conditions. In this section, we see how an approximate Jacobian for a convex composite continuous function $g \circ F$, where F is neither Gâteaux differentiable nor locally Lipschitzian, can be constructed using a G-approximate Jacobian. It must be noted that in the case where F is continuous and $\partial F(x)$ is an (unbounded) approximate Jacobian of F at x, the set $\partial g(F(x))^T \overline{\mathrm{co}}(\partial F(x))$ is, in general, not an approximate Jacobian for $g \circ F$ at x (see [14]).

We also present a method for deriving first-order conditions for the convex composite functions without the use of the upper semicontinuity of the approximate Jacobian. Recall that the convex subdifferential $\partial_C g(x)$ of a convex function $g : \mathbb{R}^n \to \mathbb{R}$ at the point x is given by

$$\partial_C g(x) = \left\{ v \in \mathbb{R}^m \mid g(y) - g(x) \geq v^T(y - x) \ \ \forall y \in \mathbb{R}^m \right\}$$

and is a closed and bounded convex subset of $\mathbb{R}^m$.

THEOREM 3.1. *Let $x \in \mathbb{R}^n$, let $F : \mathbb{R}^n \to \mathbb{R}^m$, and let $g : \mathbb{R}^m \to \mathbb{R}$ be a convex function. If $\partial F(x)$ is a G-approximate Jacobian of F at x, then for each $\epsilon > 0$ the closure of the set*

$$(\partial_C g(F(x)) + \epsilon B_m)^T \partial F(x))$$

is an approximate Jacobian of $g \circ F$ at x.

Proof. Let $\epsilon > 0$. It is sufficient to show that, for every $u \in \mathbb{R}^n$ and every $\alpha \in \mathbb{R}$ which we may assume to be nonzero,

$$\alpha(g \circ F)^+(x, u) \leq \sup_{A^T \in (\partial_C g(F(x)) + \epsilon B_m)^T (\partial F(x))} \alpha A^T u. \tag{1}$$

Let $\{t_i\}$ be a sequence of positive numbers converging to 0 such that

$$\alpha(g\circ F)^+(x,u)=\lim_{i\to\infty}\alpha\frac{(g\circ F)(x+t_iu)-(g\circ F)(x)}{t_i}. \tag{2}$$

The mean value theorem of convex analysis gives us that there exists $y_i\in(F(x),\,F(x+t_iu))$, such that

$$\alpha g(F(x+t_iu))-\alpha g(F(x))\in\alpha\partial_C g(y_i)^T(F(x+t_iu)-F(x)).$$

Since $\partial F(x)$ is a G-approximate Jacobian of F at x, it follows that there exists $M_i\in\partial F(x)$ such that

$$F(x+t_iu)-F(x)=M_it_iu+\mathrm{o}(t_i),$$

where, as $t_i\to 0$, $\frac{\mathrm{o}(t_i)}{t_i}\to 0$. So

$$\alpha g(F(x+t_iu))-\alpha g(F(x))\in\alpha\partial_C g(y_i)^T(M_it_iu+\mathrm{o}(t_i)).$$

This yields by the upper semicontinuity of $\partial_C g$ that for sufficiently large i,

$$\alpha g(F(x+t_iu))-\alpha g(F(x))\in\alpha(\partial_C g(F(x))+\epsilon B_m)^T(M_it_iu+\mathrm{o}(t_i)).$$

Dividing by t_i, we get that

$$\frac{1}{t_i}\left(\alpha g(F(x+t_iu))-\alpha g(F(x))\right)\in\alpha(\partial g(F(x))+\epsilon B_m)^T\left(M_iu+\frac{\mathrm{o}(t_i)}{t_i}\right),$$

which yields that

$$\begin{aligned}\frac{1}{t_i}\left(\alpha g(F(x+t_iu))-\alpha g(F(x))\right)&\le\sup_{v\in\partial_C g(F(x))+\epsilon B_m}\alpha v^T\left(M_iu+\frac{\mathrm{o}(t_i)}{t_i}\right)\\&\le\sup_{v\in\partial_C g(F(x))+\epsilon B_m,\ M\in\partial F(x)}\alpha v^TMu+\frac{\alpha v^T\mathrm{o}(t_i)}{t_i}.\end{aligned}$$

Since $\partial_C g(F(x))+\epsilon B_m$ is bounded, it follows by letting $t_i\to 0$ that

$$\alpha(g\circ F)^+(x,u)\le\sup_{A\in(\partial_C g(F(x))+\epsilon B_m)^T\partial F(x)}\alpha Au. \qquad \square$$

COROLLARY 3.1. *Let $x\in\mathbb{R}^n$, let $F:\mathbb{R}^n\to\mathbb{R}^m$ be a continuous map, and let $g:\mathbb{R}^m\to\mathbb{R}$ be a convex function. If ∂F is an approximate Jacobian map of F which is upper semicontinuous at x, then for each $\epsilon>0$ the closure of the set*

$$(\partial_C g(F(x))+\epsilon B_m)^T\overline{\mathrm{co}}(\partial F(x))$$

is an approximate Jacobian of $g\circ F$ at x.

Proof. The conclusion follows from the previous theorem and Proposition 2.2 by noting that $\overline{\mathrm{co}}(\partial F(x))$ is a G-approximate Jacobian of F at x. $\square$

COROLLARY 3.2. *Let $g:\mathbb{R}^m\to\mathbb{R}$ be a convex function, and let $F:\mathbb{R}^n\to\mathbb{R}^m$. If $\partial F(x)$ is a bounded G-approximate Jacobian of F at $x\in\mathbb{R}^n$, then $\partial g(F(x))^T\partial F(x)$ is an approximate Jacobian of the composite function $g\circ F$ at x.*

Proof. In this case, for each $\epsilon > 0$ the bounded set

$$(\partial_C g(F(x)) + \epsilon B_m)^T \partial F(x)$$

is a decreasing sequence of bounded approximate Jacobians of $g \circ F$ at x. Hence,

$$\partial_C g(F(x))^T \partial F(x) = \bigcap_{\epsilon>0} (\partial_C g(F(x)) + \epsilon B_m)^T \partial F(x)$$

is also an approximate Jacobian of $g \circ f$ at x. □

It is worth noting that the conclusion of this theorem continues to hold for locally Lipschitzian function g by replacing the convex subdifferential $\partial_C g(F(x))$ by the locally bounded Clarke subdifferential $\partial_C g(F(x))$.

COROLLARY 3.3. *Let $g : \mathbb{R}^m \to \mathbb{R}$ be a convex function, and let $F : \mathbb{R}^n \to \mathbb{R}^m$ be a locally Lipschitz map at $x \in \mathbb{R}^n$. Then $\partial_C g(F(x))^T \partial_C F(x)$ is an approximate Jacobian of the composite function $g \circ F$ at x. Moreover,*

$$\partial_C (g \circ F)(x) \subset \text{co}(\partial_C g(F(x))^T \partial_C F(x)).$$

Proof. Since $\partial_C F(x)$ is a bounded G-approximate Jacobian for a locally Lipschitz map F at x, it follows from the previous theorem that $\partial_C g(F(.))^T \partial_C F(.)$ is an approximate Jacobian of $g \circ F$ which is upper semicontinuous at x. The inclusion follows from the fact that $\partial_C (g \circ F)(x)$ is the smallest convex-valued approximate Jacobian of $g \circ F$ which is upper semicontinuous at x. □

COROLLARY 3.4. *Let $g : \mathbb{R}^m \to \mathbb{R}$ be a convex function, and let $F : \mathbb{R}^n \to \mathbb{R}^m$ be Gâteaux differentiable at x with the Gâteaux derivative $\nabla F(x)$ at $x \in \mathbb{R}^n$. Then $\partial_C g(F(x))^T \ \nabla F(x)$ is an approximate Jacobian of the composite function $g \circ F$ at x.*

Proof. The conclusion follows from Corollary 3.2, as $\{\nabla F(x)\}$ is a bounded G-approximate Jacobian of F at x. □

Observe that if g is convex and Gâteaux differentiable at $F(x)$ and if F is Gâteaux differentiable at x, then $g \circ F$ is Gâteaux differentiable at x and $\nabla(g \circ F)(x) = \nabla g(F(x)^T \nabla F(x)$. This observation continues to hold for a locally Lipschitz Gâteaux differentiable function g.

Now, by using the G-approximate Jacobian of F, we establish necessary and sufficient conditions for optimality of the convex composite function $g \circ F$, where F is neither Gâteaux differentiable nor locally Lipschitz. The *recession cones* [21, 22, 24] pave the way for describing the optimality conditions involving unbounded G-approximate Jacobians. Note that a vector $u \in \mathbb{R}^l$ is said to be a *recession direction* of a nonempty set A in $\mathbb{R}^l$ if there is a sequence of positive numbers $\{t_j\}$ converging to 0 and a sequence $\{a_j\}$ of elements of A such that $u = \lim_{j\to\infty} t_j a_j$. The set of all recession directions of A is called the *recession cone of A* and is denoted by A_∞. Observe that a set is unbounded if and only if its recession cone is nontrivial.

THEOREM 3.2 (necessary condition). *Let $x \in \mathbb{R}^n$, let $F : \mathbb{R}^n \to \mathbb{R}^m$ be a map, and let $g : \mathbb{R}^m \to \mathbb{R}$ be a convex function. Assume that $\partial F(x)$ is a G-approximate Jacobian of F at x. If x is a local minimizer of the composite function $g \circ F$, then*

$$0 \in \overline{\text{co}}\partial_C g(F(x))^T \partial F(x) \cup \text{co}(\ \partial_C g(F(x))^T ((\partial F(x))_\infty \backslash \{0\})).$$

Proof. It follows from Theorem 3.1 that for each $\epsilon > 0$ the closure of the set

$$(\partial_C g(F(x)) + \epsilon B_m)^T \partial F(x))$$

is an approximate Jacobian of $g \circ F$ at x. Since $g \circ F$ attains a local minimum at x, the necessary optimality condition (see section 2) gives us that

$$0 \in \overline{\text{co}}(\partial_C g(F(x)) + \epsilon B_m)^T \partial F(x).$$

Take $\epsilon = \frac{1}{k}$, $k \geq 1$. Then by Caratheodory's theorem we can represent 0 as

$$0 = \sum_{j=1}^{n+1} \mu_k^j \left(a_{jk} + \frac{1}{k} b_{jk} \right)^T c_{jk} + \frac{1}{k} f_k, \tag{3}$$

where

$$\mu_k^j \geq 0, \quad \sum_{i=1}^{k+1} \mu_k^j = 1, \quad a_{jk} \in \partial_C g(F(x)), \quad b_{jk} \in B_m,$$
$$c_{jk} \in \partial F(x), \quad j = 1, \ldots, n+1, \quad \text{and } f_k \in B_m.$$

Let

$$J := \{1, 2, \ldots, n+1\}, \quad J_1 := \{j \in J : \{c_{jk}\}_{k \geq 1} \text{ is bounded}\}, \quad \text{and } J_2 := J \setminus J_1.$$

Then (3) can be rewritten as

$$0 = \left(\sum_{j \in J_1} \mu_k^j \left(a_{jk} + \frac{1}{k} b_{jk} \right)^T c_{jk} + \sum_{j \in J_2} \mu_k^j \left(a_{jk} + \frac{1}{k} b_{jk} \right)^T c_{jk} \right) + \frac{1}{k} f_k. \tag{4}$$

We may now assume, without loss of generality, that

$$\mu_k^j \to \mu^j \in [0,1] \text{ and } \sum_{j=1}^{n+1} \mu^j = 1,$$
$$a_{jk} \to a_j \in \partial_C g(F(x)), \quad b_{jk} \to b_j \in B_m, \quad j = 1, \ldots, n+1,$$
$$c_{jk} \to c_j \in \partial F(x), \quad j \in J_1, \quad \text{and } f_k \to f \in B_m.$$

Case 1. $J_2 = \phi$. In this case, by letting $k \to \infty$, (4) yields

$$0 = \sum_{j=1}^{n+1} \mu^j a_j^T c_j \in \text{co}\, \partial_C g(F(x))^T \partial F(x).$$

Case 2. $J_2 \neq \phi$.

Case 2(a). Assume that $\{\mu_k^j c_{jk}\}_{k \geq 1}$ is bounded for every $j \in J_2$. Then $\mu^j = 0$ for all $j \in J_2$. Hence $\sum_{j \in J_1} \mu^j = 1$. So we may assume that

$$\mu_k^j c_{jk} \to c_j \in (\partial F(x))_\infty, \quad j \in J_2.$$

Passing (4) to limit, we get

$$\begin{aligned} 0 &\in \sum_{j \in J_1} \mu^j a_j^T c_j + \sum_{j \in J_2} a_j^T c_j \\ &\in (\text{co}\, \partial_C g(F(x))^T \partial F(x) + \text{co}\, (\partial_C g(F(x))^T (\partial F(x))_\infty) \\ &\subset \overline{\text{co}}\, \partial_C g(F(x))^T \partial F(x), \end{aligned}$$

since $\operatorname{co}\partial_C g(F(x))^T\partial F(x) + \operatorname{co}(\partial_C g(F(x))^T(\partial F(x))_\infty) \subset \overline{\operatorname{co}}\,\partial_C g(F(x))^T\partial F(x)$. This inclusion follows from the fact that

$$\partial_C g(F(x))^T(\partial F(x))_\infty \subset (\partial_C g(F(x))^T\partial F(x))_\infty \subset (\operatorname{co}\partial_C g(F(x))^T\partial F(x))_\infty$$

and that

$$\begin{aligned}&\operatorname{co}\partial_C g(F(x))^T\partial F(x) + \operatorname{co}(\,\partial g(F(x))^T(\partial F(x))_\infty)\\ &\subset \overline{\operatorname{co}}\,\partial_C g(F(x))^T\partial F(x) + (\overline{\operatorname{co}}(\,\partial_C g(F(x))^T\partial F(x))_\infty\\ &= \overline{\operatorname{co}}\,\partial_C g(F(x))^T\partial F(x).\end{aligned}$$

Case 2(b). Assume that there exists $j \in J_2$ such that $\{\mu_k^j c_{jk}\}_{k\geq 1}$ is unbounded. Then by taking subsequences instead we may assume there exists $j_0 \in J_2$ such that

$$\|\mu_k^{j_0} c_{j_0 k}\| \geq \|\mu_k^j c_{jk}\| \;\forall j \in J_2, \quad k \geq 1.$$

Then $\frac{\mu_k^j c_{jk}}{\|\mu_k^{j_0} c_{j_0 k}\|} \to c_j \in (\partial F(x))_\infty,\ j \in J_2$. Put $J_3 := \{j \in J_2 : c_j \neq 0\}$. Then $J_3 \neq \phi$, since $j_0 \in J_3$. Now, by dividing (4) by $\|\mu_k^{j_0} c_{j_0 k}\|$ and passing to limit for $k \to \infty$, we obtain

$$0 = \sum_{j\in J_3} a_j^T c_j \in \operatorname{co}(\partial_C g(F(x))^T((\partial F(x))_\infty \backslash \{0\})). \qquad \square$$

The example below shows that the optimality conditions for a convex composite locally Lipschitz function, expressed in terms of G-approximate Jacobians, are sharper than the corresponding conditions of the Clarke subdifferential.

EXAMPLE 3.1. *Consider the function* $F : \mathbb{R}^2 \to \mathbb{R}^2$ *and* $g : \mathbb{R}^2 \to \mathbb{R}$ *defined, respectively, by*

$$F(x,y) = (2|x| - |y|, \quad 2|y| - |x|)^T \quad \textit{and} \quad g(x,y) = x + y.$$

Then F *is locally Lipschitz,* g *is convex, and* $(g \circ F)(x,y) = |x| + |y|$ *. It is easy to verify that a G-approximate Jacobian of* F *at* 0 *is given by*

$$\partial F(0) = \left\{ \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}, \ \begin{pmatrix} -2 & 1 \\ 1 & -2 \end{pmatrix}, \ \begin{pmatrix} 2 & 1 \\ -1 & -2 \end{pmatrix}, \ \begin{pmatrix} -2 & -1 \\ 1 & 2 \end{pmatrix} \right\}.$$

Clearly, 0 *is a minimizer of* $g \circ F$ *and*

$$0 \in \operatorname{co}(\partial_C g(F(0))^T\partial F(0)) = \operatorname{co}\{(1\ 1), (-1\ -1), (1\ -1), (-1\ 1)\} \subset \partial_C(g \circ F)(0),$$

where $\partial_C(g \circ F)(0) = \{(\alpha, \beta) : \alpha, \beta \in [-1, 1]\}$.

It is also worth observing that the above optimality condition yields a sharper condition for convex composite functions $g \circ F$ in the case where F is locally Lipschitz and Gateaux differentiable than the corresponding conditions for the Clarke generalized Jacobian.

THEOREM 3.3 (sufficient condition). *Let* $F : \mathbb{R}^n \to \mathbb{R}^m$ *be a continuous map, let* $g : \mathbb{R}^m \to \mathbb{R}$ *be a convex function, and let* $a \in \mathbb{R}^n$ *. Suppose that* F *admits a G-approximate Jacobian around* a*. If there exist a convex neighborhood* $N(a)$ *of* a *such that for each* $x \in N(a) \backslash \{a\}$ *there exists* $\epsilon > 0$ *satisfying*

$$w^T(x - a) \geq 0 \qquad \forall w \in co[(\partial_C g(F(x)) + \epsilon B_m)^T\partial F(x)],$$

then a is a minimizer of $g \circ F$ on $N(a)$.

Proof. Suppose to the contrary that a is not a minimizer of $g \circ F$ on $N(a)$. Then there exists $x_0 \in N(a)$ such that $(g \circ F)(x_0) < (g \circ F)(a)$. For each $x \in (a, x_0)$, let $\epsilon_x > 0$ be such that

$$w^T(x-a) \geq 0 \qquad \forall w \in \mathrm{co}[(\partial_C g(F(x)) + \epsilon_x B_m)^T \partial F(x)] \ .$$

By Theorem 3.1, the closure of the set

$$(\partial_C g(F(x)) + \epsilon_x B_m)^T \partial F(x)$$

is an approximate Jacobian of $g \circ F$ at x. It now follows from the generalized mean value theorem (see section 2) that there exists $\hat{x} \in (a, x_0)$ such that

$$(g \circ F)(x_0) - (g \circ F)(a) \in \overline{\mathrm{co}}\left[(\partial_C g(F(\hat{x})) + \epsilon_{\hat{x}} B_m)^T \partial F(\hat{x})(x_0 - a)\right].$$

Hence, we can find $w \in \mathrm{co}[(\partial_C g(F(\hat{x})) + \epsilon_{\hat{x}} B_m)^T \partial F(\hat{x})]$ such that $w^T(x_0 - a) < 0$, and so $w^T(\hat{x} - a) < 0$. This contradicts the hypothesis of the theorem. ∎

4. Second-order conditions. In this section, we prove second-order results for the convex composite function $g \circ F$, where $g : \mathbb{R}^m \to \mathbb{R}$ is convex and $F : \mathbb{R}^n \to \mathbb{R}^m$ is Gâteaux differentiable. Here we extend the approach developed in [1, 3, 10] by using our generalized variational calculus from previous sections.

The Lagrangian function corresponding to the convex composite function $g \circ F$ is defined by

$$L(x, y^*) = y^{*T} F(x) - g^*(y^*), \qquad x \in \mathbb{R}^n,\ y^* \in \mathbb{R}^m,$$

where g^* is the *Fenchel conjugate function* of g, which is given by

$$g^*(y^*) = \sup\{\langle y^*, y\rangle - g(y) : y \in \mathbb{R}^m\}, \qquad y^* \in \mathbb{R}^m.$$

Recall that the *ϵ-subdifferential* of g at y is given by

$$\partial_\epsilon g(y) = \{y^* \in \mathbb{R}^m : g(z) \geq g(y) + y^{*T}(z - y) - \epsilon\ \forall z \in \mathbb{R}^m\}.$$

Let $h : \mathbb{R}^n \to \mathbb{R}$. A real-valued function $\phi(x, u)$ defined on $\mathbb{R}^n \times \mathbb{R}^n$ is said to be a Levitin–Miljutin–Osmolovskii (LMO)-*approximation* for h at z in the sense of Ioffe [9] if $\phi(z, 0) = h(z)$, for any x in a neighborhood of z, the function $u \to \phi(x, u)$ is convex, and

$$\liminf_{y \to z, u \to 0} \|u\|^{-1}(\phi(y, u) - h(y + u)) \geq 0.$$

Using this LMO-approximation, Ioffe established the following characterizations of a local minimum of a locally Lipschitz function.

Lemma 4.1 (Proposition 5 in Ioffe [9]). *Assume that h is a locally Lipschitz on $\mathbb{R}^n$ and $z \in \mathbb{R}^n$ and that $\phi(x, u)$ is an LMO-approximation of h at z. Let $\beta_\xi(x) = -\min\{\phi^*(x, u^*) : \|u^*\| \leq \xi\}$ for any fixed $\xi > 0$. Then the following conditions are equivalent:*

(i) *h attains a local minimum at z;*
(ii) *$0 \in \partial\phi(z, 0)$ and β_ξ attains a local minimum at z for any $\xi > 0$;*
(iii) *$0 \in \partial\phi(z, 0)$ and β_ξ attains a local minimum at z for some $\xi > 0$.*

Observe from the convexity of g and Gâteaux differentiability F that the composite function $f := g \circ F$ is directionally differentiable and its directional derivative at x is given by

$$f'(x,d) = g'(F(x), \nabla F(x)d).$$

Let

$$K(x) := \{u \in \mathbb{R}^n : \ g(F(x) + t\nabla F(x)u) \leq g(F(x)) \ \text{ for some } t > 0\},$$

and let

$$D(x) := \{u \in \mathbb{R}^n : g'(F(x), \nabla F(x)u) \leq 0\}.$$

For $z \in \mathbb{R}^n$, define

$$M_0(z) = \{y^* \in \mathbb{R}^m : y^* \in \partial_C g(F(z)), \ y^{*T}\nabla F(z) = 0\}.$$

Then clearly $M_0(z) \neq \emptyset$, provided $0 \in \partial_C g(F(z))^T \nabla F(z)$. Now we state the second-order optimality conditions for the function $g \circ F$.

THEOREM 4.1 (necessary condition). *Let $a \in \mathbb{R}^n$. Assume that g is a convex function and F is Gâteaux differentiable at a. Suppose that for each $y^* \in \mathbb{R}^m$, $\partial^2 L(a, y^*)$ is a G-approximate Hessian of $L(., y^*)$ at a and that $\partial^2 L(a, \ .)$ is upper semicontinuous on $\mathbb{R}^m$. If a is a local minimizer of $g \circ F$, then*

$$\sup\{u^T M u : M \in \partial^2 L(a, y^*), \ y^* \in M_0(a)\} \geq 0 \quad \forall u \in K(a).$$

Proof. Let $u \in K(a)$. First observe from Corollary 3.4 that

$$0 \in \partial_C g(F(a))\nabla F(a)$$

as $g \circ F$ attains a local minimum at a. This yields $M_0(a) \neq \emptyset$. Now let $\epsilon > 0$. Then it follows from Proposition 1 in Ioffe [9] that the function

$$\rho_\epsilon(x; u) = g_\epsilon(\nabla F(a)u + F(x))$$

is an LMO-approximation of f at a, where $g_\epsilon(y) = \sup\{y^{*T}y - g^*(y^*) : y^* \in \partial_\epsilon g(F(x))\}$. Let $\eta > 0$, and define the function $\phi_{\eta\epsilon}$ by

$$\phi_{\eta\epsilon}(x) = \max\{L(x, y^*) : y^* \in M_{\eta\epsilon}(a)\},$$

where

$$M_{\eta\epsilon}(a) = \{y^* \in \mathbb{R}^m : y^* \in \partial_\epsilon g(F(a)), \ \|y^{*T}\nabla F(a)\| \leq \eta\}.$$

If we show that

$$\phi_{\eta\epsilon}(x) = -\min\{\rho^*_\epsilon(x, u^*) : \|u^*\| \leq \eta\}, \tag{5}$$

where $\rho^*_\epsilon(x, u^*) = \sup\{u^{*T}u - \rho_\epsilon(x, u) : u \in \mathbb{R}^n\}$ is the Fenchel conjugate of $\rho_\epsilon(x, \cdot)$, then it follows from Lemma 4.1 that $\phi_{\eta\epsilon}$ attains a local minimum at a. To show (5), note that

$$\rho^*_\epsilon(x, h^*) = \inf\{g^*(y^*) + \delta(y^*|\partial_\epsilon g(F(a)) - y^{*T}F(x) \ : \ h^* = y^{*T}\nabla F(a)\}.$$

Then

$$\begin{aligned}
&-\min\{\rho_\epsilon^*(x,h^*) : \|h^*\| \le \eta\}\\
&= \max\{-\inf\{g^*(y^*) + \delta(y^*|\partial_\epsilon g(F(a)) - y^{*T}F(x) \ : \ h^* = y^{*T}\nabla F(a)\} - \delta(h^*|\eta B)\}\\
&= \max\{y^{*T}F(x) - g^*(y^*) - \delta(y^*|\partial_\epsilon g(F(a)) - \delta(L_x(a,y^*)|\eta B) \ : \ y^* \in \mathbb{R}^m\}\\
&= \max\{L(x,y^*) \ : \ y^* \in \partial_\epsilon g(F(a)), ||L_x(a,y^*)|| \le \eta\}\\
&= \phi_{\eta\epsilon}(x),
\end{aligned}$$

where B is the unit ball. Now from the classical mean value theorem and the definition of G-approximate Hessian we get that for t sufficiently small positive,

$$\begin{aligned}
&g(F(a)) \le \phi_{\eta\epsilon}(a+tu)\\
&= \sup\{L(a+tu,y^*) : y^* \in M_{\eta\epsilon}(a)\}\\
&= \sup\{y^{*T}F(a+tu) - g^*(y^*) : y^* \in M_{\eta\epsilon}(a)\}\\
&= \sup\{y^{*T}F(a) + y^{*T}\nabla F(a+su)(tu) - g^*(y^*) : y^* \in M_{\eta\epsilon}(a)\} \ \text{ for some } s \in (0,t)\\
&= \sup\{y^{*T}F(a) + y^{*T}\nabla F(a)(tu) + (su)^T M(tu) + \mathrm{o}(s)(tu) - g^*(y^*) : \ y^* \in M_{\eta\epsilon}(a)\},
\end{aligned}$$

where $M \in \partial^2 L(a,y^*)$. Since $u \in K(a)$ and g is convex, there exists $t_0 > 0$ such that

$$g(F(a) + t\nabla F(a)u) \le g(F(a)) \qquad \forall t \in [0,t_0].$$

The basic properties of the Fenchel conjugate function of g give us

$$y^{*T}(F(a) + t\nabla F(a)u) - g^*(y^*) \le g(F(a) + t\nabla F(a)u) \le g(F(a))\forall t \in [0,t_0].$$

So, for sufficiently small $t > 0$,

$$\sup \ \{(st)(u^T M u + \mathrm{o}(s)(tu) : y^* \in M_{\eta\epsilon}(a), \ M \in \partial^2 L(a, \ y^*)\} \ge 0.$$

Thus,

$$\sup \ \left\{u^T M u + \frac{\mathrm{o}(s)u}{s} : y^* \in M_{\eta\epsilon}(a), \ M \in \partial^2 L(a, \ y^*)\right\} \ge 0.$$

As $t \downarrow 0$, $\frac{\mathrm{o}(s)}{s} \to 0$, and so we obtain

$$\sup \ \{u^T M u : y^* \in M_{\eta\epsilon}(a), \ M \in \partial^2 L(a, \ y^*)\} \ge 0.$$

This and the upper semicontinuity of $\partial^2 L(a, \ .)$ yield the conclusion by noting that

$$\bigcap_{\eta>0,\epsilon>0} M_{\eta\epsilon}(a) = M_0(a). \qquad \square$$

COROLLARY 4.1. *Let $a \in \mathbb{R}^n$. Assume that g is a convex function and F is Gâteaux differentiable at a. Suppose that for each $y^* \in \mathbb{R}^m$, $\partial^2 L(a,y^*)$ is a bounded G-approximate Hessian of $L(.,y^*)$ at a and that $\partial^2 L(a, \ .)$ is upper semicontinuous on $\mathbb{R}^m$. If a is a local minimizer of $g \circ F$, then*

$$\sup\{u^T M u : M \in \partial^2 L(a,y^*), \ y^* \in M_0(a)\} \ge 0 \ \ \forall u \in \overline{K(a)}.$$

Proof. We need only to notice that the conditions of the previous theorem are now true for any $u \in \overline{K(a)}$ since $\partial^2 L(a, y^*)$ is bounded for each $y^* \in M_0(a)$. □

THEOREM 4.2 (sufficient conditions). *Let $a \in \mathbb{R}^n$. Assume that g is a convex function and F is continuously Gâteaux differentiable. Suppose that for each $y^* \in \mathbb{R}^m$, $\partial^2 L(., y^*)$ is an approximate Hessian of $L(., y^*)$. If $M_0(a) \neq \emptyset$ and if for each $u \in D(a)\backslash\{0\}$ there exist $\epsilon > 0$ and $\delta > 0$ satisfying*

$$\inf_{v \in u+\delta B_n} \sup_{y^* \in M_0(a)} \inf_{M \in \overline{\text{co}}\partial^2 L(a+\epsilon B_n, y^*)} v^T M v > 0,$$

then a is a strict local minimizer of order 2 for the function $g \circ F$.

Proof. Suppose to the contrary that a is not a strict local minimizer of order 2 for $g \circ F$. Then there exist $\{x_k\} \subseteq \mathbb{R}^n$, $x_k \to a$, and $\epsilon_k \downarrow 0$ as $k \to +\infty$ such that for each k,

$$f(x_k) \leq f(a) + \epsilon_k \|x_k - a\|^2.$$

We may assume that $u_k := \frac{x_k - a}{\|x_k - a\|} \to u \in D(a)\backslash\{0\}$ as $k \to +\infty$. It now follows from the definition of conjugate function that

$$\begin{aligned} g(F(x_k)) &= \sup\{{y^*}^T F(x_k) - g^*(y^*) : y^* \in \mathbb{R}^n\} \\ &\geq \sup\{{y^*}^T F(a + t_k u_k) - g^*(y^*) : y^* \in M_0(a)\}, \end{aligned}$$

where $t_k = ||x_k - a|| \to 0$ as $k \to \infty$. Now, by the generalized Taylor's expansion (see section 2, (vii)), there exists $s_k > 0$ with $t_k > s_k$ and $M_k \in \overline{\text{co}}\partial^2 L(a + s_k u_k, y^*)$ such that

$$\begin{aligned} &{y^*}^T F(a + t_k u_k) - g^*(y^*) \\ &= {y^*}^T F(a) - g^*(y^*) + {y^*}^T \nabla F(a) t_k u_k + \frac{1}{2}(t_k u_k)^T M_k (t_k u_k) + \text{o}(t_k^2 ||u_k||^2), \end{aligned}$$

where $\frac{\text{o}(t_k^2 ||u_k||^2)}{t_k^2} \to 0$, as $k \to \infty$. From the conjugate duality theory and the assumption that $M_0(a)$ is nonempty we get $g(F(a)) = {y^*}^T F(a) - g^*(y^*)$ and ${y^*}^T \nabla F(a) = 0$, for $y^* \in M_0(a)$, we obtain that

$$\epsilon_k \geq \sup_{y^* \in M_0(a)} \left\{ \frac{1}{2} {u_k}^T M_k u_k + \frac{\text{o}(t_k^2 ||u_k||^2)}{t_k^2} \right\},$$

where $M_k \in \overline{\text{co}}\partial^2 L(a + s_k u_k, y^*)$. Let $\alpha > 0$ be a constant such that

$$\sup_{y^* \in M_0(a)} \inf_{M \in \overline{\text{co}}\partial^2 L(a+\epsilon B_n, y^*)} v^T M v \geq \alpha > 0 \quad \forall v \in u + \delta B_n.$$

Let k_0 be an integer sufficiently large such that $u_k \in u + \delta B_n$ and $M_k \in \overline{\text{co}}\partial^2 L(a + \epsilon B_n, y^*)$ for $k \geq k_0$. Let k_1 be another integer such that

$$\epsilon_k - \frac{\text{o}(t_k^2 ||u_k||^2)}{t_k^2} \leq \frac{\alpha}{4} \text{ for } k \geq k_1.$$

Hence, we get that

$$\frac{\alpha}{4} \geq \sup_{y^* \in M_0(a)} \frac{1}{2} {u_k}^T M_k u_k \geq \frac{\alpha}{2},$$

which contradicts the hypothesis, and so the conclusion follows. □

COROLLARY 4.2. *Let $a \in \mathbb{R}^n$. Assume that g is a convex function and F is $C^{1,1}$ at a (i.e., F is continuously Gâteaux differentiable at a with locally Lipschitz ∇F). Then,*

(i) *if a is a local minimum of $g \circ F$, then*

$$\max\{L^{oo}(a, y^*; u, u) : y^* \in M_0(a)\} \geq 0 \forall u \in \overline{K(a)}; \tag{6}$$

(ii) *if $M_0(a) \neq \emptyset$ and if*

$$\max\{-L^{oo}(a, y^*; u, -u) : y^* \in M_0(a)\} > 0 \forall u \in D(a) \backslash \{0\},$$

then a is a strict local minimum of order 2 for $g \circ F$.

Proof. Now $L(x, y^*)$ is $C^{1,1}$, since F is $C^{1,1}$. We choose $\partial^2 L(x, y^*) = \partial_H^2 L(x, y^*)$ at each x and y^*, where $\partial_H^2 L(., y^*)$ is the Clarke generalized Hessian which is a bounded G-approximate Hessian of $L(., y^*)$ at x and is upper semicontinuous at x. Moreover, for each $u \in \mathbb{R}^n$,

$$L^{\circ\circ}(a, y^*; u, u) = \sup_{M \in \partial_H^2 L(a,\ y^*)} u^T M u,$$

$$-L^{\circ\circ}(a, y^*; u, -u) = -\sup_{M \in \partial_H^2 L(a,\ y^*)} u^T M(-u) = \inf_{M \in \partial_H^2 L(a,\ y^*)} u^T M u.$$

Hence, the conclusion of (i) follows easily from the Corollary 4.1 and that of (ii) from Theorem 4.2 using the upper semicontinuity of $\partial_H^2 L(., y^*)$. □

Acknowledgment. The authors are grateful to the referees for their detailed comments and suggestions which have contributed to the final preparation of the paper.

REFERENCES

[1] J. V. BURKE, *Second order necessary and sufficient conditions for convex composite NDO*, Math. Programming, 38 (1987), pp. 287–302.
[2] J. V. BURKE AND M. L. OVERTON, *Variational analysis of non-Lipschitz spectral functions*, Math. Program., 90 (2001), pp. 317–357.
[3] J. V. BURKE AND R. A. POLIQUIN, *Optimality conditions for nonfinite valued convex composite functions*, Math. Programming, 57 (1992), pp. 103–120.
[4] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, 1983.
[5] V. F. DEMYANOV AND A. M. RUBINOV, *Constructive Nonsmooth Analysis*, Verlag Peter Lang, Frankfurt, Germany, 1995.
[6] A. FISCHER, V. JEYAKUMAR, AND D. T. LUC, *Solution point characterizations and convergence analysis of a descent algorithm for nonsmooth continuous complementarity problems*, J. Optim. Theory Appl., 110 (2001), pp. 493–513.
[7] R. FLETCHER, *Practical Methods of Optimization*, John Wiley, New York, 1987.
[8] J. B. HIRIART-URRUTY, J. J. STRODIOT, AND V. HIEN NGUYEN, *Generalized Hessian matrix and second-order optimality conditions for problems with $C^{1,1}$ data*, Appl. Math. Optim., 11 (1984), pp. 43–56.
[9] A. D. IOFFE, *Necessary and sufficient conditions for a local minimum.* 2: *Conditions of Levitin–Miljutin–Osmolovskii type*, SIAM J. Control Optim., 17 (1979), pp. 251–265.
[10] 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.
[11] V. JEYAKUMAR, *Convex Composite Optimization*, in Encyclopedia of Optimization, Vol. 1, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2001, pp. 307–310.

[12] V. Jeyakumar, *Composite nonsmooth programming with Gâteaux differentiability*, SIAM J. Optim., 1 (1991), pp. 30–41.
[13] V. Jeyakumar and D. T. Luc, *Approximate Jacobian matrices for nonsmooth continuous maps and C^1-optimization*, SIAM J. Control Optim., 36 (1998), pp. 1815–1832.
[14] V. Jeyakumar and D. T. Luc, *Nonsmooth calculus, minimality and monotonicity of convexificators*, J. Optim. Theory Appl., 101 (1999), pp. 599–621.
[15] V. Jeyakumar and D. T. Luc, *An open mapping theorem using unbounded generalized Jacobians*, Nonlinear Anal., 50 (2002), pp. 647–663.
[16] V. Jeyakumar, D. T. Luc, and S. Schaible, *Characterizations of generalized monotone nonsmooth continuous maps using approximate Jacobians*, J. Convex Anal., 5 (1998), pp. 119–132.
[17] V. Jeyakumar and X. Wang, *Approximate Hessian matrices and second-order optimality conditions for nonlinear programming problems with C^1-data*, J. Austral. Math. Soc. Ser. B, 40 (1999), pp. 403–420.
[18] V. Jeyakumar and X. Q. Yang, *Convex composite multi-objective nonsmooth programming*, Math. Programming, 59 (1993), pp. 325–343.
[19] V. Jeyakumar and X. Q. Yang, *Convex composite minimization with $C^{1,1}$ functions*, J. Optim. Theory Appl., 86 (1995), pp. 631–648.
[20] D. T. Luc, *A Fuzzy Chain Rule for Approximate Jacobians of Continuous Functions*, Preprint N2001/5, Institute of Mathematics, Hanoi, Vietnam, 2001.
[21] D. T. Luc, *Recession maps and applications*, Optimization, 27 (1993), pp. 1–15.
[22] D. T. Luc, *Theory of Vector Optimization*, Lecture Notes in Econom. and Math. Systems 319, Springer-Verlag, Berlin, 1989.
[23] J. P. Penot, *Optimality conditions in mathematical programming and composite optimization*, Math. Programming, 67 (1994), pp. 225–245.
[24] R. T. Rockafellar and R. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
[25] M. Studniarski and V. Jeyakumar, *A generalized mean-value theorem and optimality conditions in composite nonsmooth minimization*, Nonlinear Anal., 24 (1995), pp. 883–894.
[26] X. Wang and V. Jeyakumar, *A sharp Lagrange multiplier rule for nonsmooth mathematical programming problems involving equality constraints*, SIAM J. Optim., 10 (2000), pp. 1136–1148.
[27] X. Q. Yang, *Second-order global optimality conditions for convex composite optimization*, Math. Programming, 81 (1998), pp. 327–347.
[28] X. Q. Yang and V. Jeyakumar, *First and second-order optimality conditions for convex composite multi-objective optimization*, J. Optim. Theory Appl., 95 (1997), pp. 209–224.

SIAM J. OPTIM.
Vol. 13, No. 3, pp. 921–937

AN OPTIMIZATION APPROACH FOR RADIOSURGERY TREATMENT PLANNING*

MICHAEL C. FERRIS†, JINHO LIM‡, AND DAVID M. SHEPARD§

Abstract. We outline a new approach for radiosurgery treatment planning, based on solving a series of optimization problems. We consider a specific treatment planning problem for a specialized device known as the *gamma knife*, which provides an advanced stereotactic approach to the treatment of tumors, vascular malformations, and pain disorders within the head. The sequence of optimization problems involves nonlinear and mixed integer programs whose solution is required in a given planning time (typically less than 30 minutes). This paper outlines several modeling decisions that result in more efficient and robust solutions. Furthermore, it outlines a new approach for determining starting points for the nonlinear programs, based on a skeletonization of the target volume. Treatment plans generated for real patient data show the efficiency of the approach.

Key words. radiation therapy, optimization, treatment planning, gamma knife

AMS subject classifications. 90C50, 65K05, 90C30

PII. S105262340139745X

1. Introduction.

Radiation therapy is the treatment of cancer with ionizing radiation. This radiation, in the form of X-rays and gamma rays, damages the DNA of the cells in the area being treated, interfering with their ability to divide and grow. Cancerous cells are unable to repair this damage, and thus their growth is curtailed and the tumor shrinks. Healthy cells may also be damaged by the radiation, but they are more able to repair the damage and return to normal function. Radiation therapy may be used to treat solid tumors, such as cancers of the skin, brain, and breast. It can attack cancer cells both on the surface of the body and deep within. It can be used as the sole form of treatment, or in conjunction with surgery (to shrink the tumor before surgery, or to kill remaining cancer cells after surgery) or chemotherapy.

Devices for delivering the radiation allow a significant amount of control over the characteristics of the radiation. Treatment plans, which specify the shapes of the applied radiation beams, times of exposure, etc., should be designed in a way that delivers a specified dose to the tumor while avoiding an excessive dose to the surrounding healthy tissue and, in particular, to any important nearby organs. The full potential of these devices to deliver optimal treatment plans has yet to be realized, due to the complexity of the treatment design process. This paper describes how to use advanced modeling techniques and state-of-the-art optimization algorithms for the design of treatment plans that fully exploit the capabilities of this new generation of technology.

*Received by the editors November 6, 2001; accepted for publication (in revised form) July 29, 2002; published electronically February 27, 2003. This material is based on research supported in part by the National Science Foundation grant CCR-9972372, the Air Force Office of Scientific Research Grant F49620-01-1-0040, Microsoft Corporation, and the Guggenheim Foundation.
http://www.siam.org/journals/siopt/13-3/39745.html
†Oxford University Computing Laboratory, Wolfson Building, Parks Road, Oxford, OX1 3QD, UK. Permanent address: Computer Sciences Department, University of Wisconsin, 1210 West Dayton Street, Madison, WI 53706 (ferris@cs.wisc.edu).
‡Department of Industrial Engineering, 1513 University Ave., University of Wisconsin, Madison, WI 53706 (jin-ho@cs.wisc.edu).
§Department of Radiation Oncology, University of Maryland School of Medicine, 22 South Green Street, Baltimore, MD 21201 (dshep001@umaryland.edu).

FIG. 1.1. *Gamma knife treatment unit.*

Specifically, we consider treatment planning for a specialized device known as the *gamma knife*, which provides an advanced stereotactic approach to the treatment of tumors, vascular malformations, and pain disorders within the head [7]; see Figure 1.1. Inside a shielded treatment unit, beams from 201 cobalt-60 radioactive sources are focused so that they intersect at a certain point in space, producing an ellipsoidal region of high radiation dose referred to as a *shot*. A typical treatment consists of a number of shots, of possibly different sizes and different durations, centered at different locations in the tumor, whose cumulative effect is to deliver a certain dose to the treatment area while minimizing the effect on surrounding tissue.

Treatment goals can vary from one neurosurgeon to the next. Therefore, a treatment planning tool must be able to accommodate several different requirements. Three typical such requirements are homogeneity, conformity, and avoidance. Homogeneity requires that the complete target volume must be covered by a dose that has intensity at least $\beta\%$ of the maximum delivered dosage. The conformity requirement minimizes the dose to the nontarget volume. Avoidance requirements limit the amount of dosage that is delivered to certain critical structures near to the target area. There are standard rules established by the American Medical Association that determine minimum homogeneity and conformity requirements.

The motivation for this problem, and the approaches that form the basis of this work have appeared elsewhere [5, 6, 14]. The key contributions of this paper are as follows:

1. The description and implementation of a heuristic approach to generate a good starting point for the nonlinear programs used to model the treatment planning approach (see section 3). The approach is based on skeletonization ideas from computational graphics, is augmented using various optimization subproblems, and leads to improved speed and quality of solutions (see section 4).
2. Some practically motivated changes to the underlying models to improve robustness of the solution process and quality of the resulting treatment plan.

In particular, several nonlinear programs have been replaced by a single (easy to solve) mixed integer program, some "hard constraints" have been remodeled using inexact penalization, and least squares optimization has been used for parameter estimation (see section 2).

3. Tuning of the model parameters to improve solution speed and robustness (see section 4).

The resulting tool provides solutions to the problems that are currently under study at the University of Maryland Medical School. The work described here has enabled the simple prototype to be enhanced to the state in which it is usable without the intervention of an optimization expert, as a mechanism for robustly improving the operation of a complex medical system.

2. Models and solution process. The first step in building a treatment planning tool is to model the dose delivered to the patient by a given shot that is centered at a given location. A nonlinear least squares model for this was developed in [5]. The total dose delivered to a voxel (i, j, k) from a given set of shots can be calculated as

$$Dose(i,j,k) = \sum_{(s,w)\in\mathcal{S}\times\mathcal{W}} t_{s,w} D_w(x_s, y_s, z_s, i, j, k), \tag{2.1}$$

where $\mathcal{S} \in \{1, 2, \ldots, n\}$ denotes the set of n shots considered in the optimization, $w \in \mathcal{W}$ denotes the discrete width of a shot, $t_{s,w}$ is the time for which each shot (s, w) is exposed, and $D_w(x_s, y_s, z_s, i, j, k)$ is the dose delivered to the voxel (i, j, k) by the shot of size w that is centered at (x_s, y_s, z_s):

$$D_w(x_s, y_s, z_s, i, j, k) = \sum_{p=1}^{2} \lambda_p \left(1 - \operatorname{erf}\left(\frac{\sqrt{(i-x_s)^2 + \mu_p^y (j-y_s)^2 + \mu_p^z (k-z_s)^2} - r_p}{\sigma_p} \right) \right).$$

The notation $\operatorname{erf}(x)$ represents the integral of the standard normal distribution from $-\infty$ to x. We fit the ten parameters λ_p, μ_p^y, μ_p^z, r_p, and σ_p to the measured data via least squares, with different values for each shot width (see [5] for details). These values were then fixed at their computed values, and the expression for dose given in (2.1) was used as the core of the optimization models described in the remainder of this paper

2.1. Basic model and formulation. The basic optimization problem is to determine *a set of coordinates* (x_s, y_s, z_s), *a discrete set of collimator sizes* w, and *radiation exposure times* $t_{s,w}$. The main models used in the treatment planning process are nonlinear and mixed integer programs, defined over a (grid) subset $\mathcal{G}$ of the voxels in the target $\mathcal{T}$.

At the core of the model lie the requirements for homogeneity, conformity, and avoidance. Since these requirements are conflicting, a variety of techniques can be used to balance their relative imposition. It is easy to specify homogeneity in the models simply by imposing lower and upper bounds on the dose delivered to voxels in the target $\mathcal{T}$ and minimizing the dose outside the target. Similar bounding techniques can be used for avoidance requirements. Typically, however, the imposition of rigid bounds leads to plans that are overly homogeneous and not conformal enough; that

is, they provide too much dose outside the target. To overcome this problem, the notion of "underdose" was suggested in [5]:

$$UnderDose(i,j,k) := \max\{0, \theta - Dose(i,j,k)\}. \tag{2.2}$$

Informally, underdose measures how much the delivered dose is below the prescribed dose θ on the target voxels. Our basic model attempts to minimize the sum of the underdose on $\mathcal{G}$ subject to constraints on conformity, homogeneity, and avoidance.

To this point, our discussion has omitted the fact that we can use only a certain number of size/location combinations in the treatment plan. Choosing the particular shot size at each location is a discrete optimization problem that is treated by approximating the step function

$$H(t) = \begin{cases} 1 & \text{if } t > 0, \\ 0 & \text{if } t = 0 \end{cases}$$

by a nonlinear function,

$$H(t) \approx H_\alpha(t) := \frac{2\arctan(\alpha t)}{\pi} \quad .$$

For increasing values of α, H_α becomes a closer approximation to the step function H for $t \geq 0$. This process is typically called smoothing.

The set of shot sizes for a given number of shots n is chosen by imposing the constraint

$$n = \sum_{(s,w)\in\mathcal{S}\times\mathcal{W}} H_\alpha(t_{s,w}). \tag{2.3}$$

This states that the total number of size/location combinations to be used is n.

The basic model attempts to minimize the underdose to the target, subject to (2.3) and a constraint that the conformity of the plan exceed a certain (specified) value:

$$\begin{aligned}
&\min && \sum_{(i,j,k)\in\mathcal{G}} UnderDose(i,j,k), \\
&\text{subject to} && Dose(i,j,k) = \sum_{(s,w)\in\mathcal{S}\times\mathcal{W}} t_{s,w} D_w(x_s, y_s, z_s, i, j, k), \\
& && \theta \leq UnderDose(i,j,k) + Dose(i,j,k), \\
& && 0 \leq UnderDose(i,j,k), \\
& && 0 \leq Dose(i,j,k) \leq 1 \quad \forall (i,j,k) \in \mathcal{G}, \\
& && C\frac{\mathcal{N}_\mathcal{G}}{\mathcal{N}} \leq \frac{\sum_{(i,j,k)\in\mathcal{G}} Dose(i,j,k)}{\sum_{(s,w)\in\mathcal{S}\times\mathcal{W}} \bar{D}_w t_{s,w}}, \\
& && n = \sum_{(s,w)\in\{1,\ldots,n\}\times\mathcal{W}} H_\alpha(t_{s,w}), \\
& && 0 \leq t_{s,w} \leq \bar{t}.
\end{aligned} \tag{2.4}$$

The constraints involving $UnderDose$ coupled with the objective function enforce the definition given in (2.2).

C is an input parameter that specifies the conformity—it is multiplied by $\mathcal{N}_\mathcal{G}/\mathcal{N}$ to account for the fact that the number of target voxels in the grid $\mathcal{N}_\mathcal{G}$ is typically smaller than the total number of voxels $\mathcal{N}$ in the target. In practice, for solution performance, the constraint involving C is rearranged as a linear constraint by rationalizing the denominator. The conformity index value C must be given in advance. We describe how to estimate the value of C and the data $\bar{D}_w$ for a specific tumor in section 2.2.

This model is essentially the same as described in [5], except that an upper bound has been applied to the exposure times. While this upper bound was motivated by application-specific considerations, it also helps increase solution robustness.

The mechanism for updating both $\mathcal{G}$ and α is described in section 2.3.

2.2. Conformity estimation. The conformity of the plan is harder to deal with since it involves voxels outside of the target, of which there may be many. Furthermore, a reasonable conformity for a given patient plan is very hard to estimate a priori since it depends critically on the number of shots allowed and how the volume of the target interacts with the volumes of the allowable shots.

The conformity index C is an estimate of the ratio of the dose delivered to the target divided by the total dose delivered to the patient. The latter quantity is estimated by summing the (measured) dose delivered ($\bar{D}_w$) by a shot of size w for length $t_{s,w}$ to a "phantom." Thus C is calculated by the following expression:

$$C = \frac{\sum_{(i,j,k)\in\mathcal{T}} Dose(i,j,k)}{\sum_{(s,w)\in\mathcal{S}\times\mathcal{W}} \bar{D}_w t_{s,w}}.$$

Note that there are standard rules established by various professional and advisory groups that specify acceptable conformity requirements. In previous work [5], we attempted to estimate C by minimizing the total dose to the target, subject to hard constraints on the amount of dose delivered at each voxel in the target. However, instead of enforcing these hard constraints, we now propose the following optimization model as a mechanism for determining C:

$$\begin{aligned}
&\min && \sum_{(s,w)\in\mathcal{S}\times\mathcal{W}} \bar{D}_w t_{s,w},\\
&\text{subject to} && Dose(i,j,k) = \sum_{(s,w)\in\mathcal{S}\times\mathcal{W}} t_{s,w} D_w(x_s,y_s,z_s,i,j,k),\\
&&& \theta \le UnderDose(i,j,k) + Dose(i,j,k),\\
&&& 0 \le UnderDose(i,j,k),\\
&&& 0 \le Dose(i,j,k) \le 1 \quad \forall (i,j,k)\in\mathcal{T},\\
&&& \sum_{(i,j,k)\in\mathcal{T}} UnderDose(i,j,k) \le \mathcal{N}\mathcal{P}_\mathcal{U},\\
&&& n = \sum_{(s,w)\in\{1,\dots,n\}\times\mathcal{W}} H_\alpha(t_{s,w}),\\
&&& 0 \le t_{s,w} \le \bar{t}.
\end{aligned} \tag{2.5}$$

The crucial constraint is the one involving both $\mathcal{N}$, the number of voxels in the target, and $\mathcal{P}_\mathcal{U}$, a user-supplied estimate of the "average percentage" underdose allowable on the target. By increasing the value of $\mathcal{P}_\mathcal{U}$, the user is able to relax the homogeneity requirement, thereby reducing the total dose delivered to the patient.

TABLE 2.1
Comparison of conformity estimation models.

Patient	Old conformity model			New conformity model		
	C	Obj.val.	time	C	Obj.val.	time
Patient 5	0.296	28.89	106.1	0.296	25.68	77.4
	(0.007)	(13.93)	(32.9)	(0.005)	(12.93)	(17.3)
Patient 6	0.246	17.81	397.0	0.247	14.89	358.3
	(0.011)	(14.54)	(90.5)	(0.009)	(13.21)	(56.2)
Patient 8	0.323	3.33	195.2	0.323	2.86	167.6
	(0.007)	(2.73)	(60.8)	(0.003)	(1.79)	(56.3)

(The model from [5] forced the underdose to be zero at every voxel in the target.) Notice that reducing the total dose delivered to the patient typically increases C. Thus, C is essentially a monotone function of $\mathcal{P}_{\mathcal{U}}$. The upper bound on exposure time $\bar{t}$ is typically chosen as a large fraction of the maximum dose delivered to $\mathcal{T}$ (here assumed to be 1) for the purposes of improving solver performance.

Table 2.1 indicates the motivation for this change. For a variety of patients, the estimate of C is essentially the same, but it has smaller standard deviation (indicated in parentheses) and smaller computing times. (For each of the patients, the starting point for the conformity problem was randomly perturbed by up to two voxels in each coordinate direction to generate the sample. The variance is calculated over a set of 30 runs.) Furthermore, it seems clear that the final objective values arising from the subsequent solves are better if these solves are seeded with the new conformity estimation model solutions.

2.3. Solution process. A series of the following five optimization problems are solved to determine the treatment plan. The reason the basic model in section 2.1 is solved iteratively (steps 2, 3, and 4) is an effort to reduce the total time required to find the solution. Our experience shows that combining those three steps into one increases the time to converge at least three-fold, which is often not clinically acceptable.

1. Conformity estimation. In order to avoid calculating the dose delivered outside of the target, we first solve an optimization problem on the target to estimate an "ideal" conformity for the particular patient for a given number of shots; details can be found in section 2.2. The conformity estimate C is passed to the basic model as an input parameter.
2. Coarse grid estimate. Given the estimate of conformity C, we then specify a series of optimization problems whose purpose is to minimize the total underdose on the target for the given conformity. In order to reduce the computational time required to determine the plan, we first solve (2.4) on a coarse grid subset of the target voxels. We have found it beneficial to use in the model one or two more shot locations than the number requested by the user, that is, $\mathcal{S} := \{1, \ldots, n+2\}$, and allow the optimization not only to choose useful sizes but also to discard the extraneous shot locations.
3. Refined grid estimate. To keep the number of voxels in the optimization as small as possible, we add to the coarse grid only those voxels on a finer grid for which the homogeneity (bound) constraints are violated. This procedure

improves the quality of the plan without greatly increasing the execution time.

Note that it is possible for the solution from a previous optimization in this sequence to suggest that multiple shots be centered at the same location (i.e., for a given s there are several nonzero $t_{s,w}$). If, in addition, there are other locations s' that are not used at all in the solution at hand, we shift as many of the multiple shots as possible to these unused locations. This maintains the objective value of the current solution while giving any subsequent solves the ability to move the different size shots independently. In our automatic procedure we shift the largest value of $t_{s,w}$ to the unused location.

4. Shot reduction problem. In the solution steps given above, we use a small value of α, typically 6, to impose the constraint (2.3) in an approximate manner. In the fourth solve, we increase the value of α to 100 in an attempt to force the planning system to choose which size/location pairs to use. At the end of this solve, there may still exist some size/location pairs that have very small exposure times t. Also note that our solution technique does not guarantee that the shots are centered at locations within the target.
5. Fixed location model. The computed solution may have more shots used than the user requested and furthermore may not be implementable on the gamma knife since the coordinate locations cannot be keyed into the machine. Our approach to refining the optimization solution in order to generate implementable coordinates for the shot locations is to round the shot location values and then fix them. Once these locations are fixed, the problem becomes linear in the intensity values t. We reoptimize these values and force the user-requested number of size/location pairs precisely, using a mixed integer program. Further details can be found in section 2.4.

Note that the starting point for each of the models is the solution point of the previous model. Details on how to generate an effective starting point for the first model are given in section 3. All the optimization models are written in the GAMS [3] modeling language and solved using CONOPT [4] or CPLEX [10].

2.4. Fixed location model. In order to implement the solution on the gamma knife, we round the location values from the fourth solve and fix them at $\bar{x}_s$, $\bar{y}_s$, and $\bar{z}_s$, respectively. The values of $D_w(\bar{x}_s, \bar{y}_s, \bar{z}_s, i, j, k)$ can then be calculated at each location (i, j, k) as data. The final optimization involves the following mixed integer linear optimization problem:

$$
\begin{aligned}
\text{min} \quad & \sum_{(i,j,k)\in\mathcal{G}} UnderDose(i,j,k), \\
\text{subject to} \quad & Dose(i,j,k) = \sum_{(s,w)\in\mathcal{S}\times\mathcal{W}} t_{s,w} D_w(\bar{x}_s, \bar{y}_s, \bar{z}_s, i, j, k), \\
& \theta \le UnderDose(i,j,k) + Dose(i,j,k), \\
& 0 \le UnderDose(i,j,k), \\
& 0 \le Dose(i,j,k) \le 1 \quad \forall (i,j,k) \in \mathcal{G}, \\
& C\frac{\mathcal{N}_{\mathcal{G}}}{\mathcal{N}} \sum_{(s,w)\in\mathcal{S}\times\mathcal{W}} \bar{D}_w t_{s,w} \le \sum_{(i,j,k)\in\mathcal{G}} Dose(i,j,k), \\
& 0 \le t_{s,w} \le \psi_{s,w}\bar{t},
\end{aligned}
\tag{2.6}
$$

$$\sum_{(s,w)\in\mathcal{S}\times\mathcal{W}} \psi_{s,w} \le n,$$
$$\psi_{s,w} \in \{0,1\}.$$

This model was adapted from the work described in [6]. The key observation is the use of the binary variable $\psi_{s,w}$ to indicate whether a shot of size w is used at location s. The penultimate constraint in the model ensures that no more than n shots are used, while the upper bound on t ensures that no exposure time occurs if the corresponding shot is not used. In previous work [5], we had used increasing values of α coupled with the removal of small shots in a nonlinear programming approach. The current scheme is guaranteed to outperform this.

It may, of course, be possible to extend this model to include more locations, but this was not deemed necessary for our work. Furthermore, it could be argued that the basic model should use integer variables to enforce the discrete size choices. Our investigations found such approaches to be impractical and not as robust as the scheme outlined above.

3. Starting point generation. A good starting point is very important for nonlinear programs, especially if the problem is not convex. This section will explore some techniques for finding an initial starting solution for our solution process. The main focus is to find a set of good shot locations and their corresponding sizes. We propose a shot location and size determination (SLSD) process based on three-dimensional (3D) medial axis transformation. Our results show that it takes no more than 6 seconds to produce a good starting solution for all the 3D data considered in our research.

Our targets are collections of 3D voxels. For the large scale problems of interest, the data manipulation and optimization solution times are much larger than allowable (typically 20–40 minutes is allowed for planning), and we must resort to data compression. One technique used extensively in computer vision and pattern recognition is the notion of a skeleton, a series of connected lines providing a simple representation of the object at hand [1, 8, 11, 15, 18]. Skeletons have been used by physicians and scientists to explore virtual human body organs with noninvasive techniques [9, 17]. The term skeleton was proposed in [1] to describe the axis of symmetry, based on the physical analogy of grassfire propagation, namely, the locus of centers of maximal disks (balls) contained in a two- (three-) dimensional shape.

Some applications require that the original object be reconstructed from the compact representation, and hence the normal measure of goodness is the error between the original and reconstructed object. However, in our case, we will just use the skeleton to quickly generate good starting shot locations for the nonlinear program. Thus we adapt techniques from the literature to achieve these goals.

Our process occurs in three stages. First we generate the skeleton, then we place shots and choose their sizes along the skeleton to maximize a measure of our objective. After this, we choose the initial exposure times using a simple linear program. Finally, we apply the five-stage optimization process outlined in section 2 to improve upon the starting points found.

3.1. Skeleton generation. In this section, we introduce a 3D skeleton algorithm that follows procedures similar to those of [17]. The first step in the skeleton generation is to compute the contour map containing distance information from each voxel to a nearest target boundary. The ideal distance metric is Euclidean, but this is too time-consuming to implement in a 3D environment.

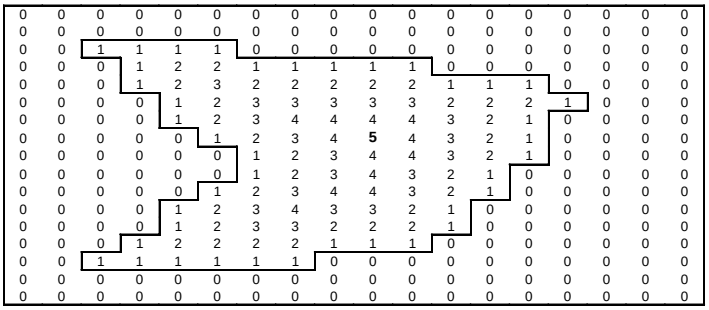


FIG. 3.1. *A contour map on a two-dimensional example.*

To describe our simpler scheme, we first introduce some terminology.

DEFINITION 3.1. *Considering a voxel i as a 3D box, an adjacent voxel j is called an F-neighbor of i if j shares a face with i, an E-neighbor of i if j shares an edge with i, and a V-neighbor of i if j shares a vertex with i.*

Our procedure is as follows:

1. Assign 0 to the nontarget area, and let $v = 0$.
2. Assign $v+1$ to any voxel that is unassigned and has an F-neighbor with value v.
3. Increment v by 1 and repeat until all voxels in the target area are assigned.

An example of a two-dimensional (2D) contour map generated through this procedure is shown in Figure 3.1.

Note that if the maximum height in the contour map is less than 2, we terminate the skeleton generation process.

Extracting an initial skeleton. Based on the contour map, there are several known skeleton extraction methods in the literature [17]: *boundary peeling* (also called thinning) [12], *distance coding* (distance transformation) [13], and *polygon-based Voronoi methods* [2]. Because it is simple and fast, we use the distance transformation method to generate a skeleton. In our terminology, this means that we define a *skeleton point* as a voxel whose contour map value is greater than or equal to those of its E-neighbors.

Refinement for connectivity of a thin skeleton. We say that two skeleton points are *connected* if they are V-neighbors. Unfortunately, not all the skeleton points generated will be connected, and thus we use a two-stage process to connect the pieces of the skeleton together.

For example, Figure 3.2(a) shows a raw skeleton with several disconnected components. We use two algorithms to join all the disconnected components. The first algorithm is a *directional search* algorithm. The second is the *shortest path* algorithm. After these refinements, we have a connected skeleton as seen in Figure 3.2(b).

We first use depth-first search to label each skeleton point as belonging to a particular component of the skeleton. The first connection phase is a steepest ascent technique. Consider the contour map as a function f. We calculate an approximate gradient ∇f using coordinatewise central divided differences. Thus, for each voxel (i, j, k) we use the values of f at each of its F-neighbors to generate a 3D vector

$$\begin{aligned}\nabla f(i,j,k) := (&\operatorname{sgn}(f(i+1,j,k) - f(i-1,j,k)),\\ &\operatorname{sgn}(f(i,j+1,k) - f(i,j-1,k)),\\ &\operatorname{sgn}(f(i,j,k+1) - f(i,j,k-1)))\end{aligned}$$

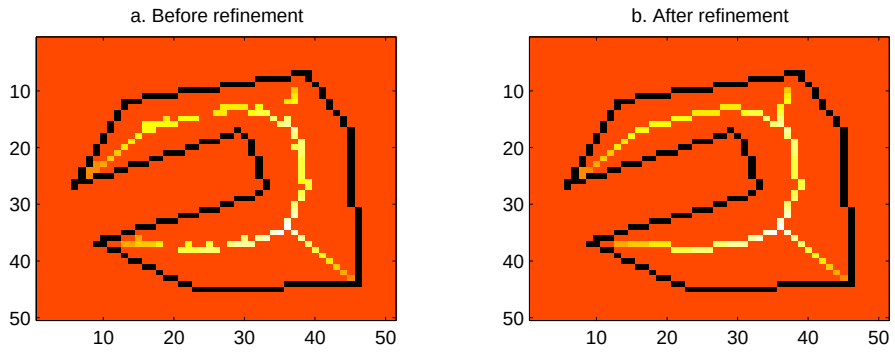


FIG. 3.2. *An example of skeleton refinement.*

and store these in a divided difference table. Given the voxel (i, j, k), we evaluate f at the V-neighbor $(i, j, k) + \nabla f(i, j, k)$ and accept the move if f does not decrease. We terminate the process if either f decreases or we move to a voxel in a different piece of the skeleton, thus connecting (i, j, k) to this piece. Including the paths generated in this fashion in the skeleton typically connects pieces that are close but not currently connected.

The directional search algorithm, while joining many of the disconnected pieces of the skeleton along ridges of the contour map, may fail in cases where the value of the contour map decreases in the gap between two disconnected pieces. Therefore, the second connection phase uses a shortest path algorithm to connect the skeleton (instead of using the *saddle point method* discussed in [17]).

Let $\mathcal{K}$ be the set of all skeletal points, divided into d disconnected components. In order to reduce the search space for the shortest path algorithm, we generate a cloud of voxels $\mathcal{C}$ in the target volume, each of which are local maxima among their F-neighbors. Note that $\mathcal{C}$ contains $\mathcal{K}$ by definition and can be thought of heuristically as a cloud of points encircling the skeleton. We will only join the disconnected components of $\mathcal{K}$ using points in $\mathcal{C}$.

Let each voxel in $\mathcal{C}$ be a node. An arc $(i, j) \in \mathcal{A} \subseteq \mathcal{C} \times \mathcal{C}$ is defined if voxels i and j are V-neighbors.

We choose an arbitrary voxel in an arbitrary component as the source node s. A representative node is chosen arbitrarily from each of the remaining components and joined to a dummy node t that will be the destination. The distance c_{ij} between voxels in a connected cluster is assigned a value of 0, whereas other V-neighbors of a given voxel are at distance 1. We attempt to send $d - 1$ units of flow from s to t. We also add an arc from s to t directly with a high cost to allow for the fact that it may not be possible to join every component through $\mathcal{C}$. If this is the case, it will be signified by flow along these final arcs. The complete formulation of our problem follows:

$$\begin{aligned}
\min \quad & \sum_{(i,j)\in\mathcal{A}} c_{ij}x_{ij}, \\
\text{subject to} \quad & \sum_{\{j|(i,j)\in\mathcal{A}\}} x_{ij} - \sum_{\{j|(j,i)\in\mathcal{A}\}} x_{ji} = \begin{cases} (d-1) & if \quad i = s, \\ -(d-1) & if \quad i = t, \\ 0 & otherwise, \end{cases} \\
& 0 \le x_{ij} \quad \forall (i,j) \in \mathcal{A}.
\end{aligned}$$

Typically, this problem is solved very quickly by standard linear programming algorithms, even though specialized network flow algorithms could be applied.

3.2. Shot placement. At this stage, we recall that our goal is to determine where to place shots and how large to make them initially; the skeleton generation is a data reduction technique to facilitate this goal. We restrict our attention to points on the skeleton. This is reasonable, since the dose delivered (2.1) looks ellipsoidal in nature, and hence being centrally located within the target (that is, on the skeleton) is preferable.

Our approach moves along the skeleton evaluating whether the current point is a good location at which to place a shot. There are two special types of skeleton points, an end point and a cross point, that help determine the shot size and the location; see Figure 3.3.

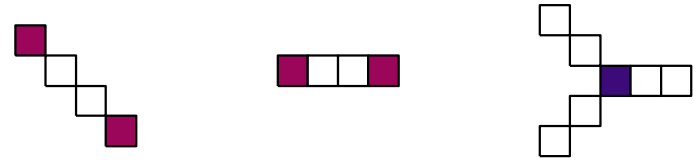


FIG. 3.3. *Examples of end points and a cross point.*

We define an end point and a cross point as follows.

DEFINITION 3.2. *A voxel is an* end point *if*

1. *it is in the skeleton,*
2. *it has only one V-neighbor in the skeleton.*

A voxel is a cross point *if*

1. *it is in the skeleton,*
2. *it has at least three V-neighbors,*
3. *it is a local maximum in the contour map.*

These points are respectively the start (end point) and finish (cross point) points for our heuristic.

Let $\mathcal{K}$ be a set of skeletal points in the target volume. The first phase of the methods determines all end points in the current skeleton. Given an end point $(x, y, z) \in \mathcal{K}$, we carry out the following steps to generate a stack for the end point:

1. Calculate a merit value at the current location. Save the location information, the best shot size, and the merit value on a stack.
2. Find all V-neighbors of the current point, in the skeleton, that are not in the stack. If there is exactly one neighbor, make the neighbor the current location and repeat these two steps. Otherwise, the neighbor is a cross point or an end point, and we terminate this process.

If the length of the stack is less than 3, then we discard these points from the skeleton. Otherwise, we choose the shot location and size determined by the smallest merit value on the stack. This shot will cover a subset of the voxels in the target; these voxels are removed from the target at this stage.

We then move to the next end point and repeat the above process. Once all end points have been processed, we attempt to generate a new skeleton based on the remaining (uncovered) voxels in the target. We then repeat the whole process with the new skeleton.

The key to this approach is the merit function. Ideally, we would like to place shots that cover the entire region, without overdosing within (or outside) of the target. Overdosing occurs outside the target if we choose a shot size that is too large for the current location, and hence the shot protrudes from the target. Overdosing also occurs within the target if we place two shots too close together for their chosen sizes.

Thus, if we label *height* as the approximate Euclidean distance from the current point to the target boundary, *spread* as the minimum distance between the current location and the end point at which we started, and w as the shot size, we would like to ensure that all three of these measures are as close as possible. Therefore, we choose an objective function that is a weighted sum of squared differences between these three quantities:

1. $\Phi_{sh}(x, y, z) := (spread(x, y, z) - height(x, y, z))^2$,
2. $\Phi_{sw}(x, y, z, w) := (spread(x, y, z) - w)^2$,
3. $\Phi_{hw}(x, y, z, w) := (height(x, y, z) - w)^2$.

The first function ensures that we pack the target volume as well as possible; that is, the current spread between shots should be close to the distance to the closest target boundary. The second function is used to choose a helmet size that fits the skeleton best for the current location. The third function favors a location that is the appropriate distance from the target boundary for the current shot size.

Our objective function Φ is defined as a linear combination (with weights λ) of these penalty functions and a fourth $(\bar{w} - w)^2$, which is designed to favor large shot sizes. Note that $\bar{w}$ is the maximum shot width at hand, typically 18mm. The weights can be adjusted based on a user's preference. In practice we use 1/3 for the first three objective weights, and 1/2 for the fourth.

3.3. Modifying the number of shots used. Often, the application expert knows based upon experience how many shots will be needed to treat a specific tumor. The planning tool accepts this information as input. However, the SLSD procedure uses only target information, and it might suggest using fewer or more shots.

If the number of shots generated by SLSD is too large, the first $n + 2$ shots are used as the starting point. We allow the nonlinear program to adjust the locations further and remove the least useful shots during the solution process.

If the number of shot locations obtained from the SLSD procedure is lower than the requested number, we add extra shot locations using the following (SemiRand) heuristic. The key idea is to spread out the shot center locations with appropriate shot sizes over the target area.

We assume that we are given ρ, an estimate of the conformity that we require from any shot. In practice, we choose this value as 0.2. We then generate k different shot/size combinations as follows. First, a random location s is generated from the target area that is not covered by the current set of shots. Second, a random shot size w for the specific location is generated within the set of different shots available $\mathcal{W}$. For each shot/size combination we calculate the fraction $f(s, w)$ of the dose that hits the target by taking the ratio of the number of voxels that it hits in the target to the total number of voxels in a shot of the given size.

We decide the location and size (s, w) to use as follows. If $\max f(s, w) \leq \rho$, then we choose the combination that maximizes $f(s, w)$. Otherwise, amongst all those combinations that are acceptable (i.e., $f(s, w) \geq \rho$), we choose the largest one (i.e., the one that maximizes w among these).

Note that the SemiRand scheme can be used in cases where the SLSD procedure fails (when a 3D volume of the target cannot be defined) and also as an alternative scheme for locating starting points. In practice we use $k = 5$.

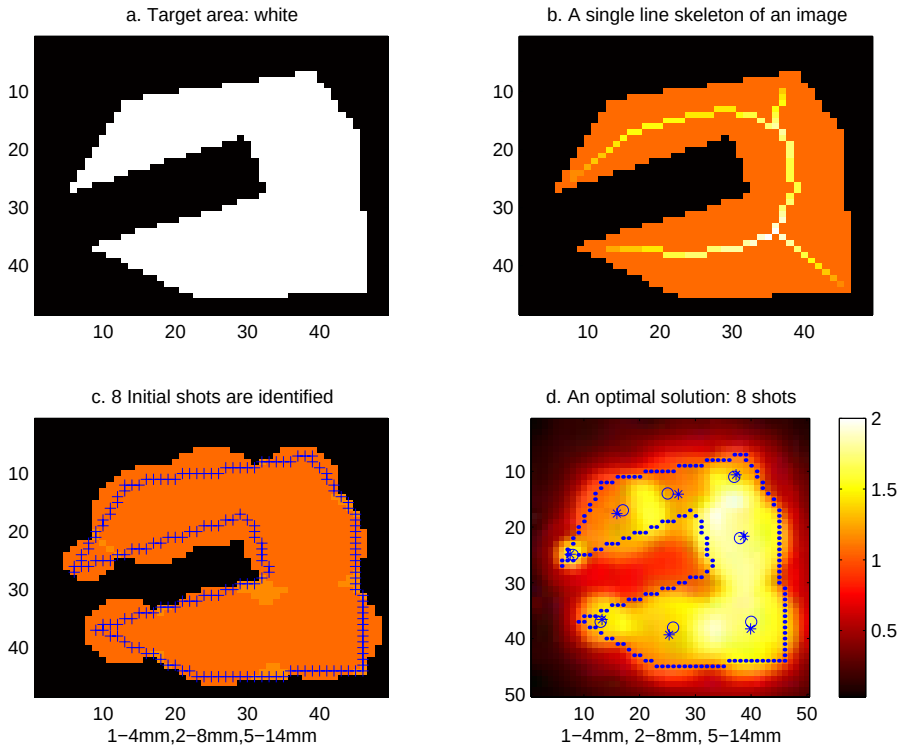


FIG. 4.1. *Computational results on a 2D example; axes represent pixel labeling.*

4. Computational results. In this section, we demonstrate how to use the techniques outlined above on 2D testing problems as well as real patient data.

4.1. Examples on 2D problems. We start with some simple 2D examples that show the types of skeletons that are produced and portray the resulting optimization solutions.

Figure 4.1(a) depicts a particular target (*tumor*) area for our problem as white space. This tumor is approximately 3 inches square. The shape is not convex: It has an indentation that makes it difficult for a normal optimization model to obtain an acceptable plan. Figure 4.1(b) shows a thin line skeleton generated from the image. The skeleton generation process takes less than 1 second on a Pentium III 800MHz workstation. We then apply the SLSD process to obtain the starting solution for the nonlinear programming (NLP) model as shown in Figure 4.1(c). Eight shots of radiation are used for this example: one 4 mm, two 8 mm, and five 14 mm width shots. We use 0.9 as the initial exposure times in the model. The solution covers the target area well. We solve the conformity estimation optimization model using the CONOPT2 interface with the starting solution, finding an optimal solution of 8 shots in 61 seconds of execution time. Figure 4.1(d) shows the resulting plot obtained using the MATLAB image toolbox. The circles are the starting solution, and the stars are the optimal solution from CONOPT. They are almost identical in shot center locations. The SLSD process outperforms a random starting solution. Given 8 shots to use, the NLP model using a random starting solution finds an optimal solution in 1122 seconds.

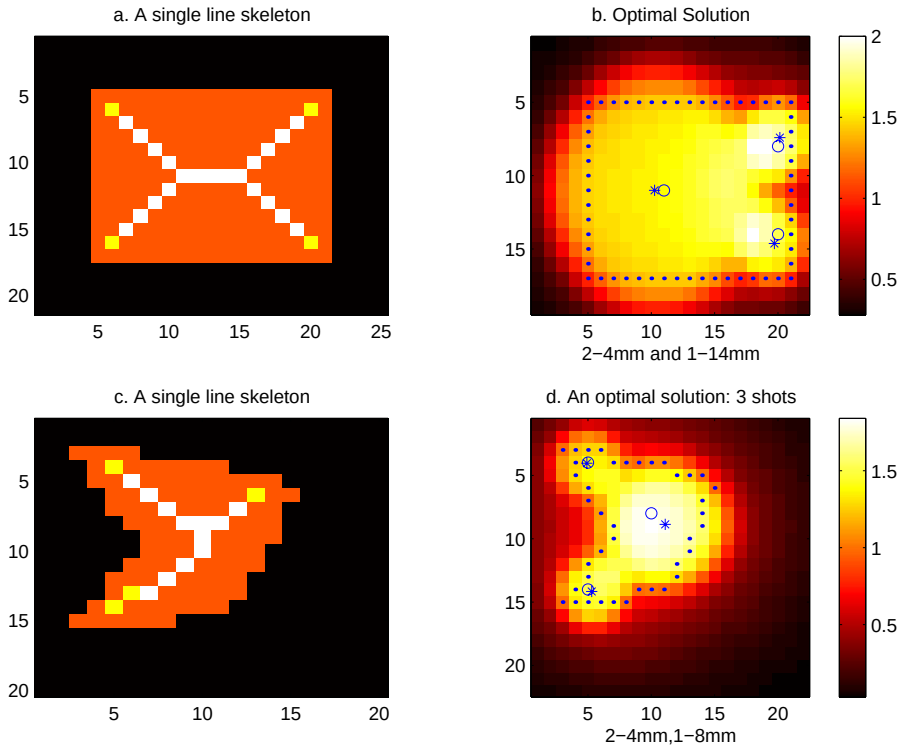


FIG. 4.2. *2D examples: a rectangular target* (a),(b) *and a small target* (c),(d).

We show two more results on other examples in Figure 4.2. Figure 4.2(a) is a rectangular target for which three shots are used. The optimization model finds the solution of two 4mm and one 14mm shots, depicted in Figure 4.2(b). The total time to produce the solution is about 15 seconds. Another example is given in Figure 4.2(c)–(d). This is a small tumor (less than 1 inch square) for which three shots are again used. The SLSD model takes 1.5 seconds to generate the starting solution. The NLP model finds an optimal solution of two 4mm and one 8mm shots in 6 seconds.

4.2. Application to real patient data. We have tested our techniques on ten targets arising from real patient cases. The ten targets are radically different in size and complexity. The tumor volumes range from 28 voxels to 36088 voxels. Since our problems are not convex, the choice of parameters in their solution can also have dramatic effects. In this section, we demonstrate how to choose good parameters for the NLP models. Some further description of the medical implications of these results is given in [14].

We generate good initial shot center locations and sizes by running SLSD. This is a starting solution for the NLP model with an exception of shot exposure times. These times $t_{s,w}$ are estimated using the following simple linear program:

$$
\begin{aligned}
&\min && \sum_{(i,j,k)\in\mathcal{G}} UnderDose(i,j,k),\\
&\text{subject to} && Dose(i,j,k) = \sum_{(s,w)\in\mathcal{S}\times\mathcal{W}} t_{s,w} D_w(\bar{x}_s,\bar{y}_s,\bar{z}_s,i,j,k),\\
& && \theta \le UnderDose(i,j,k) + Dose(i,j,k),
\end{aligned}
\tag{4.1}
$$

$$0 \le UnderDose(i,j,k),$$
$$0 \le Dose(i,j,k) \le 1 \quad \forall (i,j,k) \in \mathcal{G},$$
$$\underline{t} \le t_{s,w} \le \bar{t}.$$

Note that we fix the locations of the shots at the points suggested by SLSD and only update the exposure times. Furthermore, we ensure that every size shot has positive weight in an initial solution by enforcing a lower bound (typically 0.1) on the exposure lengths.

The procedure for varying α (controlling the enforcement of the discrete choices) can have a dramatic effect on solution quality and times. We generated solutions for a variety of patients under a number of different choices of α. These solutions were analyzed by an application expert. Based on his feedback, we suggest using initial values of α between 4 and 8.

TABLE 4.1
Average optimal objective value and solution times in seconds for different tumors.

Patient	Objective			Time		
(#voxels)	Random	SemiRand	SLSD	Random	SemiRand	SLSD
1	2.17	0.88	NA	0.3	0.3	NA
(28)	(0.86)	(0.29)	NA	(0.05)	(0.03)	NA
2	14.70	8.21	6.64	32	30	26
(2144)	(6.90)	(4.68)	(2.61)	(6)	(9)	(9)
3	27.53	19.22	14.43	89	67	52
(3279)	(19.07)	(8.87)	(14.99)	(25)	(16)	(9)
4	16.55	12.89	9.85	97	94	84
(3229)	(4.45)	(6.70)	(4.88)	(18)	(22)	(19)
5	34.87	34.53	23.85	153	128	77
(4006)	(16.36)	(17.26)	(13.84)	(40)	(30)	(17)
6	33.32	28.49	15.00	556	513	355
(6940)	(17.25)	(13.09)	(13.22)	(103)	(100)	(52)
7	35.45	29.97	31.03	590	460	343
(10061)	(12.63)	(11.16)	(13.65)	(228)	(100)	(75)
8	9.31	3.22	2.78	887	240	168
(22124)	(2.73)	(2.80)	(1.72)	(157)	(68)	(56)
9	45.05	35.18	31.05	874	629	498
(24839)	(18.10)	(7.11)	(10.25)	(425)	(166)	(99)
10	18.55	11.57	8.59	3568	937	695
(36088)	(11.20)	(11.83)	(6.71)	(589)	(108)	(79)

Table 4.1 shows average objective values of three different starting solution generation techniques: Random, SemiRand, and SLSD. The objective value represents the total average underdose of the target when the solution is applied. The numbers in parentheses are the standard deviations from a batch of 50 perturbed runs. (In each run, the set of initial solution locations (x, y, z) were perturbed voxel by voxel by a distance of no more than two voxels.) We compare the techniques based on the final objective values and the run times. By fixing $\alpha = 6$, 50 perturbed runs were made for each patient-method pair. In each run, we generated initial locations

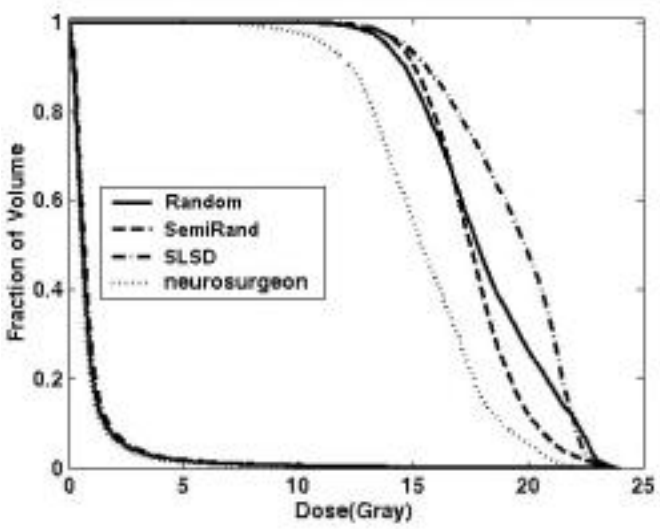


FIG. 4.3. *A dose-volume histogram for patient* 6.

randomly within the target for the random scheme, while location perturbation was used for SemiRand and SLSD. The tumor was so small for Patient 1 that SLSD failed to generate a skeleton (maximum height in the contour map was less than 2).

Using standard statistical tests, the pairwise p-value [16] between Random and SemiRand was 0.013, between Random and SLSD was 0.0006, and between SemiRand and SLSD was 0.078. This leads to the conclusion that these results are significantly different at the 90% confidence level.

Table 4.1 also shows average run times of the entire model for the seven different patients. Although a gain of speed using SLSD depends on the shape and size of the tumor, the table shows that the model execution time can be substantially reduced using SLSD over the other two techniques regardless of the size of tumor. Again, these results are significantly different at the 90% confidence level. The pairwise p-value between Random and SemiRand was 0.017, between Random and SLSD was 0.0006, and between SemiRand and SLSD was 0.063.

To conclude this section, we show a dose-volume histogram relating various plans that were generated for patient 6 (see Figure 4.3). The histogram depicts the fraction of the volume that receives a particular dose for both the skull and the target volumes. The curves on the right depict information related to the target, while on the left they refer to the skull. On the target, the curves that extend furthest to the right receive more dose. Since the target curves can be moved to the right by just delivering more dose to the patient's skull, the lines to the left show that the fraction of the skull receiving a particular dosage is essentially unchanged. The figure compares the three techniques outlined here, along with the actual plan used on the patient case. Clearly, all of the automatic plans are better than the neurosurgeon's plan, and the SLSD approach appears preferable to the other two automatic plans in quality.

5. Conclusion and future directions. We have used a variety of optimization techniques in this paper to develop an approach for solving a planning problem for medical treatment. While our approach has been tailored to the specific application, we believe the methods and approaches used here can be effectively adapted to many other problem classes.

The work described in this paper was motivated by feedback received from an initial prototype use of our planning tool at the University of Maryland Medical School. The key features that needed improvement were the speed and robustness of the process. This paper has addressed both issues by using a variety of different optimization models and computational techniques. In particular, the speed of solving the sequence of nonlinear programming models has been substantially reduced by using the skeleton-based starting point generation technique. Statistically, we have shown that SLSD outperforms two other heuristics for generating starting points. Furthermore, the use of an improved conformity estimation model, coupled with a "clean-up" mixed integer programming model, ensures that the solutions generated are clinically acceptable and conform to the input specifications of the user. The modified tool is now in use at the hospital without intervention from any of the authors.

Our future work involves predicting the number of shots that can be used for a particular patient.

REFERENCES

[1] H. Blum, *A transformation for extracting new descriptors of shape*, in Models for the Perception of Speech and Visual Form, W. Wathen-Dunn, ed., MIT Press, Cambridge, MA, 1967, pp. 362–380.

[2] J. W. Brandt and V. R. Algazi, *Continuous skeleton computation by Voroni diagram*, CVGIP: Graph Models and Image Processing, 55 (1992), pp. 329–338.

[3] A. Brooke, D. Kendrick, and A. Meeraus, *GAMS: A User's Guide*, The Scientific Press, South San Francisco, CA, 1988.

[4] A. Drud, *CONOPT: A GRG code for large sparse dynamic nonlinear optimization problems*, Math. Programming, 31 (1985), pp. 153–191.

[5] M. C. Ferris, J.-H. Lim, and D. M. Shepard, *Radiosurgery treatment planning via nonlinear programming*, Ann. Oper. Res., (2002), to appear.

[6] M. C. Ferris and D. M. Shepard, *Optimization of gamma knife radiosurgery*, in Discrete Mathematical Problems with Medical Applications, D.-Z. Du, P. Pardalos, and J. Wang, eds., DIMACS Ser. Discrete Math. Theoret. Comput. Sci. 55, American Mathematical Society, Providence, RI, 2000, pp. 27–44.

[7] J. C. Ganz, *Gamma Knife Surgery*, Springer-Verlag, Vienna, 1997.

[8] Y. Ge and J. M. Fitzpatrick, *On the generation of skeletons from discrete Euclidean distance maps*, IEEE Trans. Pattern Analysis and Machine Intelligence, 18 (1996), pp. 1055–1066.

[9] L. Hong, A. Kaufman, Y. Wei, A. Viswambharn, M. Wax, and Z. Liang, *3D virtual colonoscopy*, Proceedings of IEEE Symposium on Biomedical Visualization, Atlanta, GA, IEEE, Piscataway, NJ, 1995, pp. 26–32.

[10] ILOG CPLEX Division, *CPLEX Optimizer*, ILOG, Incline Village, NV.

[11] F. Leymarie and M. D. Levine, *Fast raster scan distance propagation on the discrete rectangular lattice*, Computer Vision Graphics Image Processing, 55 (1992), pp. 84–94.

[12] C. M. Mao and M. Sonka, *A fully parallel 3D thinning algorithm and its applications*, Computer Vision Image Understanding, 64 (1996), pp. 420–433.

[13] C. W. Niblack, P. B. Gibbons, and D. W. Capson, *Generating skeletons and centerlines from the distance transform*, CVGIP: Graph Models and Image Processing, 54 (1992), pp. 420–437.

[14] D. M. Shepard, M. C. Ferris, R. Ove, and L. Ma, *Inverse treatment planning for gamma knife radiosurgery*, Medical Phys., 27 (2000), pp. 2748–2756.

[15] Q. J. Wu and J. D. Bourland, *Morphology-guided radiosurgery treatment planning and optimization for multiple isocenters*, Medical Phys., 26 (1999), pp. 2151–2160.

[16] B. S. Yandell, *Practical Data Analysis for Designed Experiments*, Chapman and Hall, London, 1997.

[17] Y. Zhou, A. Kaufman, and A. W.Toga, *Three dimensional skeleton and centerline generation based on an approximate minimum distance field*, Visual Computers, 14 (1998), pp. 303–314.

[18] Y. Zhou and A. W. Toga, *Efficient skeletonization of volumetric objects*, IEEE Trans. Visualization and Computer Graphics, 5 (1999), pp. 196–209.

SIAM J. OPTIM.
Vol. 13, No. 3, pp. 938–945

STRONG CONVERGENCE OF A PROXIMAL-TYPE ALGORITHM IN A BANACH SPACE*

SHOJI KAMIMURA† AND WATARU TAKAHASHI‡

Abstract. In this paper, we study strong convergence of the proximal point algorithm. It is known that the proximal point algorithm converges weakly to a solution of a maximal monotone operator, but it fails to converge strongly. Then, in [*Math. Program.*, 87 (2000), pp. 189–202], Solodov and Svaiter introduced the new proximal-type algorithm to generate a strongly convergent sequence and established a convergence property for it in Hilbert spaces. Our purpose is to extend Solodov and Svaiter's result to more general Banach spaces. Using this, we consider the problem of finding a minimizer of a convex function.

Key words. proximal point algorithm, maximal monotone operator, Banach space, strong convergence

AMS subject classifications. 47H05, 47J25

PII. S105262340139611X

1. Introduction.

Let H be a real Hilbert space with inner product $\langle\cdot,\cdot\rangle$, and let $T\colon H \to 2^H$ be a maximal monotone operator. The problem of finding an element $x \in H$ such that $0 \in Tx$ is very important in the area of optimization and related fields. For example, if T is the subdifferential ∂f of a proper lower semicontinuous convex function $f\colon H \to (-\infty, \infty]$, then T is a maximal monotone operator and the equation $0 \in \partial f(x)$ is reduced to $f(x) = \min\{f(z) : z \in H\}$. One method of solving $0 \in Tx$ is the proximal point algorithm. Let I denote the identity operator on H. The proximal point algorithm generates, for any starting point $x_0 = x \in H$, a sequence $\{x_n\}$ in H by the rule

$$x_{n+1} = (I + r_n T)^{-1} x_n, \qquad n = 0, 1, 2, \ldots, \tag{1.1}$$

where $\{r_n\}$ is a sequence of positive real numbers. Note that (1.1) is equivalent to

$$0 \in Tx_{n+1} + \frac{1}{r_n}(x_{n+1} - x_n), \qquad n = 0, 1, 2, \ldots.$$

This algorithm was first introduced by Martinet [6] and generally studied by Rockafellar [10] in the framework of a Hilbert space. Later many authors studied the convergence of (1.1) in a Hilbert space; see Brézis and Lions [1], Lions [5], Passty [7], Güler [2], Solodov and Svaiter [11], and the references mentioned there. Rockafellar [10] proved that if $T^{-1}0 \neq \emptyset$ and $\liminf_{n\to\infty} r_n > 0$, then the sequence generated by (1.1) converges weakly to an element of $T^{-1}0$. Further, Rockafellar [10] posed an open question of whether the sequence generated by (1.1) converges strongly or not. This question was solved by Güler [2], who introduced an example for which the sequence generated by (1.1) converges weakly but not strongly. On the other hand, Kamimura and Takahashi [3, 4] and Solodov and Svaiter [12] recently modified the

*Received by the editors October 5, 2001; accepted for publication (in revised form) August 19, 2002; published electronically February 27, 2003.
http://www.siam.org/journals/siopt/13-3/39611.html
†Graduate School of International Corporate Strategy, Hitotsubashi University, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8439, Japan (kamimura@ics.hit-u.ac.jp).
‡Department of Mathematical and Computing Sciences, Tokyo Institute of Technology, Ohokayama, Meguro-ku, Tokyo 152-8552, Japan (wataru@is.titech.ac.jp).

proximal point algorithm to generate a strongly convergent sequence. Solodov and Svaiter [12] introduced the following algorithm:

$$\begin{cases} x_0 \in H, \\ 0 = v_n + \dfrac{1}{r_n}(y_n - x_n), \qquad v_n \in Ty_n, \\ H_n = \{z \in H : \langle z - y_n, v_n\rangle \le 0\}, \\ W_n = \{z \in H : \langle z - x_n, x_0 - x_n\rangle \le 0\}, \\ x_{n+1} = P_{H_n \cap W_n} x_0, \qquad n = 0, 1, 2, \ldots. \end{cases} \tag{1.2}$$

Here, for each $x \in H$ and each nonempty closed convex subset C of H, P_C is defined by $\|x - P_C x\| = \inf\{\|x - z\| : z \in C\}$. The mapping P_C is sometimes called the metric projection of H onto C. They proved that if $T^{-1}0 \neq \emptyset$ and $\liminf_{n\to\infty} r_n > 0$, then the sequence generated by (1.2) converges strongly to $P_{T^{-1}0}x_0$.

It is our purpose in this paper to extend Solodov and Svaiter's result to more general Banach spaces like the spaces L^p $(1 < p < \infty)$. Using this, we will then consider the problem of finding a minimizer of a convex function. The duality mapping and geometric properties of Banach spaces will play important roles in our study.

2. Preliminaries. Let E be a real Banach space with norm $\|\cdot\|$, and let E^* denote the dual of E. We denote the value of $f \in E^*$ at $x \in E$ by $\langle x, f\rangle$. When $\{x_n\}$ is a sequence in E, we denote strong convergence of $\{x_n\}$ to $x \in E$ by $x_n \to x$ and weak convergence by $x_n \rightharpoonup x$. A multivalued operator $T\colon E \to 2^{E^*}$ with domain $D(T) = \{z \in E : Tz \neq \emptyset\}$ and range $R(T) = \bigcup\{Tz : z \in D(T)\}$ is said to be monotone if $\langle x_1 - x_2, y_1 - y_2\rangle \ge 0$ for each $x_i \in D(T)$ and $y_i \in Tx_i$, $i = 1, 2$. A monotone operator T is said to be maximal if its graph $G(T) = \{(x, y) : y \in Tx\}$ is not properly contained in the graph of any other monotone operator.

A Banach space E is said to be strictly convex if $\|(x + y)/2\| < 1$ for all $x, y \in E$ with $\|x\| = \|y\| = 1$ and $x \neq y$. It is also said to be uniformly convex if $\lim_{n\to\infty} \|x_n - y_n\| = 0$ for any two sequences $\{x_n\}$, $\{y_n\}$ in E such that $\|x_n\| = \|y_n\| = 1$ and $\lim_{n\to\infty} \|(x_n + y_n)/2\| = 1$. It is known that a uniformly convex Banach space is reflexive and strictly convex. Let $U = \{x \in E : \|x\| = 1\}$. A Banach space E is said to be smooth if the limit

$$\lim_{t\to 0} \frac{\|x + ty\| - \|x\|}{t} \tag{2.1}$$

exists for all $x, y \in U$. It is also said to be uniformly smooth if the limit (2.1) is attained uniformly for $x, y \in U$. It is known that the space L^p $(1 < p < \infty)$ is a uniformly convex and uniformly smooth Banach space. The (normalized) duality mapping J from E into 2^{E^*} is defined by

$$Jx = \{v \in E^* : \langle x, v\rangle = \|x\|^2 = \|v\|^2\}$$

for $x \in E$. Notice that, in a Hilbert space, the duality mapping is the identity operator. The duality mapping J has the following properties:

1. $\|x\|^2 - \|y\|^2 \ge 2\langle x - y, j\rangle$ for all $x, y \in E$ and $j \in Jy$;
2. if E is smooth, then J is single valued;
3. if E is smooth, then J is norm-to-weak* continuous;
4. if E is uniformly smooth, then J is uniformly norm-to-norm continuous on each bounded subset of E.

Further, we know the following result, which characterizes a uniformly convex Banach space.

PROPOSITION 1 (see Xu [13]). *Let $s > 0$ and let E be a Banach space. Then E is uniformly convex if and only if there exists a continuous, strictly increasing, and convex function $g\colon [0,\infty) \to [0,\infty)$, $g(0) = 0$, such that*

$$\|x+y\|^2 \geq \|x\|^2 + 2\langle y, j\rangle + g(\|y\|)$$

for all $x, y \in \{z \in E : \|z\| \leq s\}$ and $j \in Jx$.

Next we define a real-valued function which plays a crucial role in our discussion. Let E be a smooth Banach space. The function $\phi\colon E \times E \to \mathbb{R}$ is defined by

$$\phi(x,y) = \|x\|^2 - 2\langle x, Jy\rangle + \|y\|^2$$

for $x, y \in E$. It is obvious from the definition of ϕ that

$$(\|x\| - \|y\|)^2 \leq \phi(x,y) \tag{2.2}$$

for all $x, y \in E$. Further, we can show the following two propositions.

PROPOSITION 2. *Let E be a uniformly convex and smooth Banach space and let $\{y_n\}, \{z_n\}$ be two sequences of E. If $\phi(y_n, z_n) \to 0$ and either $\{y_n\}$ or $\{z_n\}$ is bounded, then $y_n - z_n \to 0$.*

Proof. It follows from $\phi(y_n, z_n) \to 0$ that $\{\phi(y_n, z_n)\}$ is bounded. Then if one of the sequences $\{y_n\}$ and $\{z_n\}$ is bounded, so is the other because of (2.2). Therefore, by Proposition 1, there exists a continuous, strictly increasing, and convex function $g\colon [0,\infty) \to [0,\infty)$, $g(0) = 0$, such that

$$\begin{aligned} g(\|y_n - z_n\|) &\leq \|z_n + (y_n - z_n)\|^2 - \|z_n\|^2 - 2\langle y_n - z_n, Jz_n\rangle \\ &= \|y_n\|^2 - \|z_n\|^2 - 2\langle y_n, Jz_n\rangle + 2\|z_n\|^2 \\ &= \phi(y_n, z_n). \end{aligned}$$

It follows from $\phi(y_n, z_n) \to 0$ that $g(\|y_n - z_n\|) \to 0$. Then the properties of g yield that $y_n - z_n \to 0$. ∎

PROPOSITION 3. *Let E be a reflexive, strictly convex, and smooth Banach space, let C be a nonempty closed convex subset of E, and let $x \in E$. Then there exists a unique element $x_0 \in C$ such that*

$$\phi(x_0, x) = \inf\{\phi(z, x) : z \in C\}. \tag{2.3}$$

Proof. Since E is reflexive and $\|z_n\| \to \infty$ implies $\phi(z_n, x) \to \infty$, there exists $x_0 \in C$ such that $\phi(x_0, x) = \inf\{\phi(z, x) : z \in C\}$. Since E is strictly convex, $\|\cdot\|^2$ is a strictly convex function, that is, $\|\lambda x_1 + (1-\lambda)x_2\|^2 < \lambda\|x_1\|^2 + (1-\lambda)\|x_2\|^2$ for all $x_1, x_2 \in E$ with $x_1 \neq x_2$ and $\lambda \in (0,1)$. Then the function $\phi(\cdot, y)$ is also strictly convex. Therefore $x_0 \in C$ is unique. ∎

For each nonempty closed convex subset C of a reflexive, strictly convex, and smooth Banach space E and $x \in E$, we define the mapping Q_C of E onto C by $Q_C x = x_0$, where x_0 is defined by (2.3). It is easy to see that, in a Hilbert space, the mapping Q_C is coincident with the metric projection. In our discussion, instead of the metric projection, we make use of the mapping Q_C. Finally, we shall prove two results concerning Proposition 3 and the mapping Q_C. The first one is the usual analogue of a characterization of the metric projection in a Hilbert space.

PROPOSITION 4. *Let E be a smooth Banach space, let C be a convex subset of E, let $x \in E$, and let $x_0 \in C$. Then*

$$\phi(x_0, x) = \inf\{\phi(z, x) : z \in C\} \tag{2.4}$$

if and only if

$$\langle z - x_0, Jx_0 - Jx\rangle \geq 0 \quad \text{for all } z \in C. \tag{2.5}$$

Proof. First we shall show that (2.4) $\Rightarrow$ (2.5). Let $z \in C$ and let $\lambda \in (0, 1)$. It follows from $\phi(x_0, x) \leq \phi((1-\lambda)x_0 + \lambda z, x)$ that

$$\begin{aligned}
0 &\leq \|(1-\lambda)x_0 + \lambda z\|^2 - 2\langle (1-\lambda)x_0 + \lambda z, Jx\rangle + \|x\|^2 - \|x_0\|^2 + 2\langle x_0, Jx\rangle - \|x\|^2 \\
&= \|(1-\lambda)x_0 + \lambda z\|^2 - \|x_0\|^2 - 2\lambda\langle z - x_0, Jx\rangle \\
&\leq 2\lambda\langle z - x_0, J((1-\lambda)x_0 + \lambda z)\rangle - 2\lambda\langle z - x_0, Jx\rangle \\
&= 2\lambda\langle z - x_0, J((1-\lambda)x_0 + \lambda z) - Jx\rangle,
\end{aligned}$$

which implies

$$\langle z - x_0, J((1-\lambda)x_0 + \lambda z) - Jx\rangle \geq 0.$$

Tending $\lambda \downarrow 0$, since J is norm-to-weak* continuous, we obtain

$$\langle z - x_0, Jx_0 - Jx\rangle \geq 0,$$

which shows (2.5).

Next we shall show that (2.5) $\Rightarrow$ (2.4). For any $z \in C$, we have

$$\begin{aligned}
\phi(z, x) - \phi(x_0, x) &= \|z\|^2 - 2\langle z, Jx\rangle + \|x\|^2 - \|x_0\| + 2\langle x_0, Jx\rangle - \|x\|^2 \\
&= \|z\|^2 - \|x_0\|^2 - 2\langle z - x_0, Jx\rangle \\
&\geq 2\langle z - x_0, Jx_0\rangle - 2\langle z - x_0, Jx\rangle \\
&= 2\langle z - x_0, Jx_0 - Jx\rangle \\
&\geq 0,
\end{aligned}$$

which proves (2.4). ∎

PROPOSITION 5. *Let E be a reflexive, strictly convex, and smooth Banach space, let C be a nonempty closed convex subset of E, and let $x \in E$. Then*

$$\phi(y, Q_C x) + \phi(Q_C x, x) \leq \phi(y, x) \tag{2.6}$$

for all $y \in C$.

Proof. It follows from Proposition 4 that

$$\begin{aligned}
&\phi(y, x) - \phi(Q_C x, x) - \phi(y, Q_C x) \\
&\quad = \|y\|^2 - 2\langle y, Jx\rangle + \|x\|^2 - \|Q_C x\|^2 + 2\langle Q_C x, Jx\rangle - \|x\|^2 \\
&\qquad - \|y\|^2 + 2\langle y, JQ_C x\rangle - \|Q_C x\|^2 \\
&\quad = -2\langle y, Jx\rangle + 2\langle Q_C x, Jx\rangle + 2\langle y, JQ_C x\rangle - 2\|Q_C x\|^2 \\
&\quad = 2\langle y - Q_C x, JQ_C x - Jx\rangle \\
&\quad \geq 0
\end{aligned}$$

for all $y \in C$. This completes the proof. ∎

3. Main result. Throughout this section, unless otherwise stated, we assume that $T\colon E \to 2^{E^*}$ is a maximal monotone operator. In this section, we study the following algorithm in a smooth Banach space E, which is an extension of (1.2):

$$\begin{cases} x_0 \in E, \\ 0 = v_n + \dfrac{1}{r_n}(Jy_n - Jx_n), \qquad v_n \in Ty_n, \\ H_n = \{z \in E : \langle z - y_n, v_n\rangle \le 0\}, \\ W_n = \{z \in E : \langle z - x_n, Jx_0 - Jx_n\rangle \le 0\}, \\ x_{n+1} = Q_{H_n \cap W_n} x_0, \qquad n = 0, 1, 2, \ldots, \end{cases} \tag{3.1}$$

where $\{r_n\}$ is a sequence of positive real numbers.

First we investigate the condition under which the algorithm (3.1) is well defined. Rockafellar [9] proved the following theorem.

THEOREM 6. *Let E be a reflexive, strictly convex, and smooth Banach space, and let $T\colon E \to 2^{E^*}$ be a monotone operator. Then T is maximal if and only if $R(J + rT) = E^*$ for all $r > 0$.*

Using this theorem, we can show the following result.

PROPOSITION 7. *Let E be a reflexive, strictly convex, and smooth Banach space. If $T^{-1}0 \neq \emptyset$, then the sequence generated by* (3.1) *is well defined.*

Proof. It is obvious that both H_n and W_n are closed convex sets. Let $w \in T^{-1}0$. From Theorem 6, there exists $(y_0, v_0) \in E \times E^*$ such that $0 = v_0 + (Jy_0 - Jx_0)/r_0$ and $v_0 \in Ty_0$. Since T is monotone, it follows that

$$\langle y_0 - w, v_0\rangle \ge 0,$$

which implies $w \in H_0$. On the other hand, it is clear that $w \in W_0 = E$. Then $w \in H_0 \cap W_0$, and therefore $x_1 = Q_{H_0 \cap W_0} x_0$ is well defined. Suppose that $w \in H_{n-1} \cap W_{n-1}$ and x_n is well defined for some $n \ge 1$. Again by Theorem 6, we obtain $(y_n, v_n) \in E \times E^*$ such that $0 = v_n + (Jy_n - Jx_n)/r_n$ and $v_n \in Ty_n$. Then the monotonicity of T implies that $w \in H_n$. It follows from Proposition 4 that

$$\langle w - x_n, Jx_0 - Jx_n\rangle = \langle w - Q_{H_{n-1}\cap W_{n-1}} x_0, Jx_0 - JQ_{H_{n-1}\cap W_{n-1}} x_0\rangle \le 0,$$

which implies $w \in W_n$. Therefore $w \in H_n \cap W_n$, and hence $x_{n+1} = Q_{H_n \cap W_n} x_0$ is well defined. Then, by induction, the sequence generated by (3.1) is well defined for each nonnegative integer n. □

Remark 1. From the above proof, we obtain

$$T^{-1}0 \subset H_n \cap W_n$$

for each nonnegative integer n.

Now we are ready to prove the main theorem.

THEOREM 8. *Let E be a uniformly convex and uniformly smooth Banach space. If $T^{-1}0 \neq \emptyset$ and $\{r_n\} \subset (0, \infty)$ satisfies $\liminf_{n\to\infty} r_n > 0$, then the sequence $\{x_n\}$ generated by* (3.1) *converges strongly to $Q_{T^{-1}0} x_0$.*

Proof. It follows from the definition of W_n and Proposition 4 that $Q_{W_n} x_0 = x_n$. Further, from $x_{n+1} \in W_n$ and Proposition 5, we have

$$\phi(x_{n+1}, Q_{W_n} x_0) + \phi(Q_{W_n} x_0, x_0) \le \phi(x_{n+1}, x_0)$$

and hence

$$\phi(x_{n+1}, x_n) + \phi(x_n, x_0) \le \phi(x_{n+1}, x_0). \tag{3.2}$$

Therefore $\lim_{n\to\infty} \phi(x_n, x_0)$ exists and, in particular, $\{\phi(x_n, x_0)\}$ is bounded. Then, by (2.2), $\{x_n\}$ is also bounded. This implies that there exists a subsequence $\{x_{n_i}\}$ of $\{x_n\}$ such that $x_{n_i} \rightharpoonup w$ for some $w \in E$. We shall show that $w \in T^{-1}0$. It follows from (3.2) that $\phi(x_{n+1}, x_n) \to 0$. On the other hand,

$$\begin{aligned}\phi(Q_{H_n}x_n, x_n) - \phi(y_n, x_n) &= \|Q_{H_n}x_n\|^2 - \|y_n\|^2 + 2\langle y_n - Q_{H_n}x_n, Jx_n\rangle \\ &\ge 2\langle Q_{H_n}x_n - y_n, Jy_n\rangle + 2\langle y_n - Q_{H_n}x_n, Jx_n\rangle \\ &= 2\langle y_n - Q_{H_n}x_n, Jx_n - Jy_n\rangle.\end{aligned}$$

Since $Q_{H_n}x_n \in H_n$ and $v_n = (Jx_n - Jy_n)/r_n$, it follows that $\langle y_n - Q_{H_n}x_n, Jx_n - Jy_n\rangle \ge 0$ and therefore that $\phi(Q_{H_n}x_n, x_n) \ge \phi(y_n, x_n)$. Further, from $x_{n+1} \in H_n$, we have $\phi(x_{n+1}, x_n) \ge \phi(Q_{H_n}x_n, x_n)$, which yields $\phi(x_{n+1}, x_n) \ge \phi(Q_{H_n}x_n, x_n) \ge \phi(y_n, x_n)$. Then it follows from $\phi(x_{n+1}, x_n) \to 0$ that $\phi(y_n, x_n) \to 0$. Consequently, by Proposition 2, we have $x_n - y_n \to 0$, which implies $y_{n_i} \rightharpoonup w$. Moreover, since J is uniformly norm-to-norm continuous on bounded subsets and $\liminf_{n\to\infty} r_n > 0$, we obtain

$$v_n = \frac{1}{r_n}(Jx_n - Jy_n) \to 0.$$

It follows from $v_n \in Ty_n$ and the monotonicity of T that

$$\langle z - y_n, z' - v_n\rangle \ge 0$$

for all $z \in D(T)$ and $z' \in Tz$. This implies that

$$\langle z - w, z'\rangle \ge 0$$

for all $z \in D(T)$ and $z' \in Tz$. Therefore, from the maximality of T, we obtain $w \in T^{-1}0$.

Let $w^* = Q_{T^{-1}0}x_0$. From $x_{n+1} = Q_{H_n\cap W_n}x_0$ and $w^* \in T^{-1}0 \subset H_n \cap W_n$, we have $\phi(x_{n+1}, x_0) \le \phi(w^*, x_0)$. Then

$$\begin{aligned}\phi(x_n, w^*) &= \phi(x_n, x_0) + \phi(x_0, w^*) - 2\langle x_n - x_0, Jw^* - Jx_0\rangle \\ &\le \phi(w^*, x_0) + \phi(x_0, w^*) - 2\langle x_n - x_0, Jw^* - Jx_0\rangle,\end{aligned}$$

which yields

$$\limsup_{i\to\infty} \phi(x_{n_i}, w^*) \le \phi(w^*, x_0) + \phi(x_0, w^*) - 2\langle w - x_0, Jw^* - Jx_0\rangle.$$

From Proposition 4,

$$\begin{aligned}&\phi(w^*, x_0) + \phi(x_0, w^*) - 2\langle w - x_0, Jw^* - Jx_0\rangle \\ &\quad= 2(\|w^*\|^2 - \langle w^*, Jx_0\rangle - \langle w, Jw^*\rangle + \langle w, Jx_0\rangle) \\ &\quad= 2\langle w - w^*, Jx_0 - Jw^*\rangle \\ &\quad\le 0.\end{aligned}$$

Then we obtain $\limsup_{i\to\infty} \phi(x_{n_i}, w^*) \le 0$ and hence $\phi(x_{n_i}, w^*) \to 0$. It follows from Proposition 2 that $x_{n_i} \to w^*$. This means that the whole sequence $\{x_n\}$ converges weakly to w^* and that each weakly convergent subsequence of $\{x_n\}$ converges strongly to w^*. Therefore $\{x_n\}$ converges strongly to $w^* = Q_{T^{-1}0}x_0$. ∎

4. Application. Let $f\colon E \to (-\infty, \infty]$ be a proper convex lower semicontinuous function. Then the subdifferential ∂f of f is defined by

$$\partial f(z) = \{v \in E^* : f(y) \geq f(z) + \langle y - z, v\rangle, \forall y \in E\} \quad \text{for all } z \in E.$$

Using Theorem 8, we consider the problem of finding a minimizer of the function f.

THEOREM 9. *Let E be a uniformly convex and uniformly smooth Banach space, and let $f\colon E \to (-\infty, \infty]$ be a proper convex lower semicontinuous function. Assume that $\{r_n\} \subset (0, \infty)$ satisfies $\liminf_{n\to\infty} r_n > 0$, and let $\{x_n\}$ be the sequence generated by*

$$\begin{cases} x_0 \in E, \\ y_n = \operatorname*{argmin}_{z\in E} \left\{ f(z) + \dfrac{1}{2r_n}\|z\|^2 \ - \dfrac{1}{r_n}\langle z, Jx_n\rangle \right\}, \\ 0 = v_n + \dfrac{1}{r_n}(Jy_n - Jx_n), \ v_n \in \partial f(y_n), \\ H_n = \{z \in E : \langle z - y_n, v_n\rangle \leq 0\}, \\ W_n = \{z \in E : \langle z - x_n, Jx_0 - Jx_n\rangle \leq 0\}, \\ x_{n+1} = Q_{H_n \cap W_n} x_0, \quad n = 0, 1, 2, \ldots . \end{cases}$$

If $(\partial f)^{-1}0 \neq \phi$, then $\{x_n\}$ converges strongly to the minimizer of f.

Proof. Since $f\colon E \to (-\infty, \infty]$ is a proper convex lower semicontinuous function, by Rockafellar [8], the subdifferential ∂f of f is a maximal monotone operator. We also know that

$$y_n = \operatorname*{argmin}_{z\in E} \left\{ f(z) + \frac{1}{2r_n}\|z\|^2 - \frac{1}{r_n}\langle z, Jx_n\rangle \right\}$$

is equivalent to

$$0 \in \partial f(y_n) + \frac{1}{r_n} Jy_n - \frac{1}{r_n} Jx_n.$$

Thus, we have $v_n \in \partial f(y_n)$ such that $0 = v_n + (Jy_n - Jx_n)/r_n$. Using Theorem 8, we get the conclusion. □

Acknowledgment. The authors would like to express their sincere thanks to the anonymous referee for his careful reading of the manuscript and his corrections and suggestions.

REFERENCES

[1] H. BRÉZIS AND P. L. LIONS, *Produits infinis de resolvants*, Israel J. Math., 29 (1978), pp. 329–345.
[2] O. GÜLER, *On the convergence of the proximal point algorithm for convex minimization*, SIAM J. Control Optim., 29 (1991), pp. 403–419.
[3] S. KAMIMURA AND W. TAKAHASHI, *Approximating solutions of maximal monotone operators in Hilbert spaces*, J. Approx. Theory, 106 (2000), pp. 226–240.
[4] S. KAMIMURA AND W. TAKAHASHI, *Weak and strong convergence of solutions to accretive operator inclusions and applications*, Set-Valued Anal., 8 (2000), pp. 361–374.
[5] P. L. LIONS, *Une methode iterative de resolution d'une inequation variationnelle*, Israel J. Math., 31 (1978), pp. 204–208.
[6] B. MARTINET, *Regularisation d'inequations variationnelles par approximations successives*, Rev. Franc. Inform. Rech. Oper., 4 (1970), pp. 154–159.

[7] 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.
[8] R. T. Rockafellar, *Characterization of the subdifferentials of convex functions*, Pacific J. Math., 17 (1966), pp. 497–510.
[9] R. T. Rockafellar, *On the maximality of sums of nonlinear monotone operators*, Trans. Amer. Math. Soc., 149 (1970), pp. 75–88.
[10] R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.
[11] M. V. Solodov and B. F. Svaiter, *A hybrid projection–proximal point algorithm*, J. Convex Anal., 6 (1999), pp. 59–70.
[12] M. V. Solodov and B. F. Svaiter, *Forcing strong convergence of proximal point iterations in a Hilbert space*, Math. Program., 87 (2000), pp. 189–202.
[13] H. K. Xu, *Inequalities in Banach spaces with applications*, Nonlinear Anal., 16 (1991), pp. 1127–1138.

SIAM J. OPTIM.
Vol. 13, No. 4, pp. 947–959

CHARACTERIZING SET CONTAINMENTS INVOLVING INFINITE CONVEX CONSTRAINTS AND REVERSE-CONVEX CONSTRAINTS*

V. JEYAKUMAR†

Abstract. Dual characterizations of the containment of a closed convex set, defined by infinite convex constraints, in an arbitrary polyhedral set, in a reverse-convex set, defined by convex constraints, and in another convex set, defined by finite convex constraints, are given. A special case of these dual characterizations has played a key role in generating knowledge-based support vector machine classifiers which are powerful tools in data classification and mining. The conditions in these dual characterizations reduce to simple nonasymptotic conditions under Slater's constraint qualification.

Key words. set containment, infinite convex constraints, reverse-convex set, knowledge-based classifier, conjugate function

AMS subject classifications. 49A52, 90C30, 26A24

PII. S1052623402401944

1. Introduction.

Dual conditions, which characterize the containment of a closed convex set, defined by linear or convex constraints, in a closed half-space, have played an important role in optimization and mathematical programming (see [4, 8, 9, 16, 17, 21]). Such conditions, which appear in the generalizations of Farkas' lemma [9, 15, 21] and in solvability theorems [17, 18], have been used to develop Lagrange multipliers [8, 9, 10], dual optimization problems [4, 18], and minimax theories in optimization [11, 17, 24]. Recently, these dual characterizations have been employed in knowledge-based data classification [3, 22].

Data classification is one of the primary methods in data mining [1] which addresses the question of how best to use historical data to improve the process of making decisions and to discover general regularities [23]. There has been widespread interest in support vector machines (SVMs) [25], which are powerful tools for data classification [2, 23]. The principal aim in data classification is to accurately discriminate between two training sets of data by means of a linear separating plane. In the case where the training data are linearly inseparable, the SVM approach attempts the discrimination by solving a mathematical programming problem. The SVM mathematical programs are usually formulated as convex quadratic programming problems. The knowledge-based SVM formulation generates separating planes by training on data and utilizing prior knowledge (see [3]). Using a dual characterization of the containment of a polyhedral set in a closed half-space, Fung, Mangasarian, and Shavlik [3] have incorporated prior knowledge (represented by a polyhedral set) into a SVM mathematical program that can be solved efficiently.

Motivated by more general nonpolyhedral knowledge-based data classification, Mangasarian [22] has recently established elegant dual characterizations of the containment of a polyhedral set in an arbitrary polyhedral set, and of a general closed

*Received by the editors February 4, 2002; accepted for publication (in revised form) September 13, 2002; published electronically March 5, 2003. This research was partially supported by a grant from the Australian Research Council.

http://www.siam.org/journals/siopt/13-4/40194.html

†Department of Applied Mathematics, University of New South Wales, Sydney 2052, Australia (jeya@maths.unsw.edu.au).

convex set, defined by finite convex constraints, in a reverse-convex set [21], defined by convex constraints. Stimulated by the work of Mangasarian [22], we establish in this paper dual characterizations of the containment of a closed convex set, defined by infinite convex constraints, in an arbitrary polyhedral set, in a reverse-convex set, defined by convex constraints, and in another convex set, defined by finite convex constraints. The dual characterizations are given in terms of epigraphs of conjugate functions.

The outline of the paper is as follows. In section 2, we present definitions and preliminary results that will be used later in the paper. In section 3, we derive (asymptotic) dual characterizations of the set containment of a closed convex set, defined by infinite convex constraints, in an arbitrary polyhedral set. In section 4, we provide general characterizations of the set containment of a closed convex set, defined by infinite convex constraints, in a reverse-convex set, and in a not necessarily polyhedral convex set. In the appendix, we provide technical results that ensure nonasymptotic dual conditions characterizing the set containments.

2. Notation and preliminaries. In this section, we describe our notation and present preliminary results. Throughout the paper, all vectors will be column vectors. A column vector will be transposed to a raw vector by a prime $'$. The inner product of two vectors u and x in the n-dimensional real space $\mathbb{R}^n$ will be denoted by $u(x) := u'x = \langle u, x\rangle$. The null vector in $\mathbb{R}^n$ will be denoted by 0. For a set $D \subseteq \mathbb{R}^n$ we shall denote the *closure* and *convex hull* of D by $\operatorname{cl} D$ and $\operatorname{co} D$, respectively. Similarly, we shall denote the *cone* generated by the set D and the *closed convex cone* generated by the set D by $\operatorname{cone} D = \bigcup_{\alpha\geq 0} \alpha D$ and $\operatorname{cl}(\operatorname{coneco} D) = \operatorname{cl}(\bigcup_{\alpha\geq 0} \alpha \operatorname{co} D)$, respectively. For the set D, the *support function* σ_D is defined by

$$\sigma_D(u) = \sup_{x\in D} u(x)$$

and the indicator function δ_D is defined by

$$\delta_D(x) = \begin{cases} 0, & x \in D, \\ +\infty, & x \notin D. \end{cases} \tag{2.1}$$

Unless stated otherwise, we assume throughout that $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a proper convex function. Then the *conjugate function* of f (in particular, the Fenchel–Moreau conjugate), $f^* : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, is defined by

$$f^*(u) = \sup_{x\in \operatorname{dom} f} \{u(x) - f(x)\},$$

where the domain of f is given by $\operatorname{dom} f := \{x \in \mathbb{R}^n : f(x) < +\infty\}$. The *epigraph* of f, $\operatorname{epi} f$, is defined by

$$\operatorname{epi} f = \{(x, r) \in \mathbb{R}^n \times \mathbb{R} : x \in \operatorname{dom} f,\ f(x) \leq r\}.$$

Recall that, for $\varepsilon \geq 0$, the *ε-subdifferential* of f at $a \in \operatorname{dom} f$ is defined as the nonempty closed convex set

$$\partial_\varepsilon f(a) = \{v \in \mathbb{R}^n \mid f(x) - f(a) \geq v(x - a) - \varepsilon\ \forall x \in \operatorname{dom} f\}.$$

Note that

$$\bigcap_{\varepsilon>0} \partial_\varepsilon f(a) = \partial f(a).$$

If dom $f = \mathbb{R}^n$ and if f is actually *sublinear* (i.e., convex and positively homogeneous of degree one), then $\partial_\varepsilon f(0) = \partial f(0)$ for all $\varepsilon \geq 0$, where $\partial f(0)$ is the usual convex subdifferential of f at 0.

The following elementary result, given recently in [14], illustrates the connection between the ϵ-subdifferential of a convex function f and the epigraph of its conjugate f^*.

PROPOSITION 2.1. *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous convex function, and let $a \in \operatorname{dom} f$. Then*

$$\operatorname{epi} f^* = \bigcup_{\epsilon \geq 0} \{(v,\ \epsilon + v(a) - f(a)) : v \in \partial_\epsilon f(a)\}.$$

Proof. Let $(u,\ r) \in \operatorname{epi} f^*$. Then $f^*(u) \leq r$. From the definition of conjugate function, for each $x \in \operatorname{dom} f$, $f^*(u) \geq u(x) - f(x)$; thus, for each $x \in \operatorname{dom} f$, $u(x) - f(x) \leq r$. Let $\varepsilon_0 = r + f(a) - u(a) \geq 0$. So $r = \varepsilon_0 - f(a) + u(a)$. Now, for each $x \in \operatorname{dom} f$,

$$f(x) - f(a) \geq u(x) - r - f(a) = u(x - a) - \varepsilon_0;$$

thus, $u \in \partial_{\varepsilon_0} f(a)$. Hence,

$$\operatorname{epi}\ f^* \subset K := \bigcup_{\varepsilon \geq 0} \{(v,\ \varepsilon + v(a) - f(a))\ :\ v \in \partial_\varepsilon f(a)\}.$$

Conversely, let $(u,\ r) \in K$. Then there exists $\varepsilon_0 \geq 0$ such that $u \in \partial_{\varepsilon_0} f(a)$ and $r = -f(a) + u(a) + \varepsilon_0$. This gives us $f^*(u) + f(a) - u(a) \leq \varepsilon_0$, which means that $f^*(u) \leq \varepsilon_0 + u(a) - f(a)$; thus, $f^*(u) \leq r$, and so $(u,\ r) \in \operatorname{epi} f^*$. □

It is easy to see from Proposition 2.1 that if dom $f = \mathbb{R}^n$ and if f is sublinear, then $\operatorname{epi} f^* = \partial f(0) \times \mathbb{R}_+$, where $\mathbb{R}_+$ is the set of all nonnegative numbers in $\mathbb{R}$. Moreover, if $\tilde{f}\ (x) = f(x) - k$, $x \in \mathbb{R}^n$, $k \in \mathbb{R}$, then $\operatorname{epi} \tilde{f}^* = \partial f(0) \times [k,\ \infty)$.

It is also worth noting that if $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a lower semicontinuous convex function, then $\{0\} \times \mathbb{R}_+ \subset \operatorname{cl}(\operatorname{cone} \operatorname{epi} f^*)$. Indeed, $(0, 1) \in \operatorname{cl}(\operatorname{cone} \operatorname{epi} f^*)$. Otherwise, by the (Hahn–Banach) separation theorem there is an $(x, \alpha) \in \mathbb{R}^n \times \mathbb{R}$ such that

$$\alpha < 0 \ \text{ and } \ (\forall (\mathrm{u}, \gamma) \in \operatorname{cl}(\operatorname{cone} \operatorname{epi} \mathrm{f}^*)) \ \ \mathrm{u(x)} + \gamma\alpha \geq 0.$$

Now, for each $u \in \operatorname{dom} f$ and each $\epsilon > 0$, $(u, f^*(u) + \epsilon) \in \operatorname{cl}(\operatorname{cone} \operatorname{epi} f^*)$, and so $u(x) + (f^*(u) + \epsilon)\alpha \geq 0$; thus, $\frac{1}{\epsilon}(u(x) + f^*(u)) + \alpha \geq 0$. Letting $\epsilon \to \infty$, we get that $\alpha \geq 0$, which is a contradiction.

For a detailed discussion of conjugate functions and *ε-subdifferentials*, see [12, 13]. See also [8, 16, 18] for recent applications of these concepts in global optimization.

3. Containment of a convex set in a polyhedral set. In this section, we present a characterization of the containment of a closed convex set, defined by infinite convex constraints, in an arbitrary polyhedral set. We begin by deriving the following technical result, which plays a key role in characterizing set containments later in the paper.

LEMMA 3.1. *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous convex function, and let $A = \{x \in \mathbb{R}^n : f(x) \leq 0\}$. Then the following statements hold:*

(a) $A \neq \emptyset \Leftrightarrow\ (0,\ -1) \notin \operatorname{cl}(\operatorname{cone}\ \operatorname{epi}\ f^*)$.

(b) $A \neq \emptyset\ \Rightarrow\ \operatorname{epi}\ \sigma_A = \operatorname{cl}(\operatorname{cone}\ \operatorname{epi}\ f^*)$.

Proof. (a) Observe first that if A is nonempty, then it follows from the definitions of A and δ_A that $\delta_A(x) \geq f(x)$ for each $x \in \mathbb{R}^n$, and so $\delta_A^*(u) \leq f^*(u)$ for each $u \in \mathbb{R}^n$. Since δ_A^* is a lower semicontinuous sublinear function and $\delta_A^* = \sigma_A$, we obtain the inclusion that cl (cone epi f^*) $\subseteq$ epi σ_A.

Now, by the definition of epi σ_A, $(0, -1) \notin$ epi σ_A, and so by the above inclusion $(0, -1) \notin$ cl (cone epi f^*). Conversely, suppose that $(0, -1) \notin$ cl (cone epi f^*). Then by the separation theorem there is an $(x, \alpha) \in \mathbb{R}^n \times \mathbb{R}$ such that

$$-\alpha < 0 \text{ and } (\forall (u, \gamma) \in \text{cl (cone epi f}^*)) \text{ u(x)} + \gamma\alpha \geq 0.$$

Let $\bar{x} = x/\alpha$. Then for each $(u, \gamma) \in$ cl (cone epi f^*), $u(\bar{x}) + \gamma \geq 0$. Thus, for any $u \in$ dom f^*, $u(\bar{x}) + f^*(u) \geq 0$; thus, $u(-\bar{x}) - f^*(u) \leq 0$. Hence, $f(-\bar{x}) = \sup_u [u(-\bar{x}) - f^*(u)] \leq 0$, and so $A \neq \emptyset$.

(b) We have already established in part (a) that cl (cone epi f^*) $\subseteq$ epi σ_A. To see the converse inclusion, let $(u, \alpha) \notin$ cl (cone epi f^*). Since $A \neq \emptyset$, $(0, -1) \notin$ cl (cone epi f^*). Then

$$B \cap (\text{cl (cone epi } f^*)) = \emptyset,$$

where

$$B := \{\delta(u, \alpha) + (1 - \delta)(0, -1) \in \mathbb{R}^n \times \mathbb{R} \mid \delta \in [0, 1]\}$$

is the convex compact set which is the segment connecting the points (u, α) and $(0, -1)$. Otherwise, there is $\delta \in (0, 1)$ such that

$$\delta(u, \alpha) + (1 - \delta)(0, -1) \in \text{cl (cone epi } f^*);$$

thus, $(\delta u, \delta\alpha - (1 - \delta)) \in$ cl (cone epi f^*). Since $\{0\} \times \mathbb{R}_+ \subset$ cl (cone epi f^*), it follows that

$$(\delta u, \delta\alpha) = (\delta u, \delta\alpha - (1 - \delta)) + (0, 1 - \delta) \in \text{cl (cone epi } f^*),$$

which implies that

$$(u, \alpha) = \frac{1}{\delta}(\delta u, \delta\alpha) \in \text{cl (cone epi } f^*).$$

This is a contradiction.

Now by the separation theorem there is $(x, \beta) \in \mathbb{R}^n \times \mathbb{R}$, $(x, \beta) \neq (0, 0)$, such that

$$[\delta(u, \alpha) + (1 - \delta)(0, -1)](x, \beta) < 0 \ \forall \delta \in [0, 1],$$

$$v(x) + \gamma\beta \geq 0 \ \forall (v, \gamma) \in \text{cl (cone epi } f^*).$$

By letting $\delta = 0$ we get that $\beta > 0$, and by letting $\delta = 1$ we obtain $u(x) + \alpha\beta < 0$; thus, $u(-\frac{x}{\beta}) > \alpha$. Moreover, for each $v \in$ dom f^*, one has

$$v\left(-\frac{x}{\beta}\right) - f^*(v) \leq 0 \ \ \forall v \in \text{dom } f^*,$$

since $(v, f^*(v)) \in$ cl (cone epi f^*). Hence,

$$f\left(-\frac{x}{\beta}\right) = f^{**}\left(-\frac{x}{\beta}\right) = \sup_v \left[v\left(-\frac{x}{\beta}\right) - f^*(v)\right] \leq 0,$$

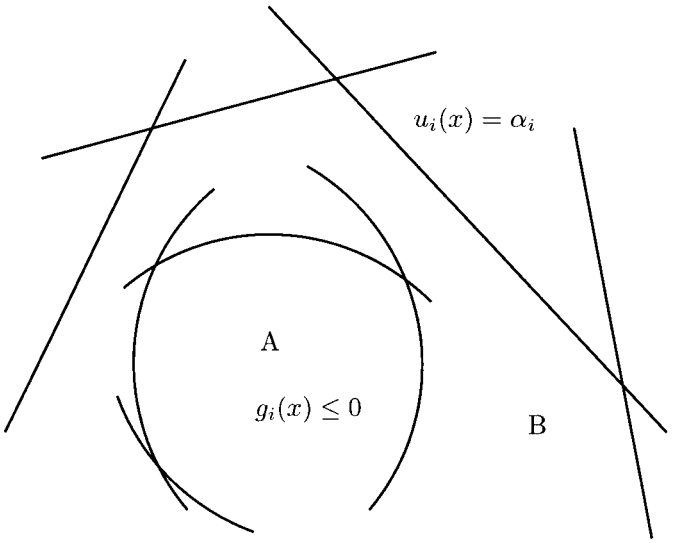


FIG. 1. *Containment of the convex set* $A = \{x \mid g_i(x) \leq 0,\ i \in I\}$ *in the polyhedral set* $B = \{x \mid u_i(x) \leq \alpha_i,\ i = 1, 2, \ldots, m\}$, *where* $g_i : \mathbb{R}^n \to \mathbb{R}$ *is a convex function and* $u_i \in \mathbb{R}^n$ *and* $\alpha_i \in \mathbb{R}$.

which means that $-\frac{x}{\beta} \in A$. This, together with the condition that $u(-\frac{x}{\beta}) > \alpha$, gives $(u, \alpha) \notin \text{epi } \sigma_A$, and so conclusion (b) follows. □

It should be noted that the conditions required in Lemma 3.1 do not, in general, guarantee that the set cone epi f^* is closed. To ensure closure we require additional regularity conditions on f. As shown in the appendix, if $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function with $\{x \in \mathbb{R}^n : f(x) < 0\}$ nonempty and $f^*(0) < +\infty$, then cone epi f^* is closed.

We now apply Lemma 3.1 to obtain a dual characterization of a set containment of a nonempty closed convex set, defined by infinite convex constraints, in an arbitrary polyhedral set, depicted in Figure 1.

THEOREM 3.2. *Let* I *be an arbitrary index set. For each* $i \in I$, *let* $g_i : \mathbb{R}^n \to \mathbb{R}$ *be a convex function, and for* $j = 1, 2, \ldots, m$, *let* $u^j \in \mathbb{R}^n$ *and* $\alpha^j \in \mathbb{R}$. *Let* $\{x \in \mathbb{R}^n : (\forall i \in I)\, g_i(x) \leq 0\}$ *be nonempty. Then the following statements are equivalent:*

(i) $\{x \in \mathbb{R}^n : (\forall i \in I)\, g_i(x) \leq 0\} \subseteq \{x \in \mathbb{R}^n : u^j(x) \leq \alpha^j,\ j = 1, 2, \ldots, m\}$.

(ii) *For* $j = 1, 2, \ldots, m$,

$$(u^j, \alpha^j) \in \text{cl} \left(\text{coneco} \bigcup_{i \in I} \text{epi } g_i^* \right).$$

Proof. Let $f = \sup_{i \in I} g_i$. Then f is clearly lower semicontinuous and convex. Let $A = \{x \in \mathbb{R}^n : f(x) \leq 0\}$. Then (i) is equivalent to the condition that for $j = 1, 2, \ldots, m$, $(u^j, \alpha^j) \in \text{epi } \sigma_A$, which is in turn equivalent to the inclusion that, for $j = 1, 2, \ldots, m$, $(u^j, \alpha^j) \in \text{cl}\,(\text{cone epi } f^*)$ by Lemma 3.1(b). The conclusion will follow if we show that

$$\text{epi } f^* = \text{cl co} \bigcup_{i \in I} \text{epi } g_i^*.$$

Now from [13, Chap. X, Thms. 2.4.4 and 1.3.5] we see that

$$f^* = (\sup_i g_i)^* = \overline{\text{co}}\,(\inf g_i^*),$$

where $\overline{\text{co}}\,(\inf g_i^*)$ is the *closed convex hull* of $\inf g_i^*$ (i.e., the largest lower semicontinuous convex function minorizing $\inf g_i^*$) and

$$\text{epi}\, f^* = \text{epi}\,(\overline{\text{co}}\,(\inf g_i^*)) = \text{cl co}\,(\text{epi}\,(\inf g_i^*)) = \text{cl co}\bigcup_{i\in I}\text{epi}\, g_i^*. \qquad \square$$

We now obtain from Theorem 3.2 a useful dual characterization of the containment of a closed convex set defined by infinite linear constraints in an arbitrary polyhedral set. Recall that the *characteristic cone* M [4, 6], generated by the set of affine functions $a_i(x) - b_i$, $i \in I$, where $a_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}$, is given by

$$M = \text{coneco}\left(\left\{\begin{pmatrix} a_i \\ b_i \end{pmatrix} : i \in I\right\} \cup \begin{pmatrix} 0 \\ 1 \end{pmatrix}\right).$$

The characteristic cone is closed if, for instance, the set $\{x \in \mathbb{R}^n : (\forall i \in I)\, a_i(x) < b_i\}$ is nonempty and the set

$$\left\{\begin{pmatrix} a_i \\ b_i \end{pmatrix} : i \in I\right\}$$

is compact (see Cor. 2.4.2 of [20]). For general conditions which ensure closure of the characteristic cone, see [6, 7].

COROLLARY 3.3. *Let I be an arbitrary index set. For each $i \in I$, let $a_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}$. For $j = 1, 2, \ldots, m$, let $u^j \in \mathbb{R}^n$ and $\alpha^j \in \mathbb{R}$. Let $\{x \in \mathbb{R}^n : (\forall i \in I)\, a_i(x) \le b_i\}$ be nonempty. Then the following statements are equivalent:*

(i) $\{x \in \mathbb{R}^n : (\forall i \in I)\, a_i(x) \le b_i\} \subseteq \{x \in \mathbb{R}^n : u^j(x) \le \alpha^j,\ j = 1, 2, \ldots, m\ \}$.

(ii) *For $j = 1, 2, \ldots, m$,*

$$\begin{pmatrix} u^j \\ \alpha^j \end{pmatrix} \in \text{cl}\left(\text{coneco}\left(\left\{\begin{pmatrix} a_i \\ b_i \end{pmatrix} : i \in I\right\} \cup \begin{pmatrix} 0 \\ 1 \end{pmatrix}\right)\right).$$

Proof. Let $g_i(x) = a_i(x) - b_i$, and let $f = \sup_{i\in I} g_i$. Then the set

$$A := \{x \in \mathbb{R}^n : f(x) \le 0\} = \{x \in \mathbb{R}^n : (\forall i \in I)\, a_i(x) \le b_i\}.$$

It is known (see, for instance, [5, Thm. 2.1, (iii)]) that

$$\text{epi}\ \sigma_A = \text{cl}(M).$$

Now Lemma 3.1(b) gives us that $\text{cl}\,(\text{cone epi}\ f^*) = \text{cl}(M)$, and so

$$\text{cl}\left(\text{coneco}\bigcup_{i\in I}\text{epi}\, g_i^*\right) = \text{cl}(M). \tag{3.1}$$

Hence, the conclusion follows from Theorem 3.2. $\square$

It is worth noting that, in the case where $g_i(x) = a_i(x) - b_i$,

$$\text{cl}\left(\text{coneco}\bigcup_{i\in I}\text{epi}\, g_i^*\right) = \text{cl}\left(\text{coneco}\left(\left\{\begin{pmatrix} a_i \\ b_i \end{pmatrix} : i \in I\right\} \cup \begin{pmatrix} 0 \\ 1 \end{pmatrix}\right)\right).$$

However, the preceding equality may not be valid without the closure on each side of the equality. To see this, let $I = \{1\}$ and define $g_1 : \mathbb{R} \to \mathbb{R}$ by $g_1(x) = a_1(x) - b_1 = x - 1$. Then $\operatorname{epi} g_1^* = \{(1, \alpha) \ : \ \alpha \geq 1\}$ and

$$\begin{pmatrix} 0 \\ 1 \end{pmatrix} \notin \operatorname{conecoepi} g_1^* = \operatorname{coneco} \bigcup_{i \in I} \operatorname{epi} g_i^*.$$

In this particular case,

$$\operatorname{coneco} \bigcup_{i \in I} \operatorname{epi} g_i^* \neq \operatorname{coneco} \left(\left\{ \begin{pmatrix} a_i \\ b_i \end{pmatrix} : i \in I \right\} \cup \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right).^1$$

Now from Corollary 3.3 we deduce Mangasarian's [22] matrix version of the characterization of the containment of a nonempty polyhedral set in an arbitrary polyhedral set.

COROLLARY 3.4. *Let $B \in \mathbb{R}^{k \times n}$, $C \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^k$, and $\alpha \in \mathbb{R}^m$. Let $\{x \in \mathbb{R}^n : Bx \leq b\}$ be nonempty. Then the following statements are equivalent:*

(i) $\{x \in \mathbb{R}^n : Bx \leq b\} \subseteq \{x \in \mathbb{R}^n : Cx \geq \alpha\}$.

(ii) *There exists a matrix $\Lambda \in \mathbb{R}^{m \times k}$ such that $C + \Lambda B = 0$, $\alpha + \Lambda b \leq 0$, $\Lambda \geq 0$.*

Proof. Let $I = \{1, 2, \ldots, k\}$, and let

$$B = \begin{bmatrix} a_1' \\ .. \\ a_k' \end{bmatrix}, \quad C = \begin{bmatrix} c_1' \\ .. \\ c_m' \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ .. \\ b_k \end{bmatrix}, \text{ and } \alpha = \begin{bmatrix} \alpha_1 \\ .. \\ \alpha_m \end{bmatrix},$$

where, for each $i \in I$ and $j = 1, 2, \ldots, m$, $a_i \in \mathbb{R}^n$, $c_j \in \mathbb{R}^n$, $b_i \in \mathbb{R}$, and $\alpha_j \in \mathbb{R}$. Then,

$$Bx = \begin{bmatrix} a_1(x) \\ .. \\ a_k(x) \end{bmatrix}, \quad Cx = \begin{bmatrix} c_1(x) \\ .. \\ c_m(x) \end{bmatrix},$$

$$\{x \in \mathbb{R}^n : Bx \leq b\} = \{x \in \mathbb{R}^n : (\forall i \in I)\, a_i(x) \leq b_i\},$$

and

$$\{x \in \mathbb{R}^n : Cx \geq \alpha\} = \{x \in \mathbb{R}^n : -c_j(x) \leq -\alpha_j, \ j = 1, 2, \ldots, m\}.$$

So (i) is equivalent to the inclusion

$$\{x \in \mathbb{R}^n : (\forall i \in I) \ \ a_i(x) \leq b_i\} \subseteq \{x \in \mathbb{R}^n : -c_j(x) \leq -\alpha_j, \ \ j = 1, 2, \ldots, m \ \}.$$

Since in this case, where I is finite, the characteristic cone

$$\operatorname{coneco} \left(\left\{ \begin{pmatrix} a_i \\ b_i \end{pmatrix} : i \in I \right\} \cup \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right)$$

is closed, it follows from Corollary 3.3 that (i) is equivalent to the condition that for $j = 1, 2, \ldots, m$,

$$\begin{pmatrix} -c_j \\ -\alpha_j \end{pmatrix} \in \operatorname{coneco} \left(\left\{ \begin{pmatrix} a_i \\ b_i \end{pmatrix} : i \in I \right\} \cup \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right).$$

[1] Thanks to an anonymous referee for providing the example.

This means that for each $j = 1, 2, \ldots, m$, there exist $\mu_j \geq 0$, $\lambda_j^1 \geq 0, \ldots, \lambda_j^{k+1} \geq 0$ such that $\sum_{r=1}^{k+1} \lambda_j^r = 1$ and

$$\begin{array}{lll} -c_j & = \mu_j \lambda_j^1 a_1 + & + \mu_j \lambda_j^k a_k + \mu_j \lambda_j^{k+1}.0, \\ -\alpha_j & = \mu_j \lambda_j^1 b_1 + & + \mu_j \lambda_j^k b_k + \mu_j \lambda_j^{k+1}.1; \end{array}$$

that is, for each $j = 1, 2, \ldots, m$,

$$\begin{array}{lll} c_j + \mu_j \lambda_j^1 a_1 + & + \mu_j \lambda_j^k a_k & = 0, \\ \alpha_j + \mu_j \lambda_j^1 b_1 + & + \mu_j \lambda_j^k b_k & = -\mu_j \lambda_j^{k+1} \leq 0. \end{array}$$

Defining $\Lambda \in \mathbb{R}^{m \times k}$ by

$$\Lambda = \begin{bmatrix} \mu_1 \lambda_1^1 & . & . & \mu_1 \lambda_1^k \\ . & . & . & . \\ . & . & . & . \\ \mu_m \lambda_m^1 & . & . & \mu_m \lambda_m^k \end{bmatrix},$$

we see that the preceding system is equivalent to the statement that there exists $\Lambda \geq 0$ such that $C + \Lambda B = 0$, $\alpha + \Lambda b \leq 0$. Hence, the statements (i) and (ii) are equivalent. □

As illustrated in Mangasarian [22], it is easy to see from linear programming duality that the statements (i) and (ii) are also equivalent to the condition that for each $j = 1, 2, \ldots, m$, the m linear programs are solvable and satisfy

$$\min_x \{(c_j(x) - \alpha_j) \ : \ Bx \leq b\} \geq 0.$$

Note that the dual conditions developed in Theorem 3.2 are in general asymptotic conditions. In the following, by application of a Slater-type regularity condition, we will obtain a nonasymptotic condition characterizing the set containment.

THEOREM 3.5. *Let $I = \{1, 2, \ldots, n\}$. For each $i \in I$, let $g_i : \mathbb{R}^n \to \mathbb{R}$ be a convex function. For $j = 1, 2, \ldots, m$, let $0 \neq u^j \in \mathbb{R}^n$ and $\alpha^j \in \mathbb{R}$. Let the set*

$$\{x \in \mathbb{R}^n \ : \ (\forall i \in I)\, g_i(x) < 0\}$$

be nonempty. Then the following statements are equivalent:

(i) $\{x \in \mathbb{R}^n : (\forall i \in I)\, g_i(x) \leq 0\} \subseteq \{x \in \mathbb{R}^n : u^j(x) \leq \alpha^j,\ j = 1, 2, \ldots, m\ \}$.

(ii) *For $j = 1, 2, \ldots, m$, $(u^j, \alpha^j) \in \operatorname{coneco} \bigcup_{i \in I} \operatorname{epi} g_i^*$.*

Proof ($(i) \Leftrightarrow (ii)$). Let $f = \sup_{i \in I} g_i$, and let $A = \{x \in \mathbb{R}^n : f(x) \leq 0\}$. Then by the finiteness of I, f is continuous and finite-valued. In addition,

$$\operatorname{epi} f^* = \operatorname{co} \bigcup_{i \in I} \operatorname{epi} g_i^*.$$

Clearly, (i) is equivalent to $(u^j, \alpha^j) \in \operatorname{epi} \sigma_A$ for $j = 1, 2, \ldots, m$. Now from Lemma 3.1 we see that $\operatorname{epi} \sigma_A = \operatorname{clcone} \operatorname{epi} f^*$, and so $(u^j, \alpha^j) \in \operatorname{clcone} \operatorname{epi} f^*$. Since $u^j \neq 0$, it follows from Proposition 6.1 in the appendix that

$$(u^j, \alpha^j) \in \operatorname{clcone} \operatorname{epi} f^* \Longleftrightarrow (u^j, \alpha^j) \in \operatorname{cone} \operatorname{epi} f^*.$$

This shows that the statement (i) is equivalent to the condition that for $j = 1, 2, \ldots, m$,

$$(u^j, \alpha^j) \in \operatorname{coneco} \bigcup_{i \in I} \operatorname{epi} g_i^*. \qquad \square$$

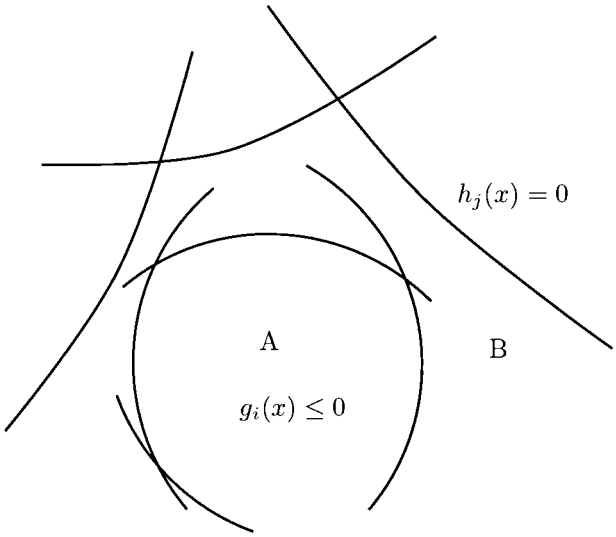


FIG. 2. *Containment of the convex set* $A = \{x \mid g_i(x) \leq 0,\ i \in I\}$ *in the reverse-convex set* $B = \{x \mid h_j(x) \geq 0,\ j = 1, 2, \ldots, m\}$, *where* $g_i,\ h_j : \mathbb{R}^n \to \mathbb{R}$ *are convex functions.*

4. Containment of a convex set in a reverse-convex set. In this section, we present dual characterizations of set containments involving not necessarily polyhedral sets. First, we examine a characterization of the set containment of a nonempty closed convex set, defined by convex constraints, in a reverse-convex set, defined by convex constraints, depicted in Figure 2 as follows.

THEOREM 4.1. *Let* I *be an arbitrary index set. For each* $i \in I$, *let* $g_i : \mathbb{R}^n \to \mathbb{R}$ *be a convex function with* $\{x \in \mathbb{R}^n : (\forall i \in I)\, g_i(x) \leq 0\}$ *nonempty. For each* $j = 1, 2, \ldots, m$, *let* $h_j : \mathbb{R}^n \to \mathbb{R}$ *be a convex function. Then the following statements are equivalent:*

(i) $\{x \in \mathbb{R}^n : (\forall i \in I)\, g_i(x) \leq 0\} \subseteq \{x \in \mathbb{R}^n : h_j(x) \geq 0,\ j = 1, 2, \ldots, m\}$.

(ii) *For* $j = 1, 2, \ldots, m$, $0 \in \text{epi}\ h_j^* + \text{cl}\,(\text{coneco} \bigcup_{i \in I} \text{epi}\ g_i^*)$.

Proof ((ii)⇒(i)). Let $f = \sup_{i \in I} g_i$, and let $A = \{x \in \mathbb{R}^n : f(x) \leq 0\}$. Then A is a closed and convex set. Since, by Lemma 3.1, epi $\sigma_A = \text{cl}\,(\text{cone}\,\text{epi}\ f^*)$, for each j, there exists $(u_j, a_j) \in \text{epi}\,\sigma_A$ such that $-(u_j, a_j) \in \text{epi}\ h_j^*$. This gives us if $x \in A$, then $u_j(x) \leq \sigma_A(u_j) \leq a_j$ and $-a_j \geq -u_j(x) - h_j(x)$. So $h_j(x) \geq 0$.

(i)⇒(ii). Let $H_j = \{x \in \mathbb{R}^n : h_j(x) \geq 0\}$. Clearly, $A \subseteq H_j$ if and only if $h_j + \delta_A \geq 0$. Then it follows from the definition of epi $(h_j + \delta_A)^*$ and the inequality, $h_j + \delta_A \geq 0$, that

$$0 \in \text{epi}\ (h_j + \delta_A)^*. \tag{4.1}$$

Since h_j is continuous and convex on $\mathbb{R}^n$ and A is nonempty, it follows that

$$(h_j + \delta_A)^* = h_j^* \overset{+}{\vee} \delta_A^*,$$

where $\overset{+}{\vee}$ denotes the inf-convolution (see Thm. 2.3.2, Chap. X of [12, 13]). Moreover, for each $x \in \text{dom}\,(h_j^* \overset{+}{\vee} \delta_A^*)$, there exist $x_1,\ x_2 \in \mathbb{R}^n$ and $x_1 + x_2 = x$ such that

$$(h_j^* \overset{+}{\vee} \delta_A^*)(x) = h_j^*(x_1) + \delta_A^*(x_2).$$

Now it is easy to show that

$$\text{epi } (h_j^* \overset{+}{\vee} \delta_A^*) = \text{epi } h_j^* + \text{epi } \delta_A^*,$$

and so

$$\text{epi } (h_j + \delta_A)^* = \text{epi } h_j^* + \text{epi } \delta_A^*.$$

As $\delta_A^* = \sigma_A$, it follows from Lemma 3.1 that

$$\text{epi } \delta_A^* = \text{cl}\left(\text{coneco } \bigcup_{i\in I} \text{epi}\, g_i^*\right). \tag{4.2}$$

Hence, for each $j = 1, 2, \ldots, m$, if $A \subseteq H_j$, then

$$0 \in \text{epi}\, h_j^* + \text{cl}\left(\text{coneco } \bigcup_{i\in I} \text{epi}\, g_i^*\right),$$

and so (ii) holds. □

If, for instance, for each j, $h_j : \mathbb{R}^n \to \mathbb{R}$ is a subaffine function of the form

$$h_j(x) = s_j(x) - \alpha_j,$$

where s_j is a sublinear function and $\alpha_j \in \mathbb{R}$, then condition (ii) of Theorem 4.1 collapses to the following:

$$\text{For } j = 1, 2, \ldots, m,\ 0 \in \partial s_j(0) \times [\alpha_j, \infty) + \text{cl}\left(\text{coneco} \bigcup_{i\in I} \text{epi}\, g_i^*\right).$$

The following theorem extends the characterization of the set containment of a convex set, defined by infinite convex constraints, in a reverse-convex set to a set involving difference of convex functions.

THEOREM 4.2. *Let I be an arbitrary index set. For each $i \in I$, let $g_i : \mathbb{R}^n \to \mathbb{R}$ be a convex function with $\{x \in \mathbb{R}^n : (\forall i \in I)\, g_i(x) \le 0\}$ nonempty. For each $j = 1, 2, \ldots, m$, let f_j, $h_j : \mathbb{R}^n \to \mathbb{R}$ be convex functions. Then the following statements are equivalent:*

(i) $\{x \in \mathbb{R}^n : (\forall i \in I)\, g_i(x) \le 0\} \subseteq \{x \in \mathbb{R}^n : f_j(x) - h_j(x) \le 0,\ \ j = 1, 2, \ldots, m\}$.

(ii) *For* $j = 1, 2, \ldots, m$, $\text{epi } f_j^* \subseteq \text{epi } h_j^* + cl\big(\text{coneco } \bigcup_{i\in I} \text{epi } g_i^*\big)$.

Proof. Since for each $x \in \mathbb{R}^n$, $f_j(x) = f_j^{**}(x)$, it follows from the definition of $\text{epi}\, f_j^*$ that (i) is equivalent to the implication that, for each $j = 1, 2, \ldots, m$, and for each $(u_j, \alpha_j) \in \text{epi}\, f_j^*$,

$$(\forall i \in I)\, g_i(x) \le 0 \Rightarrow\ h_j(x) + \alpha_j - u_j(x) \ge 0.$$

Now, for $(u_j, \alpha_j) \in \text{epi}\, f_j^*$, define the convex function w by $w(x) = h_j(x) + \alpha_j - u_j(x)$. Thus, (i) is equivalent to the inclusion that for $(u_j, \alpha_j) \in \text{epi}\, f_j^*$,

$$\{x \in \mathbb{R}^n : (\forall i \in I)\, g_i(x) \le 0\} \subseteq \{x \in \mathbb{R}^n :\ w(x) \ge 0\}.$$

Hence, Theorem 4.1 gives us that the preceding inclusion is equivalent to the following statement: For $j = 1, 2, \ldots, m$,

$$\forall (u_j, \alpha_j) \in \text{epi}\, f_j^*,\ (u_j, \alpha_j) \in \text{epi}\, h_j^* + \text{cl}\left(\text{coneco } \bigcup_{i\in I} \text{epi}\, g_i^*\right),$$

which is equivalent to (ii). □

As a special case of Theorem 4.2, we obtain a characterization for the set containment of a convex set, defined by infinite convex constraints, in another convex set, defined by finite convex constraints.

COROLLARY 4.3. *Let I be an arbitrary index set. For each $i \in I$, let $g_i : \mathbb{R}^n \to \mathbb{R}$ be a convex function with $\{x \in \mathbb{R}^n : (\forall i \in I)\, g_i(x) \leq 0\}$ nonempty. For $j = 1, 2, \ldots, m$, let $f_j : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Then the following statements are equivalent:*

(i) $\{x \in \mathbb{R}^n : (\forall i \in I)\, g_i(x) \leq 0\} \subseteq \{x \in \mathbb{R}^n : f_j(x) \leq 0,\ j = 1, 2, \ldots, m\}$.

(ii) $\left(\bigcup_{j=1}^m epi\, f_j^*\right) \subseteq \text{cl}\left(\text{coneco}\ \bigcup_{i\in I} \text{epi}\ g_i^*\right)$.

Proof. The equivalence of (i) and (ii) follows from the conclusion of the previous theorem by taking $h_j = 0$ for each $j = 1, 2, \ldots, m$. ∎

5. Conclusion and further research. In this paper, we have presented geometric dual conditions characterizing the set containment of a closed convex set, defined by infinite convex constraints, in an arbitrary polyhedral set, in a reverse-convex set, defined by convex constraints, and in another convex set, defined by finite convex constraints. Our approach, which employs epigraphs of conjugate functions, provides a unified scheme for characterizing set containment properties for polyhedral, convex, and certain nonconvex sets. Moreover, the containment of a convex set defined by infinite linear constraints in an arbitrary polyhedral set is characterized in terms of the characteristic cone generated by the affine functions involved in the constraints. The characteristic cone has played a central role in semi-infinite linear programming. These results with the application of semi-infinite linear programming duality may lead to more computationally tractable characterizations of the set containment by the solution of semi-infinite linear programs and merit further research.

On the other hand, characterizations of set containment properties were motivated by the recent work of knowledge-based SVM classifiers (see [3, 22]), which generate separating planes [19] by training on labeled data and utilizing prior knowledge. Using dual conditions which characterize the containment of a polyhedral set in a closed half-space, Fung, Mangasarian, and Shavlik [3] have incorporated prior knowledge in the form of polyhedral sets into SVM classifiers by adding the dual conditions as constraints to the SVM mathematical program. The results of this paper may possibly lead to SVM classifiers which incorporate more general knowledge sets represented by infinite system of linear (or convex) constraints. The applications of our dual characterizations to data classification will be treated elsewhere.

6. Appendix: Regularity conditions. In this section, we provide regularity conditions that ensure nonasymptotic dual conditions characterizing the set containments. The following proposition plays a useful role in establishing such regularity conditions.

PROPOSITION 6.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function, and let $A = \{x \in \mathbb{R}^n : f(x) \leq 0\}$. If the interior of A, $\text{int}(A)$, is nonempty, then*

$$\text{clcone epi}\ f^* \backslash \{0 \times \mathbb{R}\} = \text{cone epi}\ f^* \backslash \{0 \times \mathbb{R}\}.$$

Proof. The conclusion will follow if we show that $(u, \alpha) \in \text{cl cone epi}\, f^*$ with $u \neq 0$ implies $(u, \alpha) \in \text{cone epi}\, f^*$. To see this, let $(u_n, \alpha_n) \in \text{cone epi}\, f^*$ and $(u_n, \alpha_n) \to (u, \alpha)$. Thus there are $\lambda_n \geq 0$ and (h_n, μ_n) in epi f^* such that $u_n = \lambda_n h_n$ and $\alpha_n = \lambda_n \mu_n$. We now consider the following cases.

Case 1. $0 < \inf \lambda_n \leq \sup \lambda_n < +\infty$. Without loss of generality, we can assume $\lambda_n \to \lambda$ with $0 < \lambda < +\infty$. Clearly, there are $h \in \mathbb{R}^n$ and $\mu \in \mathbb{R}$ such that

$$h_n \to h \ \text{ and } \ \mu_n \to \mu.$$

Since f^* is lower semicontinuous, it follows that $(h, \mu) \in \operatorname{epi} f^*$ and $(u, \alpha) = \lambda(h, \mu) \in \operatorname{cone} \operatorname{epi} f^*$.

Case 2. $\sup \lambda_n = +\infty$. Without loss of generality, we can assume that $\lambda_n \to +\infty$. In this case, $h_n \to 0$ and by the lower semicontinuous of f^* we have

$$\liminf \mu_n \geq \liminf f^*(h_n) \geq f^*(0).$$

The existence of $x_0 \in \mathbb{R}^n$ with $f(x_0) < 0$ (i.e., $x_0 \in \operatorname{int}(A)$) ensures that $f^*(0) > 0$. So $\liminf \mu_n > 0$. Hence we have $\alpha_n = \lambda_n \mu_n \to +\infty$. However, $\alpha_n \to \alpha < +\infty$, and we have a contradiction. Thus $\sup \lambda_n = +\infty$ is impossible.

Case 3. $\lambda_n \to 0$. Since $u \neq 0$ and $\lambda_n h_n \to u$, it follows that $||h_n|| \to +\infty$. Now, by the finiteness of f, we have that f^* is 1-coercive (see [12, 13]); hence

$$\alpha_n = \lambda_n \mu_n = \mu_n \frac{\|u_n\|}{\|h_n\|} \geq \frac{\|u_n\| f^*(h_n)}{\|h_n\|} \to +\infty.$$

However, $\alpha_n \to \alpha < +\infty$, and once again we have a contradiction. Hence $\lambda_n \not\to 0$. □

PROPOSITION 6.2. *If $f : \mathbb{R}^n \to \mathbb{R}$ is convex with $\{x \in \mathbb{R}^n : f(x) < 0\}$ nonempty and if $f^*(0) < +\infty$, then* $\operatorname{cone} \operatorname{epi} f^*$ *is closed.*

Proof. Let $(u_n, \alpha_n) \in \operatorname{cone} \operatorname{epi} f^*$ and $(u_n, \alpha_n) \to (u, \alpha)$. If $u \neq 0$, then it follows that $(u, \alpha) \in \operatorname{cone} \operatorname{epi} f^*$ by Proposition 6.1. Hence consider the case where $u = 0$. In this case, if $\alpha = 0$, then $(u, \alpha) = (0, 0) \in \operatorname{cone} \operatorname{epi} f^*$, so we first consider the case in which $\alpha > 0$. Since $f^*(0)$ is finite, $(0, f^*(0)) \in \operatorname{epi} f^*$. Moreover, since $f^*(0)$ is positive, $(0, \alpha) \in \operatorname{cone} \operatorname{epi} f^*$ for any $\alpha > 0$. Thus $(u, \alpha) \in \operatorname{cone} \operatorname{epi} f^*$. Now consider the case in which $\alpha < 0$. Since $(0, \alpha) \in \operatorname{cl} \operatorname{cone} \operatorname{epi} f^*$ and this is a cone, $(0, -1) \in \operatorname{cl} \operatorname{cone} \operatorname{epi} f^*$ in this case. However, by Lemma 3.1, this condition is equivalent to the inconsistency of $f(x) \leq 0$. This is a contradiction so that $\alpha < 0$ is not possible. Hence $\operatorname{cone} \operatorname{epi} f^*$ is closed. □

Note that the existence of $x_0 \in \mathbb{R}^n$ with $f(x_0) < 0$ implies that $0 < f^*(0)$. Further, the assumption that $f^*(0) < +\infty$ (i.e., $0 \in \operatorname{dom} f^*$) is equivalent to the assumption that f is bounded below on $\mathbb{R}^n$. Thus if $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function with $\{x \in \mathbb{R}^n : f(x) < 0\}$ nonempty and f is bounded below on $\mathbb{R}^n$, then $\operatorname{cone} \operatorname{epi} f^*$ is closed. For related results on regularity conditions, see [8, 10, 12, 13].

Acknowledgment. The author is grateful to the anonymous referees for their constructive comments and valuable suggestions which have contributed to the final preparation of this paper.

REFERENCES

[1] P. S. BRADLEY, U. M. FAYYAD, AND O. L. MANGASARIAN, *Data mining: Overview and optimization opportunities*, INFORMS J. Comput., 11 (1999), pp. 217–238.

[2] C. J. C. BURGES, *A tutorial on support vector machines for pattern recognition*, Data Mining and Knowledge Discovery, 2 (1998), pp. 121–167.

[3] G. FUNG, O. L. MANGASARIAN, AND J. SHAVLIK, *Knowledge-Based Support Vector Machine Classifiers*, Technical Report 01-09, Data Mining Institute, Computer Sciences Department, University of Wisconsin, Madison, WI, 2001, ftp: //ftp.cs.wisc.edu/pub/dmi/tech-reports/01-09.ps; also available online from http://www.cs.wisc.edu/math-prog/tech-reports/.

[4] M. A. GOBERNA AND M. A. LOPEZ, *Linear Semi-Infinite Optimization*, Wiley Ser. Math. Methods Pract. 2, John Wiley and Sons, Chichester, UK, 1998.

[5] M. A. GOBERNA AND M. A. LOPEZ, *Optimal value function in semi-infinite programming*, J. Optim. Theory Appl., 59 (1988), pp. 261–279.

[6] M. A. Goberna and M. A. Lopez, *Conditions for the closedness of the characteristic cone associated to an infinite linear system*, in Infinite Programming, E. J. Anderson and A. Philpott, eds., Springer-Verlag, Berlin, 1985, pp. 16–28.
[7] M. A. Goberna, M. A. Lopez, and J. Pastor, *Farkas-Minkowski systems in semi-infinite programming*, Appl. Math. Optim., 7 (1981), pp. 295–308.
[8] B. M. Glover, Y. Ishizuka, V. Jeyakumar, and H. D. Tuan, *Complete characterizations of global optimality for problems involving the pointwise minimum of sublinear functions*, SIAM J. Optim., 6 (1996), pp. 362–372.
[9] B. M. Glover, V. Jeyakumar, and W. Oettli, *A Farkas lemma for difference sublinear systems and quasidifferentiable programming*, Math. Program., 63 (1994), pp. 109–125.
[10] B. M. Glover, V. Jeyakumar, and A. M. Rubinov, *Dual conditions characterizing optimality for convex multi-objective programs*, Math. Program., 84 (1999), pp. 201–217.
[11] C.-W. Ha, *On systems of convex inequalities*, J. Math. Anal. Appl., 68 (1979), pp. 25–34.
[12] J.-B. Hiriart-Urruty and C. Lemarechal, *Convex Analysis and Minimization Algorithms*, Vol. I, Springer-Verlag, Berlin, 1993.
[13] J.-B. Hiriart-Urruty and C. Lemarechal, *Convex Analysis and Minimization Algorithms*, Vol. II, Springer-Verlag, Berlin, 1993.
[14] V. Jeyakumar, *Asymptotic dual conditions characterizing optimality for convex programs*, J. Optim. Theory Appl., 93 (1997), pp. 153–165.
[15] V. Jeyakumar, *Farkas' Lemma: Generalizations*, in Encyclopedia of Optimization, Vol. 2, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2001, pp. 87–91.
[16] V. Jeyakumar and B. M. Glover, *Nonlinear extensions of Farkas' lemma with applications to global optimization and least squares*, Math. Oper. Res., 20 (1995), pp. 818–837.
[17] V. Jeyakumar and J. Gwinner, *Inequality systems and optimization*, J. Math. Anal. Appl., 159 (1991), pp. 51–71.
[18] 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.
[19] V. Jeyakumar and B. Waterhouse, *Data Classification via Separable Convex Programming*, Appl. Math. Research Report AMR00/13, University of New South Wales, Sydney, Australia, 2000.
[20] M. A. Lopez and E. Vercher, *Optimality conditions for nondifferentiable convex semi-infinite programming*, Math. Programming, 27 (1983), pp. 307–319.
[21] O. L. Mangasarian, *Nonlinear Programming*, Classics Appl. Math. 10, SIAM, Philadelphia, 1994.
[22] O. L. Mangasarian, *Set containment characterization*, J. Global Optim., 24 (2002), pp. 473–480.
[23] O. L. Mangasarian, *Mathematical programming in data mining*, Data Mining and Knowledge Discovery, 1 (1997), pp. 183–201.
[24] N. Shioj and W. Takahashi, *Fan's theorem concerning systems of convex inequalities and its applications*, J. Math. Anal. Appl., 135 (1988), pp. 383–393.
[25] V. N. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed., Springer, New York, 2000.

SIAM J. OPTIM.
Vol. 13, No. 4, pp. 960–985

ANALYSIS OF NONSMOOTH SYMMETRIC-MATRIX-VALUED FUNCTIONS WITH APPLICATIONS TO SEMIDEFINITE COMPLEMENTARITY PROBLEMS*

XIN CHEN†, HOUDUO QI‡, AND PAUL TSENG§

Abstract. For any function f from $\mathbb{R}$ to $\mathbb{R}$, one can define a corresponding function on the space of $n \times n$ (block-diagonal) real symmetric matrices by applying f to the eigenvalues of the spectral decomposition. We show that this matrix-valued function inherits from f the properties of continuity, (local) Lipschitz continuity, directional differentiability, Fréchet differentiability, continuous differentiability, as well as (ρ-order) semismoothness. Our analysis uses results from nonsmooth analysis as well as perturbation theory for the spectral decomposition of symmetric matrices. We also apply our results to the semidefinite complementarity problem, addressing some basic issues in the analysis of smoothing/semismooth Newton methods for solving this problem.

Key words. symmetric-matrix-valued function, nonsmooth analysis, semismooth function, semidefinite complementarity problem

AMS subject classifications. 49M45, 90C25, 90C33

PII. S1052623400380584

1. Introduction.

Let $\mathcal{X}$ denote the space of $n \times n$ block-diagonal real matrices with m blocks of size $n_1, \ldots, n_m$, respectively (the blocks are fixed). Thus, $\mathcal{X}$ is closed under matrix addition $x + y$, multiplication xy, transposition x^T, and inversion x^{-1}, where $x, y \in \mathcal{X}$. We endow $\mathcal{X}$ with the inner product and norm

$$\langle x, y \rangle := \mathrm{tr}[x^T y], \qquad \|x\| := \sqrt{\langle x, x \rangle},$$

where $x, y \in \mathcal{X}$ and $\mathrm{tr}[\cdot]$ denotes the matrix trace, i.e., $\mathrm{tr}[x] = \sum_{i=1}^n x_{ii}$. [$\|x\|$ is the Frobenius norm of x and " := " means "define"]. Let $\mathcal{O}$ denote the set of $p \in \mathcal{X}$ that are orthogonal, i.e., $p^T = p^{-1}$. Let $\mathcal{S}$ denote the subspace comprising those $x \in \mathcal{X}$ that are symmetric, i.e., $x^T = x$. This is a subspace of $\mathbb{R}^{n \times n}$ of dimension $n_1(n_1+1)/2 + \cdots + n_m(n_m+1)/2$.

For any $x \in \mathcal{S}$, its (repeated) eigenvalues $\lambda_1, \ldots, \lambda_n$ are real and it admits a spectral decomposition of the form

$$x = p \, \mathrm{diag}[\lambda_1, \ldots, \lambda_n] p^T \tag{1}$$

for some $p \in \mathcal{O}$, where $\mathrm{diag}[\lambda_1, \ldots, \lambda_n]$ denotes the $n \times n$ diagonal matrix with its ith diagonal entry λ_i. Then, for any function $f : \mathbb{R} \to \mathbb{R}$, we can define a corresponding function $f^{\square} : \mathcal{S} \to \mathcal{S}$ [1], [13] by

$$f^{\square}(x) := p \, \mathrm{diag}[f(\lambda_1), \ldots, f(\lambda_n)] p^T. \tag{2}$$

*Received by the editors November 7, 2000; accepted for publication (in revised form) July 12, 2002; published electronically March 5, 2003.
http://www.siam.org/journals/siopt/13-4/38058.html

†Operations Research Center, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Building E40-194, Cambridge, MA 02139 (xchen@mit.edu).

‡School of Mathematics, The University of New South Wales, Sydney, New South Wales 2052, Australia (hdqi@maths.unsw.edu.au). This author was supported by the Australian Research Council.

§Department of Mathematics, University of Washington, Seattle, WA 98195 (tseng@math.washington.edu). This author was supported by the National Science Foundation, grant CCR-9731273.

It is known that $f^{\square}(x)$ is well defined (independent of the ordering of $\lambda_1, \ldots, \lambda_n$ and the choice of p) and belongs to $\mathcal{S}$; see [1, Chap. V] and [13, sec. 6.2]. Moreover, a result of Daleckii and Krein showed that if f is continuously differentiable, then $f^{\square}$ is differentiable (in the Fréchet sense) and its Jacobian $\nabla f^{\square}(x)$ has a simple formula—see [1, Thm. V.3.3]; also see Proposition 4.3. In fact, in this case $f^{\square}$ is continuously differentiable—see [8, Lem. 4]; also see Proposition 4.4. Much of the studies on $f^{\square}$ has focused on conditions for it to be *operator monotone* or *operator convex*—see [1], [13], and the references cited in [1, pp. 150–151] for discussions. We note that [8] swaps p and p^T in (1)–(2), but this is only a difference in notation.

The above results show that $f^{\square}$ inherits smoothness properties from f. In this paper, we make an analogous study for properties associated with nonsmooth functions. In particular, we show that the properties of continuity, strict continuity, Lipschitz continuity, directional differentiability, differentiability, continuous differentiability, and (ρ-order) semismoothness are each inherited by $f^{\square}$ from f (see Propositions 4.1, 4.2, 4.3, 4.4, 4.6, 4.8, and 4.10). Our ρ-order semismoothness result generalizes a recent result of Sun and Sun [29] which considers the case of the absolute-value function $f(\xi) = |\xi|$ and shows that $f^{\square}(x) = (x^2)^{1/2}$ is strongly semismooth. In the case where $f = g'$ for some function g, our differentiability and continuous differentiability results can also be inferred from a recent work of Lewis and Sendov [19] on twice differentiability of spectral functions. Our proofs use a combination of results from matrix analysis and nonsmooth analysis—in particular, perturbation results for spectral decomposition [17, 28] and properties of the generalized gradient ∂f (in the Clarke sense) [9, 26], as well as a lemma from [29]. The property of semismoothness, as introduced by Mifflin [20] for functionals and scalar-valued functions and further extended by Qi and Sun [23] for vector-valued functions, is of particular interest due to the key role it plays in the superlinear convergence analysis of certain generalized Newton methods [14, 21, 23]. In section 5, we formulate the semidefinite complementarity problem (SDCP) as a nonsmooth equation

$$H(x, y) = 0,$$

where $H : \mathcal{S} \times \mathcal{S} \to \mathcal{S} \times \mathcal{S}$ is a certain semismooth function. This facilitates the development of nonsmooth Newton methods for solving the SDCP—a contrast to existing smoothing or differentiable merit function approaches [8, 27, 30, 32]. We show that H, together with the Chen–Mangasarian class of smoothing functions studied in [8], satisfies the Jacobian Consistence Property introduced in [6]. This paves a way for extending some smoothing methods for nonlinear complementarity problems (NCPs), such as those studied by Chen, Qi, and Sun [6] and later by Kanzow and Pieper [16], to the SDCP. Final remarks are given in section 6.

Our notations are, for the most part, consistent with those used in [8, 30]. If $F : \mathcal{S} \to \mathcal{S}$ is differentiable (in the Fréchet sense) at $x \in \mathcal{S}$, we denote by $\nabla F(x)$ the Jacobian of F at $x \in \mathcal{S}$, viewed as a linear mapping from $\mathcal{S}$ to $\mathcal{S}$. Throughout, $\|\cdot\|$ denotes the Frobenius norm for matrices and the 2-norm for vectors. For any linear mapping $M : \mathcal{S} \to \mathcal{S}$, we denote its operator norm $|||M||| := \max_{\|x\|=1} \|Mx\|$. For any $x \in \mathcal{S}$, we denote by x_{ij} the (i,j)th entry of x. We use $\circ$ to denote the Hardamard product, i.e.,

$$x \circ y = [x_{ij} y_{ij}]_{i,j=1}^{n}.$$

For any $x \in \mathcal{S}$ and scalar $\gamma > 0$, we denote the γ-ball around x by $\mathcal{B}(x, \gamma) := \{y \in \mathcal{S} \mid \|y - x\| \le \gamma\}$. We write $z = O(\alpha)$ (respectively, $z = o(\alpha)$), with $\alpha \in \mathbb{R}$ and $z \in \mathcal{S}$, to mean $\|z\|/|\alpha|$ is uniformly bounded (respectively, tends to zero) as $\alpha \to 0$.

2. Basic properties. In this section, we review some basic properties of vector-valued functions. These properties are continuity, (local) Lipschitz continuity, directional differentiability, continuous differentiability, as well as (ρ-order) semismoothness. We note that $\mathcal{S}$ is a vector space of dimension $n_1(n_1+1)/2+\cdots+n_m(n_m+1)/2$, so these properties apply to the symmetric-matrix-valued function $f^{\square}$ defined by (1)–(2). In what follows, we consider a function/mapping $F:\mathbb{R}^k \to \mathbb{R}^\ell$.

We say F is continuous at $x \in \mathbb{R}^k$ if

$$F(y) \to F(x) \quad \text{as} \quad y \to x;$$

and F is continuous if F is continuous at every $x \in \mathbb{R}^k$. F is strictly continuous (also called "locally Lipschitz continuous") at $x \in \mathbb{R}^k$ [26, Chap. 9] if there exist scalars $\kappa > 0$ and $\delta > 0$ such that

$$\|F(y) - F(z)\| \le \kappa \|y - z\| \quad \forall y, z \in \mathbb{R}^k \text{ with } \|y - x\| \le \delta,\ \|z - x\| \le \delta;$$

and F is strictly continuous if F is strictly continuous at every $x \in \mathbb{R}^k$. If δ can be taken to be ∞, then F is Lipschitz continuous with Lipschitz constant κ. Define the function $\mathrm{lip}F:\mathbb{R}^k \to [0,\infty]$ by

$$\mathrm{lip}F(x) := \limsup_{\substack{y,z \to x \\ y \ne z}} \frac{\|F(y) - F(z)\|}{\|y - z\|}.$$

Then F is strictly continuous at x if and only if $\mathrm{lip}F(x)$ is finite.

We say F is directionally differentiable at $x \in \mathbb{R}^k$ if

$$F'(x;h) \ := \ \lim_{t \to 0^+} \frac{F(x+th) - F(x)}{t} \quad \text{exists} \qquad \forall h \in \mathbb{R}^k;$$

and F is directionally differentiable if F is directionally differentiable at every $x \in \mathbb{R}^k$. F is differentiable (in the Fréchet sense) at $x \in \mathbb{R}^k$ if there exists a linear mapping $\nabla F(x):\mathbb{R}^k \to \mathbb{R}^\ell$ such that

$$F(x+h) - F(x) - \nabla F(x)h \ = \ o(\|h\|).$$

We say that F is continuously differentiable if F is differentiable at every $x \in \mathbb{R}^k$ and ∇F is continuous.

If F is strictly continuous, then F is almost everywhere differentiable by Rademacher's theorem—see [9] and [26, sec. 9J]. Then the generalized Jacobian $\partial F(x)$ of F at x (in the Clarke sense) can be defined as the convex hull of the generalized Jacobian $\partial_B F(x)$ (in the Bouligand sense), where

$$\partial_B F(x) := \left\{ \lim_{x^j \to x} \nabla F(x^j) \Big| F \text{ is differentiable at } x^j \in \mathbb{R}^k \right\}.$$

In [26, Chap. 9], the case of $\ell = 1$ is considered and the notations "$\bar{\nabla}$" and "$\bar{\partial}$" are used instead of, respectively, "∂_B" and "∂."

Assume $F:\mathbb{R}^k \to \mathbb{R}^\ell$ is strictly continuous. We say F is semismooth at x if F is directionally differentiable at x and, for any $V \in \partial F(x+h)$, we have

$$F(x+h) - F(x) - Vh = o(\|h\|).$$

We say F is ρ-order semismooth at x ($0 < \rho < \infty$) if F is semismooth at x and, for any $V \in \partial F(x+h)$, we have

$$F(x+h) - F(x) - Vh = O(\|h\|^{1+\rho}).$$

We say F is semismooth (respectively, ρ-order semismooth) if F is semismooth (respectively, ρ-order semismooth) at every $x \in \mathbb{R}^k$. We say F is strongly semismooth if it is 1-order semismooth. Convex functions and piecewise continuously differentiable functions are examples of semismooth functions. The composition of two (respectively, ρ-order) semismooth functions is also a (respectively, ρ-order) semismooth function. The property of semismoothness plays an important role in nonsmooth Newton methods [23] as well as in some smoothing methods mentioned in the previous section. For extensive discussions of semismooth functions, see [10, 20, 23].

3. Perturbation results for symmetric matrices. In this section, we review some useful perturbation results for the spectral decomposition of real symmetric matrices. These results will be used in the next section to analyze properties of the symmetric-matrix-valued function $f^{\square}$ given by (1)–(2). The main sources of reference for the results are Chapter 2 of the book by Kato [17] and the book by Stewart and Sun [28].

Let $\mathcal{D}$ denote the space of $n \times n$ real diagonal matrices with nonincreasing diagonal entries. For each $x \in \mathcal{S}$, define the two sets of orthonormal eigenvectors of x by

$$\mathcal{O}_x := \{p \in \mathcal{O} |\ p^T x p \in \mathcal{D}\}, \qquad \tilde{\mathcal{O}}_x := \{p \in \mathcal{O} |\ p^T x p \text{ is diagonal }\}.$$

Clearly, $\mathcal{O}_x$ and $\tilde{\mathcal{O}}_x$ are nonempty for each $x \in \mathcal{S}$. The following key lemma, proved in [8, Lem. 3] using results from [28, pp. 92 and 250], shows that $\mathcal{O}_x$ is locally upper Lipschitzian with respect to x.

LEMMA 3.1. *For any $x \in \mathcal{S}$, there exist scalars $\eta > 0$ and $\epsilon > 0$ such that*

$$\min_{p \in \mathcal{O}_x} \|p - q\| \le \eta \|x - y\| \quad \forall\ y \in \mathcal{B}(x, \epsilon),\ \forall q \in \mathcal{O}_y. \tag{3}$$

We will also need the following perturbation result of Weyl for eigenvalues of symmetric matrices—see [1, p. 63] and [12, p. 367].

LEMMA 3.2. *Let $\lambda_1 \ge \cdots \ge \lambda_n$ be the eigenvalues of any $x \in \mathcal{S}$ and $\mu_1 \ge \cdots \ge \mu_n$ be the eigenvalues of any $y \in \mathcal{S}$. Then*

$$|\lambda_i - \mu_i| \le \|x - y\| \quad \forall\ i = 1, \ldots, n.$$

Lastly, for our differential analysis, we need the following classical result [25, Thm. 1] showing that, for any $x \in \mathcal{S}$ and any $h \in \mathcal{S}$, the orthonormal eigenvectors of $x + th$ may be chosen to be analytic in t. As is remarked in [17, p. 122], the existence of such orthonormal eigenvectors depending smoothly on t is one of the most remarkable results in the analytic perturbation theory for symmetric operators.

LEMMA 3.3. *For any $x \in \mathcal{S}$ and any $h \in \mathcal{S}$, there exist $p(t) \in \tilde{\mathcal{O}}_{x+th}$, $t \in \mathbb{R}$, whose entries are power series in t, convergent in a neighborhood of $t = 0$.*

4. Continuity and differential properties of symmetric-matrix functions. In this section, we use the results from section 3 to show that if $f : \mathbb{R} \to \mathbb{R}$ has the property of continuity (respectively, strict continuity, Lipschitz continuity, directional differentiability, semismoothness, ρ-order semismoothness), then so does the symmetric-matrix-valued function $f^{\square}$ defined by (1)–(2). We begin with the continuity result below.

PROPOSITION 4.1. *For any $f : \mathbb{R} \to \mathbb{R}$, the following results hold:*

(a) *$f^{\square}$ is continuous at an $x \in \mathcal{S}$ with eigenvalues $\lambda_1, \ldots, \lambda_n$ if and only if f is continuous at $\lambda_1, \ldots, \lambda_n$.*

(b) *$f^{\square}$ is continuous if and only if f is continuous.*

Proof. (a) Fix any $x \in \mathcal{S}$ with eigenvalues $\lambda_1, \ldots, \lambda_n$. Assume without loss of generality that $\lambda_1 \geq \cdots \geq \lambda_n$.

Suppose f is continuous at $\lambda_1, \ldots, \lambda_n$. By Lemma 3.1, there exist $\eta > 0$ and $\epsilon > 0$ such that (3) holds. Then, for any $y \in \mathcal{B}(x, \epsilon)$ and any $q \in \mathcal{O}_y$, there exists $p \in \mathcal{O}_x$ satisfying

$$\|p - q\| \leq \eta \|x - y\|.$$

Moreover,

$$q^T y q = \mathrm{diag}[\mu_1, \ldots, \mu_n], \qquad p^T x p = \mathrm{diag}[\lambda_1, \ldots, \lambda_n],$$

where $\mu_1 \geq \cdots \geq \mu_n$ and $\lambda_1 \geq \cdots \geq \lambda_n$ are the eigenvalues of y and x, respectively. Since f is continuous and, by Lemma 3.2, $|\lambda_i - \mu_i| \leq \|x - y\|$ for all i, we have $f(\mu_i) \to f(\lambda_i)$ and $\|p - q\| \to 0$ as $y \to x$. Then (2) yields

$$\begin{aligned} f^{\square}(x) - f^{\square}(y) &= p\ \mathrm{diag}[f(\lambda_1), \ldots, f(\lambda_n)]p^T - q\ \mathrm{diag}[f(\mu_1), \ldots, f(\mu_n)]q^T \\ &= p\ \mathrm{diag}[f(\lambda_1) - f(\mu_1), \ldots, f(\lambda_n) - f(\mu_n)]p^T \\ &\quad + (p - q)\mathrm{diag}[f(\mu_1), \ldots, f(\mu_n)]p^T + q\ \mathrm{diag}[f(\mu_1), \ldots, f(\mu_n)](p - q)^T \\ &\to 0 \qquad \text{as} \quad y \to x. \end{aligned}$$

Thus $f^{\square}$ is continuous at x.

Suppose instead $f^{\square}$ is continuous at x. Fix any $p \in \mathcal{O}_x$. Then for each $i \in \{1, \ldots, n\}$, $p\,\mathrm{diag}[\lambda_1, \ldots, \mu_i, \ldots, \lambda_n]p^T \to x$ as $\mu_i \to \lambda_i$ so that $f^{\square}(p\,\mathrm{diag}[\lambda_1, \ldots, \mu_i, \ldots, \lambda_n]p^T) \to f^{\square}(x)$ or, equivalently, $f(\mu_i) \to f(\lambda_i)$. Thus f is continuous at λ_i for $i = 1, \ldots, n$.

(b) is an immediate consequence of (a). □

For any $\lambda = (\lambda_1, \ldots, \lambda_n)^T \in \mathbb{R}^n$, any $h \in \mathcal{S}$, and any function $f : \mathbb{R} \to \mathbb{R}$ that is directionally differentiable at $\lambda_1, \ldots, \lambda_n$, we denote by $f^{[1]}(\lambda; h)$ the $n \times n$ symmetric matrix whose (i, j)th entry is

$$f^{[1]}(\lambda; h)_{ij} := \begin{cases} \dfrac{f(\lambda_i) - f(\lambda_j)}{\lambda_i - \lambda_j} h_{ij} & \text{if } \lambda_i \neq \lambda_j, \\ f'(\lambda_i; h_{ij}) & \text{if } \lambda_i = \lambda_j. \end{cases} \tag{4}$$

By using Lemma 3.3, we have the directional differentiability result below.

PROPOSITION 4.2. *For any $f : \mathbb{R} \to \mathbb{R}$, the following results hold:*

(a) *$f^{\square}$ is directionally differentiable at an $x \in \mathcal{S}$ with eigenvalues $\lambda_1, \ldots, \lambda_n$ if and only if f is directionally differentiable at $\lambda_1, \ldots, \lambda_n$. Moreover, for any nonzero $h \in \mathcal{S}$,*

$$(f^{\square})'(x; h) = p\ f^{[1]}(\lambda; p^T h p)\ p^T \tag{5}$$

for some $p \in \mathcal{O}$ such that $(p^T h p)_{ij} = 0$ whenever $\lambda_i = \lambda_j$ and $i \neq j$.

(b) *$f^{\square}$ is directionally differentiable if and only if f is directionally differentiable.*

Proof. (a) Fix any $x \in \mathcal{S}$. By Lemma 3.3, for any nonzero $h \in \mathcal{S}$ there exist $p(t) \in \tilde{\mathcal{O}}_{x(t)}$, $t \in \mathbb{R}$, whose entries are power series in t, convergent in a neighborhood $\mathcal{I}$ of $t = 0$, where $x(t) := x + th$. Then the corresponding eigenvalues

$$\lambda_i(t) := [p(t)^T x(t) p(t)]_{ii}, \quad i = 1, \ldots, n,$$

are also power series in t, convergent for $t \in \mathcal{I}$, and satisfy

$$x(t) = p(t)\mathrm{diag}[\lambda_1(t), \ldots, \lambda_n(t)]p(t)^T. \tag{6}$$

Multiplying both sides of (6) by $p(t)^T$ from the left and then differentiating both sides with respect to t using the product rule, we obtain

$$p'(t)^T x(t) + p(t)^T x'(t) = \Lambda'(t)p(t)^T + \Lambda(t)p'(t)^T,$$

where $\Lambda(t) := \mathrm{diag}[\lambda_1(t), \ldots, \lambda_n(t)]$ and $\Lambda'(t) := \mathrm{diag}[\lambda_1'(t), \ldots, \lambda_n'(t)]$. Multiplying both sides on the right by $p(t)$ and using $x'(t) = h$, we arrive at

$$\Lambda'(t) - \hat{h}(t) = \hat{p}(t)\Lambda(t) - \Lambda(t)\hat{p}(t),$$

where $\hat{h}(t) := p(t)^T h p(t)$ and $\hat{p}(t) := p'(t)^T p(t)$. This implies

$$\hat{h}(t)_{ii} = \lambda_i'(t), \qquad i = 1, \ldots, n, \tag{7}$$

$$\hat{h}(t)_{ij} = \hat{p}(t)_{ij}(\lambda_i(t) - \lambda_j(t)) \quad \forall i \neq j. \tag{8}$$

For simplicity, let

$$\begin{array}{lll} p := p(0), & p' := p'(0), & \hat{p} := \hat{p}(0), \\ \lambda_i := \lambda_i(0), & \lambda_i' := \lambda_i'(0), & i = 1, \ldots, n. \end{array}$$

Assume f is directionally differentiable at $\lambda_1, \ldots, \lambda_n$. Then we have from $\lambda_i(t) = \lambda_i + t\lambda_i' + o(t)$ and the positive homogeneity property of $f'(\lambda_i; \cdot)$ the expansions

$$p(t) = p + tp' + o(t) \ \text{ and } \ f(\lambda_i(t)) = f(\lambda_i) + tf'(\lambda_i; \lambda_i') + o(t), \quad i = 1, \ldots, n.$$

Also, $p(\cdot)$ and $p'(\cdot)$ are continuous at $t = 0$ so that $\lim_{t\to 0} \hat{h}(t) = p^T h p$ and $\lim_{t\to 0} \hat{p}(t) = \hat{p}$. Using (2) and the above expansions, we then obtain

$$\begin{aligned} f^{\square}(x + th) &= p(t)\mathrm{diag}[f(\lambda_1(t)), \ldots, f(\lambda_n(t))]p(t)^T \\ &= p\ \mathrm{diag}[f(\lambda_1), \ldots, f(\lambda_n)]p^T + t\left(p\ \mathrm{diag}[f'(\lambda_1; \lambda_1'), \ldots, f'(\lambda_n; \lambda_n')]p^T\right) \\ &\quad + t\left(p'\mathrm{diag}[f(\lambda_1), \ldots, f(\lambda_n)]p^T + p\ \mathrm{diag}[f(\lambda_1), \ldots, f(\lambda_n)](p')^T\right) + o(t) \\ &= f^{\square}(x) + tp\ \mathrm{diag}[f'(\lambda_1; \lambda_1'), \ldots, f'(\lambda_n; \lambda_n')]p^T \\ &\quad + tp\left(\hat{p}^T\mathrm{diag}[f(\lambda_1), \ldots, f(\lambda_n)] + \mathrm{diag}[f(\lambda_1), \ldots, f(\lambda_n)]\hat{p}\right)p^T + o(t) \\ &= f^{\square}(x) + tp\ \mathrm{diag}[f'(\lambda_1; \lambda_1'), \ldots, f'(\lambda_n; \lambda_n')]p^T \\ &\quad + tp\left[(f(\lambda_i) - f(\lambda_j))\hat{p}_{ij}\right]_{i,j=1}^n p^T + o(t) \\ &= f^{\square}(x) + tp\ f^{[1]}(\lambda; p^T h p)\ p^T + o(t), \end{aligned} \tag{9}$$

where the fourth equality follows from $p(t)^T p(t) = I$ so that $p'(t)^T p(t) + p(t)^T p'(t) = 0$, implying $\hat{p}^T = -\hat{p}$; the last equality follows from (7) so that $\lambda_i' = \hat{h}(0)_{ii} = (p^T h p)_{ii}$ for $i = 1, \ldots, n$, and from (8) so that $\hat{p}_{ij} = (p^T h p)_{ij}/(\lambda_i - \lambda_j)$ whenever $\lambda_i \neq \lambda_j$ and $(p^T h p)_{ij} = 0$ whenever $\lambda_i = \lambda_j$ and $i \neq j$. It follows from (9) that

$$(f^{\square})'(x; h) = \lim_{t\to 0^+} \frac{f^{\square}(x + th) - f^{\square}(x)}{t} \ = \ p\ f^{[1]}(\lambda; p^T h p)\ p^T.$$

This proves (5).

Suppose instead $f^\square$ is directionally differentiable at x with eigenvalues $\lambda_1, \ldots, \lambda_n$. Fix any $p \in \mathcal{O}$ satisfying $x = p\ \mathrm{diag}[\lambda_1, \ldots, \lambda_n]p^T$. For each $i \in \{1, \ldots, n\}$ and each $d_i \in \mathbb{R}$, let $h := p\ \mathrm{diag}[0, \ldots, d_i, \ldots, 0]p^T$. Then, it is readily verified that $\mathrm{diag}[0, \ldots, f'(\lambda_i; d_i), \ldots, 0] = p^T(f^\square)'(x; h)p$, so $f'(\lambda_i; d_i)$ is well defined.

(b) is an immediate consequence of (a). □

We note that p in the formula for $(f^\square)'(x; h)$ depends on h as well as x. In fact, the proof of Proposition 4.2 shows that a necessary condition for $p(t)$ to comprise orthonormal eigenvectors of $x + th$ that are differentiable at $t = 0$ is that $(p^T hp)_{ij} = 0$ whenever $\lambda_i = \lambda_j$ and $i \neq j$, where $p := p(0)$. In the case of $f(\cdot) = |\cdot|$, directional differentiability of $f^\square$ has been shown by Sun and Sun [29, Lem. 4.8]. In addition, they derived a formula for the directional derivative $(f^\square)'(x; h)$ that also involves $p \in \mathcal{O}_x$ but with p *independent* of h.

For any $\lambda = (\lambda_1, \ldots, \lambda_n)^T \in \mathbb{R}^n$ and any function $f : \mathbb{R} \to \mathbb{R}$ that is differentiable at $\lambda_1, \ldots, \lambda_n$, we denote by $f^{[1]}(\lambda)$ the $n \times n$ symmetric matrix whose (i, j)th entry is

$$f^{[1]}(\lambda)_{ij} = \begin{cases} \dfrac{f(\lambda_i) - f(\lambda_j)}{\lambda_i - \lambda_j} & \text{if } \lambda_i \neq \lambda_j, \\ f'(\lambda_i) & \text{if } \lambda_i = \lambda_j. \end{cases}$$

$f^{[1]}(\lambda)$ is called the first divided difference of f at λ [1, p. 123]. The next proposition, based on Lemmas 3.1, 3.2, and the proof idea for Proposition 4.10, characterizes when $f^\square$ is differentiable (in the Fréchet sense) at an $x \in \mathcal{S}$. This characterization will be needed for computing the generalized Jacobian of a strictly continuous $f^\square$ and for analyzing semismooth property of $f^\square$. We note that the proof idea of Proposition 4.2 cannot be used here because the $p(t)$ constructed in that proof depends on h. In particular, it is not known if $\|p''(t)\|$ is uniformly bounded in $\|h\|$.

PROPOSITION 4.3. *For any $f : \mathbb{R} \to \mathbb{R}$, the following results hold:*

(a) *$f^\square$ is differentiable at an $x \in \mathcal{S}$ with eigenvalues $\lambda_1, \ldots, \lambda_n$ if and only if f is differentiable at $\lambda_1, \ldots, \lambda_n$. Moreover, $\nabla f^\square(x)$ is given by*

$$\nabla f^\square(x)h = p(f^{[1]}(\lambda) \circ (p^T hp))p^T \qquad \forall h \in \mathcal{S} \tag{10}$$

for any $p \in \mathcal{O}$ satisfying $x = p\ \mathrm{diag}[\lambda_1, \ldots, \lambda_n]p^T$, where $\lambda = (\lambda_1, \ldots, \lambda_n)^T$.

(b) *$f^\square$ is differentiable if and only if f is differentiable.*

Proof. (a) Fix any $x \in \mathcal{S}$ and let $\lambda_1, \ldots, \lambda_n$ denote the eigenvalues of x.

It is known [1] that the right-hand side of (10) is independent of the choice of $p \in \mathcal{O}$ satisfying $p^T xp = \mathrm{diag}[\lambda_1, \ldots, \lambda_n]$. This can be seen by noting that any two such p are related by a right multiplication by a block diagonal $o \in \mathcal{O}$ whose diagonal blocks correspond to the distinct eigenvalues of x, while the entries of $f^{[1]}(\lambda)$ in each of these diagonal blocks, as well as in each of the off-diagonal blocks, are equal.

Suppose $f : \mathbb{R} \to \mathbb{R}$ is differentiable at $\lambda_1, \ldots, \lambda_n$. We can without loss of generality assume that $\lambda_1 \geq \cdots \geq \lambda_n$. By Lemma 3.1, there exist scalars $\eta > 0$ and $\epsilon > 0$ such that (3) holds. We will show that, for any $h \in \mathcal{S}$ with $\|h\| \leq \epsilon$, there exists $p \in \mathcal{O}_x$ such that

$$f^\square(x + h) - f^\square(x) - p(c \circ (p^T hp))p^T = o(\|h\|), \tag{11}$$

where $c := f^{[1]}(\lambda)$ and $o(\cdot)$, $O(\cdot)$ depend on f and x only. This together with the independence of the third term on p would show that $f^\square$ is differentiable at x and $\nabla f^\square(x)$ is given by (10) for any $p \in \mathcal{O}$ satisfying $p^T xp = \mathrm{diag}[\lambda_1, \ldots, \lambda_n]$. Let

$\mu_1 \geq \cdots \geq \mu_n$ denote the eigenvalues of $x+h$, and choose any $q \in \mathcal{O}_{x+h}$. Then, there exists $p \in \mathcal{O}_x$ satisfying

$$\|p - q\| \leq \eta \|h\|.$$

For simplicity, let r denote the left-hand side of (11), i.e.,

$$r := f^{\square}(x+h) - f^{\square}(x) - p(c \circ (p^T h p))p^T,$$

and denote $\tilde{r} = p^T r p$ and $\tilde{h} := p^T h p$. Then we have from (2) that

$$\tilde{r} = o^T b o - a - c \circ \tilde{h}, \tag{12}$$

where for simplicity we also denote $a := \operatorname{diag}[f(\lambda_1), \ldots, f(\lambda_n)]$, $b := \operatorname{diag}[f(\mu_1), \ldots, f(\mu_n)]$, and $o := q^T p$.

Since $\operatorname{diag}[\lambda_1, \ldots, \lambda_n] = p^T x p = o^T \operatorname{diag}[\mu_1, \ldots, \mu_n] o - \tilde{h}$, we have

$$\sum_{k=1}^{n} o_{ki} o_{kj} \mu_k - \tilde{h}_{ij} = \begin{cases} \lambda_i & \text{if } i = j; \\ 0 & \text{else,} \end{cases} \qquad i, j = 1, \ldots, n. \tag{13}$$

Since $o = q^T p = (q-p)^T p + I$ and $\|p - q\| \leq \eta \|h\|$, it follows that

$$o_{ij} = O(\|h\|) \quad \forall i \neq j. \tag{14}$$

Since $p, q \in \mathcal{O}$, we have $o \in \mathcal{O}$ so that $o^T o = I$. This implies

$$1 = o_{ii}^2 + \sum_{k \neq i} o_{ki}^2 = o_{ii}^2 + O(\|h\|^2), \quad i = 1, \ldots, n, \tag{15}$$

$$0 = o_{ii} o_{ij} + o_{ji} o_{jj} + \sum_{k \neq i,j} o_{ki} o_{kj} = o_{ii} o_{ij} + o_{ji} o_{jj} + O(\|h\|^2) \quad \forall i \neq j. \tag{16}$$

We now show that $\tilde{r} = o(\|h\|)$ which, by $\|r\| = \|\tilde{r}\|$, would prove (11). For any $i \in \{1, \ldots, n\}$, we have from (12) and (13) that

$$\begin{aligned}
\tilde{r}_{ii} &= \sum_{k=1}^{n} o_{ki}^2 f(\mu_k) - f(\lambda_i) - f'(\lambda_i) \tilde{h}_{ii} \\
&= \sum_{k=1}^{n} o_{ki}^2 f(\mu_k) - f(\lambda_i) - f'(\lambda_i) \left(-\lambda_i + \sum_{k=1}^{n} o_{ki}^2 \mu_k \right) \\
&= o_{ii}^2 f(\mu_i) - f(\lambda_i) - f'(\lambda_i)(-\lambda_i + o_{ii}^2 \mu_i) + O(\|h\|^2) \\
&= (1 + O(\|h\|^2)) f(\mu_i) - f(\lambda_i) - f'(\lambda_i)(-\lambda_i + (1 + O(\|h\|^2))\mu_i) + O(\|h\|^2) \\
&= f(\mu_i) - f(\lambda_i) - f'(\lambda_i)(\mu_i - \lambda_i) + O(\|h\|^2),
\end{aligned}$$

where the third and fifth equalities use (14), (15), and the local boundedness of f. Since f is differentiable at $\lambda_1, \ldots, \lambda_n$ and Lemma 3.2 implies $|\mu_i - \lambda_i| \leq \|h\|$, the right-hand side is $o(\|h\|)$. For any $i, j \in \{1, \ldots, n\}$ with $i \neq j$, we have from (12) and (13) that

$$\begin{aligned}\tilde{r}_{ij} &= \sum_{k=1}^{n} o_{ki}o_{kj}f(\mu_k) - c_{ij}\tilde{h}_{ij}\\ &= \sum_{k=1}^{n} o_{ki}o_{kj}f(\mu_k) - c_{ij}\sum_{k=1}^{n} o_{ki}o_{kj}\mu_k\\ &= o_{ii}o_{ij}f(\mu_i) + o_{ji}o_{jj}f(\mu_j) - c_{ij}(o_{ii}o_{ij}\mu_i + o_{ji}o_{jj}\mu_j) + O(\|h\|^2)\\ &= (o_{ii}o_{ij} + o_{ji}o_{jj})f(\mu_i) + o_{ji}o_{jj}(f(\mu_j) - f(\mu_i))\\ &\quad - c_{ij}\left((o_{ii}o_{ij} + o_{ji}o_{jj})\mu_i + o_{ji}o_{jj}(\mu_j - \mu_i)\right) + O(\|h\|^2)\\ &= o_{ji}o_{jj}\left(f(\mu_j) - f(\mu_i) - c_{ij}(\mu_j - \mu_i)\right) + O(\|h\|^2),\end{aligned}$$

where the third and fifth equalities use (14), (16), and the local boundedness of f. Thus, if $\lambda_i = \lambda_j$, the preceding relation together with (14) and $|\mu_i - \lambda_i| \le \|h\|$, $|\mu_j - \lambda_j| \le \|h\|$ and the continuity of f at λ_i yields

$$\tilde{r}_{ij} = o(\|h\|).$$

If $\lambda_i \ne \lambda_j$, then $c_{ij} = (f(\lambda_j) - f(\lambda_i))/(\lambda_j - \lambda_i)$ and the preceding relation yields

$$\begin{aligned}\tilde{r}_{ij} &= o_{ji}o_{jj}\left(f(\mu_j) - f(\mu_i) - \frac{f(\lambda_j) - f(\lambda_i)}{\lambda_j - \lambda_i}(\mu_j - \mu_i)\right) + O(\|h\|^2)\\ &= o_{ji}o_{jj}\left(f(\mu_j) - f(\mu_i) - (f(\lambda_j) - f(\lambda_i))\left(1 + \frac{\mu_j - \mu_i - \lambda_j + \lambda_i}{\lambda_j - \lambda_i}\right)\right) + O(\|h\|^2).\end{aligned}$$

This together with (14) and $|\mu_i - \lambda_i| \le \|h\|$, $|\mu_j - \lambda_j| \le \|h\|$ and the continuity of f at λ_i and λ_j yields $\tilde{r}_{ij} = o(\|h\|)$.

Suppose $f : \mathbb{R} \to \mathbb{R}$ is not differentiable at λ_i for some $i \in \{1, \dots, n\}$. Then, either f is not directionally differentiable at λ_i or, if it is, the right- and left-directional derivatives of f at λ_i are unequal. In either case, this means there exist two sequences of nonzero scalars t^ν and τ^ν, $\nu = 1, 2, \dots$, converging to zero, such that the limits

$$\lim_{\nu\to\infty} \frac{f(\lambda_i + t^\nu) - f(\lambda_i)}{t^\nu}, \qquad \lim_{\nu\to\infty} \frac{f(\lambda_i + \tau^\nu) - f(\lambda_i)}{\tau^\nu}$$

exist (possibly $-\infty$ or ∞) and either are unequal or are both equal to ∞ or are both equal to $-\infty$. Consider any $p \in \mathcal{O}$ satisfying $x = p\ \mathrm{diag}[\lambda_1, \dots, \lambda_n]p^T$. Then, letting $h = p\mathrm{diag}[0, \dots, 1, \dots, 0]p^T$ with the 1 being in the ith diagonal, we obtain that $x + th = p\mathrm{diag}[\lambda_1, \dots, \lambda_i + t, \dots, \lambda_n]p^T$ for all $t \in \mathbb{R}$ and hence

$$\begin{aligned}\lim_{\nu\to\infty} \frac{f^{\square}(x + t^\nu h) - f^{\square}(x)}{t^\nu} &= p\ \mathrm{diag}\left[0, \dots, 0, \lim_{\nu\to\infty} \frac{f(\lambda_i + t^\nu) - f(\lambda_i)}{t^\nu}, 0, \dots, 0\right] p^T,\\ \lim_{\nu\to\infty} \frac{f^{\square}(x + \tau^\nu h) - f^{\square}(x)}{\tau^\nu} &= p\ \mathrm{diag}\left[0, \dots, 0, \lim_{\nu\to\infty} \frac{f(\lambda_i + \tau^\nu) - f(\lambda_i)}{\tau^\nu}, 0, \dots, 0\right] p^T.\end{aligned}$$

It follows that these two limits either are unequal or are both nonfinite. Thus f is not differentiable at x.

(b) is an immediate consequence of (a). □

Notice that the Jacobian formula (10) is independent of the choice of p and the ordering of $\lambda_1, \dots, \lambda_n$. This formula, together with the differentiability of $f^{\square}$, has been shown under the assumption that f is continuously differentiable—see Theorem V.3.3 and p. 150 of [1]. Proposition 4.3(b) improves on this result by assuming only

that f is differentiable. After obtaining Proposition 4.3, we learned of a closely related recent result of Lewis and Sendov [19] on twice differentiability of spectral functions. In particular, in the case where $f = g'$ for some differentiable $g : \mathbb{R} \to \mathbb{R}$, applying Theorem 3.3 in [19] to the spectral function

$$x \mapsto g(\lambda_1) + \cdots + g(\lambda_n),$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $x \in \mathcal{S}$ in nonincreasing order, yields Proposition 4.3(a). For general f, however, Proposition 4.3(a) appears to be distinct from the results in [19]. In particular, for any $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$, there exists a function $f : \mathbb{R} \to \mathbb{R}$ that is differentiable at $\lambda_1, \ldots, \lambda_n$ and yet there is no differentiable function $g : \mathbb{R} \to \mathbb{R}$ satisfying $g' = f$. One such f is

$$f(\xi) := \begin{cases} (\xi - \lambda_1)^2 & \text{if } \xi \in \{\alpha_1, \alpha_2, \ldots\}; \\ 0 & \text{else,} \end{cases}$$

where $\alpha_1, \alpha_2, \ldots$ is any sequence of points in $\mathbb{R} \backslash \{\lambda_1, \ldots, \lambda_n\}$ converging to λ_1. Here f is differentiable at $\lambda_1, \ldots, \lambda_n$, but the range of f is not an interval, so f cannot be the derivative of a differentiable function. Specifically, a theorem of Darboux says that, for any open interval $\mathcal{I}$ containing a closed interval $[\alpha, \beta]$ and any differentiable $g : \mathcal{I} \to \mathbb{R}$, either $[g'(\alpha), g'(\beta)]$ or $[g'(\beta), g'(\alpha)]$ is a subset of $\{g'(\xi) | \alpha \leq \xi \leq \beta\}$. (This can be seen by defining, for each η strictly between $g'(\alpha)$ and $g'(\beta)$, the function $h(\xi) := g(\xi) - \eta\xi$. Then h is differentiable on $[\alpha, \beta]$ and $h'(\alpha) = g'(\alpha) - \eta$, $h'(\beta) = g'(\beta) - \eta$ have opposite signs. Thus, h has an extremum at some ξ^* in (α, β), implying $h'(\xi^*) = 0$ or, equivalently, $g'(\xi^*) = \eta$.) In fact, any function that coincides with f in a neighborhood of λ_1 cannot be the derivative of a differentiable function. Also, we speculate that the proof idea for Proposition 4.3(a) may be useful for second-or-higher order analysis of spectral functions.

We next have the following continuous differentiability result based on [8, Lem. 4], which in turn was proven using Lemmas 3.1 and 3.2.

PROPOSITION 4.4. *For any $f : \mathbb{R} \to \mathbb{R}$, the matrix function $f^{\square}$ is continuously differentiable if and only if f is continuously differentiable.*

Proof. The "if" direction was proven in [8, Lem. 4]. To see the "only if" direction, suppose $f^{\square}$ is continuously differentiable. Then it follows from (10)and the definition of $f^{[1]}(\cdot)$ that $f'(\lambda_1)$ is well defined for all $\lambda_1 \in \mathbb{R}$. Moreover, $\nabla f^{\square}(\mathrm{diag}[\lambda_1, 0, \ldots, 0])$ is continuous in λ_1 or, equivalently, $f'(\lambda_1)$ is continuous in λ_1. □

Similar to Proposition 4.3, it can be seen that, in the case where $f = g'$ for some differentiable g, Proposition 4.4 is a special case of Theorem 4.2 in [19]. We next have the following result of Rockafellar and Wets [26, Thm. 9.67] which we need to analyze strict continuity and Lipschitz continuity of $f^{\square}$.

LEMMA 4.5. *Suppose $f : \mathbb{R}^k \to \mathbb{R}$ is strictly continuous. Then there exist continuously differentiable functions $f^\nu : \mathbb{R}^k \to \mathbb{R}$, $\nu = 1, 2, \ldots$, converging uniformly to f on any compact set C in $\mathbb{R}^k$ and satisfying*

$$\| |\nabla f^\nu(x)| \| \leq \sup_{x \in C} \ \mathrm{lip} f(x) \qquad \forall x \in C, \ \forall \nu.$$

Lemma 4.5 is slightly different from the original version given in [26, Thm. 9.67]. In particular, the second part of Lemma 4.5 is not contained in [26, Thm. 9.67], but it is implicit in its proof. This second part is needed to show that strict continuity and Lipschitz continuity are inherited by $f^{\square}$ from f. We note that the proof idea

of Proposition 4.1 cannot be used because eigenvectors do not behave in a (locally) Lipschitzian manner.

PROPOSITION 4.6. *For any $f : \mathbb{R} \to \mathbb{R}$, the following results hold:*

(a) *$f^{\square}$ is strictly continuous at an $x \in \mathcal{S}$ with eigenvalues $\lambda_1, \ldots, \lambda_n$ if and only if f is strictly continuous at $\lambda_1, \ldots, \lambda_n$.*

(b) *$f^{\square}$ is strictly continuous if and only if f is strictly continuous.*

(c) *$f^{\square}$ is Lipschitz continuous with constant κ if and only if f is Lipschitz continuous with constant κ.*

Proof. (a) Fix any $x \in \mathcal{S}$ with eigenvalues $\lambda_1, \ldots, \lambda_n$.

Suppose f is strictly continuous at $\lambda_1, \ldots, \lambda_n$. Then, there exist scalars $\kappa_i > 0$ and $\delta_i > 0$, $i = 1, \ldots, n$, such that

$$|f(\xi) - f(\zeta)| \le \kappa_i |\xi - \zeta| \quad \forall \xi, \zeta \in [\lambda_i - \delta_i, \lambda_i + \delta_i]$$

for all i. Let $\tilde{f} : \mathbb{R} \to \mathbb{R}$ be the function that coincides with f on

$$C := \bigcup_{i=1}^{n} [\lambda_i - \delta_i, \lambda_i + \delta_i]$$

and, on $\mathbb{R} \setminus C$, is defined by linearly extrapolating f at the boundary points of C. In other words, if $\xi < \zeta$ are two points in C such that $(\xi, \zeta) \subseteq \mathbb{R} \setminus C$, then $\tilde{f}(t\xi + (1-t)\zeta) = tf(\xi) + (1-t)f(\zeta)$ for all $t \in (0,1)$. If ξ is a point in C such that $(\xi, \infty) \subseteq \mathbb{R} \setminus C$, then $\tilde{f}(\zeta) = f(\xi)$ for all $\zeta > \xi$. Similarly, if ζ is a point in C such that $(-\infty, \zeta) \subseteq \mathbb{R} \setminus C$, then $\tilde{f}(\xi) = f(\zeta)$ for all $\xi < \zeta$. By definition, $\tilde{f}$ is Lipschitz continuous, so there exists a scalar $\kappa > 0$ such that $\mathrm{lip} f(\xi) \le \kappa$ for all $\xi \in \mathbb{R}$. Since C is compact, by Lemma 4.5, there exist continuously differentiable functions $f^\nu : \mathbb{R} \to \mathbb{R}$, $\nu = 1, 2, \ldots$, converging uniformly to $\tilde{f}$ and satisfying

$$|(f^\nu)'(\xi)| \le \kappa \qquad \forall\, \xi \in C,\ \forall \nu. \tag{17}$$

Denote $\delta := \min_{i=1,\ldots,n} \delta_i$. By Lemma 3.2, C contains all the eigenvalues of $y \in \mathcal{B}(x, \delta)$. Moreover, for any $w \in \mathcal{B}(x, \delta)$, any $q \in \mathcal{O}$, and any $\mu = (\mu_1, \ldots, \mu_n)^T \in \mathbb{R}^n$ such that $w = q\, \mathrm{diag}[\mu_1, \ldots, \mu_n] q^T$, we have

$$\begin{aligned} \|(f^\nu)^{\square}(w) - f^{\square}(w)\| &= \|q\, \mathrm{diag}[f^\nu(\mu_1), \ldots, f^\nu(\mu_n)] q^T - q\, \mathrm{diag}[f(\mu_1), \ldots, f(\mu_n)] q^T\| \\ &= \|\mathrm{diag}[f^\nu(\mu_1) - f(\mu_1), \ldots, f^\nu(\mu_n) - f(\mu_n)]\|, \end{aligned}$$

where the second equality uses $q^T q = I$ and properties of the Frobenius norm $\|\cdot\|$. Since $\{f^\nu\}_1^\infty$ converges uniformly to f on C, this shows that $\{(f^\nu)^{\square}\}_1^\infty$ converges uniformly to $f^{\square}$ on $\mathcal{B}(x, \delta)$. Moreover, it follows from (10) that, for all $w \in \mathcal{B}(x, \delta)$ and all ν, we have

$$\begin{aligned} \||\nabla (f^\nu)^{\square}(w)|\| &= \sup_{\|h\|=1} \|\nabla (f^\nu)^{\square}(w) h\| \\ &= \sup_{\|h\|=1} \|q((f^\nu)^{[1]}(\mu) \circ (q^T h q)) q^T\| \\ &= \sup_{\|h\|=1} \|(f^\nu)^{[1]}(\mu) \circ (q^T h q)\| \\ &\le \sup_{\|h\|=1} \kappa \|q^T h q\| \ = \ \kappa, \end{aligned} \tag{18}$$

where the first inequality uses (17). Fix any $y, z \in \mathcal{B}(x,\delta)$ with $y \neq z$. Since $\{(f^\nu)^\square\}_1^\infty$ converges uniformly to $f^\square$ on $\mathcal{B}(x,\delta)$, then for any $\epsilon > 0$ there exists an integer ν_0 such that for all $\nu \geq \nu_0$ we have

$$\|(f^\nu)^\square(w) - f^\square(w)\| \leq \epsilon\|y - z\| \quad \forall w \in \mathcal{B}(x,\delta).$$

Since f^ν is continuously differentiable, then Proposition 4.4 shows that $(f^\nu)^\square$ is continuously differentiable for all ν. Then, by (18) and the mean-value theorem for continuously differentiable functions, we have

$$\begin{aligned}
&\|f^\square(y) - f^\square(z)\| \\
&= \|f^\square(y) - (f^\nu)^\square(y) + (f^\nu)^\square(y) - (f^\nu)^\square(z) + (f^\nu)^\square(z) - f^\square(z)\| \\
&\leq \|f^\square(y) - (f^\nu)^\square(y)\| + \|(f^\nu)^\square(y) - (f^\nu)^\square(z)\| + \|(f^\nu)^\square(z) - f^\square(z)\| \\
&\leq 2\epsilon\|y - z\| + \|\int_0^1 \nabla(f^\nu)^\square(z + \tau(y - z))(y - z)d\tau\| \\
&\leq (\kappa + 2\epsilon)\|y - z\|.
\end{aligned}$$

Since $y, z \in \mathcal{B}(x,\delta)$ and ϵ is arbitrary, this yields

$$\|f^\square(y) - f^\square(z)\| \leq \kappa\|y - z\| \quad \forall y, z \in \mathcal{B}(x,\delta). \tag{19}$$

Thus $f^\square$ is strictly continuous at x.

Suppose instead that $f^\square$ is strictly continuous at x. Then, there exist scalars $\kappa > 0$ and $\delta > 0$ such that (19) holds. Choose any $p \in \mathcal{O}$ satisfying $x = p\,\mathrm{diag}[\lambda_1, \ldots, \lambda_n]p^T$. For any $i \in \{1, \ldots, n\}$ and any $\psi, \zeta \in [\lambda_i - \delta, \lambda_i + \delta]$, let

$$\begin{aligned}
y &:= p\,\mathrm{diag}[\lambda_1, \ldots, \lambda_{i-1}, \psi, \lambda_{i+1}, \ldots, \lambda_n]p^T, \\
z &:= p\,\mathrm{diag}[\lambda_1, \ldots, \lambda_{i-1}, \zeta, \lambda_{i+1}, \ldots, \lambda_n]p^T.
\end{aligned}$$

Then, $\|y - x\| = |\psi - \lambda_i| \leq \delta$ and $\|z - x\| = |\zeta - \lambda_i| \leq \delta$, so it follows from (2) and (19) that

$$\begin{aligned}
|f(\psi) - f(\zeta)| &= \|f^\square(y) - f^\square(z)\| \\
&\leq \kappa\|y - z\| \\
&= \kappa|\psi - \zeta|.
\end{aligned}$$

This shows that f is strictly continuous at λ_i for $i = 1, \ldots, n$.

(b) is an immediate consequence of (a).

(c) Suppose f is Lipschitz continuous with constant κ. Then $\mathrm{lip}f(\xi) \leq \kappa$ for all $\xi \in \mathbb{R}$. Fix any $x \in \mathcal{S}$ with eigenvalues $\lambda_1, \ldots, \lambda_n$. For any scalar $\delta > 0$, define the compact set C in $\mathbb{R}$ by

$$C := \bigcup_{i=1}^n [\lambda_i - \delta, \lambda_i + \delta].$$

Then, as in the proof of (a), we obtain that (19) holds. Since the choice of $\delta > 0$ was arbitrary and κ is independent of δ, this implies

$$\|f^\square(y) - f^\square(z)\| \leq \kappa\|y - z\| \qquad \forall y, z \in \mathcal{S}.$$

Hence $f^\square$ is Lipschitz continuous with constant κ.

Suppose instead that $f^{\square}$ is Lipschitz continuous with constant $\kappa > 0$. Then, for any $\xi, \zeta \in \mathbb{R}$ we have

$$\begin{aligned}|f(\xi) - f(\zeta)| &= \|f^{\square}(\mathrm{diag}[\xi, 0, \ldots, 0]) - f^{\square}(\mathrm{diag}[\zeta, 0, \ldots, 0])\| \\ &\le \kappa \|\mathrm{diag}[\xi, 0, \ldots, 0] - \mathrm{diag}[\zeta, 0, \ldots, 0]\| \\ &= \kappa |\xi - \zeta|,\end{aligned}$$

so f is Lipschitz continuous with constant κ. □

Suppose $f : \mathbb{R} \to \mathbb{R}$ is strictly continuous. Then, by Proposition 4.6, $f^{\square}$ is strictly continuous. Hence $\partial_B f^{\square}(x)$ is well defined for all $x \in \mathcal{S}$. The following lemma studies the structure of this generalized Jacobian.

LEMMA 4.7. *Let $f : \mathbb{R} \to \mathbb{R}$ be strictly continuous. Then, for any $x \in \mathcal{S}$, the generalized Jacobian $\partial_B f^{\square}(x)$ is well defined and nonempty. Moreover, for any $V \in \partial_B f^{\square}(x)$, we have*

$$Vh = p((p^T h p) \circ c) p^T \qquad \forall h \in \mathcal{S} \tag{20}$$

for some $p \in \mathcal{O}_x$, $c \in \mathcal{S}$, and $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ satisfying $x = p\ \mathrm{diag}[\lambda_1, \ldots, \lambda_n] p^T$ and

$$c_{ij} = \frac{f(\lambda_i) - f(\lambda_j)}{\lambda_i - \lambda_j} \quad \text{whenever } \lambda_i \ne \lambda_j, \qquad c_{ij} \in \partial f(\lambda_i) \quad \text{whenever } \lambda_i = \lambda_j. \tag{21}$$

Proof. Fix any $V \in \partial_B f^{\square}(x)$. According to the definition of $\partial_B f^{\square}(x)$, there exists a sequence $\{x_k\} \subseteq \mathcal{S}$ converging to x such that f is differentiable at x_k for all k and $\lim_{k\to\infty} \nabla f^{\square}(x_k) = V$. Let $\lambda_1 \ge \cdots \ge \lambda_n$ and $\lambda_1^k \ge \cdots \ge \lambda_n^k$ be the eigenvalues of x and x_k, $k = 1, 2, \ldots$, respectively. Choose any $p_k \in \mathcal{O}_{x_k}$. By Lemma 3.1, there exist η and $\tilde{p}_k \in \mathcal{O}_x$ satisfying

$$\|p_k - \tilde{p}_k\| \le \eta \|x - x_k\|$$

for all k sufficiently large. By passing to a subsequence if necessary, we assume that this holds for all k and that p_k converges. By Lemma 3.2, we have $\lambda_i^k \to \lambda_i$ for $i = 1, \ldots, n$. Denote $\lambda^k = (\lambda_1^k, \ldots, \lambda_n^k)^T$. Then we have from Proposition 4.3 that f is differentiable at $\lambda_1^k, \ldots, \lambda_n^k$ and

$$\nabla f^{\square}(x_k) h = p_k((p_k^T h p_k) \circ c^k) p_k^T \qquad \forall h \in \mathcal{S}, \tag{22}$$

where we denote $c^k := f^{[1]}(\lambda^k)$. Thus,

$$c_{ij}^k = \begin{cases} (f(\lambda_i^k) - f(\lambda_j^k))/(\lambda_i^k - \lambda_j^k) & \text{if } \lambda_i^k \ne \lambda_j^k; \\ f'(\lambda_i^k) & \text{if } \lambda_i^k = \lambda_j^k. \end{cases} \tag{23}$$

Since f is strictly continuous, then $\{c_{ij}^k\}$ is bounded for all i, j. By passing to a subsequence if necessary, we can assume that $\{c_{ij}^k\}$ converges to some $c_{ij} \in \mathbb{R}$ for all i, j. For each i, we have

$$c_{ii}^k = f'(\lambda_i^k) \to c_{ii} \in \partial_B f(\lambda_i).$$

For each $i \ne j$ such that $\lambda_i \ne \lambda_j$, we have $\lambda_i^k \ne \lambda_j^k$ for all k sufficiently large and hence

$$c_{ij}^k = \frac{f(\lambda_i^k) - f(\lambda_j^k)}{\lambda_i^k - \lambda_j^k} \to c_{ij} = \frac{f(\lambda_i) - f(\lambda_j)}{\lambda_i - \lambda_j}.$$

For each $i \neq j$ such that $\lambda_i = \lambda_j$, if $\lambda_i^k = \lambda_j^k$ for k along some subsequence, then

$$c_{ij}^k = f'(\lambda_i^k) \to c_{ii} \in \partial_B f(\lambda_i) \subseteq \partial f(\lambda_i);$$

if $\lambda_i^k \neq \lambda_j^k$ for k along some subsequence, then a mean-value theorem of Lebourg [9, Proposition 2.3.7], [26, Thm. 10.48] yields

$$c_{ij}^k = \frac{f(\lambda_i^k) - f(\lambda_j^k)}{\lambda_i^k - \lambda_j^k} \in \partial f(\hat{\lambda}_{ij}^k)$$

for some $\hat{\lambda}_{ij}^k$ in the interval between λ_i^k and λ_j^k. Since f is strictly continuous so that ∂f is upper semicontinuous [9, Proposition 2.1.5] or, equivalently, outer semicontinuous [26, Proposition 8.7], this together with $\hat{\lambda}_{ij}^k \to \lambda_i = \lambda_j$ implies the limit of $\{c_{ij}^k\}$ belongs to $\partial f(\lambda_i)$. Thus, taking limits on both sides of (22) and using the above results, we obtain (20) and (21) for some $p \in \mathcal{O}_x$ and $c \in \mathcal{S}$, which are the limit of $\{p_k\}$ and $\{f^{[1]}(\lambda^k)\}$, respectively. This proves the lemma. □

Lemma 4.7 does not, however, provide a characterization of $\partial_B f^{\square}$. It is an open question whether such a (tractable) characterization can be found for any strictly continuous f. In the special case where f is piecewise continuously differentiable (e.g., $f(\cdot) = |\cdot|$) and, more generally, where the directional derivative of f has a one-sided continuity property, a simple characterization of $\partial_B f^{\square}$ can be found as we show below. In what follows we denote the right- and left-directional derivative of $f : \mathbb{R} \to \mathbb{R}$ by

$$f'_+(\xi) := \lim_{\zeta \to \xi^+} \frac{f(\zeta) - f(\xi)}{\zeta - \xi}, \qquad f'_-(\xi) := \lim_{\zeta \to \xi^-} \frac{f(\zeta) - f(\xi)}{\zeta - \xi}.$$

PROPOSITION 4.8. *Let $f : \mathbb{R} \to \mathbb{R}$ be a strictly continuous and directionally differentiable function with the property that*

$$\lim_{\substack{\zeta,\nu \to \xi^\sigma \\ \zeta \neq \nu}} \frac{f(\zeta) - f(\nu)}{\zeta - \nu} = \lim_{\substack{\zeta \to \xi^\sigma \\ \zeta \in D_f}} f'(\zeta) = f'_\sigma(\xi) \quad \forall \xi \in \mathbb{R},\ \sigma \in \{-, +\}, \tag{24}$$

where $D_f := \{\xi \in \mathbb{R} | f$ is differentiable at $\xi\}$. Then, for any $x \in \mathcal{S}$, we have that $V \in \partial_B f^{\square}(x)$ if and only if V has the form (20) for some $p \in \mathcal{O}_x$ and $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ satisfying $x = p\ \mathrm{diag}[\lambda_1, \ldots, \lambda_n]p^T$ and c has the form

$$c_{ij} = \begin{cases} (f(\lambda_i) - f(\lambda_j))/(\lambda_i - \lambda_j) & \text{if } \lambda_i \neq \lambda_j, \\ f'_{\sigma_i}(\lambda_i) & \text{if } \lambda_i = \lambda_j \text{ and } i \in \alpha_l, j \in \beta \cup \alpha_\nu \text{ for some } l < \nu, \\ f'_{\sigma_j}(\lambda_j) & \text{if } \lambda_i = \lambda_j \text{ and } i \in \beta \cup \alpha_l, j \in \alpha_\nu \text{ for some } l > \nu, \\ (\omega_i f'_{\sigma_i}(\lambda_i) + \omega_j f'_{\sigma_j}(\lambda_j))/(\omega_i + \omega_j) & \text{if } \lambda_i = \lambda_j \text{ and } i, j \in \alpha_l \text{ for some } l, \\ f'(\lambda_i) & \text{if } \lambda_i = \lambda_j \text{ and } i, j \in \beta \end{cases} \tag{25}$$

for some partition $\alpha_1, \ldots, \alpha_\ell, \beta$ of $\{1, \ldots, n\}$ ($\ell \geq 0$) and some $\sigma_i \in \{-, +\}$ and $\omega_i \in (0, \infty)$ for $i \in \alpha_1 \cup \cdots \cup \alpha_\ell$. (Implicit in (25) is the differentiability of f at λ_i, $i \in \beta$.)

Proof. Consider any $V \in \partial_B f^{\square}(x)$. By Lemma 4.7 and its proof, V has the form (20) for some $p \in \mathcal{O}_x$ and $\lambda_1 \geq \cdots \geq \lambda_n$ satisfying $x = p\ \mathrm{diag}[\lambda_1, \ldots, \lambda_n]p^T$ and with

c being the cluster point of c^k given by (23), $k = 1, 2, \ldots$ for some $\lambda^k = (\lambda_1^k, \ldots, \lambda_n^k)^T$ converging to $\lambda = (\lambda_1, \ldots, \lambda_n)^T$. Moreover, f is differentiable at $\lambda_1^k, \ldots, \lambda_n^k$ for all k. By passing to a subsequence if necessary, we can assume that, for each $i \in \{1, \ldots, n\}$, either (i) $\lambda_i^k > \lambda_i$ for all k or (ii) $\lambda_i^k < \lambda_i$ for all k or (iii) $\lambda_i^k = \lambda_i$ for all k. Denote

$$\beta := \{i \in \{1, \ldots, n\} | \text{case (iii) holds for } i\}.$$

By further passing to a subsequence if necessary, we can assume that, for each $i, j \in \{1, \ldots, n\} \setminus \beta$,

$$\frac{|\lambda_i^k - \lambda_i|}{|\lambda_j^k - \lambda_j|} \text{ has a limit } \rho_{ij} \in [0, \infty] \text{ as } k \to \infty.$$

Then, $\{1, \ldots, n\} \setminus \beta$ may be partitioned into disjoint subsets $\alpha_1, \ldots, \alpha_\ell$ for some $\ell \geq 0$ such that

$$\begin{array}{ll} \rho_{ij} \in (0, \infty) & \text{whenever } i, j \in \alpha_l \text{ for some } l, \\ \rho_{ij} = \infty & \text{whenever } i \in \alpha_l, j \in \alpha_\nu \text{ for some } l < \nu. \end{array}$$

Moreover, for each $l \in \{1, \ldots, \ell\}$ and each $i \in \alpha_l$, the quantity

$$\omega_i^k := |\lambda_i^k - \lambda_i| / \left(\sum_{j \in \alpha_l} |\lambda_j^k - \lambda_j| \right)$$

converges to a positive limit, which we denote by ω_i. For each $i \in \{1, \ldots, n\} \setminus \beta$, set $\sigma_i = +$ if case (i) holds for i and set $\sigma_i = -$ if case (ii) holds for i. We now verify that c has the form (25). For any $i, j \in \{1, \ldots, n\}$ with $\lambda_i \neq \lambda_j$, this follows from (21). For any $i, j \in \{1, \ldots, n\}$ with $\lambda_i = \lambda_j$, we consider the following disjoint cases.

Case 1. Suppose $i \in \alpha_l$ and $j \in \alpha_\nu$ for some $l, \nu \in \{1, \ldots, \ell\}$ and $\sigma_i = \sigma_j = +$. Then $\lambda_i^k > \lambda_i$ and $\lambda_j^k > \lambda_i$ for all k. If $l = \nu$, it follows from (23) and (24) that

$$c_{ij}^k \to f'_+(\lambda_i) = (\omega_i f'_{\sigma_i}(\lambda_i) + \omega_j f'_{\sigma_j}(\lambda_j))/(\omega_i + \omega_j) = c_{ij},$$

where the last equality uses (25). If $l < \nu$, a similar argument shows that

$$c_{ij}^k \to f'_+(\lambda_i) = f'_{\sigma_i}(\lambda_i) = c_{ij}.$$

The remaining subcase of $l > \nu$ can be treated analogously.

Case 2. Suppose $i \in \alpha_l$ and $j \in \alpha_\nu$ for some $l, \nu \in \{1, \ldots, \ell\}$ and $\sigma_i = +, \sigma_j = -$. Then $\lambda_i^k > \lambda_i$ and $\lambda_j^k < \lambda_i$ for all k. If $l = \nu$, it follows from (23) and (24) that

$$\begin{aligned} c_{ij}^k &= \frac{f(\lambda_i^k) - f(\lambda_j^k)}{\lambda_i^k - \lambda_j^k} \\ &= \frac{\omega_i^k}{\omega_i^k + \omega_j^k} \frac{f(\lambda_i^k) - f(\lambda_i)}{\lambda_i^k - \lambda_i} + \frac{\omega_j^k}{\omega_i^k + \omega_j^k} \frac{f(\lambda_j^k) - f(\lambda_i)}{\lambda_j^k - \lambda_i} \\ &\to \frac{\omega_i}{\omega_i + \omega_j} f'_+(\lambda_i) + \frac{\omega_j}{\omega_i + \omega_j} f'_-(\lambda_j) \\ &= (\omega_i f'_{\sigma_i}(\lambda_i) + \omega_j f'_{\sigma_j}(\lambda_j))/(\omega_i + \omega_j) \\ &= c_{ij}, \end{aligned}$$

where the last equality uses (25). If $l < \nu$, a similar argument together with $\rho_{ij} = \infty$ shows that

$$\begin{aligned} c_{ij}^k &= \frac{|\lambda_i^k - \lambda_i|}{|\lambda_i^k - \lambda_i| + |\lambda_j^k - \lambda_j|} \frac{f(\lambda_i^k) - f(\lambda_i)}{\lambda_i^k - \lambda_i} + \frac{|\lambda_j^k - \lambda_j|}{|\lambda_i^k - \lambda_i| + |\lambda_j^k - \lambda_j|} \frac{f(\lambda_j^k) - f(\lambda_i)}{\lambda_j^k - \lambda_i} \\ &\to f'_+(\lambda_i) \\ &= c_{ij}. \end{aligned}$$

The remaining subcase of $l > \nu$ can be treated analogously.

Case 3. Suppose $i \in \alpha_l$ and $j \in \beta$ for some $l \in \{1, \ldots, \ell\}$ and $\sigma_i = +$. Then $\lambda_i^k > \lambda_i$ and $\lambda_j^k = \lambda_i$ for all k. It follows from (23) and (24) that

$$c_{ij}^k = \frac{f(\lambda_i^k) - f(\lambda_i)}{\lambda_i^k - \lambda_i} \to f'_+(\lambda_i) = c_{ij}.$$

Case 4. Suppose $i, j \in \beta$. Then $\lambda_i^k = \lambda_j^k = \lambda_i$ for all k and it follows from (23) that f is differentiable at λ_i, $i \in \beta$, and

$$c_{ij}^k = f'(\lambda_i) = c_{ij}.$$

Case 5. Suppose $i \in \alpha_l$ and $j \in \alpha_\nu$ for some $l, \nu \in \{1, \ldots, \ell\}$ and $\sigma_i = \sigma_j = -$. This case is analogous to Case 1.

Case 6. Suppose $i \in \alpha_l$ and $j \in \beta$ for some $l \in \{1, \ldots, \ell\}$ and $\sigma_i = -$. This case is analogous to Case 3.

Conversely, suppose that V has the form (20) for some $p \in \mathcal{O}_x$ and $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ satisfying $x = p \operatorname{diag}[\lambda_1, \ldots, \lambda_n] p^T$ and c has the form (25) for some partition $\alpha_1, \ldots, \alpha_\ell, \beta$ of $\{1, \ldots, n\}$ ($\ell \geq 0$) and some $\sigma_i \in \{-, +\}$ and $\omega_i \in (0, \infty)$ for $i \in \alpha_1 \cup \cdots \cup \alpha_\ell$. For each $i \in \beta$, set $d_i^k := 0$ for $k = 1, 2, \ldots$. For each $i \in \alpha_l$, $l \in \{1, \ldots, \ell\}$, let $\delta_i^k = \omega_i (1/2)^{kl}$ if $\sigma_i = +$ and let $\delta_i^k = -\omega_i (1/2)^{kl}$ if $\sigma_i = -$, $k = 1, 2, \ldots$. Since f is strictly continuous, by Rademacher's theorem (see [26, Thm. 9.60]), D_f is dense in $\mathbb{R}$. Thus, for each $i \in \alpha_1 \cup \cdots \cup \alpha_\ell$ and each index k, there exists $d_i^k \in \mathbb{R}$ satisfying

$$\lambda_i + d_i^k \in D_f \quad \text{and} \quad |d_i^k - \delta_i^k| \leq |\delta_i^k|^2.$$

Let $\lambda_i^k := \lambda_i + d_i^k$ for all i. Then, by Proposition 4.3, $f^\square$ is differentiable at

$$x^k := p \operatorname{diag}[\lambda_1^k, \ldots, \lambda_n^k] p^T$$

for all k and

$$\nabla f^\square(x^k) h = p(c^k \circ (p^T h p)) p^T \quad \forall h \in \mathcal{S},$$

where c^k is given by (23). Also, the definition of $d_1^k, \ldots, d_n^k$ yields

$$d_i^k \to 0 \ \forall i, \quad \frac{|d_i^k|}{|d_j^k|} \to \frac{\omega_i}{\omega_j} \ \forall i, j \in \alpha_l, \ l = 1, \ldots, \ell, \quad \frac{|d_i^k|}{|d_j^k|} \to \infty \ \forall i \in \alpha_l, j \in \alpha_\nu, l < \nu,$$

and $\sigma_i = +$ implies $d_i^k > 0$ for all k and $\sigma_i = -$ implies $d_i^k < 0$ for all k. Then, it is straightforward to verify that $x^k \to x$ and $c^k \to c$, implying

$$\nabla f^\square(x^k) h \to p(c \circ (p^T h p)) p^T = V h \quad \forall h \in \mathcal{S}.$$

This shows that $V \in \partial_B f^{\square}(x)$. □

Notice that a V of the form (20) is invertible if and only all entries of c are nonzero. Also, notice that the p in the formula (20) depends on V; i.e., two elements of $\partial_B f^{\square}(x)$ may have different p in their formulas. Thus $\partial f^{\square}(x)$, being the convex hull of $\partial_B f^{\square}(x)$, has a rather complicated structure.

The following lemma, proven by Sun and Sun [29, Thm. 3.6] using the definition of generalized Jacobian,[1] enables one to study the semismooth property of $f^{\square}$ by examining only those points $x \in \mathcal{S}$ where $f^{\square}$ is differentiable and thus work only with the Jacobian of $f^{\square}$, rather than the generalized Jacobian.

LEMMA 4.9. *Suppose $F : \mathcal{S} \to \mathcal{S}$ is strictly continuous and directionally differentiable in a neighborhood of $x \in \mathcal{S}$. Then, for any $0 < \rho < \infty$, the following two statements (where $O(\cdot)$ depends on F and x only) are equivalent:*

(a) *For any $h \in \mathcal{S}$ and any $V \in \partial F(x+h)$,*

$$F(x+h) - F(x) - Vh = o(\|h\|) \quad (\textit{respectively, } O(\|h\|^{1+\rho})).$$

(b) *For any $h \in \mathcal{S}$ such that F is differentiable at $x+h$,*

$$F(x+h) - F(x) - \nabla F(x+h)h = o(\|h\|) \quad (\textit{respectively, } O(\|h\|^{1+\rho})).$$

By using Lemmas 3.1, 3.2, and 4.9 and Propositions 4.2, 4.3, and 4.6, we are now ready to state and prove the last result of this section. The proof is motivated by and in some sense generalizes the proof of Lemma 4.12 in [29], though it is also simpler. The proof idea was also used for proving Proposition 4.3, with the main difference being that here $x+h$ is diagonalized rather than x.

PROPOSITION 4.10. *For any $f : \mathbb{R} \to \mathbb{R}$, the matrix function $f^{\square}$ is semismooth if and only if f is semismooth. If $f : \mathbb{R} \to \mathbb{R}$ is ρ-order semismooth $(0 < \rho < \infty)$, then $f^{\square}$ is $\min\{1, \rho\}$-order semismooth.*

Proof. Suppose f is semismooth. Then f is strictly continuous and directionally differentiable. By Propositions 4.2 and 4.6, $f^{\square}$ is strictly continuous and directionally differentiable. Let $\mathcal{D} := \{x \in \mathcal{S} | f^{\square} \text{ is differentiable at } x\}$.

Fix any $x \in \mathcal{S}$ and let $\lambda_1 \geq \cdots \geq \lambda_n$ denote the eigenvalues of x. By Lemma 3.1, there exist scalars $\eta > 0$ and $\epsilon > 0$ such that (3) holds. By taking ϵ smaller if necessary, we can assume that $\epsilon < (\lambda_i - \lambda_{i+1})/2$ whenever $\lambda_i \neq \lambda_{i+1}$. We will show that, for any $h \in \mathcal{S}$ with $x + h \in \mathcal{D}$ and $\|h\| \leq \epsilon$, we have

$$f^{\square}(x+h) - f^{\square}(x) - \nabla f^{\square}(x+h)h = o(\|h\|), \tag{26}$$

where $o(\cdot)$ and $O(\cdot)$ depend on f and x only. Then, it follows from Lemma 4.9 that $f^{\square}$ is semismooth at x. Since the choice of $x \in \mathcal{S}$ was arbitrary, $f^{\square}$ is semismooth. Let $\mu_1 \geq \cdots \geq \mu_n$ denote the eigenvalues of $x+h$, and choose any $q \in \mathcal{O}_{x+h}$. Then, there exists $p \in \mathcal{O}_x$ satisfying

$$\|p - q\| \leq \eta \|h\|.$$

For simplicity, let r denote the left-hand side of (26), i.e.,

$$r := f^{\square}(x+h) - f^{\square}(x) - \nabla f^{\square}(x+h)h,$$

[1] Sun and Sun did not consider the case of $o(\|h\|)$, but their argument readily applies to this case.

and denote $\tilde{r} = q^T r q$ and $\tilde{h} := q^T h q$. Since $x + h \in \mathcal{D}$, Proposition 4.3 implies f is differentiable at $\mu_1, \ldots, \mu_n$. Then we have from (2) and (10) that

$$\tilde{r} = b - o^T a o - c \circ \tilde{h}, \tag{27}$$

where for simplicity we also denote $a := \operatorname{diag}[f(\lambda_1), \ldots, f(\lambda_n)]$, $b := \operatorname{diag}[f(\mu_1), \ldots, f(\mu_n)]$, $c := f^{[1]}(\mu)$, and $o := p^T q$.

Since $\operatorname{diag}[\mu_1, \ldots, \mu_n] = q^T(x+h)q = o^T \operatorname{diag}[\lambda_1, \ldots, \lambda_n] o + \tilde{h}$, we have

$$\sum_{k=1}^{n} o_{ki} o_{kj} \lambda_k + \tilde{h}_{ij} = \begin{cases} \mu_i & \text{if } i = j, \\ 0 & \text{else,} \end{cases} \qquad i, j = 1, \ldots, n. \tag{28}$$

Since $o = p^T q = (p-q)^T q + I$ and $\|p - q\| \le \eta \|h\|$, it follows that

$$o_{ij} = O(\|h\|) \quad \forall i \ne j. \tag{29}$$

Since $p, q \in \mathcal{O}$, we have $o \in \mathcal{O}$ so that $o^T o = I$. This implies

$$1 = o_{ii}^2 + \sum_{k \ne i} o_{ki}^2 = o_{ii}^2 + O(\|h\|^2), \quad i = 1, \ldots, n, \tag{30}$$

$$0 = o_{ii} o_{ij} + o_{ji} o_{jj} + \sum_{k \ne i,j} o_{ki} o_{kj} = o_{ii} o_{ij} + o_{ji} o_{jj} + O(\|h\|^2) \quad \forall i \ne j. \tag{31}$$

We now show that $\tilde{r} = o(\|h\|)$ which, by $\|r\| = \|\tilde{r}\|$, would prove (26). For any $i \in \{1, \ldots, n\}$, we have from (27) and (28) that

$$\begin{aligned}
\tilde{r}_{ii} &= f(\mu_i) - \sum_{k=1}^{n} o_{ki}^2 f(\lambda_k) - f'(\mu_i) \tilde{h}_{ii} \\
&= f(\mu_i) - \sum_{k=1}^{n} o_{ki}^2 f(\lambda_k) - f'(\mu_i) \left(\mu_i - \sum_{k=1}^{n} o_{ki}^2 \lambda_k \right) \\
&= f(\mu_i) - o_{ii}^2 f(\lambda_i) - f'(\mu_i)(\mu_i - o_{ii}^2 \lambda_i) + O(\|h\|^2) \\
&= f(\mu_i) - (1 + O(\|h\|^2)) f(\lambda_i) - f'(\mu_i)(\mu_i - (1 + O(\|h\|^2))\lambda_i) + O(\|h\|^2) \\
&= f(\mu_i) - f(\lambda_i) - f'(\mu_i)(\mu_i - \lambda_i) + O(\|h\|^2),
\end{aligned}$$

where the third and fifth equalities use (29), (30), and the local boundedness of f and f'. Since f is semismooth and Lemma 3.2 implies $|\mu_i - \lambda_i| \le \|h\|$, then clearly the right-hand side is of $o(\|h\|)$. For any $i, j \in \{1, \ldots, n\}$ with $i \ne j$, we have from (27) and (28) that

$$\begin{aligned}
\tilde{r}_{ij} &= -\sum_{k=1}^{n} o_{ki} o_{kj} f(\lambda_k) - c_{ij} \tilde{h}_{ij} \\
&= -\sum_{k=1}^{n} o_{ki} o_{kj} f(\lambda_k) + c_{ij} \sum_{k=1}^{n} o_{ki} o_{kj} \lambda_k \\
&= -(o_{ii} o_{ij} f(\lambda_i) + o_{ji} o_{jj} f(\lambda_j)) + c_{ij}(o_{ii} o_{ij} \lambda_i + o_{ji} o_{jj} \lambda_j) + O(\|h\|^2) \\
&= -\left((o_{ii} o_{ij} + o_{ji} o_{jj}) f(\lambda_i) + o_{ji} o_{jj}(f(\lambda_j) - f(\lambda_i))\right) \\
&\quad + c_{ij}\left((o_{ii} o_{ij} + o_{ji} o_{jj})\lambda_i + o_{ji} o_{jj}(\lambda_j - \lambda_i)\right) + O(\|h\|^2) \\
&= -o_{ji} o_{jj}\left(f(\lambda_j) - f(\lambda_i) - c_{ij}(\lambda_j - \lambda_i)\right) + O(\|h\|^2),
\end{aligned}$$

where the third and fifth equalities use (29), (31), and the local boundedness of f and f'. Thus, if $\lambda_i = \lambda_j$, the preceding relation yields

$$\tilde{r}_{ij} = O(\|h\|^2).$$

If $\lambda_i \neq \lambda_j$, then Lemma 3.2 implies $|\mu_i - \lambda_i| \leq \|h\|$ and $|\mu_j - \lambda_j| \leq \|h\|$ so that $|\mu_i - \mu_j| = |\lambda_i - \lambda_j - (\lambda_i - \mu_i) + (\lambda_j - \mu_j)| \geq |\lambda_i - \lambda_j| - 2\|h\| > 2\epsilon - 2\|h\| \geq 0$. Hence $\mu_i \neq \mu_j$, so $c_{ij} = (f(\mu_j) - f(\mu_i))/(\mu_j - \mu_i)$ and the preceding relation yields

$$\begin{aligned}\tilde{r}_{ij} &= -o_{ji}o_{jj}\left(f(\lambda_j) - f(\lambda_i) - \frac{f(\mu_j) - f(\mu_i)}{\mu_j - \mu_i}(\lambda_j - \lambda_i)\right) + O(\|h\|^2)\\ &= -o_{ji}o_{jj}\left(f(\lambda_j) - f(\lambda_i) - (f(\mu_j) - f(\mu_i))\left(1 + \frac{\lambda_j - \lambda_i - \mu_j + \mu_i}{\mu_j - \mu_i}\right)\right) + O(\|h\|^2)\\ &= O(\|h\|^2),\end{aligned}$$

where the last equality uses (29) and the strict continuity of f at λ_i, λ_j, so that $f(\mu_i) - f(\lambda_i) = O(|\mu_i - \lambda_i|) = O(\|h\|)$ and $f(\mu_j) - f(\lambda_j) = O(|\mu_j - \lambda_j|) = O(\|h\|)$.

Suppose f is ρ-order semismooth ($0 < \rho < \infty$). Then the preceding argument shows that $\tilde{r}_{ii} = O(\max\{\|h\|^{1+\rho}, \|h\|^2\}) = O(\|h\|^{1+\min\{1,\rho\}})$ for all i while we still have $\tilde{r}_{ij} = O(\|h\|^2)$ for all $i \neq j$. This shows that $f^{\square}$ is $\min\{1, \rho\}$-order semismooth at x. Since the choice of $x \in \mathcal{S}$ was arbitrary, $f^{\square}$ is $\min\{1, \rho\}$-order semismooth.

Suppose $f^{\square}$ is semismooth. Then $f^{\square}$ is strictly continuous and directionally differentiable. By Propositions 4.2 and 4.6, f is strictly continuous and directionally differentiable. For any $\xi \in \mathbb{R}$ and any $\eta \in \mathbb{R}$ such that f is differentiable at $\xi+\eta$, Proposition 4.3 yields that $f^{\square}$ is differentiable at $x + h$, where we denote $x := \mathrm{diag}[\xi, \ldots, \xi] = \xi I$ and $h := \mathrm{diag}[\eta, \ldots, \eta] = \eta I$. Since $f^{\square}$ is semismooth, it follows from Lemma 4.9 that

$$f^{\square}(x + h) - f^{\square}(x) - \nabla f^{\square}(x + h)h = o(\|h\|),$$

which, by (2) and (10), is equivalent to

$$f(\xi + \eta) - f(\xi) - f'(\xi + \eta)\eta = o(|\eta|).$$

Then Lemma 4.9 yields that f is semismooth. □

We note that for each of the preceding global results there is a corresponding local result. This can be seen from our proofs where, in order to show that a global property of f is inherited by $f^{\square}$, we first show that this property is locally inherited from f by $f^{\square}$. For example, we can show the following local analogue of Proposition 4.4: If $f : \mathbb{R} \to \mathbb{R}$ is continuously differentiable at each of the eigenvalues of $x \in \mathcal{S}$, then $f^{\square}$ is continuously differentiable at x and $\nabla f^{\square}(x)$ is given by (10).

5. Applications to the SDCP. In this section, we consider the semidefinite complementarity problem (SDCP), which is to find, for a given function $F : \mathcal{S} \to \mathcal{S}$, an $(x, y) \in \mathcal{S} \times \mathcal{S}$ satisfying

$$x \in \mathcal{S}_+, \quad y \in \mathcal{S}_+, \quad \langle x, y\rangle = 0, \quad F(x) - y = 0, \tag{32}$$

where $\mathcal{S}_+$ denotes the convex cone comprising those $x \in \mathcal{S}$ that are positive semidefinite. We assume that F is continuously differentiable. The SDCP includes as a special case the nonlinear complementarity problem (NCP), where $n_1 = \cdots = n_m = 1$. It is also connected to eigenvalue optimization [18]. There has been much interest in the

numerical solution of the SDCP (32) using, e.g., the interior-point approach [27], the merit function approach [30, 32], and the noninterior smoothing approach [8] (also see references therein). We will consider a related approach of reformulating the SDCP as a semismooth equation and then, by applying the results of section 4, study issues relevant to the design and analysis of smoothing Newton methods based on this reformulation.

It is known [30, Proposition 2.1] that $(x, y) \in \mathcal{S} \times \mathcal{S}$ solves the SDCP if and only if it solves the equations

$$H(x,y) := \begin{pmatrix} x - [x-y]_+ \\ F(x) - y \end{pmatrix} = 0, \tag{33}$$

where $[\cdot]_+ : \mathcal{S} \to \mathcal{S}_+$ denotes the nearest-point projection onto $\mathcal{S}_+$, i.e.,

$$[x]_+ := \arg\min\{\|x - y\| \mid y \in \mathcal{S}_+\}.$$

The function H is nonsmooth due to the nonsmoothness of the matrix projection operator $[\cdot]_+$. However, it was shown by Sun and Sun [29] that $[\cdot]_+$ is strongly semismooth, so that H is semismooth. We will see that this result also follows from Proposition 4.10 and, in particular, $f^{\square}(\cdot) = [\cdot]_+$ with $f(\cdot) = \max\{0, \cdot\}$ (Proposition 5.2).

There have been many smoothing methods proposed for solving semismooth equation reformulation of the NCP—see [2, 3, 4, 5, 6, 7, 11, 16, 22, 24] and references therein. These methods are based on making accurate smooth approximation of the semismooth equations. In particular, the smoothing method studied by Chen, Qi, and Sun [6] and later studied by Kanzow and Pieper [16] have an accuracy criterion called the Jacobian Consistence Property. We will verify this property with respect to a class of smoothing functions H_μ for H, as proposed by Chen and Mangasarian [4, 5] for the case of the linear program (LP) and the NCP and recently extended in [8] to the SDCP. This property, together with semismoothness of H, allows the development of methods of the form

$$(x^{k+1}, y^{k+1}) = (x^k, y^k) - t_k \nabla H_{\mu_k}(x^k, y^k)^{-1} H(x^k, y^k), \quad k = 0, 1, \ldots,$$

with $t_k > 0$ and $\mu_k \downarrow 0$ suitably chosen, that achieve both global convergence and local superlinear convergence, assuming nonsingularity of all $V \in \partial H(x, y)$ locally; see [6, Thm. 3.2]. Such methods have the advantage of requiring only one linear equation solve per iteration, in contrast to the two (or more) linear equation solves required by other smoothing methods having similar global and local convergence properties. Thus, our study paves the way for extending methods of the above form from the NCP to the SDCP. This, for example, would improve on the methods of [8, 15] which require two linear equation solves per iteration.

Let $\mathcal{CM}$ denote the class of convex continuously differentiable functions $g : \mathbb{R} \to \mathbb{R}$ with the properties

$$\lim_{\tau \to -\infty} g(\tau) = 0, \quad \lim_{\tau \to \infty} g(\tau) - \tau = 0, \quad \text{and} \quad 0 < g'(\tau) < 1 \ \forall \tau \in \mathbb{R}.$$

Two typical examples of g are the so-called *CHKS function* $g(\tau) = ((\tau^2 + 4)^{1/2} + \tau)/2$ and the *neural network function* $g(\tau) = \ln(e^\tau + 1)$. For any $g \in \mathcal{CM}$, consider the following smooth approximation of $x - [x-y]_+$, as proposed by Chen and Mangasarian [4, 5] for the case of the LP and the NCP:

$$\phi_\mu(x, y) := x - \mu g^{\square}((x - y)/\mu), \quad \mu > 0. \tag{34}$$

It was shown in [8, Lem. 1] that the limit $\lim_{\mu\to 0}\phi_\mu(x,y)$ exists and is equal to $x-[x-y]_+$. Moreover, one has [8, Cor. 1]

$$\|\phi_\mu(x,y)-(x-[x-y]_+)\|\le\sqrt{n}g(0)\mu, \tag{35}$$

and ϕ_μ is continuously differentiable for any $\mu>0$ [8, Lem. 2]. Hence a smooth approximation of $H(x,y)$ is

$$H_\mu(x,y):=\begin{pmatrix}\phi_\mu(x,y)\\ F(x)-y\end{pmatrix},\qquad \mu>0. \tag{36}$$

We say that H_μ has the Jacobian Consistence Property relative to H if there exists a constant $\kappa>0$ such that, for any $(x,y)\in\mathcal{S}\times\mathcal{S}$, we have (i)

$$\|H_\mu(x,y)-H(x,y)\|\le\kappa\mu\qquad\forall\mu>0 \tag{37}$$

and (ii)

$$\lim_{\mu\to 0^+}\operatorname{dist}(\nabla H_\mu(x,y),\partial H(x,y))=0; \tag{38}$$

i.e., the distance between $\nabla H_\mu(x,y)$ and the set $\partial H(x,y)$ approaches zero as μ is decreased to zero. Here, we denote $\operatorname{dist}(L,\mathcal{M}):=\inf_{M\in\mathcal{M}}|||L-M|||$ for any linear mapping $L:\mathcal{S}\times\mathcal{S}\to\mathcal{S}\times\mathcal{S}$ and any nonempty collection $\mathcal{M}$ of linear mappings from $\mathcal{S}\times\mathcal{S}$ to $\mathcal{S}\times\mathcal{S}$. Also, for any $(x,y)\in\mathcal{S}\times\mathcal{S}$, we define $\|(x,y)\|=\sqrt{\|x\|^2+\|y\|^2}$. We show below that H is semismooth and H_μ has the Jacobian Consistence Property relative to H. These results facilitate the extension of the smoothing Newton methods of Chen, Qi, and Sun [6] for the NCP, later studied by Kanzow and Pieper [16], to the SDCP. Such methods are promising. For example, a smoothing method of [8], based on (34) and (36) with g being the CHKS function, is comparable to primal-dual interior-point methods in terms of the number of iterations to solve benchmark semidefinite programs with relative infeasibility and duality gap below $3\cdot 10^{-9}$. As with interior-point methods and barrier/penalty methods, the smoothing parameter μ needs to be small to obtain an accurate solution and, as μ becomes smaller, $\nabla H_\mu(x,y)$ can become more ill-conditioned. Thus, such smoothing methods could have difficulty achieving solution accuracy much greater than 10^{-9}.

We begin with the following lemma showing that the Jacobian Consistence Property is inherited by $f^\square$ and its smooth approximations from f and its smooth approximations.

LEMMA 5.1. *Let $f:\mathbb{R}\to\mathbb{R}$ be a strictly continuous function. Let $f_\mu:\mathbb{R}\to\mathbb{R}$, $\mu>0$, be differentiable functions such that there exists a scalar constant $\kappa>0$ for which*

$$|f_\mu(\zeta)-f(\zeta)|\le\kappa\mu\quad\forall\mu>0, \tag{39}$$

$$\lim_{\mu\to 0^+}\operatorname{dist}(f'_\mu(\zeta),\partial f(\zeta))=0 \tag{40}$$

for all $\zeta\in\mathbb{R}$. Then, for any $z\in\mathcal{S}$, we have

$$\|f^\square_\mu(z)-f^\square(z)\|\le\sqrt{n}\kappa\mu\quad\forall\mu>0, \tag{41}$$

$$\lim_{\mu\to 0^+}\operatorname{dist}(\nabla f^\square_\mu(z),\partial f^\square(z))=0. \tag{42}$$

Proof. Fix any $z \in \mathcal{S}$. Consider any $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ and any $p \in \mathcal{O}$ satisfying $z = p\,\mathrm{diag}[\lambda_1, \ldots, \lambda_n]p^T$.

By (1) and (2), we have

$$\begin{aligned}\|f_\mu^\square(z) - f^\square(z)\| &= \|p^T f_\mu^\square(z)p - p^T f^\square(z)p\| \\ &= \|\mathrm{diag}[f_\mu(\lambda_1) - f(\lambda_1), \ldots, f_\mu(\lambda_n) - f(\lambda_n)]\| \\ &\le \sqrt{n}\kappa\mu,\end{aligned}$$

where the last inequality uses (39). This proves (41).

We now prove (42). For any $\mu > 0$, since f_μ is differentiable, then Proposition 4.3 yields that $f_\mu^\square$ is differentiable and

$$\nabla f_\mu^\square(z)h = p(c_\mu \circ (p^T hp))p^T \qquad \forall h \in \mathcal{S}, \tag{43}$$

where $c_\mu := f_\mu^{[1]}(\lambda)$ and $\lambda := (\lambda_1, \ldots, \lambda_n)^T$. Let $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_m$ denote the *distinct* eigenvalues of z and denote $\mathcal{I}_k := \{i \in \{1, \ldots, n\} | \lambda_i = \tilde{\lambda}_k\}$, $k = 1, \ldots, m$. We have

$$(c_\mu)_{ij} = \begin{cases} (f_\mu(\tilde{\lambda}_k) - f_\mu(\tilde{\lambda}_\ell))/(\tilde{\lambda}_k - \tilde{\lambda}_\ell) & \text{if } i \in \mathcal{I}_k,\ j \in \mathcal{I}_\ell \text{ for some } k \ne \ell, \\ f'_\mu(\tilde{\lambda}_k) & \text{if } i, j \in \mathcal{I}_k \text{ for some } k. \end{cases} \tag{44}$$

By (39) and (40), for each $\epsilon > 0$ there exists $\delta > 0$ such that for each $\mu \in (0, \delta)$ we have

$$|f_\mu(\tilde{\lambda}_k) - f(\tilde{\lambda}_k)| < \epsilon \quad \text{and} \quad |f'_\mu(\tilde{\lambda}_k) - v_k| < \epsilon, \qquad k = 1, \ldots, m, \tag{45}$$

for some $v_k \in \partial f(\tilde{\lambda}_k)$ depending on μ. Letting $c \in \mathcal{S}$ denote the symmetric matrix whose (i, j)th entry is

$$c_{ij} := \begin{cases} (f(\tilde{\lambda}_k) - f(\tilde{\lambda}_\ell))/(\tilde{\lambda}_k - \tilde{\lambda}_\ell) & \text{if } i \in \mathcal{I}_k,\ j \in \mathcal{I}_\ell \text{ for some } k \ne \ell, \\ v_k & \text{if } i, j \in \mathcal{I}_k \text{ for some } k, \end{cases} \tag{46}$$

we then obtain from (39), (44), (45), and (46) that

$$|(c_\mu)_{ij} - c_{ij}| < \epsilon\beta \qquad \forall i, j = 1, \ldots, n, \tag{47}$$

where $\beta > 0$ is a scalar independent of μ and ϵ. Define the linear mapping $V : \mathcal{S} \to \mathcal{S}$ by

$$Vh := p(c \circ (p^T hp))p^T \qquad \forall h \in \mathcal{S}. \tag{48}$$

Then V depends on μ and, by (43) and (47), we have

$$\||\nabla f_\mu^\square(z) - V\|| = \sup_{\|h\|=1} \|\nabla f_\mu^\square(z)h - Vh\| = \sup_{\|h\|=1} \|(c_\mu - c) \circ (p^T hp)\| < \epsilon\beta.$$

Thus $\||\nabla f_\mu^\square(z) - V\|| \to 0$ as $\mu \to 0^+$. We now show that V belongs to $\partial f^\square(z)$. For each $k \in \{1, \ldots, m\}$, since $v_k \in \partial f(\tilde{\lambda}_k)$, there exist integer $\tau_k \ge 1$ and $v_k[\nu] \in \partial_B f(\tilde{\lambda}_k)$ and $\omega_k[\nu] \in (0, \infty)$, $\nu = 1, \ldots, \tau_k$, satisfying

$$\sum_{\nu=1}^{\tau_k} \omega_k[\nu] = 1, \qquad \sum_{\nu=1}^{\tau_k} \omega_k[\nu]\ v_k[\nu] = v_k.$$

Then, it is straightforward to verify that

$$\sum_{\nu_1=1}^{\tau_1}\cdots\sum_{\nu_m=1}^{\tau_m}\left(\prod_{k=1}^{m}\omega_k[\nu_k]\right)=1,\qquad \sum_{\nu_1=1}^{\tau_1}\cdots\sum_{\nu_m=1}^{\tau_m}\left(\prod_{k=1}^{m}\omega_k[\nu_k]\right)c[\nu_1,\ldots,\nu_m]=c,$$

where $c[\nu_1,\ldots,\nu_m]\in\mathcal{S}$ denotes the symmetric matrix whose (i,j)th entry is

$$c[\nu_1,\ldots,\nu_m]_{ij}:=\begin{cases}(f(\tilde\lambda_k)-f(\tilde\lambda_\ell))/(\tilde\lambda_k-\tilde\lambda_\ell) & \text{if } i\in\mathcal{I}_k,\ j\in\mathcal{I}_\ell \text{ for some } k\neq\ell,\\ \upsilon_k[\nu_k] & \text{if } i,j\in\mathcal{I}_k \text{ for some } k.\end{cases}$$

We now show that the linear mapping $V[\nu_1,\ldots,\nu_m]:\mathcal{S}\to\mathcal{S}$ defined by

$$V[\nu_1,\ldots,\nu_m]h:=p(c[\nu_1,\ldots,\nu_m]\circ(p^Thp))p^T\qquad\forall h\in\mathcal{S}$$

belongs to $\partial_B f^{\square}(z)$. For each $k\in\{1,\ldots,m\}$, since $\upsilon_k[\nu_k]\in\partial_B f(\tilde\lambda_k)$, there exist $\tilde\lambda_{kl}\in\mathbb{R}$, $l=1,2,\ldots$, such that f is differentiable at $\tilde\lambda_{kl}$ for all l and $\tilde\lambda_{kl}\to\tilde\lambda_k$ and $f'(\tilde\lambda_{kl})\to\upsilon_k[\nu_k]$ as $l\to\infty$. Then, letting

$$z_l:=p\,\mathrm{diag}[\lambda_{1l},\ldots,\lambda_{nl}]p^T\qquad\text{with}\qquad\lambda_{il}:=\tilde\lambda_{kl}\quad\forall i\in\mathcal{I}_k,\ k=1,\ldots,m,$$

for $l=1,2,\ldots$, we have from Proposition 4.3 that $f^{\square}$ is differentiable at z_l. Moreover, as $l\to\infty$, we have $z_l\to z$ and

$$\begin{aligned}|||\nabla f^{\square}(z_l)-V[\nu_1,\ldots,\nu_m]|||&=\sup_{\|h\|=1}\|\nabla f^{\square}(z_l)h-V[\nu_1,\ldots,\nu_m]h\|\\&=\sup_{\|h\|=1}\|(f^{[1]}(\lambda_{1l},\ldots,\lambda_{nl})-c[\nu_1,\ldots,\nu_m])\circ(p^Thp)\|\ \to\ 0.\end{aligned}$$

Hence $V[\nu_1,\ldots,\nu_m]\in\partial_B f(z)$. □

By using Lemma 5.1 together with Proposition 4.10, we can now establish the main result of this section. Part (a) of this result was already shown in [29]. Here we show that it also follows from Proposition 4.10.

PROPOSITION 5.2. *For the functions H and H_μ defined by* (33) *and* (36) *with $g\in\mathcal{CM}$, respectively, the following results hold.*

(a) *H is semismooth. If F is ρ-order semismooth $(0<\rho<\infty)$, then H is $\min\{1,\rho\}$-order semismooth.*

(b) *H_μ has the Jacobian Consistence Property relative to H.*

Proof. Let

$$f(\zeta):=\max\{0,\zeta\},\qquad f_\mu(\zeta):=\mu g(\zeta/\mu)\qquad\forall\zeta\in\mathbb{R}.\tag{49}$$

(a) It was shown in [30, Lem. 2.1] that

$$f^{\square}(z)=[z]_+\qquad\forall z\in\mathcal{S}.$$

Also, it is well known that f is piecewise linear on $\mathbb{R}$ and hence f is strongly semismooth. Then, by Proposition 4.10, $f^{\square}$ is strongly semismooth. It is known that the composition of two ρ-order semismooth functions is also ρ-order semismooth [10, Thm. 19]. Hence the composite function $(x,y)\mapsto f^{\square}(x-y)=[x-y]_+$ is strongly semismooth. Since F is semismooth, then H is semismooth. If F is ρ-order semismooth $(0<\rho<\infty)$, then H is $\min\{1,\rho\}$-order semismooth.

(b) It can be seen from (33), (35), and (36) that (37) is satisfied with $\kappa := \sqrt{n}g(0)$. Alternatively, this can be deduced by applying Lemma 5.1 and using (49). We now prove (38). It is readily seen from (49) and properties of g (see, e.g., [31]) that

$$\lim_{\mu\to 0^+} f'_\mu(\zeta) = \lim_{\mu\to 0^+} g'(\zeta/\mu) = \begin{cases} g'(0) & \text{if } \zeta = 0, \\ 1 & \text{if } \zeta > 0, \\ 0 & \text{if } \zeta < 0, \end{cases} \qquad \partial f(\zeta) = \begin{cases} [-1,1] & \text{if } \zeta = 0, \\ \{1\} & \text{if } \zeta > 0, \\ \{0\} & \text{if } \zeta < 0. \end{cases}$$

Since $g'(0) \in (0,1)$, this shows that (40) holds for all $\zeta \in \mathbb{R}$. Thus, by Lemma 5.1, (42) holds for all $z \in \mathcal{S}$. Fix any $x, y \in \mathcal{S}$. It can be seen from (33) and $f^{\square}(\cdot) = [\cdot]_+$ that

$$B \in \partial H(x,y) \quad \text{if and only if} \quad B = \begin{bmatrix} I - V & V \\ \nabla F(x) & -I \end{bmatrix} \text{ for some } V \in \partial f^{\square}(x-y).$$

Also, we have from (34) and (36) that

$$\nabla H_\mu(x,y) = \begin{bmatrix} I - \nabla f_\mu^{\square}(x-y) & \nabla f_\mu^{\square}(x-y) \\ \nabla F(x) & -I \end{bmatrix}.$$

Thus

$$\begin{aligned} \operatorname{dist}(\nabla H_\mu(x,y), \partial H(x,y)) &= \min_{V \in \partial f^{\square}(x-y)} \left\{ \max_{\|(u,v)\|=1} \|(\nabla f_\mu^{\square}(x-y) - V)(u-v)\| \right\} \\ &\le \sqrt{2}\, \operatorname{dist}(\nabla f_\mu^{\square}(x-y), \partial f^{\square}(x-y)) \\ &\to 0 \qquad \text{as} \quad \mu \to 0^+, \end{aligned}$$

where the last relation follows from (42) with $z = x - y$. This verifies (38). □

We note that, for the particular choice (49) of f and f_μ, we can obtain an explicit formula for c given by (46) and directly verify that V given by (48) belongs to $\partial f^{\square}(z)$. Specifically, for any $z \in \mathcal{S}$ and any $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ and $p \in \mathcal{O}$ satisfying $z = p\, \operatorname{diag}[\lambda_1, \ldots, \lambda_n] p^T$, define the three index sets

$$\alpha := \{i |\ \lambda_i > 0\}, \qquad \beta := \{i |\ \lambda_i = 0\}, \qquad \gamma := \{i |\ \lambda_i < 0\}.$$

Upon taking $\mu \to 0^+$ in (44) and using (49) and properties of g [31], we obtain in the limit that the (i,j)th entry of c is given by

$$c_{ij} = \lim_{\mu\to 0^+} (c_\mu)_{ij} = \begin{cases} 1 & \text{if } \ i,j \in \alpha, \\ 1 & \text{if } \ i \in \alpha, j \in \beta \ \text{ or } \ i \in \beta, j \in \alpha, \\ \lambda_i/(\lambda_i - \lambda_j) & \text{if } \ i \in \alpha, j \in \gamma, \\ \lambda_j/(\lambda_j - \lambda_i) & \text{if } \ i \in \gamma, j \in \alpha, \\ g'(0) & \text{if } \ i,j \in \beta, \\ 0 & \text{else.} \end{cases} \tag{50}$$

To see that V given by (48) belongs to $\partial f^{\square}(z)$, let ϵ_l, $l = 1, 2, \ldots$, be any sequence of positive scalars converging to 0, and define for $\sigma = -1, 1$ and $l = 1, 2, \ldots$ the symmetric matrix

$$z_l[\sigma] := z + \sigma \epsilon_l p\, \operatorname{diag}[d_1, \ldots, d_n] p^T, \qquad \text{with} \quad d_i := \begin{cases} 1 & \text{if } i \in \beta, \\ 0 & \text{else.} \end{cases}$$

For each $\sigma \in \{-1, 1\}$, it can be seen that the eigenvalues of $z_l[\sigma]$ are $\lambda_{il}[\sigma] := \lambda_i + \sigma\epsilon_l d_i$, $i = 1, \ldots, n$, which are nonzero for all l sufficiently large. Thus, f is differentiable

at $\lambda_{il}[\sigma]$, $i = 1, \ldots, n$, for all l sufficiently large. Hence, by Proposition 4.3, $f^{\square}$ is differentiable at $z_l[\sigma]$ for all l sufficiently large and

$$\nabla f^{\square}(z_l[\sigma])h = p(c_l[\sigma] \circ (p^T hp))p^T \qquad \forall h \in \mathcal{S},$$

where $c_l[\sigma] := f^{[1]}(\lambda_{1l}[\sigma], \ldots, \lambda_{nl}[\sigma]) \in \mathcal{S}$. Using (49), it can be seen that, as $l \to \infty$, $z_l[\sigma] \to z$ and $c_l[\sigma]$ converges entrywise to $c[\sigma]$ whose (i,j)th entry is

$$(c[\sigma])_{ij} := \begin{cases} 1 & \text{if } \ i,j \in \alpha, \\ 1 & \text{if } \ i \in \alpha, j \in \beta \ \text{ or } \ i \in \beta, j \in \alpha, \\ \lambda_i/(\lambda_i - \lambda_j) & \text{if } \ i \in \alpha, j \in \gamma, \\ \lambda_j/(\lambda_j - \lambda_i) & \text{if } \ i \in \gamma, j \in \alpha, \\ \max\{0, \sigma\} & \text{if } \ i,j \in \beta, \\ 0 & \text{else.} \end{cases} \tag{51}$$

Hence $\nabla f^{\square}(z_l[\sigma])$ converges in operator norm to $V[\sigma] : \mathcal{S} \to \mathcal{S}$ defined by

$$V[\sigma]h := p(c[\sigma] \circ (p^T hp))p^T \qquad \forall h \in \mathcal{S}.$$

By the definition of $\partial_B f^{\square}(z)$, we see that $V[\sigma] \in \partial_B f^{\square}(z)$. Moreover, (50) and (51) show that $c = g'(0)c[-1] + (1 - g'(0))c[1]$, and hence $V = g'(0)V[-1] + (1 - g'(0))V[1]$. This shows that $V \in \partial f^{\square}(z)$.

6. Final remarks. In this paper, we studied various continuity and differentiability properties of a class of symmetric-matrix-valued functions, which are natural extensions of real-valued functions to matrix-valued functions. Using these properties, we reformulated the SDCP as a semismooth equation based on the matrix projection operator $[\cdot]_+$. We verified the Jacobian Consistence Property for the reformulated semismooth equation and its smooth approximation based on a class of smoothing functions proposed by Chen and Mangasarian [4, 5] for the LP and NCP and extended in [8] to the SDCP. This result facilitates the extension of the smoothing method studied in [6] and [16] for the NCP to the SDCP. We stress that, apart from the Jacobian Consistence Property, there are other important issues in extending the smoothing method of [6] to the SDCP. One of them is the solvability of the smoothing Newton equations. We leave this issue for future research.

REFERENCES

[1] R. Bhatia, *Matrix Analysis*, Springer-Verlag, New York, 1997.

[2] B. Chen and X. Chen, *A global and local superlinear continuation-smoothing method for P_0 and R_0 NCP or monotone NCP*, SIAM J. Optim., 9 (1999), pp. 624–645.

[3] B. Chen and P.T. Harker, *Smoothing approximations to nonlinear complementarity problems*, SIAM J. Optim., 7 (1997), pp. 403–420.

[4] C. Chen and O.L. Mangasarian, *Smoothing methods for convex inequalities and linear complementarity problems*, Math. Programming, 71 (1995), pp. 51–69.

[5] C. Chen and O.L. Mangasarian, *A class of smoothing functions for nonlinear and mixed complementarity problems*, Comput. Optim. Appl., 5 (1996), pp. 97–138.

[6] X. Chen, L. Qi, and D. Sun, *Global and superlinear convergence of the smoothing Newton method and its application to general box constrained variational inequalities*, Math. Comp., 67 (1998), pp. 519–540.

[7] X. Chen and Y. Ye, *On homotopy-smoothing methods for box-constrained variational inequalities*, SIAM J. Control Optim., 37 (1999), pp. 589–616.

[8] X. Chen and P. Tseng, *Non-interior continuation methods for solving semidefinite complementarity problems*, Math. Programming, to appear.

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

[10] A. Fischer, *Solution of monotone complementarity problems with locally Lipschitzian functions*, Math. Programming, 76 (1997), pp. 513–532.
[11] M. Fukushima and L. Qi, eds., *Reformulation–Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods*, Kluwer Academic Publishers, Boston, 1999.
[12] R.A. Horn and C.R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.
[13] R.A. Horn and C.R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.
[14] H. Jiang and D. Ralph, *Global and local superlinear convergence analysis of Newton-type methods for semismooth equations with smooth least squares*, in Reformulation—Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods, M. Fukushima and L. Qi, eds., Kluwer Academic Publishers, Boston, 1999, pp. 181–209.
[15] C. Kanzow and C. Nagel, *Semidefinite programs: New search directions, smoothing-type methods, and numerical results*, SIAM J. Optim., 13 (2002), pp. 1–23.
[16] C. Kanzow and H. Pieper, *Jacobian smoothing methods for nonlinear complementarity problems*, SIAM J. Optim., 9 (1999), pp. 342–373.
[17] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1984.
[18] A.S. Lewis and M.L. Overton, *Eigenvalue optimization*, Acta Numer., 5 (1996), pp. 149–190.
[19] A.S. Lewis and H.S. Sendov, *Twice differentiable spectral functions*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 368–386.
[20] R. Mifflin, *Semismooth and semiconvex functions in constrained optimization*, SIAM J. Control Optim., 15 (1977), pp. 959–972.
[21] L. Qi, *Convergence analysis of some algorithms for solving nonsmooth equations*, Math. Oper. Res., 18 (1993), pp. 227–244.
[22] H.-D. Qi, *A regularized smoothing Newton method for box constrained variational inequality problems with P_0-functions*, SIAM J. Optim., 10 (1999), pp. 315–330.
[23] L. Qi and J. Sun, *A nonsmooth version of Newton's method*, Math. Programming, 58 (1993), pp. 353–367.
[24] L. Qi, D. Sun, and G. Zhou, *A new look at smoothing Newton methods for nonlinear complementarity problems and box constrained variational inequalities*, Math. Programming, 87 (2000), pp. 1–35.
[25] F. Rellich, *Perturbation Theory of Eigenvalue Problems*, Gordon and Breach, New York, 1969.
[26] R.T. Rockafellar and R.J.-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
[27] M. Shida and S. Shindoh, *Monotone Semidefinite Complementarity Problems*, Research Report 312, Department of Mathematical and Computer Sciences, Tokyo Institute of Technology, Tokyo, 1996.
[28] G.W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, New York, 1990.
[29] D. Sun and J. Sun, *Semismooth matrix valued functions*, Math. Oper. Res., 27 (2002), pp. 150–169.
[30] P. Tseng, *Merit functions for semi-definite complementarity problems*, Math. Programming, 83 (1998), pp. 159–185.
[31] P. Tseng, *Analysis of a non-interior continuation method based on Chen-Mangasarian smoothing functions for complementarity problems*, in Reformulation—Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods, M. Fukushima and L. Qi, eds., Kluwer Academic Publishers, Boston, 1999, pp. 381–404.
[32] N. Yamashita and M. Fukushima, *A new merit function and a descent method for semidefinite complementarity problems*, in Reformulation—Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods, M. Fukushima and L. Qi, eds., Kluwer Academic Publishers, Boston, 1999, pp. 405–420.

SIAM J. OPTIM.
Vol. 13, No. 4, pp. 986–1003

EPICONVERGENCE OF CONVEXLY COMPOSITE FUNCTIONS IN BANACH SPACES*

CHRISTOPHE COMBARI† AND LIONEL THIBAULT†

Abstract. This paper gives conditions ensuring the epiconvergence or Γ-convergence of sequences of convexly composite functions via the study of Painlevé–Kuratowski convergence of inverse images of sets. The convergence of multipliers of optimization problems associated with convexly composite functions is also established.

Key words. epiconvergence, Painlevé–Kuratowski convergence, qualification condition, differentiable-like mapping, convexly composite function, optimization problem, multiplier

AMS subject classifications. 49J52, 58C20

PII. S1052623499359804

1. Introduction.

One of the main topics in the theory of epiconvergence or Γ-convergence of functions is concerned with the characterization (by such convergences) of the Painlevé–Kuratowski convergence of the graphs of subdifferentials of convex or nonconvex functions. Such a characterization is recognized to be crucial for a good behavior of optimization problems with respect to parameters. Generally, it essentially holds for classes of functions satisfying some stability properties. It is known that it is the case for the class of convex functions defined on Banach spaces (see [1]) and the class of primal lower nice functions defined on Hilbert spaces (see [11], [10]). We also refer to [17] for some other classes of locally Lipschitzian functions.

A prototype of primal lower nice functions is the composition of an extended real-valued convex function f defined on a Hilbert space Y with a twice continuously differentiable mapping g from another Hilbert space X into Y satisfying a qualification condition (see [11], [16]). Such a qualification condition cannot be avoided in any optimization study about convexly composite functions. Indeed, any lower semicontinuous function defined on a Hilbert space (in fact, on some more general Banach spaces) can be represented locally as the composition of a lower semicontinuous convex function and a smooth mapping (see [12]). For a convexly composite function $f \circ g$, the qualification condition that appears to ensure that $f \circ g$ is primal lower nice near a point $\bar{x}$ such that $f \circ g(\bar{x})$ is finite is the following Robinson qualification condition:

$$\mathbb{R}_+(\operatorname{dom} f - g(\bar{x})) - \operatorname{Im} \nabla g(\bar{x}) = Y. \tag{1}$$

Here $\mathbb{R}_+$ denotes the set of nonnegative real numbers, $\nabla g(\bar{x})$ is the derivative of g at $\bar{x}$, and $\operatorname{Im}\nabla g(\bar{x}) = \nabla g(\bar{x})(X)$. When one deals with a sequence of convexly composite functions $(f_n \circ g_n)_n$, one naturally needs to consider a uniform qualification condition. To state such a uniform condition, we use another condition equivalent to (1) requiring that there exist $r > 0$ and $s > 0$ such that

$$s\mathbb{B}_Y \subset (\{f \leq r + f(g(\bar{x}))\} - g(\bar{x})) - \nabla g(\bar{x})(r\,\mathbb{B}_X), \tag{2}$$

*Received by the editors July 30, 1999; accepted for publication (in revised form) September 3, 2002; published electronically March 5, 2003.
http://www.siam.org/journals/siopt/13-4/35980.html
†Université Montpellier II, Département de Mathématiques, CC 051, Place Eugène Bataillon, 34095 Montpellier cedex 5, France (combari@math.univ-montp2.fr, thibault@math.univ-montp2.fr).

where $\mathbb{B}_Y$ denotes the closed unit ball of Y centered at the origin of Y and $\{f \leq \rho\} := \{y \in Y : f(y) \leq \rho\}$. This condition (2) can be adapted to a sequence of convexly composite functions $(f_n \circ g_n)_n$ in the following form, supposing that there exist $r > 0$ and $s > 0$ such that for all $n \in \mathbb{N}$

$$s\mathbb{B}_Y \subset (\{f_n \leq r + f(g(\bar{x}))\} - g(\bar{x})) - \nabla g(\bar{x})(r\mathbb{B}_X). \tag{3}$$

In this paper, we show that condition (3) (in addition to some other natural assumptions) ensures that the sequence $(f_n \circ g_n)_n$ epiconverges to $f \circ g$ whenever $(f_n)_n$ epiconverges to f and $(g_n)_n$ converges to g in some sense. The epiconvergence is established in Theorem 5.2 for general Banach spaces X and Y, and, as usual, one can derive from it results on the behavior of optimal value functions of optimization problems associated with convexly composite functions. In this paper, we use the epiconvergence result to study the convergence of Kuhn–Tucker multipliers for such problems. We must also point out that Theorem 5.2 is also used in [4] to show the Painlevé–Kuratowski convergence of the graphs of subdifferentials when the underlying functions are convexly composite and defined on Banach spaces. Such results can also be applied to evolution equations governed by subdifferentials of convexly composite functions in the line of [8] and [9].

Our method is based on general estimation results (similar to metric regularity inequalities) for the distance to inverse images of sequences of set-valued mappings depending on a parameter. The necessity of dealing with a parameter in such estimations appears in the proof of Theorem 5.2 and is also explained in the beginning of section 3. The method is general enough to allow us to also obtain the Painlevé–Kuratowski convergence of inverse images of sets and the Attouch–Wets convergence of the sequence $(f_n \circ g_n)_n$ under appropriate assumptions.

2. Preliminaries. Throughout the paper, X and Y will be two real Banach spaces. For a sequence $(f_n)_n$ of functions from X into $\mathbb{R} \cup \{+\infty\}$, recall that the epilimits (or Γ-limits) inferior or superior are defined (see [1], [5], and [15]) by

$$(Lif_n)(x) := \inf\{\liminf f_n(x_n) : x_n \to x\}$$

and

$$(Lsf_n)(x) := \inf\{\limsup f_n(x_n) : x_n \to x\}.$$

One says that the sequence $(f_n)_n$ epiconverges (or Γ-converges) to f around a point $\bar{x}$ when there exists a neighborhood V of $\bar{x}$ such that for all $x \in V$

$$(Lsf_n)(x) = f(x) = (Lif_n)(x), \tag{4}$$

and one writes $f_n \xrightarrow{\text{epi}} f$ over V. When $V = X$, one says that the sequence epiconverges (or Γ-converges) to f.

The term epiconvergence comes from the fact that this can be characterized by the Painlevé–Kuratowski convergence of the epigraphs. For a sequence $(C_n)_n$ of subsets of X, recall that the limits inferior and superior are given (see [1], [3], and [15]) by

$$LiC_n := \{x \in C : x = \lim x_n \quad \text{with } x_n \in C_n\}$$

and

$$LsC_n := \{x \in C : x = \lim x_{s(n)} \quad \text{with } x_{s(n)} \in C_{s(n)}\}.$$

Here $(s(n))_n$ denotes an increasing sequence of natural numbers.

One says that $(C_n)_n$ Painlevé–Kuratowski converges to C if

$$LsC_n = C = LiC_n.$$

So in terms of epigraphs $\operatorname{Epi} f_n := \{(x,r) \in X \times \mathbb{R} : f_n(x) \le r\}$ (see [1], [15]) one has

$$Ls(\operatorname{Epi} f_n) = \operatorname{Epi}(Lif_n) \quad \text{and} \quad Li(\operatorname{Epi} f_n) = \operatorname{Epi}(Lsf_n),$$

and hence $(f_n)_n$ epiconverges to f if and only if $(\operatorname{Epi} f_n)_n$ Painlevé–Kuratowski converges to $\operatorname{Epi} f$.

Another important convergence for sets and functions is the Attouch–Wets (or bounded-Hausdorff) convergence. A sequence $(C_n)_n$ of subsets of X Attouch–Wets converges (see [2] and [15]) to a subset C of X, provided that for any real number $\rho > 0$ and any $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that for all $n \ge N$

$$C_n \cap \rho\,\mathbb{B}_X \subset C + \varepsilon\,\mathbb{B}_X \quad \text{and} \quad C \cap \rho\,\mathbb{B}_X \subset C_n + \varepsilon\,\mathbb{B}_X;$$

here $\mathbb{B}_X$ denotes the closed unit ball in X centered at the origin of X. This is equivalent to saying that

$$e(C_n \cap \rho\,\mathbb{B}_X, C) \to 0 \quad \text{and} \quad e(C \cap \rho\,\mathbb{B}_X, C_n) \to 0,$$

where $e(C,D)$ denotes the excess of C over D; i.e., $e(C,D) = \sup_{c\in C} d(c,D)$. We will adopt the convention $d(a,\emptyset) = \infty$, and we will sometimes write $e_\rho(C,D)$ in place of $e(C \cap \rho\,\mathbb{B}_X, D)$.

When Epi f_n Attouch–Wets converges to Epi f, one says that the sequence of functions f_n Attouch–Wets converges to the function f, and one writes $f_n \overset{A.W}{\longrightarrow} f$.

Before ending this section, let us recall the following fixed point lemma of Dontchev and Hager [6]. The lemma will be fully used in the next section.

LEMMA 2.1. *Let Z be a complete metric space, let $T : Z \rightrightarrows Z$ be a set-valued mapping with closed values, and let $z_0 \in Z$ and r and λ be such that $0 \le \lambda < 1$, $d(z_0, T(z_0)) < r(1-\lambda)$, and T is pseudo-λ-contracting over $B(z_0,r)$ in the sense that*

$$e(T(z) \cap B(z_0,r), T(z')) \le \lambda d(z,z')$$

for all $z, z' \in B(z_0,r)$. Then T has a fixed point in $B(z_0,r)$; i.e., there exists $\bar z \in B(z_0,r)$ with $\bar z \in T(\bar z)$. Here $B(z_0,r)$ denotes the closed ball centered at z_0 with radius r.

3. Estimation results. In this section, we will establish some preparatory results that will be needed in the next sections.

Consider a sequence $(g_n)_n$ of mappings from X into Y and a sequence $(f_n)_n$ of functions from Y into $\mathbb{R} \cup \{+\infty\}$. As explained above, we are interested in the local study of epiconvergence of the sequence $(f_n \circ g_n)_n$, i.e., the epiconvergence of the sequence around a point $\bar x \in X$, when appropriate assumptions about convergences of (f_n) and (g_n) to f and g, respectively, are made. We know that the global epiconvergence of $(f_n \circ g_n)_n$ over all the space X is equivalent to the Painlevé–Kuratowski convergence of the sequences of sets $(\operatorname{Epi} f_n \circ g_n)_n$ in $X \times \mathbb{R}$. When one works with (4) around a point $\bar x$, one does not know that $f(x)$ remains near $f(\bar x)$ whenever x is near $\bar x$. Indeed, the function f is generally merely lower semicontinuous and noncontinuous. So we are naturally led to investigate for some fixed $\beta > 0$ the behavior of the sequence $(\operatorname{Epi} f_n \circ g_n)_n$ around each ball in $X \times \mathbb{R}$ with radius β and

centered at $(x,t) \in \operatorname{Epi} f \circ g$ (or centered at $(\bar{x},t)$ with $f \circ g(x) \leq t$) for any point x in some fixed appropriate neighborhood of $\bar{x}$ and such that $f \circ g(x) < \infty$. Thus, we have to consider such (x,t) as a parameter u and then study the behavior of the sequence $h_n^{-1}(\operatorname{Epi} f_n)$ around all balls $B(q(u),\beta)$, where $h_n : X \times \mathbb{R} \to Y \times \mathbb{R}$ with $h_n(x,r) = (g_n(x),r)$ and q is an appropriate mapping (see the proof of Theorem 5.2) from the set of parameters into $X \times \mathbb{R}$.

Therefore, we are first concerned with establishing some estimation results related to the inverse images of sequences of set-valued mappings also depending on a parameter. We begin with the following lemma dealing with such set-valued mappings.

LEMMA 3.1. *Let U be a nonempty set of parameters and, for each $u \in U$, let $(M_n(.,u))_n$ be a sequence of set-valued mappings with closed graphs from X into Y for which there exist two real numbers $l \geq 0$ and $\alpha > 0$ (both independent of u) such that for all $x \in \alpha\,\mathbb{B}_X$, $y \in \alpha\,\mathbb{B}_Y$, $n \in \mathbb{N}$, and $u \in U$*

$$d(x, M_n(.,u)^{-1}(y)) \leq l\,d(y, M_n(x,u)). \tag{5}$$

Let $\varepsilon \in]0, 1/l[$ and $L > l/(1-\varepsilon l)$. Let $(g_n)_n$ be any sequence of mappings from $X \times U$ into Y such that there exist $\delta > 0$ and $N \in \mathbb{N}$ (both independent of u) satisfying

$$\|g_n(x,u) - g_n(x',u)\| \leq \varepsilon \|x - x'\| \tag{6}$$

for all $x, x' \in \delta\mathbb{B}_X$, $n \geq N$, and $u \in U$. Assume that $g_n(0,u) \to 0$ uniformly with respect to $u \in U$, and consider the perturbed set-valued mappings $G_n : X \times U \rightrightarrows Y$ defined by $G_n(x,u) := g_n(x,u) + M_n(x,u)$. Then there exist an integer $N' \in \mathbb{N}$ and two positive real numbers β and δ' that depend only on α, δ, ε such that for all $x \in \delta'\,\mathbb{B}_X$, $y \in \beta\,\mathbb{B}_Y$, $n \geq N'$, and $u \in U$

$$d(x, G_n(.,u)^{-1}(y)) \leq L\; d(y, G_n(x,u) \cap \beta\,\mathbb{B}_Y). \tag{7}$$

Proof. The proof is an adaptation of techniques used in Dontchev and Hager [7]; see the proof of their Theorem 4.1.

Put

$$\delta' = \frac{1}{2}\min(\delta, \alpha, \alpha/3\varepsilon) \quad \text{and} \quad \beta = \min(\alpha/2, 2\delta'/(5L)).$$

Choose an integer $N' \geq N$ such that $\|g_n(0,u)\| \leq \alpha/6$ for all $n \geq N'$ and $u \in U$. Then for any $x \in 2\delta'\,\mathbb{B}_X$, any $y \in \beta\,\mathbb{B}_Y$, any $n \geq N'$, and any $u \in U$ one has according to (6)

$$\begin{aligned} \|y - g_n(x,u)\| &\leq \|y\| + \|g_n(0,u)\| + \|g_n(x,u) - g_n(0,u)\| \\ &\leq \tfrac{\alpha}{2} + \tfrac{\alpha}{6} + \varepsilon.\tfrac{\alpha}{3\varepsilon} = \alpha, \end{aligned} \tag{8}$$

and hence $y - g_n(x,u) \in \alpha\,\mathbb{B}_Y$. Fix $n \geq N'$, $x \in \delta'\mathbb{B}_X$, $y \in \beta\,\mathbb{B}_Y$, and $u \in U$ and consider the set-valued mapping $T_n : X \rightrightarrows X$ defined by

$$T_n(z) = M_n(.,u)^{-1}(y - g_n(z,u)) \quad \text{for all } z \in X.$$

As the graph of $M_n(.,u)$ is closed in $X \times Y$, the set $T_n(z)$ is closed in X for each $z \in X$. One may suppose $G_n(x,u) \cap \beta\,\mathbb{B}_Y \neq \emptyset$ because (7) obviously holds for any x such that $G_n(x,u) \cap \beta\,\mathbb{B}_Y = \emptyset$. Fix any $y' \in G_n(x,u) \cap \beta\,\mathbb{B}_Y$, and put $r = L\|y - y'\|$.

Let us show that the set-valued mapping T_n is pseudocontracting on the closed ball $B(x,r)$. Observe first that

$$r \leq L(\|y\| + \|y'\|) \leq 2L\beta \leq 4\delta'/5 < \delta'.$$

Then for $z' \in B(x,r)$ one has $\|z'\| \leq \|x\| + \|z'-x\| \leq \delta' + r \leq 2\delta'$, and hence it follows from (8) that $y - g_n(z',u) \in \alpha\,\mathbb{B}_Y$. So for $z, z' \in B(x,r)$ and $v \in M_n(.,u)^{-1}(y - g_n(z,u)) \cap B(x,r)$ we obtain $\|v\| \leq 2\delta' \leq \alpha$, $z, z' \in \delta\,\mathbb{B}_X$, and $y - g_n(z',u) \in \alpha\,\mathbb{B}_Y$, and hence according to (5) and (6) one has

$$\begin{array}{rcl} d(v, M_n(.,u)^{-1}(y - g_n(z',u))) & \leq & l\, d(y - g_n(z',u), M_n(v,u)) \\ & \leq & l\, \|(y - g_n(z',u)) - (y - g_n(z,u))\| \\ & \leq & l\, \varepsilon\, \|z - z'\|. \end{array}$$

Therefore for $z, z' \in B(x,r)$

$$\begin{array}{rcl} e(T_n(z) \cap B(x,r), T_n(z')) & = & e(M_n(.,u)^{-1}(y - g_n(z,u)) \cap B(x,r), \\ & & \quad M_n(.,u)^{-1}(y - g_n(z',u))) \\ & \leq & l\, \varepsilon\, \|z - z'\| \end{array}$$

and T_n is pseudocontracting on $B(x,r)$ because $l\,\varepsilon < 1$. Moreover, as $y' \in G_n(x,u)$ one has $y' - g_n(x,u) \in M_n(x,u)$ and, according to (5),

$$\begin{array}{rcl} d(x, T_n(x)) & = & d(x, M_n(.,u)^{-1}(y - g_n(x,u))) \\ & \leq & l\, d(y - g_n(x,u), M_n(x,u)) \\ & \leq & l\, \|(y - g_n(x,u)) - (y' - g_n(x,u))\| \\ & = & l\, \|y - y'\| \\ & < & L(1 - \varepsilon l)\|y' - y\| = r(1 - \varepsilon\, l). \end{array}$$

So one may apply Lemma 2.1 and get a fixed point $z \in T_n(z) \cap B(x,r)$. Therefore $y - g_n(z,u) \in M_n(z,u) = G_n(z,u) - g_n(z,u)$, and hence $z \in G_n(.,u)^{-1}(y)$ which ensures that

$$d(x, G_n(.,u)^{-1}(y)) \leq \|x - z\| \leq r = L\, \|y - y'\|.$$

Since this holds for all $y' \in G_n(x,u) \cap \beta\,\mathbb{B}_Y$, one obtains

$$d(x, G_n(.,u)^{-1}(y)) \leq L\, d(y, G_n(x,u) \cap \beta\,\mathbb{B}_Y),$$

and this completes the proof. □

We will need the following differentiability-like concept for a sequence of parametrized mappings in Proposition 3.1 and also in sections 4 and 5.

DEFINITION 3.1. *Consider a set of parameters U, a mapping $q : U \to X$, and a sequence $\{g_n : n \in \mathbb{N}\}$ of mappings from $X \times U$ into Y. We will say that this sequence is strictly differentiable-like at $(q(u))_{u\in U}$ if there exists a family of continuous linear mappings $(A_u)_{u\in U}$ from X into Y such that for each real number $\varepsilon > 0$ there exist $\delta > 0$ and $N \in \mathbb{N}$, both independent of $u \in U$, such that for all $u \in U$, $x, x' \in q(u) + \delta\,\mathbb{B}_X$, and $n \geq N$*

$$\|g_n(x,u) - g_n(x',u) - A_u(x - x')\| \leq \varepsilon \|x - x'\|.$$

When U is a singleton set, we will say that the sequence is strictly differentiable-like at the fixed point $q(u) = \bar{x}$.

If, in addition to the strict differentiable-like property, for some $u \in U$ the sequence $\{g_n(.,u) : n \in \mathbb{N}\}$ pointwise converges to $g(.,u)$ over some neighborhood of $q(u)$, then it is easily seen that the mapping $g(.,u)$ is strictly differentiable in the usual sense at the point $q(u)$ with $\nabla_1 g(q(u),u) = A_u$. Here $\nabla_1 g(x,u)$ denotes the derivative of $g(.,u)$ at the point x.

The following provides two typical examples of such sequences that are strictly differentiable-like.

First consider a point $\bar{x} \in X$ and a sequence $\{g_n : n \in \mathbb{N}\} \cup \{g\}$ of mappings from X into Y that are of class C^1 on a same neighborhood of $\bar{x}$. Assume that

$$\lim_{\substack{x \to \bar{x} \\ n \to \infty}} \nabla g_n(x) = \nabla g(\bar{x}). \tag{9}$$

Let $\varepsilon > 0$. Choose $\beta > 0$ and $N \in \mathbb{N}$ such that for all $n \in \mathbb{N}$, $x \in B(\bar{x}, 2\beta)$

$$\|\nabla g_n(x) - \nabla g(\bar{x})\| \leq \varepsilon. \tag{10}$$

Then for $x, x' \in B(\bar{x}, \beta)$ and $n \geq N$ we have

$$g_n(x) - g_n(x') - \nabla g(\bar{x})(x - x') = \int_0^1 (\nabla g_n(x' + t(x - x')) - \nabla g(\bar{x}))(x - x')dt,$$

and hence according to (10)

$$\|g_n(x) - g_n(x') - \nabla g(\bar{x})(x - x')\| \leq \varepsilon \|x - x'\|.$$

So the sequence $\{g_n : n \in \mathbb{N}\} \cup \{g\}$ is strictly differentiable-like at the fixed point $\bar{x}$.

Obviously, the assumption (9) holds whenever $(\nabla g_n)_n$ uniformly converges to ∇g on a neighborhood of $\bar{x}$. Also note that in this first example the point $q(u) = \bar{x}$ does not depend on the parameter u.

Now consider another example that we will need later in the paper. The strict differentiability-like concept has essentially been motivated by this example. Let $\{g_n : n \in \mathbb{N}\}$ be a sequence of mappings from X into Y that is strictly differentiable-like at a fixed point $\bar{x} \in X$. Let Λ be a continuous linear mapping from X into a Banach space Z, and let p be any mapping from U into Z. Define $\hat{g}_n : (X \times Z) \times U \to Y \times Z$ by putting

$$\hat{g}_n(x, z, u) = (g_n(x), \Lambda(z)),$$

and define $q : U \to X \times Z$ by setting $q(u) = (\bar{x}, p(u))$. Then for $\hat{A} : X \times Z \to Y \times Z$ with $\hat{A}(x,z) = (A(x), \Lambda(z))$ we have

$$\hat{g}_n(x, z, u) - \hat{g}_n(x', z', u) - \hat{A}(x - x', z - z') = (g_n(x) - g_n(x') - A(x - x'), 0),$$

and hence the sequence $\{\hat{g}_n : n \in \mathbb{N}\}$ is strictly differentiable-like at $(q(u))_{u \in U}$.

The following lemma is a direct consequence of the statement and the proof of Theorem 2 in Robinson [13] (see also [14]). We will use it in the proof of Proposition 3.1 below.

LEMMA 3.2. *Let X and Y be two real normed vector spaces and M be a set-valued mapping with a convex graph. Suppose that there exist two positive real numbers r and s such that*

$$s\,\mathbb{B}_Y \subset M(r\,\mathbb{B}_X).$$

Then for all $\rho > 0$, $t \in [0, s[$, $x \in \rho\,\mathbb{B}_X$, and $y \in t\,\mathbb{B}_Y$ one has

$$d(x, M^{-1}(y)) \leq \frac{r+\rho}{s-t} d(y, M(x)).$$

PROPOSITION 3.1. *Let $\{g_n : n \in \mathbb{N}\}$ be a sequence of mappings from $X \times U$ into Y, and let $(C_n)_n$ and $(D_n)_n$ be two sequences of closed convex subsets of X and Y, respectively. Assume that the sequence $\{g_n : n \in \mathbb{N}\}$ is strictly differentiable-like at $(q(u))_{u\in U}$ and that there exists $g : X \times U \to Y$ such that the sequence $(g_n(q(.), .))_n$ uniformly converges over U to $g(q(.), .)$. Also assume that there exist two positive numbers s and r such that for all $n \in \mathbb{N}$ and $u \in U$*

$$s\,\mathbb{B}_Y \subset A_u((C_n - q(u)) \cap r\,\mathbb{B}_X) - (D_n - g(q(u), u)),$$

where A_u is given by Definition 3.1. *Then for each $\rho > 0$ there exist an integer $N \in \mathbb{N}$ and two numbers δ and β that all depend only on r, ρ, and s such that*

$$d(x, C_n \cap g_n(., u)^{-1}(D_n)) \leq 3s^{-1}(r+\rho) d(g_n(x, u), D_n)$$

for all $u \in U$, $n \geq N$, and $x \in C_n \cap B(q(u), \delta)$ satisfying $g_n(x, u) \in D_n + \beta\,\mathbb{B}_Y$.

Proof. Put $M_n(x, u) = A_u(x) + g(q(u), u) - D_n$ if $x \in C_n - q(u)$ and $M_n(x, u) = \emptyset$ otherwise. Then $M_n(., u)$ is a set-valued mapping whose graph is closed and convex. As $s\,\mathbb{B}_Y \subset M_n(., u)(r\,\mathbb{B}_X)$, it follows from Lemma 3.2 that for any fixed $\rho > 0$

$$d(x, M_n^{-1}(., u)(y)) \leq 2s^{-1}(r+\rho) d(y, M_n(x, u))$$

for all $y \in \frac{s}{2}\mathbb{B}_Y$ and $x \in \rho\,\mathbb{B}_X$. Considering $\varepsilon := s/8(r+\rho)$ and $L := 3(r+\rho)/s$ and putting

$$h_n(x, u) := g_n(q(u) + x, u) - g(q(u), u) - A_u(x),$$

it is not difficult to see that $L > l/(1 - \varepsilon l)$ for $l := 2(r+\rho)/s$ and that there exist $N_1 \in \mathbb{N}$ and $\delta' > 0$ (both depending only on r, ρ, and s) such that

$$\|h_n(x, u) - h_n(x', u)\| \leq \varepsilon \|x - x'\|$$

for all $x, x' \in \delta'\,\mathbb{B}_X$, $n \geq N_1$, and $u \in U$. Further, $h_n(0, u) \to 0$ uniformly with respect to $u \in U$. Consider the set-valued mappings $G_n := h_n + M$ and observe that

$$G_n(x, u) = g_n(q(u) + x, u) - D_n \quad \text{if } x \in C_n - q(u),$$

$$G_n(x, u) = \emptyset \quad \text{otherwise.}$$

Then

$$\begin{aligned} G_n(., u)^{-1}(0) &= \{x \in X : (x + q(u)) \in C_n \cap g_n(., u)^{-1}(D_n)\} \\ &= C_n \cap g_n(., u)^{-1}(D_n) - q(u) \end{aligned}$$

and by Lemma 3.1 (with $\alpha = \min(\rho, s/2)$) there exist $N \in \mathbb{N}$, $\beta > 0$, and $\delta > 0$ all depending only on r, ρ, and s such that for all $z \in \delta\,\mathbb{B}_X$

$$d(z, G_n(., u)^{-1}(0)) \leq L\, d(0, G_n(z, u) \cap \beta\,\mathbb{B}_Y). \tag{11}$$

Fix $u \in U$ and $n \geq N$, and fix any $x \in C_n \cap B(q(u), \delta)$ with $g_n(x,u) \in D_n + \beta\,\mathbb{B}_Y$ (if any). Then $x - q(u) \in (C_n - q(u)) \cap \delta\,\mathbb{B}_X$ and for any $y \in D_n \cap (g_n(x,u) + \beta\,\mathbb{B}_Y)$ we have

$$g_n(x,u) - y \in G_n(x - q(u), u) \cap \beta\,\mathbb{B}_Y.$$

So choosing $z = x - q(u)$ in (11) we get

$$d(x, C_n \cap g_n(.,u)^{-1}(D_n)) \leq L\, d(0, g_n(x,u) - y),$$

and hence

$$d(x, C_n \cap g_n(.,u)^{-1}(D_n)) \leq L\, d(g_n(x,u), D_n \cap (g_n(x,u) + \beta\,\mathbb{B}_Y)).$$

This completes the proof because one obviously has

$$d(y, D_n \cap (y + \gamma\,\mathbb{B}_Y)) = d(y, D_n)$$

when $D_n \cap (y + \gamma\,\mathbb{B}_Y) \neq \emptyset$. □

4. Attouch–Wets convergence of convexly composite functions. With the estimations of section 3 in hand, we begin by proving the following theorem on Attouch–Wets convergence of inverse images of sets. The related convergence of convexly composite functions will be derived from this theorem.

We first need to state the following lemma of Robinson [13] (see Lemma 2 in [13]) that will be used in the proof of the theorem.

LEMMA 4.1. *Let M be a set-valued mapping with closed convex graph between X and Y. Suppose that $x_0 \in X$ and $y_0 \in Y$ are such that for some bounded convex set $S \subset Y$ and some real numbers $0 < r < s$*

$$y_0 + sS \subset M(x_0 + \mathbb{B}_X) + rS.$$

Then

$$y_0 + \ \operatorname{int}(s - r)S \subset \ \operatorname{int} M(x_0 + \mathbb{B}_X).$$

(Here int *denotes the topological interior.)*

The convergence theorem of inverse images is then the following.

THEOREM 4.1. *Let $g : X \to Y$ be strictly differentiable at a fixed point $\bar{x} \in X$, and let $g_n : X \to Y$ be such that the sequence $\{g_n : n \in \mathbb{N}\} \cup \{g\}$ is strictly differentiable-like at the fixed point $\bar{x}$. Let $\{C_n : n \in \mathbb{N}\} \cup \{C\}$ and $\{D_n : n \in \mathbb{N}\} \cup \{D\}$ be two sequences of nonempty closed convex subsets of X and Y, respectively. Assume that*

(i) *$(g_n)_n$ uniformly converges to g on a neighborhood of $\bar{x}$;*
(ii) *$(C_n)_n$ and $(D_n)_n$ Attouch–Wets converges to C and D, respectively;*
(iii) *$\bar{x} \in C \cap g^{-1}(D)$ and*

$$\mathbb{R}_+[\nabla g(\bar{x})(C - \bar{x}) - (D - g(\bar{x})] = Y.$$

Then

$$C_n \cap g_n^{-1}(D_n) \xrightarrow{A.W.} C \cap g^{-1}(D) \quad \text{around } \bar{x};$$

that is, there exists $\beta > 0$ such that for any $\varepsilon > 0$ there is $N \in \mathbb{N}$ with

$$C_n \cap g_n^{-1}(D_n) \cap B(\bar{x}, \beta) \subset C \cap g^{-1}(D) + \varepsilon\,\mathbb{B}_X \tag{12}$$

and

$$C \cap g^{-1}(D) \cap B(\bar{x}, \beta) \subset C_n \cap g_n^{-1}(D_n) + \varepsilon\, \mathbb{B}_X \tag{13}$$

for all $n \geq \mathbb{N}$.

Proof. We begin by showing the inclusion (13). First observe that, according to Theorem 1 in Robinson [13], there exist $r > 0$ and $s > 0$ such that

$$s\, \mathbb{B}_Y \subset \nabla g(\bar{x})((C - \bar{x}) \cap r\, \mathbb{B}_X) - (D - g(\bar{x})) \cap r\, \mathbb{B}_Y. \tag{14}$$

Fix a real number $\varepsilon > 0$ such that $\varepsilon' := \varepsilon(1 + \|\nabla g(\bar{x})\|) < s$, and put $r' := r + \varepsilon$. By the Attouch–Wets convergence of (C_n) and (D_n) there exists an integer n_0 such that for all $n \geq n_0$

$$(C - \bar{x}) \cap r\mathbb{B}_X \subset (C_n - \bar{x}) + \varepsilon \mathbb{B}_X$$

and

$$(D - g(\bar{x})) \cap r\mathbb{B}_Y \subset (D_n - g(\bar{x})) + \varepsilon B_Y,$$

and hence

$$(C - \bar{x}) \cap r\mathbb{B}_X \subset (C_n - \bar{x}) \cap r'\mathbb{B}_X + \varepsilon \mathbb{B}_X$$

and

$$(D - g(\bar{x})) \cap r\mathbb{B}_Y \subset (D_n - g(\bar{x})) \cap r'\mathbb{B}_Y + \varepsilon B_Y.$$

Using (14) we obtain for all $n \geq n_0$

$$s\mathbb{B}_Y \subset \nabla g(\bar{x})((C_n - \bar{x}) \cap r'\mathbb{B}_X) - (D_n - g(\bar{x})) \cap r'\mathbb{B}_Y + \varepsilon'\mathbb{B}_Y.$$

So applying Lemma 4.1 with $x_0 = (0, 0)$ and $y_0 = 0$ and with the set-valued mapping M_n between $X \times Y$ and Y given by

$$M_n(x, y) = \nabla g(\bar{x})(r'x) - r'y \quad \text{if} \quad (x, y) \in \frac{1}{r'}(C_n - \bar{x}) \times \frac{1}{r'}(D_n - g(\bar{x}))$$

$$\text{and} \quad M_n(x, y) = \emptyset \quad \text{otherwise,}$$

and fixing a real number s' with $0 < s' < s - \varepsilon'$ we get for all $n \geq n_0$

$$s'\mathbb{B}_Y \subset \nabla g(\bar{x})((C_n - \bar{x}) \cap r'\mathbb{B}_X) - (D_n - g(\bar{x})) \cap r'\mathbb{B}_Y.$$

According to Proposition 3.1, there exist an integer $n_1 \geq n_0$ and positive numbers l, δ, and β' (independent of n) such that

$$d(x, C_n \cap g_n^{-1}(D_n)) \leq l\, d(g_n(x), D_n) \tag{15}$$

for all $n \geq n_1$ and $x \in C_n \cap B(\bar{x}, \delta)$ satisfying $g_n(x) \in D_n + \beta'\mathbb{B}_Y$. Choose a positive number $\beta < \min(\beta', \delta)$ and some $\rho \geq \beta + \|\bar{x}\|$ such that $\|g(x)\| \leq \rho$ for any $x \in B(\bar{x}, \beta)$, $(g_n)_n$ uniformly converges on $B(\bar{x}, \beta)$, and

$$\|g_n(x) - g_n(x') - \nabla g(\bar{x})(x - x')\| \leq \|x - x'\|$$

for all $x, x' \in B(\bar{x}, 2\beta)$ and $n \geq n_2$ with $n_2 \geq n_1$.

Fix any real number $\lambda > 1$ and consider any $x \in g^{-1}(D) \cap C \cap B(\bar{x}, \beta)$. Then $x \in C$ and $g(x) \in D \cap \rho\,\mathbb{B}_Y$. Putting

$$\|g_n - g\|_\beta := \sup\{||g_n(x) - g(x)|| : x \in B(x, \beta)\},$$

we obtain

$$g_n(x) \in D \cap \rho\,\mathbb{B}_Y + \|g_n - g\|_\beta\,\mathbb{B}_Y \subset D_n + (\lambda\, e_\rho(D, D_n) + \|g_n - g\|_\beta)\,\mathbb{B}_Y.$$

In the same way, as $x \in C \cap \rho\,\mathbb{B}_X$, we can choose $x'_n \in C_n$ with $\|x - x'_n\| \leq \lambda\, e_\rho(C, C_n)$. Then for n large enough

$$\|g_n(x'_n) - g_n(x)\| \leq (1 + \|\nabla g(\bar{x})\|)\|x - x'_n\| \leq \lambda(1 + \|\nabla g(\bar{x})\|)e_\rho(C, C_n),$$

and hence

$$\begin{aligned} g_n(x'_n) \in D_n &+ (\|g_n - g\|_\beta + \lambda e_\rho(D, D_n) \\ &+ \lambda(1 + \|\nabla g(\bar{x})\|)e_\rho(C, C_n))\,\mathbb{B}_Y. \end{aligned} \tag{16}$$

As $\|x'_n - \bar{x}\| \leq \|x - \bar{x}\| + \|x'_n - x\| \leq \beta + \|x'_n - x\|$ and $\|x'_n - x\| \to 0$, we can choose $N \geq n_2$ such that, for all $n \geq N$, we have $\|x'_n - \bar{x}\| < \delta$ and

$$\|g_n - g\|_\beta + \lambda\, e_\rho(D, D_n) + \lambda(1 + \|\nabla g(\bar{x})\|)\; e_\rho(C, C_n) \leq \beta'.$$

Applying (15) we obtain, for any $n \geq N$, some $x_n \in C_n \cap g_n^{-1}(D_n)$ such that

$$\|x_n - x'_n\| \leq \lambda\, l\, d\,(g_n(x'_n), D_n).$$

According to (16) and the inequality $\|x - x'_n\| \leq \lambda\, e_\rho(C, C_n)$ and putting

$$t_n := \lambda(l\|g_n - g\|_\beta + \lambda\, l\, e_\rho(D, D_n) + (1 + \lambda\, l + \lambda\, l\|\nabla g(\bar{x})\|)\; e_\rho(C, C_n)),$$

we get $\|x - x_n\| \leq t_n$, and hence

$$C \cap g^{-1}(D) \cap B(\bar{x}, \beta) \subset C_n \cap g_n^{-1}(D_n) + t_n \mathbb{B}_X. \tag{17}$$

The inclusion (13) easily follows from (17).

The proof of the inclusion (12) is similar. □

As mentioned before, the following result on the Attouch–Wets convergence of convexly composite functions comes from Theorem 4.1.

THEOREM 4.2. *Let $g : X \to Y$ be strictly differentiable at a fixed point $\bar{x} \in X$, and let $g_n : X \to Y$ be such that the sequence $\{g_n : n \in \mathbb{N}\} \cup \{g\}$ is strictly differentiable-like at the fixed point $\bar{x}$. Let $f_n : Y \to \mathbb{R} \cup \{\infty\}$ be a sequence of proper lower semicontinuous convex functions that converges in the sense of Attouch–Wets to f. Assume that $f(g(\bar{x}))$ is finite, $(g_n)_n$ uniformly converges to g on a neighborhood of $\bar{x}$, and*

$$\mathbb{R}_+(\operatorname{dom} f - g(\bar{x})) - \operatorname{Im} \nabla g(\bar{x}) = Y.$$

Then

$$f_n \circ g_n \xrightarrow{A.W.} f \circ g \quad \text{around } \bar{x};$$

that is, there exists $\beta > 0$ such that for any $\varepsilon > 0$ there is $N \in \mathbb{N}$ with

$$(\operatorname{Epi} f_n \circ g_n) \cap B((\bar{x}, f \circ g(\bar{x})), \beta) \subset \operatorname{Epi} f \circ g + \varepsilon \mathbb{B}_{X \times \mathbb{R}}$$

and

$$(\operatorname{Epi} f \circ g) \cap B((\bar{x}, f \circ g(\bar{x})), \beta) \subset \operatorname{Epi} f_n \circ g_n + \varepsilon \mathbb{B}_{X \times \mathbb{R}}$$

for all $n \geq N$.

Proof. Considering $\hat{g}, \hat{g}_n : X \times \mathbb{R} \to Y \times \mathbb{R}$ with $\hat{g}(x,t) := (g(x), t)$, $\hat{g}_n(x,t) := (g_n(x), t)$ and putting $\bar{t} := f \circ g(\bar{x})$, it is not difficult to see that

$$\mathbb{R}_+[(\operatorname{Epi} f - \hat{g}(\bar{x}, \bar{t})) - \operatorname{Im} \nabla \hat{g}(\bar{x}, \bar{t})] = X \times Y.$$

Furthermore, the sequence $\{\hat{g}_n : n \in \mathbb{N}\} \cup \{\hat{g}\}$ is obviously strictly differentiable-like at $(\bar{x}, \bar{t})$ and $(\hat{g}_n)_n$ uniformly converges to $\hat{g}$ on a neighborhood of $(\bar{x}, \bar{t})$. So observing that

$$\hat{g}_n^{-1}(\operatorname{Epi} f_n) = \operatorname{Epi} f_n \circ g_n \quad \text{and} \quad \hat{g}^{-1}(\operatorname{Epi} f) = \operatorname{Epi} f \circ g$$

we obtain that the result follows from Theorem 4.1 with $C_n = C = X \times \mathbb{R}$, $D_n = \operatorname{Epi} f_n$, and $D = \operatorname{Epi} f$. ∎

According to our first example of strictly differentiable-like mappings, Theorem 4.2 admits the following corollary.

COROLLARY 4.1. *Let $\{g_n : n \in \mathbb{N}\} \cup \{g\}$ be a sequence of mappings from X into Y that are of class C^1 on a same neighborhood of $\bar{x} \in X$, and let $f_n : Y \to \mathbb{R} \cup \{+\infty\}$ be a sequence of proper lower semicontinuous convex functions that converges in the sense of Attouch–Wets to f. Assume that $f(g(\bar{x}))$ is finite and*

(i) *$(g_n)_n$ uniformly converges to g on a neighborhood of $\bar{x}$;*

(ii) *$\lim_{\substack{x \to \bar{x} \\ n \to \infty}} \nabla g_n(x) = \nabla g(\bar{x})$;*

(iii) *$\mathbb{R}_+(\operatorname{dom} f - g(\bar{x}) - \operatorname{Im} \nabla g(\bar{x})) = Y$.*

Then

$$f_n \circ g_n \xrightarrow{A.W.} f \circ g \quad \text{around } \bar{x}.$$

5. Epiconvergence of convexly composite functions. In this section, we will establish sufficient conditions ensuring the epiconvergence or Γ-convergence of convexly composite functions. These conditions are applied in [4] to study the Painlevé–Kuratowski convergence of the graphs of subdifferentials of convexly composite functions. First, we prove the following lemma.

LEMMA 5.1. *Let q be a mapping from X into Y, let $\{g_n : n \in \mathbb{N}\} \cup \{g\}$ be a sequence of mappings from X into Y that is strictly differentiable-like at $(q(u))_{u \in U}$, and let $(C_n)_n$ and $(D_n)_n$ be two sequences of closed convex subsets of X and Y, respectively. Assume that there exists some $\alpha > 0$ such that for each $u \in U$ the sequence $(g_n(.,u))_n$ pointwise converges to $g(.,u)$ over $B(q(u), \alpha)$, and assume $g_n(q(u),u) - g(q(u),u) \to 0$ uniformly with respect to $u \in U$. Assume also that there exist $r > 0$ and $s > 0$ such that for all $n \in \mathbb{N}$ and $u \in U$*

$$s \mathbb{B}_Y \subset \nabla_1 g(q(u),u)((C_n - q(u)) \cap r \mathbb{B}_X) - (D_n - g(q(u),u)).$$

Then there exists $\beta > 0$ independent of u such that for each $u \in U$

$$\begin{aligned}&[(Li C_n) \cap g(.,u)^{-1}(Li D_n)] \cap B(q(u), \beta) \\ &\subset [Li(C_n \cap g_n(.,u)^{-1}(D_n))] \cap B(q(u), \beta)\end{aligned}$$

and

$$[Ls(C_n \cap g_n(.,u)^{-1}(D_n))] \cap B(q(u),\beta) \\ \subset [(LsC_n) \cap g(.,u)^{-1}(LsD_n)] \cap B(q(u),\beta).$$

Proof. According to Proposition 3.1, there exist $N \in \mathbb{N}$ and positive real numbers l, δ, and $\beta < \alpha/2$ (independent of u) such that

$$d(x, C_n \cap g_n^{-1}(.,u)(D_n)) \leq l\, d(g_n(x,u), D_n) \tag{18}$$

for all $u \in U$, $n \geq N$, and $x \in C_n \cap B(q(u), 2\beta)$ satisfying $g_n(x,u) \in D_n + 2\beta\,\mathbb{B}_Y$. Without loss of generality, we may suppose that for all $u \in U$, x, $x' \in B(q(u), 2\beta)$, and $n \geq N$

$$\|g_n(x,u) - g_n(x',u) - \nabla_1 g(q(u),u)(x-x')\| \leq \|x - x'\|.$$

So according to the pointwise convergence of $(g_n(.,u))_n$ to $g(.,u)$ over $B(q(u),2\beta)$ we see that for each $u \in U$ and each $x \in B(q(u),2\beta)$ we have $g_n(x_n,u) \to g(x,u)$ for every $x_n \to x$. Fix $u \in U$, and consider any $x \in B(q(u),\beta)$ with $x \in (LiC_n) \cap g(.,u)^{-1}(LiD_n)$. Then there exist $x_n \in C_n$ and $y_n \in D_n$ with $x_n \to x$ and $y_n \to g(x,u)$.

First note that for n large enough $x_n \in C_n \cap B(q(u),2\beta)$ because $x \in B(q(u),\beta)$ and $x_n \to x$. As $g_n(x_n,u) \to g(x,u)$ we have $g_n(x_n,u) - y_n \to 0$, and hence

$$g_n(x_n,u) \in y_n + \|g_n(x_n,u) - y_n\|\mathbb{B}_Y \subset D_n + 2\beta\mathbb{B}_Y$$

for n large enough. We can apply (18) to get $x_n' \in C_n \cap g_n(.,u)^{-1}(D_n)$ such that

$$\|x_n - x_n'\| \leq 2ld(g_n(x_n,u), D_n) \leq 2l\|g_n(x_n,u) - y_n\|.$$

So $\|x_n - x_n'\| \to 0$, which ensures that $x_n' \to x$, and hence $x \in Li(C_n \cap g_n(.,u)^{-1}(D_n))$. This proves that

$$[(Li\, C_n) \cap g(.,u)^{-1}(Li\, D_n)] \cap B(q(u),\beta) \\ \subset [Li\,(C_n \cap g_n(.,u)^{-1}(D_n))] \cap B(q(u),\beta).$$

Let us consider the case of limit superior. Fix any

$$x \in [Ls(C_n \cap g_n(.,u)^{-1}(D_n)] \cap B(q(u),\beta).$$

By definition, there exists $x_{s(n)} \in C_{s(n)} \cap g_{s(n)}(.,u)^{-1}(D_{s(n)})$ with $x_{s(n)} \to x$. As $\|x_{s(n)}\| \leq 2\beta$ for n large enough (because $\|x\| \leq \beta$), we have by what precedes that $g_{s(n)}(x_{s(n)},u) \to g(x)$. Since $x_{s(n)} \in C_{s(n)}$ and $g_{s(n)}(x_{s(n)},u) \in D_{s(n)}$, we obtain $x \in Ls\, C_n$ and $g(x,u) \in Ls\, D_n$. So

$$x \in [(Ls\, C_n) \cap g(.,u)^{-1}(Ls\, D_n)] \cap B(q(u),\beta),$$

and hence

$$[Ls(C_n \cap g_n(.,u)^{-1}(D_n))] \cap B(q(u),\beta) \\ \subset [(Ls\, C_n) \cap g(.,u)^{-1}(Ls\, D_n)] \cap B(q(u),\beta).$$

This completes the proof. □

We can derive two important theorems from the lemma above. The first one deals with the case when $q(u)$ is independent of u, and its proof follows directly from the lemma.

THEOREM 5.1. *Let $\{g_n : n \in \mathbb{N}\} \cup \{g\}$ be a sequence of mappings from X into Y that is strictly differentiable-like at a fixed point $\bar{x} \in X$, and let $(C_n)_n$ and $(D_n)_n$ be two sequences of closed convex subsets of X and Y, respectively. Assume that $(g_n)_n$ pointwise converges to g on a neighborhood of $\bar{x}$ and that there exist $r > 0$ and $s > 0$ such that for all $n \in \mathbb{N}$*

$$s\,\mathbb{B}_Y \subset \nabla g(\bar{x})((C_n - \bar{x}) \cap r\,\mathbb{B}_X) - (D_n - g(\bar{x})).$$

Then there exists $\beta > 0$ such that

$$[(Li\,C_n) \cap g^{-1}(Li\,D_n)] \cap B(\bar{x}, \beta) \subset [Li(C_n \cap g_n^{-1}(D_n))] \cap B(\bar{x}, \beta)$$

and

$$[Ls(C_n \cap g_n^{-1}(D_n))] \cap B(\bar{x}, \beta) \subset [(Ls\,C_n) \cap g^{-1}(Ls\,D_n)] \cap B(\bar{x}, \beta).$$

The following convergence result is a direct corollary of Theorem 5.1.

COROLLARY 5.1. *Suppose in addition to the assumptions of Theorem 5.1 that $(C_n)_n$ and $(D_n)_n$ Painlevé–Kuratowski converge to C and D, respectively. Then there exists $\beta > 0$ such that $(C_n \cap g_n^{-1}(D_n))_n$ Painlevé–Kuratowski converges to $C \cap g^{-1}(D)$ over $B(\bar{x}, \beta)$ in the sense that*

$$\begin{aligned}[Li(C_n \cap g_n^{-1}(D_n))] \cap B(\bar{x}, \beta) &= [C \cap g^{-1}(D)] \cap B(\bar{x}, \beta) \\ &= [Ls(C_n \cap g_n^{-1}(D_n))] \cap B(\bar{x}, \beta).\end{aligned}$$

The second theorem is concerned with the epiconvergence of convexly composite functions. Its proof uses Lemma 5.1 in its parametric form.

THEOREM 5.2. *Let $\{g_n : n \in \mathbb{N}\} \cup \{g\}$ be a sequence of mappings from X into Y that is strictly differentiable-like at a fixed point $\bar{x} \in X$, and let $f_n : Y \to \mathbb{R} \cup \{+\infty\}$ be a sequence of proper lower semicontinuous convex functions that epiconvergences to f. Assume that $f(g(\bar{x}))$ is finite and that $(g_n)_n$ pointwise converges to g over some neighborhood of $\bar{x}$, and assume that there exist some $r > 0$ and $s > 0$ such that for all $n \in \mathbb{N}$*

$$s\,\mathbb{B}_Y \subset \nabla g(\bar{x})(r\,\mathbb{B}_X) - (\{f_n \le f(g(\bar{x})) + r\} - g(\bar{x})). \tag{19}$$

Then

$$f_n \circ g_n \xrightarrow{\text{epi}} f \circ g \quad \text{around } \bar{x},$$

in the sense that there exists some neighborhood V of $\bar{x}$ such that for every $x \in V$

$$(Li\,f_n \circ g_n)(x) = f \circ g(x) = (Ls\,f_n \circ g_n)(x).$$

Proof. Put $U := X \times \mathbb{R}$, $D_n := \operatorname{Epi} f_n$, $D := \operatorname{Epi} f$, and $\gamma(v,t) := \max(t, f \circ g(\bar{x}))$ for all $(v,t) \in U$. Define $q : U \to X \times \mathbb{R}$ and $\hat{g}$, $\hat{g}_n : (X \times \mathbb{R}) \times U \to Y \times \mathbb{R}$ by

$$q(u) = (\bar{x}, \gamma(u)), \quad \hat{g}(x,t,u) = (g(x), t), \quad \text{and} \quad \hat{g}_n(x,t,u) = (g_n(x), t).$$

Then for $\hat{A}_u(x,t) = (A(x), t)$ with $A := \nabla g(\bar{x})$ it follows from our second example of strictly differentiable-like sequences that the sequence $\{\hat{g}_n : n \in \mathbb{N}\} \cup \{\hat{g}\}$ is strictly

differentiable-like at $(q(u))_{u\in U}$. Moreover, choosing $\alpha > 0$ such that $(g_n)_n$ pointwise converges to g over $B(\bar{x},\alpha)$ we see that $((\hat{g}_n(.,u))_n$ pointwise converges to $\hat{g}(.,u)$ over $B(q(u),\alpha)$ for each $u \in U$. Also, as $\hat{g}_n(q(u),u) = (g_n(\bar{x}),\gamma(u))$, one has $\hat{g}_n(q(u),u) - \hat{g}(q(u),u) \to 0$ uniformly with respect to $u \in U$. Now put $\rho := r + s$ and consider any $u \in U$. For any $b \in \mathbb{B}_Y$ and $\lambda \in [-1,1]$ we can choose according to (19) some $b' \in \mathbb{B}_X$ and $y \in \{f_n \le f(g(\bar{x})) + r\}$ such that

$$sb = A(rb') - (y - g(\bar{x})),$$

and hence

$$s(b,\lambda) = (A(rb'), s\lambda + r) - ((y,\gamma(u)+r) - (g(\bar{x}),\gamma(u)).$$

As $|s\lambda + r| \le \rho$ and $\gamma(u) + r \ge f(g(\bar{x})) + r \ge f_n(y)$, we obtain

$$s(b,\lambda) \in \hat{A}_u(\rho\,\mathbb{B}_{X\times\mathbb{R}}) - (\mathrm{Epi}\, f_n - \hat{g}(\bar{x},\gamma(u),u)),$$

and hence

$$s\,\mathbb{B}_{Y\times\mathbb{R}} \subset \hat{A}_u(\rho\,\mathbb{B}_{X\times\mathbb{R}}) - (D_n - \hat{g}(q(u),u)).$$

Therefore we can apply Lemma 5.1 with $C_n = X \times \mathbb{R}$ to get some $\beta' > 0$ independent of $u \in U$ such that for all $u \in U$

$$\begin{aligned}[Li\,\hat{g}_n(.,u)^{-1}(D_n)] \cap B(q(u),\beta') &= \hat{g}(.,u)^{-1}(D) \cap B(q(u),\beta') \\ &= [Ls\,\hat{g}_n(.,u)^{-1}(D_n)] \cap B(q(u),\beta');\end{aligned}$$

that is,

$$\begin{aligned}[Li(\mathrm{Epi} f_n \circ g_n)] \cap B(q(u),\beta') &= (\mathrm{Epi} f \circ g) \cap B(q(u),\beta') \\ &= [Ls(\mathrm{Epi} f_n \circ g_n)] \cap B(q(u),\beta').\end{aligned} \tag{20}$$

According to the lower semicontinuity of $f \circ g$, there exists some positive number $\beta < \beta'$ such that for all $x \in B(\bar{x},\beta)$

$$f \circ g(\bar{x}) - \beta' \le f \circ g(x).$$

For any $x \in B(\bar{x},\beta)$ with $f \circ g(x) < +\infty$ we have

$$(x, f\circ g(x)) \in (\mathrm{Epi} f \circ g) \cap B(q(x, f\circ g(x)),\beta')$$

which, according to (20), ensures that

$$(x, f\circ g(x)) \in Li(\mathrm{Epi} f_n \circ g_n) = \mathrm{Epi}(Ls\, f_n \circ g_n),$$

and hence $(Ls\, f_n \circ g_n)(x) \le f \circ g(x)$. So we have for any $x \in B(\bar{x},\beta)$

$$(Ls\, f_n \circ g_n)(x) \le f \circ g(x). \tag{21}$$

Now consider the epilimit inferior. The study of this case can be done in a much more direct way. By assumptions, we may fix some $\beta'' > 0$ and $N \in \mathbb{N}$ such that the sequence $(g_n)_n$ pointwise converges to g over $B(\bar{x},2\beta'')$ and for all $n \ge N$ and $x, x' \in B(\bar{x},2\beta'')$

$$\|g_n(x) - g_n(x') - A(x - x')\| \le \|x - x'\|,$$

where A is the continuous linear mapping given by Definition 3.1. (In fact, $A = \nabla g(\bar{x})$ here because of the pointwise convergence near $\bar{x}$ of $(g_n)_n$ to g.) Fix any $x \in B(\bar{x}, \beta'')$, and consider any sequence $x_n \to x$. Then the last inequality above and the pointwise convergence property imply $g_n(x_n) \to g(x)$, and hence

$$f(g(x)) \leq \liminf f_n(g_n(x_n)) = \liminf f_n \circ g_n(x_n)$$

because $(f_n)_n$ epiconverges to f. It follows that $f \circ g(x) \leq (Lif_n \circ g_n)(x)$, and the proof is complete. □

According to our first example of a strictly differentiable-like sequence of mappings, we get the following corollary.

COROLLARY 5.2. *Let $\{g_n : n \in \mathbb{N}\} \cup \{g\}$ be a sequence of mappings from X into Y that are of class C^1 on a same neighborhood of $\bar{x} \in X$, and let f_n be a sequence of proper lower semicontinuous convex functions from Y into $\mathbb{R} \cup \{+\infty\}$ that epiconverges to f. Assume that $f(g(\bar{x}))$ is finite and*

(i) *$(g_n)_n$ pointwise converges to g on a neighborhood of $\bar{x}$;*
(ii) *$\lim_{\substack{x \to \bar{x} \\ n \to \infty}} \nabla g_n(x) = \nabla g(\bar{x})$;*
(iii) *there exist some $r > 0$ and $s > 0$ such that for all $n \in \mathbb{N}$*

$$s\,\mathbb{B}_Y \subset (\{f_n \leq r + f(g(\bar{x}))\} - g(\bar{x})) - \nabla g(\bar{x})(r\,\mathbb{B}_X).$$

Then

$$f_n \circ g_n \xrightarrow{\text{epi}} f \circ g \quad \text{around } \bar{x}.$$

6. Convergence of multipliers. Generally, when one uses the approximation of an optimization problem with a sequence of other problems (often simpler), one needs to know whether the sequence of multipliers will converge to a multiplier for the original problem. We consider, in this section, optimization problems associated with convexly composite functions.

Let $h_0 : X \to \mathbb{R} \cup \{+\infty\}$ and $h_i = X \to \mathbb{R}$ for $i = 1, \ldots, p$ be convexly composite functions with $h_i = f_i \circ g_i$. For each $n \in \mathbb{N}$, also let $h_{0,n} : X \to \mathbb{R} \cup \{+\infty\}$ and $h_{i,n} : X \to \mathbb{R}$ for $i = 1, \ldots, p$ be convexly composite functions with $h_{i,n} = f_{i,n} \circ g_{i,n}$ for $i = 0, \ldots, p$. Consider the optimization problem

$$\text{(P)} \qquad \text{Minimize } h_0(x) \quad \text{subject to } h_i(x) \leq 0 \quad \text{for all } i = 1, \ldots, p.$$

Recall that for a local solution $\bar{x}$ of (P), a vector $(\lambda_1, \ldots, \lambda_p) \in \mathbb{R}^p$ is a multiplier of Kuhn–Tucker type if $\lambda_i \geq 0$, $\lambda_i h_i(\bar{x}) = 0$, and

$$0 \in \partial h_0(\bar{x}) + \lambda_1 \partial h_1(\bar{x}) + \cdots + \lambda_p \partial h_p(\bar{x}). \tag{22}$$

Denote by (P_n) the corresponding optimization problem associated with the functions $h_{i,n}$ for $i = 0, \ldots, p$.

Similar problems have been also considered in Zolezzi [17] with equilower semidifferentiability assumptions for the functions defining the problems and under the separability of X. Convergence of multipliers with convexly composite functions h_i has already been studied in [4] under the equi-Lipschitz behavior (for $n \in \mathbb{N}$) of the constraint functions $h_{i,n}$ for $i = 1, \ldots, p$ as well as the equi-Lipschitz behavior of the objective functions. The convergence in [4] concerns multipliers of Fritz–John type and is derived from a convergence theorem of subdifferentials. Here we deal with multipliers of Kuhn–Tucker type, but we allow the objective functions to take the

value $+\infty$, and hence set constraints are implicitly incorporated, which makes a big difference from [4]. Another difference is provided by the approach that we use here. Instead of a convergence result of subdifferentials, we proceed in a direct way which allows us to drop the assumption in [4] of the weak-star sequential compactness of the closed unit ball of X^*. The theorem below will strongly use the results of the preceding section as well as some techniques of [4].

THEOREM 6.1. *Let $\bar{x}$ be a local solution of* (P). *Assume that for each $i = 0, 1, \ldots, p$ the assumptions of Corollary* 5.2 *are satisfied with the functions f_i, $f_{i,n}$, g_i and $g_{i,n}$. Let (x_n) be a sequence of local solutions of* ($\mathrm{P_n}$) *converging to $\bar{x}$. Then any limit of Kuhn–Tucker multipliers of* ($\mathrm{P_n}$) *is a Kuhn–Tucker multiplier of* (P).

Proof. Fix a Kuhn–Tucker multiplier $\lambda^n = (\lambda_1^n, \ldots, \lambda_p^n)$ of ($\mathrm{P_n}$) at x_n that converges to some $\lambda = (\lambda_1, \ldots, \lambda_p)$. According to (22), there exists, for each $i = 0, \ldots, p$, some $\zeta_i^n \in \partial f_i \circ g_i(x_n)$ such that

$$0 = \zeta_0^n + \lambda_1^n \zeta_1^n + \cdots + \lambda_p^n \zeta_p^n. \tag{23}$$

The equi-Lipschitz assumption entails that the sequence $(\zeta_1^n, \ldots, \zeta_p^n)$ is bounded, and hence by (23) the sequence (ζ_0^n) is also bounded. By the epiconvergence of $(f_{i,n} \circ g_{i,n})$ ensured by Corollary 5.2, we have for the positive number r (of (iii) in Corollary 5.2) some integer N_1 such that for $i = 0, \ldots, p$

$$f_i(g_i(\bar{x})) \le f_{i,n}(g(x_n)) + r \quad \text{and} \quad f_{i,n}(g_{i,n}(x_n)) < +\infty \text{ for all } n \ge N_1.$$

Using Lemma 4.1 we obtain, as in the first part of the proof of Theorem 4.1, some positive numbers r' and s' independent of n and an integer $N \ge N_1$ such that for all $n \ge N$ and $i = 0, \ldots, p$

$$s'\mathbb{B}_Y \subset (\{f_{i,n} \le r' + f_{i,n}(g_{i,n}(x_n))\} - g_{i,n}(x_n)) - \nabla g_{i,n}(x_n)(r'\mathbb{B}_X),$$

which entails (see (1.2) in [4]) that

$$\partial(f_{i,n} \circ g_{i,n})(x_n) = \{\xi \circ \nabla g_{i,n}(x_n) : \xi \in \partial f_{i,n}(g_{i,n}(x_n))\}. \tag{24}$$

So we may choose for each $n \ge N$ some $\xi_i^n \in \partial f_{i,n}(g_{i,n}(x_n))$ such that $\zeta_i^n = \xi_i^n \circ \nabla g_{i,n}(x_n)$. By (24) and Proposition 1.1 in [4] one has $s'\|\xi_i^n\| \le r'(1 + \|\zeta_i^n\|)$, and hence the sequence $(\xi_0^n, \ldots, \xi_p^n)$ is bounded. Let $(\zeta_0, \ldots, \zeta_p, \zeta_0, \ldots, \zeta_i^n)$, and hence the sequence $(\zeta_0^n, \ldots, \zeta_p^n)$ is bounded. Let $(\zeta_0, \ldots, \zeta_p, \xi_0, \ldots, \xi_p)$ be a weak-star cluster point of the sequence $(\zeta_0^n, \ldots, \zeta_p^n, \xi_p^n, \ldots, \xi_p^n)$. Fixing $i \in \{0, \ldots, p\}$ and $y \in Y$ and taking any sequence (y_n) converging to y, we have

$$\langle \xi_i^n, y_n - g_{i,n}(x_n)\rangle + f_{i,n}(g_{i,n}(x_n)) \le f_{i,n}(y_n),$$

and hence, taking the epiconvergence of $f_{i,n} \circ g_{i,n}$ into account,

$$\begin{aligned}
&\langle \xi_i, y - g_i(\bar{x})\rangle + f_i(g_i(\bar{x})) \le \langle \xi_i, y - g_i(\bar{x})\rangle + \liminf_{n\to\infty} f_{i,n}(g_{i,n}(x_n)) \\
&\le \limsup_{n\to\infty} \langle \xi_i^n, y_n - g_{i,n}(x_n)\rangle + \liminf_{n\to\infty} f_{i,n}(g_{i,n}(x_n)) \\
&\le \limsup_{n\to\infty} [\langle \xi_i^n, y_n - g_{i,n}(x_n)\rangle + f_{i,n}(g_{i,n}(x_n))] \\
&\le \limsup_{n\to\infty} f_{i,n}(y_n).
\end{aligned}$$

Taking the infimum over all sequence (y_n) converging to y we get, according to the epiconvergence of f_n to f,

$$\langle \xi_i, y - g_i(\bar{x})\rangle + f_i(g_i(\bar{x})) \le f_i(y).$$

This entails that $\xi_i \in \partial f_i\ (g_i(\bar{x})$, noting that the equality $\zeta_i^n = \xi_i^n \circ \nabla g_{i,n}(x_n)$ and the assumption (ii) in Corollary 5.2 imply that $\zeta_i = \xi_i \circ \nabla g(\bar{x})$. So by (1.2) in [4] we have $\zeta_i \in \partial(f_i \circ g_i)(\bar{x})$ and by (23) we also have

$$0 = \zeta_0 + \lambda_1\zeta_1 + \cdots + \lambda_p\zeta_p,$$

and hence

$$0 \in \partial h_0(\bar{x}) + \lambda_1\partial h_1(\bar{x}) + \cdots + \lambda_p\partial h_p(\bar{x}).$$

As $\lambda_i \geq 0$ (because $\lambda_i^n \geq 0$), it remains to show that $\lambda_i h_i(\bar{x}) = 0$ for each $i = 1, \ldots, p$. Fix any $i \in \{1, \ldots, p\}$. As $\bar{x}$ is an admissible point for the problem (P) we have $\lambda_i h_i(\bar{x}) \leq 0$. Now choose, according to the epiconvergence of $f_{i,n}$, some sequence $y_i^n \to g_i(\bar{x})$ with $f_{i,n}(y_i^n) \to f_i(g(\bar{x}))$. Since $\lambda_i^n \geq 0$ and $f_{i,n}$ is convex we also have

$$0 = \lambda_i^n f_{i,n}(g_{i,n}(x_n)) = \lambda_i^n f_{i,n}(y_i^n) - \lambda_i^n \langle \xi_i^n, y_i^n - g_{i,n}(x_n)\rangle. \tag{25}$$

Further, for n large enough

$$g_{i,n}(x_n) - g_i(\bar{x}) = g_{i,n}(\bar{x}) - g_i(\bar{x}) + \int_0^1 \nabla g_{i,n}(\bar{x} + t(x_n - \bar{x}))(x_n - \bar{x})dt,$$

which entails, according to assumptions (i) and (ii) in Corollary 5.2, that $g_{i,n}(x_n) \to g_i(\bar{x})$. As ξ_i is a weak-star cluster point of the bounded sequence (ξ_i^n), it follows from (25) that

$$0 \leq \lambda_i f_i(g(\bar{x})) = \lambda_i h_i(\bar{x}), \quad \text{and hence} \quad \lambda_i h_i(\bar{x}) = 0.$$

The proof is then complete. □

REFERENCES

[1] H. Attouch, *Variational Convergence for Functions and Operators*, Pitman, Boston, London, Melbourne, 1984.

[2] H. Attouch and R. J.-B. Wets, *Quantitative stability of variational systems.* I. *The epigraphical distance*, Trans. Amer. Math. Soc., 328 (1991), pp. 695–729.

[3] J. P. Aubin and H. Frankowska, *Set-Valued Analysis*, Birkhauser, Boston, 1990.

[4] C. Combari, R. A. Poliquin, and L. Thibault, *Convergence of subdifferentials of convexly composite functions*, Canad. J. Math., 51 (1999), pp. 250–265.

[5] G. Dal Maso, *An Introduction to Gamma-Convergence*, Birkhauser, Boston, 1993.

[6] A. L. Dontchev and W. W. Hager, *An inverse mapping theorem for set-valued maps*, Proc. Amer. Math. Soc., 121 (1994), pp. 481–489.

[7] A. L. Dontchev and W. W. Hager, *Implicit functions, Lipschitz maps, and stability in optimization*, Math. Oper. Res., 19 (1994), pp. 753–768.

[8] S. Guillaume, *Evolution equations governed by the subdifferential of a convex composite function in finite dimensional equations*, Discrete Contin. Dynam. Systems, 2 (1996), pp. 23–52.

[9] S. Guillaume, *Problèmes d'optimisation et d'évolution en analyse nonconvexe de type convexe composite*, Thesis, Université Montpellier II, Montpellier cedex, France, 1996.

[10] A. B. Levy, R. A. Poliquin, and L. Thibault, *Partial extensions of Attouch theorem and applications to proto-derivatives of subgradient mappings*, Trans. Amer. Math. Soc., 347 (1995), pp. 1269–1294.

[11] R. A. Poliquin, *An extension of Attouch's theorem and its application to second order epidifferentiability of convexly composite functions*, Trans. Amer. Math. Soc., 332 (1992), pp. 861–874.

[12] R. A. Poliquin, J. Vanderwerff, and V. Zizler, *Renormings and convex composite representation of functions*, Bull. Polish Acad. Sci. Math., 42 (1994), pp. 9–19.

[13] S. M. Robinson, *Regularity and stability for convex multivalued functions*, Math. Oper. Res., 1 (1976), pp. 130–143.
[14] S. M. Robinson, *Stability theory for systems of inequalities, Part* II: *Differentiable nonlinear systems*, SIAM J. Numer. Anal., 13 (1976), pp. 497–513.
[15] R. T. Rockafellar and R. J.-B Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
[16] L. Thibault and D. Zagrodny, *Integration of subdifferentials of lower semicontinuous functions*, J. Math. Anal. Appl., 189 (1995), pp. 33–58.
[17] T. Zolezzi, *Convergence of generalized gradients*, Set-Valued Anal., 2 (1994), pp. 381–393.

SIAM J. OPTIM.
Vol. 13, No. 4, pp. 1004–1013

ON THE PRIMAL-DUAL GEOMETRY OF LEVEL SETS IN LINEAR AND CONIC OPTIMIZATION*

ROBERT M. FREUND†

Abstract. For a conic optimization problem

$$P: \quad \begin{array}{ll} \text{minimize}_x & c^T x \\ \text{s.t.} & Ax = b, \\ & x \in C \end{array}$$

and its dual

$$D: \quad \begin{array}{ll} \text{supremum}_{y,s} & b^T y \\ \text{s.t.} & A^T y + s = c, \\ & s \in C^*, \end{array}$$

we present a geometric relationship between the primal objective function level sets and the dual objective function level sets, which shows that the maximum norms of the primal objective function level sets are nearly inversely proportional to the maximum inscribed radii of the dual objective function level sets.

Key words. convex optimization, conic optimization, duality, level sets

AMS subject classifications. 90C, 90C05, 90C22, 90C25, 90C46

PII. S1052623401393645

1. Introduction and motivation. This paper is concerned with the interrelated geometry of the primal objective function level sets and the dual objective function level sets of the following conic convex optimization primal and dual pair:

$$P: \quad \begin{array}{ll} \text{minimum}_x & c^T x \\ \text{s.t.} & Ax = b, \\ & x \in C \end{array}$$

and

$$D: \quad \begin{array}{ll} \text{supremum}_{y,s} & b^T y \\ \text{s.t.} & A^T y + s = c, \\ & s \in C^*, \end{array}$$

where C is a closed convex cone in a finite-dimensional normed vector space X. We present a geometric relationship between the primal objective function level sets and the dual objective function level sets, namely, that the maximum norms of the primal objective function level sets are nearly inversely proportional to the maximum inscribed radii of the dual objective function level sets.

To provide motivation without yet becoming encumbered by details, consider the case when C is the nonnegative orthant, i.e., $C = \Re^n_+ := \{x \in \Re^n \mid x \geq 0\}$, in which case P and D are simply linear programming (LP) primal and dual problems. Below we list and comment on two well-known properties of LP:

*Received by the editors August 10, 2001; accepted for publication (in revised form) September 22, 2002; published electronically March 5, 2003. This research has been partially supported through the Singapore-MIT Alliance.

http://www.siam.org/journals/siopt/13-4/39364.html

†MIT Sloan School of Management, 50 Memorial Drive, Cambridge, MA 02142-1347 (rfreund@mit.edu).

Property 1. Suppose that P and D are both feasible. Then the set of optimal solutions of P is unbounded if and only if there is no strictly feasible solution of D; that is, $A^T y + s = c, s \geq 0$ implies $s \not> 0$. This property is easily proved via LP duality, for example, and is part of the folklore of optimization. Put another way, Property 1 can be stated as follows:

> "The set of primal optimal solutions is unbounded if and only if every dual feasible s lies on the boundary of $\Re^n_+$."

Property 2. If P and D each have feasible solutions that satisfy all inequalities strictly, then the central trajectory exists, whereby for each $\mu > 0$ there exists unique feasible solutions $x(\mu)$ of P and $(y(\mu), s(\mu))$ of D for which $x_j(\mu) \cdot s_j(\mu) = \mu, j = 1, \ldots, n$. This is an elementary consequence of the optimality conditions for the logarithmic barrier functions appended to a linear program; see Wright [7], for example. Now notice here that for a given value $\mu > 0$, the norm $\|x(\mu)\|$ is large if and only if $\text{dist}(s_j(\mu), \partial\Re^n_+)$ is small. In fact, a little basic arithmetic manipulation easily shows that

$$\mu \leq \|x(\mu)\|_1 \cdot \min_j\{s_j(\mu)\} \leq n\mu,$$

which can then be used to assert the following:

> "For a given duality gap $\theta > 0$, there exists a primal feasible x and a dual feasible (y, s) with duality gap at most θ and with the property that $\theta/n \leq \|x\|_1 \cdot \text{dist}(s, \partial\Re^n_+) \leq \theta$."

This brief discussion points to an interrelationship between the norms of certain primal feasible solutions x and the distances of certain dual feasible solutions s to the boundary of the nonnegative orthant. In section 2 we make this interrelationship precise for the case of linear optimization in Theorem 2.1, which shows that the maximum norms of primal objective level sets are almost exactly inversely proportional to the maximum distances to the boundary of dual objective level sets. In fact, just as linear optimization is a special case of more general conic convex optimization, Theorem 2.1 is a special case of a more general theorem that demonstrates an inverse proportional relationship between the maximum norms of primal objective level sets and the maximum distances to the boundary of dual objective level sets in conic convex optimization. This more general result is presented in section 3 as Theorem 3.2 and is the main result of this paper. Section 4 discusses several aspects of cone geometry that arise in our development, and section 5 contains proofs.

Notation. We denote real n-dimensional space and the nonnegative n-dimensional orthant by $\Re^n$ and $\Re^n_+$, respectively. Let $e = (1, \ldots, 1)^T$ denote the vector of 1's in $\Re^n$.

2. Primal-dual geometry of level sets for linear optimization. Consider the following dual pair of linear optimization problems:

$$\begin{array}{lll} LP: & \text{minimize} & c^T x \\ & \text{s.t.} & Ax = b, \\ & & x \geq 0 \end{array}$$

and

$$\begin{array}{lll} LD: & \text{maximize} & b^T y \\ & \text{s.t.} & A^T y + s = c, \\ & & s \geq 0, \end{array}$$

whose common optimal value is z^*. For $\epsilon > 0$ and $\delta > 0$, define the ϵ- and δ-level sets for the primal and dual problems as follows:

$$P_\epsilon := \left\{ x \mid Ax = b, x \geq 0, c^T x \leq z^* + \epsilon \right\}$$

and

$$D_\delta := \left\{ s \mid \exists\ y \text{ satisfying } A^T y + s = c, s \geq 0, b^T y \geq z^* - \delta \right\}.$$

Define

$$\begin{array}{rlll} R_\epsilon := & \max & \|x\|_1 \\ & \text{s.t.} & Ax = b, \\ & & c^T x \leq z^* + \epsilon, \\ & & x \geq 0 \end{array} \tag{2.1}$$

and

$$\begin{array}{rlll} r_\delta := & \max & \min_j \{s_j\} \\ & \text{s.t.} & A^T y + s = c, \\ & & b^T y \geq z^* - \delta, \\ & & s \geq 0. \end{array} \tag{2.2}$$

The quantity R_ϵ is simply the size of the largest vector x in the primal level set P_ϵ, measured in the L_1-norm. The quantity r_δ can be interpreted as the positivity of the most positive vector s in the dual level set D_δ or, equivalently, as the maximum distance to the boundary of the nonnegative orthant over all points s in D_δ. The following theorem presents a reciprocal relationship between R_ϵ and r_δ.

THEOREM 2.1. *Suppose that z^* is finite. If R_ϵ is positive and finite, then*

$$\min\{\epsilon, \delta\} \ \leq \ R_\epsilon \cdot r_\delta \ \leq \ \epsilon + \delta.$$

Otherwise, $R_\epsilon = 0$ if and only if $r_\delta = +\infty$, and $R_\epsilon = +\infty$ if and only if $r_\delta = 0$.

Theorem 2.1 bounds the size of the largest vector in P_ϵ and the positivity of the most positive vector in D_δ from above and below, and shows that these quantities are almost exactly inversely proportional. In fact, taking $\delta = \epsilon$, the result states that $R_\epsilon \cdot r_\epsilon$ lies in the interval $[\epsilon, 2\epsilon]$. The proof of this theorem follows as a special case of a more general result for convex conic optimization, namely Theorem 3.2 in section 3.

Remark 2.1. If $R_\epsilon < \infty$, then

$$R_{\epsilon'} \leq \left(\frac{\epsilon'}{\epsilon} \right) R_\epsilon \tag{2.3}$$

for all $\epsilon' \geq \epsilon$. □

Proof. If $R_\epsilon = 0$, the result follows trivially, since then $R_{\epsilon'} = 0$ for all $\epsilon' > 0$. So suppose that $0 < R_\epsilon < +\infty$. Let x^* be an optimal solution of LP, and let $x' \in P_{\epsilon'}$ be given. Then $x := \frac{\epsilon}{\epsilon'} x' + \frac{\epsilon' - \epsilon}{\epsilon'} x^*$ satisfies $x \in P_\epsilon$, whereby $\|x\|_1 \leq R_\epsilon$. Now notice that $\|x'\|_1 = e^T x' = \frac{\epsilon'}{\epsilon} e^T x - \frac{\epsilon' - \epsilon}{\epsilon} e^T x^* \leq \frac{\epsilon'}{\epsilon} e^T x = \frac{\epsilon'}{\epsilon} \|x\|_1 \leq \frac{\epsilon'}{\epsilon} R_\epsilon$. Therefore $R_{\epsilon'} \leq \frac{\epsilon'}{\epsilon} R_\epsilon$, proving the result. □

Remark 2.1 bounds the rate of growth of $R_{\epsilon'}$ as ϵ' increases and shows that $R_{\epsilon'}$ grows at most linearly in ϵ' and at a rate no greater than $\frac{R_\epsilon}{\epsilon}$. There is a version of (2.3) for r_δ and $r_{\delta'}$, namely

$$r_{\delta'} \geq \left(\frac{\delta'}{\delta}\right) r_\delta \tag{2.4}$$

for all $0 \leq \delta' \leq \delta$, which is true as an elementary consequence of the convexity of the feasible region of LD.

By exchanging the roles of the primal and dual problems, we obviously can construct analogous results for the most positive vector x in P_ϵ as well as for the size of the largest vector s in D_δ.

3. Conic optimization with a norm on X. We now consider the generalization of linear optimization to convex optimization in conic linear form:

$$\begin{array}{lrll} P: & z^* := \text{minimum}_x & c^T x \\ & \text{s.t.} & Ax = b, \\ & & x \in C \end{array}$$

and its dual

$$\begin{array}{lrll} D: & v^* := \text{supremum}_{y,s} & b^T y \\ & \text{s.t.} & A^T y + s = c, \\ & & s \in C^*, \end{array}$$

where $C \subset X$ is a closed convex cone in the (finite) n-dimensional linear vector space X, and b lies in the (finite) m-dimensional vector space Y. This format for convex optimization dates back at least to Duffin [2]. Strong duality results can be found in [2] as well as in Ben-Israel, Charnes, and Kortanek [1].

For $\epsilon > 0$ and $\delta > 0$, we define the ϵ- and δ-level sets for the primal and dual problems as follows:

$$P_\epsilon := \left\{x \mid Ax = b, x \in C, c^T x \leq z^* + \epsilon\right\}$$

and

$$D_\delta := \left\{s \mid \exists\ y \text{ satisfying } A^T y + s = c, s \in C^*, b^T y \geq v^* - \delta\right\}.$$

We make the following assumption.

Assumption A. z^* is finite. The cone C satisfies $C \neq \{0\}$, and C contains no line (whereby C^* has an interior).

Suppose that X is endowed with a norm $\|\cdot\|$, and so X^* is endowed with the dual norm $\|\cdot\|_*$. Let $B(x,r)$ and $B^*(s,r)$ denote the balls of radius r centered at $x \in X$ and $s \in X^*$, respectively, defined for the appropriate norms.

We denote the maximum norm of P_ϵ by R_ϵ, defined as

$$\begin{array}{lrll} R_\epsilon := & \max_x & \|x\| \\ & \text{s.t.} & x \in P_\epsilon. \end{array} \tag{3.1}$$

We denote by r_δ the inscribed size of D_δ, defined as

$$\begin{array}{lrll} r_\delta := & \max_{s,r} & r \\ & \text{s.t.} & s \in D_\delta, \\ & & B^*(s,r) \subset C^*. \end{array} \tag{3.2}$$

As in the case of linear optimization, r_δ measures the distance of the most interior point of the dual level set D_δ to the boundary of the cone C^*. Put another way, r_δ measures the "interiorness" (with respect to C^*) of the most interior point in D_δ.

Before presenting the version of Theorem 2.1 for convex conic optimization, we first review the concept of the min-width of a cone. We use the following definition of the min-width.

DEFINITION 3.1. *Let $K \subset X$ be a closed convex cone in the normed linear vector space X satisfying* (i) *K has a nonempty interior and* (ii) *$K \neq X$. The min-width of K is defined as*

$$\tau_K \ := \ \max_{x \in \text{int}K} \left\{ \frac{\text{dist}\,(x, \partial K)}{\|x\|} \right\} \ = \max_{x \neq 0} \left\{ \frac{r}{\|x\|} \ \middle| \ B(x,r) \subset K \right\} .$$

Note that τ_K measures the maximum ratio of the radius to the norm of the center of an inscribed ball in K, and so larger values of τ_K correspond to an intuitive notion of greater minimum width of K. The quantity τ_K was called the "inner measure" of K for Euclidean norms in Goffin [5] and has been used more recently for general norms in analyzing condition measures for conic convex optimization; see [3]. Note that $\tau_K \in (0,1]$, since K has a nonempty interior and $K \neq X$, and τ_K is attained for some $x^0 \in \text{int}K$ satisfying $\|x^0\| = 1$, as well as along the ray αx^0 for all $\alpha > 0$. Let τ_{K^*} be defined similarly for the dual cone K^*.

The following is analogous to Theorem 2.1 for conic problems.

THEOREM 3.2. *Suppose that Assumption A holds. If R_ϵ is positive and finite, then $z^* = v^*$ and*

$$\tau_{C^*} \cdot \min\{\epsilon, \delta\} \ \leq \ R_\epsilon \cdot r_\delta \ \leq \ \epsilon + \delta. \tag{3.3}$$

If $R_\epsilon = 0$, then $z^* = v^*$ and $r_\delta = +\infty$; else if $R_\epsilon = +\infty$ and v^* is finite, then $r_\delta = 0$.

Here we have had to introduce the min-width τ_{C^*} into the left inequality of (3.3), somewhat weakening the result. In the next section we show that the left inequality can be tight. We also show how to define a family of cone-based norms for which $\tau_{C^*} = 1$, and we show that for norms induced by a ϑ-normal barrier function on C the min-width constant τ_{C^*} satisfies $\tau_{C^*} \geq 1/\sqrt{\vartheta}$. Theorem 3.2 is proved in section 5. Here we use Theorem 3.2 to prove Theorem 2.1.

Proof of Theorem 2.1. Note that LP is a special case of P with $X = \Re^n$ and $C = \Re^n_+$, whereby $C^* = \Re^n_+$. Endow X with the L_1-norm $\|\cdot\| = \|\cdot\|_1$, whose dual norm on X^* is the L_∞-norm $\|\cdot\|_* = \|\cdot\|_\infty$. To prove the theorem it suffices to show that $\tau_{C^*} = 1$, which we do now. Let $s^0 = e$, and note that $\|s^0\|_\infty = 1$, and that $B^*(s^0, 1) = \{s \mid \|s - e\|_\infty \leq 1\} \subset \Re^n_+ = C^*$, whereby $\tau_{C^*} \geq 1$. However, $\tau_{C^*} \leq 1$ because C^* is a pointed cone, and so $\tau_{C^*} = 1$, completing the proof. ∎

The following remark, analogous to Remark 2.1, is proved in section 5.

Remark 3.1. If $R_\epsilon < \infty$, then

$$R_{\epsilon'} \leq \left(\frac{\epsilon'}{\epsilon} \right) \left(\frac{1}{\tau_{C^*}} \right) R_\epsilon$$

for all $\epsilon' \geq \epsilon$.

4. On the min-width constant.

4.1. The min-width constant can be tight. Here we show by example that the left inequality in (3.3) can be tight, and so the constant τ_{C^*} cannot be replaced by a larger quantity. Let $X = \Re^n$ and $C = \Re^n_+$ (whereby $C^* = \Re^n_+$), and let X be endowed with the L_p-norm $\|x\|_p := (\sum_{j=1}^n |x_j|^p)^{\frac{1}{p}}$ for $1 \le p \le +\infty$, whose dual norm is $\|s\|_* = \|s\|_q$, where $\frac{1}{p} + \frac{1}{q} = 1$ with appropriate limits for $p, q = 1$ and/or ∞. Then it is straightforward to show that $\tau_C = n^{-\frac{1}{p}}$ and $\tau_{C^*} = n^{-\frac{1}{q}}$. Consider the following LP primal and dual instance:

$$\begin{array}{rccc} \tilde{P}: & \min_x & 0^T x & \qquad \tilde{D}: \quad \max_{y,s} \quad e^T y \\ & \text{s.t.} & Ix = e, & \qquad \text{s.t.} \quad Iy + s = 0, \\ & & x \in \Re^n_+, & \qquad s \in \Re^n_+, \end{array}$$

whose common optimal value is $z^* = 0$. Then $R_\epsilon = \|e\|_p = n^{\frac{1}{p}}$, and $r_\delta = \frac{\delta}{n}$ for all $\epsilon, \delta > 0$. Let $\epsilon := \delta$, whereby $R_\epsilon \cdot r_\delta = n^{\frac{1}{p}} \cdot \frac{\delta}{n} = \delta \cdot n^{\left(\frac{1}{p}-1\right)} = \delta \cdot n^{-\frac{1}{q}} = \delta \cdot \tau_{C^*} = \min\{\epsilon, \delta\}\tau_{C^*}$, which shows that the left inequality of (3.3) can indeed be tight.

4.2. Min-widths for the family of norms induced by a ϑ-normal barrier. In this subsection we assume that C is a regular cone; i.e., C is pointed and has an interior. Suppose that $F(\cdot) : \text{int}C \to \Re$ is a ϑ-normal barrier for C; see [6]. Then $F^*(\cdot) : \text{int}C^* \to \Re$, the conjugate function of $F(\cdot)$, is also a ϑ-normal barrier for C^*; see [6] as well.

Let $s^0 \in \text{int}C^*$ be given. The norm induced by the ϑ-normal barrier $F(\cdot)$ at s^0 is defined as follows:

$$\|s\|_{*,s^0} := \sqrt{s^T H^*(s^0) s},$$

where $H^*(s^0)$ is the Hessian of $F^*(\cdot)$ evaluated at s^0. It then follows from Theorem 2.1.1 of [6] that $B^*(s^0, 1) \subset C^*$ and from Proposition 2.3.4 of [6] that $\|s^0\|_{*,s^0} = \sqrt{\vartheta}$. Therefore under the dual norm $\|s\|_* := \|s\|_{*,s^0}$ we have $\tau_{C^*} \ge 1/\sqrt{\vartheta}$.

4.3. A family of norms on X for which $\tau_{C^*} = 1$. In this subsection we also assume that C is a regular cone. For every $s^0 \in \text{int}C^*$, there is a norm analogous to the L_∞-norm for the nonnegative orthant for which the associated min-width is $\tau_{C^*} = 1$. To see this, consider a given interior point $s^0 \in \text{int}C^*$, and define the following norm:

$$\begin{array}{rccc} \|s\|_* & := & \min_\alpha & \alpha \\ & & \text{s.t.} & s + \alpha s^0 \in C^*, \\ & & & -s + \alpha s^0 \in C^*. \end{array}$$

It is a straightforward exercise to verify that $\|\cdot\|_*$ is indeed a norm, and its dual norm turns out to be

$$\begin{array}{rccc} \|x\| & := & \min_{x^1, x^2} & (s^0)^T (x^1 + x^2) \\ & & \text{s.t.} & x^1 - x^2 = x, \\ & & & x^1 \in C, \\ & & & x^2 \in C. \end{array}$$

Under $\|\cdot\|_*$, it is easily shown that $\|s^0\|_* = 1$ and $\tau_{C^*} = 1$.

In the case when $X = \Re^n$, $C = C^* = \Re^n_+$, and $s^0 = e$, we recover the L_∞-norm as $\|s\|_*$ for $s \in X^* = \Re^n$ and the L_1-norm as $\|x\|$ for $x \in X = \Re^n$.

5. Proofs of main results. We start by pointing out a fact about strong duality in general conic convex optimization that we will use in our proof of Theorem 3.2. Suppose we have a primal and dual pair of conic convex optimization problems

$$\hat{P}: \quad \hat{z}^* := \begin{array}{ll} \inf_x & f^T x \\ \text{s.t.} & Mx = g, \\ & x \in K, \end{array} \qquad \hat{D}: \quad \hat{v}^* := \begin{array}{ll} \sup_{y,s} & g^T y \\ \text{s.t.} & M^T y + s = f, \\ & s \in K^*, \end{array}$$

where $K \subset X$ is a closed convex cone in the (finite) n-dimensional linear vector space X, and g lies in the (finite) m-dimensional vector space Y. The following lemma presents a sufficient condition for this pair to exhibit strong duality.

LEMMA 5.1. *Assume that $\hat{z}^*$ is finite and for some $\epsilon > 0$ the level set $\hat{P}_\epsilon := \{x \mid Mx = g, x \in K, f^T x \le \hat{z}^* + \epsilon\}$ is bounded. Then $\hat{P}$ attains its optimum and $\hat{z}^* = \hat{v}^*$.*

Proof. Note that $\hat{P}$ attains its optimum, since $\hat{P}_\epsilon$ is bounded. The boundedness of $\hat{P}_\epsilon$ also implies that

$$\{0\} = \{x \in X \mid Mx = 0, x \in K, f^T x \le 0\}. \tag{5.1}$$

It is elementary to show that $\hat{z}^* \ge \hat{v}^*$. Suppose that $\hat{z}^* > \hat{v}^*$, let $\bar{\epsilon}$ be such that $0 < \bar{\epsilon} < \hat{z}^* - \hat{v}^*$, and let

$$S = \{(w, \alpha) | \ \exists y \in Y^*, s \in K^* \text{ satisfying } w = M^T y + s - f, g^T y \ge \hat{v}^* + \bar{\epsilon} - \alpha\}.$$

Then S is a nonempty convex set in $X^* \times \Re$, and $(0,0) \notin S$, whereby there exists $(x, \theta) \ne 0$ satisfying $x^T w + \theta\alpha \ge 0$ for all $(w, \alpha) \in S$. Therefore

$$x^T \left(M^T y + s - f\right) + \theta \left(-g^T y + \hat{v}^* + \bar{\epsilon} + \eta\right) \ge 0 \quad \forall y \in Y^*, \forall s \in K^*, \forall \eta \ge 0. \tag{5.2}$$

This implies that $Mx = g\theta$, $\theta \ge 0$, and $x \in K$. We now have two cases.

Case 1. $\theta > 0$. Without loss of generality we can assume that $\theta = 1$. Therefore x is feasible for $\hat{P}$, and (5.2) also implies that $\hat{z}^* \le f^T x \le \hat{v}^* + \bar{\epsilon} < \hat{z}^*$, which is a contradiction.

Case 2. $\theta = 0$. In this case $x \ne 0$, $x \in K$, $Mx = 0$, and (5.2) implies that $f^T x \le 0$, contradicting (5.1).

In both cases we have a contradiction, and so $\hat{z}^* = \hat{v}^*$. ∎

We next state some properties of norms and the min-width. The following is a special case of the Hahn–Banach theorem; for a short proof of this proposition based on the subdifferential operator, see Proposition 2 of [4].

PROPOSITION 5.2. *For every $x \in X$, there exists $\bar{x} \in X^*$ with the property that $\|\bar{x}\|_* = 1$ and $\|x\| = \bar{x}^T x$.* ∎

The following exhibits some useful properties of the min-width of a cone.

PROPOSITION 5.3. *Suppose K^* is a convex cone whose min-width τ_{K^*} is attained at some point $s^0 \in \text{int}K^*$ satisfying $\|s^0\|_* = 1$. Then*

(i) *$\tau_{K^*}\|x\| \le (s^0)^T x \le \|x\|$ for all $x \in K$, and*

(ii) *if $s - \lambda s^0 \in K^*$, then $B^*(s, \lambda\tau_{K^*}) \subset K^*$.*

Proof. For a given $x \in K \subset X$, there exists $\bar{x} \in X^*$ for which $\|\bar{x}\|_* = 1$ and $\|x\| = \bar{x}^T x$ from Proposition 5.2. By construction of s^0 we have $B^*(s^0, \tau_{K^*}) \subset K^*$, and so $s^0 - \tau_{K^*}\bar{x} \in K^*$. Therefore $\|x\| = \|x\|\|s^0\|_* \ge (s^0)^T x = (s^0 - \tau_{K^*}\bar{x} + \tau_{K^*}\bar{x})^T x \ge \tau_{K^*}\bar{x}^T x = \tau_{K^*}\|x\|$, proving (i). To prove (ii), let $u := s - \lambda s^0$. Then $s = u + \lambda s^0$, where $u \in K^*$ and $B^*(s^0, \tau_{K^*}) \subset K^*$, whereby it follows that $B^*(s, \lambda\tau_{K^*}) \subset K^*$. ∎

We are now ready to prove Theorem 3.2, which we do by proving the following four statements:

(i) If R_ϵ is positive and finite, then $z^* = v^*$ and $R_\epsilon \cdot r_\delta \le \epsilon + \delta$.
(ii) If R_ϵ is positive and finite, then $R_\epsilon \cdot r_\delta \ge \tau_{C^*} \min\{\epsilon, \delta\}$.
(iii) If $R_\epsilon = 0$, then $z^* = v^*$ and $r_\delta = +\infty$.
(iv) If $R_\epsilon = +\infty$ and v^* is finite, then $r_\delta = 0$.

Proof of (i). Since R_ϵ is finite, it follows that P_ϵ is bounded, and so $z^* = v^*$ from Lemma 5.1. Let $x \in P_\epsilon$ be given, and let $\bar{x}$ satisfy $\|\bar{x}\|_* = 1$ and $\|x\| = \bar{x}^T x$; see Proposition 5.2. Now suppose that $s \in D_\delta$ satisfies $B^*(s, r) \subset C^*$ for some $r \ge 0$. It follows that $\epsilon + \delta \ge c^T x - z^* - b^T y + v^* = c^T x - b^T y = x^T s = x^T(s - r\bar{x} + r\bar{x}) \ge rx^T\bar{x} = r\|x\|$. As this is true for all $x \in P_\epsilon$ and all $s \in D_\delta$ satisfying $B^*(s, r) \subset C^*$, it follows that $\epsilon + \delta \ge R_\epsilon \cdot r_\delta$. ∎

Proof of (ii). Let s^0 satisfy $\|s^0\|_* = 1$ and $B^*(s^0, \tau_{C^*}) \subset C^*$, and consider the following conic convex dual programs:

$$\begin{array}{lllll} \bar{P}: & \bar{R}_\epsilon := & \max_x & (s^0)^T x & \\ & & \text{s.t.} & Ax = b, & \\ & & & c^T x \le z^* + \epsilon, & \\ & & & x \in C, & \end{array} \qquad \begin{array}{lllll} \bar{D}: & \bar{Q} := & \inf_{y,s,\theta} & -b^T y + (z^* + \epsilon)\theta & \\ & & \text{s.t.} & A^T y + s = \theta c, & \\ & & & s - s^0 \in C^*, & \\ & & & \theta \ge 0. & \end{array}$$

From Proposition 5.3 it follows that $\tau_{C^*}\|x\| \le (s^0)^T x \le \|x\|$ for any $x \in C$, whereby $\tau_{C^*} R_\epsilon \le \bar{R}_\epsilon \le R_\epsilon$, and, in particular, the level sets of $\bar{P}$ are bounded. Then we can invoke Lemma 5.1 on the pair $\bar{P}, \bar{D}$ and assert that $\bar{P}$ attains its optimum and $\bar{R}_\epsilon = \bar{Q}$.

For $\alpha \in (0, \min\{\epsilon, \delta\})$ we show below that

$$r_\delta \ge \frac{\tau_{C^*}}{\bar{R}_\epsilon + \alpha} \left(\min\{\epsilon, \delta - \alpha\}\right) \tag{5.3}$$

and letting $\alpha \to 0$ will complete the proof since (5.3) and $\alpha \to 0$ imply that $R_\epsilon \cdot r_\delta \ge \bar{R}_\epsilon \cdot r_\delta \ge \tau_{C^*} \min\{\epsilon, \delta\}$. For $\alpha \in (0, \min\{\epsilon, \delta\})$ let (y, s, θ) be a feasible solution of $\bar{D}$ satisfying

$$-b^T y + (z^* + \epsilon)\theta \le \bar{Q} + \alpha = \bar{R}_\epsilon + \alpha, \tag{5.4}$$

and define $w := s - s^0 \in C^*$. We prove (5.3) by considering three cases.

Case 1. $\theta = 0$. In this case $A^T y + s = 0$ and $-b^T y \le \bar{R}_\epsilon + \alpha$. Let $(\bar{y}, \bar{s})$ be any feasible solution of D satisfying $b^T\bar{y} \ge z^* - \alpha$, and define

$$(\hat{y}, \hat{s}) := (\bar{y}, \bar{s}) + \frac{\delta - \alpha}{\bar{R}_\epsilon + \alpha}(y, s).$$

Then $(\hat{y}, \hat{s})$ is feasible for D, and

$$b^T\hat{y} = b^T\bar{y} + \frac{\delta - \alpha}{\bar{R}_\epsilon + \alpha} b^T y \ge z^* - \alpha - \delta + \alpha = z^* - \delta.$$

Also, $\hat{s} - \frac{\delta-\alpha}{\bar{R}_\epsilon+\alpha} s^0 = \frac{\delta-\alpha}{\bar{R}_\epsilon+\alpha} w + \bar{s} \in C^*$, whereby $\hat{s} \in D_\delta$ and $B^*(\hat{s}, \frac{\delta-\alpha}{\bar{R}_\epsilon+\alpha}\tau_{C^*}) \subset C^*$ from Proposition 5.3. This then implies that $r_\delta \ge \frac{\delta-\alpha}{\bar{R}_\epsilon+\alpha}\tau_{C^*}$, which implies (5.3).

Case 2. $\theta > 0$ and $\frac{\bar{R}_\epsilon+\alpha}{\theta} - \epsilon \le \delta$. Define

$$(\hat{y}, \hat{s}) = \frac{1}{\theta}(y, s),$$

whereby $(\hat{y}, \hat{s})$ satisfies $\hat{s} \in C^*$, $A^T\hat{y} + \hat{s} = c$, and

$$b^T\hat{y} = \frac{1}{\theta} b^T y \geq -\frac{\bar{R}_\epsilon + \alpha}{\theta} + z^* + \epsilon \geq z^* - \delta,$$

which shows that $\hat{s} \in D_\delta$. Furthermore, $\hat{s} = \frac{s^0}{\theta} + \frac{w}{\theta}$, $w \in C^*$, and so $\hat{s} - \frac{1}{\theta}s^0 \in C^*$. Now it follows from Proposition 5.3 that $B^*\left(\hat{s}, \frac{\tau_{C^*}}{\theta}\right) \subset C^*$, and so $r_\delta \geq \frac{\tau_{C^*}}{\theta}$. However,

$$z^* \geq b^T\hat{y} \geq -\frac{\bar{R}_\epsilon + \alpha}{\theta} + z^* + \epsilon,$$

and so $\frac{1}{\theta} \geq \frac{\epsilon}{\bar{R}_\epsilon + \alpha}$, whereby $r_\delta \geq \frac{\tau_{C^*}}{\theta} \geq \frac{\epsilon}{\bar{R}_\epsilon + \alpha}\tau_{C^*}$, which then implies (5.3).

Case 3. $\theta > 0$ and $\frac{\bar{R}_\epsilon + \alpha}{\theta} - \epsilon \geq \delta$. Let $(\bar{y}, \bar{s})$ be any feasible solution of D satisfying

$$b^T\bar{y} \geq z^* - \alpha, \tag{5.5}$$

and define

$$(\hat{y}, \hat{s}) = \lambda\left(\frac{(y, s)}{\theta}\right) + (1 - \lambda)(\bar{y}, \bar{s}),$$

where

$$\lambda = \frac{\delta - \alpha}{\frac{\bar{R}_\epsilon + \alpha}{\theta} - \epsilon - \alpha}.$$

Then $\lambda \in [0, 1]$ for $\alpha \in (0, \delta)$, and so $(\hat{y}, \hat{s})$ is a convex combination of $\frac{(y,s)}{\theta}$ and $(\bar{y}, \bar{s})$ and so satisfies $A^T\hat{y} + \hat{s} = c$, $\hat{s} \in C^*$. It also follows from (5.4) and (5.5) that $b^T\hat{y} \geq z^* - \delta$, whereby $\hat{s} \in D_\delta$. Finally, $\hat{s} - \frac{\lambda}{\theta}s^0 \in C^*$, and so from Proposition 5.3 we have $B^*(\hat{s}, \frac{\lambda\tau_{C^*}}{\theta}) \subset C^*$. Therefore

$$r_\delta \geq \frac{\lambda\tau_{C^*}}{\theta} = \frac{\delta - \alpha}{\bar{R}_\epsilon + \alpha - \alpha\theta - \epsilon\theta}\tau_{C^*} \geq \frac{\delta - \alpha}{\bar{R}_\epsilon + \alpha}\tau_{C^*},$$

from which (5.3) follows.

Therefore (5.3) is true in all cases, and the proof is complete. □

Proof of (iii). Since $R_\epsilon = 0$ it follows that $P_\epsilon = \{0\}$ is bounded, and so $z^* = v^*$ from Lemma 5.1. It then follows that $b = 0$, and so $z^* = v^* = 0$. To prove that $r_\delta = +\infty$ it suffices to prove that there exists $(\tilde{y}, \tilde{s})$ satisfying

$$A^T\tilde{y} + \tilde{s} = 0 \quad \text{and} \quad \tilde{s} \in \text{int}C^*. \tag{5.6}$$

Let $s^0, \bar{P}$, and $\bar{D}$ be exactly as in the proof of (ii), and the same logic as in the proof of (ii) yields $\bar{R}_\epsilon = \bar{Q} = 0$; notice that because $b = 0$ and $z^* = 0$ it follows that the objective function of $\bar{D}$ is simply $\epsilon\theta$. If $\bar{D}$ attains its optimal value $\bar{Q} = 0$, then any optimal solution (y^*, s^*, θ^*) of $\bar{D}$ will satisfy $\theta^* = 0$, and so (5.6) will be satisfied by setting $(\tilde{y}, \tilde{s}) = (y^*, s^*)$. Alternatively, if $c = 0$, then the (y, s) variables of any feasible solution (y, s, θ) of $\bar{D}$ will satisfy (5.6). It remains to consider the case when $\bar{D}$ does not attain its optimum and $c \neq 0$. Let $\alpha := \frac{\epsilon \cdot \tau_{C^*}}{2\|c\|_*}$, and let (y, s, θ) be a feasible solution of $\bar{D}$ satisfying $\epsilon\theta = -b^Ty + (z^* + \epsilon)\theta \leq \bar{R}_\epsilon + \alpha = \frac{\epsilon \cdot \tau_{C^*}}{2\|c\|_*}$; then, in particular, $\theta \leq \frac{\tau_{C^*}}{2\|c\|_*}$. Define $w := s - s^0 \in C^*$. Let $(\tilde{y}, \tilde{s}) = (y, s^0 + w - \theta c)$. Then $A^T\tilde{y} + \tilde{s} = A^Ty + s - \theta c = 0$, and $\tilde{s} = s^0 + w - \theta c = w + \frac{1}{2}s^0 + \frac{1}{2}(s^0 - 2\theta c)$. Notice

that $\|2\theta c\|_* \leq \tau_{C^*}$, and so $s^0 - 2\theta c \in C^*$, and also $w \in C^*$ and $s^0 \in \text{int}C^*$, whereby it follows that $\tilde{s} \in \text{int}C^*$, validating (5.6). ∎

Proof of (iv). Because $R_\epsilon = +\infty$ it follows that there exists $x \neq 0$ satisfying $x \in C, Ax = 0$, and $c^T x = 0$. From Proposition 5.2 there also exists $\bar{x} \in X^*$ for which $\|\bar{x}\|_* = 1$ and $\|x\| = \bar{x}^T x$. Now suppose that v^* is finite, and let $\hat{s} \in D_\delta$ satisfy $B^*(\hat{s}, r) \subset C^*$ for some $r \geq 0$. Then there exists $\hat{y}$ for which $A^T\hat{y} + \hat{s} = c$, and so $x^T\hat{s} = x^T(c - A^T\hat{y}) = 0 - 0 = 0$. Also, $\hat{s} - r\bar{x} \in C^*$, and $x \in C$ implies that $0 \leq x^T(\hat{s} - r\bar{x}) = -rx^T\bar{x} = -r\|x\|$, whereby $r = 0$. This then implies that $r_\delta = 0$. ∎

Proof of Remark 3.1. If $R_\epsilon = 0$, the result follows trivially, since then $R_{\epsilon'} = 0$ for all $\epsilon' > 0$. So suppose that $0 < R_\epsilon < +\infty$. Let x^* be an optimal solution of P (P attains its optimum; see Lemma 5.1), and let $x' \in P_{\epsilon'}$ be given. Then $x := \frac{\epsilon}{\epsilon'}x' + \frac{\epsilon' - \epsilon}{\epsilon'}x^*$ satisfies $x \in P_\epsilon$, whereby $\|x\| \leq R_\epsilon$. Let s^0 satisfy $\|s^0\|_* = 1$ and $B^*(s^0, \tau_{C^*}) \subset C^*$. Then from Proposition 5.3 we have $\tau_{C^*}\|x'\| \leq (s^0)^T x' = \frac{\epsilon'}{\epsilon}(s^0)^T x - \frac{\epsilon' - \epsilon}{\epsilon}(s^0)^T x^* \leq \frac{\epsilon'}{\epsilon}(s^0)^T x \leq \frac{\epsilon'}{\epsilon}\|x\| \leq \frac{\epsilon'}{\epsilon}R_\epsilon$. Therefore $\|x'\| \leq \frac{\epsilon'}{\epsilon}\frac{1}{\tau_{C^*}}R_\epsilon$ for all $x' \in P'_\epsilon$, and so $R_{\epsilon'} \leq \frac{\epsilon'}{\epsilon}\frac{1}{\tau_{C^*}}R_\epsilon$, proving the result. ∎

Acknowledgment. The author is grateful to the referees for suggesting several improvements in the results as well as in the presentation.

REFERENCES

[1] A. Ben-Israel, A. Charnes, and K. O. Kortanek, *Duality and asymptotic solvability over cones*, Bull. Amer. Math. Soc., 75 (1969), pp. 318–324.

[2] R. J. Duffin, *Infinite Programs*, Ann. of Math. Stud. 38, H.W. Kuhn and A.W. Tucker, eds., Princeton University Press, Princeton, NJ, 1956, pp. 157–170.

[3] R. M. Freund and J. R. Vera, *Condition-based complexity of convex optimization in conic linear form via the ellipsoid algorithm*, SIAM J. Optim., 10 (1999), pp. 155–176.

[4] R. M. Freund and J. R. Vera, *Some characterizations and properties of the "distance to ill-posedness" and the condition measure of a conic linear system*, Math. Program., 86 (1999), pp. 225–260.

[5] J. L. Goffin, *The relaxation method for solving systems of linear inequalities*, Math. Oper. Res., 5 (1980), pp. 388–414.

[6] Y. Nesterov and A. Nemirovskii, *Interior Point Polynomial Algorithms in Convex Programming*, SIAM, Philadelphia, 1994.

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

SIAM J. OPTIM.
Vol. 13, No. 4, pp. 1014–1028

COMPUTATIONAL EXPERIENCE WITH STABLE SET RELAXATIONS*

GERALD GRUBER† AND FRANZ RENDL‡

Abstract. We investigate relaxations for the maximum stable set problem based on the Lovász number $\vartheta(G)$ as an initial upper bound. We strengthen this relaxation by adding two classes of cutting planes, odd circuit and triangle inequalities. We present computational results using this tighter model on many classes of graphs.

Key words. stable set problem, theta function, semidefinite programming

AMS subject classifications. 90C22, 90C27, 90C59

PII. S1052623401394092

1. The stable set problem. Let $G = (V(G), E(G))$ denote an undirected graph with $|V(G)| = n$ and $|E(G)| = m$. An edge joining nodes i and j is represented by (ij). A stable set S (also called an independent set) in G is by definition a subset of $V(G)$ such that no two vertices in S are joined by an edge in $E(G)$. The *maximum stable set problem* is the problem of finding a stable set of maximum cardinality. We will denote the size of a largest stable set in G by $\alpha(G)$.

Some notation. We will denote by $S(n)$ the space of $n \times n$ symmetric matrices and write $\mathrm{diag}(X)$ to denote the vector formed from the diagonal elements of a square matrix X. Let $\mathrm{trace}(X) = \sum_i x_{ii}$. The symbol $\succeq$ denotes the Löwner partial order, i.e., $A \succeq B$ if $A - B$ is positive semidefinite. $A \succ 0$ means A is positive definite. Let e denote the vector of all ones, J the matrix of all ones, and let $E_{ij} = e_i e_j^T + e_j e_i^T$, where e_i stands for the ith column of the identity matrix of appropriate size.

The stable set problem can be formulated in several ways as an integer optimization problem. Let us introduce the binary variables x_i with $x_i = 1$ if i is to be contained in some stable set. Then we get the following integer linear program for the stable set problem:

$$\text{(STAB)}\quad \alpha(G) = \max\left\{\sum_{i=1}^{n} x_i : x_i + x_j \le 1\ \forall (ij) \in E(G),\ x_i \in \{0,1\}\ \forall i \in V(G)\right\}.$$

The linear formulation (STAB) is the starting point for the polyhedral approach to the stable set problem, going back to the 1970s. We will review some of the main results of this approach in subsection 2.1.

The condition that for $(ij) \in E(G)$ at most one of the nodes i and j can be taken into a stable set can be modeled not only as $x_i + x_j \le 1$, $x_i, x_j \in \{0,1\}$, but also as $x_i x_j = 0$, $x_i, x_j \in \{0,1\}$. This leads to the following integer quadratic optimization formulation of the stable set problem:

$$\text{(1.1)}\quad \text{(STAB)}\quad \alpha(G) = \max\{x^T x : x_i x_j = 0\ \forall (ij) \in E(G),\ x_i^2 = x_i\ \forall i \in V(G)\}.$$

*Received by the editors August 23, 2001; accepted for publication (in revised form) August 12, 2002; published electronically March 19, 2003. Partial financial support from the Austrian Science Foundation, FWF project P12660-MAT, is gratefully acknowledged.

http://www.siam.org/journals/siopt/13-4/39409.html

†Carinthia Tech Institute, Department of Geoinformation, Europastrasse 4, A-9524 Villach, Austria (g.gruber@cti.ac.at).

‡Universität Klagenfurt, Institut für Mathematik, Universitätsstraße 65-67, A-9020 Klagenfurt, Austria (franz.rendl@uni-klu.ac.at).

This formulation will give rise to the well-known semidefinite relaxation introduced by Lovász in 1979; see [17]. We summarize the main results of this approach in subsection 2.2. Section 2 reviews linear and semidefinite relaxations of the stable set problem. We recall the interconnection between (2.9) and (2.7) below. In section 2.4 we review the lift-and-project idea. In section 3 we discuss a computationally feasible way to approximate $N_+(FRAC(G))$, and conclude with computational results on many classes of graphs, both taken from the literature and randomly generated.

Our contributions. We first investigate the computational effort required to compute the ϑ function introduced by Lovász. We will show that computing the ϑ function using (2.9) is far more efficient than using (2.7). Unfortunately, it is not clear how the model (2.9) can be combined with purely linear relaxations of the stable set problem, while this is straightforward for the model (2.7). As another contribution, we show how the optimal solution to (2.9) can be used to generate an optimal solution to (2.7), which can be further strengthened by introducing additional cutting planes. Finally, we propose an iterative procedure that leads to a computationally efficient way to approximate a theoretically very strong relaxation of the stable set problem, namely, optimizing over $N_+(FRAC(G))$. Our computational results indicate that this relaxation often gives very tight approximations to $\alpha(G)$.

2. Tractable relaxations.

2.1. Linear relaxations.

Linear programming-based relaxations of a combinatorial optimization problem are based on the study of the convex hull of its integer solutions. Let us denote by $STAB(G)$ the convex hull of incidence vectors of stable sets in G, i.e.,

$$STAB(G) = \text{conv}\{x \in \{0,1\}^n : x_i + x_j \leq 1 \ \forall (ij) \in E(G)\}. \tag{2.1}$$

The following inequalities are all valid for $STAB(G)$:

$$x_i \leq 1 \qquad \text{for isolated nodes,}$$

$$x_i \geq 0, \qquad i \in V(G), \tag{2.2}$$

(nonnegativity constraints)

$$x_i + x_j \leq 1, \qquad (ij) \in E(G), \tag{2.3}$$

(edge constraints)

$$\sum_{i \in V(C)} x_i \ \leq \ \frac{1}{2}(|V(C)| - 1), \qquad C \text{ odd cycle in } G, \tag{2.4}$$

(odd-cycle constraints)

$$\sum_{i \in Q} x_i \leq 1, \qquad Q \text{ clique in } G. \tag{2.5}$$

(clique constraints)

There are many more classes of inequalities which are valid for $STAB(G)$, such as the odd-antihole constraints, the wheel constraints, and so on. We refer to Chapter 9 of [8] for a thorough treatment of linear relaxations of the stable set problem.

Several subsets of the constraint classes above were given names. The fractional stable set polytope is given by

$$FRAC(G) = \{x : x \text{ satisfies (2.2) and (2.3)}\}.$$

The odd-cycle polytope is given by

$$CSTAB(G) = \{x : x \text{ satisfies (2.2), (2.3), and (2.4)}\}.$$

The clique-constraints are collected in the clique-polytope

$$QSTAB(G) = \{x : x \text{ satisfies (2.2) and (2.5)}\}.$$

Here we recall a few of the most fundamental facts which are relevant for our approach. The simplest constraints (2.2) and (2.3) suffice for only the trivial case of bipartite graphs.

FACT 2.1 (see [8]). *$STAB(G) = FRAC(G)$ if and only if G is bipartite and has no isolated nodes.*

While the separation problem for $CSTAB(G)$ can be done in polynomial time via shortest path computations in graphs with nonnegative edge weights, it is likely to be difficult to optimize over $QSTAB(G)$.

FACT 2.2 (see [7]). *Any linear function can be optimized over $CSTAB(G)$ in polynomial time.*

FACT 2.3 (see [6]). *It is NP-complete to optimize a linear objective function over $QSTAB(G)$.*

$STAB(G)$ is in general a proper subset of $QSTAB(G)$. The question of characterizing graphs G, where $QSTAB(G) = STAB(G)$, leads to one of the most intriguing topics of graph theory related to the stable set problem, the study of perfect graphs. For a formal definition of perfect graphs, we refer to [6]. In the 1970s Fulkerson [5] and Chvátal [4] characterized the graphs for which $QSTAB(G) = STAB(G)$. They showed the following.

FACT 2.4. *$STAB(G) = QSTAB(G)$ if and only if G is perfect.*

2.2. Theta function. In this section we recall the stable set relaxation introduced by Lovász; see [17]. There are many different ways to derive this relaxation. Most of them were already analyzed in the original paper [17].

A quick way to obtain this relaxation was proposed by Lovász and Schrijver [19], and goes as follows. Suppose that $x \in \{0,1\}^n$ is the characteristic vector of a stable set in a graph G with $n := |V(G)|$. Consider the rank-one matrix $Y = \binom{x}{1}\binom{x}{1}^T \in S(n+1)$. If we partition this matrix as $Y = \left(\begin{smallmatrix} X & x \\ x^T & 1 \end{smallmatrix}\right)$, then the main diagonal of $X = (x_{ij})$ is contained in the following set $TH(G)$:

$$TH(G) = \left\{ x \in \mathbb{R}^n : \exists Y = \begin{pmatrix} X & x \\ x^T & 1 \end{pmatrix},\ Y \succeq 0,\ \operatorname{diag}(X) = x,\ x_{ij} = 0\ \forall (ij) \in E(G) \right\}. \tag{2.6}$$

Therefore optimizing over $TH(G)$ gives a relaxation of $\alpha(G)$, which we denote by $\vartheta(G)$:

$$\vartheta(G) = \max\left\{ \sum_i x_i : x \in TH(G) \right\} \geq \alpha(G). \tag{2.7}$$

REMARK 2.5. *It is well known that $\left(\begin{smallmatrix} X & x \\ x^T & 1 \end{smallmatrix}\right) \succeq 0$ if and only if $X - xx^T \succeq 0$.*

$\vartheta(G)$ is obtained as the solution of a semidefinite programming (SDP) problem. It can be computed in polynomial time to some fixed prescribed precision; see [7, 22]. It is pointed out in [23] that the relaxation leading to $\vartheta(G)$ can alternatively be derived by taking the Lagrangian dual of the stable set formulation as a quadratically constrained quadratic problem in binary variables (1.1). This is also observed more recently in [16]. The derivation through the Lagrangian dual is quite general, and this approach applies to many other discrete optimization problems; see, e.g., [23]. The

set $TH(G)$ can be viewed as the projection of the set of $(n+1)\times(n+1)$ matrices $Y \in S(n+1)$ satisfying

$$y_{ij} = 0 \text{ for all } (ij) \in E(G),\ y_{ii} = y_{i,n+1} \text{ for all } i \in V(G),\ y_{n+1,n+1} = 1,\ Y \succeq 0 \tag{2.8}$$

onto its main diagonal entries 1 to n. Note also that there are positive definite matrices $Y \in S(n+1)$ satisfying (2.8), so that the SDP defining $\vartheta(G)$ satisfies the Slater constraint qualification.

Even though the derivation of the relaxation (2.7) is quite natural, there are several seemingly different relaxations of $\alpha(G)$, which all give the same value $\vartheta(G)$. We review some basic results which are useful for our purposes.

Let $x \neq 0$ again be the characteristic vector of some stable set in G. Then $e^T x = x^T x$ and the matrix $X := \frac{1}{x^T x} x x^T$ has the following properties:

$$X \succeq 0, \quad x_{ij} = 0 \text{ for all } (ij) \in E(G), \quad \operatorname{trace}(X) = 1.$$

Moreover, $\operatorname{trace}(JX) = e^T x$. Thus

$$z(G) = \max\{\operatorname{trace}(JX) : \operatorname{trace}(X) = 1,\ x_{ij} = 0\ \forall (ij) \in E(G),\ X \succeq 0\} \geq \alpha(G). \tag{2.9}$$

The relaxation $z(G)$ is among the formulations of $\vartheta(G)$ investigated in [17]. It is shown, e.g., in [19] that indeed $z(G) = \vartheta(G)$. Since this result is important for our computational approach, we include the following proof, which differs from the argument in [19]. We start out with some simple observations.

LEMMA 2.6. *Let $X \in S(n)$, $X \succeq 0$, and assume that $\operatorname{trace}(JX) =: z > 0$. Then there exists a matrix $B = (b_1, \ldots)$ such that $X = BB^T$ and $Xe = \sqrt{z} b_1$.*

Proof. If $X \succeq 0$, then there exists C such that $X = CC^T$. We set $x = \frac{1}{\sqrt{z}} C^T e$ and note that $x^T x = \frac{1}{z} e^T C C^T e = \frac{1}{z} \operatorname{trace}(JX) = 1$. Thus there exists an orthogonal matrix $Q = (x, \ldots)$ with x as the first column. The matrix $CQ = B = (b_1, \ldots)$ satisfies $BB^T = CQQ^T C^T = X$ and $Xe = BQ^T C^T e = \sqrt{z} BQ^T x = \sqrt{z} B e_1 = \sqrt{z} b_1$. ∎

LEMMA 2.7. *Let X be optimal for (2.9) with value $z(G) = \operatorname{trace}(JX)$. Then $Xe = z(G)\operatorname{diag}(X) \in TH(G)$.*

Proof. The dual of (2.9) is the following SDP (see, e.g., [12]):

$$\min \lambda \text{ such that } S = \lambda I - J + \sum_{(ij)\in E(G)} y_{ij} E_{ij} \succeq 0.$$

Let X be optimal for (2.9), and $\lambda, \{y_{ij} : (ij) \in E(G)\}$ be optimal for the dual problem. Strong duality holds for these problems, because both problems satisfy the Slater constraint qualification. Therefore $z(G) = \lambda$ and $SX = 0$. Note that $z(G) \geq 1$ since $\frac{1}{n} I$ is feasible in (2.9) with objective value 1. Thus

$$\begin{aligned} 0 = (SX)_{ii} &= s_{ii} x_{ii} + \sum_{(ij)\in E(G)} s_{ij} x_{ij} + \sum_{i\neq j,\,(ij)\notin E(G)} s_{ij} x_{ij} \\ &= (z(G) - 1) x_{ii} - \sum_{i\neq j,\,(ij)\notin E(G)} x_{ij}, \end{aligned}$$

because $s_{ii} = z(G) - 1$, $x_{ij} = 0$ for $(ij) \in E(G)$, and $s_{ij} = -1$ for $(ij) \notin E(G)$. This shows that $z(G) x_{ii} = (Xe)_i$.

From Lemma 2.6 we get $X = BB^T$ and $d := z(G)\operatorname{diag}(X) = Xe = \sqrt{z(G)} b_1$. Therefore $z(G)X - dd^T = z(G)(BB^T - b_1 b_1^T) = z(G)\sum_{i>1} b_i b_i^T \succeq 0$. Moreover, $x_{ij} = 0$ for $(ij) \in E(G)$, showing that $d \in TH(G)$. ∎

LEMMA 2.8. *Let $x \in TH(G)$, i.e., $\exists X = (x_{ij})$, $x = \operatorname{diag}(X)$, $x_{ij} = 0$ if $(ij) \in E(G)$, $X - xx^T \succeq 0$. Suppose that $\sum_i x_i = \gamma > 0$. Then $Y = \frac{1}{\gamma}X$ is feasible for* (2.9) *and* $\operatorname{trace}(JY) \geq \gamma$.

Proof. Feasibility of Y for (2.9) is obvious. Since $X - xx^T \succeq 0$, we get $e^T X e \geq (e^T x)^2 = \gamma^2$. Therefore $e^T Y e \geq \gamma$. ∎

Using these results, we show now how optimal solutions of (2.9) and (2.7) relate to each other.

THEOREM 2.9. *Let $x \in TH(G)$ be optimal for* (2.7), *i.e., $\exists X \in S(n)$, $X = (x_{ij})$ with $x = \operatorname{diag}(X)$, $x_{ij} = 0$ if $(ij) \in E(G)$, $X - xx^T \succeq 0$, and objective value $\vartheta(G) = \sum_i x_i$. Further, let Y be optimal for* (2.9) *with $z(G) = \operatorname{trace}(JY)$. Then*

$$z(G)\operatorname{diag}(Y) \text{ is optimal for } (2.7),$$

$$\frac{1}{\vartheta(G)}X \text{ is optimal for } (2.9) \text{ and } z(G) = \vartheta(G).$$

Proof. Let x, X, Y satisfy the hypotheses of the theorem. Lemma 2.7 shows that $\bar{z} = z(G)\operatorname{diag}(Y)$ is feasible for (2.7) and $z(G) = \sum_i \bar{z}_i \leq \sum_i x_i = \vartheta(G)$. Lemma 2.8 shows that $\frac{1}{\vartheta}X$ is feasible for (2.9) and $\vartheta(G) \leq e^T(\frac{1}{\vartheta(G)}X)e \leq z(G)$. ∎

There is some asymmetry between the two transformations. While it is straightforward to transform feasible solutions from (2.7) into feasible solutions of (2.9) with the same objective value or higher (see Lemma 2.8) it is not clear how arbitrary feasible solutions of (2.9) can be transformed into feasible solutions of (2.7). Lemma 2.7 makes essential use of X being optimal.

We close this section with some basic facts about the set $TH(G)$.

FACT 2.10 (see [6, 7]).

1. $STAB(G) \subseteq TH(G) \subseteq QSTAB(G)$.
2. $TH(G)$ *is polyhedral if and only if* G *is perfect.*
3. $STAB(G) = TH(G) = QSTAB(G)$ *if and only if* G *is perfect.*
4. *We can optimize linear functions over* $TH(G)$ *in polynomial time.*

REMARK 2.11. *It is easy to see that adding the rank-one constraint to matrices satisfying* (2.8) *produces precisely the characteristic vectors of stable sets.*

$$\{Y : Y \text{ satisfies } (2.8),\ \operatorname{rank}(Y) = 1\}$$
$$= \left\{ \begin{pmatrix} x \\ 1 \end{pmatrix} \begin{pmatrix} x \\ 1 \end{pmatrix}^T : x \text{ characteristic vector of a stable set in } G \right\}.$$

We leave it to the reader to verify that adding the rank-one constraint to the relaxation (2.9) *again gives* $\alpha(G)$.

$$\alpha(G) = \max\{\operatorname{trace}(JX) : \operatorname{trace}(X) = 1,\ x_{ij} = 0\ \forall (ij) \in E(G),\ X \succeq 0,\ \operatorname{rank}(X) = 1\}.$$

It should be noted, however, that not all feasible rank-one matrices of (2.9) *lead to multiples of characteristic vectors of stable sets.*

2.3. Optimizing over $TH(G)$ and $CSTAB(G)$ in practice. In view of Theorem 2.9, we have two SDP models available to compute $\vartheta(G)$. Theoretically, they are equivalent, but the practical computational effort is much smaller for (2.9). This should not be too surprising as there are, aside from the constraints $x_{ij} = 0$, $(ij) \in E(G)$, which appear in both models, $n+1$ additional equations fixing the main diagonal of Y to its last column in (2.8), while there is only one additional equation $\operatorname{trace}(X) = 1$ in (2.9).

TABLE 2.1
Comparison of running times for computing $\vartheta(G)$.

				CPU time [sec]		
File	n	m	$\vartheta(G)$	(2.9)	(2.7)	CSDP 3.2
g100	100	497	32.8792	7	68	-
g150	150	1105	42.1660	43	645	70
g200	200	1948	50.3212	176	4172	311
g250	250	3027	57.2323	589	no attempt	956
g300	300	4374	63.4471	1524	no attempt	2929

TABLE 2.2
Numerical results on triangulated planar graphs.

			optimizing over			
n	m	$\alpha(G)$	$TH(G)$	CPU time [sec]	$CSTAB(G)$	CPU time [sec]
100	294	44	44	2.3	44.6667	0.9
200	594	87	87	16.6	87.3333	4.5
300	894	126	126	53.7	127.0000	13.0
400	1194	174	174	155.8	174.6667	29.7
500	1494	220	220	335.4	220.3333	64.3
600	1794	270	270	606.9	270.0000	59.0

In Table 2.1 we include computation times for both models and also compare our results with those from the software package CSDP 3.2 from Brian Borchers[1] written in C, which also contains a module to compute (2.9). The test problems are taken from this software package. The column labeled n gives the number of vertices, column m the number of edges, in the graph G. Both relaxations are computed with Matlab using some interfaces in C; CSDP 3.2 is written in C. The computations for (2.9) and (2.7) are done on a Pentium II 400 MHz computer, the results for CSDP 3.2 are taken from Mittelmann's web page,[2] where he reports computation times using a Pentium II 450 MHz computer.

Since both CSDP 3.2 and our code for (2.9) solve the same problem, the difference in computation times is attributed to the highly efficient implementation of matrix operations in Matlab. To allow a comparison, our Matlab routine is available on the web.[3]

We also report results from optimizing over the linear relaxation $CSTAB(G)$ of the stable set polytope. ILOG CPLEX 6.5 is used to solve the linear programs. Note that $x = \frac{1}{3}e$, $x \in \mathbb{R}^n$, is contained in $CSTAB(G)$. Therefore the optimal objective value is at least $\frac{n}{3}$. In our extensive experiments it has been observed that optimizing over $CSTAB(G)$ often yields $x = \frac{1}{3}e$ if $\alpha(G) < \frac{n}{3}$. Therefore optimizing over $CSTAB(G)$ is only meaningful if $\alpha(G) > \frac{n}{3}$. In Table 2.2 we compare the computational effort to optimize over $TH(G)$, which involves solving an SDP, with the effort to solve the linear relaxation given by $CSTAB(G)$. We use triangulated planar graphs, which are known to be perfect, hence $\alpha(G) = \vartheta(G)$. It is interesting to see that the relaxation

[1] http://www.nmt.edu/~borchers/csdp.html
[2] ftp://plato.la.asu.edu/pub/sdplib.txt
[3] http://www-sci.uni-klu.ac.at/math-or/home/publications/theta-ml.tar.gz

given by *CSTAB* is also rather strong for this class of problems.

2.4. Matrix relaxations: Lift-and-project. The derivation of $\vartheta(G)$ can be viewed as going from $\mathbb{R}^n$ into the space of symmetric $n \times n$ matrices, and then back to $\mathbb{R}^n$ again, by projecting onto the main diagonal. This idea is further elaborated in [19] and was independently used by many other researchers; see, e.g., [1, 18, 26]. For recent developments on lift-and-project, see Laurent [14]. We now recall the approach investigated in [19]. Given a polytope $P = \{x : Ax - b \geq 0\}$, let P_I denote the convex hull of its 0-1 integral points. Multiplying each inequality defining P by x_k and $1 - x_k$ results in a system of quadratic inequalities. This system is linearized by substituting y_{ij} for $x_i x_j$. 0-1 solutions also satisfy $x_i^2 = x_i$; hence x_i^2 is replaced by y_{ii}. Imposing further that $y_{ij} = y_{ji}$ leads to inequalities which are now linear in y_{ij}. We denote by $M(P)$ the set of all symmetric matrices Y satisfying these linearized constraints. Projection onto the main diagonal gives the set $N(P)$,

$$N(P) = \{x : \exists Y \in M(P) \text{ such that } x = \operatorname{diag}(Y)\}.$$

It is clear from the definition that $P_I \subseteq N(P) \subseteq P$. We now recall the results of [19] in the case $P = FRAC(G)$. Multiplying $x_i \geq 0$ by $x_k \geq 0$ and $1 - x_k \geq 0$ and linearizing gives

$$0 \leq y_{ik} \leq y_{ii}.$$

Similarly we multiply $1 - x_i - x_j \geq 0$ with x_k and $1 - x_k$ and get for all $(ij) \in E(G)$ and for all k

$$y_{kk} - y_{ik} - y_{jk} \geq 0 \quad \text{and} \quad 1 - y_{ii} - y_{jj} - y_{kk} + y_{ik} + y_{jk} \geq 0.$$

It should be noted that $y_{ik} \leq y_{kk}$ is implied by the second set of inequalities. Moreover, it is easy to see that $y_{ij} = 0$ for $(ij) \in E(G)$. We conclude (see [19])

$$M(FRAC(G)) = \Big\{ Y = (y_{ij}) : Y \geq 0,\ y_{ii} + y_{jj} + y_{kk} - 1 \leq y_{ik} + y_{jk} \leq y_{kk} \\ \forall (ij) \in E(G),\ \forall k \Big\}.$$

Since $M(FRAC(G))$ is given by an explicit system of inequalities (of polynomial size), we can optimize in polynomial time over $N(FRAC(G))$.

We close this section with yet another view of lift-and-project. It is well known that the quadric Boolean polytope $\operatorname{conv}\{xx^T : x \in \{0,1\}^n\}$ is contained in the metric polytope $MET(n)$; see, e.g., [2, 27]. By definition, a matrix X satisfies

$$(2.10)\quad X \in MET(n) :\Leftrightarrow \begin{cases} 0 \leq x_{ij} \leq x_{ii}, \\ x_{ii} + x_{jj} - x_{ij} \leq 1, \\ -x_{kk} - x_{ij} + x_{ik} + x_{jk} \leq 0, \\ x_{kk} + x_{ii} + x_{jj} - x_{ij} - x_{ik} - x_{jk} \leq 1,\ 1 \leq i, j, k \leq n. \end{cases}$$

Let us define $MET(G)$ to consist of those matrices in $MET(n)$ where $x_{ij} = 0$ for $(ij) \in E(G)$,

$$MET(G) = \{X \in MET(n) : x_{ij} = 0 \text{ if } (ij) \in E(G)\}.$$

We observe that $MET(G) \subseteq M(FRAC(G))$, because $M(FRAC(G))$ contains all the inequalities of $MET(G)$ for which $(ij) \in E(G)$. On the other hand, $MET(G)$ contains

the additional inequalities $-x_{kk}-x_{ij}+x_{ik}+x_{jk} \leq 0$ and $x_{kk}+x_{ii}+x_{jj}-x_{ij}-x_{ik}-x_{jk} \leq 1$ for stable sets i, j, k, which clearly are not contained in $M(FRAC(G))$. It turns out, however, that these constraints do not further tighten the relaxation.

THEOREM 2.12.

$$N(FRAC(G)) = CSTAB(G) = N(MET(G)).$$

We omit a formal proof, as the result follows from Theorem 2.3 in [19] and Proposition 6 from [15]. We note, however, that the inclusions

$$N(MET(G)) \subseteq N(FRAC(G)) \subseteq CSTAB(G)$$

can be easily verified: the first one follows from the fact that $MET(G) \subseteq M(FRAC(G))$, and the second one is verified in [19]. Proposition 6 from [15] shows that a vector d belongs to $CSTAB(G)$ if and only if there exists a matrix $X \in MET(G)$ such that $d_i = x_{ii}$ for all $i \in V$. This implies that $CSTAB(G)$ is contained in $N(MET(G))$ and, therefore, equality holds throughout in the above chain of inclusions. It is worth mentioning this since this gives an alternative proof for the inclusion $CSTAB(G) \subseteq N(MET(G))$, which is the hardest part of the result from [19].

Lovász and Schrijver propose a further refinement of $M(P)$ and $N(P)$ by observing that matrices Y in $M(P)$ satisfy the following semidefiniteness condition:

$$Y - \operatorname{diag}(Y) \operatorname{diag}(Y)^T \succeq 0. \tag{2.11}$$

Hence they define $M_+(P) := \{Y \in M(P) : Y \text{ satisfies (2.11)}\}$, and similarly $N_+(P)$ to be the projection of $M_+(P)$ onto the main diagonal.

In case of $P = FRAC(G)$ we get

$$M_+(FRAC(G)) = \{Y : Y \in M(FRAC(G)),\ Y \text{ satisfies (2.11)}\}. \tag{2.12}$$

Optimizing over $N_+(FRAC(G))$ can still be done in polynomial time [19], but practical implementations are currently not available. Optimizing over $N_+(FRAC(G))$ amounts to solving an SDP with $m = |E(G)|$ equality constraints ($x_{ij} = 0$ $(ij) \in E(G)$) and $O(mn)$ inequalities, given by $M(FRAC(G))$. Solving this SDP directly with current algorithms is therefore out of reach for graphs of interesting sizes, say $n \geq 30$, $m \geq 100$.

In view of Theorem 2.12 it is clear that when optimizing over $N_+(FRAC(G))$, we can ignore the odd cycle constraints, as they are automatically satisfied. We will show below that these constraints are nonetheless quite useful in the case where we approximate the optimization over $N_+(FRAC(G))$ by an iterative procedure, where only some of the linear inequalities defining $N_+(FRAC(G))$ are considered.

3. Towards approximating $N_+(FRAC(G))$. In Table 2.1 we gave some idea about the problem sizes for which it is feasible to optimize over $TH(G)$ yielding $\vartheta(G)$. Now we are interested in investigating by how much we can improve $\vartheta(G)$, when (at least approximately) optimizing over $N_+(FRAC(G))$. Directly including all $O(nm)$ inequalities from (2.12) in the SDP is intractable for graphs of reasonable size. Therefore we use the approach from [12, 11], where an iterative scheme is proposed in which only a small subset of (2.10) is selected carefully to be added to

$$F(G) := \{X \in S(n) : \operatorname{diag}(X) \in TH(G)\}.$$

TABLE 3.1
Average running times on randomly generated sparse graphs.

		CPU time [sec] from optimizing over				
n	*density*[%] $-$ m	$F(G)$	$F(G) \cap C(G)$		$F(G) \cap C(G) \cap MET(G)$	
50	10 - 125	0.7	4.3	(10)	8.8	(8/47)
100	10 - 500	5.1	41.4	(9)	141.8	(2/180)
150	5 - 559	10.1	150.7	(42)	354.3	(22/266)
200	3 - 597	16.1	309.6	(94)	958.8	(63/383)
250	2 - 623	25.5	424.7	(95)	1458.3	(98/297)

Since we are interested in tightening the relaxation $TH(G)$ by adding further linear inequalities, we need a model in which those constraints can conveniently be included. In the case of (2.7) this is straightforward. It is not clear how the odd circuit constraints, for instance, or the inhomogeneous triangle inequalities from (2.10) could be included in (2.9). We therefore use the fast model (2.9) to generate the optimal solution X which we transform with Theorem 2.9 into an optimal solution of (2.7). Starting from this point, we include further violated cutting planes, if there are any.

In [11] some preliminary computational experience is reported for optimizing over $F(G) \cap MET(G)$. The main message seems to be that it is computationally very demanding to optimize over this set, as the number of inequalities defining $MET(G)$ grows cubically with n.

It should be mentioned that Schrijver [25] proposed a strengthening of the Lovász bound $\vartheta(G)$ by adding the nonnegativity constraints. These are only a subclass of the constraints of $MET(G)$, and there does not seem to be any reason to prefer them over the general constraints defining $MET(G)$. Optimizing over $F(G) \cap MET(G)$ clearly dominates the Schrijver model.

In our approach one refinement of $TH(G)$ with problem-specific cutting planes consists of intersecting $F(G)$ with

$$C(G) := \{X \in S(n) : \operatorname{diag}(X) \in CSTAB(G)\}.$$

We recall that the separation problem for $CSTAB(G)$ reduces to shortest path computations and hence is easy.

Ultimately we want to get improved bounds on $\alpha(G)$ as quickly as possible. Practical experience has shown that, at the beginning of our iterative process, the inclusion of a moderate number of triangles and odd circuits induces a significant improvement of the bound on $\alpha(G)$, contrary to adding the most violated triangles only. The cycle inequalities get redundant as more and more triangle inequalities are included. Further computational details on this effect can be found in [9].

To get a feeling for the computational effort to solve these relaxations (at least approximately), we include in Table 3.1 some computation times on several randomly generated problems of various size. This table compares computation times from optimizing over the different relaxations. Column n gives the respective order of the graphs. Each line lists the average CPU time to solve 10 problems of given size and density. The values in the parenthesis indicate the (average number of) active constraints at termination. We also provide the average number m of edges and the density. It should be noted that optimizing over $MET(G)$ is not done exactly. The iterations are stopped once the maximum violation of $MET(G)$ has fallen below a

threshold of 50% of the maximum initial violation. Optimizing over $F(G) \cap C(G)$ is computationally feasible and so is done exactly.

4. Interior-point implementation. In this section we discuss some implementation details of our cutting plane technique. Our algorithm involves solving a number of semidefinite problems. These programs are solved using primal-dual path-following interior-point methods. The search direction used in our implementation was proposed independently by Helmberg et al. [12] and Kojima, Shindoh, and Hara [13]; see also [21]. All our codes are implemented in Matlab with interfaces in C.

4.1. Iteratively adding cuts. As already indicated in subsection 2.2, we first compute the optimal solution of (2.9) to generate an optimal solution of (2.7), which can be used to add violated triangle and odd cycle constraints. Thus our initial upper bound on $\alpha(G)$ is always $\vartheta(G)$.

The initial semidefinite bound $\vartheta(G)$ (in general the upper bound of the current model) can be improved now by identifying and including violated odd circuit constraints and/or triangle inequalities. In general there are a lot more violated cutting planes than we are willing to add to the current relaxation. The computational effort to reoptimize grows drastically with the number of constraints. Therefore we add only the most violated constraints.

Since interior-point techniques are used for solving the relaxations, we have to assure that the initial starting point for the subsequent iteration is an interior point. It does not have to be feasible, though.

Hence the current point has to be modified for restarting. A detailed description of this modification is given in section 4.2.

After having added new violated constraints, the interior-point code is restarted. We stop it after it has reached both primal and dual feasibility. At this point we look again for violated constraints, add them, and iterate as before.

After several such rounds of collecting inequalities, we solve the present model to optimality. This allows us to remove constraints that have become inactive.

After having purged the model from inactive constraints, we proceed to the next phase of adding cutting planes, until some stopping condition is satisfied. We refer the reader to [9, Chapter 5] for extensive implementation details.

4.2. Restarting. Let the linear operator $\mathcal{B}(\cdot)$ be defined such that all added cutting planes with complying right-hand side b may be written as $\mathcal{B}(X) \leq b$. The current optimal X is no longer an interior point, because it violates the most recently added cuts. Since the interior-point machinery requires an initial $X \succ 0$ that strictly satisfies these inequalities, we use a simple trick in formulating the current semidefinite model. We introduce a primal slack variable $s = b - \mathcal{B}(X) \geq 0$. We then introduce an artificial dual slack variable $t \geq 0$ for the dual variable r corresponding to s. This way we get the additional dual constraint $t = r$. This restatement of the semidefinite model simplifies the update of the variables used in the interior-point code. To get the initial variables $(X^\circ, y^\circ, Z^\circ)$ we implemented a simple backtracking strategy from the current (X, y, Z) in the direction to a well centered point in the interior of the feasible set of (2.7). This is a standard technique, and therefore we do not restate it here precisely. The initial variables r°, s°, and t° are constructed as follows. We set $(r_i, s_i, t_i) = (0, 0.1, 0.1)$ for every component i corresponding to a newly added inequality. We take the current variables r, s, and t for updating the remaining

components. We set

$$r_i^\circ = \begin{cases} r_i & \text{if } r_i \geq 0.1, \\ 0 & \text{otherwise,} \end{cases} \quad s_i^\circ = \begin{cases} s_i & \text{if } s_i \geq 0.1, \\ 0.1 & \text{otherwise,} \end{cases} \quad t_i^\circ = \begin{cases} t_i & \text{if } t_i \geq 0.1, \\ 0.1 & \text{otherwise.} \end{cases}$$

The point constructed in this way is strictly interior but not feasible for the current semidefinite model.

4.3. Rounding. Finally, we use the main diagonal of our primal matrix X to generate a stable set. A large entry x_{ii} indicates that vertex i is likely to be contained in a large stable set. Hence we run through all the vertices of G in the order given by the main diagonal of X sorted nonincreasingly. We select the vertices to be contained in a stable set with probability proportional to x_{ii}. If vertex j is accepted to be in a stable set, we exclude all the neighbors of j. The final stable set is locally improved in a second phase of the rounding heuristic; the complete process is repeated several times. We include the size of the largest stable set found this way in the tables of section 5 under the heading "lower bound on $\alpha(G)$."

5. Computational results. In this section we report computational experience on different classes of test problems. It is the primary objective of this section to compare the quality of the different relaxations described above. We have already discussed the computational effort in section 3. Therefore running times are omitted. All our computational tests were performed on a DEC Alpha workstation.

5.1. Test problems. Our numerical experiments were carried out on the following sets of graphs:

(i) random graphs with given density,
(ii) Sanchis graphs (see [24, 10]),
(iii) Mannino graphs (see [20]),
(iv) Johnson graphs (see [10]),
(v) triangulated planar graphs,
(vi) graphs with high girth,
(vii) geometric graphs.

Sanchis graphs are hard instances for the minimum vertex cover problem (and hence for the equivalent maximum clique problem). Originally they were designed to test algorithms for these problems. These graphs are constructed in such a way that $\alpha(G)$ is known; see [24]. The independence number is also known for the Mannino graphs and the Johnson graphs. The Mannino graphs are obtained from the set covering formulation of the Steiner triple problem [20], and the Johnson graphs emerge in coding problems [10]. For our tests on Sanchis, Mannino, and Johnson graphs we take the complement graphs and indicate this by the suffix the _c in the tables. These graphs are part of the DIMACS benchmark instances and can be downloaded from the web.

Triangulated planar graphs are perfect, and thus $\vartheta(G) = \alpha(G)$; see [8]. To generate graphs with high girth, we start with a random graph G and randomly replace edges of G by odd paths of prescribed maximum length.

Geometric graphs are constructed as follows. First we randomly generate points in the unit square. If the Euclidean distance between two points is less than a prescribed distance δ, then there is an edge connecting these two points.

5.2. Computations. The computational results on these problem sets are reported in Tables 5.1–5.4. Column n gives the number of vertices, column m the

TABLE 5.1

Numerical results on randomly generated sparse graphs. The relative error in % with respect to the values given in column lbd on $\alpha(G)$ is separated from the upper bound by a dash. The numbers given in parenthesis represent the number of active inequality constraints of the final relaxation. Two numbers separated by a slash within one parenthesis show remaining odd circuit and triangle inequalities separately.

		lbd on	optimizing over				
n	m	$\alpha(G)$	$F(G)$	$F(G) \cap C(G)$		$F(G) \cap C(G) \cap MET(G)$	
50	251	14*	15.0016 - 7.15	14.9990 - 7.14	(1)	14.6960 - 4.97	(1/130)
80	320	27*	28.1924 - 4.42	28.1053 - 4.09	(20)	27.5501 - 2.04	(8/315)
100	393	35	37.0451 - 5.84	36.8770 - 5.36	(21)	36.2010 - 3.43	(7/320)
100	434	33	35.1823 - 6.61	35.0532 - 6.22	(21)	34.3682 - 4.15	(3/361)
150	451	59	62.2523 - 5.51	61.4944 - 4.23	(77)	60.6595 - 2.81	(37/694)
150	552	55	58.0287 - 5.51	57.6335 - 4.79	(49)	56.5200 - 2.76	(21/710)
170	383	77*	79.3290 - 3.02	77.6756 - 0.88	(60)	77.5463 - 0.71	(44/318)
170	423	75*	76.7735 - 2.36	75.6992 - 0.93	(56)	75.4973 - 0.66	(40/305)
200	418	93*	94.5638 - 1.68	93.2256 - 0.24	(65)	93.0111 - 0.01	(52/401)
200	563	81	85.5002 - 5.56	84.3313 - 4.11	(99)	83.1449 - 2.65	(39/1059)
300	894	121	129.3327 - 6.89	127.7402 - 5.57	(153)	126.4218 - 4.48	(55/1195)
350	994	146	152.5927 - 4.52	150.3067 - 2.95	(183)	150.1405 - 2.84	(122/213)
400	835	191*	193.4866 - 1.30	191.0000 - 0.00	(132)	191.1058 - 0.06	(72/5)

TABLE 5.2

Numerical results on Sanchis graphs, Mannino graphs, and Johnson graphs.

						optimizing over
File	n	m	$\alpha(G)$	BPG	lbd on $\alpha(G)$	$F(G)$
san200_0.9_1.clq_c	200	1990	70	46	70	70.0000
san200_0.9_2.clq_c	200	1990	60	36	60	60.0000
san200_0.9_3.clq_c	200	1990	44	33	44	44.0000
sanr200_0.9.clq_c	200	2037	-	41	42	49.2735
MANN_a9.clq_c	45	72	16	16	16	17.4750
MANN_a27.clq_c	378	702	126	121	126	132.7629
MANN_a45.clq_c	1035	1980	345	336	343	356.0488
johnson8-2-4.clq_c	28	168	4	-	4	4.0
johnson8-4-4.clq_c	70	560	14	-	14	14.0
johnson8-16-2-4.clq_c	120	1680	8	-	8	8.0

number of edges in the graph G. The column labeled *lbd on* $\alpha(G)$ stands for "lower bound on $\alpha(G)$." This column gives the size of the largest stable set which we have found using our heuristic method sketched in the previous section. The remaining columns contain the upper bounds found by our code. In addition to the value of the upper bound, we also provide the gap in % with respect to the values given in column *lbd on* $\alpha(G)$.

We first note that for triangulated planar graphs, which are perfect, the computation of α is easy; see Table 2.2. The running times to obtain $\vartheta(G)$ are reasonably moderate for these graphs but still much larger than the time needed to optimize over

TABLE 5.3
Numerical results on graphs with long odd circuits. The relative error in % with respect to the values given in column lbd on $\alpha(G)$ is separated from the upper bound by a dash. The numbers given in parenthesis represent the number of active inequality constraints of the final relaxation. Two numbers separated by a slash within one parenthesis show remaining odd circuit and triangle inequalities separately.

		lbd on	optimizing over				
n	m	$\alpha(G)$	$F(G)$	$F(G) \cap C(G)$		$F(G) \cap C(G) \cap MET(G)$	
118	152	57*	58.6107 - 2.83	57.4000 - 0.70	(18)	57.3252 - 0.57	(14/83)
292	365	141*	144.2935 - 2.34	142.6667 - 1.18	(86)	141.9742 - 0.69	(10/181)
280	394	134*	137.1680 - 2.36	136.6476 - 1.98	(94)	134.4599 - 0.34	(3/215)
282	339	136	139.5288 - 2.59	137.6923 - 1.24	(19)	137.7490 - 1.29	(16/120)
372	534	179*	181.2776 - 1.27	179.0000 - 0.00	(96)	179.0000 - 0.00	(104/370)
448	465	220	223.8245 - 1.74	222.3333 - 1.06	(20)	222.3333 - 1.06	(19/1)
520	947	245*	250.1892 - 2.12	250.1473 - 2.10	(7)	245.9738 - 0.40	(0/354)
544	688	264	269.0532 - 1.91	267.0000 - 1.14	(104)	266.1425 - 0.81	(24/254)

TABLE 5.4
Numerical results on geometric graphs.

			lbd on	optimizing over
n	m	δ	$\alpha(G)$	$F(G)$
50	128	0.2	16	16.0000
50	476	0.4	7	7.0000
50	698	0.6	4	4.0000
100	185	0.1	39	39.0000
100	561	0.2	21	21.0000
100	1065	0.3	13	13.0000
150	327	0.1	53	53.0000
150	1137	0.2	22	22.0000
150	2358	0.3	13	13.3710

$CSTAB(G)$. We note that already the linear bound on $\alpha(G)$ is often tight.

The results in Tables 5.1 and 5.3 illustrate that optimizing over $F(G) \cap C(G)$ leads to substantial improvements of $\vartheta(G)$. This optimization is carried out exactly; i.e., no cycle inequalities are violated at the final solution. The inclusion of additional inequalities from $MET(G)$ leads to further improvements, as can be seen from the last columns in these tables; however, these are associated with an enormous increase of computational cost (see Table 3.1). We did not iterate until all triangle inequalities were satisfied, because this is computationally too involved. The results in the column labeled $F(G) \cap C(G) \cap MET(G)$ are obtained by the following stopping conditions. First, all cycle inequalities are satisfied. Second, the maximum violation of triangle inequalities has fallen below a threshold of .03 (in the test sets of dimension 50 to 300 in Table 5.1). We also provide the number of active odd cycle and triangle constraints at termination. It should be noted that only a few odd cycle constraints are active in the final solution. We have found the optimal solution for many of the test problems with this approach. This is denoted by an asterisk in column *lbd on* $\alpha(G)$.

In Table 5.2 we look at some of the DIMACS test graphs. It is interesting to see

that the class of Sanchis graphs seems to be easy to approximate through $\vartheta(G)$. This is in contrast to several heuristic approaches, which have difficulties in finding large stable sets in these graphs; see [3]. Table 5.2 also shows our experiments on the class of Mannino graphs. We observed that no odd circuits get violated by the optimal $\bar{X}$ to (2.7). Hence the relaxation $TH(G)$ cannot be further strengthened by considering the odd circuits. The violation of the triangle inequalities is so small that a substantial improvement of the upper bound on $\alpha(G)$ through including the violated ones may not be expected. Therefore our experiments on the class of Mannino graphs are restricted to optimizing over $TH(G)$. Anyhow, in the leading two cases our rounding heuristic applied to the optimal $\bar{X}$ to (2.7) yields a stable set with cardinality $\alpha(G)$. The size of the largest stable set in the remaining instance is also very close to $\alpha(G)$. This is again an improvement compared to the heuristic approaches proposed in [3]. In Table 5.2 the column headed with BPG contains the size of the largest clique obtained by Bomze, Pelillo, and Giacomini [3]. Finally, Table 5.2 also shows that optimizing over $TH(G)$ gives tight bounds on $\alpha(G)$ for the Johnson graphs.

We ran extensive tests on geometric graphs of different sizes and density. It turned out that $\vartheta(G)$ was tight in all instances. We present a choice of test results for this class of graphs in Table 5.4. The column headed by δ in this table contains the distances needed for the construction of these graphs; see section 5.1.

6. Concluding comments. In this paper, we have presented a semidefinite interior-point cutting plane approach for approximating the maximum stable set problem. In section 2.2 we have investigated two different semidefinite relaxations for the maximum stable set problem. We have provided a simple argument for the well known fact that these two models are equivalent. In particular, we have shown how optimal solutions from one formulation can be transformed into optimal solutions of the other formulation.

Our computational tests give rise to the following conclusions. The essential message is that the process of adding and dropping cutting planes works robustly and can significantly improve the quality of the final solution. This was demonstrated on many different categories of test instances. Optimizing over $F(G) \cap C(G) \cap MET(G)$ produces the widest progress. The computational effort, however, increases dramatically with the number of cutting planes. Moreover, no practically efficient method is known to optimize exactly over this set, even though in principle this can be done in polynomial time. In contrast, optimizing over $F(G) \cap C(G)$ can be done exactly and fast. Optimizing over this set is not much harder than computing $\vartheta(G)$, which can be done routinely for graphs with up to several thousand edges.

A disadvantage of our approach is the large memory requirement for large scale problems. This comes from the Cholesky factorization used in our implementation. A promising direction for future research would be to examine the relaxations in the context of branch-and-bound methods or bundle methods.

Acknowledgments. We thank the referees and the associate editor for the thoughtful comments and constructive suggestions to get this paper into its present form.

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] F. Barahona, *On cuts and matchings in planar graphs*, Math. Programming, 36 (1993), pp. 53–68.
[3] I. M. Bomze, M. Pelillo, and R. Giacomini, *Evolutionary approach to the maximum clique problem: Empirical evidence on a larger scale*, in Developments in Global Optimization, I. M. Bomze, T. Csendes, R. Horst, and P. M. Pardalos, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1997, pp. 95–108.
[4] V. Chvátal, *On certain polytopes associated with graphs*, J. Combin. Theory Ser. B, 18 (1975), pp. 138–154.
[5] D. R. Fulkerson, *On the perfect graph theorem*, in Mathematical Programming, T. C. Hu and S. M. Robinson, eds., Academic Press, New York, 1973, pp. 69–76.
[6] M. Grötschel, L. Lovász, and A. Schrijver, *The ellipsoid method and its consequences in combinatorial optimization*, Combinatorica, 1 (1981), pp. 169–197.
[7] M. Grötschel, L. Lovász, and A. Schrijver, *Relaxations of vertex packing*, J. Combin. Theory Ser. B, 40 (1986), pp. 330–343.
[8] M. Grötschel, L. Lovász, and A. Schrijver, *Geometric Algorithms and Combinatorial Optimization*, Springer-Verlag, New York, 1988.
[9] G. Gruber, *On Semidefinite Programming and Applications in Combinatorial Optimization*, Ph.D. thesis, University of Technology, Graz, Austria, Shaker Verlag, Aachen, Germany, Maastricht, The Netherlands, 2000.
[10] J. Hasselberg, P. M. Pardalos, and G. Vairaktarakis, *Test case generators and computational results for the maximum clique problem*, J. Global Optim., 3 (1993), pp. 463–482.
[11] C. Helmberg, S. Poljak, F. Rendl, and H. Wolkowicz, *Combining semidefinite and polyhedral relaxations for integer programs*, in Integer Programming and Combinatorial Optimization (Proceedings of the 4th International IPCO Conference, Copenhagen, Denmark, 1995), Lecture Notes in Comput. Sci. 920, E. Balas and J. Clausen, eds., Springer-Verlag, Berlin, 1995, pp. 124–134.
[12] C. Helmberg, F. Rendl, R. J. Vanderbei, and H. Wolkowicz, *An interior-point method for semidefinite programming*, SIAM J. Optim., 6 (1996), pp. 342–361.
[13] M. Kojima, S. Shindoh, and S. Hara, *Interior-point methods for the monotone semidefinite linear complementarity problem in symmetric matrices*, SIAM J. Optim., 7 (1997), pp. 86–125.
[14] M. Laurent, *A Comparison of the Sherali–Adams, Lovász–Schrijver and Lasserre Relaxations for* 0-1 *Programming*, Technical report, Centrum voor Wiskunde en Informatica (CWI), Amsterdam, 2001.
[15] M. Laurent, S. Poljak, and F. Rendl, *Connections between semidefinite relaxations of the max-cut and stable set problems*, Math. Programming, 77 (1997), pp. 225–246.
[16] C. Lemaréchal and F. Oustry, *Semidefinite Relaxations and Lagrangian Duality with Application to Combinatorial Optimization*, Technical report, Institut National de Recherche en Informatique et en Automatique (INRIA), St. Martin, France, 1999.
[17] L. Lovász, *On the Shannon capacity of a graph*, IEEE Trans. Inform. Theory, 25 (1979), pp. 1–7.
[18] L. Lovász, *Semidefinite Programs and Combinatorial Optimization*, lecture notes, 1995.
[19] L. Lovász and A. Schrijver, *Cones of matrices and set-functions and* 0-1 *optimization*, SIAM J. Optim., 1 (1991), pp. 166–190.
[20] C. Mannino and A. Sassano, *Edge projection and the maximum cardinality stable set problem*, in Cliques, Coloring, and Satisfiability, DIMACS Ser. Discrete Math. Theoret. Comput. Sci. 26, D. S. Johnson and M. A. Trick, eds., AMS, Providence, RI, 1996, pp. 205–219.
[21] R. D. C. Monteiro, *Primal–dual path-following algorithms for semidefinite programming*, SIAM J. Optim., 7 (1997), pp. 663–678.
[22] Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM Stud. Appl. Math. 13, SIAM, Philadelphia, 1994.
[23] S. Poljak, F. Rendl, and H. Wolkowicz, *A recipe for semidefinite relaxation for* $(0,1)$*-quadratic programming*, J. Global Optim., 7 (1995), pp. 51–73.
[24] L. A. Sanchis, *Test case construction for the vertex cover problem*, in Computational Support for Discrete Mathematics, DIMACS Ser. Discrete Math. Theoret. Comput. Sci. 15, AMS, Providence, RI, 1994, pp. 1052–1798.
[25] A. Schrijver, *A comparison of the Delsarte and Lovász bounds*, IEEE Trans. Inform. Theory, 25 (1979), pp. 425–429.
[26] 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.
[27] C. De Simone, *The cut polytope and the Boolean quadric polytope*, Discrete Math., 79 (1989), pp. 71–75.

SIAM J. OPTIM.
Vol. 13, No. 4, pp. 1029–1053

THE ANALYTIC CENTER CUTTING PLANE METHOD WITH SEMIDEFINITE CUTS*

MOHAMMAD R. OSKOOROUCHI† AND JEAN-LOUIS GOFFIN‡

Abstract. We analyze an analytic center cutting plane algorithm for convex feasibility problems with semidefinite cuts. The problem of interest seeks a feasible point in a bounded convex set, which contains a full-dimensional ball with ε (< 1) radius and is contained in a compact convex set described by matrix inequalities, known as the set of localization. At each iteration, an approximate analytic center of the set of localization is computed. If this point is not in the solution set, an oracle is called to return a p-dimensional semidefinite cut. The set of localization is then updated by adding the semidefinite cut through the center. We prove that the analytic center is recovered after adding a p_k-dimensional semidefinite cut in $O(p_k \log(p_k + 1))$ damped Newton's iteration and that the algorithm stops with a point in the solution set when the dimension of the accumulated block diagonal cut matrix reaches the bound of $O^*(\frac{p^2 m^3}{\mu^2 \varepsilon^2})$, where p is the maximum dimension of the cut matrices and $\mu > 0$ is a condition number of the field of cuts.

Key words. analytic center, cutting plane method, semidefinite programming, semidefinite cut

AMS subject classifications. 90C22, 90C25

PII. S1052623400374148

1. Introduction.

Semidefinite relaxations arising from combinatorial applications can often be too large to be handled by classical interior point methods. On the other hand, many such problems are well structured and have sparse matrix coefficients. Some algorithms that exploit the sparsity of problems of these types have been developed in the past few years. Benson, Ye, and Zhang [3] propose a dual scaling algorithm for the problems with rank one matrix coefficients. Helmberg and Rendl [11] transform the dual semidefinite problem into an eigenvalue optimization problem and apply a spectral bundle method to solve it as a convex nondifferentiable optimization problem in the cone of semidefinite matrices. The idea of the latter paper is of interest to us.

An alternative technique for nonsmooth optimization is the analytic center cutting plane method (ACCPM). This method was introduced by Sonnevend [21], Ye [28], and Goffin, Haurie, and Vial [6].

For the purpose of proving complexity results, the ACCPM is more clearly described in the context of a convex feasibility problem: find a point in a bounded convex set Ω^* with a nonempty interior. The solution set Ω^* is assumed to contain a ball $\mathcal{N}_\varepsilon$ with radius $\varepsilon < 1$ and is contained in a compact convex set described by matrix inequalities. At each iteration the analytic center of the set of localization is computed and a separation oracle is called: the oracle determines if either the center is in Ω^*, thus solving the problem, or it returns a cut which cuts off the current point

*Received by the editors June 20, 2000; accepted for publication (in revised form) October 3, 2002; published electronically March 19, 2003. This work was supported by Natural Sciences and Engineering Research Council of Canada grant OPG0004152, a strategic grant from the FCAR of Quebec, and a major fellowship from the Faculty of Graduate Studies at McGill University.
http://www.siam.org/journals/siopt/13-4/37414.html

†College of Business Administration, California State University San Marcos, San Marcos, CA 92096-0001 (moskooro@csusm.edu).

‡GERAD/Faculty of Management, McGill University, 1001 Sherbrooke West, Montreal, QC, H3A 1G5, Canada (jlg@crt.umontreal.ca).

and contains the solution set. A special updating step is then needed to get as close as possible to the next analytic center, as first suggested by Mitchell and Todd [15].

The ACCPM has been successfully implemented in a wide variety of applications, as for instance in [5] and [9].

The complexity of the method has been analyzed in the case of single cuts by Atkinson and Vaidya [2], Nesterov [16], and Goffin, Luo, and Ye [7], in the case of multiple cuts by Ye [29] and Goffin and Vial [8], and in the case of quadratic cuts by Luo and Sun [12], Lüthi and Büeler [13], and Sharifi Mokhtarian and Goffin [20].

In this paper we propose an analytic center cutting plane algorithm for convex feasibility problem with *semidefinite cuts*. A semidefinite cut contains as special cases single and multiple linear cuts, as well as quadratic cuts. At each step of the algorithm, an oracle returns a p-dimensional semidefinite cut. We add the cut at the center and derive an updating direction to compute the next analytic center by maximizing the "log det" of the new slack matrix. This is an extension of the direction obtained by Goffin and Vial [8] for the multiple linear cuts to the semidefinite cuts. For alternative approaches to solving determinant maximization problems, see [24, 27].

The restoration procedure is discussed in detail. We prove that the number of Newton steps needed to recover the analytic center from the interior point obtained by the updating direction is of the order of $p\log(p+1)$. Moreover, we show that the analytic center cutting plane algorithm stops with a point in the solution set when the dimension of the accumulated block diagonal cut matrix reaches the bound of $O^*(\frac{p^2m^3}{\mu^2\varepsilon^2})$, where p is the maximum dimension of the cut matrices and $\mu > 0$ is a condition number of the field of cuts. Furthermore, we prove that the Newton method finds the optimal updating direction in at most $O(\frac{p\log\frac{1}{\varepsilon^*}+\log\frac{1}{\mu}}{\beta-\log(1+\beta)})$ iterations, where β is the Newton decrement and $\varepsilon^* = \frac{(1-\theta)\varepsilon}{(1+\theta)(1+n)}$ and θ is a positive constant less than 1.

Independent of our work, recently there have been a few papers dealing with the semidefinite feasibility problem [22, 25, 4]. They differ in that they work with the semidefinite cone directly.

The paper is organized as follows. In section 2 we review the most important properties of the analytic center of a convex set of linear matrix inequalities. This includes the primal, dual, and primal-dual potential functions, the optimality conditions, and a dual algorithm for the computation of the analytic center. We introduce the semidefinite cuts in section 3 and derive the optimal updating direction to restore the analytic center after adding a semidefinite cut. Section 4 deals with the complexity of the restoration algorithm. In section 5 we present the ACCPM algorithm for the convex feasibility problem with semidefinite cuts. In section 6 we derive the complexity of the algorithm and finally in section 7 an upper bound on the number of damped Newton steps to compute the optimal updating direction is established.

Notation. We use the following notation: Lowercase letters are used to show vectors and uppercase letters are used for matrices. I and I_n are identity matrices of appropriate size or of size n. The ith column of I is shown by e_i, and $\mathbf{Diag}(e_i)$ is a diagonal matrix with e_i on its main diagonal.

We refer to the space of $n \times n$ symmetric matrices by $\mathcal{S}^n$, positive semidefinite matrices by $\mathcal{S}^n_+$, and positive definite matrices by $\mathcal{S}^n_{++}$. We denote the jth eigenvalue of a symmetric matrix A by $\lambda_j(A)$ in decreasing order.

For a square matrix A, $\mathbf{tr}(A)$ is the trace of A, $\mathbf{diag}(A)$ is a column vector made

up of the diagonal elements of A, the Frobenius-norm of A is defined via

$$\|A\|^2 = \mathbf{tr} A^T A = \sum_{j=1}^{n} (\lambda_j(A^T A)),$$

and if A is symmetric, then the ∞-norm of A is defined by

$$\|A\|_\infty = \max |\lambda_j(A)|, \qquad j = 1, \ldots, n.$$

The operator "$\bullet$" indicates the inner product of two matrices:

$$A \bullet B = \mathbf{tr} A^T B = \sum_{i,j} a_{ij} b_{ij}.$$

For symmetric matrices A_i, $i = 1, \ldots, m$, we define the m-vector a_q^l by

$$a_q^l = ((A_1)_{lq}, (A_2)_{lq}, \ldots, (A_m)_{lq}). \tag{1.1}$$

This vector is denoted by b_q^l when we deal with symmetric matrices B_i.

The Löwner partial order on the symmetric matrices is defined by $A \succeq B$ ($A \succ B$) if $A - B \in \mathcal{S}_+^n$ ($A - B \in \mathcal{S}_{++}^n$).

2. Analytic center and its properties. We start with an important lemma which plays a key role in the interior point algorithms [1].

LEMMA 2.1. *Let $X \in \mathcal{S}^n$; then*

$$\log \det X \leq I \bullet (X - I)$$

with equality iff $X = I$. Moreover, if $\|X - I\|_\infty < 1$, then

$$\log \det X \geq I \bullet (X - I) - \frac{\|X - I\|^2}{2(1 - \|X - I\|_\infty)}.$$

Consider the following set:

$$\Omega_D = \{y \in R^m : \mathcal{A}^T y \preceq C\},$$

where $\mathcal{A}$ is a linear operator from $\mathcal{S}^n$ to m-vector $\mathcal{R}^m$ defined by $(\mathcal{A}X)_i = A_i \bullet X$, for $A_i \in \mathcal{S}^n$, $i = 1, \ldots, m$, and $\mathcal{A}^T : \mathcal{R}^m \to \mathcal{S}^n$ is its adjoint operator defined by $\mathcal{A}^T y = \sum_i y_i A_i$. Note that

$$\langle \mathcal{A}X, y \rangle = \langle \mathcal{A}^T y, X \rangle.$$

We assume that Ω_D is a convex compact set and therefore A_i are linearly independent. We also assume that Ω_D has a strictly feasible point. That is,

$$\Omega_D^\circ = \{y \in R^m : \mathcal{A}^T y \prec C\}$$

is nonempty. Given a point y in Ω_D°, the dual potential function is defined via

$$\begin{aligned} \phi_D(y) &= \log \det (C - \mathcal{A}^T y)^{-1} \\ &= \log \det S^{-1}, \end{aligned}$$

where $S(y) = C - \mathcal{A}^T y$ is the slack matrix. We denote $S(y)$ by S when there is no ambiguity. The minimizer of the dual potential function is called the analytic center:

$$y^a = \arg\min \phi_D(y). \tag{2.1}$$

Since $\phi_D(y)$ is strictly convex on Ω_D°, the analytic center is well defined and unique. By the KKT optimality conditions, a point y^a is the analytic center of Ω_D iff there exist matrices $S^a \succ 0$ and $X^a \succ 0$ such that

$$\begin{aligned} \mathcal{A}X^a &= 0, \\ \mathcal{A}^T y^a + S^a &= C, \\ X^a S^a &= I. \end{aligned} \tag{2.2}$$

Abusing notation somewhat, we also denote the dual potential function by $\phi_D(S)$ and the analytic center by (X^a, y^a, S^a).

The analytic center can also be derived by primal characterization. Let

$$\Omega_P = \{X \in \mathcal{S}^n : \mathcal{A}X = 0, X \succeq 0\},$$

and let there exist $X \in \Omega_P$ with $X \succ 0$. One can verify that the minimizer of the primal potential function

$$\phi_P(X) = C \bullet X - \log\det X$$

over the primal feasible region satisfies the optimality conditions (2.2).

The analytic center can alternatively be characterized as the minimizer of the primal-dual potential function

$$\begin{aligned} \phi_{PD}(X,S) &= \phi_P(X) + \phi_D(S) \\ &= C \bullet X - \log\det XS \\ &= X \bullet S - \log\det XS \end{aligned}$$

over $\Omega_{PD} = \Omega_D \times \Omega_P$. Let us show that (2.2) is the optimality condition for this minimization. First observe that

$$\begin{aligned} \phi_{PD}(X^a, S^a) &= X^a \bullet S^a - \log\det X^a S^a \\ &= \mathbf{tr} I - \log\det I \\ &= n. \end{aligned}$$

On the other hand, from Lemma 2.1

$$\phi_{PD}(X,S) \geq X \bullet S - \mathbf{tr}(XS - I) = n$$

for all $(X,S) \in \Omega_{PD}$, with equality iff $XS = I$. Thus (X^a, y^a, S^a) is the (unique) minimizer of $\phi_{PD}(X,S)$. Approximate analytic centers are defined for computational reasons. A θ-approximate analytic center is denoted by $(\bar{X}, \bar{y}, \bar{S})$ and defined via

$$\begin{aligned} \mathcal{A}\bar{X} &= 0, \\ \mathcal{A}^T \bar{y} + \bar{S} &= C, \\ \|\bar{X}\bar{S} - I\| &\leq \theta < 1. \end{aligned} \tag{2.3}$$

Let $\tilde{y} \in \Omega_D^\circ$, and let $\tilde{S}$ be its slack matrix. An ellipsoid centered at $\tilde{y}$, and contained in the interior of Ω_D, i.e., the set

$$\{y : \|\tilde{S}^{-.5}\mathcal{A}^T(y - \tilde{y})\tilde{S}^{-.5}\| \leq 1\},$$

is called the dual Dikin ellipsoid. Similarly, we can define the Dikin ellipsoid for the primal feasible set. Let $\tilde{X} \succ 0$ be in Ω_P. Then the primal Dikin ellipsoid centered at $\tilde{X}$ is

$$\{X : \|\tilde{X}^{-.5}(X - \tilde{X})\tilde{X}^{-.5}\| \leq 1\}.$$

The next lemma gives lower and upper bounds on the primal potential function at a θ-approximate center. Similar bounds can be established for the dual and primal-dual potentials.

LEMMA 2.2. *Let $(\bar{X}, \bar{y}, \bar{S})$ be a θ-approximate center. Then*

$$\phi_P(X^a) \leq \phi_P(\bar{X}) \leq \phi_P(X^a) + \frac{\theta^2}{2(1-\theta)}.$$

Proof. The left-hand side inequality is trivial. We prove the upper bound. From Lemma 2.1

$$\begin{aligned}\phi_{PD}(\bar{X}, \bar{S}) &\leq \bar{X} \bullet \bar{S} - I \bullet (\bar{X}\bar{S} - I) + \frac{\|\bar{X}\bar{S} - I\|^2}{2(1 - \|\bar{X}\bar{S} - I\|)} \\ &\leq n + \frac{\theta^2}{2(1-\theta)},\end{aligned}$$

and therefore

$$\phi_{PD}(\bar{X}, \bar{S}) - \phi_{PD}(X^a, S^a) \leq \frac{\theta^2}{2(1-\theta)}$$

or

$$(\phi_P(\bar{X}) - \phi_P(X^a)) + (\phi_D(\bar{y}) - \phi_D(y^a)) \leq \frac{\theta^2}{2(1-\theta)}.$$

Since y^a is the minimizer of ϕ_D over Ω_D, we have

$$\phi_D(\bar{y}) - \phi_D(y^a) \geq 0.$$

Thus,

$$\phi_P(\bar{X}) - \phi_P(X^a) \leq \frac{\theta^2}{2(1-\theta)}.$$

The lemma follows. □

Computational algorithms for the analytic center of a polytope have been devised in primal, dual, and primal-dual settings based on the Newton method. These algorithms can be extended to compute the analytic center of a convex body described by matrix inequalities [26, 18]. In the rest of this section we discuss the extension of the dual algorithm to compute an approximate analytic center. We refer the reader to [30, Chap. 3] for a comprehensive analysis of the computational algorithms for the analytic center in the linear case.

The Newton method is used to solve (2.1). Let $y \in \Omega_D^\circ$ and dy be the dual direction. Consider the following quadratic approximation of $\phi_D(y)$:

$$\phi_D(y + dy) \approx \phi_D(y) + (\mathcal{A}S^{-1})^T dy + \frac{1}{2} dy^T (\mathcal{A}_D \mathcal{A}_D^T) dy,$$

where $\mathcal{A}_D : \mathcal{S}^n \to \Re^m$ is a linear operator and $\mathcal{A}_D^T : \Re^m \to \mathcal{S}^n$ is its adjoint operator, defined via

$$\mathcal{A}_D X = \begin{pmatrix} S^{-.5}A_1S^{-.5} \bullet X \\ \vdots \\ S^{-.5}A_mS^{-.5} \bullet X \end{pmatrix} \quad \text{and} \quad \mathcal{A}_D^T y = \sum_{i=1}^{m} y_i S^{-.5}A_iS^{-.5}.$$

Note that $(\mathcal{A}_D\mathcal{A}_D^T) \in \mathcal{S}^m$ with $(\mathcal{A}_D\mathcal{A}_D^T)_{ij} = \mathbf{tr}A_iS^{-1}A_jS^{-1}$. Since A_i are linearly independent, then $\mathcal{A}_D\mathcal{A}_D^T \succ 0$.

By minimizing the quadratic approximation of $\phi_D(y)$,

$$\min_{dy \in R^m} \quad (\mathcal{A}S^{-1})^T dy + \tfrac{1}{2} dy^T (\mathcal{A}_D\mathcal{A}_D^T) dy,$$

we have

$$dy = -(\mathcal{A}_D\mathcal{A}_D^T)^{-1}\mathcal{A}S^{-1}.$$

Let $y^+ = y + dy$ be the updated dual point. y^+ should be a feasible point for Ω_D. That is, $\mathcal{A}^T y^+ + S^+ = C$ with $S^+ = S + dS$ and thus $dS = -\mathcal{A}^T dy$.

Now let

$$X(S) = S^{-1}\left(\mathcal{A}^T dy + S\right) S^{-1}.$$

The following lemma shows that $X(S)$ is the solution of a least squares problem.

LEMMA 2.3. *Let $y \in \Omega_D^\circ$, and let S be the slack matrix. Then the primal solution $X(S)$ is the minimizer of the least squares problem*

$$\begin{array}{ll} \min & \left|\left|S^{.5}XS^{.5} - I\right|\right| \\ s.t. & \\ & \mathcal{A}X = 0. \end{array}$$

Proof. The KKT condition for this problem is

$$2SXS - 2S - \mathcal{A}^T v = 0, \tag{2.4}$$

where $v \in R^m$. By multiplying (2.4) from the right side and from the left side by S^{-1} and then applying the operator $\mathcal{A}$ and noting that $\mathcal{A}X = 0$, we have

$$-2\mathcal{A}S^{-1} - (\mathcal{A}_{\mathcal{D}}\mathcal{A}_{\mathcal{D}}{}^T)v = 0$$

or

$$\begin{aligned} v &= -2(\mathcal{A}_D\mathcal{A}_D^T)^{-1}\mathcal{A}S^{-1} \\ &= 2dy. \end{aligned}$$

The proof follows from (2.4). ∎

Now let

$$P(S) = S^{.5}X(S)S^{.5} - I.$$

The following lemma shows that if $\|P(S)\| < 1$, then the updated slack matrix S^+ is strictly feasible and the dual algorithm converges to an approximate analytic center quadratically.

LEMMA 2.4. *If* $\|P(S)\| < 1$ *for some interior point* $y \in \Omega_D^\circ$ *and its slack matrix* S*, then*

$$S^+ \succ 0 \quad \textit{and} \quad \|P(S^+)\| \le \|P(S)\|^2 < 1.$$

Proof. Note that $S^+ = S^{.5}(I - P(S))S^{.5}$ and since $\|P(S)\| < 1$, then $\lambda_j(P(S)) < 1$ for $j = 1, \ldots, n$, which implies that $S^+ \succ 0$.

To prove the second part of the lemma, first observe that from Lemma 2.3

$$\begin{aligned} \|P(S^+)\| &= \|(S^+)^{.5} X(S^+)(S^+)^{.5} - I\| \\ &\le \|(S^+)^{.5} X(S)(S^+)^{.5} - I\|. \end{aligned} \tag{2.5}$$

On the other hand,

$$S^+ = S - \mathcal{A}^T dy = 2S - \left(\mathcal{A}^T dy + S\right) = 2S - SX(S)S. \tag{2.6}$$

Now from (2.5) and (2.6) (in what follows we denote $X(S)$ by X)

$$\begin{aligned} \|P(S^+)\|^2 &\le \|X^{.5} S^+ X^{.5} - I\|^2 \\ &= \|X^{.5}(2S - SXS)X^{.5} - I\|^2 \\ &= \|(X^{.5} S X^{.5} - I)^2\|^2 \\ &= \mathbf{tr}(X^{.5} S X^{.5} - I)^4 \\ &= \sum (\lambda_j(X^{.5} S X^{.5}) - 1)^4 \\ &\le \left(\sum (\lambda_j(X^{.5} S X^{.5}) - 1)^2\right)^2 \\ &= \left(\|X^{.5} S X^{.5} - I\|^2\right)^2 \\ &= \|P(S)\|^4. \end{aligned}$$

The lemma now follows. ∎

When a strict interior point (y, S) with $\|P(S)\| \ge 1$ is available, the direction dy with a step size $\alpha/\|P(S)\|$, $0 < \alpha < 1$, is taken. One can prove that, in this case, the potential function is reduced by a constant amount $\delta > 0$ at each iteration, i.e.,

$$\phi_D(y^+) \le \phi_D(y) - \delta,$$

and that after a finite number of iterations $\|P(S)\|$ satisfies the desired condition (< 1). The general complexity of the algorithm can be obtained from the fact that the potential function at the analytic center is a lower bound for $\phi_D(y)$. This implies that after at most $O(\phi_D(y^0) - \phi_D(y^a))$ iterations the algorithm stops with an approximation of the analytic center.

Primal and primal-dual algorithms also give the same result, and the analysis is more or less similar to the dual case. The complexity result, however, for the primal-dual case is more specific since the potential function at the center is known in advance. That is, the primal-dual algorithm stops after $O(\phi_{PD}(X^0, y^0, S^0) - n)$ iterations of the Newton method.

3. Semidefinite cut. In this section we discuss the recentering process. Let us formally define a semidefinite cut.

DEFINITION 3.1. *A p-dimensional semidefinite cut is a cut of the form*

$$\mathcal{B}^T y \preceq D,$$

where $D \in \mathcal{S}^p$ and $\mathcal{B} : \mathcal{S}^p \to \Re^m$ is a linear operator defined by $(\mathcal{B}X)_i = B_i \bullet X$ with $B_i \in \mathcal{S}^p$, and $\mathcal{B}^T y = \sum_{i=1}^m y_i B_i$ is its adjoint operator. The matrices B_i are called the cut matrices, and if $D = \mathcal{B}^T \bar{y}$, where $\bar{y}$ is an approximate center of Ω_D, then the cut is called a central semidefinite cut.

Notice that the semidefinite cut $\mathcal{B}^T y \preceq D$ is a generalization of linear, multiple, and quadratic cuts. If B_i and D are scalar, then $\mathcal{B}^T y \preceq D$ is reduced to a single cut $b^T y \leq b_0$, and if they are diagonal matrices, then the cut is reduced to a set of multiple linear cuts $B^T y \leq \mathbf{diag}(D)$, where the columns of matrix B are $\mathbf{diag}(B_i)$. Furthermore, if the cut matrices B_i and the constant matrix D are of the form

$$B_i = \begin{pmatrix} \mathbf{0} & -b_i \\ -b_i^T & q_i \end{pmatrix}, \quad D = \begin{pmatrix} \mathbf{I} & 0 \\ 0 & d \end{pmatrix},$$

then $\mathcal{B}^T y \preceq D$ is reduced to a quadratic cut $y^T(B^T B)y + q^T y \leq d$, where the vectors b_i form the columns of matrix B.

In our analysis, from now on, we assume that the cuts are central. The updated set of localization $\Omega_D^+ \subset \Omega_D$ after adding the cut is

$$\Omega_D^+ = \{y \in R^m : \mathcal{A}^T y \preceq C,\ \mathcal{B}^T y \preceq \mathcal{B}^T \bar{y}\}.$$

To compute an approximate center of the updated set of localization, we need a strict interior point of Ω_D^+. We start from $\bar{y}$ and choose the direction $dy = y - \bar{y}$ towards the interior of the set of localization as the maximizer of the determinant of the new slack matrix to the boundary of the dual Dikin ellipsoid centered at $\bar{y}$:

$$\begin{array}{ll} \min & -\log\det\Lambda \\ \text{s.t.} & \\ & \|\bar{S}^{-.5}\mathcal{A}^T dy \bar{S}^{-.5}\| \leq 1, \\ & \mathcal{B}^T dy + \Lambda = 0, \\ & \Lambda \succeq 0. \end{array} \tag{3.1}$$

We call the optimal solution of problem (3.1) the *optimal updating direction*. The new point will then be used as an initial point to restore the analytic center. Problem (3.1) can be reformulated as follows:

$$\begin{array}{ll} \min & -\log\det(-\mathcal{B}^T dy) \\ \text{s.t.} & \\ & dy^T(\mathcal{A}_D\mathcal{A}_D^T)dy \leq 1. \end{array}$$

By the KKT optimality conditions and since the interior of the feasible region is nonempty, $\tilde{dy}$ and $\tilde{\Lambda}$ are optimal iff there exists unique multiplier $\sigma > 0$ such that

$$\mathcal{B}\tilde{\Lambda}^{-1} + \sigma(\mathcal{A}_D\mathcal{A}_D^T)\tilde{dy} = 0, \tag{3.2}$$

$$\mathcal{B}^T\tilde{dy} + \tilde{\Lambda} = 0. \tag{3.3}$$

From (3.2)

$$\tilde{dy} = -\frac{1}{\sigma}(\mathcal{A}_D\mathcal{A}_D^T)^{-1}\mathcal{B}\tilde{\Lambda}^{-1} \tag{3.4}$$

and from (3.3)

$$\tilde{\Lambda} = \frac{1}{\sigma}\mathcal{B}^T(\mathcal{A}_D\mathcal{A}_D^T)^{-1}\mathcal{B}\tilde{\Lambda}^{-1}.$$

Define the operator $\mathcal{V} : \mathcal{S}^p \to \mathcal{S}^p$ by $\mathcal{V} = \mathcal{B}^T(\mathcal{A}_D\mathcal{A}_D^T)^{-1}\mathcal{B}$; then

$$\tilde{\Lambda} = \frac{1}{\sigma}\mathcal{V}\tilde{\Lambda}^{-1}. \tag{3.5}$$

If $\mathcal{V}$ is nonsingular, then the dual direction $\tilde{\Lambda}$ can be uniquely computed by solving the following optimization problem:

$$\min_{\Lambda \succeq 0} \quad \tfrac{\sigma}{2}\mathbf{tr}\Lambda\mathcal{V}^{-1}\Lambda - \log\det\Lambda.$$

The correct value of the Lagrange multiplier $\sigma \geq 0$ is known in advance:

$$\begin{aligned}
\|S^{-.5}\mathcal{A}^T\tilde{dy}S^{-.5}\|^2 &= \tilde{dy}^T(\mathcal{A}_D\mathcal{A}_D^T)\tilde{dy} \\
&= \frac{1}{\sigma^2}(\mathcal{B}\tilde{\Lambda}^{-1})^T(\mathcal{A}_D\mathcal{A}_D^T)^{-1}\mathcal{B}\tilde{\Lambda}^{-1} && \text{(from (3.4))} \\
&= \frac{1}{\sigma^2}\mathbf{tr}\tilde{\Lambda}^{-1}\mathcal{B}^T(\mathcal{A}_D\mathcal{A}_D^T)^{-1}\mathcal{B}\tilde{\Lambda}^{-1} \\
&= \frac{1}{\sigma^2}\mathbf{tr}\tilde{\Lambda}^{-1}\mathcal{V}\tilde{\Lambda}^{-1} \\
&= \frac{1}{\sigma}\mathbf{tr}\tilde{\Lambda}^{-1}\tilde{\Lambda} && \text{(from (3.5))} \\
&= \frac{p}{\sigma}.
\end{aligned}$$

On the other hand, $\|\bar{S}^{-.5}\mathcal{A}^T\tilde{dy}\bar{S}^{-.5}\| = 1$ and thus $\sigma = p$. Consequently,

$$\tilde{dy} = -\frac{1}{p}(\mathcal{A}_D\mathcal{A}_D^T)^{-1}\mathcal{B}\tilde{\Lambda}^{-1} \quad \text{and} \quad \tilde{dS} = -\mathcal{A}^T\tilde{dy},$$

where

$$\tilde{\Lambda} = \arg\min_{\Lambda \succeq 0}\left\{\frac{p}{2}\mathbf{tr}\Lambda\mathcal{V}^{-1}\Lambda - \log\det\Lambda\right\}.$$

To update the primal direction, observe that the updated primal feasible region Ω_P^+ is

$$\Omega_P^+ = \left\{\begin{pmatrix} X & \\ & T \end{pmatrix} \succeq 0 : \mathcal{A}X + \mathcal{B}T = 0\right\},$$

and the primal direction $\tilde{dX}$ is obtained by maximizing $\log\det T$ while respecting primal feasibility and remaining in the primal Dikin ellipsoid centered at $\bar{X}$:

$$\begin{aligned}
&\min && -\log\det T \\
&\text{s.t.} && \\
&&& \mathcal{A}dX + \mathcal{B}T = 0, \\
&&& \|\bar{S}dX\| \leq 1, \\
&&& T \succeq 0.
\end{aligned} \tag{3.6}$$

The optimality conditions of problem (3.6) are

$$-\tilde{T}^{-1} + \mathcal{B}^T v = 0, \tag{3.7}$$

$$\mathcal{A}^T v + \sigma'\bar{S}(\tilde{dX})\bar{S} = 0, \tag{3.8}$$

$$\sigma'(1 - \|\bar{S}\tilde{dX}\|) = 0, \tag{3.9}$$

$$\mathcal{A}(\tilde{dX}) + \mathcal{B}\tilde{T} = 0, \tag{3.10}$$

where $\sigma' \geq 0$ is the Lagrange multiplier associated with the norm constraint. By multiplying (3.8) from the left and from the right by $\bar{S}^{-1}$ and then applying the operator $\mathcal{A}$ we have

$$(\mathcal{A}_D\mathcal{A}_D^T)v + \sigma'\mathcal{A}(d\tilde{X}) = 0,$$

using (3.10)

$$v = \sigma'(\mathcal{A}_D\mathcal{A}_D^T)^{-1}\mathcal{B}\tilde{T},$$

and again from (3.8)

$$\begin{aligned} d\tilde{X} &= -\frac{1}{\sigma'}\bar{S}^{-1}(\mathcal{A}^T v)\bar{S}^{-1} \\ &= -\bar{S}^{-1}\mathcal{A}^T(\mathcal{A}_D\mathcal{A}_D^T)^{-1}\mathcal{B}\tilde{T}\bar{S}^{-1}. \end{aligned}$$

Since $\mathcal{A}^T(\mathcal{A}_D\mathcal{A}_D^T)^{-1}\mathcal{B}\tilde{T}$ is symmetric, then $d\tilde{X}$ is symmetric. Finally, from (3.7)

$$\begin{aligned} \tilde{T}^{-1} &= \mathcal{B}^T v \\ &= \sigma'\mathcal{B}^T(\mathcal{A}_D\mathcal{A}_D^T)^{-1}\mathcal{B}\tilde{T} \\ &= \sigma'\mathcal{V}\tilde{T}, \end{aligned} \tag{3.11}$$

and $\tilde{T}$ is the unique solution of the following optimization problem:

$$\tilde{T} = \arg\min_{T\succeq 0}\left\{\frac{\sigma'}{2}\mathbf{tr}T\mathcal{V}T - \log\det T\right\}. \tag{3.12}$$

Let us find the Lagrange multiplier σ':

$$\begin{aligned} \|\bar{S}(d\tilde{X})\|^2 &= \mathbf{tr}\bar{S}^{.5}(d\tilde{X})\bar{S}(d\tilde{X})\bar{S}^{.5} \\ &= \mathbf{tr}\bar{S}^{-1}\mathcal{A}^T\underbrace{(\mathcal{A}_D\mathcal{A}_D^T)^{-1}\mathcal{B}\tilde{T}}_{u}\bar{S}^{-1}\mathcal{A}^T\underbrace{(\mathcal{A}_D\mathcal{A}_D^T)^{-1}\mathcal{B}\tilde{T}}_{u} \\ &= u^T(\mathcal{A}_D\mathcal{A}_D^T)u \\ &= (\mathcal{B}\tilde{T})^T(\mathcal{A}_D\mathcal{A}_D^T)^{-1}\mathcal{B}\tilde{T} \\ &= \mathbf{tr}\tilde{T}\mathcal{V}\tilde{T} \\ &= \frac{p}{\sigma'}. \end{aligned}$$

Hence $\sigma' = p$.

Now for $\alpha < 1 - \theta$, let $y^+ = \bar{y} + \alpha\tilde{dy}$, and let

$$X^+ = \begin{pmatrix} \bar{X} + \alpha d\tilde{X} & 0 \\ 0 & \alpha\tilde{T} \end{pmatrix}, \quad S^+ = \begin{pmatrix} \bar{S} + \alpha\tilde{dS} & 0 \\ 0 & \alpha\tilde{\Lambda} \end{pmatrix}$$

be the updated iteration. Since $\tilde{T}$ is uniquely defined, from (3.5) and (3.11) one can easily prove that $\tilde{T}\tilde{\Lambda} = \frac{1}{p}I$ and therefore computing $\tilde{T}$ suffices to update the iteration. We postpone the complexity analysis of the recentering direction (problem (3.12)) to section 7.

Notice that the nonsingularity of operator $\mathcal{V}$ establishes a symmetry between the primal and the dual updating directions. In other words, if $\mathcal{V}$ is invertible, then the primal direction can be computed from the dual direction and vice versa. On the

other hand, if $\mathcal{V}$ is singular, then $\Lambda \succ 0$ does not guarantee the existence of $T \succ 0$ with $\Lambda = \mathcal{V}T$. In such a case, the dual direction is obtained as a by-product of the primal direction via $\Lambda = \mathcal{V}T$ and (3.11).

The following lemma guarantees the strict feasibility of the iteration.

LEMMA 3.2. *The updated points X^+ and (y^+, S^+) are strictly feasible for Ω_P^+ and Ω_D^+, respectively, and therefore they can be used as the starting point to recover the analytic center.*

Proof. First observe that

$$\begin{aligned}
\|\bar{X}^{-1}d\tilde{X}\|^2 &= \|(\bar{S}^{-.5}\bar{X}^{-1}\bar{S}^{-.5})(\bar{S}^{.5}d\tilde{X}\bar{S}^{.5})\|^2 \\
&= \mathbf{tr}(\bar{S}^{-.5}\bar{X}^{-1}\bar{S}^{-.5})^2(\bar{S}^{.5}d\tilde{X}\bar{S}^{.5})^2 \\
&\le \sum_{j=1}^{n} \lambda_j^2(\bar{S}^{-.5}\bar{X}^{-1}\bar{S}^{-.5})\lambda_j^2(\bar{S}^{.5}d\tilde{X}\bar{S}^{.5}) \qquad (3.13)\\
&\le \lambda_1^2(\bar{S}^{-.5}\bar{X}^{-1}\bar{S}^{-.5})\sum_{j=1}^{n} \lambda_j^2(\bar{S}^{.5}d\tilde{X}\bar{S}^{.5}) \\
&= \|\bar{S}^{-1}\bar{X}^{-1}\|_\infty^2\|\bar{S}d\tilde{X}\|^2, \qquad (3.14)
\end{aligned}$$

where inequality (3.13) is due to Theobald [23] (see also Marshall and Olkin [14]). From (3.14) and noting that $d\tilde{X}$ is optimal for problem (3.6), and $\bar{X}$ and $\bar{S}$ are approximate centers, one has

$$\|\bar{X}^{-1}d\tilde{X}\| \le \frac{1}{1-\theta}.$$

On the other hand, $\alpha < 1 - \theta$, and

$$\bar{X} + \alpha d\tilde{X} = \bar{X}^{.5}(I + \alpha\bar{X}^{-.5}d\tilde{X}\bar{X}^{-.5})\bar{X}^{.5}.$$

Thus $\bar{X} + \alpha d\tilde{X} \succ 0$. Moreover, $\tilde{T}$ is positive definite by construction, and hence $X^+ \succ 0$. Since $\bar{X}$ is primal feasible, then $\mathcal{A}(\bar{X} + \alpha d\tilde{X}) + \mathcal{B}(\alpha\tilde{T}) = 0$. That is, X^+ is strictly feasible for the updated primal set Ω_P^+.

To prove the strict feasibility of the dual iteration, we have

$$\bar{S} + \alpha d\tilde{S} = \bar{S}^{.5}(I - \alpha\bar{S}^{-.5}\mathcal{A}^T d\tilde{y}\bar{S}^{-.5})\bar{S}^{.5},$$

and since $d\tilde{y}$ is optimal for problem (3.1), then $\|\bar{S}^{-.5}\mathcal{A}^T d\tilde{y}\bar{S}^{-.5}\| = 1$. Thus $\bar{S}+\alpha d\tilde{S} \succ 0$. ∎

4. Analysis of restoration. Before getting started, we state a lemma similar to Lemma 2.1.

LEMMA 4.1. *Let $S \in \mathcal{S}^n$ be such that $\|S\| < 1$. Then*

$$\log\det(I + S) \ge I \bullet S + \|S\| + \log(1 - \|S\|).$$

Proof. The following inequality is well known (for a proof, see Roos, Terlaky, and Vial [19, p. 439]):

$$\sum_{j=1}^{n} \log(1 + s_j) \ge e^T s + \|s\| + \log(1 - \|s\|).$$

The lemma follows by letting $\lambda_j(S) = s_j$. ∎

The following lemma bounds the potential functions at the new point.

LEMMA 4.2. *Let $(\bar{X}, \bar{y}, \bar{S})$ be a θ-approximate analytic center. Then*

$$\phi_D(S^+) \le \phi_D(\bar{S}) - \alpha(1-\theta) - \log(1-\alpha) - \log\det\alpha\tilde{\Lambda}, \tag{4.1}$$

$$\phi_P(X^+) \le \phi_P(\bar{X}) - \alpha(1-\theta) - \log(1-\alpha) - \log\det\alpha\tilde{T}, \tag{4.2}$$

and

$$\phi_{PD}(X^+, S^+) \le \phi_{PD}(\bar{X}, \bar{S}) - 2\alpha(1-\theta) - 2\log(1-\alpha) - 2p\log\alpha + p\log p. \tag{4.3}$$

Proof.

$$\begin{aligned}\phi_D(S^+) &= -\log\det\bar{S}(I + \alpha\bar{S}^{-1}d\tilde{S}) - \log\det\alpha\tilde{\Lambda}\\ &= \phi_D(\bar{S}) - \log\det(I + \alpha\bar{S}^{-1}d\tilde{S}) - \log\det\alpha\tilde{\Lambda}.\end{aligned}$$

By Lemma 4.1

$$\begin{aligned}&\phi_D(S^+) \le \phi_D(\bar{S})\\ &+ (-I \bullet \alpha\bar{S}^{-1}d\tilde{S}) - \|\alpha\bar{S}^{-1}d\tilde{S}\| - \log(1 - \|\alpha\bar{S}^{-1}d\tilde{S}\|) - \log\det\alpha\tilde{\Lambda}.\end{aligned} \tag{4.4}$$

Since $\bar{X} \bullet d\tilde{S} = -(\mathcal{A}\bar{X})^T d\tilde{y} = 0$,

$$\begin{aligned}\left|I \bullet \alpha\bar{S}^{-1}d\tilde{S}\right| &= \alpha\left|(\bar{S}^{-1} - \bar{X}) \bullet d\tilde{S}\right|\\ &= \alpha\left|(I - \bar{S}^{.5}\bar{X}\bar{S}^{.5}) \bullet \bar{S}^{-.5}(d\tilde{S})\bar{S}^{-.5}\right|\\ &\le \alpha\|\bar{S}^{.5}\bar{X}\bar{S}^{.5} - I\|\|\bar{S}^{-.5}(d\tilde{S})\bar{S}^{-.5}\|\\ &\le \alpha\theta.\end{aligned}$$

The first inequality thus follows from the inequality (4.4), the above fact, and noting that $f(t) = -t - \log(1-t)$ is an increasing function over its domain.

To prove the second inequality note that

$$\phi_P(X^+) = \phi_P(\bar{X}) + \alpha C \bullet d\tilde{X} + \alpha\bar{y}^T\mathcal{B}\tilde{T} - \log\det(I + \alpha\bar{X}^{-1}d\tilde{X}) - \log\det\alpha\tilde{T};$$

again by Lemma 4.1

$$\begin{aligned}\phi_P(X^+) \quad\le\quad & \phi_P(\bar{X}) + \alpha C \bullet d\tilde{X} + \alpha\bar{y}^T\mathcal{B}\tilde{T} - I \bullet \alpha\bar{X}^{-1}d\tilde{X}\\ & -\|\alpha\bar{X}^{-1}d\tilde{X}\| - \log(1 - \|\alpha\bar{X}^{-1}d\tilde{X}\|) - \log\det\alpha\tilde{T}.\end{aligned} \tag{4.5}$$

On the other hand, from $\mathcal{A}d\tilde{X} + \mathcal{B}\tilde{T} = 0$

$$\begin{aligned}\left|\alpha C \bullet d\tilde{X} + \alpha\bar{y}^T\mathcal{B}\tilde{T} - \alpha\bar{X}^{-1} \bullet d\tilde{X}\right| &= \left|\alpha\bar{S} \bullet d\tilde{X} - \alpha\bar{X}^{-1} \bullet d\tilde{X}\right|\\ &= \alpha\left|(\bar{X}^{.5}\bar{S}\ \bar{X}^{.5} - I) \bullet \bar{X}^{-.5}d\tilde{X}\bar{X}^{-.5}\right|\\ &\le \alpha\|\bar{X}^{.5}\bar{S}\ \bar{X}^{.5} - I\|\|\bar{X}^{-.5}d\tilde{X}\bar{X}^{-.5}\|\\ &\le \alpha\theta.\end{aligned}$$

The primal inequality therefore follows from the inequality (4.5), the above fact, and the property of increasing function $f(t)$. Finally, the last inequality is obtained by adding (4.1) and (4.2). ∎

The following theorem gives the complexity of updating the analytic center after adding a p-dimensional semidefinite cut.

THEOREM 4.3. *Starting from the strict interior point (X^+, S^+), the number of iterations to update an approximate analytic center is bounded by $O(p\log(p+1))$, where p is the dimension of the central semidefinite cut $\mathcal{B}^T y \preceq \mathcal{B}^T \bar{y}$.*

Proof. Since $(\bar{X}, \bar{S})$ is a θ-approximate center of the current set of localization, from Lemma 2.1 we have

$$\begin{aligned}\phi_{PD}(\bar{X}, \bar{S}) &\le \bar{X} \bullet \bar{S} - I \bullet (\bar{X}\bar{S} - I) + \frac{\|\bar{X}\bar{S} - I\|^2}{2(1 - \|\bar{X}\bar{S} - I\|)} \\ &\le n + \frac{\theta^2}{2(1-\theta)}.\end{aligned}$$

Now let the analytic center of the new convex body be $((X^a)^+, (S^a)^+)$. Since $\phi_{PD}((X^a)^+, (S^a)^+) = n + p$, from (4.3)

$$\phi_{PD}(X^+, S^+) - \phi_{PD}\left((X^a)^+, (S^a)^+\right) \le \kappa(\alpha, \theta, p) + p\log p,$$

where

$$\kappa(\alpha, \theta, p) = \frac{\theta^2}{2(1-\theta)} - 2\alpha(1-\theta) - 2\log(1-\alpha) - 2p\log\alpha - p.$$

At each iteration of the Newton method the potential function is reduced by a constant amount δ. Therefore after at most

$$\left\lceil \frac{\kappa(\alpha, \theta, p) + p\log p}{\delta} \right\rceil \sim O(p\log(p+1))$$

iterations the algorithm stops with an updated analytic center. ∎

In the next section we present an ACCPM algorithm for the convex feasibility problem with semidefinite cuts.

5. The ACCPM algorithm. The ACCPM algorithm attempts to find a feasible point in $\Omega^* \subset \Omega_D$, where Ω^* is the solution set and contains a full-dimensional ball $\mathcal{N}_\varepsilon$ with radius ε. We make the following assumptions.

ASSUMPTION 1. $\Omega_D \subset [0,1]^m$.

ASSUMPTION 2. *Ω_D is described by an oracle. That is, the oracle determines if either the center is in Ω^*, thus solving the problem, or it returns a p-dimensional semidefinite cut which contains Ω^*.*

ASSUMPTION 3. *For the semidefinite cut $\mathcal{B}^T$, we assume that*

$$\max_{i,l,q} \left(\mathbf{tr} B_i, \|b_q^l\|\right) = 1, \tag{5.1}$$

where b_q^l is the m-vector defined in (1.1).

Assumptions 1 and 3 are made for simplicity and they can be satisfied by scaling. Assumption 2 guarantees the existence of an oracle to return semidefinite cuts at each iteration. In practice, when the ACCPM is applied to optimize a nonsmooth function, such an oracle returns the subgradients of the function at the current point if it is not the optimal solution.

For the next assumption we need to define a condition number on the semidefinite cut.

DEFINITION 5.1. *At any point $z \notin \Omega^*$, let $\mathcal{B}_z^T y \preceq \mathcal{B}_z^T z$ be the cut generated by the oracle. The condition number of the cut $\mathcal{B}_z^T$ is defined via*

$$\mu_z = \max\{\det \mathcal{B}_z^T u : \ \mathcal{B}_z^T u \succeq 0, \ \|u\| \leq 1\} \tag{5.2}$$

and the condition number of the field of cuts $\{\mathcal{B}_z^T$ for all $z \notin \Omega^*\}$ is defined by

$$\mu = \inf_{z \notin \Omega^*} \mu_z. \tag{5.3}$$

ASSUMPTION 4.

$$\mu > 0.$$

Now let $S_z(y) = \mathcal{B}_z^T(z - y) \succeq 0$ be the slack matrix corresponding to one of the cuts, let y^c be the center of $\mathcal{N}_\varepsilon$, and let u be a vector such that $\|u\| \leq 1$. Then

$$S_z(y^c + \varepsilon u) = S_z(y^c) - \varepsilon \mathcal{B}_z^T u,$$

and as $y^c + \varepsilon u \in \mathcal{N}_\varepsilon \subset \Omega^*$,

$$S_z(y^c + \varepsilon u) \succeq 0,$$

and thus

$$S_z(y^c) \succeq \varepsilon \mathcal{B}_z^T u.$$

In view of Assumption 4 now the following lemma is clear.

LEMMA 5.2. *For any $z \notin \Omega^*$,*

$$\det S_z(y^c) \geq \varepsilon^p \mu,$$

where p is the dimension of the cut.

Now we present the ACCPM algorithm.

ALGORITHM 1. *Given $\Omega_D^0 = \{y \in R^m : (\mathcal{A}^0)^T y \preceq C^0\}$, where $(\mathcal{A}^0)^T y = \sum_i y_i A_i^0$, with*

$$A_i^0 = \begin{pmatrix} \mathbf{Diag}(e_i) & 0 \\ 0 & -\mathbf{Diag}(e_i) \end{pmatrix} \quad \text{and} \quad C^0 = \begin{pmatrix} I_m & 0 \\ 0 & 0_m \end{pmatrix}.$$

Let $k = 0$

1. *Compute an approximate center $\bar{y}^k$ for Ω_D^k.*
2. *If $\bar{y}^k \in \Omega^*$, stop.*
3. *Otherwise, call the oracle for the p_k-dimensional cut $(\mathcal{B}^k)^T y \preceq (\mathcal{B}^k)^T \bar{y}^k$.*
4. *Update the set of localization: $\Omega_D^{k+1} = \{y \in R^m : (\mathcal{A}^{k+1})^T y \preceq C^{k+1}\}$, where $(\mathcal{A}^{k+1})^T y = \sum_i y_i A_i^{k+1}$, with*

$$A_i^{k+1} = \begin{pmatrix} A_i^k & 0 \\ 0 & B_i^k \end{pmatrix} \quad \text{and} \quad C^{k+1} = \begin{pmatrix} C^k & 0 \\ 0 & (\mathcal{B}^k)^T \bar{y}^k \end{pmatrix}. \tag{5.4}$$

Set $k = k + 1$ and go to step 1.

It is worth mentioning that at each iteration k we enlarge the dimension of the cut matrices A_i by p_k when adding the semidefinite cut as a block diagonal. That is, for all k

$$\dim(A_i^k) = 2m + n_k = 2m + \sum_{i=0}^{k-1} p_i$$

and $n_0 = 0$.

6. Convergence of the algorithm. Let us bound the potential function at the new center. We first define the min-potential functions.

Let Ω_P and Ω_D be the current primal and dual feasible sets, respectively. The primal (dual) min-potential function denoted by $\mathcal{P}(\Omega_P)$ ($\mathcal{D}(\Omega_D)$) is the value of the primal (dual) potential function at the analytic center of Ω_P (Ω_D). We have the following theorem.

THEOREM 6.1. *Let $\mathcal{D}(\Omega_D)$ be the dual min-potential function at the current set of localization Ω_D, and let Ω_D^+ be the updated set after adding the p-dimensional semidefinite cut $\mathcal{B}^T y \preceq \mathcal{B}^T \bar{y}$ at a θ-approximate center $\bar{y}$. Then*

$$\mathcal{D}(\Omega_D^+) \geq \mathcal{D}(\Omega_D) - \sum_{i=1}^{p} \log t_i - \mathcal{C}(p,\theta,\alpha), \tag{6.1}$$

where

$$\mathcal{C}(p,\theta,\alpha) = \frac{\theta^2}{2(1-\theta)} - \alpha(1-\theta) - \log(1-\alpha) - p(1+\log\alpha) + p\log p,$$

and

$$t_i = \sqrt{(b_i^i)^T(\mathcal{A}_D\mathcal{A}_D^T)^{-1}b_i^i}, \tag{6.2}$$

where b_i^i is the m-vector defined in (1.1).

Proof. Let $\mathcal{P}(\Omega_P)$ be the primal min-potential function at Ω_P, and let Ω_P^+ be the updated primal feasible set after adding the cut. From the properties of the primal-dual potential function and (4.2)

$$\begin{aligned}\mathcal{D}(\Omega_D^+) &= n + p - \mathcal{P}(\Omega_P^+)\\ &\geq n + p - \phi_P(\bar{X}) + \alpha(1-\theta) + \log(1-\alpha) + \log\det\alpha\tilde{T}.\end{aligned}$$

In view of Lemma 2.2 and the above inequality

$$\mathcal{D}(\Omega_D^+) \geq \mathcal{D}(\Omega_D) + p - \frac{\theta^2}{2(1-\theta)} + \alpha(1-\theta) + \log(1-\alpha) + \log\det\alpha\tilde{T}. \tag{6.3}$$

Recall that

$$\tilde{T} = \arg\min\left\{\frac{p}{2}\mathbf{tr}T\mathcal{V}T - \log\det T\right\} \quad \text{and} \quad \mathbf{tr}\tilde{T}\mathcal{V}\tilde{T} = 1.$$

Thus $\log\det\tilde{T} \geq \log\det T'$ for any positive semidefinite matrix T' with $\mathbf{tr}T'\mathcal{V}T' = 1$. Let

$$T' = \frac{T^{-1}}{\sqrt{\mathbf{tr}T^{-1}\mathcal{V}T^{-1}}},$$

where T is a diagonal matrix made up of $t_k > 0$, defined in (6.2). First we prove that $\mathbf{tr}T^{-1}\mathcal{V}T^{-1} \leq p^2$:

$$\begin{aligned}\mathbf{tr}T^{-1}\mathcal{V}T^{-1} &= \mathbf{tr}T^{-1}\mathcal{B}^T(\mathcal{A}_D\mathcal{A}_D^T)^{-1}\mathcal{B}T^{-1}\\ &= (\mathcal{B}T^{-1})^T(\mathcal{A}_D\mathcal{A}_D^T)^{-1}\mathcal{B}T^{-1}\\ &= \sum_{i,j=1}^{m}(\mathbf{tr}B_iT^{-1})(\mathbf{tr}B_jT^{-1})\left((\mathcal{A}_D\mathcal{A}_D^T)^{-1}\right)_{ij}\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i,j=1}^{m} \left(\sum_{k=1}^{p} \frac{(B_i)_{kk}}{t_k} \right) \left(\sum_{k=1}^{p} \frac{(B_j)_{kk}}{t_k} \right) \left((\mathcal{A}_D \mathcal{A}_D^T)^{-1} \right)_{ij} \\
&= \sum_{l,q=1}^{p} \frac{1}{t_l t_q} \sum_{i,j=1}^{m} (B_i)_{ll} (B_j)_{qq} \left((\mathcal{A}_D \mathcal{A}_D^T)^{-1} \right)_{ij} \\
&= \sum_{l,q} \frac{1}{t_l t_q} (b_l^l)^T (\mathcal{A}_D \mathcal{A}_D^T)^{-1} b_q^q;
\end{aligned}
$$

clearly $(b_l^l)^T (\mathcal{A}_D \mathcal{A}_D^T)^{-1} b_q^q \leq t_l t_q$. Thus

$$
\mathbf{tr} T^{-1} \mathcal{V} T^{-1} \leq p^2. \tag{6.4}
$$

Now

$$
\begin{aligned}
\log\det \tilde{T} &\geq -p \log \sqrt{\mathbf{tr} T^{-1} \mathcal{V} T^{-1}} - \log\det T \\
&\geq -p \log p - \sum_{i=1}^{p} \log t_i.
\end{aligned}
$$

From (6.3), the inequality (6.1) is immediate now. ∎

Theorem 6.1 establishes a bound on the potential function at the new center in terms of p as well as θ and α. Since the values of θ and α are arbitrary within their limit, we can simplify the bound by choosing fixed values for them. Let $\theta = 0.01$ and $\alpha = 0.9$. One can check that

$$
\mathcal{C}(p, \theta, \alpha) \leq p \log(p+1),
$$

and therefore the inequality (6.1) is reduced to

$$
\mathcal{D}(\Omega_D^+) \geq \mathcal{D}(\Omega_D) - p \log(p+1) - \sum_{i=1}^{p} \log t_i. \tag{6.5}
$$

We note that (6.1) is valid for moderate values of θ and α; i.e., for θ close to zero, one should not choose α very close to 1 (e.g., $\alpha < 0.9$ does the job).

At iteration k, let $p = \max\{p_i, i = 1, \ldots, k\}$ and $\mathcal{D}(\Omega_D^k)$ be the dual min-potential function at Ω_D^k; using (6.5), we have

$$
\begin{aligned}
\mathcal{D}(\Omega_D^{k+1}) &\geq \mathcal{D}(\Omega_D^k) - p_k \log(p_k + 1) - \sum_{i=1}^{p_k} \log t_i \\
&\geq \mathcal{D}(\Omega_D^k) - p_k \log(p+1) - \sum_{i=1}^{p_k} \log t_i \\
&\vdots \\
&\geq \mathcal{D}(\Omega_D^0) - n_{k+1} \log(p+1) - \sum_{i=1}^{n_{k+1}} \log t_i. \qquad (6.6)
\end{aligned}
$$

Now we state a series of technical lemmas to construct a bound on the summation term in (6.6).

LEMMA 6.2. *Let $\mathcal{A}\mathcal{A}^T \in \mathcal{S}^m$, with $(\mathcal{A}\mathcal{A}^T)_{ij} = A_i \bullet A_j$, where $A_i \in \mathcal{S}^n$. Then*

$$\mathcal{A}\mathcal{A}^T = \sum_{l,q=1}^{n} a_q^l (a_q^l)^T,$$

where a_q^l is the m-vector defined in (1.1).

Proof. First observe that

$$A_i \bullet A_j = \sum_{q=1}^{n} (\bar{a}_q^i)^T \bar{a}_q^j,$$

where $A_i = (\bar{a}_1^i, \bar{a}_2^i, \ldots, \bar{a}_n^i)$, $\bar{a}_q^i \in \mathcal{R}^n$, $i = 1, \ldots, m$. Now consider n Gram matrices G^q, $q = 1, \ldots, n$, defined by $G_{ij}^q = (\bar{a}_q^i)^T \bar{a}_q^j$. Thus

$$\begin{aligned} \mathcal{A}\mathcal{A}^T &= \sum_{q=1}^{n} G^q \\ &= \sum_{q=1}^{n} \bar{A}_q^T \bar{A}_q, \end{aligned}$$

where $\bar{A}_q = \left(\bar{a}_q^1, \bar{a}_q^2, \ldots, \bar{a}_q^m\right)$. $\bar{A}_q^T \bar{A}_q$ can alternatively be expressed by the summation of a number of rank one matrices:

$$G^q = \sum_{l=1}^{n} a_q^l (a_q^l)^T,$$

where a_q^l is the row l of $\bar{A}_q$, i.e., $a_q^l = ((A_1)_{lq}, (A_2)_{lq}, \ldots, (A_m)_{lq})$. ∎

LEMMA 6.3.

$$\mathcal{A}_D \mathcal{A}_D^T \succeq \frac{1}{(\mathbf{tr}S)^2} \mathcal{A}\mathcal{A}^T.$$

Proof. Consider the quadratic form associated with $\mathcal{A}_D\mathcal{A}_D^T - \frac{1}{(\mathbf{tr}S)^2}\mathcal{A}\mathcal{A}^T$.

For $y \in R^m$

$$f(y) = y^T \left(\mathcal{A}_D \mathcal{A}_D^T - \frac{1}{(\mathbf{tr}S)^2} \mathcal{A}\mathcal{A}^T \right) y.$$

Let $\mathcal{A}_D^T y = W$; then $\mathcal{A}^T y = S^{.5} W S^{.5}$ and we have

$$\begin{aligned} f(y) &= W \bullet W - \frac{1}{(\mathbf{tr}S)^2} (S^{.5} W S^{.5}) \bullet (S^{.5} W S^{.5}) \\ &= \|W\|^2 - \frac{1}{(\mathbf{tr}S)^2} \|S^{.5} W S^{.5}\|^2 \\ &\geq \|W\|^2 - \frac{1}{(\mathbf{tr}S)^2} \|S^{.5}\|^4 \|W\|^2 \\ &= 0. \end{aligned}$$

Therefore the quadratic form is nonnegative for any $y \in R^m$. ∎

LEMMA 6.4. *At the kth iteration of the ACCPM algorithm*

$$\mathcal{A}_D^k(\mathcal{A}_D^k)^T \succeq 8I + \frac{1}{m^2}\sum_{l,q=1}^{n_k} b_q^l (b_q^l)^T, \tag{6.7}$$

where $b_q^l = ((B_1)_{lq}, (B_2)_{lq}, \ldots, (B_m)_{lq})$ *and matrices* B_i *are block diagonal matrices composed of cut matrices* B_i^r *for* $r = 0, 1, \ldots, k-1$.

Proof. From Algorithm 1 since $\Omega_D^0 = [0,1]^m$

$$\mathcal{A}_D^0(\mathcal{A}_D^0)^T \succeq 8I,$$

and after adding $k+1$ semidefinite cuts $(\mathcal{B}^r)^T y \preceq (\mathcal{B}^r)^T y^r$, $r = 0, 1, \ldots, k$, we have

$$\begin{aligned}\mathcal{A}_D^{k+1}(\mathcal{A}_D^{k+1})^T &= \mathcal{A}_D^k(\mathcal{A}_D^k)^T + \mathcal{B}_D^k(\mathcal{B}_D^k)^T \\ &\succeq 8I + \sum_{r=0}^{k} \mathcal{B}_D^r(\mathcal{B}_D^r)^T, \end{aligned}\tag{6.8}$$

where $(\mathcal{B}_D^r(\mathcal{B}_D^r)^T)_{ij} = \mathbf{tr}(S^r)^{-1}B_i^r(S^r)^{-1}B_j^r$ and $S^r = (\mathcal{B}^r)^T(y^r - y)$.

On one hand, from Lemma 6.3

$$\mathcal{B}_D^r(\mathcal{B}_D^r)^T \succeq \frac{1}{(\mathbf{tr}S^r)^2}\mathcal{B}^r(\mathcal{B}^r)^T,$$

where $(\mathcal{B}^r(\mathcal{B}^r)^T)_{ij} = \mathbf{tr}B_i^r B_j^r$.

From (5.1), we have $\mathbf{tr}S^r = \sum_{i=1}^m (y^r - y)_i \mathbf{tr}B_i^r \leq m$ and therefore

$$\mathcal{B}_D^r(\mathcal{B}_D^r)^T \succeq \frac{1}{m^2}\mathcal{B}^r(\mathcal{B}^r)^T. \tag{6.9}$$

On the other hand, by Lemma 6.2

$$\mathcal{B}^r(\mathcal{B}^r)^T = \sum_{l,q=1}^{p_r} b_q^l (b_q^l)^T. \tag{6.10}$$

The lemma follows now from (6.8)–(6.10). ∎

The next lemma is essential to bound (6.6). This lemma is due to Ye [29] with some changes to suit our case.

LEMMA 6.5. *If* $p \leq m$, *then*

$$\sum_{i=1}^{n_{k+1}} t_i^2 \leq 2m^3 \log\left(8 + \frac{n_{k+1}^2}{m^3}\right) \tag{6.11}$$

for t_i *defined in* (6.2).

Proof. Define

$$\mathcal{H}^{k+1} = \mathcal{H}^k + \frac{1}{m^2}\sum_{i,j=1}^{p_k} b_j^i (b_j^i)^T,$$

where b_j^i is the m-vector defined in (1.1), and let $\mathcal{H}^0 = 8I$.

$$\det \mathcal{H}^{k+1} = \det\left(\mathcal{H}^k + \frac{1}{m^2}\sum_{i,j\in\mathcal{I}_1} b_j^i (b_j^i)^T\right)\left(1 + \frac{r^2}{m^2}\right), \tag{6.12}$$

where $\mathcal{I}_1 = \{i, j = 1, \ldots, p_k \backslash (i, j) = (1, 1)\}$ and

$$r^2 = (b_1^1)^T \left(\mathcal{H}^k + \frac{1}{m^2} \sum_{i,j \in \mathcal{I}_1} b_j^i (b_j^i)^T \right)^{-1} b_1^1.$$

Now we establish a lower bound on r. To this end we study the eigenvalues of

$$\mathcal{G} = I + \frac{1}{m^2} \sum_{i,j \in \mathcal{I}_1} (\mathcal{H}^k)^{-.5} b_j^i (b_j^i)^T (\mathcal{H}^k)^{-.5}.$$

Let $x \in R^m$ with $\|x\| = 1$; then

$$\begin{aligned} x^T \mathcal{G} x &= \|x\|^2 + \frac{1}{m^2} \sum_{i,j \in \mathcal{I}_1} (x^T (\mathcal{H}^k)^{-.5} b_j^i)^2 \\ &\leq \|x\|^2 + \frac{1}{m^2} \sum_{i,j \in \mathcal{I}_1} \|x\|^2 \|(\mathcal{H}^k)^{-.5} b_j^i\|^2 \\ &= 1 + \frac{1}{m^2} \sum_{i,j \in \mathcal{I}_1} (b_j^i)^T (\mathcal{H}^k)^{-1} b_j^i. \end{aligned}$$

Since $\mathcal{H}^k \succeq 8I$

$$x^T \mathcal{G} x \leq 1 + \frac{1}{m^2} \sum_{i,j \in \mathcal{I}_1} \frac{1}{8} \|b_j^i\|^2.$$

From assumption (5.1)

$$x^T \mathcal{G} x \leq 1 + \frac{p_k^2 - 1}{8m^2},$$

and since $p_k \leq p \leq m$, then

$$x^T \mathcal{G} x \leq \frac{9}{8}.$$

That is, $\mathcal{G}^{-1} \succeq (8/9) I$ and therefore

$$\begin{aligned} r^2 &= (b_1^1)^T (\mathcal{H}^k)^{-.5} \mathcal{G}^{-1} (\mathcal{H}^k)^{-.5} b_1^1 \\ &\geq (8/9) r_{11}^2, \end{aligned}$$

where $r_{11}^2 = (b_1^1)^T (\mathcal{H}^k)^{-1} b_1^1$. Now from (6.12)

$$\det \mathcal{H}^{k+1} \geq \left(1 + \frac{8 r_{11}^2}{9m^2} \right) \det \left(\mathcal{H}^k + \frac{1}{m^2} \sum_{i,j \in \mathcal{I}_1} b_j^i (b_j^i)^T \right).$$

By repeating this procedure for each i and j one has

$$\det \mathcal{H}^{k+1} \geq \prod_{i,j=1}^{p_k} \left(1 + \frac{8 r_{ij}^2}{9m^2} \right) \det \mathcal{H}^k,$$

where $r_{ij}^2 = (b_j^i)^T(\mathcal{H}^k)^{-1}b_j^i$.

By taking logarithm from both sides of the above inequality we have

$$\begin{aligned}\log\det\mathcal{H}^{k+1} &\geq \sum_{i,j=1}^{p_k}\log\left(1+\frac{8r_{ij}^2}{9m^2}\right)+\log\det\mathcal{H}^k\\ &\geq \sum_{i=1}^{p_k}\log\left(1+\frac{8r_{ii}^2}{9m^2}\right)+\log\det\mathcal{H}^k.\end{aligned}$$

Since $r_{ii}^2 \leq 1/8$, then $\frac{8r_{ii}^2}{9m^2} \leq \frac{1}{9}$ and

$$\log\left(1+\frac{8r_{ii}^2}{9m^2}\right) \geq \frac{8r_{ii}^2}{9m^2} - \frac{\left(\frac{8r_{ii}^2}{9m^2}\right)^2}{2\left(1-\frac{8r_{ii}^2}{9m^2}\right)} \geq \frac{r_{ii}^2}{2m^2}$$

and therefore

$$\begin{aligned}\log\det\mathcal{H}^{k+1} &\geq \sum_{i=1}^{p_k}\frac{r_{ii}^2}{2m^2}+\log\det\mathcal{H}^k\\ &\geq \sum_{i=1}^{n_{k+1}}\frac{r_{ii}^2}{2m^2}+\log\det\mathcal{H}^0.\end{aligned} \tag{6.13}$$

On the other hand, using the arithmetic-geometric inequality

$$\log\det\mathcal{H}^{k+1} = \log\prod_{j=1}^{m}\lambda_j(\mathcal{H}^{k+1}) \leq m\log\frac{\mathbf{tr}\mathcal{H}^{k+1}}{m}$$

and from the definition of $\mathcal{H}^k$ and assumption (5.1)

$$\mathbf{tr}\mathcal{H}^{k+1} = \mathbf{tr}\left(8I+\frac{1}{m^2}\sum_{i,j=1}^{n_{k+1}}b_j^i(b_j^i)^T\right) \leq 8m+\frac{n_{k+1}^2}{m^2};$$

thus

$$\log\det\mathcal{H}^{k+1} \leq m\log\left(8+\frac{n_{k+1}^2}{m^3}\right).$$

This inequality together with (6.13) gives

$$\sum_{l=1}^{n_{k+1}}r_{ii}^2 \leq 2m^3\log\left(8+\frac{n_{k+1}^2}{m^3}\right). \tag{6.14}$$

In view of Lemma 6.4, $\mathcal{A}_D^k(\mathcal{A}_D^k)^T \succeq \mathcal{H}^k$ and thus

$$(b_i^i)^T\left(\mathcal{A}_D^k(\mathcal{A}_D^k)^T\right)^{-1}b_i^i \leq (b_i^i)^T(\mathcal{H}^k)^{-1}b_i^i$$

or $t_i^2 \leq r_{ii}^2$. The proof follows from (6.14) now. ∎

In the next theorem we prove the main result of this paper; i.e., we derive a bound on n_k, the dimension of the accumulated block diagonal cut matrix.

THEOREM 6.6. *The ACCPM algorithm stops with a solution in* Ω^* *when*

$$n_k \sim O^*\left(\frac{p^2m^3}{\mu^2\varepsilon^2}\right),$$

where in O^* *the lower-order terms are ignored.*

Proof. Consider the kth iteration of the algorithm. Since the analytic center is the minimizer of the dual potential function and since $\bar{y}^j \notin \Omega^*$ for $j = 0, 1, \ldots, k-1$, in view of Lemma 5.2 and (5.2) we have

$$\begin{aligned}\mathcal{D}(\Omega_D^k) &\le -\log\det(C^k - (\mathcal{A}^k)^T y^c)\\ &= -\log\det(C^0 - (\mathcal{A}^0)^T y^c) - \sum_{j=0}^{k-1}\log\det((\mathcal{B}^j)^T(\bar{y}^j - y^c))\\ &\le -(2m + n_k)\log\varepsilon - k\log\mu.\end{aligned}$$

Notice that if $\mu \ge 1$, this parameter can simply be eliminated from the above inequality. We therefore consider the worst-case complexity where $\mu < 1$. Now from inequality (6.6)

$$(2m + n_{k+1})\log\mu\varepsilon \le 2m\log\frac{1}{2} + n_{k+1}\log(p+1) + \frac{1}{2}\sum_{i=1}^{n_{k+1}}\log t_i^2$$

or

$$\begin{aligned}\log\mu\varepsilon - \log(p+1) &\le \frac{1}{2(2m+n_{k+1})}\left(2m\log\frac{1}{4} + \sum_{i=1}^{n_{k+1}}\log t_i^2\right)\\ &\le \frac{1}{2}\log\frac{\frac{m}{2} + \sum_{i=1}^{n_{k+1}} t_i^2}{2m + n_{k+1}}.\end{aligned}$$

Note that the second inequality is due to the arithmetic-geometric mean. Finally, by Lemma 6.5

$$\left(\frac{\mu\varepsilon}{p+1}\right)^2 \le \frac{\frac{m}{2} + 2m^3\log\left(8 + \frac{n_{k+1}^2}{m^3}\right)}{2m + n_{k+1}}.$$

The algorithm stops with a solution in Ω^* when this inequality is violated. In other words, it stops when $n_k \sim O^*(p^2m^3/\mu^2\varepsilon^2)$. ∎

We complete our analysis by bounding the number of damped Newton steps needed to solve problem (3.12).

7. Complexity of the recentering direction. Let $\bar{y}$ be an approximate center of Ω_D and consider a p-dimensional semidefinite cut at $\bar{y}$. Let

$$F(T) = \frac{p}{2}\mathbf{tr}T\mathcal{V}T - \log\det T.$$

Recall that the optimal restoration direction is obtained by minimizing this function over the positive semidefinite cone. In this section we analyze the behavior of the Newton method as applied to F.

We first prove that the (dual) feasible region is contained in an enlarged Dikin ellipsoid. This result is used to construct an upper bound on the functional gap of F at its optimal and initial points.

LEMMA 7.1. *Let $(\bar{X}, \bar{y}, \bar{S})$ be a θ-approximate analytic center of Ω_D. Then*

$$\|\bar{S}^{-.5}\mathcal{A}^T(y-\bar{y})\bar{S}^{-.5}\| \leq \frac{1+\theta}{1-\theta}(n+1)$$

for any $y \in \Omega_D$. In other words, the current set of localization is contained in a Dikin ellipsoid centered at $\bar{y}$ and enlarged by a factor of $\frac{(1+\theta)(n+1)}{1-\theta}$.

Proof. Let $y \in \Omega_D$ be dual feasible, and let $S = C - \mathcal{A}^T y$. From the properties of matrix norm one can prove that (see the proof of (3.14))

$$\|\bar{S}^{-1}(S-\bar{S})\| \leq \|\bar{S}^{-1}\bar{X}^{-1}\|_\infty \|\bar{X}(S-\bar{S})\|. \tag{7.1}$$

Since $\bar{X} \bullet (S - \bar{S}) = 0$,

$$\|\bar{X}(S-\bar{S}) + I\|^2 = \|\bar{X}(S-\bar{S})\|^2 + n,$$

and therefore

$$\|\bar{X}(S-\bar{S})\| \leq \|\bar{X}S\| + \|\bar{X}\bar{S} - I\|. \tag{7.2}$$

However,

$$\begin{aligned}
\|\bar{X}S\|^2 &= \mathbf{tr}(\bar{X}^{.5}S\bar{X}^{.5})^2 \\
&= \sum_i \lambda_i^2(\bar{X}^{.5}S\bar{X}^{.5}) \\
&\leq \Big(\sum \lambda_i(\bar{X}^{.5}S\bar{X}^{.5})\Big)^2 \\
&= (\bar{X} \bullet S)^2,
\end{aligned}$$

and since $\mathcal{A}\bar{X} = 0$, then $\|\bar{X}S\| = \bar{X} \bullet \bar{S}$.

Now from $\|\bar{X}\bar{S} - I\| \leq \theta$ and

$$\begin{aligned}
\|\bar{X}\bar{S} - I\|^2 &= \|\bar{S}^{.5}\bar{X}\bar{S}^{.5} - I\|^2 \\
&= \sum_i \lambda_i^2(\bar{S}^{.5}\bar{X}\bar{S}^{.5} - I),
\end{aligned}$$

one has

$$1 - \theta \leq \lambda_i(\bar{X}\bar{S}) \leq 1 + \theta,$$

and thus

$$\bar{X} \bullet \bar{S} \leq (1+\theta)n \quad \text{and} \quad \|\bar{S}^{-1}\bar{X}^{-1}\|_\infty \leq \frac{1}{1-\theta}.$$

The above inequalities along with (7.1) and (7.2) prove the lemma. ∎

In the next theorem we derive a bound on the number of iterations of the Newton method as applied to $F(T)$.

THEOREM 7.2. *Let $T^0 = \frac{T^{-1}}{\sqrt{\mathbf{tr}T^{-1}\mathcal{V}T^{-1}}}$, where T is the diagonal matrix defined in Theorem* 6.1. *Then starting from T^0 the Newton method finds the optimal updating direction in at most*

$$O\left(\frac{p\log\frac{1}{\varepsilon^*} + \log\frac{1}{\mu}}{\beta - \log(1+\beta)}\right)$$

iterations, where

$$\varepsilon^* = \frac{(1-\theta)\varepsilon}{(1+\theta)(1+n)},$$

β is the Newton decrement, and $\mu > 0$ is the condition number for the field of cuts defined by (5.3).

Proof. Let $\tilde{T}$ and $\tilde{\Lambda}$ be the optimal solutions of problems (3.12) and (3.1), respectively. We first derive an upper bound on the functional gap of F at T^0 and $\tilde{T}$. Observe that

$$\begin{aligned} F(T^0) &= \frac{p}{2}\mathbf{tr}T^0\mathcal{V}T^0 - \log\det T^0 \\ &= \frac{p}{2} + p\log\sqrt{\mathbf{tr}T^{-1}\mathcal{V}T^{-1}} - \log\det T^{-1}, \end{aligned}$$

and from (6.4) and the definition of T

$$F(T^0) \le \frac{p}{2} + p\log p + \sum_{i=1}^{p}\log t_i.$$

Since $\mathcal{A}_D\mathcal{A}_D^T \succeq 8I$, then $\sum \log t_i \le (p/2)\log(1/8) \le 0$. Thus

$$F(T^0) \le \frac{p}{2} + p\log p. \tag{7.3}$$

On the other hand, recall that $p\tilde{T}\tilde{\Lambda} = I$; thus

$$F(\tilde{T}) - \log\det\tilde{\Lambda} = \frac{p}{2} + p\log p. \tag{7.4}$$

Let us construct an upper bound on $-\log\det\tilde{\Lambda}$. From Lemma 7.1, the updated set of localization Ω_D^+ is contained in a Dikin ellipsoid enlarged by a factor of $\frac{(1+\theta)(n+1)}{1-\theta}$. By shrinking the Dikin ellipsoid with a factor of $\frac{1-\theta}{(1+\theta)(n+1)}$ at $\bar{y}$ and noting that Ω_D^+ contains a ball with radius ε, one can prove that

$$\Omega_D^+ \cap \{y \in \mathcal{R}^m : \|\bar{S}^{-.5}\mathcal{A}^T(y-\bar{y})\bar{S}^{-.5}\| \le 1\}$$

contains a ball $\mathcal{N}_{\varepsilon^*}$ with radius $\varepsilon^* = \frac{(1-\theta)\varepsilon}{(1+\theta)(n+1)}$. This set is the feasible region of problem (3.1). Let y^* be the center of $\mathcal{N}_{\varepsilon^*}$. Then $y^* + \varepsilon^* u \in \mathcal{N}_{\varepsilon^*}$ for any u such that $\|u\| \le 1$. In view of Assumption 4, following the same line of argument as in Lemma 5.2, we have

$$-\log\det\tilde{\Lambda} \le p\log\frac{1}{\varepsilon^*} + \log\frac{1}{\mu}.$$

Now from (7.3), (7.4), and the above inequality, one has

$$F(T^0) - F(\tilde{T}) \le p\log\frac{1}{\varepsilon^*} + \log\frac{1}{\mu}. \tag{7.5}$$

Now observe that $F(T)$ is composed of a self-concordant barrier and a convex quadratic function and due to the stability of the self-concordant functions under summation [17, Prop. 2.1.1], $F(T)$ is a self-concordant function on $\mathcal{S}_+^p$. Using Theorem

2.2.3 in [17] one can prove that the Newton algorithm with step size $\frac{1}{1+\beta}$ reduces the value of $F(T)$ by a constant amount $(\beta - \log(1+\beta))$ at each iteration, where $\beta \geq 1$ is the Newton decrement; and the convergence rate becomes quadratic when the iteration is close to the optimal solution.

Thus, we have

$$F(T^+) \leq F(T) - (\beta - \log(1+\beta)), \tag{7.6}$$

where $T^+ = T + \frac{1}{1+\beta} dT$, and $\beta \geq 1$.

The theorem now follows from (7.5) and (7.6). □

8. Conclusion. In this paper we introduced a nonpolyhedral model to the ACCPM by means of semidefinite cuts. The ingredients of the ACCPM, such as computing the analytic center, recentering procedure, and restoration direction, were modified accordingly.

For the purpose of the complexity result, we analyzed the ACCPM in the context of convex feasibility problem with p-dimensional semidefinite cuts. We derived the optimal updating direction dy after adding cuts by maximizing the log det of the new slack matrix. The current center along with dy generate an interior point of the updated set of localization. Starting from this point, we proved that $O(p \log(p+1))$ Newton iterations suffice to recover the analytic center.

A semidefinite cut contains as special cases linear and quadratic cuts. However, the complexity of recentering is recovered by this generalization only in case of linear cuts. More precisely, this complexity when $p = 1$ (single linear cut) is $O(1)$, and when the cut matrices are diagonal (p linear cuts) it is $O(p \log(p+1))$. On the other hand, the reduction of a p-dimensional semidefinite cut to a single quadratic cut does not give the best complexity of recentering [20].

The implementation of the algorithm to an optimization problem and the case of the deep cuts are yet to be tested.

Acknowledgments. The authors gratefully acknowledge discussion with Zhi-Quan Luo on Assumption 4 and Lemma 5.2. We would also like to thank an anonymous reviewer for pointing out the error in the upper bound of the dual min-potential function in the earlier version of this paper.

REFERENCES

[1] F. Alizadeh, *Interior point methods in semidefinite programming with applications to combinatorial optimization*, SIAM J. Optim., 5 (1995), pp. 13–51.

[2] D. S. Atkinson and P. M. Vaidya, *A cutting plane algorithm that uses analytic centers*, Math. Programming, 69 (1995), pp. 1–43.

[3] S. J. Benson, Y. Ye, and X. Zhang, *Solving large-scale sparse semidefinite programs for combinatorial optimization*, SIAM J. Optim., 10 (2000), pp. 443–461.

[4] S. K. Chua, K. C. Toh, and G. Y. Zhao, *An Analytic Center Cutting Plane Method with Deep Cuts for Semidefinite Feasibility Problems*, Technical report, Department of Mathematics, National University of Singapore, Republic of Singapore, 2002.

[5] J.-L. Goffin, J. Gondzio, R. Sarkissian, and J.-P. Vial, *Solving nonlinear multicommodity flow problems by the analytic center cutting plane method*, Math. Programming, 76 (1997), pp. 131–154.

[6] J.-L. Goffin, A. Haurie, and J.-P. Vial, *Decomposition and nondifferentiable optimization with the projective algorithm*, Management Science, 38 (1992), pp. 284–302.

[7] 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.

[8] J.-L. Goffin and J.-P. Vial, *Multiple cuts in the analytic center cutting plane method*, SIAM J. Optim., 11 (2000), pp. 266–288.

[9] J. Gondzio, *Warm start of the primal–dual method applied in the cutting plane scheme*, Math. Programming, 83 (1998), pp. 125–143.
[10] C. Helmberg, *Semidefinite Programming for Combinatorial Optimization*, Konrad-Zuse-Zentrum für Informationstechnik, Berlin, 2000.
[11] C. Helmberg and F. Rendl, *A spectral bundle method for semidefinite programming*, SIAM J. Optim., 10 (2000) pp. 673–696.
[12] Z.-Q. Luo and J. Sun, *An analytic center based column generation algorithm for convex quadratic feasibility problems*, SIAM J. Optim., 9 (1998) pp. 217–235.
[13] H.-J. Lüthi and B. Büeler, *The analytic center quadratic cut method for strongly monotone variational inequality problems*, SIAM J. Optim., 10 (2000), pp. 415–426.
[14] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications*, Academic Press, San Diego, 1979.
[15] J. E. Mitchell and M. J. Todd, *Solving combinatorial optimization problems using Karmarkar's algorithm*, Math. Programming, 56 (1992), pp. 245–284.
[16] Y. Nesterov, *Cutting plane algorithms from analytic centers: Efficiency estimates*, Math. Programming, 69 (1995), pp. 149–176.
[17] Y. Nesterov and A. Nemirovskii, *Interior Point Polynomial Algorithms in Convex Programming*, SIAM Stud. Appl. Math. 13, SIAM, Philadelphia, 1994.
[18] I. S. Pressman and S. Jibrin, *A weighted analytic center for linear matrix inequalities*, JIPAM J. Ineq. Pure Appl. Math., 2 (2001), article 29.
[19] C. Roos, T. Terlaky, and J.-P. Vial, *Theory and Algorithms for Linear Optimization*, John Wiley, Chichester, England, 1997.
[20] F. Sharifi Mokhtarian and J.-L. Goffin, *An analytic center quadratic cut method for the convex quadratic feasibility problem*, Math. Program., 93 (2002), pp. 305–325.
[21] G. Sonnevend, *New Algorithms in Convex Programming Based on a Notation of Center and on Rational Extrapolations*, Internat. Schriftenreihe Numer. Math. 84, Birkhäuser-Verlag, Basel, Switzerland, 1988, pp. 311–327.
[22] J. Sun, K. C. Toh, and G. Y. Zhao, *An analytic center cutting plane method for semidefinite feasibility problems*, Math. Oper. Res., 27 (2002), pp. 332–346.
[23] C. M. Theobald, *An inequality for the trace of the product of two symmetric matrices*, Math. Proc. Cambridge Philos. Soc., 77 (1975), pp. 265–267.
[24] K. C. Toh, *Primal-dual path-following algorithms for determinant maximization problems with linear matrix inequalities*, Comput. Optim. Appl., 14 (1999), pp. 309–330.
[25] K.-C. Toh, G. Zhao, and J. Sun, *A multiple-cut analytic center cutting plane method for semidefinite feasibility problems*, SIAM J. Optim., 12 (2002), pp. 1126–1146.
[26] L. Vandenberghe and S. Boyd, *Semidefinite programming*, SIAM Rev., 38 (1996), pp. 49–95.
[27] L. Vandenberghe, S. Boyd, and S.-P. Wu, *Determinant maximization with linear matrix inequality constraints*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 499–533.
[28] Y. Ye, *A potential reduction algorithm allowing column generation*, SIAM J. Optim., 2 (1992), pp. 7–20.
[29] Y. Ye, *Complexity analysis of the analytic center cutting plane method that uses multiple cuts*, Math. Programming, 78 (1997), pp. 85–104.
[30] Y. Ye, *Interior Point Algorithms, Theory and Analysis*, John Wiley, New York, 1997.

SIAM J. OPTIM.
Vol. 13, No. 4, pp. 1054–1079

A VARIANT OF THE VAVASIS–YE LAYERED-STEP INTERIOR-POINT ALGORITHM FOR LINEAR PROGRAMMING*

RENATO D. C. MONTEIRO† AND TAKASHI TSUCHIYA‡

Abstract. In this paper we present a variant of Vavasis and Ye's layered-step path-following primal-dual interior-point algorithm for linear programming. Our algorithm is a predictor–corrector-type algorithm which uses from time to time the layered least squares (LLS) direction in place of the affine scaling (AS) direction. It has the same iteration-complexity bound of Vavasis and Ye's algorithm, namely $\mathcal{O}(n^{3.5}\log(\bar{\chi}_A+n))$, where n is the number of nonnegative variables and $\bar{\chi}_A$ is a certain condition number associated with the constraint matrix A. Vavasis and Ye's algorithm requires explicit knowledge of $\bar{\chi}_A$ (which is very hard to compute or even estimate) in order to compute the layers for the LLS direction. In contrast, our algorithm uses the AS direction at the current iterate to determine the layers for the LLS direction, and hence does not require the knowledge of $\bar{\chi}_A$. A variant with similar properties and with the same complexity has been developed by Megiddo, Mizuno, and Tsuchiya [*Math. Programming*, 82 (1998), pp. 339–355]. However, their algorithm needs to compute n LLS directions on every iteration, while ours computes at most one LLS direction on any given iteration.

Key words. interior-point algorithms, primal-dual algorithms, path-following, central path, layered steps, condition number, polynomial complexity, predictor-corrector, affine scaling, strongly polynomial, linear programming

AMS subject classifications. 65K05, 68Q25, 90C05, 90C51, 90C60

PII. S1052623401388926

1. Introduction.

We consider the linear programming (LP) problem

$$\begin{array}{ll} \text{minimize}_x & c^T x \\ \text{subject to} & Ax = b, \quad x \geq 0, \end{array} \tag{1}$$

and its associated dual problem

$$\begin{array}{ll} \text{maximize}_{(y,s)} & b^T y \\ \text{subject to} & A^T y + s = c, \quad s \geq 0, \end{array} \tag{2}$$

where $A \in \Re^{m\times n}$, $c \in \Re^n$, and $b \in \Re^m$ are given, and the vectors $x, s \in \Re^n$, and $y \in \Re^m$ are the unknown variables. This paper proposes a primal-dual layered-step predictor-corrector interior-point algorithm that is a variant of the Vavasis–Ye layered-step interior-point algorithm proposed in [26, 27].

Karmarkar in his seminal paper [5] proposed the first polynomially convergent interior-point method with an $\mathcal{O}(nL)$ iteration-complexity bound, where L is the size

*Received by the editors May 4, 2001; accepted for publication (in revised form) October 2, 2002; published electronically March 19, 2003.
http://www.siam.org/journals/siopt/13-4/38892.html

†School of ISyE, Georgia Institute of Technology, Atlanta, GA 30332 (monteiro@isye.gatech.edu). This author was supported in part by NSF grants CCR-9902010, CCR-0203113, and INT-9910084, and ONR grant N00014-03-1-0401.

‡The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-Ku, Tokyo, 106-8569, Japan (tsuchiya@sun312.ism.ac.jp). This author was supported in part by Japan-US Joint Research Projects of Japan Society for the Promotion of Science "Algorithms for linear programs over symmetric cones" and the Grant-in-Aid for Scientific Research (C) 08680478 of the Ministry of Science, Technology, Education, and Culture of Japan.

of the LP instance (1). The first path-following interior-point algorithm was proposed by Renegar in his breakthrough paper [16]. Renegar's method closely follows the primal central path and exhibits an $\mathcal{O}(\sqrt{n}L)$ iteration-complexity bound. The first path-following algorithm which simultaneously generates iterates in both the primal and dual spaces has been proposed by Kojima, Mizuno, and Yoshise [6] and Tanabe [18], based on ideas suggested by Megiddo [9]. In contrast to Renegar's algorithm, Kojima, Mizuno, and Yoshise's algorithm has an $\mathcal{O}(nL)$ iteration-complexity bound. A primal-dual path-following with an $\mathcal{O}(\sqrt{n}L)$ iteration-complexity bound was subsequently obtained by Kojima, Mizuno, and Yoshise [7] and Monteiro and Adler [13, 14] independently. Following these developments, many other primal-dual interior-point algorithms for linear programming have been proposed.

An outstanding open problem in optimization is whether there exists a strongly polynomial algorithm for linear programming, that is, one whose complexity is bounded by a polynomial of m and n only. A major effort in this direction is due to Tardos [19], who developed a polynomial-time algorithm whose complexity is bounded by a polynomial of m, n, and L_A, where L_A denotes the size of A. Such an algorithm gives a strongly polynomial method for the important class of LP problems where the entries of A are either 1, -1, or 0, e.g., LP formulations of network flow problems. Tardos's algorithm consists of solving a sequence of "low-sized" LP problems by a standard polynomially convergent LP method and using their solutions to obtain the solution of the original LP problem.

The development of a method which works entirely in the context of the original LP problem and whose complexity is also bounded by a polynomial of m, n, and L_A is due to Vavasis and Ye [26]. Their method is a primal-dual path-following interior-point algorithm similar to the ones mentioned above except that it uses from time to time a crucial step, namely the layered least squares (LLS) direction. They showed that their method has an $\mathcal{O}(n^{3.5}(\log \bar{\chi}_A + \log n))$ iteration-complexity bound, where $\bar{\chi}_A$ is a condition number associated with A having the property that $\log \bar{\chi}_A = \mathcal{O}(L_A)$. The number $\bar{\chi}_A$ was first introduced implicitly by Dikin and Zorkalcev [1] in the study of primal affine scaling algorithms and was later studied by several researchers including Vanderbei and Lagarias [25], Todd [20], and Stewart [17]. Properties of $\bar{\chi}_A$ are studied in [3, 23, 24].

The complexity analysis of Vavasis and Ye's algorithm is based on the notion of a crossover event, a combinatorial event concerning the central path. Intuitively, a crossover event occurs between two variables when one of them is larger than the other at a point in the central path and then becomes smaller asymptotically as the optimal solution set is approached. Vavasis and Ye showed that there can be at most $n(n-1)/2$ crossover events and that a distinct crossover event occurs every $\mathcal{O}(n^{1.5}(\log \bar{\chi}_A + \log n))$ iterations, from which they deduced the overall $\mathcal{O}(n^{3.5}(\log \bar{\chi}_A + \log n))$ iteration-complexity bound. In [12], an LP instance is given where the number of crossover events is $\Theta(n^2)$.

One disadvantage of Vavasis and Ye's method is that it requires the explicit knowledge of $\bar{\chi}_A$ in order to determine a partition of the variables into layers used in the computation of the LLS step. This difficulty was remedied in a variant proposed by Megiddo, Mizuno, and Tsuchiya [10] which does not require the explicit knowledge of the number $\bar{\chi}_A$. They observed that at most n types of partitions arise as $\bar{\chi}_A$ varies from 1 to ∞ and that one of these can be used to compute the LLS step. Based on this idea, they developed a variant which computes the LLS steps for all these partitions and picks the one that yields the greatest duality gap reduction at

the current iteration. Moreover, using the argument that once the first LLS step is computed the other ones can be cheaply computed by performing rank-one updates, they show that the overall complexity of their algorithm is exactly the same as Vavasis and Ye's algorithm.

In this paper, we propose another variant of Vavasis and Ye's algorithm which has the same complexity as theirs and computes only one LLS step per iteration without any explicit knowledge of $\bar{\chi}_A$. Our algorithm is a predictor–corrector-type algorithm like the one described in [11] except that at the predictor stage it takes a step along either the primal-dual affine scaling (AS) step or the LLS step. More specifically, first the AS direction is computed and a test involving this direction is performed to determine whether the LLS step is needed. If the LLS direction is not needed, a step along the AS direction is taken as usual. Otherwise, the AS direction is used to determine a partition of the variables into layers, and the LLS step with respect to this partition is computed. The algorithm then takes a step along the direction (either the AS or the LLS) which yields the largest duality gap reduction.

It is worth noting that our algorithm computes the LLS step only when a step along the AS direction has the potential to yield a significant duality gap decrease. In such a case, the LLS direction seems to be even better suited and it is used whenever the current iteration permits it. Another advantage of the LLS step is that it possesses the ability to determine an exact primal-dual optimal solution, and hence imply finite termination of the algorithm.

The organization of the paper is as follows. Section 2 consists of five subsections. In subsection 2.1, we review the notion of the primal-dual central path and its associated two-norm neighborhoods. Subsection 2.2 introduces the notion of the condition number $\bar{\chi}_A$ of a matrix A and describes the properties of $\bar{\chi}_A$ that will be useful in our analysis. Subsection 2.3 reviews the AS step and the corrector (or centrality) step which are the basic ingredients of several well-known interior-point algorithms. Subsection 2.4 describes the LLS step. Subsection 2.5 describes our algorithm in detail and states the main convergence result of this paper. Section 3, which consists of three subsections, introduces some basic tools which are used in our convergence analysis. Subsection 3.1 discusses the notion of crossover events. Subsection 3.2 states an approximation result that provides an estimation of the closeness between the AS direction and the LLS direction. Subsection 3.3 reviews from a different perspective an important result from Vavasis and Ye [26], which basically provides sufficient conditions for the occurrence of crossover events. Section 4 is dedicated to the proof of the main result stated in subsection 2.5. Section 5 gives some concluding remarks. Finally, the appendix gives the proof of the approximation result between the AS and the LLS directions stated in subsection 3.2.

The following notation is used throughout our paper. We denote the vector of all ones by e. Its dimension is always clear from the context. The symbols $\Re^n$, $\Re^n_+$, and $\Re^n_{++}$ denote the n-dimensional Euclidean space, the nonnegative orthant of $\Re^n$, and the positive orthant of $\Re^n$, respectively. The set of all $m \times n$ matrices with real entries is denoted by $\Re^{m\times n}$. If J is a finite index set, then $|J|$ denotes its cardinality, that is, the number of elements of J. For $J \subseteq \{1,\dots,n\}$ and $w \in \Re^n$, we let w_J denote the subvector $[w_i]_{i\in J}$; moreover, if E is an $m \times n$ matrix, then E_J denotes the $m \times |J|$ submatrix of E corresponding to J. For a vector $w \in \Re^n$, we let $\max(w)$ and $\min(w)$ denote the largest and the smallest component of w, respectively, $\mathrm{Diag}(w)$ denote the diagonal matrix whose ith diagonal element is w_i for $i = 1,\dots,n$, and w^{-1} denote the vector $[\mathrm{Diag}(w)]^{-1}e$ whenever it is well-defined. For two vectors $u, v \in \Re^n$, uv

denotes their Hadamard product, i.e., the vector in $\Re^n$ whose ith component is $u_i v_i$. The Euclidean norm, the 1-norm, and the ∞-norm are denoted by $\|\cdot\|$, $\|\cdot\|_1$, and $\|\cdot\|_\infty$, respectively. For a matrix E, $\mathrm{Im}(E)$ denotes the subspace generated by the columns of E and $\mathrm{Ker}(E)$ denotes the subspace orthogonal to the rows of E. The superscript ^{T} denotes transpose.

2. Problem and primal-dual predictor-corrector interior-point algorithms. In this section we describe the proposed feasible interior-point primal-dual predictor-corrector algorithm for solving the pair of LP problems (1) and (2). We also present the main convergence result which establishes a polynomial iteration-complexity bound for the algorithm that depends only on the constraint matrix A.

This section is divided into five subsections. In subsection 2.1, we describe the primal-dual central path and its associated two-norm neighborhoods. In subsection 2.2, we describe the notion of the condition number of a matrix and describe the properties of the condition number that will be useful in our analysis. In subsection 2.3, we review the AS step and the corrector (or centrality) step which are the basic ingredients of several well-known interior-point algorithms. We also derive a lower bound on the stepsize along the AS step. In subsection 2.4, we describe an alternative step, namely the LLS step, which is sometimes used in place of the AS direction by our algorithm. In subsection 2.5, we describe our algorithm in detail and state the main convergence result of this paper.

2.1. The problem, the central path, and its associated neighborhoods. In this subsection we introduce the pair of dual linear programs and the assumptions used in our development. We also describe the associated primal-dual central path and its corresponding two-norm neighborhoods.

Given $A \in \Re^{m\times n}$, $c \in \Re^n$, and $b \in \Re^m$, consider the pairs of linear programs (1) and (2), where $x \in \Re^n$ and $(y, s) \in \Re^m \times \Re^n$ are their respective variables. The set of strictly feasible solutions for these problems are

$$\mathcal{P}^{++} \equiv \{x \in \Re^n : Ax = b,\ x > 0\},$$
$$\mathcal{D}^{++} \equiv \{(y, s) \in \Re^{m\times n} : A^T y + s = c,\ s > 0\},$$

respectively. Throughout the paper we make the following assumptions on the pair of problems (1) and (2).

A.1 $\mathcal{P}^{++}$ and $\mathcal{D}^{++}$ are nonempty.

A.2 The rows of A are linearly independent.

Under the above assumptions, it is well known that for any $\nu > 0$ the system

$$xs = \nu e, \tag{3}$$
$$Ax = b, \quad x > 0, \tag{4}$$
$$A^T y + s = c, \quad s > 0, \tag{5}$$

has a unique solution (x, y, s), which we denote by $(x(\nu), y(\nu), s(\nu))$. The central path is the set consisting of all these solutions as ν varies in $(0, \infty)$. As ν converges to zero, the path $(x(\nu), y(\nu), s(\nu))$ converges to a primal-dual optimal solution (x^*, y^*, s^*) for problems (1) and (2). Given a point $w = (x, y, s) \in \mathcal{P}^{++} \times \mathcal{D}^{++}$, its duality gap and its normalized duality gap are defined as $x^T s$ and $\mu = \mu(x, s) \equiv x^T s/n$, respectively, and the point $(x(\mu), y(\mu), s(\mu))$ is said to be the central point associated with w. Note that $(x(\mu), y(\mu), s(\mu))$ also has normalized duality gap μ. We define the proximity

measure of a point $w = (x, y, s) \in \mathcal{P}^{++} \times \mathcal{D}^{++}$ with respect to the central path by

$$\eta(w) \equiv \|xs/\mu - e\|.$$

Clearly, $\eta(w) = 0$ if and only if $w = (x(\mu), y(\mu), s(\mu))$ or, equivalently, w coincides with its associated central point. The two-norm neighborhood of the central path with opening $\beta > 0$ is defined as

$$\mathcal{N}(\beta) \equiv \{w = (x, y, s) \in \mathcal{P}^{++} \times \mathcal{D}^{++} : \eta(w) \le \beta\}.$$

Finally, for any point $w = (x, y, s) \in \mathcal{P}^{++} \times \mathcal{D}^{++}$ we define

$$\delta(w) \equiv s^{1/2} x^{-1/2} \in \Re^n. \tag{6}$$

The following proposition provides important estimates which are used throughout our analysis.

PROPOSITION 2.1. *Let $w = (x, y, s) \in \mathcal{N}(\beta)$ for some $\beta \in (0, 1)$ be given and define $\delta \equiv \delta(w)$. Let $w(\mu) = (x(\mu), y(\mu), s(\mu))$ be the central point associated with w. Then*

$$\frac{1-\beta}{1+\beta}\, s \le s(\mu) \le \frac{1}{1-\beta}\, s, \qquad \frac{1-\beta}{1+\beta}\, x \le x(\mu) \le \frac{1}{1-\beta}\, x, \tag{7}$$

$$\frac{1-\beta}{(1+\beta)^{1/2}}\, \delta \le \frac{s(\mu)}{\sqrt{\mu}} \le \frac{(1+\beta)^{1/2}}{1-\beta}\, \delta, \tag{8}$$

$$\frac{(1-\beta)^2}{(1+\beta)}\, \frac{\delta_i}{\delta_j} \le \frac{s_i(\mu)}{s_j(\mu)} \le \frac{(1+\beta)}{(1-\beta)^2}\, \frac{\delta_i}{\delta_j} \quad \forall i, j \in \{1, \ldots, n\}. \tag{9}$$

Proof. The second and fourth inequalities in (7) follow from Lemma 2.4(ii) of Gonzaga [2]. Using these two inequalities together with $xs \le (1+\beta)\mu e$ and $x(\mu)s(\mu) = \mu e$, we obtain the other two inequalities in (7). Using the definition of $\delta = \delta(w)$ in (6) together with the relations $xs \le (1+\beta)\mu e$ and $x(\mu)s(\mu) = \mu e$, we easily see that the first and second inequalities of (8) follow from the fourth and second inequalities of (7), respectively. Finally, (9) immediately follows from (8). □

2.2. Condition number. In this subsection we define a certain condition number associated with the constraint matrix A and state the properties of $\bar{\chi}_A$ which will play an important role in our analysis.

Let $\mathcal{D}$ denote the set of all positive definite $n \times n$ diagonal matrices and define

$$\begin{aligned} \bar{\chi}_A &\equiv \sup\{\|A^T(A\tilde{D}A^T)^{-1}A\tilde{D}\| \ : \tilde{D} \in \mathcal{D}\} \\ &= \sup\Big\{ \tfrac{\|A^T y\|}{\|c\|} : \ y = \operatorname*{argmin}_{\tilde{y} \in \Re^n} \|\tilde{D}^{1/2}(A^T\tilde{y} - c)\| \text{ for some } 0 \ne c \in \Re^n \text{ and } \tilde{D} \in \mathcal{D}\Big\}. \end{aligned} \tag{10}$$

The parameter $\bar{\chi}_A$ plays a fundamental role in the complexity analysis of algorithms for linear programming and least squares problems (see [26] and references therein). Its finiteness has been established first by Dikin and Zorkalcev [1]. Other authors have also given alternative derivations of the finiteness of $\bar{\chi}_A$ (see, for example, Stewart [17], Todd [20], and Vanderbei and Lagarias [25]).

We summarize in the next proposition a few important facts about the parameter $\bar{\chi}_A$.

PROPOSITION 2.2. *Let $A \in \Re^{m \times n}$ with full row rank be given. Then the following statements hold:*

(a) $\bar{\chi}_{GA} = \bar{\chi}_A$ *for any nonsingular matrix* $G \in \Re^{m\times m}$.
(b) $\bar{\chi}_A = \max\{\|G^{-1}A\| : G \in \mathcal{G}\}$, *where* $\mathcal{G}$ *denote the set of all* $m\times m$ *nonsingular submatrices of* A.
(c) *If the entries of* A *are all integers, then* $\bar{\chi}_A$ *is bounded by* $2^{\mathcal{O}(L_A)}$, *where* L_A *is the input bit length of* A.
(d) $\bar{\chi}_A = \bar{\chi}_F$ *for any* $F \in \Re^{(n-m)\times n}$ *such that* $\mathrm{Ker}(A) = \mathrm{Im}(F^T)$.
(e) *If the* $m \times m$ *identity matrix is a submatrix of* A *and* $\tilde{A}$ *is an* $r \times n$ *submatrix of* A, *then* $\|\tilde{G}^{-1}\tilde{A}\| \le \bar{\chi}_A$ *for every* $r \times r$ *nonsingular submatrix* $\tilde{G}$ *of* $\tilde{A}$.

Proof. Statement (a) readily follows from the definition (10). The inequality $\bar{\chi}_A \ge \max\{\|G^{-1}A\| : G \in \mathcal{G}\}$ is established in Lemma 3 of [26], while the proof of the reverse inequality is given in [20] (see also Theorem 1 of [21]). Hence, (b) holds. The proof of (c) can be found in Lemma 24 of [26]. A proof of (d) can be found in [3].

We now consider (e). Using the assumption that the $m \times m$ identity matrix is a submatrix of A, we easily see that A has an $m \times m$ nonsingular submatrix G which, after some symmetric permutation of its rows and columns, can be put into the form

$$\begin{bmatrix} \tilde{G} & 0 \\ \tilde{E} & I \end{bmatrix}$$

for some matrix $\tilde{E} \in \Re^{(n-r)\times r}$. Since the inverse of the above matrix is

$$\begin{bmatrix} \tilde{G}^{-1} & 0 \\ -\tilde{E}\tilde{G}^{-1} & I \end{bmatrix},$$

we easily see that $\tilde{G}^{-1}\tilde{A}$ is a submatrix of $G^{-1}A$. Hence, $\|\tilde{G}^{-1}\tilde{A}\| \le \|G^{-1}A\| \le \bar{\chi}_A$, where the last inequality is due to (b). ∎

We now state a Hoffman-type result for a system of linear equalities that will be used in the proof of an approximation result given in the appendix.

LEMMA 2.3. *Let* $A \in \Re^{m\times n}$ *with full row rank be given, and let* $(\mathcal{K}, \mathcal{L})$ *be an arbitrary bipartition of the index set* $\{1, \ldots, n\}$. *Assume that* $\bar{w} \in \Re^{|\mathcal{L}|}$ *is an arbitrary vector such that the system* $A_{\mathcal{K}}u = A_{\mathcal{L}}\bar{w}$ *is feasible. Then this system has a feasible solution* $\bar{u}$ *such that* $\|\bar{u}\| \le \bar{\chi}_A\|\bar{w}\|$.

Proof. Due to Proposition 2.2(a), it is sufficient to establish the lemma for a matrix of the form GA, where G is an $m \times m$ nonsingular matrix. Hence, we may assume that A contains the $m \times m$ identity matrix. Eliminating some redundant rows from $A_{\mathcal{K}}u = A_{\mathcal{L}}\bar{w}$ and some variables from u, we obtain a nonsingular system

$$\tilde{G}\tilde{u} = \tilde{H}\bar{w},$$

where $\tilde{G}$ is a square submatrix of $A_{\mathcal{K}}$ such that $\mathrm{rank}(\tilde{G}) = \mathrm{rank}(A_{\mathcal{K}})$, $\tilde{H}$ is the corresponding submatrix of $A_{\mathcal{L}}$, and $\tilde{u}$ is a subvector of u. Clearly, the solution $\tilde{u}$ of this system satisfies $\|\tilde{u}\| \le \|\tilde{G}^{-1}\tilde{H}\|\,\|\bar{w}\| \le \bar{\chi}_A\|\bar{w}\|$, where the last inequality follows from Proposition 2.2(e). We can augment $\tilde{u}$ to a solution $\bar{u}$ of $A_{\mathcal{K}}u = A_{\mathcal{L}}\bar{w}$ by setting the components of u not in $\tilde{u}$ to zero. ∎

2.3. Predictor-corrector step and its properties. In this subsection we describe the well-known predictor-corrector (P-C) iteration which is used by several interior-point algorithms (see, for example, Mizuno, Todd, and Ye [11]). We also describe the properties of this iteration which will be used in our analysis.

The P-C iteration consists of two steps, namely the predictor (or AS) step and the corrector (or centrality) step. The search direction used by either step from a current point in $(x, y, s) \in \mathcal{P}^{++} \times \mathcal{D}^{++}$ is the solution of the following linear system of equations:

$$\begin{aligned} S\Delta x + X\Delta s &= \sigma\mu e - xs, \\ A\Delta x &= 0, \\ A^T\Delta y + \Delta s &= 0, \end{aligned} \tag{11}$$

where $\mu = \mu(x,s)$ and $\sigma \in \Re$ is a prespecified parameter, commonly referred to as the centrality parameter. When $\sigma = 0$, we denote the solution of (11) by $(\Delta x^{\rm a}, \Delta y^{\rm a}, \Delta s^{\rm a})$ and refer to it as the primal-dual AS direction at w; it is the direction used in the predictor step of the P-C iteration. When $\sigma = 1$, we denote the solution of (11) by $(\Delta x^{\rm c}, \Delta y^{\rm c}, \Delta s^{\rm c})$ and refer to it as the corrector direction at w; it is the direction used in the corrector step of the P-C iteration.

We are now ready to describe the entire P-C iteration. Suppose that a constant $\beta \in (0, 1/4]$ and a point $w = (x,y,s) \in \mathcal{N}(\beta)$ is given. The P-C iteration generates another point $(x^+, y^+, s^+) \in \mathcal{N}(\beta)$ as follows. It first moves along the direction $(\Delta x^{\rm a}, \Delta y^{\rm a}, \Delta s^{\rm a})$ until it hits the boundary of the enlarged neighborhood $\mathcal{N}(2\beta)$. More specifically, it computes the point $w^{\rm a} = (x^{\rm a}, y^{\rm a}, s^{\rm a}) \equiv (x,y,s) + \alpha_{\rm a}(\Delta x^{\rm a}, \Delta y^{\rm a}, \Delta s^{\rm a})$, where

$$\alpha_{\rm a} \equiv \sup\{\alpha \in [0,1] : (x,y,s) + \alpha(\Delta x^{\rm a}, \Delta y^{\rm a}, \Delta s^{\rm a}) \in \mathcal{N}(2\beta)\}. \tag{12}$$

Next, the P-C iteration generates a point inside the smaller neighborhood $\mathcal{N}(\beta)$ by taking a unit step along the corrector direction $(\Delta x^{\rm c}, \Delta y^{\rm c}, \Delta s^{\rm c})$ at the point $w^{\rm a}$; that is, it computes the point $(x^+, y^+, s^+) \equiv (x^{\rm a}, y^{\rm a}, s^{\rm a}) + (\Delta x^{\rm c}, \Delta y^{\rm c}, \Delta s^{\rm c}) \in \mathcal{N}(\beta)$. The successive repetition of this iteration leads to the so-called Mizuno–Todd–Ye (MTY) P-C algorithm (see [11]).

Our method is very similar to the algorithm of [11] except that it sometimes replaces the AS step by the LLS step described in the next subsection. The insertion of the LLS step in the above MTY P-C algorithm guarantees that the modified method has the finite termination property. Hence, the LLS step can be viewed as a termination procedure which is performed only when some "not-so-likely-to-occur" conditions are met. Moreover, the LLS step is taken only when it yields a point with a smaller duality gap than the one obtained from the AS step as described above.

In the remainder of this subsection, we discuss some properties of the P-C iteration and the primal-dual AS direction. For a proof of the next two propositions, we refer the reader to [11].

PROPOSITION 2.4 (predictor step). *Suppose that $w = (x,y,s) \in \mathcal{N}(\beta)$ for some constant $\beta \in (0, 1/2]$. Let $\Delta w^{\rm a} = (\Delta x^{\rm a}, \Delta y^{\rm a}, \Delta s^{\rm a})$ denote the AS direction at $w^{\rm a}$ and let $\alpha_{\rm a}$ be the stepsize computed according to* (12)*. Then the following statements hold:*

(a) *the point $w + \alpha\Delta w^{\rm a}$ has normalized duality gap $\mu(\alpha) = (1-\alpha)\mu$ for all $\alpha \in \Re$;*

(b) *$\alpha_{\rm a} \geq \sqrt{\beta/n}$, and hence $\mu(\alpha_{\rm a})/\mu \leq 1 - \sqrt{\beta/n}$.*

PROPOSITION 2.5 (corrector step). *Suppose that $w = (x,y,s) \in \mathcal{N}(2\beta)$ for some constant $\beta \in (0, 1/4]$, and let $(\Delta x^{\rm c}, \Delta y^{\rm c}, \Delta s^{\rm c})$ denote the corrector step at w. Then $w + \Delta w^{\rm c} \in \mathcal{N}(\beta)$. Moreover, the (normalized) duality gap of $w + \Delta w^{\rm c}$ is the same as that of w.*

For the purpose of future comparison with the LLS step, we mention the following alternative characterization of the primal-dual AS direction whose verification is straightforward:

$$\Delta x^{\rm a} \equiv \operatorname*{argmin}_{p \in \Re^n} \left\{ \|\delta(x+p)\|^2 : Ap = 0 \right\}, \tag{13}$$

$$(\Delta y^{\rm a}, \Delta s^{\rm a}) \equiv \operatorname*{argmin}_{(r,q) \in \Re^m \times \Re^n} \left\{ \|\delta^{-1}(s+q)\|^2 : A^T r + q = 0 \right\}, \tag{14}$$

where $\delta \equiv \delta(w)$. For a search direction $(\Delta x, \Delta y, \Delta s)$ at a point (x, y, s), the quantity

$$(Rx, Rs) \equiv \left(\frac{\delta(x + \Delta x)}{\sqrt{\mu}}, \frac{\delta^{-1}(s + \Delta s)}{\sqrt{\mu}} \right) = \left(\frac{x^{1/2}s^{1/2} + \delta \Delta x}{\sqrt{\mu}}, \frac{x^{1/2}s^{1/2} + \delta^{-1}\Delta s}{\sqrt{\mu}} \right) \tag{15}$$

appears quite often in our analysis. We refer to it as the *residual* of $(\Delta x, \Delta y, \Delta s)$. Note that if $(Rx^{\mathrm{a}}, Rs^{\mathrm{a}})$ is the residual of $(\Delta x^{\mathrm{a}}, \Delta y^{\mathrm{a}}, \Delta s^{\mathrm{a}})$, then

$$Rx^{\mathrm{a}} = -\frac{1}{\sqrt{\mu}}\delta^{-1}\Delta s^{\mathrm{a}}, \qquad Rs^{\mathrm{a}} = -\frac{1}{\sqrt{\mu}}\delta \Delta x^{\mathrm{a}}, \tag{16}$$

and

$$Rx^{\mathrm{a}} + Rs^{\mathrm{a}} = \frac{x^{1/2}s^{1/2}}{\sqrt{\mu}}, \tag{17}$$

due to the fact that $(\Delta x^{\mathrm{a}}, \Delta y^{\mathrm{a}}, \Delta s^{\mathrm{a}})$ satisfies the first equation in (11) with $\sigma = 0$. The following quantity is used in the test to determine when the LLS step should be used in place of the AS step:

$$\varepsilon^{\mathrm{a}}_{\infty} \equiv \max_i \left\{ \min \left\{ |Rx^{\mathrm{a}}_i| , |Rs^{\mathrm{a}}_i| \right\} \right\}. \tag{18}$$

We end this section by providing some estimates involving the residual of the AS direction.

LEMMA 2.6. *Suppose that $w = (x, y, s) \in \mathcal{N}(\beta)$ for some $\beta \in (0, 1/4]$. Then, for all $i = 1, \ldots, n$, we have*

$$\max \left\{ |Rx^{\mathrm{a}}_i|, |Rs^{\mathrm{a}}_i| \right\} \geq \frac{\sqrt{1-\beta}}{2} \geq \frac{1}{4}.$$

Proof. Assume for contradiction that for some $i \in \{1, \ldots, n\}$, $\max \{|Rx^{\mathrm{a}}_i|, |Rs^{\mathrm{a}}_i|\} < \sqrt{1-\beta}/2$. Then, using (17), we obtain the following contradiction:

$$\frac{x_i^{1/2}s_i^{1/2}}{\sqrt{\mu}} = Rx^{\mathrm{a}}_i + Rs^{\mathrm{a}}_i \leq |Rx^{\mathrm{a}}_i| + |Rs^{\mathrm{a}}_i| < \sqrt{1-\beta} \leq \frac{x_i^{1/2}s_i^{1/2}}{\sqrt{\mu}}. \qquad \square$$

2.4. The LLS step. In this subsection we describe the other type of step used in our algorithm, namely the LLS step. This step was first introduced by Vavasis and Ye in [26].

Let $w = (x, y, s) \in \mathcal{P}^{++} \times \mathcal{D}^{++}$ and a partition $(J_1, \ldots, J_p)$ of the index set $\{1, \ldots, n\}$ be given and define $\delta \equiv \delta(w)$. The primal LLS direction $\Delta x^{\mathrm{ll}} = (\Delta x^{\mathrm{ll}}_{J_1}, \ldots, \Delta x^{\mathrm{ll}}_{J_p})$ at w with the respect to J is defined recursively according to the order $\Delta x^{\mathrm{ll}}_{J_p}, \ldots, \Delta x^{\mathrm{ll}}_{J_1}$ as follows. Assume that the components $\Delta x^{\mathrm{ll}}_{J_p}, \ldots, \Delta x^{\mathrm{ll}}_{J_{k+1}}$ have been determined. Let $\Pi_{J_k} : \Re^n \to \Re^{J_k}$ denote the projection map defined as $\Pi_{J_k}(u) = u_{J_k}$ for all $u \in \Re^n$. Then $\Delta x^{\mathrm{ll}}_{J_k} \equiv \Pi_{J_k}(L^x_k)$, where L^x_k is given by

$$\begin{aligned} L^x_k &\equiv \operatorname*{Argmin}_{p \in \Re^n} \{ \|\delta_{J_k}(x_{J_k} + p_{J_k})\|^2 : p \in L^x_{k-1} \} \\ &= \operatorname*{Argmin}_{p \in \Re^n} \{ \|\delta_{J_k}(x_{J_k} + p_{J_k})\|^2 : p \in \mathrm{Ker}(A),\ p_{J_i} = \Delta x^{\mathrm{ll}}_{J_i} \quad \forall i = k+1, \ldots, p \}, \end{aligned} \tag{19}$$

with the convention that $L^x_0 = \mathrm{Ker}(A)$. The slack component $\Delta s^{\mathrm{ll}} = (\Delta s^{\mathrm{ll}}_{J_1}, \ldots, \Delta s^{\mathrm{ll}}_{J_p})$ of the dual LLS direction $(\Delta y^{\mathrm{ll}}, \Delta s^{\mathrm{ll}})$ at w with the respect to J is defined recursively

as follows. Assume that the components $\Delta s^{\rm ll}_{J_1}, \ldots, \Delta s^{\rm ll}_{J_{k-1}}$ have been determined. Then $\Delta s^{\rm ll}_{J_k} \equiv \Pi_{J_k}(L^s_k)$, where L^s_k is given by

$$\begin{aligned} L^s_k &\equiv \operatorname*{Argmin}_{q\in\Re^n} \{\|\delta_{J_k}^{-1}(s_{J_k}+q_{J_k})\|^2 : q \in L^s_{k-1}\} \\ &= \operatorname*{Argmin}_{q\in\Re^n} \{\|\delta_{J_k}^{-1}(s_{J_k}+q_{J_k})\|^2 : q \in \mathrm{Im}(A^T),\ q_{J_i} = \Delta s^{\rm ll}_{J_i} \quad \forall i = 1,\ldots,k-1\}, \end{aligned} \tag{20}$$

with the convention that $L^s_0 = \mathrm{Im}(A^T)$. Finally, once $\Delta s^{\rm ll}$ has been determined, the component $\Delta y^{\rm ll}$ is determined from the relation $A^T\Delta y^{\rm ll} + \Delta s^{\rm ll} = 0$.

Note that (13) and (14) imply that the AS direction is a special LLS direction, namely the one with respect to the only partition in which $p = 1$. Clearly, the LLS direction at a given $w \in \mathcal{P}^{++} \times \mathcal{D}^{++}$ depends on the partition $J = (J_1, \ldots, J_p)$ used.

A partition $J = (J_1, \ldots, J_p)$ is said to be *ordered* at a point $w = (x, y, s) \in \mathcal{P}^{++} \times \mathcal{D}^{++}$ if $\max(\delta_{J_i}) \le \min(\delta_{J_{i+1}})$ for all $i = 1, \ldots, p-1$. In this case, the gap of J, denoted by $\mathrm{gap}(J)$, is defined as

$$\mathrm{gap}(J) \;=\; \min_{1\le i\le p-1}\left\{\frac{\min(\delta_{J_{i+1}})}{\max(\delta_{J_i})}\right\} \;=\; \frac{1}{\max_{1\le i\le p-1}\left(\|\delta_{J_i}\|_\infty \|\delta^{-1}_{J_{i+1}}\|_\infty\right)} \;\ge\; 1,$$

with the convention that $\mathrm{gap}(J) = \infty$ if $p = 1$.

The LLS step used by our algorithm is computed with respect to a specific partition which is ordered at the current iterate $w \in \mathcal{P}^{++} \times \mathcal{D}^{++}$. We now describe the construction of this ordered partition. First, with the aid of the AS direction at w, we compute the bipartition (B, N) of $\{1, \ldots, n\}$ according to

$$B \equiv \{i : |Rs^{\rm a}_i| \le |Rx^{\rm a}_i|\,\}, \qquad N \equiv \{i : |Rs^{\rm a}_i| > |Rx^{\rm a}_i|\,\}. \tag{21}$$

Note that this definition and (18) imply that

$$\varepsilon^{\rm a}_\infty = \max\left\{\|Rx^{\rm a}_N\|_\infty,\, \|Rs^{\rm a}_B\|_\infty\right\}. \tag{22}$$

Next, an order $(i_1, \ldots, i_n)$ of the index variables is chosen such that $\delta_{i_1} \le \cdots \le \delta_{i_n}$. Then the first block of consecutive indices in the n-tuple $(i_1, \ldots, i_n)$ lying in the same set B or N are placed in the first layer $\mathcal{J}_1$, the next block of consecutive indices lying in the other set is placed in $\mathcal{J}_2$, and so on. As an example, assume that $(i_1, i_2, i_3, i_4, i_5, i_6, i_7) \in B\times B\times N\times B\times B\times N\times N$. In this case, we have $\mathcal{J}_1 = \{i_1, i_2\}$, $\mathcal{J}_2 = \{i_3\}$, $\mathcal{J}_3 = \{i_4, i_5\}$, and $\mathcal{J}_4 = \{i_6, i_7\}$. A partition obtained according to the above construction is clearly ordered at w. We refer to it as an *ordered* (B, N)-*partition* and denote it by $\mathcal{J} = \mathcal{J}(w)$. The LLS step with respect to an ordered (B, N)-partition is sometimes used as a replacement for the primal-dual AS direction in the predictor step of our algorithm.

Note that an ordered (B, N)-partition is not uniquely determined since there can be more than one n-tuple $(i_1, \ldots, i_n)$ satisfying $\delta_{i_1} \le \cdots \le \delta_{i_n}$. This situation happens exactly when there are two or more indices i with the same value for δ_i. If these tying indices do not all belong to the same set B or N, then there will be more than one way to generate an ordered (B, N)-partition $\mathcal{J}$.

We say that the bipartition (B, N) is *regular* if there do not exist $i \in B$ and $j \in N$ such that $\delta_i = \delta_j$. Observe that there exists a unique ordered (B, N)-partition if and only if (B, N) is regular. When (B, N) is not regular, our algorithm avoids the computation of an ordered (B, N)-partition and hence of any LLS direction with respect to such a partition. Thus, there is no ambiguity in our algorithm.

2.5. Algorithm and the main convergence result. In this subsection, we describe our algorithm and state the main result of this paper which guarantees the convergence of the method in a strong sense. More specifically, we establish an iteration-complexity bound for our method which depends only on the constraint matrix A. This bound is exactly the same as the one obtained in Vavasis and Ye [26].

P-C Layered Algorithm.

Let $0 < \beta \leq 1/4$, $\varepsilon_0 > 0$, and $w^0 \in \mathcal{N}(\beta)$ be given. Set $k = 0$.

1. Set $w = w^k$ and compute the AS direction $(\Delta x^{\mathrm{a}}, \Delta y^{\mathrm{a}}, \Delta s^{\mathrm{a}})$ at w;
2. Compute the quantities $\varepsilon_\infty^{\mathrm{a}}$ and α_{a} as in (18) and (12), and the bipartition (B, N) according to (21);
3. If $\varepsilon_\infty^{\mathrm{a}} > \varepsilon_0$ or (B, N) is not regular, then set $w \leftarrow w + \alpha_{\mathrm{a}} \Delta w^{\mathrm{a}}$ and go to step 7;
4. Otherwise, determine the ordered (B, N)-partition $\mathcal{J} = (\mathcal{J}_1, \ldots, \mathcal{J}_r)$ and compute the LLS step $\Delta w^{\mathrm{l}} = (\Delta x^{\mathrm{l}}, \Delta y^{\mathrm{l}}, \Delta s^{\mathrm{l}})$ at w with respect to $\mathcal{J}$;
5. Let $w^{\mathrm{l}} = w + \alpha_{\mathrm{l}} \Delta w^{\mathrm{l}}$, where $\alpha_{\mathrm{l}} \equiv \sup\{\alpha \in [0, 1] : w + \alpha \Delta w^{\mathrm{l}} \in \mathcal{N}(2\beta)\}$;
6. If $\mu(w^{\mathrm{l}}) < (1 - \alpha_{\mathrm{a}})\mu$, then set $w \leftarrow w^{\mathrm{l}}$, else set $w \leftarrow w + \alpha_{\mathrm{a}} \Delta w^{\mathrm{a}}$;
7. If $\mu(w) = 0$, then **stop**; in this case w is an optimal solution;
8. Compute the corrector step Δw^{c} at w and set $w \leftarrow w + \Delta w^{\mathrm{c}}$;
9. Set $w^{k+1} = w$, increment k by 1 and go to step 1.

End

We now make a few comments about the above algorithm. Step 2 followed by step 8 is a standard P-C iteration of the type described in subsection 2.3. This iteration is always performed in those iterations for which $\varepsilon_\infty^{\mathrm{a}} > \varepsilon_0$ or (B, N) is not regular. In the other iterations, the algorithm performs either a standard P-C iteration or a layered-corrector iteration, depending on which of the two iterations gives the lowest reduction of the duality gap. This test is performed in step 6 since the term $(1 - \alpha_{\mathrm{a}})\mu$ is the normalized duality gap obtained when the AS step is taken (see Proposition 2.4(a)).

The following convergence theorem is the main result of the paper.

Theorem 2.7. *The P-C layered algorithm described above finds a primal-dual optimal solution $(x^\infty, s^\infty, y^\infty)$ of problems* (1) *and* (2) *satisfying strict complementarity (i.e., $x^\infty + s^\infty > 0$) in at most $\mathcal{O}(n^{3.5} \log(\bar{\chi}_A + n + \varepsilon_0^{-1}))$ iterations. In particular, if $\varepsilon_0 = \Omega(1/n^\tau)$ for some constant τ, then the iteration-complexity bound reduces to $\mathcal{O}(n^{3.5} \log(\bar{\chi}_A + n))$.*

3. Basic tools. In this section we introduce the basic tools that will be used in the proof of Theorem 2.7. The analysis heavily relies on the notion of crossover events due to Vavasis and Ye [26]. Subsection 3.1 below gives the definition of a crossover event which is slightly different than the one used in [26] and discusses some of its properties. In subsection 3.2, we state an approximation result that provides an estimation of the closeness between the LLS direction with respect to a partition J of $\{1, \ldots, n\}$ and the AS direction. Subsection 3.3 reviews from a different perspective an important result from [26], namely Lemma 17 of [26], that essentially guarantees the occurrence of crossover events. Since this result is stated in terms of the residual of an LLS step, the use of the approximation result of subsection 3.2 between the AS and LLS steps allows us to obtain a similar result stated in terms of the residual of the AS direction.

3.1. Crossover events. In this subsection we discuss the notion of crossover event which plays a fundamental role in our convergence analysis.

Definition. For two indices $i, j \in \{1, \ldots, n\}$ and a constant $\mathcal{C} \geq 1$, a *$\mathcal{C}$-crossover event* for the pair (i, j) is said to occur on the interval $(\nu', \nu]$ if

$$\begin{aligned} &\text{there exists } \nu_0 \in (\nu', \nu] \text{ such that } \frac{s_j(\nu_0)}{s_i(\nu_0)} \leq \mathcal{C} \\ &\text{and } \frac{s_j(\tilde{\nu})}{s_i(\tilde{\nu})} > \mathcal{C} \quad \forall \tilde{\nu} \leq \nu'. \end{aligned} \tag{23}$$

Moreover, the interval $(\nu', \nu]$ is said to contain a $\mathcal{C}$-crossover event if (23) holds for some pair (i, j).

Hence, the notion of a crossover event is independent of any algorithm and is a property of the central path only. Note that in view of (3), condition (23) can be reformulated into an equivalent condition involving only the primal variable. For our purposes, we will use only (23).

We have the following simple but crucial result about crossover events.

PROPOSITION 3.1. *Let $\mathcal{C} > 0$ be a given constant. There can be at most $n(n-1)/2$ disjoint intervals of the form $(\nu', \nu]$ containing $\mathcal{C}$-crossover events.*

The notion of $\mathcal{C}$-crossover events can be used to define the notion of $\mathcal{C}$-crossover events between two iterates of the P-C layered algorithm as follows. We say that a $\mathcal{C}$-crossover event occurs between two iterates w^k and w^l, $k < l$, generated by the P-C layered algorithm if the interval $(\mu(w^l), \mu(w^k)]$ contains a $\mathcal{C}$-crossover event. Note that in view of Proposition 3.1, there can be at most $n(n-1)/2$ intervals of this type. We will show in the remainder of this paper that there exists a constant $\mathcal{C} > 0$ with the following property: for any index k, there exists an index $l > k$ such that $l - k = \mathcal{O}(n^{1.5} \log(\bar{\chi}_A + n + \varepsilon_0^{-1}))$ and a $\mathcal{C}$-crossover event occurs between the iterates w^k and w^l of the P-C layered algorithm. Proposition 3.1 and a simple argument then show that the P-C layered algorithm must terminate within $\mathcal{O}(n^{3.5} \log(\bar{\chi}_A + n + \varepsilon_0^{-1}))$ iterations.

3.2. Relation between the LLS and AS directions. In this subsection, we describe how the LLS step provides a good approximation of the AS direction, a result that will be important in our convergence analysis. Another result along this direction has also been obtained by Vavasis and Ye [28]. However, our result is more general and better suited for the development of this paper.

The approximation result below can be proved using the projection decomposition techniques developed in [22]. However, we give a simpler proof using instead the techniques developed in [15]. The result essentially states that the larger the gap of J is, the closer the AS direction and the LLS direction with respect to J will be to one another.

THEOREM 3.2. *Let $w = (x, y, s) \in \mathcal{P}^{++} \times \mathcal{D}^{++}$ and an ordered partition $J = (J_1, \ldots, J_p)$ at w be given. Define $\delta \equiv \delta(w)$, and let $\Delta w^{\mathrm{a}} = (\Delta x^{\mathrm{a}}, \Delta y^{\mathrm{a}}, \Delta s^{\mathrm{a}})$ and $\Delta w^{\mathrm{ll}} = (\Delta x^{\mathrm{ll}}, \Delta y^{\mathrm{ll}}, \Delta s^{\mathrm{ll}})$ denote the AS direction at w and the LLS direction at w with respect to J, respectively. If the gap of J satisfies $\mathrm{gap}(J) \geq 4\, p\, \bar{\chi}_A$, then*

$$\max\left\{ \left\|\delta(\Delta x^{\mathrm{a}} - \Delta x^{\mathrm{ll}})\right\|_\infty , \left\|\delta^{-1}(\Delta s^{\mathrm{a}} - \Delta s^{\mathrm{ll}})\right\|_\infty \right\} \leq \frac{12\sqrt{n\mu}\ \bar{\chi}_A}{\mathrm{gap}(J)}.$$

In particular, if $(Rx^{\mathrm{ll}}, Rs^{\mathrm{ll}})$ denote the residual for the LLS direction Δw^{ll}, then

$$\max\left\{ \left\|Rx^{\mathrm{a}} - Rx^{\mathrm{ll}}\right\|_\infty , \left\|Rs^{\mathrm{a}} - Rs^{\mathrm{ll}}\right\|_\infty \right\} \leq \frac{12\sqrt{n}\ \bar{\chi}_A}{\mathrm{gap}(J)}.$$

Proof. Using the characterization (13) of $\Delta x^{\rm a}$ and the definition (19) of $\Delta x^{\rm ll}$, we see that the vectors $d^0 = (d_1^0, \ldots, d_p^0) \equiv (\delta_{J_p} \Delta x^{\rm a}_{J_p}, \ldots, \delta_{J_1} \Delta x^{\rm a}_{J_1})$ and $\tilde{d}^0 = (\tilde{d}_1^0, \ldots, \tilde{d}_p^0)$ $\equiv (\delta_{J_p} \Delta x^{\rm ll}_{J_p}, \ldots, \delta_{J_1} \Delta x^{\rm ll}_{J_1})$ satisfy the assumptions of Theorem 6.1 with $g = 0$, F_{p+1-i} $= A_{J_i}$, $h_{p+1-i} = (\delta x)_{J_i} = (x^{1/2} s^{1/2})_{J_i}$, and $z_{p+1-i} = \delta_{J_i}^{-1}$ for all $i = 1, \ldots, p$. Hence, by the conclusion of Theorem 6.1, we conclude that

$$\left\| \delta(\Delta x^{\rm a} - \Delta x^{\rm ll}) \right\|_\infty \leq \frac{12\, \bar{\chi}_F \left\| x^{1/2} s^{1/2} \right\|}{\mathrm{gap}(J)} = \frac{12\, \bar{\chi}_A \sqrt{n\mu}}{\mathrm{gap}(J)}.$$

Now let G be an $(n - m) \times n$ full row rank matrix such that $AG^T = 0$. Clearly, we have $\mathrm{Ker}(A) = \mathrm{Im}(G^T)$, and hence $\bar{\chi}_A = \bar{\chi}_G$ in view of Proposition 2.2(d). Using the characterization (14) of $\Delta s^{\rm a}$ and the definition (20) of $\Delta s^{\rm ll}$, we see that the vectors $d^0 = (d_1^0, \ldots, d_p^0) \equiv (\delta_{J_1}^{-1} \Delta s^{\rm a}_{J_1}, \ldots, \delta_{J_p}^{-1} \Delta s^{\rm a}_{J_p})$ and $\tilde{d}^0 = (\tilde{d}_1^0, \ldots, \tilde{d}_p^0) \equiv$ $(\delta_{J_1}^{-1} \Delta s^{\rm ll}_{J_1}, \ldots, \delta_{J_p}^{-1} \Delta s^{\rm ll}_{J_p})$ satisfy the assumptions of Theorem 6.1 with $g = 0$, $G_i = F_{J_i}$, $h_i = (\delta^{-1} s)_{J_i} = (x^{1/2} s^{1/2})_{J_i}$, and $z_i = \delta_{J_i}$ for all $i = 1, \ldots, p$. Hence, by the conclusion of Theorem 6.1, we conclude that

$$\left\| \delta^{-1}(\Delta s^{\rm a} - \Delta s^{\rm ll}) \right\|_\infty \leq \frac{12\, \bar{\chi}_F \left\| x^{1/2} s^{1/2} \right\|}{\mathrm{gap}(J)} = \frac{12\, \bar{\chi}_A \sqrt{n\mu}}{\mathrm{gap}(J)}.$$

Hence, the first inequality of the theorem follows. The second inequality follows immediately from the first one and the definition of residual of a direction $(\Delta x, \Delta y, \Delta s)$. ∎

In view of the above result, the AS direction can be well approximated by LLS directions with respect to ordered partitions J which have large gaps. The LLS direction with $p = 1$, which is the AS direction, provides the perfect approximation to the AS direction itself. However, this kind of trivial approximation is not useful for us due to the need of keeping the "spread" of some layers J_k under control. For an ordered partition J at w, the spread of the layer J_k, denoted by $\mathrm{spr}(J_k)$, is defined as

$$\mathrm{spr}(J_k) \equiv \frac{\max(\delta_{J_k})}{\min(\delta_{J_k})} \quad \forall k = 1, \ldots, p.$$

We now describe a special ordered partition introduced by Vavasis and Ye [26] which plays a crucial role in our analysis. Given a point $w \in \mathcal{P}^{++} \times \mathcal{D}^{++}$ and a parameter $\bar{g} \geq 1$, this partition, which we refer to as the *VY $\bar{g}$-partition* at w, is defined as follows. Let $(i_1, \ldots, i_n)$ be an ordering of $\{1, \ldots, n\}$ such that $\delta_{i_1} \leq \cdots \leq \delta_{i_n}$, where $\delta = \delta(w)$. For $k = 2, \ldots, n$, let $r_k \equiv \delta_{i_k} / \delta_{i_{k-1}}$ and define $r_1 \equiv \infty$. Let $k_1 < \cdots < k_p$ be all the indices k such that $r_k > \bar{g}$ for all $j = 1, \ldots, p$. The VY $\bar{g}$-partition J is then defined as $J = (J_1, \ldots, J_p)$, where $J_q \equiv \{i_{k_q}, i_{k_q+1}, \ldots, i_{k_{q+1}-1}\}$ for all $q = 1, \ldots, p$. More generally, given a subset $I \subset \{1, \ldots, n\}$, we can similarly define the *VY $\bar{g}$-partition* of I at w by taking an ordering $(i_1, \ldots, i_m)$ of I satisfying $\delta_{i_1} \leq \cdots \leq \delta_{i_m}$, where $m = |I|$, defining the ratios $r_1, \ldots, r_m$ as above, and proceeding exactly as in the construction above to obtain the partition $J = (J_1, \ldots, J_p)$ of I.

It is easy to see that the following result holds for the partition J described in the previous paragraph.

PROPOSITION 3.3. *Given a subset $I \subseteq \{1, \ldots, n\}$, a point $w \in \mathcal{P}^{++} \times \mathcal{D}^{++}$, and a constant $\bar{g} \geq 1$, the VY $\bar{g}$-partition $J = (J_1, \ldots, J_p)$ of I at w satisfies* $\mathrm{gap}(J) > \bar{g}$ *and* $\mathrm{spr}(J_q) \leq \bar{g}^{|J_q|} \leq \bar{g}^n$ *for all $q = 1, \ldots, p$.*

3.3. Relation between crossover events and search directions. Using Lemma 17 of [26], we derive in this section an upper bound on the number of iterations needed to guarantee the occurrence of a crossover event which depends on the size of the residual of the LLS step and the stepsize at the initial iterate. Under suitable conditions, we derive with the aid of Theorem 3.2 another upper bound on the number of iterations needed to guarantee the occurrence of a crossover event which depends only on the size of the residual of the AS direction at the initial iterate.

Even though Lemma 17 of Vavasis and Ye [26] is stated and proved in a very advanced stage of their paper, one does not need to go through the whole material preceding it. In order to fully understand this result, it is recommended that one read only the material of section 4 of [26], followed by Lemma 16 and finally Lemma 17.

LEMMA 3.4. *Let $w = (x, y, s) \in \mathcal{N}(\beta)$ for some $\beta \in (0, 1)$ and an ordered partition $J = (J_1, \ldots, J_p)$ at w be given. Let $\delta \equiv \delta(w)$, $\mu = \mu(w)$, and $(Rx^{\rm ll}, Rs^{\rm ll})$ denote the residual of the LLS direction $(\Delta x^{\rm ll}, \Delta y^{\rm ll}, \Delta s^{\rm ll})$ at w with respect to J. Then the following statements hold for every $q = 1, \ldots, p$:*

(a) *There exists $i \in J_1 \cup \cdots \cup J_q$ such that*

$$s_i(\nu) \geq \frac{\sqrt{\mu}\ \|Rs^{\rm ll}_{J_q}\|_\infty \min(\delta_{J_q})}{n^{1.5}\bar{\chi}_A} \quad \forall \nu \in (0, \mu].$$

(b) *There exists $j \in J_q \cup \cdots \cup J_p$ such that*

$$x_j(\nu) \geq \frac{\sqrt{\mu}\ \|Rx^{\rm ll}_{J_q}\|_\infty}{n^{1.5}\bar{\chi}_A \max(\delta_{J_q})} \quad \forall \nu \in (0, \mu].$$

(c) *For any $\mathcal{C}_q \geq (1+\beta)\,\mathrm{spr}(J_q)/(1-\beta)^2$ and for any $\mu' \in (0, \mu)$ such that*

$$\frac{\mu'}{\mu} \leq \frac{\|Rx^{\rm ll}_{J_q}\|_\infty \|Rs^{\rm ll}_{J_q}\|_\infty}{n^3 \mathcal{C}_q^2 \bar{\chi}_A^2},$$

the interval $(\mu', \mu]$ contains a $\mathcal{C}_q$-crossover event.

Proof. Noting that our definition of δ is the one used in [26] divided by $\sqrt{\mu}$, we easily see that statements (a) and (b) follow directly from Lemma 17 of [26]. We now prove (c). Let i and j be as in statements (a) and (b). First note that by Proposition 2.1 we have

$$\frac{s_i(\mu)}{s_j(\mu)} \leq \frac{1+\beta}{(1-\beta)^2}\frac{\delta_i}{\delta_j} \leq \frac{1+\beta}{(1-\beta)^2}\frac{\max(\delta_{J_q})}{\min(\delta_{J_q})} = \frac{1+\beta}{(1-\beta)^2}\mathrm{spr}(J_q) \leq \mathcal{C}_q.$$

Now, by (b) and (3), we have

$$\frac{1}{s_j(\nu)} \geq \frac{\sqrt{\mu}\ \|Rx^{\rm ll}_{J_q}\|_\infty}{\nu n^{1.5}\bar{\chi}_A \max(\delta_{J_q})} \quad \forall \nu \in (0, \mu].$$

Using the last relation, the relation in (a), the fact that J is an ordered partition for w, and the conditions on $\mathcal{C}_q$ and μ', we obtain for every $\nu \in (0, \mu']$ that

$$\frac{s_i(\nu)}{s_j(\nu)} \ \geq\ \frac{\mu\,\|Rx^{\rm ll}_{J_q}\|_\infty \|Rs^{\rm ll}_{J_q}\|_\infty}{\nu\, n^3 \bar{\chi}_A^2\, \mathrm{spr}(J_q)} \ >\ \frac{\mu\,\|Rx^{\rm ll}_{J_q}\|_\infty \|Rs^{\rm ll}_{J_q}\|_\infty}{\mu'\, n^3 \bar{\chi}_A^2\, \mathcal{C}_q} \ \geq\ \mathcal{C}_q.$$

We have thus shown that a crossover event for the pair (i, j) occurs in the interval $(\nu', \nu]$. ∎

An immediate consequence of Lemma 3.4(c) which has implications in the analysis of the P-C layered algorithm is as follows.

LEMMA 3.5. *Let $w = (x, y, s) \in \mathcal{N}(\beta)$ for some $\beta \in (0, 1/4]$ and an ordered partition $J = (J_1, \ldots, J_p)$ at w be given. Define $\delta \equiv \delta(w)$ and $\mu = \mu(w)$, and let $(Rx^{\mathrm{ll}}, Rs^{\mathrm{ll}})$ denote the residual of the LLS direction $(\Delta x^{\mathrm{ll}}, \Delta y^{\mathrm{ll}}, \Delta s^{\mathrm{ll}})$ at w with respect to J. Then, for every $q \in \{1, \ldots, p\}$ and every $\mathcal{C}_q \geq (1+\beta)\mathrm{spr}(J_q)/(1-\beta)^2$, the following statements hold:*

(a) *The P-C layered algorithm started from the point w will either generate an iterate $\hat{w}$ with a $\mathcal{C}_q$-crossover event occurring between w and $\hat{w}$ or terminate in*

$$\mathcal{O}\left(\sqrt{n}\left(\log(\bar{\chi}_A + n) + \log \mathcal{C}_q + \log\left(\frac{\mu_+/\mu}{\|Rx^{\mathrm{ll}}_{J_q}\|_\infty \|Rs^{\mathrm{ll}}_{J_q}\|_\infty}\right)\right)\right) \tag{24}$$

iterations, where μ_+ is the normalized duality gap attained immediately after the first iteration.

(b) *If, in addition,*

$$\mathrm{gap}(J) \geq \max\left\{4\, n\bar{\chi}_A\,,\ \frac{24\sqrt{n}\bar{\chi}_A}{\varepsilon^{\mathrm{a}}_{J_q}}\right\}, \tag{25}$$

where $\varepsilon^{\mathrm{a}}_{J_q} \equiv \min\{\|Rx^{\mathrm{a}}_{J_q}\|_\infty\,,\ \|Rs^{\mathrm{a}}_{J_q}\|_\infty\}$, then (24) is bounded above by

$$\mathcal{O}\left(\sqrt{n}\left(\log(\bar{\chi}_A + n) + \log \mathcal{C}_q + \log(\varepsilon^{\mathrm{a}}_{J_q})^{-1}\right)\right). \tag{26}$$

Proof. To prove (a), it is sufficient to show that a $\mathcal{C}_q$-crossover event will occur if the algorithm does not terminate in a number of iterations bounded above by (24). Lemma 3.4(c) guarantees that a $\mathcal{C}_q$-crossover event occurs between w and another iterate $\hat{w}$ whenever

$$\frac{\mu(\hat{w})}{\mu(w)} \leq \frac{\|Rx^{\mathrm{ll}}_{J_q}\|_\infty \|Rs^{\mathrm{ll}}_{J_q}\|_\infty}{n^3 \mathcal{C}_q^2 \bar{\chi}_A^2}. \tag{27}$$

Observe that the duality gap is reduced by a factor of μ_+/μ in the first iteration and by a factor of at least $1 - \sqrt{\beta/n}$ in subsequent iterations due to Proposition 2.4(b). Thus, an iterate $\hat{w}$ satisfying (27) will be generated in at most $N_0 + 1$ iterations, where N_0 is the smallest integer satisfying

$$\log\left(\frac{\mu_+}{\mu}\right) + N_0 \log\left(1 - \sqrt{\frac{\beta}{n}}\right) \leq \log\left[\frac{\|Rx^{\mathrm{ll}}_{J_q}\|_\infty \|Rs^{\mathrm{ll}}_{J_q}\|_\infty}{n^3 \mathcal{C}_q^2 \bar{\chi}_A^2}\right].$$

The first part of the lemma now immediately follows by rearranging this inequality and using the fact that $\log(1+x) < x$ for any $x > -1$.

We now prove (b). We will show that (24) is bounded above by (26) when (25) holds. By Theorem 3.2 and (25), it follows that

$$\max\left\{\ \left\|Rx^{\mathrm{a}} - Rx^{\mathrm{ll}}\right\|_\infty\,,\ \left\|Rs^{\mathrm{a}} - Rs^{\mathrm{ll}}\right\|_\infty\ \right\} \ \leq\ \frac{12\sqrt{n}\ \bar{\chi}_A}{\mathrm{gap}(J)}\ \leq \frac{\varepsilon^{\mathrm{a}}_{J_q}}{2}.$$

Hence, we have

$$\begin{aligned}
&\min\left\{\|Rx^{\rm ll}_{J_q}\|_\infty, \|Rs^{\rm ll}_{J_q}\|_\infty\right\} \\
&\quad \geq \min\left\{ \left\|Rx^{\rm a}_{J_q}\right\|_\infty - \left\|Rx^{\rm a} - Rx^{\rm ll}\right\|_\infty , \left\|Rs^{\rm a}_{J_q}\right\|_\infty - \left\|Rs^{\rm a} - Rs^{\rm ll}\right\|_\infty \right\} \\
&\quad \geq \min\left\{\left\|Rx^{\rm a}_{J_q}\right\|_\infty , \left\|Rs^{\rm a}_{J_q}\right\|_\infty\right\} - \frac{\varepsilon^{\rm a}_{J_q}}{2} \;=\; \varepsilon^{\rm a}_{J_q} - \frac{\varepsilon^{\rm a}_{J_q}}{2} \;=\; \frac{\varepsilon^{\rm a}_{J_q}}{2}.
\end{aligned}$$

Using this estimate in (24) together with the fact that $\mu_+/\mu \leq 1$, we conclude that (24) is bounded above by (26). ∎

4. Convergence analysis of the P-C layered algorithm. In this section, we give the proof of Theorem 2.7.

Lemma 3.5 gives a good idea of the effort which will be undertaken in this section, namely, to show that for each $w \in \mathcal{N}(\beta)$ there exist an ordered partition $J = (J_1, \ldots, J_p)$ and an index $q = 1, \ldots, p$ such that the sum of two last logarithms in (24) can be bounded above by $\mathcal{O}(n \log(\bar{\chi}_A + n + {\varepsilon_0}^{-1}))$. The analysis of this claim will be broken into two cases, namely (i) $\varepsilon^{\rm a}_\infty \geq \varepsilon_0$ and (ii) $\varepsilon^{\rm a}_\infty \leq \varepsilon_0$, where $\varepsilon^{\rm a}_\infty$ is given by (18). The first result below considers the case $\varepsilon^{\rm a}_\infty \geq \varepsilon_0$ for which the VY $\bar{g}$-partition at w is quite suitable. We introduce the following global constants which will be used in the remainder of this paper:

$$\mathcal{C} \equiv \frac{(1+\beta)}{(1-\beta)^2}\,\bar{g}^n, \qquad \bar{g} \equiv 24\,\bar{\chi}_A\,\sqrt{n}\,\max\left\{\varepsilon_0^{-1}\,,\,\frac{4(1+2\beta)\sqrt{n}}{\beta - 2\beta^2}\right\}. \tag{28}$$

LEMMA 4.1. *Suppose that $w = (x, y, s) \in \mathcal{N}(\beta)$ for some $\beta \in (0, 1/4]$ and that $\varepsilon^{\rm a}_\infty \geq \varepsilon_0$ for some constant $\varepsilon_0 > 0$. Then the P-C layered algorithm started from the point w will either generate an iterate $\hat{w}$ with a $\mathcal{C}$-crossover event occurring between w and $\hat{w}$ or terminate in $\mathcal{O}(n^{1.5}\log(\bar{\chi}_A + n + \varepsilon_0^{-1}))$ iterations.*

Proof. The assumption that $\varepsilon^{\rm a}_\infty \geq \varepsilon_0$ and definition (18) imply the existence of an index $i = 1, \ldots, n$ such that $\min\{|Rx^{\rm a}_i|, |Rs^{\rm a}_i|\} \geq \varepsilon_0$. Now let $J = (J_1, \ldots, J_p)$ be a VY $\bar{g}$-partition at w, and let J_q be the layer containing the index i above. Clearly, we have

$$\varepsilon^{\rm a}_{J_q} \equiv \min\{\|Rx^{\rm a}_{J_q}\|_\infty, \|Rs^{\rm a}_{J_q}\|_\infty\} \geq \varepsilon_0. \tag{29}$$

Using the above inequality, the fact that $\mathrm{gap}(J) \geq \bar{g}$, and (28), we easily see that (25) holds. Since by Proposition 3.3 the spread of every layer of a VY $\bar{g}$-partition at w is bounded above by $\bar{g}^n$, we conclude that $\mathrm{spr}(J_q) \leq \bar{g}^n$. Hence, we may set $\mathcal{C}_q = \mathcal{C} \equiv (1+\beta)\bar{g}^n/(1-\beta)^2$ in Lemma 3.5, from which it follows that

$$\log(\mathcal{C}_q) = \mathcal{O}(n \log \bar{g}) = \mathcal{O}(n \log(\bar{\chi}_A + n + \varepsilon_0^{-1})), \tag{30}$$

where the last equality is due to (28). The result now follows from Lemma 3.5(b) by noting that (26) is $\mathcal{O}(n^{1.5}\log(\bar{\chi}_A + n + \varepsilon_0^{-1}))$ in view of (29) and (30). ∎

We now consider the case in which $\varepsilon^{\rm a}_\infty \leq \varepsilon_0$ and show that a $\mathcal{C}$-crossover also happens within $\mathcal{O}(n^{1.5}\log(\bar{\chi}_A + n + \varepsilon_0^{-1}))$ iterations of the P-C layered algorithm (if it does not terminate). From now on, we let $\mathcal{J} = (\mathcal{J}_1, \ldots, \mathcal{J}_r)$ denote an ordered (B, N)-partition at w. We will split the analysis of this case into two subcases, namely (i) $\mathrm{gap}(\mathcal{J}) \leq \bar{g}$ and (ii) $\mathrm{gap}(\mathcal{J}) \geq \bar{g}$. The next result takes care of the case in which $\mathrm{gap}(\mathcal{J}) \leq \bar{g}$, without assuming anything about $\varepsilon^{\rm a}_\infty$.

LEMMA 4.2. *Suppose that $w = (x, y, s) \in \mathcal{N}(\beta)$ for some $\beta \in (0, 1/4]$. Let $\bar{g}$ and $\mathcal{C}$ be the constants defined in* (28). *Let $\mathcal{J} = (\mathcal{J}_1, \ldots, \mathcal{J}_r)$ be an ordered (B, N)-partition*

at w, where (B,N) is the bipartition defined in (21)*, and assume that* $\text{gap}(\mathcal{J}) < \bar{g}$*. Then the P-C layered algorithm started from the point w will either generate an iterate $\hat{w}$ with a $\mathcal{C}$-crossover event occurring between w and $\hat{w}$ or terminate in $\mathcal{O}\left(n^{1.5}\log(\bar{\chi}_A + n + \varepsilon_0^{-1})\right)$ iterations.*

Proof. Assume that $\text{gap}(\mathcal{J}) < \bar{g}$. Let $J = (J_1, \ldots, J_p)$ be a VY $\bar{g}$-partition at w. Using the assumption that $\text{gap}(\mathcal{J}) < \bar{g}$, it is easy to see that there exist two indices i, j of different types, say $i \in B$ and $j \in N$, both lying in some layer J_q of J. By Lemma 2.6 and the definition of (B,N) given in (21), it follows that $|Rx_i^{\text{a}}| \geq 1/4$ and $|Rs_j^{\text{a}}| \geq 1/4$, and hence that

$$\varepsilon_{J_q}^{\text{a}} \equiv \min\{\|Rx_{J_q}^{\text{a}}\|_\infty\,,\, \|Rs_{J_q}^{\text{a}}\|_\infty\} \geq \frac{1}{4}. \tag{31}$$

Using this inequality and the fact that $\text{gap}(J) \geq \bar{g} \geq 96\bar{\chi}_A n$, where the last inequality is due to (28), we easily see that (25) holds. Since by Proposition 3.3 the spread of every layer of a VY $\bar{g}$-partition at w is bounded above by $\bar{g}^n$, we conclude that $\text{spr}(J_q) \leq \bar{g}^n$. Hence, we may set $\mathcal{C}_q = \mathcal{C} \equiv (1+\beta)\bar{g}^n/(1-\beta)^2$ in Lemma 3.5, from which it follows that (30) holds. The result now follows from Lemma 3.5(b) by noting that (26) is $\mathcal{O}(n^{1.5}\log(\bar{\chi}_A + n + \varepsilon_0^{-1}))$ in view of (30) and (31). ∎

The next result considers the case in which $\text{gap}(\mathcal{J}) \geq \bar{g}$ and derives an upper bound on the number of iterations for a $\mathcal{C}$-crossover event to occur. As in Lemma 3.5, nothing is assumed about $\varepsilon_\infty^{\text{a}}$.

LEMMA 4.3. *Suppose that $w = (x,y,s) \in \mathcal{N}(\beta)$ for some $\beta \in (0, 1/4]$. Let $\bar{g}$ and $\mathcal{C}$ be the constants defined in* (28)*. Let $\mathcal{J} = (\mathcal{J}_1, \ldots, \mathcal{J}_r)$ be the (B,N)-partition at w, where (B,N) is the bipartition defined in* (21)*, and assume that* $\text{gap}(\mathcal{J}) \geq \bar{g}$*. Let $(Rx^{\text{l}}, Rs^{\text{l}})$ denote the residual of the LLS direction at w with respect to $\mathcal{J}$. Then the P-C layered algorithm started from the point w will either generate an iterate $\hat{w}$ with a $\mathcal{C}$-crossover event occurring between w and $\hat{w}$ or terminate in*

$$\mathcal{O}\left(n^{1.5}\log\left(\bar{\chi}_A + n + {\varepsilon_0}^{-1}\right) + \sqrt{n}\,\log\left(\frac{\mu_+/\mu}{\varepsilon_\infty^{\text{l}}}\right)\right) \tag{32}$$

iterations, where μ_+ is the normalized duality gap attained immediately after the first iteration, and

$$\varepsilon_\infty^{\text{l}} \equiv \max\left\{\left\|Rx_N^{\text{l}}\right\|_\infty, \left\|Rs_B^{\text{l}}\right\|_\infty\right\}.$$

Proof. Assume without loss of generality that $\varepsilon_\infty^{\text{l}} = \|Rx_N^{\text{l}}\|_\infty$; the case in which $\varepsilon_\infty^{\text{l}} = \|Rs_B^{\text{l}}\|_\infty$ can be proved similarly. Then $\varepsilon_\infty^{\text{l}} = |Rx_i^{\text{l}}|$ for some $i \in N$. Let $\mathcal{J}_t$ be the layer of $\mathcal{J}$ containing the index i and note that

$$\varepsilon_\infty^{\text{l}} \;=\; |Rx_i^{\text{l}}| \;=\; \|Rx_{\mathcal{J}_t}^{\text{l}}\|_\infty \;\leq\; \|Rx_{\mathcal{J}_t}^{\text{l}}\|. \tag{33}$$

Now let $\mathcal{I} = (\mathcal{I}_1, \ldots, \mathcal{I}_p)$ be the VY $\bar{g}$-partition of $\mathcal{J}_t$ at w and consider the ordered partition $\mathcal{J}'$ defined as

$$\mathcal{J}' \equiv (\mathcal{J}_1, \ldots, \mathcal{J}_{t-1}, \mathcal{I}_1, \ldots, \mathcal{I}_p, \mathcal{J}_{t+1}, \ldots, \mathcal{J}_r).$$

Let $(Rx^{\text{ll}}, Rs^{\text{ll}})$ denote the residual of the LLS direction at w with respect to $\mathcal{J}'$. Using the definition of the LLS step, it is easy to see that $Rx_{\mathcal{J}_j}^{\text{l}} = Rx_{\mathcal{J}_j}^{\text{ll}}$ for all $j = t+1, \ldots, r$. Moreover, we have $\|Rx_{\mathcal{J}_t}^{\text{l}}\| \leq \|Rx_{\mathcal{J}_t}^{\text{ll}}\|$ since $\|Rx_{\mathcal{J}_t}^{\text{l}}\|$ is the optimal

value of the least squares problem which determines the $\Delta x^{\rm l}_{\mathcal{J}_t}$-component of the LLS step with respect to $\mathcal{J}$, whereas $\|Rx^{\rm ll}_{\mathcal{J}_t}\|$ is the objective value at a certain feasible solution for the same least squares problem. Hence, for some $q \in \{1,\ldots,p\}$ we have

$$\|Rx^{\rm ll}_{\mathcal{I}_q}\|_\infty = \|Rx^{\rm ll}_{\mathcal{J}_t}\|_\infty \geq \frac{1}{\sqrt{|\mathcal{J}_t|}}\,\|Rx^{\rm ll}_{\mathcal{J}_t}\| \geq \frac{1}{\sqrt{n}}\,\|Rx^{\rm ll}_{\mathcal{J}_t}\| \geq \frac{1}{\sqrt{n}}\,\|Rx^{\rm l}_{\mathcal{J}_t}\|. \tag{34}$$

Combining (33) and (34), we then obtain

$$\|Rx^{\rm ll}_{\mathcal{I}_q}\|_\infty \geq \frac{1}{\sqrt{n}}\,\varepsilon^{\rm l}_\infty. \tag{35}$$

Let us now bound the quantity $\|Rs^{\rm ll}_{\mathcal{I}_q}\|_\infty$ from below. Using triangle inequality for norms, Lemma 2.6, Theorem 3.2, and the fact that $\mathrm{gap}(\mathcal{J}') \geq \bar{g} \geq 96\bar{\chi}_A n$, where the second inequality is due to (28), we obtain

$$\|Rs^{\rm ll}_{\mathcal{I}_q}\|_\infty \;\geq\; \|Rs^{\rm a}_{\mathcal{I}_q}\|_\infty - \|Rs^{\rm ll}_{\mathcal{I}_q} - Rs^{\rm a}_{\mathcal{I}_q}\|_\infty \;\geq\; \frac{1}{4} - \frac{12\sqrt{n}\,\bar{\chi}_A}{\mathrm{gap}(\mathcal{J}')} \;\geq\; \frac{1}{4} - \frac{1}{8} \;\geq\; \frac{1}{8}. \tag{36}$$

Also note that by (28) and Proposition 3.3 we have

$$\mathcal{C} = \frac{1+\beta}{(1-\beta)^2}\,\bar{g}^n \geq \frac{1+\beta}{(1-\beta)^2}\,\mathrm{spr}(\mathcal{I}_q) \tag{37}$$

and

$$\log \mathcal{C} = \mathcal{O}\left(n \log\left(\bar{\chi}_A + n + \varepsilon_0^{-1}\right)\right). \tag{38}$$

Hence, from Lemma 3.5(a) with $J = \mathcal{J}'$ and $\mathcal{C}_q = \mathcal{C}$ and the estimates (35)–(38), it follows that the P-C layered algorithm started from w will find an iterate $\hat{w}$ with a $\mathcal{C}$-crossover event occurring between w and $\hat{w}$ in

$$\begin{aligned}
&\mathcal{O}\left(n^{1.5}\log\left(\bar{\chi}_A + n + \varepsilon_0^{-1}\right) + \sqrt{n}\,\log\left(\frac{\mu_+/\mu}{\|Rx^{\rm l}_{\mathcal{I}_q}\|_\infty \|Rs^{\rm l}_{\mathcal{I}_q}\|_\infty}\right)\right) \\
&= \mathcal{O}\left(n^{1.5}\log\left(\bar{\chi}_A + n + \varepsilon_0^{-1}\right) + \sqrt{n}\,\log\left(\frac{\mu_+/\mu}{\varepsilon^{\rm l}_\infty}\right)\right)
\end{aligned}$$

iterations. □

Our goal now will be to estimate the second logarithm that appears in the iteration-complexity bound (32). It is exactly in this estimation process that we will need to assume that $\varepsilon^{\rm a}_\infty \leq \varepsilon_0$. Under this condition, we know that the duality gap reduction μ^+/μ obtained in the first iteration from w is the smaller between the two duality gap reductions obtained by taking an AS step and an LLS step. Hence, μ^+/μ is majorized by the duality gap reduction obtained from an LLS step from w. Lemma 4.6 below provides an estimation of the duality gap reduction obtained from an LLS step. The two lemmas that precede it, namely Lemmas 4.4 and 4.5, are just technical results which are used in its proof.

LEMMA 4.4. *Let $w = (x,y,s) \in \mathcal{P}^{++} \times \mathcal{D}^{++}$ be given and assume that $\|xs - \nu e\| \leq \tau\nu$ for some constants $\tau \in (0,1)$ and $\nu > 0$. Then $(1-\tau/\sqrt{n})\nu \leq \mu(w) \leq (1+\tau/\sqrt{n})\nu$ and $w \in \mathcal{N}(\tau/(1-\tau))$.*

Proof. We have

$$|\mu(w)-\nu| = \left|\frac{x^Ts-n\nu}{n}\right| = \left|\frac{e^T(xs-\nu e)}{n}\right| \le \frac{\|e\|\,\|xs-\nu e\|}{n} \le \frac{\tau}{\sqrt{n}}\nu,$$

from which the two inequalities of the lemma follow. Since $\tilde{\nu}=\mu(w)$ is the constant which minimizes $\|xs-\tilde{\nu}e\|$, we have

$$\|xs-\mu(w)e\| \le \|xs-\nu e\| \le \tau\nu \le \frac{\tau}{1-\tau/\sqrt{n}}\,\mu(w) \le \frac{\tau}{1-\tau}\,\mu(w),$$

showing that $w \in \mathcal{N}(\tau/(1-\tau))$. ∎

The following lemma is well known (see [4] or [8], for example).

LEMMA 4.5. *Let $\{w^k\}=\{(x^k,y^k,s^k)\}$ be a sequence of points in $\mathcal{P}^{++}\times\mathcal{D}^{++}$ such that $\lim_{k\to\infty}\mu_k=0$ and, for some $\gamma>0$, $x^ks^k\ge\gamma\mu_k e$ for all k, where $\mu_k\equiv\mu(w^k)$. Then every accumulation point $w^\infty=(x^\infty,y^\infty,s^\infty)$ of the sequence $\{w^k\}$ is a primal-dual optimal solution of (1) and (2) satisfying the strict complementarity condition, namely $(x^\infty)^Ts^\infty=0$ and $x^\infty+s^\infty>0$.*

The following lemma gives an estimate of the duality gap reduction obtained by taking an LLS step.

LEMMA 4.6. *Suppose that $w\in\mathcal{N}(\beta)$ for some $\beta\in(0,1/2)$. Let $J=(J_1,\ldots,J_p)$ be an ordered partition at w, and let $\Delta w^{\rm ll}=(\Delta x^{\rm ll},\Delta y^{\rm ll},\Delta s^{\rm ll})$ denote the LLS direction at w with respect to J. Define*

$$\varepsilon^{\rm ll}_\infty \equiv \max\left\{\left\|Rx^{\rm ll}_N\right\|_\infty, \left\|Rs^{\rm ll}_B\right\|_\infty\right\}, \tag{39}$$

$$\alpha_{\rm ll} \equiv \sup\{\alpha\in[0,1] : w+\alpha\Delta w^{\rm ll}\in\mathcal{N}(2\beta)\},$$

where $(Rx^{\rm ll},Rs^{\rm ll})$ is the residual of $\Delta w^{\rm ll}$. Then the following statements hold:

(a) *If $\mathrm{gap}(J)>\max\{4p\bar{\chi}_A\,,\,24\sqrt{n}\bar{\chi}_A\}$, then $x^T\Delta s^{\rm ll}+s^T\Delta x^{\rm ll}<0$, and hence $\mu(w+\alpha\Delta w^{\rm ll})$ is a strictly decreasing affine function of α.*

(b) *If $\mathrm{gap}(J)\ge 96n\bar{\chi}_A/\eta$, where $\eta\equiv(\beta-2\beta^2)/(1+2\beta)$, then*

$$\frac{\mu(w+\alpha_{\rm ll}\Delta w^{\rm ll})}{\mu(w)} \le \frac{4\sqrt{n}\,\varepsilon^{\rm ll}_\infty(\varepsilon^{\rm ll}_\infty+4)}{\eta}.$$

Proof. We first show (a). From the first equation in (11), we easily see that $s^T\Delta x^{\rm a}+x^T\Delta s^{\rm a}=-n\mu$, where $\mu\equiv\mu(w)$. Using this fact, the definition of the residual of a direction, Theorem 3.2, and the assumption that $\mathrm{gap}(J)>\max\{4\,p\,\bar{\chi}_A\,,\,24\sqrt{n}\bar{\chi}_A\}$, we obtain

$$\begin{aligned}
s^T\Delta x^{\rm ll}+x^T\Delta s^{\rm ll} &= s^T\Delta x^{\rm a}+x^T\Delta s^{\rm a}+s^T\left(\Delta x^{\rm ll}-\Delta x^{\rm a}\right)+x^T\left(\Delta s^{\rm ll}-\Delta s^{\rm a}\right)\\
&= -n\mu+\mu\left(\frac{x^{1/2}s^{1/2}}{\sqrt{\mu}}\right)^T\left[\left(Rx^{\rm ll}-Rx^{\rm a}\right)+\left(Rs^{\rm ll}-Rs^{\rm a}\right)\right]\\
&\le -n\mu+\mu\sqrt{n}\left\|\frac{x^{1/2}s^{1/2}}{\sqrt{\mu}}\right\|\left(\left\|Rx^{\rm ll}-Rx^{\rm a}\right\|_\infty+\left\|Rs^{\rm ll}-Rs^{\rm a}\right\|_\infty\right)\\
&\le -n\mu+\mu n\,\frac{24\sqrt{n}\bar{\chi}_A}{\mathrm{gap}(J)} \;=\; -n\mu\left(1-\frac{24\sqrt{n}\bar{\chi}_A}{\mathrm{gap}(J)}\right) \;<\; 0,
\end{aligned}$$

from which (a) follows.

To prove (b), assume that $\mathrm{gap}(J) \geq 96n\bar{\chi}_A/\eta$. Define $v(\alpha) \equiv (x+\alpha\Delta x^{\mathrm{ll}})(s+\alpha\Delta s^{\mathrm{ll}})$ for all $\alpha \in \Re$. We claim that

$$\|v(\alpha)-(1-\alpha)\mu e\| \leq \frac{2\beta}{1+2\beta}(1-\alpha)\mu \ \text{ for every } \ 0 \leq \alpha \leq 1-\frac{2\sqrt{n}\varepsilon_\infty^{\mathrm{ll}}(\varepsilon_\infty^{\mathrm{ll}}+4)}{\eta}. \tag{40}$$

Using this claim, (b) can be proved as follows. By Lemma 4.4 with $w = w+\alpha\Delta w^{\mathrm{ll}}$, $\nu = (1-\alpha)\mu$, and $\tau = 2\beta/(1+2\beta)$ we conclude from the claim that for any $0 \leq \alpha \leq 1-2\sqrt{n}\,\varepsilon_\infty^{\mathrm{ll}}(\varepsilon_\infty^{\mathrm{ll}}+4)/\eta$, we have $w+\alpha\Delta w^{\mathrm{ll}} \in \mathcal{N}(2\beta)$ and

$$\mu(w+\alpha\Delta w^{\mathrm{ll}}) \leq \left(1+\frac{2\beta}{\sqrt{n}(1+2\beta)}\right)(1-\alpha)\mu \leq 2(1-\alpha)\mu. \tag{41}$$

By the definition of α_{ll}, we then conclude that $\alpha_{\mathrm{ll}} \geq \alpha_* \equiv 1-2\sqrt{n}\,\varepsilon_\infty^{\mathrm{ll}}(\varepsilon_\infty^{\mathrm{ll}}+4)/\eta$. Setting $\alpha = \alpha_*$ in (41) and using the fact that $\alpha_{\mathrm{ll}} \geq \alpha_*$ and $\mu(w+\alpha\Delta w^{\mathrm{ll}})$ is a decreasing function of α, we obtain

$$\mu(w+\alpha_{\mathrm{ll}}\Delta w^{\mathrm{ll}}) \leq \mu(w+\alpha_*\Delta w^{\mathrm{ll}}) \leq 2(1-\alpha_*)\mu = \frac{4\sqrt{n}\,\varepsilon_\infty^{\mathrm{ll}}(\varepsilon_\infty^{\mathrm{ll}}+4)}{\eta}\mu;$$

that is, (b) holds. In the remainder of the proof, we show that (40) holds. It is easy to see that

$$\begin{aligned} v(\alpha)-(1-\alpha)\mu e &= (x+\alpha\Delta x^{\mathrm{ll}})(s+\alpha\Delta s^{\mathrm{ll}})-(1-\alpha)\mu e \\ &= (1-\alpha)(xs-\mu e)+\alpha h^1+\alpha(1-\alpha)h^2+\alpha^2 h^3, \end{aligned} \tag{42}$$

where h^1, h^2, and h^3 are vectors in $\Re^n$ defined as

$$\begin{pmatrix} h_B^1 \\ h_N^1 \end{pmatrix} \equiv \begin{pmatrix} x_B(s_B+\Delta s_B^{\mathrm{ll}}) \\ s_N(x_N+\Delta x_N^{\mathrm{ll}}) \end{pmatrix} = \mu \begin{pmatrix} w_B p_B \\ w_N p_N \end{pmatrix}, \tag{43}$$

$$\begin{pmatrix} h_B^2 \\ h_N^2 \end{pmatrix} \equiv \begin{pmatrix} s_B\Delta x_B^{\mathrm{ll}} \\ x_N\Delta s_N^{\mathrm{ll}} \end{pmatrix} = \mu \begin{pmatrix} w_B q_B \\ w_N q_N \end{pmatrix}, \tag{44}$$

$$\begin{pmatrix} h_B^3 \\ h_N^3 \end{pmatrix} \equiv \begin{pmatrix} \Delta x_B^{\mathrm{ll}}(s_B+\Delta s_B^{\mathrm{ll}}) \\ \Delta s_N^{\mathrm{ll}}(x_N+\Delta x_N^{\mathrm{ll}}) \end{pmatrix} = \mu \begin{pmatrix} p_B q_B \\ p_N q_N \end{pmatrix}. \tag{45}$$

Here the vectors p, q, and w appearing in the second alternative expressions for h^1, h^2, and h^3 are defined as

$$\begin{pmatrix} p_B \\ p_N \end{pmatrix} \equiv \begin{pmatrix} Rs_B^{\mathrm{ll}} \\ Rx_N^{\mathrm{ll}} \end{pmatrix}, \qquad \begin{pmatrix} q_B \\ q_N \end{pmatrix} \equiv \begin{pmatrix} \delta_B\Delta x_B^{\mathrm{ll}}/\sqrt{\mu} \\ \delta_N^{-1}\Delta s_N^{\mathrm{ll}}/\sqrt{\mu} \end{pmatrix}, \qquad w \equiv \frac{x^{1/2}s^{1/2}}{\sqrt{\mu}}.$$

Clearly, we have

$$\|p\|_\infty = \varepsilon_\infty^{\mathrm{ll}}, \qquad \|w\|_\infty \leq \sqrt{1+\beta} \leq 2, \qquad \|w\| = \sqrt{n}. \tag{46}$$

We will now derive an upper bound for $\|q\|$. Using the definition of $(Rx^{\mathrm{ll}}, Rs^{\mathrm{ll}})$ and (17), we obtain

$$\frac{\delta_B\Delta x_B^{\mathrm{ll}}}{\sqrt{\mu}} = Rx_B^{\mathrm{ll}}-w_B = -Rs_B^{\mathrm{ll}}+(Rx_B^{\mathrm{ll}}-Rx_B^{\mathrm{a}})+(Rs_B^{\mathrm{ll}}-Rs_B^{\mathrm{a}})$$

and

$$\frac{\delta_N^{-1}\Delta s_N^{\mathrm{ll}}}{\sqrt{\mu}} = Rs_N^{\mathrm{ll}} - w_N = -Rx_N^{\mathrm{ll}} + (Rs_N^{\mathrm{ll}} - Rs_N^{\mathrm{a}}) + (Rx_N^{\mathrm{ll}} - Rx_N^{\mathrm{a}}),$$

from which it follows that

$$q = -p + (Rx^{\mathrm{ll}} - Rx^{\mathrm{a}}) + (Rs^{\mathrm{ll}} - Rs^{\mathrm{a}}).$$

Hence, using the triangle inequality for norms, Theorem 3.2, and the assumption that $\mathrm{gap}(J) \geq 96n\bar{\chi}_A/\eta \geq 4\,p\,\bar{\chi}_A$, we obtain

$$\|q\| \leq \|p\| + \|Rx^{\mathrm{ll}} - Rx^{\mathrm{a}}\| + \|Rs^{\mathrm{ll}} - Rs^{\mathrm{a}}\| \leq \sqrt{n}\,\varepsilon_\infty^{\mathrm{ll}} + \frac{24n\bar{\chi}_A}{\mathrm{gap}(J)} \leq \sqrt{n}\,\varepsilon_\infty^{\mathrm{ll}} + \frac{\eta}{4}. \tag{47}$$

Using (43), (44), (45), (46), and (47), we obtain

$$\begin{aligned}
\|h^1\| \leq \mu\|w\|\,\|p\|_\infty &\leq \mu\sqrt{n}\,\varepsilon_\infty^{\mathrm{ll}},\\
\|h^2\| \leq \mu\|w\|_\infty\,\|q\| &\leq 2\mu\left(\sqrt{n}\,\varepsilon_\infty^{\mathrm{ll}} + \frac{\eta}{4}\right),\\
\|h^3\| \leq \mu\|p\|_\infty\,\|q\| &\leq \mu\varepsilon_\infty^{\mathrm{ll}}\left(\sqrt{n}\,\varepsilon_\infty^{\mathrm{ll}} + \frac{\eta}{4}\right) \leq \mu\sqrt{n}\,\varepsilon_\infty^{\mathrm{ll}}\left(\varepsilon_\infty^{\mathrm{ll}} + 1\right).
\end{aligned}$$

Using (42), the triangle inequality for norms, and the three estimates above, we then obtain

$$\begin{aligned}
\|v(\alpha) - (1-\alpha)\mu e\| &\leq (1-\alpha)\|xs - \mu e\| + \alpha\|h^1\| + \alpha(1-\alpha)\|h^2\| + \alpha^2\|h^3\|\\
&\leq (1-\alpha)\left(\|xs - \mu e\| + \|h^2\|\right) + \|h^1\| + \|h^3\|\\
&\leq \left[(1-\alpha)\left(\beta + 2\sqrt{n}\,\varepsilon_\infty^{\mathrm{ll}} + \frac{\eta}{2}\right) + \sqrt{n}\,\varepsilon_\infty^{\mathrm{ll}} + \sqrt{n}\,\varepsilon_\infty^{\mathrm{ll}}\left(\varepsilon_\infty^{\mathrm{ll}} + 1\right)\right]\mu\\
&\leq \left[\left(\beta + \frac{\eta}{2}\right)(1-\alpha) + \sqrt{n}\,\varepsilon_\infty^{\mathrm{ll}}\left(\varepsilon_\infty^{\mathrm{ll}} + 4\right)\right]\mu\\
&\leq (\beta+\eta)(1-\alpha)\mu \;=\; \frac{2\beta}{1+2\beta}(1-\alpha)\mu
\end{aligned}$$

for all $0 \leq \alpha \leq 1 - 2\sqrt{n}\,\varepsilon_\infty^{\mathrm{ll}}(\varepsilon_\infty^{\mathrm{ll}} + 4)/\eta$. Hence, the validity of the claim follows. ∎

We are now ready to prove the main result of this paper, namely Theorem 2.7.

Proof of Theorem 2.7. Let $\mathcal{C}$ and $\bar{g}$ be the constant defined in (28). We claim that the P-C layered algorithm started from any $w \in \mathcal{N}(\beta)$ either terminates (at step 7) or generates an iterate $\hat{w}$ with a $\mathcal{C}$-crossover event occurring between w and $\hat{w}$ in $\mathcal{O}(n^{1.5}\log(\bar{\chi}_A + n + \varepsilon_0^{-1}))$ iterations. Since by Proposition 3.1 there can be at most $n(n+1)/2$ $\mathcal{C}$-crossover events of the above type, we conclude that the P-C layered algorithm must ultimately terminate in $\mathcal{O}(n^{3.5}\log(\bar{\chi}_A + n + \varepsilon_0^{-1}))$ iterations. To show the above claim, let $\mathcal{J} = (\mathcal{J}_1, \ldots, \mathcal{J}_r)$ denote an ordered (B,N)-partition at w, where (B,N) is the bipartition defined in (21). We split the proof into three possible cases: (1) $\varepsilon_\infty^{\mathrm{a}} > \varepsilon_0$, (2) $\mathrm{gap}(\mathcal{J}) \leq \bar{g}$, and (3) $\varepsilon_\infty^{\mathrm{a}} \leq \varepsilon_0$ and $\mathrm{gap}(\mathcal{J}) > \bar{g}$. The claim clearly holds for the first two cases due to Lemmas 4.1 and 4.2. Moreover, Lemma 4.3 implies that the claim also holds in the third case as long as we can show that the quantity $(\mu_+/\mu)/\varepsilon_\infty^{\mathrm{l}}$ appearing in (32) is $\mathcal{O}(\sqrt{n})$. Indeed, let α_{l} be defined as in step 5 of the P-C layered algorithm. Since in case (3) the LLS step is computed and step 6 of the P-C layered algorithm is performed, we must have $\mu_+ \leq \mu(w + \alpha_{\mathrm{l}}\Delta w^{\mathrm{l}})$. Hence,

the second statement of Lemma 4.6 applied to the partition $\mathcal{J}$ and the fact that $\text{gap}(\mathcal{J}) > \bar{g} \geq 96n\bar{\chi}_A/\eta$, where the second inequality is due to (28), imply

$$\frac{\mu_+}{\mu} \leq \frac{\mu(w + \alpha_1 \Delta w^1)}{\mu} \leq \frac{4\sqrt{n}\,\varepsilon_\infty^1(\varepsilon_\infty^1 + 4)}{\eta}.$$

Hence, we conclude that $(\mu_+/\mu)/\varepsilon_\infty^1 = \mathcal{O}(\sqrt{n})$ whenever $\varepsilon_\infty^1 \leq 1$. If, on the other hand, $\varepsilon_\infty^1 > 1$, then we have $(\mu_+/\mu)/\varepsilon_\infty^1 \leq 1$ since $\mu_+/\mu \leq 1$.

It remains to show that when the method terminates at step 7 of the P-C layered algorithm it always finds a strictly complementary optimal solution. Indeed, let $\hat{w}$ be the iterate satisfying the stopping criterion of step 7. Clearly, $\mu(\hat{w}) = 0$ and $\hat{w} = w + \bar{\alpha}\Delta w$ for some $w \in \mathcal{N}(\beta)$, primal-dual feasible direction Δw, and stepsize $\bar{\alpha} > 0$ satisfying the property that $w + \alpha\Delta w \in \mathcal{N}(\beta)$ for all $\alpha \in [0, \bar{\alpha})$. Using Lemma 4.5, we conclude that $\bar{w}$ is a strictly complementary optimal solution. ∎

5. Concluding remarks. We consider our algorithm from the point of view of scaling-invariance. If one considers the change of variables $x = D\tilde{x}$, where D is a positive diagonal matrix, then the LP problem (1) is equivalent to

$$\min\{(Dc)^T\tilde{x} : AD\tilde{x} = b,\ \tilde{x} \geq 0\}.$$

It turns out that the sequence of points generated by the P-C layered algorithm when applied to (1) does not necessarily correspond (under the transformation $x = D\tilde{x}$) to the one obtained by applying it to the above LP problem. Algorithms with this desirable property are called scaling-invariant. The lack of scaling-invariance of the P-C layered algorithm, as well as the algorithms of Megiddo, Mizuno, and Tsuchiya [10] and Vavasis and Ye [26], is due to the fact that the choice of the layered partition used in the LLS step is not scaling-invariant. The construction of this partition is based on comparing the magnitudes of different components of δ, which per se is not a scaling-invariant quantity.

An interesting open problem is whether there exists a scaling-invariant algorithm whose complexity depends only on m, n, and $\bar{\chi}_A$. Note that if such an algorithm exists, its complexity will in fact depend only on m, n, and the quantity $\inf\{\bar{\chi}_{AD} : D \in \mathcal{D}\}$.

As in [26] and [10], we developed our algorithm for LP problems in which a well-centered interior feasible solution is given in advance. General LP problems can also be solved by the same algorithm applied to a suitably constructed artificial LP problem, and the resulting computational complexity can be shown to be the same as the one obtained in this paper. We refer the reader to section 10 of [26] and section 5 of [10] for more details.

6. Appendix. In this section we give the proof of Theorem 3.2. We start by stating the following result which yields Theorem 3.2 almost as an immediate consequence.

THEOREM 6.1. *Let $g \in \Re^m$, $F_i \in \Re^{m\times n_i}$, $h_i \in \Re^{n_i}$, $z_i \in \Re^{n_i}_{++}$, $i = 1, \ldots, l$, be given and assume that $g \in \text{Im}([F_1, \ldots, F_l])$. Define $d^0 = (d_1^0, \ldots, d_l^0) \in \Re^{n_1} \times \cdots \times \Re^{n_l}$ as*

$$(d_1^0, \ldots, d_l^0) \equiv \operatorname*{argmin}_{(d_1,\ldots,d_l)\in\Re^{n_1}\times\ldots\times\Re^{n_l}} \left\{ \sum_{i=1}^l \|d_i - h_i\|^2 : \sum_{i=1}^l F_i Z_i d_i = g \right\}, \tag{48}$$

and define $\tilde{d}^0 = (\tilde{d}_1^0, \ldots, \tilde{d}_l^0) \in \Re^{n_1} \times \cdots \times \Re^{n_l}$ *recursively starting from* $k = 1$ *upwards as*

$$\tilde{d}_k^0 \equiv \operatorname*{argmin}_{\tilde{d}_k \in \Re^{n_k}} \left\{ \|\tilde{d}_k - h_k\|^2 : F_k Z_k \tilde{d}_k = g - \sum_{i=1}^{k-1} F_i Z_i \tilde{d}_i^0 + \operatorname{Im}(\tilde{F}_{k+1}) \right\}$$

for every $k = 1, \ldots, l-1$, *where* $Z_k \equiv \operatorname{Diag}(z_k)$ *and* $\tilde{F}_k \equiv [F_k, \ldots, F_l] \in \Re^{m \times (n_k + \cdots + n_l)}$. *If the quantity* $\Delta \equiv \max\{\Delta_i : i = 1, \ldots, l-1\}$, *where* $\Delta_i \equiv \|z_i\|_\infty \|z_{i+1}^{-1}\|_\infty$ *for all* $i = 1, \ldots, l-1$, *satisfies* $\bar{\chi}_F \Delta \leq 1/\sqrt{2}$, *then*

$$\|d^0 - \tilde{d}^0\|_\infty \leq 4\bar{\chi}_F \Delta (1 + 4\bar{\chi}_F \Delta)^{l-2} \|d^0 - h\|, \tag{49}$$

where $h \equiv (h_1, \ldots, h_l)$ *and* $F = \tilde{F}_1$. *In particular, if* $g = 0$ *and* $4\bar{\chi}_F \Delta \leq 1/l$, *then*

$$\|d^0 - \tilde{d}^0\|_\infty \leq 12\bar{\chi}_F \Delta \|h\|. \tag{50}$$

The proof of Theorem 6.1 will be given at the end of this section after some preliminary results are derived. Note that when $g = 0$ in Theorem 6.1 the point d^0 is the projection of h onto the null space of the matrix $[F_1 Z_1, \ldots, F_l Z_l] \in \Re^{m \times (n_1 + \cdots + n_l)}$ and the point $\tilde{d}^0$ is the layered projection of h onto the null space of $[F_1 Z_1, \ldots, F_l Z_l]$ according to the partition of variables $(z_1, \ldots, z_l)$.

The proof of Theorem 6.1 will be done by induction on the number l. A crucial step in this induction proof is the validity of certain proximity bounds for the case in which $l = 2$. Hence, as a preliminary step we will derive a special result for the case in which $l = 2$.

PROPOSITION 6.2. *Let* $g \in \Re^m$, $F_i \in \Re^{m \times n_i}$, $h_i \in \Re^{n_i}$, $z_i \in \Re^{n_i}_{++}$, $i = 1, 2$, *be given and assume that* $g \in \operatorname{Im}([F_1, F_2])$. *Consider the points* $d^0 = (d_1^0, d_2^0)$ *and* $\tilde{d}^0 = (\tilde{d}_1^0, \tilde{d}_2^0)$ *determined as*

$$(d_1^0, d_2^0) \equiv \operatorname*{argmin}_d \{ \|d_1 - h_1\|^2 + \|d_2 - h_2\|^2 : F_1 Z_1 d_1 + F_2 Z_2 d_2 = g \}, \tag{51}$$

$$\tilde{d}_1^0 \equiv \operatorname*{argmin}_{d_1} \{ \|d_1 - h_1\|^2 : F_1 Z_1 d_1 \in g + \operatorname{Im}(F_2) \}, \tag{52}$$

$$\tilde{d}_2^0 \equiv \operatorname*{argmin}_{d_2} \{ \|d_2 - h_2\|^2 : F_2 Z_2 d_2 = g - F_1 Z_1 \tilde{d}_1^0 \}, \tag{53}$$

where $Z_1 \equiv \operatorname{Diag}(z_1)$ *and* $Z_2 \equiv \operatorname{Diag}(z_2)$. *Let* $\Delta \equiv \|z_1\|_\infty \|(z_2)^{-1}\|_\infty$ *and assume that* $\bar{\chi}_F \Delta \leq 1/\sqrt{2}$, *where* $F \equiv [F_1, F_2]$. *Then the following estimates of the proximity between* d^0 *and* $\tilde{d}^0$ *hold:*

$$\|d_1^0 - \tilde{d}_1^0\| \leq 4\bar{\chi}_F \Delta \|d_2^0 - h_2\|, \qquad \|d_2^0 - \tilde{d}_2^0\| \leq 4\bar{\chi}_F^2 \Delta^2 \|d_2^0 - h_2\|.$$

Before giving the proof of the above proposition, we first state and prove the following result which characterizes the displacements $\delta_1^0 \equiv d_1^0 - \tilde{d}_1^0$ and $\delta_2^0 \equiv d_2^0 - \tilde{d}_2^0$ as optimal solutions of certain optimization problems.

LEMMA 6.3. *Let* g, F_i, Z_i, $i = 1, 2$, *be as defined in Proposition* 6.2. *Then the following statements hold:*

(a) *The vector* $\delta_2^0 \equiv d_2^0 - \tilde{d}_2^0$ *is the unique optimal solution of the problem*

$$\begin{array}{ll} \textit{minimize}_{\delta_2} & \frac{1}{2}\|\delta_2\|^2 \\ \textit{subject to} & F_2 Z_2 \delta_2 = -F_1 Z_1 \delta_1^0. \end{array} \tag{54}$$

(b) *The pair* (δ_1^0, d_2^0)*, where* $\delta_1^0 \equiv d_1^0 - \tilde{d}_1^0$*, is the unique optimal solution of the problem*

$$\begin{array}{ll} \textit{minimize}_{(\delta_1,d_2)} & \frac{1}{2}\|\delta_1\|^2 + \frac{1}{2}\|d_2 - h_2\|^2 \\ \textit{subject to} & F_1 Z_1 \delta_1 + F_2 Z_2 d_2 = g - F_1 Z_1 \tilde{d}_1^0. \end{array} \tag{55}$$

Proof. We first show (a). Since d^0 and $\tilde{d}_2^0$ are optimal solutions of (51) and (53), respectively, we have

$$\begin{pmatrix} d_1^0 - h_1 \\ d_2^0 - h_2 \end{pmatrix} \in \mathrm{Im} \begin{pmatrix} Z_1 F_1^T \\ Z_2 F_2^T \end{pmatrix}, \qquad F_1 Z_1 d_1^0 + F_2 Z_2 d_2^0 = g, \tag{56}$$

$$\tilde{d}_2^0 - h_2 \in \mathrm{Im}(Z_2 F_2^T), \qquad F_1 Z_1 \tilde{d}_1^0 + F_2 Z_2 \tilde{d}_2^0 = g, \tag{57}$$

and hence

$$d_2^0 - \tilde{d}_2^0 \in \mathrm{Im}(Z_2 F_2^T), \qquad F_2 Z_2 \delta_2^0 = -F_1 Z_1 \delta_1^0. \tag{58}$$

This shows that $\delta_2^0 = d_2^0 - \tilde{d}_2^0$ satisfies the optimality conditions for problem (54). Since (54) is a strictly convex quadratic program, its optimal solution is unique and hence (a) follows. We next show (b). Since $\tilde{d}_1^0$ is the optimal solution of (52), we have

$$\begin{pmatrix} \tilde{d}_1^0 - h_1 \\ 0 \end{pmatrix} \in \mathrm{Im}(ZF^T),$$

which, together with (56) and the definition of δ_1^0, yields

$$\begin{pmatrix} \delta_1^0 \\ d_2^0 - h_2 \end{pmatrix} \in \mathrm{Im}(ZF^T), \qquad F_1 Z_1 \delta_1^0 + F_2 Z_2 d_2^0 = g - F_1 Z_1 \tilde{d}_1^0. \tag{59}$$

This shows that (δ_1^0, d_2^0) satisfies the optimality conditions for (55). Since (55) is a strictly convex quadratic program, its optimal solution is unique and hence (b) holds. ∎

Using the above lemma, we now give a proof of Proposition 6.2.

Proof of Proposition 6.2. By (58), we have that $F_1 Z_1 \delta_1^0 \in \mathrm{Range}(F_2)$. Hence, by Lemma 2.3, there exists a vector v_2^0 such that

$$F_2 v_2^0 = F_1 Z_1 \delta_1^0, \qquad \|v_2^0\| \le \bar{\chi}_F \|Z_1 \delta_1^0\| \le \bar{\chi}_F \|z_1\|_\infty \|\delta_1^0\|. \tag{60}$$

Relation (59) and (60) imply that $F_2[Z_2 d_2^0 + v_2^0] = g - F_1 Z_1 \tilde{d}_1^0$, and hence that the pair $(d_2^0 + Z_2^{-1} v_2^0,\, 0)$ is feasible for (55). This together with Lemma 6.3(b) implies that

$$\|d_2^0 - h_2\|^2 + \|\delta_1^0\|^2 \le \|d_2^0 + Z_2^{-1} v_2^0 - h_2\|^2.$$

Rearranging this expression and using relation (60) and the inequality $\|r\|^2 - \|u\|^2 \le \|r - u\|\,\|r + u\|$ for any $r, u \in \Re^p$, we obtain

$$\begin{aligned} \|\delta_1^0\|^2 &\le \left\{ \|d_2^0 + Z_2^{-1} v_2^0 - h_2\|^2 - \|d_2^0 - h_2\|^2 \right\} \\ &\le \|Z_2^{-1} v_2^0\| \,\|\, 2\,(d_2^0 - h_2) + Z_2^{-1} v_2^0 \,\| \\ &\le \|z_2^{-1}\|_\infty \|v_2^0\| \left\{ 2\,\|d_2^0 - h_2\| + \|z_2^{-1}\|_\infty \|v_2^0\| \right\} \\ &\le \bar{\chi}_F\, \Delta\, \|\delta_1^0\| \left\{ 2\,\|d_2^0 - h_2\| + \bar{\chi}_F\, \Delta\, \|\delta_1^0\| \right\}, \end{aligned}$$

from which it follows that

$$\|\delta_1^0\| \le \frac{2\,\bar{\chi}_F\,\Delta\,\|d_2^0 - h_2\|}{1 - \bar{\chi}_F^2\,\Delta^2} \le 4\bar{\chi}_F\,\Delta\,\|d_2^0 - h_2\|, \tag{61}$$

where the last inequality is due to the assumption that $\bar{\chi}_F\Delta \le 1/\sqrt{2}$. The first relation in (60) implies that $-Z_2^{-1}v_2^0$ is a feasible solution of problem (54). Hence, by Lemma 6.3(a), the second relation in (60), and relation (61), it follows that

$$\|\delta_2^0\| \le \|Z_2^{-1}v_2^0\| \le \|z_2^{-1}\|_\infty\|v_2^0\| \le \bar{\chi}_F\Delta\|\delta_1^0\| \le 4\bar{\chi}_F^2\,\Delta^2\,\|d_2^0 - h_2\|. \qquad \square$$

We are now ready to give the proof of Theorem 6.1.

Proof of Theorem 6.1. During the proof, we refer to $\tilde{d}^0$ as the l-layer point associated with problem (48). We prove the inequality (49) by induction on l. Using Proposition 6.2 and noting that $\bar{\chi}_F\Delta \le 1/\sqrt{2}$ by assumption, we obtain

$$\|d^0 - \tilde{d}^0\|_\infty \le \max\{\|d_1^0 - \tilde{d}_1^0\|\,,\,\|d_2^0 - \tilde{d}_2^0\|\} \le 4\bar{\chi}_F\Delta\|d_2^0 - h_2\| \le 4\bar{\chi}_F\Delta\|d^0 - h\|,$$

from which we conclude that (49) holds for $l = 2$. Assume now that $l \ge 3$ and inequality (49) holds for $l-1$. Consider the solution $(p_2^0, \ldots, p_l^0)$ of the problem

$$(p_2^0, \ldots, p_l^0) \equiv \operatorname*{argmin}_{(p_2,\ldots,p_l)} \left\{ \sum_{i=2}^{l} \|p_i - h_i\|^2 : \sum_{i=2}^{l} F_i Z_i p_i = g - F_1 Z_1 \tilde{d}_1^0 \right\} \tag{62}$$

and note that $\tilde{d}_1^0$ and $(p_2^0, \ldots, p_l^0)$ are the optimal solutions of problems (52) and (53) in which F_1, F_2, z_1, and z_2 in Proposition 6.2 are identified with F_1, $\tilde{F}_2$, z_1, and $(z_2, \ldots, z_l)$, respectively. Hence, it follows from Proposition 6.2 that

$$\|d_1^0 - \tilde{d}_1^0\| \le 4\bar{\chi}_F\Delta\|d^0 - h\|, \tag{63}$$

$$\|(d_2^0 - p_2^0, \ldots, d_l^0 - p_l^0)\| \le 4\bar{\chi}_F^2\Delta^2\|d^0 - h\|. \tag{64}$$

Note also that $(\tilde{d}_2^0, \ldots, \tilde{d}_l^0)$ is the $(l-1)$-layer point associated with the problem (62). Hence, it follows from the induction hypothesis, i.e., that inequality (49) holds for $l-1$, that

$$\|(p_2^0 - \tilde{d}_2^0, \ldots, p_l^0 - \tilde{d}_l^0)\|_\infty \le 4\bar{\chi}_F\Delta(1 + 4\bar{\chi}_F\Delta)^{l-3}\|(p_2^0 - h_2, \ldots, p_l^0 - h_l)\|.$$

Using the triangle inequality for norms and (64), we obtain

$$\begin{aligned}\|(p_2^0 - h_2, \ldots, p_l^0 - h_l)\| &\le \|(d_2^0 - p_2^0, \ldots, d_l^0 - p_l^0)\| + \|(d_2^0 - h_2, \ldots, d_l^0 - h_l)\| \\ &\le 4\bar{\chi}_F^2\Delta^2\|d^0 - h\| + \|d^0 - h\| \;=\; (4\bar{\chi}_F^2\Delta^2 + 1)\|d^0 - h\|.\end{aligned}$$

Combining the two last inequalities yields

$$\|(p_2^0 - \tilde{d}_2^0, \ldots, p_l^0 - \tilde{d}_l^0)\|_\infty \le 4\bar{\chi}_F\Delta(1 + 4\bar{\chi}_F\Delta)^{l-3}(4\bar{\chi}_F^2\Delta^2 + 1)\|d^0 - h\|.$$

Using the triangle inequality for norms again, the last inequality, and (64), we obtain

$$\begin{aligned}\|(d_2^0 - \tilde{d}_2^0, \ldots, d_l^0 - \tilde{d}_l^0)\|_\infty &\le \|(d_2^0 - p_2^0, \ldots, d_l^0 - p_l^0)\|_\infty + \|(p_2^0 - \tilde{d}_2^0, \ldots, p_l^0 - \tilde{d}_l^0)\|_\infty \\ &\le \left[4\bar{\chi}_F^2\Delta^2 + 4\bar{\chi}_F\Delta(1 + 4\bar{\chi}_F\Delta)^{l-3}(4\bar{\chi}_F^2\Delta^2 + 1)\right]\,\|d^0 - h\| \\ &\le 4\bar{\chi}_F\Delta(1 + 4\bar{\chi}_F\Delta)^{l-3}(\bar{\chi}_F\Delta + 4\bar{\chi}_F^2\Delta^2 + 1)\,\|d^0 - h\| \\ &\le 4\bar{\chi}_F\Delta(1 + 4\bar{\chi}_F\Delta)^{l-3}(1 + (1 + 2\sqrt{2})\bar{\chi}_F\Delta)\|d^0 - h\| \\ &\le 4\bar{\chi}_F\Delta(1 + 4\bar{\chi}_F\Delta)^{l-2}\,\|d^0 - h\|.\end{aligned}$$

The last inequality together with (63) implies that inequality (49) holds for l. It then follows by an induction argument that inequality (49) holds for any l.

We now prove that (50) holds when $g = 0$ and $4\bar{\chi}_F\Delta \leq 1/l$. Indeed, when $g = 0$, (48) implies that the vector d^0 is the orthogonal projection of h onto a subspace. Hence, $\|d^0 - h\| \leq \|h\|$. Also, $4\bar{\chi}_F\Delta \leq 1/l$ implies that $(1 + 4\bar{\chi}_F\Delta)^{l-2} \leq (1 + 1/l)^l \leq e \approx 2.718$. Substituting these two bounds into (49), we obtain (50). □

Acknowledgments. We are grateful to two anonymous referees for their valuable comments and suggestions which have helped us to improve the presentation of the paper.

REFERENCES

[1] I. I. Dikin and V. I. Zorkalcev, *Iterative Solution of Mathematical Programming Problems: Algorithms for the Method of Interior Points*, Nauka, Novosibirsk, USSR, 1980 (in Russian).

[2] C. C. Gonzaga, *Path-following methods for linear programming*, SIAM Rev., 34 (1992), pp. 167–224.

[3] C. C. Gonzaga and H. J. Lara, *A note on properties of condition numbers*, Linear Algebra Appl., 261 (1997), pp. 269–273.

[4] O. Güler and Y. Ye, *Convergence behavior of interior-point algorithms*, Math. Programming, 60 (1993), pp. 215–228.

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

[6] M. Kojima, S. Mizuno, and A. Yoshise, *A primal-dual interior-point algorithm for linear programming*, in Progress in Mathematical Programming, Interior-Point and Related Methods, N. Megiddo, ed., Springer-Verlag, New York, 1989, pp. 29–47.

[7] M. Kojima, S. Mizuno, and A. Yoshise, *A polynomial-time algorithm for a class of linear complementarity problems*, Math. Programming, 44 (1989), pp. 1–26.

[8] L. McLinden, *The complementarity problem for maximal monotone multifunction*, in Variational Inequalities and Complementarity Problems, R.W. Cottle, F. Giannessi, and J.L. Lions, ed., John Wiley, New York, 1980, pp. 251–270.

[9] N. Megiddo, *Pathways to the optimal set in linear programming*, in Progress in Mathematical Programming, Springer-Verlag, New York, Berlin, 1989, pp. 131–158.

[10] N. Megiddo, S. Mizuno, and T. Tsuchiya, *A modified layered-step interior-point algorithm for linear programming*, Math. Programming, 82 (1998), pp. 339–355.

[11] 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.

[12] S. Mizuno, N. Megiddo, and T. Tsuchiya, *A linear programming instance with many crossover events*, J. Complexity, 12 (1996) pp. 474–479.

[13] R. D. C. Monteiro and I. Adler, *Interior path-following primal-dual algorithms. Part* I: *Linear programming*, Math. Programming, 44 (1989), pp. 27–41.

[14] R. D. C. Monteiro and I. Adler, *Interior path-following primal-dual algorithms. Part* II: *Convex quadratic programming*, Math. Programming, 44 (1989), pp. 43–66.

[15] R. D. C. Monteiro and T. Tsuchiya, *Global convergence of the affine scaling algorithm for convex quadratic programming*, SIAM J. Optim., 8 (1998), pp. 26–58.

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

[17] G. W. Stewart, *On scaled projections and pseudoinverses*, Linear Algebra Appl., 112 (1989), pp. 189–193.

[18] K. Tanabe, *Centered Newton method for mathematical programming*, in System Modeling and Optimization, M. Iri and K. Yajima, eds., Springer-Verlag, Berlin, 1988, pp. 197–206.

[19] É. Tardos, *A strongly polynomial algorithm to solve combinatorial linear programs*, Oper. Res., 34 (1986), pp. 250–256.

[20] M. J. Todd, *A Dantzig-Wolfe-like variant of Karmarkar's interior-point linear programming algorithm*, Oper. Res., 38 (1990), pp. 1006–1018.

[21] M. J. Todd, L. Tunçel, and Y. Ye, *Characterizations, bounds, and probabilistic analysis of two complexity measures for linear programming problems*, Math. Programming, 90 (2001), pp. 59–69.

[22] T. Tsuchiya, *Global convergence property of the affine scaling method for primal degenerate linear programming problems*, Math. Oper. Res., 17 (1992), pp. 527–557.
[23] L. Tunçel, *Approximating the complexity measure of Vavasis-Ye algorithm is NP-hard*, Math. Program., 86 (1999), pp. 219–223.
[24] L. Tunçel, *On the condition numbers for polyhedra in Karmarkar's form*, Oper. Res. Lett., 24 (1999), pp. 149–155.
[25] R. J. Vanderbei and J. C. Lagarias, *I. I. Dikin's Convergence Result for the Affine-Scaling Algorithm*, Contemp. Math. 114, AMS, Providence, RI, 1990, pp. 109–119.
[26] S. Vavasis and Y. Ye, *A primal-dual accelerated interior-point method whose running time depends only on A*, Math. Programming, 74 (1996), pp. 79–120.
[27] S. Vavasis and Y. Ye, *A simplification to "A primal-dual interior-point method whose running time depends only on the constraint matrix,"* in High Performance Optimization, Appl. Optim. 33, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000, pp. 233–243.
[28] S. Vavasis and Y. Ye, *On the relationship between layered least squares and affine scaling steps*, in The Mathematics of Numerical Analysis, Lectures in Appl. Math. 32, AMS, Providence, RI, 1996, pp. 857–865.

SIAM J. OPTIM.
Vol. 13, No. 4, pp. 1080–1097

INEXACT VARIANTS OF THE PROXIMAL POINT ALGORITHM WITHOUT MONOTONICITY*

A. N. IUSEM†, T. PENNANEN‡, AND B. F. SVAITER†

Abstract. This paper studies convergence properties of inexact variants of the proximal point algorithm when applied to a certain class of nonmonotone mappings. The presented algorithms allow for constant relative errors, in the line of the recently proposed hybrid proximal-extragradient algorithm. The main convergence result extends a recent work of the second author, where exact solutions for the proximal subproblems were required. We also show that the linear convergence property is preserved in the case when the inverse of the operator is locally Lipschitz continuous near the origin. As an application, we give a convergence analysis for an inexact version of the proximal method of multipliers for a rather general family of problems which includes variational inequalities and constrained optimization problems.

Key words. proximal point algorithms, inexact iterates, hybrid proximal-extragradient algorithms, hypomonotone operators, multiplier methods

AMS subject classifications. 90C25, 90C30

PII. S1052623401399587

1. Introduction.

We deal in this paper with methods for finding zeroes of point-to-set operators in Hilbert spaces; i.e., given a Hilbert space H and an operator $T : H \to \mathcal{P}(H)$, we intend to find some $x^* \in H$ such that $0 \in T(x^*)$.

The proximal point algorithm, whose origins can be traced back to [9], was born in the 1960s (see, e.g., [12], [10]) and attained its current formulation in the works of Rockafellar [14], [15], where its connection with the augmented Lagrangian method for constrained nonlinear optimization was established. Basically, given a sequence $\{\gamma_n\}$ of positive real numbers bounded away from zero, the algorithm generates a sequence $\{x^n\} \subset H$, starting from some $x^0 \in H$, through the iteration

$$x^{n+1} \in (I + \gamma_n T)^{-1}(x^n). \tag{1}$$

When T is monotone, i.e.,

$$\langle x - y, u - v \rangle \geq 0 \tag{2}$$

for all $x, y \in H$, all $u \in T(x)$, and all $v \in T(y)$, and furthermore maximal monotone, i.e., $T = T'$ whenever $T' : H \to \mathcal{P}(H)$ is monotone and $T(x) \subset T'(x)$ for all $x \in H$, it follows from Minty's theorem (see [11]) that $I + \gamma T$ is onto and $(I + \gamma T)^{-1}$ is single-valued for all positive $\gamma \in \mathbb{R}$ so that the sequence defined by (1) is well defined. It has been proved in [14] that maximal monotonicity of T also ensures the weak convergence of the sequence $\{x^n\}$ defined by (1) to a zero of T when T has zeroes, and its unboundedness otherwise. Such weak convergence is global; i.e., the result just announced holds in fact for any $x^0 \in H$.

*Received by the editors December 12, 2001; accepted for publication (in revised form) October 21, 2002; published electronically March 19, 2003.
http://www.siam.org/journals/siopt/13-4/39958.html

†Instituto de Matemática Pura e Aplicada (IMPA), Estrada Dona Castorina 110, Rio de Janeiro, RJ, CEP 22460-320, Brazil (iusp@impa.br, benar@impa.br). The work of the first author was partially supported by CNPq grant 301280/86.

‡Department of Management Sciences, Helsinki School of Economics and Business Administration, PL 1210, 00101, Helsinki, Finland (pennanen@hkkk.fi).

The situation becomes considerably more complicated when T fails to be monotone. Augmented Lagrangian methods for minimization of nonconvex functions, a particular instance of the proximal point method for finding zeroes of nonmonotone operators, have been studied in [1], [6], and [20]. A survey of results on convergence of the proximal algorithm without monotonicity up to 1997 can be found in [8]. A new approach to the subject was taken in [13], which deals with a class of nonmonotone operators that, when restricted to a neighborhood of the solution set, are not far from being monotone. More precisely, it was assumed that, for some $\rho > 0$, the mapping $T^{-1} + \rho I$ is monotone when restricted to a neighborhood of $\hat{S}^* \times \{0\}$, where $\hat{S}^*$ is a nonempty connected component of the solution set $S^* = T^{-1}(0)$. When this happens, the main convergence result of [13] states that a "localized" version of (1) generates a sequence that converges to a point in $\hat{S}^*$, provided x^0 is close enough to $\hat{S}^*$ and $\inf \gamma_n > 2\rho$.

The issue of convergence of the algorithm under inexact computation of the iterates; i.e., when (1), or equivalently the inclusion

$$x^n - x^{n+1} \in \gamma_n T(x^{n+1}), \tag{3}$$

is solved only approximately, comes up immediately when dealing with the proximal algorithm for at least two reasons. First, it is generally imposible to find an exact value for x^{n+1} given by (1), or (3), particularly when T is nonlinear; second, it is clearly inefficient to spend too much effort in the computation of a given iterate x^n when only the limit of the sequence has the desired properties. Thus, the issue was dealt with even in the early treatment of the subject, e.g., in [14], but always, as far as we know, for the case of a monotone T. For instance, it has been proved in [14] that convergence is preserved when an error e^n is committed when performing the iteration given by (3), i.e., when (3) is replaced by

$$e^n + x^n - x^{n+1} \in \gamma_n T(x^{n+1}), \tag{4}$$

as long as

$$\sum_{n=0}^{\infty} \|e^n\| < \infty. \tag{5}$$

Other related conditions, but always including the summability of some measure of the error, can be found, e.g., in [14], [5]. These criteria are somewhat undesirable, because they impose increasing precision along the iterative process.

Recently, new related procedures have been presented in [17] and [18] which allow for constant relative error in the sense, e.g., that the norm of the error e^n in (4) must be smaller than a given fraction of the distance from the current iterate to the previous one. The price to be paid for this less stringent error criterion is that the resulting point (i.e., x^{n+1} in (4)) is not the next iterate, but rather an intermediate point which determines a hyperplane separating x^n from the solution set, and thus a direction pointing from x^n to this set; the actual next iterate is obtained by taking a certain step from x^n in this direction. More precisely, taking y^n as the intermediate point, and defining

$$\gamma_n v^n = e^n + x^n - y^n, \tag{6}$$

inclusion (4) becomes

$$v^n \in T(y^n) \tag{7}$$

and the error criterion is

$$\|e^n\| \leq \sigma \max\{\gamma_n \|v^n\|, \|y^n - x^n\|\}, \tag{8}$$

with $\sigma \in [0, 1)$. Indeed, the vector v^n gives the desired direction so that

$$x^{n+1} = x^n - \eta_n v^n \tag{9}$$

for some appropriate $\eta_n > 0$ (e.g., $\eta_n = \langle v^n, x^n - y^n \rangle / [\|v^n\|^2]$). It is important to emphasize that the additional cost of computing x^{n+1} once v^n and y^n have been determined (i.e., the cost of (9)) is negligible as compared to the solution, even in an inexact way, of the inclusion (4) (or the pair (6)–(7)). Algorithms of this kind have been called "hybrid" due to the presence of the additional step (9), in addition to the proximal step given by (6)–(7). We also remark that (8) can be seen as a sort of stopping criterion in an internal iterative procedure for the solution of (3): given candidate points (y^n, v^n) computed by such a procedure, if e^n as given by (6) satisfies (8), then v^n is accepted and x^{n+1} is computed according to (9); otherwise the procedure must continue, generating a new pair. In this sense, we can say that error criteria like (8) are particularly appropriate for computer implementation. In [17] it has been proved that the sequence generated by (6)–(9) is globally convergent to a zero of T in the weak topology, under the only assumptions of the existence of zeroes and the monotonicity of T (besides boundedness away from 0 of $\{\gamma_n\}$). Other related error criteria for the proximal point algorithm, allowing also for constant relative error, can be found in [19], [3], and [4].

In this paper we will consider the following inexact procedure for finding zeroes of an operator $T : H \to \mathcal{P}(H)$ whose inverse is maximal ρ-hypomonotone on a set $U \times V \subset H \times H$ (see Definition 1 below). Given $x^n \in H$, find $(y^n, v^n) \in U \times V$ such that

$$v^n \in T(y^n), \tag{10}$$

$$\gamma_n v^n + y^n - x^n = e^n, \tag{11}$$

where the error term e^n satisfies either

$$\|e^n\| \leq \sigma \left(\frac{\hat{\gamma}}{2} - \rho \right) \|v^n\| \tag{12}$$

or

$$\|e^n\| \leq \nu \|y^n - x^n\|, \tag{13}$$

with

$$\nu = \frac{\sqrt{\sigma + (1 - \sigma)(2\rho/\hat{\gamma})^2} - 2\rho/\hat{\gamma}}{1 + 2\rho/\hat{\gamma}}, \tag{14}$$

where $\sigma \in [0, 1)$, $\hat{\gamma} = \inf\{\gamma_n\}$, and ρ is the hypomonotonicity constant of T^{-1}. Then, under any of our two error criteria, the next iterate x^{n+1} is given by

$$x^{n+1} = x^n - \gamma_n v^n. \tag{15}$$

From now on, Algorithm 1 refers to the algorithm given by (10)–(12) and (15), and Algorithm 2 refers to the one given by (10), (11), and (13)–(15). We will prove that, when ρ-hypomonotonicity of T^{-1} holds on the whole space (i.e., $U = V = H$), both Algorithm 1 and Algorithm 2 generate sequences which are weakly convergent to a zero of T, starting from any $x^0 \in H$, under the assumptions of existence of zeroes of T and $2\rho < \hat{\gamma} = \inf\{\gamma_n\}$. For the case in which the set $U \times V$ where T^{-1} is ρ-hypomonotone is an appropriate neighborhood of $\hat{S}^* \times \{0\} \subset H \times H$, where $\hat{S}^*$ is a connected component of the set S^* of zeroes of T, we still get a local convergence result, meaning weak convergence of $\{x^n\}$ to a zero of T, but requiring additionally that x^0 be sufficiently close to $\hat{S}^* \cap U$, in a sense which is presented in a precise way in section 4.

We remark that when the tolerance σ vanishes, we get $e^n = 0$ from either (12) or (13)–(14), and then $x^n - y^n = \gamma_n v^n$ from (11), so that $x^{n+1} = y^n$ from (15). Thus, with $\sigma = 0$ our algorithm reduces to the exact algorithm in [13]. When comparing our analysis for this exact case with those in [13] it is worthwhile to point out the following difference: in [13] the proposed algorithm is studied by assuming first that T is locally monotone, and convergence is proved by showing that the resulting sequence coincides with the one generated by the algorithm applied to some maximal monotone operator. Then, the algorithm applied to a mapping whose inverse is locally ρ-hypomonotone is shown to be equivalent to an overrelaxed version of the proximal algorithm applied to the locally monotone operator $(T^{-1} + \rho I)^{-1}$, with a different sequence of regularization parameters, and convergence is finally obtained by invoking results in [5] on the convergence of such an overrelaxed variant of the proximal point algorithm. Our approach is less convoluted: we prove directly the Fejér monotonicity properties of $\{x^n\}$, which have as a consequence the weak convergence of $\{x^n\}$ to a zero of T. The issue of overrelaxation of the proximal point applied to the Yosida regularization $(T^{-1} + \rho I)^{-1}$ of T is confined to a lemma, also proved from scratch (up to an invocation of Minty's theorem), on the issue of existence of the iterates. Thus, our proof is (almost) self-contained and, in the exact case, it can be seen as a streamlined version of the analysis in [13].

2. Hypomonotone operators. From now on we will identify, in a set theoretic fashion, a point-to-set operator $T : H \to \mathcal{P}(H)$ with its graph, i.e., with $\{(x, v) \in H \times H : v \in T(x)\}$. Thus, $(x, v) \in T$ has the same meaning as $v \in T(x)$. We emphasize that (x, v) is seen here as an ordered pair, i.e., $(x, v) \in T$ (or equivalently $(v, x) \in T^{-1}$) is not the same as $(v, x) \in T$.

DEFINITION 1. *Given a positive $\rho \in \mathbb{R}$ and a subset W of $H \times H$, an operator $T : H \to \mathcal{P}(H)$ is said to be*

(a) *ρ-hypomonotone if and only if $\langle x-y, u-v\rangle \geq -\rho \|x - y\|^2$ for all $(x, u), (y, v) \in T$;*
(b) *maximal ρ-hypomonotone if and only if T is ρ-hypomonotone and additionally $T = T'$ whenever $T' \subset H \times H$ is ρ-hypomonotone and $T \subset T'$;*
(c) *ρ-hypomonotone in W if and only if $T \cap W$ is ρ-hypomonotone;*
(d) *maximal ρ-hypomonotone in W if and only if T is ρ-hypomonotone in W and additionally $T \cap W = T' \cap W$ whenever $T' \in H \times H$ is ρ-hypomonotone and $T \cap W \subset T' \cap W$.*

It follows from 13.33 and 13.36 of [16] that if a function $f : H \to \mathbb{R} \cup \{\infty\}$ can be written as $g - h$ in a neighborhood of a point $x \in H$, where g is finite and h is $\mathcal{C}^2$, then the subdifferential ∂f of f is ρ-hypomonotone for some $\rho > 0$ in a neighborhood of any point $(x, v) \in H \times H$ with $v \in \partial f(x)$. It is also easy to check that a locally

Lipschitz continuous mapping is hypomonotone for every ρ greater than the Lipschitz constant. In particular, if H is finite dimensional and $T : H \to \mathcal{P}(H)$ is such that T^{-1} is point-to-point and differentiable in a neighborhood of some $v \in H$, then T is ρ-hypomonotone in a neighborhood of (x, v) for any x such that $v \in T(x)$, and for any ρ larger than the absolute value of the most negative eigenvalue of $J + J^t$, where J is the Jacobian matrix of T^{-1} at v. In other words, local ρ-hypomonotonicity for some $\rho > 0$ is to be expected of any T which is not too badly behaved.

Note that a mapping T is ρ-hypomonotone if and only if $T + \rho I$ is monotone. We also have the following.

PROPOSITION 1. *If $T : H \to \mathcal{P}(H)$ is ρ-hypomonotone, then there exists a maximal ρ-hypomonotone $\hat{T} : H \to \mathcal{P}(H)$ such that $T \subset \hat{T}$.*

Proof. The proof is a routine application of Zorn's lemma, with exactly the same argument as the one used to prove that any monotone operator is contained in a maximal monotone one. ∎

Next we introduce in a slightly different way the Yosida regularization of an operator. For $\rho \geq 0$, define $Y_\rho : H \times H \to H \times H$ (Y for Yosida) as

$$Y_\rho(x, v) = (x + \rho v, v). \tag{16}$$

Observe that Y_ρ is a bijection, and $(Y_\rho)^{-1}(y, u) = (y - \rho u, u)$. Note also that

$$Y_\rho(T) = (T^{-1} + \rho I)^{-1}. \tag{17}$$

PROPOSITION 2. *Take $\rho \geq 0$, $T : H \to \mathcal{P}(H)$, and Y_ρ as in* (16). *Then*

(i) *T^{-1} is ρ-hypomonotone if and only if $Y_\rho(T)$ is monotone;*

(ii) *T^{-1} is maximal ρ-hypomonotone if and only if $Y_\rho(T)$ is maximal monotone.*

Proof.

(i) Monotonicity of the Yosida regularization means that $(T^{-1} + \rho I)^{-1}$ is monotone, which is equivalent to the monotonicity of $T^{-1} + \rho I$.

(ii) Assume that T^{-1} is maximal ρ-hypomonotone. We prove the maximal monotonicity of $Y_\rho(T)$. The monotonicity follows from item (a). Assume that $Y_\rho(T) \subset Q$ for some monotone $Q \subset H \times H$. Note that $Q = Y_\rho(T_Q)$ for some T_Q because Y_ρ is a bijection. It follows, in view of (i) and the monotonicity of Q, that T_Q^{-1} is ρ-hypomonotone, and therefore, using again the bijectivity of Y_ρ, we have $T^{-1} \subset T_Q^{-1}$. Since T^{-1} is maximal ρ-hypomonotone, we conclude that $T^{-1} = T_Q^{-1}$, i.e., $T = T_Q$, so that $Q = Y_\rho(T') = Y_\rho(T)$, proving that $Y_\rho(T)$ is maximal monotone. The converse statement is proved with a similar argument. ∎

We continue with an elementary result on the Yosida regularization $Y_\rho(T)$.

PROPOSITION 3. *For all $T : H \to \mathcal{P}(H)$ and all $\rho \geq 0$, $0 \in T(x)$ if and only if $0 \in [Y_\rho(T)](x)$.*

Proof. The result follows immediately from (17). ∎

Remark 1. It is well known that the set of zeroes of a monotone operator is closed and convex. In view of Propositions 2 and 3, the same holds for mappings whose inverses are ρ-hypomonotone. Thus, though reasonably well-behaved operators can be expected to be locally ρ-hypomonotone for some $\rho > 0$, as discussed above, global ρ-hypomonotonicity is not at all generic; looking for instance at point-to-point operators in $\mathbb{R}$, we observe that polynomials with more than one real root, or analytic functions like $T(x) = \sin x$, are not ρ-hypomonotone for any $\rho > 0$.

Next we establish local demiclosedness of maximal locally ρ-hypomonotone operators, with a proof which mirrors the one on demiclosedness of maximal monotone operators.

PROPOSITION 4. *Assume that $T^{-1} : H \to \mathcal{P}(H)$ is maximal ρ-hypomonotone in W^{-1} for some $W \subset H \times H$, and consider a sequence $\{(x^n, v^n)\} \subset T \cap W$. If $\{v^n\}$ is strongly convergent to $\bar{v}$, $\{x^n\}$ is weakly convergent to $\bar{x}$, and $(\bar{x}, \bar{v}) \in W$, then $\bar{v} \in T(\bar{x})$.*

Proof. Define $T' : H \to \mathcal{P}(H)$ as $T' = T \cup \{(\bar{x}, \bar{v})\}$. We claim that $(T')^{-1}$ is ρ-hypomonotone in W^{-1}. Since T^{-1} is ρ-hypomonotone in W^{-1}, clearly it suffices to prove that

$$-\rho \|\bar{v} - v\|^2 \leq \langle \bar{x} - x, \bar{v} - v \rangle \tag{18}$$

for all $(x, v) \in T \cap W$. Observe that, for all $(x, v) \in T \cap W$,

$$-\rho \|v^n - v\|^2 \leq \langle x^n - x, v^n - v \rangle. \tag{19}$$

Since $\{v^n\}$ is strongly convergent to $\bar{v}$ and $\{x^n\}$ is weakly convergent to $\bar{x}$, taking limits in (19) as $n \to \infty$ we obtain (18), and the claim is established. Since $T \subset T'$, $(T')^{-1}$ is ρ-hypomonotone in W^{-1}, and T^{-1} is maximal ρ-hypomonotone in W^{-1}, we have that $T \cap W = T' \cap W$ by Definition 1(d). Since $\bar{v} \in T'(\bar{x})$ and $(\bar{x}, \bar{v}) \in W$, we conclude that $\bar{v} \in T(\bar{x})$. ∎

We close this section with a result on convexity and weak closedness of some sets related to the set of zeroes of operators whose inverses are ρ-hypomonotone. We use the usual notation for sums of sets; i.e., for $A, B \subset H$, $A + B \subset H$ is defined as $A + B = \{x + y : x \in A, y \in B\}$. Also, for $x \in H$ and $\delta > 0$, $B(x, \delta)$ will denote the closed ball of radius δ centered at x.

PROPOSITION 5. *Assume that $T^{-1} : H \to \mathcal{P}(H)$ is maximal ρ-hypomonotone in a subset $V \times U \subset H \times H$, where U is convex and $0 \in V$. Let $S^* \subset H$ be the set of zeroes of T. Then*

(i) *$S^* \cap U$ is convex;*

(ii) *if $S^* \cap U$ is closed, then $(S^* \cap U) + B(0, \delta)$ is weakly closed for all $\delta \geq 0$.*

Proof.

(i) By Proposition 1, $T \cap (U \times V) \subset \hat{T}$ for some $\hat{T} : H \to \mathcal{P}(H)$ such that $\hat{T}^{-1}$ is maximal ρ-hypomonotone. Let $\hat{S}^*$ be the set of zeroes of $\hat{T}$. By Proposition 3, $\hat{S}^*$ is also the set of zeroes of $Y_\rho(\hat{T})$, which is maximal monotone by Proposition 2(ii). Since the set of zeroes of a maximal monotone operator is convex (e.g., 12.8(a) and (c) in [16]), we conclude that $\hat{S}^*$ is convex, and therefore $\hat{S}^* \cap U$ is convex, because U is convex. Since T^{-1} is maximal ρ-hypomonotone in $V \times U$, $\hat{T}^{-1}$ is ρ-hypomonotone, and $T \subset \hat{T}$, we have that $\hat{T} \cap (U \times V) = T \cap (U \times V)$, and then, since $0 \in V$, it follows easily that $\hat{S}^* \cap U = S^* \cap U$. The result follows.

(ii) Since H is a Hilbert space, $B(0, \delta)$ is weakly compact (e.g., Theorem III.8 in [2]), and $S^* \cap U$, being closed by assumption and convex by item (i), is weakly closed. Thus $(S^* \cap U) + B(0, \delta)$ is weakly closed, being the sum of a closed and a compact set, both with respect to the weak topology. ∎

3. Existence results. The issue of existence of iterates for proximal algorithms applied to nonmonotone operators is delicate. The main tool used for establishing existence in the monotone case, namely Minty's theorem, does not work without monotonicity. Overcoming this obstacle requires some technicalities, where the notion of ρ-hypomonotonicity becomes crucial.

Note that if a pair (y^n, v^n) satisfies (10)–(12), or (10)–(11) together with (13)–(14), with $\sigma = 0$ (which in the second case implies $\nu = 0$), then such a pair satisfies those conditions with any $\sigma > 0$. Since $\sigma = 0$ also implies that the error term e^n vanishes, existence of exact iterates is enough to settle the existence issue for our inexact schemes. Now, we already mentioned that our scheme reduces, in the absence of errors, to the algorithm studied in [13], and therefore we could refer to the existence results in this reference without further discussion. But since we dressed our setting somewhat differently from the one in [13] (e.g., the definition of local ρ-hypomonotonicity), we prefer to offer a full proof, which also contributes to making this paper more self-contained. The technicalities will be encapsulated in the following lemma, where, for $x \in H$ and $A \subset H$, $d(x, A)$ will denote the distance from x to A, i.e., $d(x, A) = \inf_{y \in A} \|x - y\|$.

LEMMA 1. *Let $T : H \to \mathcal{P}(H)$ be an operator such that T^{-1} is maximal ρ-hypomonotone in a subset $V \times U$ of $H \times H$. Assume that T has a nonempty set of zeroes S^*, that U is convex, and that*

(i) *$S^* \cap U$ is nonempty and closed;*
(ii) *there exists $\beta > 0$ such that $B(0, \beta) \subset V$;*
(iii) *there exists $\delta > 0$ such that $(S^* \cap U) + B(0, \delta) \subset U$.*

Take any $\gamma > 2\rho$ and define $\varepsilon = \min\{\delta, \beta\gamma/2\}$. If $x \in H$ is such that $d(x, S^* \cap U) \leq \varepsilon$, then there exists $y \in H$ such that $\gamma^{-1}(x - y) \in T(y)$ and $d(y, S^* \cap U) \leq \varepsilon$.

Proof. By Definition 1(c) and (d), $T^{-1} \cap (V \times U)$ is ρ-hypomonotone. By Proposition 1, there exists a maximal ρ-hypomonotone $\hat{T}^{-1} \subset H \times H$ such that

$$[T^{-1} \cap (V \times U)] \subset \hat{T}^{-1}. \tag{20}$$

By Proposition 2(ii), $Y_\rho(\hat{T})$ is maximal monotone, with Y_ρ as defined in (16). Let $\hat{\gamma} = \gamma - \rho$. Since $\hat{\gamma} > 0$ by assumption, it follows from Minty's theorem (see [11]) that the operator $[I + \hat{\gamma} Y_\rho(\hat{T})]^{-1}$ is onto (and also one-to-one, but this does not concern us) so that there exists $z \in H$ such that $x \in [I + \hat{\gamma} Y_\rho(\hat{T})]^{-1}(z)$, or equivalently

$$\hat{\gamma}^{-1}(x - z) \in \left[Y_\rho(\hat{T})\right](z). \tag{21}$$

Letting

$$v := \hat{\gamma}^{-1}(x - z), \tag{22}$$

we can rewrite (21) as $(z, v) \in Y_\rho(\hat{T})$, which is equivalent, in view of (16), to

$$(z - \rho v, v) \in \hat{T}. \tag{23}$$

Let now $y = z - \rho v$. In view of (22) and the definition of $\hat{\gamma}$, (23) is in turn equivalent to

$$(y, \hat{\gamma}^{-1}(x - z)) \in \hat{T}. \tag{24}$$

It follows easily from (22) and the definitions of y and $\hat{\gamma}$ that

$$\hat{\gamma}^{-1}(x - z) = (\gamma - \rho)^{-1}(x - z) = \gamma^{-1}(x - y) = \rho^{-1}(z - y). \tag{25}$$

We conclude from (24) and (25) that

$$\gamma^{-1}(x - y) \in \hat{T}(y). \tag{26}$$

Note that (26) looks pretty much like the statement of the lemma, except that we have $\hat{T}$ instead of T. The operators T and $\hat{T}$ do coincide on $U \times V$, as we will see, but in order to use this fact we must first establish that $(y, \gamma^{-1}(x-y))$ belongs indeed to $U \times V$, which will result from the assumption on $d(x, S^* \cap U)$. The analysis in the following paragraph is tantamount to establishing Fejér monotonicity of the iterates of an overrelaxed proximal algorithm applied to a maximal monotone operator, which can be found in [5].

Take any $\bar{x} \in S^* \cap U$, nonempty by condition (i), and z as in (21). Note that $\bar{x}$ is a zero of $T \cap (U \times V)$, because it belongs to $S^* \cap U$ and $0 \in V$ by condition (ii). Thus $\bar{x}$ is a zero of $\hat{T}$, which contains $T \cap (U \times V)$. By Proposition 3, $\bar{x}$ is a zero of $Y_\rho(T)$. Then

$$\|x - \bar{x}\|^2 = \|x - z\|^2 + \|z - \bar{x}\|^2 + 2\langle x - z, z - \bar{x}\rangle$$

$$= \|x - z\|^2 + \|z - \bar{x}\|^2 + 2\hat{\gamma}\langle \hat{\gamma}^{-1}(x - z) - 0, z - \bar{x}\rangle \geq \|x - z\|^2 + \|z - \bar{x}\|^2, \tag{27}$$

using (21), the monotonicity of $Y_\rho(\hat{T})$, the nonnegativity of $\hat{\gamma}$, and the fact that $\bar{x}$ is a zero of $Y_\rho(\hat{T})$ in the inequality. Take now y as defined after (23). Then

$$\begin{aligned}
\|y - \bar{x}\|^2 &= \|y - z\|^2 + \|z - \bar{x}\|^2 - 2\langle y - z, \bar{x} - z\rangle \\
&= \left(\frac{\rho}{\gamma - \rho}\right)^2 \|z - x\|^2 + \|z - \bar{x}\|^2 - \frac{2\rho}{\gamma - \rho}\langle z - x, \bar{x} - z\rangle \\
&\leq \|x - \bar{x}\|^2 - \left[1 - \left(\frac{\rho}{\gamma - \rho}\right)^2\right] \|z - x\|^2 - 2\rho\langle 0 - (\gamma - \rho)^{-1}(x - z), \bar{x} - z\rangle \\
&\leq \|x - \bar{x}\|^2 - \left[1 - \left(\frac{\rho}{\gamma - \rho}\right)^2\right] \|z - x\|^2 = \|x - \bar{x}\|^2 - \frac{\gamma(\gamma - 2\rho)}{(\gamma - \rho)^2} \|z - x\|^2 \leq \|x - \bar{x}\|^2,
\end{aligned} \tag{28}$$

using (25) in the first equality, (27) in the first inequality, (21) and the monotonicity of $Y_\rho(\hat{T})$ in the second inequality, and the assumption that $\gamma > 2\rho$ in the third inequality. It follows from (28) that $\|y - \bar{x}\| \leq \|x - \bar{x}\|$ for all $\bar{x} \in S^* \cap U$, in particular when $\bar{x}$ is the orthogonal projection of x onto $S^* \cap U$, which exists because $S^* \cap U$ is closed by condition (i) and convex by Proposition 5(i). For this choice of $\bar{x}$ we have that

$$\|y - \bar{x}\| \leq \|x - \bar{x}\| = d(x, S^* \cap U) \leq \varepsilon = \min\{\delta, \beta\gamma/2\} \leq \delta, \tag{29}$$

where the second inequality holds by the assumption on x. Since $\bar{x}$ belongs to $S^* \cap U$, we get from (29) that

$$y \in (S^* \cap U) + B(0, \delta) \subset U, \tag{30}$$

using condition (iii) in the inclusion.

Observe now that, with the same choice of $\bar{x}$,

$$\gamma^{-1}\|x - y\| \leq \gamma^{-1}(\|x - \bar{x}\| + \|y - \bar{x}\|) \leq 2\varepsilon\gamma^{-1} \leq \beta, \tag{31}$$

using (29) and the assumption on x in the second inequality, and the fact that $\varepsilon = \min\{\delta, \beta\gamma/2\}$ in the third inequality. It follows from (30), (31), and condition (ii) that

$$(y, \gamma^{-1}(x - y)) \in U \times V. \tag{32}$$

Since T is maximal ρ-hypomonotone in $U \times V$ and $\hat{T}$ is ρ-hypomonotone, it follows from (20) and Definition 1(d) that $T \cap (U \times V) = \hat{T} \cap (U \times V)$. In view of (26), we conclude from (32) that $\gamma^{-1}(x - y) \in T(y)$. Finally, using (29), $d(y, S^* \cap U) \leq \|y - \bar{x}\| \leq \varepsilon$, completing the proof. ∎

Remark 2. As we mentioned along the proof of Lemma 1, the vector z satisfying (21) is unique by Minty's theorem, and thus it is easy to check that the vector y in the statement of the lemma is also unique. This is not too relevant for our inexact algorithms: the iterates are unique for $\sigma = 0$ (i.e., in the exact case) but hopefully not so for other values of σ. We emphasize that uniqueness of the iterates is no blessing for inexact algorithms; it is rather catastrophic.

COROLLARY 1. *Consider either Algorithm* 1 *or Algorithm* 2 *applied to an operator* $T : H \to \mathcal{P}(H)$ *which is* ρ*-hypomonotone on a subset* $U \times V$ *of* $H \times H$ *satisfying conditions* (i)–(iii) *of Lemma* 1. *If* $d(x^n, S^* \cap U) \leq \varepsilon$, *with* ε *as in the statement of Lemma* 1, *and* $\gamma_n > 2\rho$, *then there exists a pair* $(y^n, v^n) \in U \times V$ *satisfying* (10) *and* (11), *and consequently a vector* x^{n+1} *satisfying* (15).

Proof. Apply Lemma 1 with $x = x^n$, $\gamma = \gamma_n$. Take y^n as the vector y whose existence is ensured by the lemma and $v^n = \gamma_n^{-1}(x^n - y^n)$. Then y^n and v^n satisfy (10) and (11) with $e^n = 0$ so that (12) or (13)–(14) hold for any $\sigma \geq 0$. Once a pair (y^n, v^n) exists, the conclusion about x^{n+1} is obvious, since (15) raises no existence issues. ∎

In order to ensure existence of the iterates, we still have to prove, in view of Corollary 1, that the whole sequence $\{x^n\}$ is contained in $B(\bar{x}, \varepsilon)$, where $\bar{x}$ is the orthogonal projection of x^0 onto $S^* \cap U$ and ε is as in Lemma 1. This will be a consequence of the Fejér monotonicity properties of $\{x^n\}$, which we will establish in the following section.

4. Convergence analysis. The next lemma establishes the Fejér monotonicity property of sequences generated by our inexact algorithm. We have not yet proved the existence of such sequences, but the lemma is phrased so as to circumvent the existential issue for the time being.

LEMMA 2. *Let* $\{x^n\} \subset H$ *be a sequence generated by either Algorithm* 1 *or Algorithm* 2 *applied to an operator* $T : H \to \mathcal{P}(H)$ *such that* T^{-1} *is* ρ*-hypomonotone in a subset* W^{-1} *of* $H \times H$, *and take* x^* *in the set* S^* *of zeroes of* T. *If* $2\rho < \hat{\gamma} = \inf\{\gamma_n\}$ *and both* $(x^*, 0)$ *and* (y^n, v^n) *belong to* W, *then*

(i)

$$\left\|x^{n+1} - x^*\right\|^2 \leq \|x^n - x^*\|^2 - (1 - \sigma)\gamma_n(\hat{\gamma} - 2\rho)\,\|v^n\|^2$$

for Algorithm 1 *and*

(ii)

$$\left\|x^* - x^{n+1}\right\|^2 \leq \|x^* - x^n\|^2 - (1 - \sigma)\left(1 - \frac{2\rho}{\hat{\gamma}}\right)\|y^n - x^n\|^2$$

for Algorithm 2.

Proof. We start with the following elementary algebraic equality:

(33)
$$\|x^* - x^n\|^2 - \left\|x^* - x^{n+1}\right\|^2 - \|y^n - x^n\|^2 + \left\|y^n - x^{n+1}\right\|^2 = 2\langle x^* - y^n, x^{n+1} - x^n\rangle.$$

Using first (15) in the right-hand side of (33), and then ρ-hypomonotonicity of T^{-1} in W^{-1}, together with the fact that both $(x^*, 0)$ and (y^n, v^n) belong to $T \cap W$, by

(10) and the assumptions of the lemma we get

$$\|x^* - x^n\|^2 - \left\|x^* - x^{n+1}\right\|^2 - \|y^n - x^n\|^2 + \left\|y^n - x^{n+1}\right\|^2$$

$$= 2\gamma_n \langle x^* - y^n, 0 - v^n \rangle \geq -2\rho\gamma_n \|v^n\|^2 . \tag{34}$$

From this point the computations differ according to the error criterion. We start with the one given by (12). It follows from (11) and (15) that $y^n - x^n = e^n - \gamma_n v^n$ and $y^n - x^{n+1} = e^n$. Substituting these two equalities in (34) we get

$$\|x^* - x^n\|^2 - \left\|x^* - x^{n+1}\right\|^2 \geq \gamma_n^2 \|v^n\|^2 - 2\gamma_n \langle v^n, e^n \rangle - 2\rho\gamma_n \|v^n\|^2$$

$$\geq \gamma_n^2 \|v^n\|^2 - 2\gamma_n \|v^n\| \, \|e^n\| - 2\rho\gamma_n \|v^n\|^2 = \gamma_n \|v^n\| \left[(\gamma_n - 2\rho) \|v^n\| - 2 \|e^n\| \right]$$

$$\geq \gamma_n \|v^n\| \left[(\gamma_n - 2\rho) \|v^n\| - \sigma(\hat{\gamma} - 2\rho) \|v^n\| \right] \geq \gamma_n \|v^n\| \left[(1 - \sigma)(\hat{\gamma} - 2\rho) \|v^n\| \right]$$

$$= (1 - \sigma)\gamma_n (\hat{\gamma} - 2\rho) \|v^n\|^2 , \tag{35}$$

using (12) in the third inequality and the definition of $\hat{\gamma}$ in the last inequality. The results follows immediately from (35).

Now we look at the error criterion given by (13)–(14). Using again (11) and (15), we can replace $y^n - x^{n+1}$ by e^n and $-v^n$ by $\gamma_n^{-1}(y^n - x^n - e^n)$ in (34), obtaining

$$\|x^* - x^n\|^2 - \left\|x^* - x^{n+1}\right\|^2 \geq \|y^n - x^n\|^2 - \left(\|e^n\|^2 + 2\rho\gamma_n^{-1} \|y^n - x^n - e^n\|^2 \right)$$

$$\geq \|y^n - x^n\|^2 - \left[\|e^n\|^2 + 2\rho\gamma_n^{-1} \left(\|y^n - x^n\| + \|e^n\| \right)^2 \right] . \tag{36}$$

Using now (13) in (36) we get

$$\|x^* - x^n\|^2 - \left\|x^* - x^{n+1}\right\|^2 \geq \left[1 - \nu^2 - \frac{2\rho}{\gamma_n}(1 + \nu)^2 \right] \|y^n - x^n\|^2$$

$$\geq \left[1 - \nu^2 - \frac{2\rho}{\hat{\gamma}}(1 + \nu)^2 \right] \|y^n - x^n\|^2 . \tag{37}$$

It follows from (14), after some elementary algebra, that

$$\left[1 - \nu^2 - \frac{2\rho}{\hat{\gamma}}(1 + \nu)^2 \right] = (1 - \sigma) \left(1 - \frac{2\rho}{\hat{\gamma}} \right) . \tag{38}$$

Replacing (38) in (37), we obtain

$$\|x^* - x^n\|^2 - \left\|x^* - x^{n+1}\right\|^2 \geq (1 - \sigma) \left(1 - \frac{2\rho}{\hat{\gamma}} \right) \|y^n - x^n\|^2 , \tag{39}$$

and the results follows immediately from (39). □

Next we combine the results of Lemmas 1 and 2 in order to obtain our convergence theorem.

THEOREM 1. *Let $T : H \to \mathcal{P}(H)$ so that T^{-1} is maximal ρ-hypomonotone in a subset $V \times U$ of $H \times H$ satisfying*

(i) *$S^* \cap U$ is nonempty and closed;*

(ii) *there exists $\beta > 0$ such that $B(0, \beta) \subset V$;*

(iii) *there exists* $\delta > 0$ *such that* $(S^* \cap U) + B(0, \delta) \subset U$*;*
(iv) U *is convex,*

where S^* *is the set of zeroes of* T*. Take a sequence* $\{\gamma_n\}$ *of positive real numbers such that* $2\rho < \hat{\gamma} = \inf\{\gamma_n\}$*. Define* $\varepsilon = \min\{\delta, \beta\hat{\gamma}/2\}$*. If* $d(x^0, S^* \cap U) \leq \varepsilon$*, then, for both Algorithm* 1 *and Algorithm* 2*,*

(a) *for all* n *there exist* $y^n, v^n, e^n, x^{n+1} \in H$ *satisfying* (10)–(12) *and* (15)*, in the case of Algorithm* 1*, and* (10)–(11) *and* (13)–(15)*, in the case of Algorithm* 2*, and such that* $(y^n, v^n) \in U \times V$*,* $d(x^{n+1}, S^* \cap U) \leq \varepsilon$*;*
(b) *for any sequence as in* (a)*, we have that* (x^n) *converges weakly to a point in* $S^* \cap U$*.*

Proof.

(a) We proceed by induction. Take any $n \geq 0$. We have that

$$d(x^n, S^* \cap U) \leq \varepsilon, \tag{40}$$

by inductive hypothesis, if $n \geq 1$, and by assumption if $n = 0$. We are within the hypotheses of Corollary 1, which indicates that the desired vectors exist and that $(y^n, v^n) \in U \times V$. It remains to establish that $d(x^{n+1}, S^* \cap U) \leq \varepsilon$. Let $\bar{x}$ be the orthogonal projection of x^n onto $S^* \cap U$, which exists by condition (i) and (iv) and Proposition 5. Note that both $(\bar{x}, 0)$ and (y^n, v^n) belong to $U \times V$. Thus we are within the hypotheses of Lemma 2, with $W = U \times V$, and both for Algorithm 1 and Algorithm 2 we get from either Lemma 2(i) or Lemma 2(ii) that

$$\left\|x^* - x^{n+1}\right\| \leq \|x^* - x^n\| \tag{41}$$

for all $x^* \in S^* \cap U$. By (41) with $\bar{x}$ instead of x^*,

$$d(x^{n+1}, S^* \cap U) \leq \left\|\bar{x} - x^{n+1}\right\| \leq \|\bar{x} - x^n\| = d(x^n, S^* \cap U) \leq \varepsilon,$$

using (40) in the last inequality.

(b) We follow here with minor variations the standard convergence proof for the proximal point algorithm; see, e.g., [14]. In view of (41), for all $x^* \in S^* \cap U$ the sequence $\{\|x^n - x^*\|\}$ is nonincreasing, and certainly nonnegative, and hence convergent. Also, since $\|x^n - x^*\| \leq \left\|x^0 - x^*\right\|$ for all n, we get that $\{x^n\}$ is bounded.
Now we consider separately both algorithms. In the case of Algorithm 1, we get from Lemma 2(i)

$$(1 - \sigma)(\hat{\gamma} - 2\rho)\gamma_n \|v^n\|^2 \leq \|x^n - x^*\|^2 - \left\|x^{n+1} - x^*\right\|^2. \tag{42}$$

Since the right-hand side of (42) converges to 0, we conclude that $\lim_{n\to\infty} \gamma_n \|v^n\| = 0$, and therefore, since $\gamma_n \geq \hat{\gamma} > 0$ for all n,

$$\lim_{n\to\infty} v^n = 0, \tag{43}$$

which implies, in view of (12), that $\lim_{n\to\infty} e^n = 0$, and therefore, by (11),

$$\lim_{n\to\infty} (y^n - x^n) = 0. \tag{44}$$

In the case of Algorithm 2, we get from Lemma 2(ii)

$$(1 - \sigma)\left(1 - \frac{2\rho}{\hat{\gamma}}\right)\|y^n - x^n\|^2 \leq \|x^* - x^n\|^2 - \left\|x^* - x^{n+1}\right\|^2. \tag{45}$$

Again, the right-hand side of (45) converges to 0, and thus (44) also holds in this case, so that, in view of (13), $\lim_{n\to\infty} e^n = 0$, which gives, in view of (44) and (11), $\lim_{n\to\infty} \gamma_n v^n = 0$, so that in this case we also have (43). We have proved that (43) and (44) hold both for Algorithm 1 and Algorithm 2, and we proceed from now on with an argument which holds for both algorithms. Since $\{x^n\}$ is bounded, it has weak cluster points. Let $\tilde{x}$ be any weak cluster point of $\{x^n\}$; i.e., $\tilde{x}$ is the weak limit of a subsequence $\{x^{k_n}\}$ of $\{x^n\}$. By (44), $\tilde{x}$ is also the weak limit of $\{y^{k_n}\}$. We claim that $(\tilde{x}, 0)$ belongs to $U \times V$. In view of condition (ii), it suffices to check that $\tilde{x} \in U$. Note that $\{x^n\} \subset (S^* \cap U) + B(0, \varepsilon)$ by item (a). Since U is convex by condition (iv) and $S^* \cap U$ is closed by condition (i), we can apply Proposition 5(ii) to conclude that $(S^* \cap U) + B(0, \varepsilon)$ is weakly closed. Thus, the weak limit $\tilde{x}$ of $\{x^{n_k}\}$ belongs to $(S^* \cap U) + B(0, \varepsilon)$, and henceforth to U, in view of condition (ii) and the fact that $\varepsilon \leq \delta$. The claim holds, and we are within the hypotheses of Proposition 4: $\{v^{k_n}\}$ is strongly convergent to 0 by (43), $\{x^{k_n}\}$ is weakly convergent to $\tilde{x}$, and $(0, \tilde{x})$ belongs to $V \times U$, where T^{-1} is maximal ρ-hypomonotone. Then $0 \in T(\tilde{x})$, i.e., $\tilde{x} \in S^* \cap U$.

Finally we establish uniqueness of the weak cluster point of $\{x^n\}$, with the standard argument (e.g., [14]) which we include just in order to keep our self-containment policy. Let $\tilde{x}$, $\hat{x}$ be two weak cluster points of $\{x^n\}$, say the weak limits of $\{x^{k_n}\}$, $\{x^{j_n}\}$, respectively. We have just proved that both $\tilde{x}$ and $\hat{x}$ belong to $S^* \times U$, and thus, by (41), both $\{\|\hat{x} - x^n\|\}$ and $\{\|\tilde{x} - x^n\|\}$ are nonincreasing, and hence convergent, say, to $\hat{\alpha} \geq 0$ and to $\tilde{\alpha} \geq 0$, respectively. Now,

$$\|\hat{x} - x^n\|^2 = \|\hat{x} - \tilde{x}\|^2 + \|\tilde{x} - x^n\|^2 + 2\langle \hat{x} - \tilde{x}, \tilde{x} - x^n \rangle. \tag{46}$$

Taking limits in (46) as $n \to \infty$ along the subsequence $\{x^{k_n}\}$, we get

$$\|\hat{x} - \tilde{x}\|^2 = \hat{\alpha}^2 - \tilde{\alpha}^2. \tag{47}$$

Reversing now the roles of $\tilde{x}$, $\hat{x}$ in (46), and taking limits along the subsequence $\{x^{j_n}\}$, we get

$$\|\hat{x} - \tilde{x}\|^2 = \tilde{\alpha}^2 - \hat{\alpha}^2. \tag{48}$$

It follows from (47) and (48) that $\tilde{x} = \hat{x}$, and thus the whole sequence $\{x^n\}$ has a weak limit which is a zero of T and belongs to U. ∎

The next corollary states the global result for the case in which T^{-1} is ρ-hypomonotone in the whole $H \times H$.

COROLLARY 2. *Assume that $T : H \to \mathcal{P}(H)$ has a nonempty set of zeroes S^* and that T^{-1} is maximal ρ-hypomonotone. Take a sequence $\{\gamma_n\}$ of positive real numbers such that $2\rho < \hat{\gamma} = \inf\{\gamma_n\}$. Then, for both Algorithm 1 and Algorithm 2, given any $x^0 \in H$,*

(a) *for all n there exist $y^n, v^n, e^n, x^{n+1} \in H$ satisfying (10)–(12) and (15), in the case of Algorithm 1, and (10)–(11) and (13)–(15), in the case of Algorithm 2;*

(b) *any sequence generated by either Algorithm 1 or Algorithm 2 is weakly convergent to a point in S^*.*

Proof. This is just Theorem 1 for the case of $U = V = H$. In this case all the assumptions above hold trivially. Regarding condition (i), note that S^* is closed

because, by Proposition 3, it is also the set of zeroes of the maximal monotone operator $Y_\rho(T)$, which is closed (see, e.g., 12.8(a) and (c) in [16]). Conditions (ii) and (iii) hold for any $\beta, \delta > 0$ so that the result will hold for any $\varepsilon > 0$, in particular for $\varepsilon > d(x^0, S^*)$. □

We close this section with a restatement of Theorem 1, which is needed in our section on multiplier methods.

COROLLARY 3. *Let $Z \subset H$ be a linear subspace, and consider Algorithms* 1 *and* 2*, with the additional requirement that e^n, besides satisfying* (12) *or* (13)–(14)*, belong to Z. Then, the results of Theorem* 1 *still hold.*

Proof. The inductive step in the proof of Theorem 1(a), based on Lemma 1, essentially consists of establishing, for all n, the existence of exact iterates, i.e., with $e^n = 0$, which certainly belongs to Z, so that item (a) does hold with the additional requirement. The proof of item (b), depending on the results of Lemma 2, is valid for all sequences $\{x^n\}$ as in item (a), and hence in particular for those sequences such that $e^n \in Z$. □

5. Convergence rate results. We prove in this section that our inexact algorithm still enjoys the convergence rate results which are already classical for proximal point algorithms, namely, at least a linear convergence rate when T^{-1} is locally Lipschitz at 0 (see [14] for the monotone case and [13] for the nonmonotone one with exact iterates). We will say that $Q : H \to \mathcal{P}(H)$ is Lipschitz continuous at $W \subset H$ if there exists a constant $\lambda \geq 0$ such that $\|v - v'\| \leq \lambda \|x - x'\|$ for all, $x, x' \in W$, all $v \in Q(x)$, and all $v' \in Q(x')$. Our convergence rate result is the following.

THEOREM 2. *Under all the assumptions of Theorem* 1*, suppose furthermore that T^{-1} is Lipschitz continuous, with constant λ, in a neighborhood $W \subset H$ of* 0*. Let x^* be the weak limit point of the sequence $\{x^n\}$. Then there exists n_0 such that the following inequalities hold for all $n \geq n_0$:*

(i)

$$\left\|x^{n+1} - x^*\right\| \leq \frac{\lambda + \mu}{\sqrt{(\lambda + \mu)^2 + \theta_n}} \|x^n - x^*\| \tag{49}$$

for Algorithm 1*, where*

$$\mu = \sigma\left(\frac{\hat{\gamma}}{2} - \rho\right), \quad \theta_n = (1 - \sigma)\gamma_n(\hat{\gamma} - 2\rho); \tag{50}$$

(ii)

$$\left\|x^{n+1} - x^*\right\| \leq \frac{\omega_n}{\sqrt{\omega_n^2 + \xi}} \|x^n - x^*\| \tag{51}$$

for Algorithm 2*, where*

$$\xi = (1 - \sigma)\left(1 - \frac{2\rho}{\hat{\gamma}}\right), \quad \omega_n = \nu + (1 + \nu)\frac{\lambda}{\gamma_n}, \tag{52}$$

with ν as in (13)*.*

Proof. By (10), $y^n \in T^{-1}(v^n)$. By Theorem 1(b), $x^* \in T^{-1}(0)$. By (43), $\lim_{n\to\infty} v^n = 0$, and so there exists n_0 such that $v^n \in W$ for $n \geq n_0$. By Lipschitz continuity of T^{-1} in W, for $n \geq n_0$,

$$\|y^n - x^*\| \leq \lambda \|v^n\|. \tag{53}$$

By (11) and (15), $x^{n+1} = y^n - e^n$. Thus,

$$\left\|x^{n+1} - x^*\right\| = \|y^n - x^* - e^n\| \leq \|y^n - x^*\| + \|e^n\| \leq \lambda \|v^n\| + \|e^n\|, \tag{54}$$

using (53) in the second inequality.

Now we consider separately both algorithms.

(i) Combining Lemma 2(i), (50), (54), and (12) we get

$$\left\|x^{n+1} - x^*\right\|^2 \leq \|x^n - x^*\|^2 - \theta_n \|v^n\|^2 \leq \|x^n - x^*\|^2 - \frac{\theta_n}{(\lambda+\mu)^2} \left\|x^{n+1} - x^*\right\|^2, \tag{55}$$

and the conclusion of item (i) follows directly from (55).

(ii) By (11), $v^n = \gamma_n^{-1}(x^n - y^n + e^n)$. Thus

$$\lambda \|v^n\| \leq \lambda\gamma_n^{-1} \|x^n - y^n + e^n\| \leq \lambda\gamma_n^{-1} (\|x^n - y^n\| + \|e^n\|). \tag{56}$$

Combining (54) and (56)

$$\begin{aligned} \left\|x^{n+1} - x^*\right\| &\leq \lambda\gamma_n^{-1} (\|x^n - y^n\| + \|e^n\|) + \|e^n\| \\ &\leq \left[\nu + (1+\nu)\frac{\lambda}{\gamma_n}\right] \|x^n - y^n\| = \omega_n \|x^n - y^n\|, \end{aligned} \tag{57}$$

using (13) in the last inequality and (52) in the equality. By Lemma 2(ii), (52), and (57),

$$\left\|x^{n+1} - x^*\right\|^2 \leq \|x^n - x^*\|^2 - \xi \|x^n - y^n\|^2 \leq \|x^n - x^*\|^2 - \frac{\xi}{\omega_n^2} \left\|x^{n+1} - x^*\right\|^2, \tag{58}$$

and the conclusion of item (ii) follows immediately from (58). □

COROLLARY 4. *Under the assumptions of Theorem* 2*, the sequences generated by Algorithms* 1 *and* 2 *converge at least linearly, with asymptotic error constants given by* $\frac{\lambda+\mu}{\sqrt{(\lambda+\mu)^2+\bar{\theta}}}$, $\frac{\bar{\omega}}{\sqrt{\bar{\omega}^2+\xi}}$, *respectively, where* $\bar{\theta} = (1-\sigma)\hat{\gamma}(\hat{\gamma} - 2\rho)$, $\bar{\omega} = \nu + (1+\nu)\frac{\lambda}{\hat{\gamma}}$, *and superlinearly, when* $\lim_{n\to\infty} \gamma_n = \infty$.

Proof. Note that $\omega_n \leq \bar{\omega}$, $\theta_n \geq \bar{\theta}$ for all n. Thus (55), and consequently (49), hold with $\bar{\theta}$ instead of θ_n. By the same token, (58), and consequently (51), hold with $\bar{\omega}$ instead of ω_n, establishing the asymptotic error constants. The statement on superlinear convergence follows directly from Theorem 2, observing that $\lim_{n\to\infty} \gamma_n = \infty$ implies that $\lim_{n\to\infty} \theta_n = \infty$ and $\lim_{n\to\infty} \omega_n = 0$. □

Of course, a caveat is in order in connection with the result on superlinear convergence in Corollary 4, as is the case with similar results for other variants of the proximal point algorithm (e.g., [7]). Proximal procedures, in general, replace the inversion of T for a sequence of subproblems, each one of which demands inversion of $I + \gamma_n T$, or equivalently of $\gamma_n^{-1} I + T$. On one hand, when γ_n becomes very large, $\gamma_n^{-1} I + T$ gets very close to T, and thus an arbitrarily high convergence rate can be achieved by making γ_n go fast enough to ∞ (in the limit, for $\gamma_n = \infty$ the algorithm would find a zero of T in one iteration). But this high convergence rate is deceiving in the following sense. Replacement of the inversion of T by a sequence of subproblems is recommended basically when the inversion of T is hard, i.e., when T is somewhat ill-conditioned. The properties of T imply that $\gamma_n^{-1} I + T$ is instead

theoretically well-conditioned (for all $\gamma_n > 0$ when T is monotone, for $\gamma_n > 2\rho$ when T is ρ-hypomonotone), but, for very large γ_n, $\gamma_n^{-1} I + T$ becomes numerically almost as ill-conditioned as T, in which case the regularization properties of the algorithm are lost (think, e.g., of the case of linear and singular T). In fact, one of the main advantages of the proximal point algorithm with respect to other regularization schemes is that it works without requiring that the regularization coefficients go to ∞, i.e., for constant γ_n, for instance. In other words, when γ_n increases, a tradeoff takes place between the improvement in the convergence rate and the deterioration of the numerical stability.

6. Inexact proximal method of multipliers. Let X and Y be Hilbert spaces. For an arbitrary $S : X \to \mathcal{P}(X)$, a maximal monotone $T : Y \to \mathcal{P}(Y)$, and a $\mathcal{C}^2$ function $F : X \to Y$, we consider the problem of finding a solution to the inclusion

$$S(x) + \nabla F(x)^* T(F(x)) \ni 0, \tag{P}$$

where $\nabla F(x)^*$ is the adjoint of the Jacobian of F at a point $x \in X$. This provides a flexible model for various applications, and it has an associated duality theory that can be seen as a generalization of the traditional convex programming duality theory. Combining dualization with the proximal point algorithm leads to multiplier methods for a wide class of problems much like in Rockafellar [15], where convex programs were treated. In [13], this approach was extended to problems of the form (P), and multiplier methods for variational inequalities and nonlinear convex programs were obtained as special cases. The purpose of this section is to derive an inexact version of the proximal method of multipliers for (P).

We reproduce here those parts of the duality theory which are needed in what follows (see [13] for a full exposition). Denote

$$F_0(x) = S(x) + \nabla F(x)^* T(F(x))$$

so that (P) can be written as $F_0(x) \ni 0$. Define the *Lagrangian* $L : X \times Y \to \mathcal{P}(X \times Y)$ by

$$L(x, y) = (\nabla F(x)^* y, -F(x)) + S(x) \times T^{-1}(y),$$

and consider the *primal-dual problem*

$$L(x, y) \ni (0, 0). \tag{PD}$$

The mapping F_0 is related to L by

$$\begin{aligned} F_0(x) &= \{S(x) + \nabla F(x)^* y \mid \exists y \in Y : \ y \in T(F(x))\} \\ &= \left\{S(x) + \nabla F(x)^* y \;\middle|\; \exists y \in Y : \ 0 \in -F(x) + T^{-1}(y)\right\} \\ &= \{v \in X \mid \exists y \in Y : \ (v, 0) \in L(x, y)\}. \end{aligned} \tag{59}$$

The following is immediate.

LEMMA 3. *We have $F_0(x) \ni e$ if and only if there exists a y such that $L(x, y) \ni (e, 0)$.*

We will also need to reformulate Algorithm 2. Eliminating v^n, and denoting y^n by $\tilde{z}^n$ and x^n by z^n, we can write it as follows.

METHOD 1.

1. *Given z^n, find a $\tilde{z}^n$ such that*

$$\gamma_n T(\tilde{z}^n) + \tilde{z}^n - z^n \ni e^n$$

for some $e^n \in Z$ satisfying

$$\|e^n\| \leq \nu \|\tilde{z}^n - z^n\|.$$

2. *Set*

$$z^{n+1} = \tilde{z}^n - e^n,$$

$n = n + 1$, and go to 1.

In step 1, Z denotes a subspace of H. Corollary 3 can now be stated in the following equivalent form.

THEOREM 3. *Under the assumptions of Theorem* 1,

(a) *there exists an infinite sequence $\{z^n\} \subset H$, conforming to Method* 1, *and satisfying*

$$\begin{aligned} \tilde{z}^n &\in U, \\ e^n - \tilde{z}^n + z^n &\in \gamma_n V, \\ d(x^{n+1}, S^* \cap U) &\leq \varepsilon; \end{aligned}$$

(b) *any sequence as in* (a) *converges weakly to a point in $S^* \cap U$.*

If we apply Method 1 with $Z = X \times \{0\}$ to (PD), we get the following.

METHOD 2.

1. *Given $(x^n, y^n) \in X \times Y$, find a $(\tilde{x}^n, \tilde{y}^n) \in X \times Y$ such that*

$$\gamma_n L(\tilde{x}^n, \tilde{y}^n) + [(\tilde{x}^n, \tilde{y}^n) - (x^n, y^n)] \ni (e^n, 0)$$

for some e^n satisfying

$$\|e^n\| \leq \nu \|(\tilde{x}^n, \tilde{y}^n) - (x^n, y^n)\|.$$

2. *Set*

$$\begin{aligned} x^{n+1} &= \tilde{x}^n - e^n, \\ y^{n+1} &= \tilde{y}^n, \end{aligned}$$

$n = n + 1$, and go to 1.

The inclusion in step 1 can be written as

$$L_n(\tilde{x}^n, \tilde{y}^n) \ni \gamma_n^{-1}(e^n, 0), \tag{60}$$

where

$$\begin{aligned} L_n(x, y) &= L(x, y) + \gamma_n^{-1}(x - x^n, y - y^n) \\ &= (\nabla F(x)^* y, -F(x)) + [S(x) + \gamma_n^{-1}(x - x^n)] \times [T^{-1}(y) + \gamma_n^{-1}(y - y^n)] \\ &= (\nabla F(x)^* y, -F(x)) + S_n(x) \times T_n^{-1}(y), \end{aligned}$$

with $S_n(x) = S(x) + \gamma_n^{-1}(x - x^n)$, and $T_n(u) = (I + \gamma_n T^{-1})^{-1}(y^n + \gamma_n u)$. We thus get from (60) that $\tilde{y}^n \in T_n(F(\tilde{x}^n))$. But since T_n is single-valued by the maximal monotonicity of T, we see that the value of $\tilde{x}^n$ determines the value of $\tilde{y}^n$ uniquely by

$$\tilde{y}^n = T_n(F(\tilde{x}^n)).$$

Since L_n is in the format of the general duality framework, we have by Lemma 3 that $(\tilde{x}^n, \tilde{y}^n)$ satisfies (60) if and only if

$$S(\tilde{x}^n) + \gamma_n^{-1}(\tilde{x}^n - x^n) + \nabla F(\tilde{x}^n)^* T_n(F(\tilde{x}^n)) \ni \gamma_n^{-1} e^n$$
$$\tilde{y}^n = T_n(F(\tilde{x}^n)).$$

Method 2 can thus be written as follows.

METHOD 3 (proximal method of multipliers).

1. *Given $(x^n, y^n) \in X \times Y$, find an $\tilde{x}^n \in X$ such that*

$$S(\tilde{x}^n) + \gamma_n^{-1}(\tilde{x}^n - x^n) + \nabla F(\tilde{x}^n)^* T_n(F(\tilde{x}^n)) \ni \gamma_n^{-1} e^n$$

for some e^n satisfying

$$\|e^n\| \le \nu \|(\tilde{x}^n, T_n(F(\tilde{x}^n))) - (x^n, y^n)\|.$$

2. *Set*

$$x^{n+1} = \tilde{x}^n - e^n,$$
$$y^{n+1} = T_n(F(\tilde{x}^n)),$$

$n = n + 1$, and go to 1.

Theorem 3 can be herefore restated in the following form.

THEOREM 4. *Assume that L^{-1} is maximal ρ-hypomonotone in a subset $V \times U$, where $U \subset X \times Y$ and $V \subset X \times Y$ satisfy*

(i) *$S^* \cap U$ is nonempty and closed;*
(ii) *there exists $\beta > 0$ such that $B(0, \beta) \subset V$;*
(iii) *there exists $\delta > 0$ such that $(S^* \cap U) + B(0, \delta) \subset U$;*
(iv) *U is convex,*

where S^* is the set of zeroes of L. Take a sequence $\{\gamma_n\}$ of positive real numbers such that $2\rho < \hat{\gamma} = \inf\{\gamma_n\}$. Define $\varepsilon = \min\{\delta, \beta\gamma/2\}$. If $d(x^0, y^0, S^* \cap U) \le \varepsilon$, then

(a) *there exists an infinite sequence $\{(x^n, y^n)\} \subset X \times Y$, conforming to Method* 3*, such that for all n*

$$(\tilde{x}^n, T_n(F(\tilde{x}^n))) \in U,$$
$$(e^n, 0) - (\tilde{x}^n, T_n(F(\tilde{x}^n))) + (x^n, y^n) \in \gamma_n V,$$
$$d((x^{n+1}, y^{n+1}), S^* \cap U) \le \varepsilon;$$

(b) *any sequence as in* (a) *converges to a point in $S^* \cap U$.*

REFERENCES

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

[2] H. BREZIS, *Analyse Fonctionelle, Théorie et Applications*. Masson, Paris, 1983.

[3] R.S. BURACHIK, S. SCHEIMBERG, AND B.F. SVAITER, *Robustness of the hybrid extragradient proximal point algorithm*, J. Optim. Theory Appl., 111 (2001), pp. 117–136.

[4] R.S. BURACHIK AND B.F. SVAITER, *A relative error tolerance for a family of generalized proximal point methods*, Math. Oper. Res., 26 (2001), pp. 816–831.

[5] J. ECKSTEIN AND D. BERTSEKAS, *On the Douglas–Ratford splitting method and the proximal point algorithm for maximal monotone operators*, Math. Programming, 55 (1992), pp. 293–318.

[6] J. Eckstein and M. Ferris, *Smooth methods of multipliers for complementarity problems*, Math. Program., 86 (1999), pp. 65–90.
[7] A.N. Iusem and M. Teboulle, *On the convergence rate of entropic proximal optimization algorithms*, Comput. Appl. Math., 12 (1993), pp. 153–168.
[8] A. Kaplan and R. Tichatschke, *Proximal point methods and nonconvex optimization*, J. Global Optim., 13 (1998), pp. 389–406.
[9] M.A. Krasnoselskii, *Two observations about the method of successive approximations*, Uspekhi Mat. Nauk, 10 (1955), pp. 123–127.
[10] B. Martinet, *Régularisation d'inéquations variationelles par approximations successives*, Rev. Française Informat. Recherche Opérationnelle, 4 (1970), pp. 154–158.
[11] G. Minty, *A theorem on monotone sets in Hilbert spaces*, J. Math. Anal. Appl., 11 (1967), pp. 434–439.
[12] J. Moreau, *Proximité et dualité dans un espace hilbertien*, Bull. Soc. Math. France, 93 (1965), pp. 273–299.
[13] T. Pennanen, *Local convergence of the proximal point algorithm and multiplier methods without monotonicity*, Math. Oper. Res., 27 (2002), pp. 170–191.
[14] R.T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.
[15] R.T. Rockafellar, *Augmented Lagrangians and applications of the proximal point algorithm in convex programming*, Math. Oper. Res., 1 (1976), pp. 97–116.
[16] R.T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer, Berlin, 1998.
[17] M.V. Solodov and B.F. Svaiter, *A hybrid projection–proximal point algorithm*, J. Convex Anal., 6 (1999), pp. 59–70.
[18] M.V. Solodov and B.F. Svaiter, *A hybrid approximate extragradient-proximal point algorithm using the enlargement of a maximal monotone operator*, Set-Valued Anal., 7 (1999), pp. 323–345.
[19] M.V. Solodov and B.F. Svaiter, *An inexact hybrid generalized proximal point algorithms and some new results on the theory of Bregman functions*, Math. Oper. Res., 51 (2000), pp. 479–494.
[20] J.E. Spingarn, *Submonotone mappings and the proximal point algorithm*. Numer. Funct. Anal. Optim., 4 (1981), pp. 123–150.

SIAM J. OPTIM.
Vol. 13, No. 4, pp. 1098–1119

A SEQUENTIAL QUADRATICALLY CONSTRAINED QUADRATIC PROGRAMMING METHOD FOR DIFFERENTIABLE CONVEX MINIMIZATION*

MASAO FUKUSHIMA†, ZHI-QUAN LUO‡, AND PAUL TSENG§

Abstract. This paper presents a sequential quadratically constrained quadratic programming (SQCQP) method for solving smooth convex programs. The SQCQP method solves at each iteration a subproblem that involves convex quadratic inequality constraints as well as a convex quadratic objective function. Such a quadratically constrained quadratic programming problem can be formulated as a second-order cone program, which can be solved efficiently by using interior point methods. We consider the following three fundamental issues on the SQCQP method: the feasibility of subproblems, the global convergence, and the quadratic rate of convergence. In particular, we show that the Maratos effect is avoided without any modification to the search direction, even though we use an ordinary ℓ_1 exact penalty function as the line search merit function.

Key words. convex programming, quadratically constrained quadratic programming, Maratos effect, quadratic convergence, global convergence

AMS subject classifications. 90C25, 90C55, 65K05

PII. S1052623401398120

1. Introduction.

This paper presents a sequential quadratically constrained quadratic programming (SQCQP) method for solving smooth convex programs. Unlike sequential quadratic programming (SQP) methods [3, 4, 8, 14], the SQCQP method solves at each iteration a subproblem that involves convex quadratic inequality constraints and a convex quadratic objective function. Such a quadratically constrained quadratic programming problem can be formulated as a second-order cone program [12, 16] and be solved efficiently by using interior point methods [1, 12, 15, 16, 17, 26].

In this paper, we consider the following three fundamental issues concerning the SQCQP method: the feasibility of subproblems, the global convergence, and the quadratic rate of convergence. First, we note that a straightforward quadratic approximation of the constraints may yield an infeasible subproblem even though we are dealing with a convex program. This is in contrast with SQP methods, in which linearized constraints in a subproblem always admit a feasible solution as long as the original convex program is feasible. To maintain feasibility of the subproblem, we propose a strategy of switching between the linear and the quadratic approximation for each constraint, taking into account the constraint violation and curvature at the current iterate. It is shown that the proposed strategy always yields a feasible subproblem and that the quadratic approximation will eventually be used for every

*Received by the editors November 15, 2001; accepted for publication (in revised form) November 7, 2002; published electronically April 30, 2003. This research was supported by a Grant-in-Aid for Scientific Research (B) from the Ministry of Education, Science, Sports and Culture of Japan.
http://www.siam.org/journals/siopt/13-4/39812.html

†Department of Applied Mathematics and Physics, Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan (fuku@i.kyoto-u.ac.jp).

‡Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON, L8S 4L7, Canada (luozq@mcmaster.ca). This author was supported by a grant from NSERC and by the Canada Research Chair Program.

§Department of Mathematics, University of Washington, Seattle, WA 98195 (tseng@math.washington.edu). This author was supported by National Science Foundation grant CCR-9731273.

constraint. To ensure global convergence, we adopt a standard line search technique using the ℓ_1 penalty function. A major difference between the SQCQP method and the classical SQP method lies in that the former uses information up to second-order for both the constraints and the objective function in generating a search direction. As a result, the SQCQP method not only enjoys local quadratic convergence, but it is also free of the usual Maratos effect known to cause difficulty in an SQP method employing a nondifferentiable exact penalty function as the line search merit function [22]. More importantly, the elimination of the Maratos effect is achieved without performing any additional correction step in the SQCQP method as required in the SQP-type methods [8, 14]. Interestingly, Fukushima [8] had already noted earlier that the Maratos effect can be avoided by using quadratically constrained quadratic programming subproblems. However, due to the lack of efficient tools to solve such subproblems at the time, an auxiliary quadratic programming subproblem (thus the name SQP) was introduced in [8] to circumvent the Maratos effect. Earlier, Coleman and Conn [5] had considered similar quadratically constrained subproblems in motivating an exact penalty function method, rather than an SQP method, with global and local superlinear convergence properties. As with [8], the quadratically constrained subproblems are used to motivate the method but are not used in the method itself.

Related SQCQP-type methods have been considered by several authors. Panin [19] studied an SQCQP method and proved its global and superlinear convergence under the somewhat restrictive assumptions that the objective and constraint functions are strongly convex functions with uniformly bounded Hessians and that the sequence of Lagrange multiplier estimates generated by the method is uniformly bounded (see also [20] for a related work). Kruk and Wolkowicz [10, 11] cite two unpublished reports from 1984 and 1985, by Dixon, Hersom, and Maany and Maany, respectively, in which an SQCQP method was developed for some highly nonlinear orbital trajectory problems. Polak, Mayne, and Higgins [21] proposed, for semi-infinite minimax problems, an iterative method that solves at each iteration a minimax subproblem constructed from quadratic approximations of component functions. Global convergence and local superlinear convergence of order 1.5 for this method were shown under the assumption that the component functions are strongly convex with uniformly bounded Hessians. Wiest and Polak [27] proposed a phase I-phase II SQCQP method in which the quadratic objective and constraint functions involved in the subproblem have the identical Hessian which is a positive multiple of the identity matrix. Global convergence and local linear convergence for this method were shown under the Mangasarian–Fromovitz constraint qualification (MFCQ) plus the assumptions of second-order sufficiency and strict complementarity. More recently, Kruk and Wolkowicz [10, 11] proposed a trust-region SQCQP method for convex and nonconvex problems, with each subproblem solved (approximately) by semidefinite programming relaxation. Feasibility of subproblem is ensured by relaxing a homogenization constraint. The reference [11, section 7] gives some discussion of the Maratos effect, as well as global and local quadratic convergence. However, the use of semidefinite programming relaxation may restrict the applicability of the approach to small and medium size problems. As this paper was being written, we learned of a recent work of Anitescu [2] proposing a trust-region SQCQP method for solving (nonconvex) nonlinear minimization problems. Local superlinear convergence of order 1.5 for this method is shown under the MFCQ plus a mild quadratic growth condition.

We remark that our SQCQP method is applicable to general smooth convex

programs. For the special class of convex conic programs admitting an efficiently computable self-concordant barrier function, one can use the well-developed primal-dual interior point methods [16] which have attractive polynomial complexity. Such convex conic programs include linear programs, second-order cone programs, and semidefinite programs. In the absence of an easily computable self-concordant barrier function, interior point methods for smooth convex programming can still be developed by applying Newton's method to a perturbed Karush–Kuhn–Tucker (KKT) system [6, 23], although polynomial complexity is no longer assured. Global convergence and local superlinear convergence of such interior point methods have been established under Slater condition, strict complementarity, constant Hessian rank assumption, and second-order sufficiency [23, Assumptions 1–7]. In [6, Assumptions (A1)–(A5)], Slater condition and constant Hessian rank assumption are replaced by the linear independence constraint qualification (LICQ), which is a more stringent assumption for convex programs. In contrast, global and local quadratic convergence for our SQCQP method assume a Slater condition and uniform positive definiteness of the Lagrangian Hessian (Assumptions 1 and 2) but not strict complementarity. The practical performance of our SQCQP method, as compared with SQP methods and interior point methods for smooth convex programming, is a topic for further study. Our SQCQP method seems well suited for problems whose constraints are quadratic or admit good quadratic approximation. For a motivation of this, consider the special case of a quadratically constrained quadratic program. In this case, we can choose the subproblem to coincide with the original problem so that a single iteration suffices to solve the original problem. Moreover, a polynomial-time interior point method [1, 12, 15, 16, 17, 26] can be used to find a solution of relative accuracy $\epsilon > 0$ in time that is polynomial in n, m, and $\log(1/\epsilon)$. In contrast, SQP methods would successively linearize the quadratic constraints, possibly leading to inefficiency. The same is true with interior point methods for smooth convex programming [6, 23]. In particular, none of these methods would solve the problem in polynomial time.

Throughout this paper, $\Re^n$ denotes the set of n-dimensional column vectors, $\mathcal{S}^n$ denotes the set of $n \times n$ symmetric real matrices, and ^{T} denotes transpose. For $x \in \Re^n$, $\|x\| = \sqrt{x^T x}$. For A and B in $\mathcal{S}^n$, $A \succeq B$ if and only if $A - B \succeq O$ (the zero matrix); i.e., $A - B$ is positive semidefinite. Also, ":=" means "define."

2. Algorithm description. We are interested in solving the following convex minimization problem:

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & c_i(x) \le 0, \quad i = 1, \ldots, m, \end{array} \tag{2.1}$$

where $f : \Re^n \to \Re$ and $c_i : \Re^n \to \Re$, $i = 1, \ldots, m$, are twice continuously differentiable convex functions. For simplicity, we will confine ourselves to the inequality constrained problem, although the subsequent argument can be extended naturally to problems involving additional linear equality constraints. Throughout we make the following assumptions on problem (2.1).

ASSUMPTION 1.

(a) *Problem* (2.1) *has a nonempty optimal solution set.*

(b) *There exists a* $\bar{x} \in \Re^n$ *satisfying the Slater condition* $c_i(\bar{x}) < 0$, $i = 1, \ldots, m$.

Notice that, for the convex problem (2.1), the Slater condition is equivalent to

the MFCQ. Define the ℓ_1 exact penalty function $F_r : \Re^n \to \Re$ for problem (2.1) as

$$F_r(x) := f(x) + r \sum_{i=1}^{m} [c_i(x)]_+,$$

where $r > 0$ is a penalty parameter and $[\cdot]_+$ denotes the projection onto the set of nonnegative real numbers, i.e., $[\cdot]_+ = \max\{0, \cdot\}$. Under Assumption 1, for any $r > 0$ sufficiently large, the set of unconstrained minimizers of F_r coincides with the solution set of problem (2.1) [3, section 5.5].

At each iteration of the SQCQP method, given a current iterate $x \in \Re^n$, we solve the following subproblem in d:

$$\begin{array}{ll} \text{minimize} & g(x)^T d + \dfrac{1}{2} d^T B d \\ \text{subject to} & c_i(x) + g_i(x)^T d + \dfrac{\alpha_i}{2} d^T G_i(x) d \le 0, \quad i = 1, \ldots, m, \end{array} \tag{2.2}$$

where $\alpha_i \in [0, 1]$, $g(x) := \nabla f(x) \in \Re^n$, $g_i(x) := \nabla c_i(x) \in \Re^n$, $G_i(x) := \nabla^2 c_i(x) \in \mathcal{S}^n$, $i = 1, \ldots, m$, and $B \in \mathcal{S}^n$ is positive semidefinite. Subproblem (2.2) is a convex quadratically constrained quadratic program and, as such, can be formulated as a second-order cone program [12, 16][1] and be solved efficiently by using interior point methods [1, 12, 15, 16, 26].

The nonnegative parameters $(\alpha_i)_{i=1}^m$ are introduced to ensure the feasibility of the subproblem (2.2). Intuitively, if $\alpha_i = 1$, then the local quadratic approximation $c_i(x) + g_i(x)^T d + \frac{1}{2} d^T G_i(x) d$ of $c_i(x + d)$ may be too "aggressive" to admit a feasible solution for (2.2). As an example, for $m = n = 1$ and $c_1(x) = x^4 - \frac{1}{4}$, it is readily checked that (2.2) with $x = 1$ and $\alpha_1 = 1$ has no feasible solution. The less-than-1 weighting factor α_i relaxes the local quadratic cut in order to preserve feasibility. The lemma below quantifies those α_i that ensure the feasibility of (2.2). In particular, at every x, the linearized constraints admit a solution; and if x is nearly feasible or the constraint functions have small curvature at x, then quadratic constraints can be used. Here, each weight α_i is set to be either 0 or 1. We can alternatively set α_i to be the largest value in $[0, 1]$ for which (2.7) below holds (with t chosen as in (a) or (b) or (c) in Lemma 2.1). This makes the subproblems harder to solve (more quadratic constraints), but the number of subproblems solved may be fewer.

LEMMA 2.1. *Let $\bar{x}$ be the Slater point of Assumption* 1(b). *Fix any $\theta \in [0, 1)$ and $\vartheta \in (\theta, 1)$. For each $x \in \Re^n$, define $\kappa_i := (\bar{x} - x)^T G_i(x)(\bar{x} - x)$ for all i and*

$$\mathcal{J} := \{i \mid \theta c_i(\bar{x}) \le c_i(x)\}, \tag{2.3}$$

$$s_1 := \max_{i : c_i(x) > 0} \frac{c_i(x)}{c_i(x) - \vartheta c_i(\bar{x})}, \tag{2.4}$$

$$s_2 := \min \left\{ \min_{i \in \mathcal{J}} \frac{c_i(x) - \vartheta c_i(\bar{x})}{\kappa_i}, 1 \right\}, \tag{2.5}$$

$$s_3 := \min \left\{ s_2, \min_{i \notin \mathcal{J}} \frac{-2(\vartheta - \theta) c_i(\bar{x})}{\kappa_i} \right\}, \tag{2.6}$$

where s_1 is understood to be $-\infty$ if there is no i such that $c_i(x) > 0$, i.e., if x is a feasible solution of problem (2.1), *and the fractions in* (2.5) *and* (2.6) *are understood*

[1] In particular, as noted in [16, p. 221], the quadratic inequality $\|y\|^2 \le z$ can be rewritten as $(\|2y\|^2 + (1 - z)^2)^{1/2} \le 1 + z$ or, equivalently, $(1 + z, 1 - z, 2y)$ belongs to the second-order cone $\{(t, w) \mid \|w\| \le t\}$.

to be $+\infty$ *if* $\kappa_i = 0$. *Then the vector* $d := t(\bar{x} - x)$ *satisfies*

$$c_i(x) + g_i(x)^T d + \frac{\alpha_i}{2} d^T G_i(x) d \;\le\; t(1-\vartheta) c_i(\bar{x}), \quad i = 1, \ldots, m, \tag{2.7}$$

under any of the following choices of t *and* $\alpha_1, \ldots, \alpha_m$*:*

(a) *Set* $t = 1$ *and* $\alpha_i = 0$ *for all* $i = 1, \ldots, m$.

(b) *If* $2s_1 < s_2$, *set* $t = s_2$ *and* $\alpha_i = 1$ *for* $i \in \mathcal{J}$ *and* $\alpha_i = 0$ *for* $i \notin \mathcal{J}$.

(c) *If* $2s_1 < s_3$, *set* $t = s_3$ *and* $\alpha_i = 1$ *for all* $i = 1, \ldots, m$.

Proof. From (2.5) and (2.6), we see that $0 < s_3 \le s_2 \le 1$ so that $0 < t \le 1$ for any choice of t in (a) or (b) or (c). The convexity of functions c_i, $i = 1, \ldots, m$, then yields

$$\begin{aligned} c_i(x) + g_i(x)^T d + \frac{\alpha_i}{2} d^T G_i(x) d &= (1-t) c_i(x) + t(c_i(x) + g_i(x)^T(\bar{x} - x)) + t^2 \frac{\alpha_i \kappa_i}{2} \\ &\le (1-t)\eta_i + t\bar{\eta}_i + t^2 \frac{\alpha_i \kappa_i}{2}, \end{aligned}$$

where for simplicity we denote $\eta_i := c_i(x)$ and $\bar{\eta}_i := c_i(\bar{x})$. Let $R_i(t, \alpha_i)$ denote the right-hand side quadratic polynomial in the above inequality. We will show that $R_i(t, \alpha_i) \le t(1-\vartheta)\bar{\eta}_i$ for any of the choices of t and α_i's in (a)–(c). Clearly, this, together with the above inequality, immediately proves the lemma.

We need to first develop two inequalities. Fix some $i \in \mathcal{J}$. It follows from (2.3) that $\eta_i - \vartheta\bar{\eta}_i > \eta_i - \theta\bar{\eta}_i > 0$. Moreover, we have

$$\begin{aligned} R_i(t, \alpha_i) - t(1-\vartheta)\bar{\eta}_i &= (1-t)\eta_i + t\bar{\eta}_i + t^2 \frac{\alpha_i \kappa_i}{2} - t(1-\vartheta)\bar{\eta}_i \\ &= \eta_i - t(\eta_i - \vartheta\bar{\eta}_i) + t^2 \frac{\alpha_i \kappa_i}{2} \\ &= (\eta_i - \vartheta\bar{\eta}_i)\left[\frac{\eta_i}{\eta_i - \vartheta\bar{\eta}_i} - t + t^2 \frac{\alpha_i \kappa_i}{2(\eta_i - \vartheta\bar{\eta}_i)}\right] \\ &\le (\eta_i - \vartheta\bar{\eta}_i)\left[s_1 - t + t^2 \frac{\alpha_i}{2 s_2}\right] \quad \forall i \in \mathcal{J}, \end{aligned} \tag{2.8}$$

where the last step follows from (2.4) and (2.5). Next we consider any $i \notin \mathcal{J}$. By (2.3) we have $\eta_i < \theta\bar{\eta}_i$. Then, for $t \in (0, 1]$, we have

$$\begin{aligned} R_i(t, \alpha_i) - t(1-\vartheta)\bar{\eta}_i &= (1-t)\eta_i + t\bar{\eta}_i + t^2 \frac{\alpha_i \kappa_i}{2} - t(1-\vartheta)\bar{\eta}_i \\ &\le (1-t)\theta\bar{\eta}_i + t\bar{\eta}_i + t^2 \frac{\alpha_i \kappa_i}{2} - t(1-\vartheta)\bar{\eta}_i \\ &\le -t\theta\bar{\eta}_i + t\bar{\eta}_i + t^2 \frac{\alpha_i \kappa_i}{2} - t(1-\vartheta)\bar{\eta}_i \\ &= t(\vartheta - \theta)\bar{\eta}_i + t^2 \frac{\alpha_i \kappa_i}{2} \\ &\le -(\vartheta - \theta)\bar{\eta}_i \left[-t + t^2 \frac{\alpha_i}{s_3}\right] \quad \forall i \notin \mathcal{J}, \end{aligned} \tag{2.9}$$

where the last step is due to (2.6). We now consider the three choices of t and α_i's in (a)–(c).

(a) By (2.3) and (2.4), we have $s_1 < 1$. Setting $t = 1$ and $\alpha_i = 0$ in both (2.8) and (2.9), we see that $R_i(1, 0) - (1-\vartheta)\bar{\eta}_i < 0$ for all i.

(b) Suppose $2s_1 < s_2$, and let $t = s_2 \in (0, 1]$ and $\alpha_i = 1$ for $i \in \mathcal{J}$ and $\alpha_i = 0$ for $i \notin \mathcal{J}$. Then we substitute these choices of t and α_i's into the inequality (2.8) to

obtain

$$R_i(s_2,1)-s_2(1-\vartheta)\bar{\eta}_i \le (\eta_i-\vartheta\bar{\eta}_i)\left[s_1 - s_2 + s_2^2\frac{1}{2s_2}\right] = (\eta_i-\vartheta\bar{\eta}_i)\left[s_1 - \frac{s_2}{2}\right] < 0 \ \ \forall i \in \mathcal{J},$$

where we used $2s_1 < s_2$, and into the inequality (2.9) to obtain

$$R_i(s_2,0) - s_2(1-\vartheta)\bar{\eta}_i \le -(\vartheta-\theta)\bar{\eta}_i\,[-s_2] < 0 \ \ \forall i \notin \mathcal{J}.$$

(c) Suppose $2s_1 < s_3$, and let $t = s_3 \in (0,1]$ and $\alpha_i = 1$ for all i. Then we substitute these choices of t and α_i's into the inequality (2.8) to obtain

$$R_i(s_3,1)-s_3(1-\vartheta)\bar{\eta}_i \le (\eta_i-\vartheta\bar{\eta}_i)\left[s_1 - s_3 + s_3^2\frac{1}{2s_2}\right] \le (\eta_i-\vartheta\bar{\eta}_i)\left[s_1 - \frac{s_3}{2}\right] < 0 \ \ \forall i \in \mathcal{J},$$

where we used $s_3 \le s_2$ and $2s_1 < s_3$, and into the inequality (2.9) to obtain

$$R_i(s_3,1) - s_3(1-\vartheta)\bar{\eta}_i \le -(\vartheta-\theta)\bar{\eta}_i\,[-s_3+s_3] = 0 \ \ \forall i \notin \mathcal{J}.$$

This completes the proof of the lemma. □

We remark that Lemma 2.1 offers a specific way of picking the parameters α_i to ensure the feasibility of subproblem (2.2). In particular, given an iterate x, we compute the values of s_1, s_2 and s_3 by (2.4)–(2.6) and determine α_i as follows:

$$\begin{array}{lll} s_2 \le 2s_1 & \Longrightarrow & \alpha_i = 0, \ i = 1,\ldots,m, \\ s_3 \le 2s_1 < s_2 & \Longrightarrow & \alpha_i = \left\{ \begin{array}{ll} 1, & i \in \mathcal{J}, \\ 0, & i \notin \mathcal{J}, \end{array} \right. \\ s_3 > 2s_1 & \Longrightarrow & \alpha_i = 1, \ i = 1,\ldots,m. \end{array} \tag{2.10}$$

Then, by Lemma 2.1, subproblem (2.2) is feasible at every iteration with α_i chosen according to the rule (2.10). In general, linear constraints with $\alpha_i = 0$ may have to be used (case (a)) in the early stage of the SQCQP method in order to maintain subproblem feasibility. However, as we will argue later (see Lemma 4.1), quadratic constraints with $\alpha_i = 1$ (case (c)) can eventually be adopted when the iterates get close to the optimal solution. The intermediate case of mixed linear and quadratic constraints (case (b)) provides a mechanism for quadratic constraints to be gracefully phased into the SQCQP method, thus potentially improving the convergence of this method in practice. However, we can alternatively set $\alpha_i = 0$ for *all* i in this intermediate case without affecting the subsequent theoretical convergence analysis of the SQCQP method.

Given its feasibility, subproblem (2.2) has an optimal solution if the objective function is bounded from below over the feasible region; see [13]. This condition is obviously satisfied if either B or at least one of $G_i(x)$, $i = 1,\ldots,m$, is positive definite. In our global convergence analysis (Theorem 4.7), we will essentially assume that B is positive definite.

Let d be an optimal solution of subproblem (2.2). Then the next iterate x^{new} is generated by

$$x^{new} := x + \beta d,$$

where $\beta \in \Re$ is the step size. To ensure global convergence, we choose $\beta > 0$ small enough to satisfy

$$F_r(x+\beta d) - F_r(x) \le \sigma\beta\left(\bar{F}_r(x,d,\alpha) - F_r(x)\right), \tag{2.11}$$

where $\sigma \in (0,1)$ is a user chosen constant, $\alpha = (\alpha_i)_{i=1}^m$, and

$$\begin{aligned}\bar{F}_r(x,d,\alpha) := f(x) + g(x)^T d + \frac{1}{2} d^T G(x) d \\ + r\sum_{i=1}^m \left[c_i(x) + g_i(x)^T d + \frac{\alpha_i}{2} d^T G_i(x) d\right]_+,\end{aligned} \tag{2.12}$$

with $G(x) := \nabla^2 f(x)$. Notice that the weights $\alpha_1, \ldots, \alpha_m$ given by (2.10) depend on the current iterate x as well as the Slater point $\bar{x}$. Since $c_i(x) + g_i(x)^T d + \frac{1}{2}\alpha_i d^T G_i(x) d \le 0$, we have

$$\bar{F}_r(x,d,\alpha) - F_r(x) = g(x)^T d + \frac{1}{2} d^T G(x) d - r\sum_{i=1}^m [c_i(x)]_+. \tag{2.13}$$

By (2.7), subproblem (2.2) has a feasible solution satisfying the Slater condition. Then it is known [25, Theorem 28.2] that there exists a vector of Lagrange multipliers $v = (v_1, \ldots, v_m)^T$ satisfying the KKT conditions for subproblem (2.2), namely,

$$\begin{aligned} &Bd + g(x) + \sum_{i=1}^m v_i(\alpha_i G_i(x) d + g_i(x)) = 0, \\ &c_i(x) + g_i(x)^T d + \frac{\alpha_i}{2} d^T G_i(x) d \le 0, \quad v_i \ge 0, \\ &v_i\Big(c_i(x) + g_i(x)^T d + \frac{\alpha_i}{2} d^T G_i(x) d\Big) = 0, \quad i = 1, \ldots, m. \end{aligned} \tag{2.14}$$

We call (d, v) a *KKT pair* of subproblem (2.2).

We state our method formally below.

SQCQP Method.

Step 0. Choose an initial point $x^1 \in \Re^n$, an initial penalty parameter $r^0 \in (0, +\infty)$, constants $\gamma \in (0,1)$, $\delta \in (0, +\infty)$, $\theta \in [0,1)$, $\vartheta \in (\theta, 1)$ and $\sigma \in (0,1)$, and a point $\bar{x} \in \Re^n$ such that $c_i(\bar{x}) < 0$, $i = 1, \ldots, m$. Initialize $k = 1$.

Step 1. Choose a positive semidefinite matrix $B^k \in \mathcal{S}^n$. Determine $\alpha^k = (\alpha_i^k)_{i=1}^m$ according to the rule (2.10) with $x = x^k$ and solve subproblem (2.2) associated with x^k, B^k, α^k to obtain a KKT pair $(d^k, v^k) \in \Re^n \times \Re^m$. If $d^k = 0$, terminate.

Step 2. Update the penalty parameter as

$$r^k := \begin{cases} r^{k-1} & \text{if } r^{k-1} \ge \max_i v_i^k + \delta; \\ \max_i v_i^k + 2\delta & \text{otherwise.} \end{cases} \tag{2.15}$$

Let β^k be the largest $\beta \in \{1, \gamma, (\gamma)^2, (\gamma)^3, \ldots\}$ satisfying

$$F_{r^k}(x^k + \beta d^k) - F_{r^k}(x^k) \le \sigma\beta\left(\bar{F}_{r^k}(x^k, d^k, \alpha^k) - F_{r^k}(x^k)\right). \tag{2.16}$$

Update $x^{k+1} := x^k + \beta^k d^k$, increment k by 1, and go to Step 1.

The updating rule (2.15) maintains the penalty parameter r^k large enough to guarantee d^k to be a descent direction for the function F_{r^k}, while allowing for r^k to eventually stay constant as the iterates x^k converge to a solution; see Lemma 3.1 below.

The termination criterion in Step 1 is justified by the following lemma.

Lemma 2.2. *For any $x \in \Re^n$ and $\alpha_i \ge 0$, $i = 1, \ldots, m$, suppose that subproblem* (2.2) *has a KKT pair (d, v). If $d = 0$, then (x, v) is a KKT pair for problem* (2.1)*, and hence x is an optimal solution of problem* (2.1).

Proof. This can easily be verified by substituting $d = 0$ in (2.14). □

To execute our SQCQP method, it is necessary to find a point $\bar{x}$ satisfying the Slater condition. This can be accomplished in various ways. For example, we can use the following simple iterative procedure:

Step 0. Choose an initial point $x^1 \in \Re^n$ and a sequence $\{\epsilon_k\}$ of positive numbers. Set $k := 1$.

Step 1. If $c_i(x^k) < 0$, $i = 1, \ldots, m$, then exit. Otherwise, solve the convex quadratic program

$$\min\{\|p\|^2 \mid c_i(x^k) + g_i(x^k)^T p + \epsilon_k \leq 0,\ i = 1, \ldots, m\}, \tag{2.17}$$

and let p^k be the optimal solution, if it exists.

Step 2. If (2.17) is infeasible, then put $x^{k+1} := x^k$. Otherwise, put $x^{k+1} := x^k + p^k$. Set $k := k + 1$ and go to Step 1.

We note that the subproblem (2.17) has a unique optimal solution as long as it is feasible. Assuming the existence of a Slater point, it can be shown [7] that the above procedure indeed produces a Slater point in a finite number of iterations, provided that $\{\epsilon_k\}$ is chosen to be a strictly decreasing sequence that tends to zero at a rate slower than any linearly convergent sequence. A possible choice of $\{\epsilon_k\}$ would be to let $\epsilon_k = \epsilon/k$ for all k, where ϵ is a positive constant. Alternatively, we can solve approximately the minimax problem

$$\text{minimize} \quad \max_{i=1,\ldots,m} c_i(x),$$

for which many methods are available. In particular, it suffices to find a point x with negative objective value, which is easier than solving the minimax problem to optimality. We can also reformulate the minimax problem as

$$\text{minimize} \quad \xi \quad \text{subject to} \quad c_i(x) - \xi \leq 0, \quad i = 1, \ldots, m,$$

which is a special case of (2.1) with readily known Slater points. Our SQCQP method can thus be applied to this "phase I" problem, terminating whenever a point (x, ξ) with $c_i(x) < 0$, $i = 1, \ldots, m$, is found. Finite termination can be shown by assuming the existence of a Slater point and the boundedness of the feasible set for (2.1). The resulting two-phase SQCQP method is reminiscent of the two-phase simplex method for linear programming, with strict feasibility replacing feasibilty as the goal of phase I.

As an alternative to finding a Slater point a priori, it suffices that we have estimates $\bar{c}$ and $\bar{\kappa}$ satisfying $\max_{i=1,\ldots,m} c_i(\bar{x}) \leq \bar{c} < 0$ and $\bar{\kappa} \geq \max_{i=1,\ldots,m}(\bar{x} - x^k)^T G_i(x^k)(\bar{x} - x^k)$ for all k for some (unknown) Slater point $\bar{x}$. Then, in determining α^k, we can replace $c_i(\bar{x})$ and κ_i in (2.4)–(2.6) by $\bar{c}$ and $\bar{\kappa}$, respectively. It can be verified that this replacement does not affect the global and local convergence properties of our SQCQP method. If such estimates $\bar{c}$ and $\bar{\kappa}$ are not known a priori, we can make an initial guess of them and use an infeasibility-detecting method to solve the subproblem (2.2) associated with x^k, B^k, α^k. In particular, there are methods that, for a given tolerance $\epsilon > 0$, can detect in a finite number of iterations whether the subproblem has a strictly feasible solution with slacks of at least ϵ [24, Theorem 3.1]. If it is detected that no such strictly feasible solution exists, we increase $\bar{c}$ and $\bar{\kappa}$ by constant factors, adjust α^k accordingly, and resolve the subproblem. The number of times in which $\bar{c}$ and $\bar{\kappa}$ are increased is finite, provided ϵ is taken sufficiently small.

3. Global convergence. In this section, we analyze the global convergence of the SQCQP method. We begin with the following lemma giving conditions for an optimal solution d of (2.2) to be a descent direction for the penalty function F_r at x.

LEMMA 3.1. *For any $x \in \Re^n$ and $\alpha = (\alpha_i)_{i=1}^m \geq 0$, suppose that subproblem (2.2) has a KKT pair (d, v). Also, suppose there exists a $\mu > 0$ such that*

$$2B - G(x) + \sum_{i=1}^m \alpha_i v_i G_i(x) \succeq \mu I \tag{3.1}$$

and $r \geq \max_i v_i$. Then

$$\bar{F}_r(x, d, \alpha) - F_r(x) \leq -\frac{1}{2}\mu \|d\|^2. \tag{3.2}$$

Proof. We have from (2.14) that

$$\begin{aligned} g(x)^T d + \frac{1}{2} d^T G(x) d &= -\left(Bd + \sum_{i=1}^m v_i(\alpha_i G_i(x)d + g_i(x)) \right)^T d + \frac{1}{2} d^T G(x) d \\ &= -d^T \left(B + \sum_{i=1}^m \alpha_i v_i G_i(x) - \frac{1}{2} G(x) \right) d - \sum_{i=1}^m v_i g_i(x)^T d \\ &= -\frac{1}{2} d^T \left(2B + \sum_{i=1}^m \alpha_i v_i G_i(x) - G(x) \right) d + v^T c(x), \end{aligned} \tag{3.3}$$

where $c(x) = (c_1(x), \ldots, c_m(x))^T$. Then, from (2.13) we obtain

$$\begin{aligned} \bar{F}_r(x, d, \alpha) - F_r(x) &= g(x)^T d + \frac{1}{2} d^T G(x) d - r \sum_{i=1}^m [c_i(x)]_+ \\ &= -\frac{1}{2} d^T \left(2B - G(x) + \sum_{i=1}^m \alpha_i v_i G_i(x) \right) d + v^T c(x) - r \sum_{i=1}^m [c_i(x)]_+ \\ &\leq -\frac{1}{2}\mu \|d\|^2 - \sum_{i=1}^m (r - v_i)[c_i(x)]_+ \\ &\leq -\frac{1}{2}\mu \|d\|^2, \end{aligned}$$

where the second equality follows from (3.3), the first inequality is due to (3.1) and $v^T c(x) \leq \sum_i v_i [c_i(x)]_+$, and the last inequality is due to $r \geq \max_i v_i$. ∎

Since each $G_i(x)$ is positive semidefinite and v_i and α_i are nonnegative, condition (3.1) is satisfied if B is chosen to satisfy

$$2B \succeq G(x) + \mu I$$

for some constant $\mu > 0$.

The next lemma shows that the line search for the penalty function F_r at x along the direction d is well defined.

LEMMA 3.2. *Suppose that the conditions in Lemma 3.1 are satisfied. If $d \neq 0$, then the descent condition (2.11) holds for all $\beta > 0$ small enough.*

Proof. Since each $G_i(x)$ is positive semidefinite, then

$$c_i(x) + g_i(x)^T d \;\leq\; c_i(x) + g_i(x)^T d + \frac{\alpha_i}{2} d^T G_i(x) d \;\leq\; 0, \qquad i = 1, \ldots, m.$$

Thus, for any $\beta \in [0,1]$, we have

$$c_i(x) + \beta g_i(x)^T d \le (1-\beta)c_i(x),$$

implying that

$$[c_i(x) + \beta g_i(x)^T d]_+ \le (1-\beta)[c_i(x)]_+, \qquad i = 1, \ldots, m.$$

As a result, we obtain

$$\sum_{i=1}^m \left([c_i(x) + \beta g_i(x)^T d]_+ - [c_i(x)]_+\right) \le -\beta \sum_{i=1}^m [c_i(x)]_+. \tag{3.4}$$

Then, using the Taylor series approximation of $f(x+\beta d)$ and $c_i(x+\beta d)$ at x, we have

$$\begin{aligned}
&F_r(x+\beta d) - F_r(x)\\
&= \beta g(x)^T d + \frac{\beta^2}{2} d^T G(x) d + o(\beta^2\|d\|^2)\\
&\quad + r\sum_{i=1}^m \left[c_i(x) + \beta g_i(x)^T d + O(\beta^2\|d\|^2)\right]_+ - r\sum_{i=1}^m [c_i(x)]_+\\
&= \beta g(x)^T d + \frac{\beta^2}{2} d^T G(x) d + r\sum_{i=1}^m \left([c_i(x) + \beta g_i(x)^T d]_+ - [c_i(x)]_+\right) + O(\beta^2\|d\|^2)\\
&\le \beta g(x)^T d + \frac{\beta^2}{2} d^T G(x) d - \beta r\sum_{i=1}^m [c_i(x)]_+ + O(\beta^2\|d\|^2),
\end{aligned}$$

where the second equality uses the Lipschitzian property of $[\cdot]_+$ and the inequality uses (3.4). Since $G(x)$ is positive semidefinite and $0 < \beta \le 1$, the above relation implies that

$$\begin{aligned}
F_r(x+\beta d) - F_r(x) &\le \beta\left(g(x)^T d + \frac{1}{2} d^T G(x) d - r\sum_{i=1}^m [c_i(x)]_+\right) + O(\beta^2\|d\|^2)\\
&= \beta\Big(\bar{F}_r(x,d,\alpha) - F_r(x)\Big) + O(\beta^2\|d\|^2),
\end{aligned}$$

where the equality follows from (2.13). Then by (3.2) and $\sigma \in (0,1)$, we have

$$F_r(x+\beta d) - F_r(x) \le \sigma\beta\left(\bar{F}_r(x,d,\alpha) - F_r(x)\right)$$

for all sufficiently small $\beta > 0$. □

The next lemma shows that a KKT pair (d,v) obtained by solving subproblem (2.2) is bounded when (x,B) lies in a bounded set.

LEMMA 3.3. *Let X be any nonempty bounded subset of $\Re^n$. Let $\Omega := \{B \in \mathcal{S}^n \mid \mu_1 I \succeq B \succeq \mu_2 I\}$ for some constants $\mu_1 \ge \mu_2 > 0$. Then there exists a bounded subset $D \times V$ of $\Re^n \times \Re^m$ such that, for any $(x,B) \in X \times \Omega$ and any KKT pair (d,v) of subproblem (2.2) with $(\alpha_i)_{i=1}^m$ given by (2.10), we have $(d,v) \in D \times V$.*

Proof. First we show the boundedness of d. Fix any $x \in X$ and $B \in \Omega$. Since $(\alpha_i)_{i=1}^m$ is given by (2.10), we see from Lemma 2.1 that there exists a $t \in (0,1]$ such

that $t(\bar{x}-x)$ is a feasible solution of subproblem (2.2). So d satisfies

$$\begin{aligned}-\|g(x)\|\|d\|+\frac{\mu_2}{2}\|d\|^2 &\le g(x)^T d+\frac{1}{2}d^T Bd\\ &\le tg(x)^T(\bar{x}-x)+\frac{t^2}{2}(\bar{x}-x)^T B(\bar{x}-x)\\ &\le \|g(x)\|\,\|\bar{x}-x\|+\frac{\mu_1}{2}\|\bar{x}-x\|^2,\end{aligned}$$

where the first and last inequalities follow from $B\in\Omega$ and $t\in(0,1]$. Notice that the right-hand side of this inequality, as well as $\|g(x)\|$ on the left-hand side, are bounded for $x\in X$. This implies that d is bounded for $x\in X$. Also, $B\in\Omega$ is bounded since B is symmetric and its eigenvalues are bounded.

Next we show the boundedness of v. Suppose to the contrary that there exists a sequence of KKT pair $\{(d^k,v^k)\}$ of subproblem (2.2) associated with some $(x^k,B^k)\in X\times\Omega$ and $(\alpha_i^k)_{i=1}^m$ given by (2.10) for x^k, $k=1,2,\ldots$, such that $\|v^k\|\to+\infty$. By passing to a subsequence if necessary, we may assume that (x^k,B^k,d^k) converges to some $(x^\infty,B^\infty,d^\infty)$, and $v^k/\|v^k\|\to w$ with $\|w\|=1$ and $w\ge 0$. By the KKT conditions (2.14), we have for each k that

$$\begin{aligned}&B^k d^k+g(x^k)+\sum_{i=1}^m v_i^k\nabla q_i^k(d^k)=0,\\ &q_i^k(d^k)\le 0,\quad v_i^k\ge 0,\quad v_i^k q_i^k(d^k)=0,\qquad i=1,\ldots,m,\end{aligned}\tag{3.5}$$

where $q_i^k:\Re^n\to\Re$ are quadratic functions defined by

$$q_i^k(d):=c_i(x^k)+g_i(x^k)^T d+\frac{\alpha_i^k}{2}d^T G_i(x^k)d.$$

By further passing to a subsequence if necessary, we may assume that α_i^k converges to some $\alpha_i^\infty\in\{0,1\}$ for each i. Define the quadratic functions $q_i^\infty:\Re^n\to\Re$, $i=1,\ldots,m$, by

$$q_i^\infty(d)=c_i(x^\infty)+g_i(x^\infty)^T d+\frac{\alpha_i^\infty}{2}d^T G_i(x^\infty)d.$$

Then, upon dividing both sides of the first equality in (3.5) by $\|v^k\|$ and taking limit, we obtain

$$\sum_{i=1}^m w_i\nabla q_i^\infty(d^\infty)=0.\tag{3.6}$$

Moreover, the rest of (3.5) implies that $q_i^\infty(d^\infty)\le 0$, $i=1,\ldots,m$, and

$$w_i>0\quad\Longrightarrow\quad q_i^\infty(d^\infty)=0.\tag{3.7}$$

Note that there is at least one index i such that $w_i>0$.

By Lemma 2.1, for each k, there exists $t^k\in\{1,s_2^k,s_3^k\}$ such that $\hat{d}^k:=t^k(\bar{x}-x^k)$ satisfies

$$q_i^k(\hat{d}^k)\le t^k(1-\vartheta)c_i(\bar{x}),\quad i=1,\ldots,m,$$

where s_2^k, s_3^k are given by (2.3), (2.5), (2.6) with $x = x^k$. It is readily seen that

$$s_2^k \geq s_3^k \geq \min\left\{\min_{i=1,\ldots,m} \frac{-(\vartheta-\theta)c_i(\bar{x})}{\kappa_i^k}, 1\right\},$$

where $\kappa_i^k := (\bar{x}-x^k)^T G_i(x^k)(\bar{x}-x^k)$ for all i. Since $\{x^k\}$ converges, $\{\kappa_i^k\}$ also converges for all i, so the above inequalities show that $\{s_2^k\}$ and $\{s_3^k\}$ are uniformly bounded away from zero. This in turn implies that $\{t^k\}$ is uniformly bounded away from zero. Then any accumulation point $(\hat{d}^\infty, t^\infty)$ of $\{(\hat{d}^k, t^k)\}$ satisfies

$$q_i^\infty(\hat{d}^\infty) \leq t^\infty(1-\vartheta)c_i(\bar{x}) < 0, \quad i = 1, \ldots, m.$$

Then, for every i such that $w_i > 0$, (3.7) implies that $q_i^\infty(d^\infty) = 0$ so that the convexity of q_i^∞ yields

$$\nabla q_i^\infty(d^\infty)^T(\hat{d}^\infty - d^\infty) \leq q_i^\infty(\hat{d}^\infty) - q_i^\infty(d^\infty) < 0.$$

Since $w \geq 0$ and there is at least one index i such that $w_i > 0$, this implies that

$$\sum_{i=1}^m w_i \nabla q_i^\infty(d^\infty)^T(\hat{d}^\infty - d^\infty) < 0.$$

This contradicts (3.6), thus completing the proof. □

It is well known that, for a convex program, the Slater condition implies the boundedness of the Lagrange multipliers. Thus, Lemma 3.3 can be viewed as a generalization of this result to the uniform boundedness of the multipliers for all the subproblems which are convex and have nonempty interiors. Now we are ready to establish a global convergence result for the SQCQP method.

THEOREM 3.4. *Let $\{x^k\}, \{B^k\}, \{\alpha^k\}, \{v^k\}$ be generated by the SQCQP method. Suppose that $\{x^k\}$ is bounded. Also, suppose that*

$$2B^k - G(x^k) + \sum_{i=1}^m \alpha_i^k v_i^k G_i(x^k) \succeq \mu I \quad \text{and} \quad \mu_1 I \succeq B^k \succeq \mu_2 I$$

for all k, where $\mu > 0$ and $\mu_1 \geq \mu_2 > 0$ are some constants. Then the SQCQP method either terminates at an optimal solution of problem (2.1) *or generates an infinite sequence $\{x^k\}$ of which every accumulation point is an optimal solution of problem* (2.1).

Proof. For each k, by Lemma 2.1 and the choice of α^k, subproblem (2.1) associated with x^k, B^k, α^k is feasible and, by B^k being positive definite, has a unique optimal solution d^k. Since $r^k \geq \max_i v_i^k$ by (2.15), Lemma 3.2 implies that β^k in the line search is well defined. If the method terminates at some iteration k, then Lemma 2.2 implies that the iterate x^k is an optimal solution of problem (2.1). Suppose that the method generates an infinite sequence $\{x^k\}$. Since $\{x^k\}$ is assumed to be bounded and $\mu_1 I \succeq B^k \succeq \mu_2 I$ for all k, Lemma 3.3 implies that $\{(d^k, v^k)\}$ is also bounded, which in particular implies that the penalty parameter r^k is constant for all k sufficiently large. Moreover, we have from Lemma 3.1 that

$$\bar{F}_{r^k}(x^k, d^k, \alpha^k) - F_{r^k}(x^k) \leq -\frac{1}{2}\mu\|d^k\|^2 \tag{3.8}$$

for all k.

We consider two cases: (i) The sequence of step sizes $\{\beta^k\}$ is bounded away from zero, say by $\hat{\beta} > 0$; and (ii) $\{\beta^k\}$ contains a subsequence converging to zero.

In case (i), since (2.16) holds with $\beta = \beta^k$, (3.8) yields

$$F_{r^k}(x^k + \beta^k d^k) - F_{r^k}(x^k) \leq -\frac{1}{2}\sigma\hat{\beta}\mu\|d^k\|^2$$

for all k. Since $\{x^k\}$ is bounded so that $\{F_{r^k}(x^k)\}$ is bounded from below, this inequality implies that $\{d^k\}$ tends to 0. Since (d^k, v^k) satisfies the KKT conditions (2.14) for subproblem (2.2) associated with x^k, B^k, α^k, by taking the limit as $k \to \infty$ and using the boundedness of $\{x^k\}$, $\{B^k\}$, $\{v^k\}$ and $\{\alpha^k\}$, we see that any accumulation point (x^*, v^*) of $\{(x^k, v^k)\}$ satisfies the KKT conditions for problem (2.1). Since (2.1) is a convex program, x^* is an optimal solution of (2.1).

In case (ii), there exists a subsequence $\{\beta^k\}_{k\in K}$ converging to 0, where $K \subset \{0, 1, \ldots\}$. Since $\{(x^k, d^k, \alpha^k)\}$ is bounded and $\{r^k\}$ has a constant tail, by further passing to a subsequence if necessary, we can assume that $\{(x^k, d^k, \alpha^k)\}_{k\in K}$ converges to some limit (x^*, d^*, α^*) and $r^k = r$, $\beta^k < 1$ for all $k \in K$, where $r > 0$ is a constant. For each $k \in K$, since $\beta^k < 1$ and $r^k = r$, the Armijo-type line search rule in Step 2 of the SQCQP method implies that

$$F_r(x^k + (\beta^k/\gamma)d^k) - F_r(x^k) > \sigma(\beta^k/\gamma)\left(\bar{F}_r(x^k, d^k, \alpha^k) - F_r(x^k)\right). \tag{3.9}$$

Using the simple property $[u_1 + u_2]_+ - [u_1]_+ \leq [u_2]_+$ for any real numbers u_1, u_2, we deduce

$$\begin{aligned}[c_i(x + \beta d)]_+ - [c_i(x)]_+ &= [c_i(x) + \beta\nabla c_i(x)^T d + o_i(x; \beta d)]_+ - [c_i(x)]_+ \\ &\leq \beta[\nabla c_i(x)^T d]_+ + [o_i(x; \beta d)]_+\end{aligned}$$

for any i, x, d, and $\beta > 0$, with $\lim_{(x,d,\beta)\to(x^*,d^*,0^+)} o_i(x; \beta d)/\beta = 0$. Dividing both sides of the above inequality by β and taking the limit give

$$[\nabla c_i(x^*)^T d^*]_+ \geq \limsup_{(x,d,\beta)\to(x^*,d^*,0^+)} ([c_i(x + \beta d)]_+ - [c_i(x)]_+)/\beta \quad \forall i.$$

Notice that the directional derivative of $[c_i(x)]_+$ at x^* in the direction of d^* is $[\nabla c_i(x^*)^T d^*]_+$ if $c_i(x^*) = 0$ and is zero if $c_i(x^*) < 0$. Combining this with the above relation, we obtain

$$\begin{aligned}F_r'(x^*; d^*) &= \nabla f(x^*)^T d^* + r \sum_{\{i: c_i(x^*)=0\}} [\nabla c_i(x^*)^T d^*]_+ \\ &\geq \limsup_{(x,d,\beta)\to(x^*,d^*,0^+)} (F_r(x + \beta d) - F_r(x))/\beta,\end{aligned}$$

where $F_r'(x^*; d^*)$ denotes the directional derivative of F_r at x^* in the direction d^*. Thus, dividing both sides of (3.9) by β^k/γ and taking the limit as $k \in K$, $k \to \infty$, yield

$$F_r'(x^*; d^*) \geq \sigma\left(\bar{F}_r(x^*, d^*, \alpha^*) - F_r(x^*)\right). \tag{3.10}$$

On the other hand, we note from (2.12) that $\bar{F}_r(x^*, 0, \alpha^*) = F_r(x^*)$ and $\bar{F}_r(x^*, d, \alpha^*)$ is convex in d and its directional derivative at $d = 0$ in the direction d^* coincides with $F_r'(x^*; d^*)$. Therefore

$$F_r'(x^*; d^*) \leq \bar{F}_r(x^*, d^*, \alpha^*) - F_r(x^*). \tag{3.11}$$

Combining (3.10) and (3.11), and using $\sigma \in (0,1)$, we obtain

$$\bar{F}_r(x^*, d^*, \alpha^*) - F_r(x^*) \geq 0.$$

Also, using $r^k = r$ and passing to the limit in (3.8) as $k \in K$, $k \to \infty$, yield

$$\bar{F}_r(x^*, d^*, \alpha^*) - F_r(x^*) \leq -\frac{1}{2}\mu\|d^*\|^2.$$

The last two inequalities imply that $d^* = 0$. Hence, by the same reasoning as in case (i), we obtain that x^* is an optimal solution of problem (2.1). □

4. Local quadratic rate of convergence. In this section, we analyze the local quadratic convergence of the SQCQP method under the following further assumptions.

ASSUMPTION 2.

(a) *Problem* (2.1) *has an optimal solution* x^* *satisfying*

$$H(x^*, v^*) \succeq \mu^* I \quad \forall v^* \in V^*$$

for some constant $\mu^* > 0$, *where* $H(x,v) := G(x) + \sum_{i=1}^m v_i G_i(x)$ *and* $V^* := \{v^* \in \Re^m \,|\, (x^*, v^*)$ *is a KKT pair of problem* (2.1)$\}$.[2]

(b) G *and* $G_1, \ldots, G_m$ *are Lipschitz continuous in a neighborhood of* x^*.

In addition, we will choose $B = G(x)$ in the SQCQP method, so subproblem (2.2) becomes

$$\begin{aligned} &\text{minimize} && g(x)^T d + \frac{1}{2} d^T G(x) d \\ &\text{subject to} && c_i(x) + g_i(x)^T d + \frac{\alpha_i}{2} d^T G_i(x) d \leq 0, \quad i = 1, \ldots, m. \end{aligned} \tag{4.1}$$

Assumption 2(a) implies that x^* is the unique solution of the convex program (2.1).

The next lemma shows that, for x near x^*, all constraints in subproblem (4.1) use the second-order information.

LEMMA 4.1. *If* x *is sufficiently close to* x^*, *then* $(\alpha_i)_{i=1}^m$ *given by* (2.10) *satisfies* $\alpha_i = 1$ *for all* i *and subproblem* (4.1) *has at least one KKT pair.*

Proof. Since $\alpha_1, \ldots, \alpha_m$ are chosen by (2.10), Lemma 2.1 implies that (4.1) is feasible for any x. Suppose x is close to the optimal solution x^* (thus nearly feasible). Let s_1, s_3 be given by (2.4)–(2.6). Then s_1 is close to 0 if x is infeasible, and $s_1 = -\infty$ if x is feasible. Moreover, in a neighborhood of x^*, s_3 is bounded from below by a positive constant. Therefore there exists a neighborhood X of x^* such that, for $x \in X$, we have $2s_1 \leq s_3$ and the rule (2.10) yields $\alpha_i = 1$ for all i.

Fix any $v^* \in V^* \neq \emptyset$. Assumption 2(a) and the continuity property of G, $G_1, \ldots, G_m$ and their eigenvalues imply $H(x, v^*)$ is positive definite for all $x \in X$ sufficiently close to x^*. Then $d = 0$ is the only solution of $G(x)d = 0$, $G_i(x)d = 0$, $i = 1, \ldots, m$, from which we can deduce that the objective function of (4.1) is bounded from below on the feasible region. This implies that (4.1) has an optimal solution [13]. In addition, Lemma 2.1 ensures that the Slater condition (2.7) holds with $t = s_3$ and $d = s_3(\bar{x} - x)$, so (4.1) has at least one Lagrange multiplier vector by [25, Theorem 28.2]. □

By Lemma 4.1, for all x sufficiently close to x^*, subproblem (4.1) reduces to

$$\begin{aligned} &\text{minimize} && g(x)^T d + \frac{1}{2} d^T G(x) d \\ &\text{subject to} && c_i(x) + g_i(x)^T d + \frac{1}{2} d^T G_i(x) d \leq 0, \quad i = 1, \ldots, m. \end{aligned} \tag{4.2}$$

[2] V^* is nonempty and bounded by Assumption 1 and [25, Theorem 28.2].

Then (d, v) forms a KKT pair for (4.2) if and only if

$$g(x) + G(x)d + \sum_{i \in \mathcal{I}(x;d)} v_i(g_i(x) + G_i(x)d) = 0, \tag{4.3}$$

$$c_i(x) + g_i(x)^T d + \frac{1}{2} d^T G_i(x) d = 0, \quad v_i \geq 0, \quad i \in \mathcal{I}(x;d), \tag{4.4}$$

$$c_i(x) + g_i(x)^T d + \frac{1}{2} d^T G_i(x) d < 0, \quad v_i = 0, \quad i \notin \mathcal{I}(x;d), \tag{4.5}$$

for some $\mathcal{I}(x;d) \subseteq \{1, \ldots, m\}$.

LEMMA 4.2. *Problem* (4.2) *with* $x = x^*$ *has the unique optimal solution* $d = 0$.

Proof. Fix any $v^* \in V^* \neq \emptyset$. Then it is readily verified using (4.3)–(4.5) that $d = 0$ together with v^* satisfies the KKT conditions for subproblem (4.2) with $x = x^*$. Since (4.2) is a convex program, this shows $d = 0$ is an optimal solution. To show uniqueness, consider any optimal solution d^* of (4.2) with $x = x^*$. Then d^* is a feasible solution with objective function value of zero so that

$$g(x^*)^T d^* + \frac{1}{2}(d^*)^T G(x^*) d^* = 0, \tag{4.6}$$

$$g_i(x^*)^T d^* + \frac{1}{2}(d^*)^T G_i(x^*) d^* \leq 0, \quad i \in \mathcal{I}^*, \tag{4.7}$$

where $\mathcal{I}^* := \{i \,|\, c_i(x^*) = 0\}$. Multiplying (4.7) by v_i^* (which is nonnegative) and adding to (4.6) yield

$$\begin{aligned} 0 &\geq g(x^*)^T d^* + \frac{1}{2}(d^*)^T G(x^*) d^* + \sum_{i \in \mathcal{I}^*} v_i^* \left(g_i(x^*)^T d^* + \frac{1}{2}(d^*)^T G_i(x^*) d^* \right) \\ &= \frac{1}{2}(d^*)^T H(x^*, v^*) d^*, \end{aligned}$$

where the equality uses $g(x^*) + \sum_{i=1}^m v_i^* g_i(x^*) = 0$ and $v_i^* = 0$ for $i \notin \mathcal{I}^*$. Since $H(x^*, v^*) \succ 0$ by Assumption 2(a), this implies that $d^* = 0$. □

LEMMA 4.3. *There exist a neighborhood* X *of* x^* *and a constant* $\rho_1 > 0$ *such that, for each* $x \in X$, *problem* (4.2) *has a KKT pair and each such KKT pair* (d, v) *satisfies* $\|d\| \leq \rho_1$ *and* $\|v\| \leq \rho_1$.

Proof. By Lemma 4.1, there exists a neighborhood X_0 of x^* such that, for each $x \in X_0$, (4.2) has a KKT pair. Moreover, the proof of Lemma 4.1 shows that the vector $s_3(\bar{x} - x)$ is a feasible solution of (4.2), where s_3 is given by (2.5)–(2.6). For each KKT pair (d, v) of (4.2), since d is an optimal solution of (4.2), we then have

$$\begin{aligned} g(x)^T d + \frac{1}{2} d^T G(x) d &\leq s_3 g(x)^T (\bar{x} - x) + \frac{(s_3)^2}{2} (\bar{x} - x)^T G(x)(\bar{x} - x) \\ &\leq \|g(x)\| \|\bar{x} - x\| + \frac{1}{2}(\bar{x} - x)^T G(x)(\bar{x} - x), \end{aligned}$$

where the last inequality follows from $0 < s_3 \leq 1$. Moreover, the feasibility of d for (4.2) implies that

$$c_i(x) + g_i(x)^T d + \frac{1}{2} d^T G_i(x) d \leq 0, \quad i = 1, \ldots, m.$$

Fix any $v^* \in V^* \neq \emptyset$. Then the above inequalities and $v^* \geq 0$ yield

$$\sum_{i=1}^{m} v_i^* c_i(x) + \left(g(x) + \sum_{i=1}^{m} v_i^* g_i(x) \right)^T d + \frac{1}{2} d^T H(x, v^*) d \leq \|g(x)\| \|\bar{x} - x\| + \frac{1}{2} (\bar{x} - x)^T G(x) (\bar{x} - x). \tag{4.8}$$

By Assumption 2(a) and the continuity property of $G, G_1, \ldots, G_m$ and their eigenvalues, there exists a neighborhood $X \subseteq X_0$ of x^* such that

$$H(x, v^*) \succeq \frac{\mu^*}{2} I$$

for every $x \in X$. Thus, the left-hand side of (4.8) is a convex quadratic function of d with bounded coefficients and uniformly positive definite Hessians. Since the right-hand side of (4.8) is clearly bounded for $x \in X$, this implies the boundedness of d.

By using Lemma 4.1, the boundedness of v can be proved in the same manner as in Lemma 3.3 and thus is omitted. □

In the next lemma, we show that the solution d of subproblem (4.2) tends to zero as the iterate x approaches x^*. In what follows, for any nonempty closed set $V \subseteq \Re^m$, we denote $\mathrm{dist}(v, V) := \min_{v' \in V} \|v - v'\|$.

LEMMA 4.4. *There exist a neighborhood X of x^* and a constant $\rho_2 > 0$ such that for each $x \in X$ and each KKT pair (d, v) of* (4.2)*, conditions* (4.3)–(4.5) *hold for some $\mathcal{I}(x; d) \subseteq \{1, \ldots, m\}$ and*

$$\max\{\|d\|, \mathrm{dist}(v, V^*)\} \leq \rho_2 \|x - x^*\|. \tag{4.9}$$

Proof. Let X be the neighborhood from Lemma 4.3. Consider any sequence $\{x^k\} \subset X$ converging to x^*. Let (d^k, v^k) be any KKT pair of (4.2) with $x = x^k$. By Lemma 4.3, $\{(d^k, v^k)\}$ is bounded. Then it follows from (4.3)–(4.5) that every accumulation point of $\{(d^k, v^k)\}$ is a KKT pair for (4.2) with $x = x^*$. By Lemma 4.2, $d = 0$ is the unique optimal solution of (4.2) with $x = x^*$. Hence it follows that $d^k \to 0$. For any subset $\mathcal{I}$ of $\{1, \ldots, m\}$, let $V_{\mathcal{I}}$ denote the set of vectors v satisfying

$$g(x^*) + \sum_{i \in \mathcal{I}} v_i g_i(x^*) = 0, \quad v_i \geq 0, \quad i \in \mathcal{I}, \quad v_i = 0, \quad i \notin \mathcal{I}. \tag{4.10}$$

Then we can deduce that, if $\mathcal{I}(x^k; d^k) = \mathcal{I}$ holds for an infinite number of k's, then $\mathcal{I} \subseteq \mathcal{I}^* := \{i \mid c_i(x^*) = 0\}$ and any accumulation point v of $\{v^k\}$ along this subsequence belongs to $V_{\mathcal{I}}$. Consequently, by taking X small enough, we can assume that

$$\mathcal{I}(x; d) \subseteq \mathcal{I}^* \quad \text{and} \quad V_{\mathcal{I}(x;d)} \neq \emptyset \tag{4.11}$$

for every $x \in X$ and every optimal solution d of subproblem (4.2).

We claim that

$$\mathrm{dist}(v, V^*) = O(\|d\| + \|x - x^*\|) \tag{4.12}$$

for all $x \in X$ and all (d, v) satisfying the KKT conditions (4.3)–(4.5) for problem (4.2). To see this, note that $V^* \neq \emptyset$ coincides with the solution set of the following linear system in v:

$$g(x^*) + \sum_{i \in \mathcal{I}^*} v_i g_i(x^*) = 0, \quad v_i \geq 0, \quad i \in \mathcal{I}^*, \quad v_i = 0, \quad i \notin \mathcal{I}^*. \tag{4.13}$$

Since (d, v) satisfies (4.3)–(4.5) and $\mathcal{I}(x; d) \subseteq \mathcal{I}^*$ by (4.11), v satisfies the last two conditions of (4.13). Moreover, v satisfies the first equation of (4.13) approximately in the sense that

$$\begin{aligned} g(x^*) + \sum_{i\in\mathcal{I}^*} v_i g_i(x^*) &= g(x^*) - g(x) - G(x)d + \sum_{i\in\mathcal{I}^*} v_i(g_i(x^*) - g_i(x)) - \sum_{i\in\mathcal{I}^*} v_i G_i(x)d \\ &= O(\|d\| + \|x - x^*\|), \end{aligned}$$

where the first equality follows from (4.3) and the second equality follows from the local Lipschitz continuity of g and g_i, $i = 1, \ldots, m$ (cf. Assumption 2(b)) and the uniform boundedness of (x, v) for $x \in X$ (cf. Lemma 4.3). Thus, by applying a well-known Hoffman's error bound [9] to the linear system (4.13) in v, we obtain (4.12), where the constant in the big "O" notation is independent of x and (d, v).

We now show that, by taking a smaller neighborhood X if necessary, we further have that

$$\|d\| = O(\|x - x^*\|),$$

which together with (4.12) would complete the proof. To show this, suppose to the contrary that there exist a sequence $\{x^k\} \subset X$ converging to x^* and a sequence $\{(d^k, v^k)\}$ satisfying (4.3)–(4.5) with $x = x^k$ such that $\|d^k\|/\|x^k - x^*\| \to \infty$. By passing to a subsequence if necessary, we can assume that $\mathcal{I}(x^k; d^k)$ in (4.3)–(4.5) is fixed at some $\mathcal{I}$ for all k. Since $x^k \to x^*$, we have $d^k \to 0$ as was shown earlier. Moreover, by (4.11), we have $\mathcal{I} \subseteq \mathcal{I}^*$ and $V_{\mathcal{I}} \neq \emptyset$. By replacing $\mathcal{I}^*$ by $\mathcal{I}$ in the proof of (4.12), we have that

$$\operatorname{dist}(v, V_{\mathcal{I}}) = O(\|d\| + \|x - x^*\|)$$

for all $x \in X$ and all (d, v) satisfying (4.3)–(4.5) with $\mathcal{I}(x; d) = \mathcal{I}$. Since $\mathcal{I}(x^k; d^k) = \mathcal{I}$, this together with the assumption $\|d^k\|/\|x^k - x^*\| \to \infty$ implies that $\|v^k - \hat{v}^k\| = O(\|d^k\|)$, where $\hat{v}^k$ is the element of $V_{\mathcal{I}}$ nearest to v^k. Since $\hat{v}^k \in V_{\mathcal{I}}$, there holds

$$g(x^*) + \sum_{i\in\mathcal{I}} \hat{v}_i^k g_i(x^*) = 0, \quad \hat{v}_i^k \geq 0, \quad i \in \mathcal{I}, \quad \hat{v}_i^k = 0, \quad i \notin \mathcal{I}.$$

Then it follows from (4.3) that

$$\begin{aligned} g(x^k) - g(x^*) + G(x^k)d^k + \sum_{i\in\mathcal{I}} v_i^k(g_i(x^k) - g_i(x^*) + G_i(x^k)d^k) \\ = -\sum_{i\in\mathcal{I}} (v_i^k - \hat{v}_i^k) g_i(x^*). \end{aligned} \tag{4.14}$$

Also, $\mathcal{I}(x^k; d^k) = \mathcal{I}$ together with (4.4) implies that

$$c_i(x^k) - c_i(x^*) + g_i(x^k)^T d^k + \frac{1}{2}(d^k)^T G_i(x^k) d^k = 0, \quad i \in \mathcal{I}.$$

By Lemma 4.3, $\{v^k\}$ is bounded. Then, by further passing to a subsequence if necessary, we can assume that $\{(d^k/\|d^k\|, (v^k - \hat{v}^k)/\|d^k\|, v^k)\}$ converges to some (u, z, w) with $\|u\| = 1$. Dividing (4.14) and the above equality by $\|d^k\|$, and using $(x^k, d^k) \to (x^*, 0)$, $\|x^k - x^*\|/\|d^k\| \to 0$, and local Lipschitz continuity of g and c_i, g_i, $i \in \mathcal{I}$, yield in the limit that

$$\begin{aligned} G(x^*)u + \sum_{i\in\mathcal{I}} w_i G_i(x^*)u = -\sum_{i\in\mathcal{I}} z_i g_i(x^*), \\ g_i(x^*)^T u = 0, \quad i \in \mathcal{I}. \end{aligned}$$

The above two equalities imply that $u^T(G(x^*) + \sum_{i\in\mathcal{I}} w_i G_i(x^*))u = 0$. Since it is readily seen from $\mathcal{I} \subseteq \mathcal{I}^*$ that $w \in V_{\mathcal{I}} \subseteq V^*$, Assumption 2(a) implies that $G(x^*) + \sum_{i\in\mathcal{I}} w_i G_i(x^*) \succ 0$, and hence $u = 0$, contradicting $\|u\| = 1$. □

Like the ordinary SQP method, the SQCQP method is locally convergent at a quadratic rate if the unit step size is used.

LEMMA 4.5. *For any x sufficiently close to x^* and any KKT pair (d, v) of (4.2), we have $H(x, v) \succeq \frac{1}{2}\mu^* I$ and*

$$\|x + d - x^*\| = O(\|x - x^*\|^2), \quad \operatorname{dist}(v, V^*) = O(\|x - x^*\|^2). \tag{4.15}$$

Proof. For each x and each KKT pair (d, v) of (4.2), we have that (d, v) satisfies (4.3) and

$$c_i(x) + g_i(x)^T d + \frac{1}{2} d^T G_i(x) d = 0, \quad i \in \mathcal{I}(x; d), \tag{4.16}$$

and $v_i = 0$ for $i \notin \mathcal{I}(x; d)$, for some $\mathcal{I}(x; d) \subseteq \{1, \ldots, m\}$.

Let X be a sufficiently small neighborhood of x^* so that Lemmas 4.3 and 4.4 apply. As was shown in the proof of Lemma 4.4, by taking X smaller if necessary, we can further assume that (4.11) holds for every $x \in X$ and every KKT pair (d, v) of (4.2), where $V_{\mathcal{I}(x;d)}$ denotes the set of vectors v satisfying (4.10) with $\mathcal{I}$ being replaced by $\mathcal{I}(x; d)$. Since $\mathcal{I}(x; d)$ takes on only a finite number of distinct values, it suffices to prove the lemma for those x and d such that $\mathcal{I}(x; d) = \mathcal{I}$ for some fixed index set $\mathcal{I} \subseteq \{1, \ldots, m\}$. Then we have from (4.11) that $\mathcal{I} \subseteq \mathcal{I}^*$ and $V_{\mathcal{I}} \neq \emptyset$. Let v^* be an arbitrary element of $V_{\mathcal{I}}$. Then we have

$$g(x^*) + \sum_{i\in\mathcal{I}} v_i^* g_i(x^*) = 0, \tag{4.17}$$

$$c_i(x^*) = 0, \quad i \in \mathcal{I}. \tag{4.18}$$

It follows from (4.3) and (4.17) that

$$\begin{aligned}
0 &= g(x) - g(x^*) + G(x)d + \sum_{i\in\mathcal{I}} \Big(v_i g_i(x) - v_i^* g_i(x^*) + v_i G_i(x) d \Big) \\
&= \left(G(x) + \sum_{i\in\mathcal{I}} v_i G_i(x) \right) (x + d - x^*) + \sum_{i\in\mathcal{I}} (v_i - v_i^*) g_i(x^*) \\
&\quad + g(x) - g(x^*) - G(x)(x - x^*) \\
&\quad + \sum_{i\in\mathcal{I}} v_i \Big(g_i(x) - g_i(x^*) - G_i(x)(x - x^*) \Big) \\
&= \left(G(x) + \sum_{i\in\mathcal{I}} v_i G_i(x) \right) (x + d - x^*) \\
&\quad + \sum_{i\in\mathcal{I}} (v_i - v_i^*) g_i(x^*) + O(\|x - x^*\|^2),
\end{aligned} \tag{4.19}$$

where the last equality follows from Assumption 2(b) and the uniform boundedness of v shown by Lemma 4.3. On the other hand, by (4.16) and (4.18), we have for each $i \in \mathcal{I}$ that

$$
\begin{aligned}
0 &= c_i(x) + g_i(x)^T d + \frac{1}{2} d^T G_i(x) d - c_i(x^*) \\
&= g_i(x^*)^T (x + d - x^*) + c_i(x) - c_i(x^*) - g_i(x^*)^T (x - x^*) \\
&\quad + (g_i(x) - g_i(x^*))^T d + \frac{1}{2} d^T G_i(x) d \\
&= g_i(x^*)^T (x + d - x^*) + O(\|x - x^*\|^2) + O(\|x - x^*\| \cdot \|d\|) + O(\|d\|^2) \\
&= g_i(x^*)^T (x + d - x^*) + O(\|x - x^*\|^2), \qquad (4.20)
\end{aligned}
$$

where the third equality follows from Assumption 2(b) and the last equality follows from Lemma 4.4 and (4.9). Since $v_i = 0$ for all $i \notin \mathcal{I}$, then

$$H(x, v) \;=\; G(x) + \sum_{i \in \mathcal{I}} v_i G_i(x).$$

Let $A_{\mathcal{I}}$ be the $n \times |\mathcal{I}|$ matrix whose columns are $g_i(x^*)$, $i \in \mathcal{I}$. Then (4.19) and (4.20) yield

$$\begin{pmatrix} H(x, v) & A_{\mathcal{I}} \\ A_{\mathcal{I}}^T & 0 \end{pmatrix} \begin{pmatrix} x + d - x^* \\ v_{\mathcal{I}} - v_{\mathcal{I}}^* \end{pmatrix} = O(\|x - x^*\|^2), \qquad (4.21)$$

where $v_{\mathcal{I}}^* := (v_i^*)_{i \in \mathcal{I}}$ and $v_{\mathcal{I}} := (v_i)_{i \in \mathcal{I}}$. Let $A_{\mathcal{I}_1} \in \Re^{n \times |\mathcal{I}_1|}$, with $\mathcal{I}_1 \subseteq \mathcal{I}$, be any submatrix of $A_{\mathcal{I}}$ whose columns are linearly independent and span the column space of $A_{\mathcal{I}}$. Then $A_{\mathcal{I}}(v_{\mathcal{I}} - v_{\mathcal{I}}^*) = A_{\mathcal{I}_1} u$ for some $u \in \Re^{|\mathcal{I}_1|}$, so that (4.21) implies that

$$\begin{pmatrix} H(x, v) & A_{\mathcal{I}_1} \\ A_{\mathcal{I}_1}^T & 0 \end{pmatrix} \begin{pmatrix} x + d - x^* \\ u \end{pmatrix} = O(\|x - x^*\|^2). \qquad (4.22)$$

Also, Assumption 2(a) together with Lemma 4.4 implies that $H(x, v) \succeq \frac{1}{2}\mu^* I$ for all x sufficiently close to x^*. Thus, by taking X sufficiently small, we can assume that $H(x, v) \succeq \frac{1}{2}\mu^* I$ for all $x \in X$ and all KKT pairs (d, v) of (4.2). Since the matrix $A_{\mathcal{I}_1}^T$ is constant with full column rank, this implies that the coefficient matrix in (4.22) is nonsingular and its inverse is uniformly bounded in operator norm for $x \in X$. Then it follows from (4.22) that

$$\|x + d - x^*\| = O(\|x - x^*\|^2), \qquad \|u\| = O(\|x - x^*\|^2).$$

Consequently, we have

$$\sum_{i \in \mathcal{I}} (v_i - v_i^*) g_i(x^*) = A_{\mathcal{I}}(v_{\mathcal{I}} - v_{\mathcal{I}}^*) = A_{\mathcal{I}_1} u = O(\|x - x^*\|^2). \qquad (4.23)$$

Recall that v^* is an arbitrary element of the solution set $V_{\mathcal{I}}$ of the linear system (4.10) in v. Then (4.23) shows that v is an approximate solution of the first equation in (4.10) with residual being $O(\|x - x^*\|^2)$. Since v satisfies the remaining equations and inequalities in (4.10), by invoking Hoffman's error bound [9] and using $V_{\mathcal{I}} \subseteq V^*$, we obtain

$$\operatorname{dist}(v, V^*) \leq \operatorname{dist}(v, V_{\mathcal{I}}) = O(\|x - x^*\|^2),$$

as desired. □

The next lemma shows that the unit step size is accepted whenever an iterate x is sufficiently close to the optimal solution x^*. In other words, the Maratos effect will not occur.

LEMMA 4.6. *Let X and ρ_1 be given by Lemma 4.3. Fix $r^0 > 0$ and $\delta > 0$. Let $\rho_3 := \max\{r^0, \rho_1 + 2\delta\}$. If $x \in X$ is sufficiently close to x^* and (d, v) is any KKT pair of (4.2) satisfying $\max_i v_i \le r \le \rho_3$, then the step size $\beta = 1$ is accepted by the line search criterion (2.11) with $\alpha_i = 1$ for all i.*

Proof. Since $x \in X$, Lemma 4.3 implies that (4.2) has a KKT pair. Any KKT pair (d, v) of (4.2) satisfies $c_i(x) + g_i(x)^T d + \frac{1}{2} d^T G_i(x) d \le 0$ for all i. Thus

$$\begin{aligned} |[c_i(x+d)]_+| &= \left| [c_i(x+d)]_+ - [c_i(x) + g_i(x)^T d + \frac{1}{2} d^T G_i(x) d]_+ \right| \\ &\le \left| c_i(x+d) - \left(c_i(x) + g_i(x)^T d + \frac{1}{2} d^T G_i(x) d \right) \right| \\ &= o(\|d\|^2) \end{aligned}$$

for all i so that

$$\begin{aligned} F_r(x+d) - F_r(x) &= g(x)^T d + \frac{1}{2} d^T G(x) d - r \sum_{i=1}^m [c_i(x)]_+ + r \cdot o(\|d\|^2) \\ &= \left(\bar{F}_r(x, d, \alpha) - F_r(x) \right) + r \cdot o(\|d\|^2), \end{aligned} \tag{4.24}$$

where $\alpha_i = 1$ for all i and the second equality follows from (2.13). Moreover, Lemma 4.5 implies that, for x sufficiently close to x^*, condition (3.1) with $B = G(x)$ is satisfied with $\mu = \frac{1}{2}\mu^*$. Since $r \ge \max_i v_i$, Lemma 3.1 yields that (3.2) holds. Let $\sigma \in (0, 1)$ be the constant in the line search criterion (2.11). Then combining (3.2) with (4.24) yields

$$F_r(x+d) - F_r(x) \le \sigma \left(\bar{F}_r(x, d, \alpha) - F_r(x) \right),$$

provided $r \cdot o(\|d\|^2) \le (1 - \sigma)\mu \|d\|^2/2$. Since $r \le \rho_3$, the latter holds whenever $\|d\|$ is sufficiently small. Notice that $\max_i v_i \le \rho_1 < \rho_3$ so that r exists. Since Lemma 4.4 shows that $\|d\|$ becomes arbitrarily small as x approaches x^*, this implies that (2.11) is satisfied by $\beta = 1$ whenever $x \in X$ is sufficiently close to x^*. ∎

We can now establish the local quadratic convergence of the SQCQP method.

THEOREM 4.7. *There exists a neighborhood X of x^* (depending on r^0 and δ) such that the sequence $\{x^k\}$ generated by the SQCQP method, with $x^1 \in X$ and $B^k := G(x^k)$ for all k, converges to x^* at a Q-quadratic rate. In this case, we also have $\operatorname{dist}(v^k, V^*) \to 0$ at an R-quadratic rate.*[3]

Proof. By Lemma 4.1, there exists a neighborhood X of x^* such that, whenever $x^k \in X$, we have $\alpha_i^k = 1$ for all i and (d^k, v^k) is a well defined KKT pair of (4.2) with $x = x^k$. By Lemmas 4.3 and 4.6, we can assume by taking X smaller if necessary that there exists a $\rho_1 > 0$ such that, whenever $x^k \in X$ and $\max_i v_i^k \le r^k \le \rho_3 := \max\{r^0, \rho_1 + 2\delta\}$, we have $\|v^k\| \le \rho_1$ and $\beta^k = 1$. By Lemma 4.5, we can further assume by taking X smaller if necessary that, whenever $x^k \in X$, we have $x^k + d^k \in X$ and $\|x^k + d^k - x^*\| \le \frac{1}{2}\|x^k - x^*\|$.

Then, whenever

$$x^k \in X \quad \text{and} \quad r^{k-1} \le \rho_3, \tag{4.25}$$

we have that $\alpha_i^k = 1$ for all i and the KKT pair (d^k, v^k) is well defined. Moreover, $\|v^k\| \le \rho_1$ so the updating rule (2.15) and the definition of ρ_3 imply that $r^k \le \rho_3$

[3]For the definition of Q-quadratic and R-quadratic rate of convergence, see [18].

as well as $r^k \geq \max_i v_i^k$. This then implies that $\beta^k = 1$, and hence $x^{k+1} \in X$ and $\|x^{k+1} - x^*\| \leq \frac{1}{2}\|x^k - x^*\|$. Since $x^1 \in X$ and $r^0 \leq \rho_3$, it follows by induction on k that (4.25) holds for all k and that $x^k \to x^*$. Moreover, Lemma 4.5 implies that $\|x^{k+1} - x^*\| = O(\|x^k - x^*\|^2)$ and $\mathrm{dist}(v^k, V^*) = O(\|x^k - x^*\|^2)$ so that $x^k \to x^*$ at a Q-quadratic rate and $\mathrm{dist}(v^k, V^*) \to 0$ at an R-quadratic rate. □

If $G(x) \succeq \mu I$ for all x, where $\mu > 0$ is some constant, then Assumption 2(a) holds automatically and $B^k = G(x^k)$ satisfies the assumptions of Theorem 3.4. Thus, in this case, we conclude from Theorems 3.4 and 4.7 that if $\{x^k\}$ generated by the SQCQP method is bounded (in addition to Assumptions 1 and 2(b)), then $\{x^k\}$ either terminates finitely at the unique optimal solution x^* or converges to x^* at a quadratic rate locally.

We remark that Lemmas 4.1–4.6 and Theorem 4.7 still hold if Assumption 2(a) is replaced by the weaker assumption that, for each x sufficiently close to x^* and each $u \in \Re^n$, there exists a Lagrange multiplier vector v of (4.2) satisfying

$$u^T H(x, v)u \quad \geq \quad \mu^* \|u\|^2$$

for some constant $\mu^* > 0$. This assumption is more restrictive than the quadratic growth assumption made by Anitescu [2, equations (1.1), (1.12)], although our conclusion of quadratic rate of convergence is stronger than the conclusion of superlinear rate of convergence obtained in [2, Theorem 3.6].

5. Conclusion. In this paper we have presented a sequential quadratically constrained quadratic programming (SQCQP) method for solving smooth convex programs. At each iteration, this method solves a single convex quadratically constrained quadratic minimization subproblem. The latter can be formulated as a second-order cone program and solved efficiently. A key advantage of the SQCQP method as compared to the classical SQP methods is that the Maratos effect can be avoided when using an ordinary ℓ_1 exact penalty function for the line search.

As future work, numerical experimentation with the SQCQP method is needed in order to assess the computational saving (if any) that can be achieved over classical SQP methods. Also, it would be interesting to see if some of the assumptions (such as the Slater condition and Assumption 2) used in the convergence analysis can be relaxed.

Acknowledgments. The authors thank Jim Burke and Steve Wright for bringing to their attention the references [21] and [2], respectively. They also thank two referees for helpful comments and suggestions on the original version of the paper.

REFERENCES

[1] F. Alizadeh and S. Schmieta, *Symmetric cones, potential reduction methods*, in Handbook of Semidefinite Programming, H. Wolkowicz, R. Saigal, and L. Vandenberghe, eds., Kluwer Academic Publishers, Boston, MA, 2000, pp. 195–233.

[2] M. Anitescu, *A superlinearly convergent sequential quadratically constrained quadratic programming algorithm for degenerate nonlinear programming*, SIAM J. Optim., 12 (2002), pp. 949–978.

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

[4] B.T. Boggs and J.W. Tolle, *Sequential Quadratic Programming*, Acta Numer. 4, Cambridge University Press, Cambridge, UK, 1995, pp. 1–51.

[5] T.F. Coleman and A.R. Conn, *Nonlinear programming via an exact penalty function: Global analysis*, Math. Programming, 24 (1982), pp. 137–161.

[6] A.S. El-Bakry, R.A. Tapia, T. Tsuchiya, and Y. Zhang, *On the formulation and theory of the Newton interior-point method for nonlinear programming*, J. Optim. Theory Appl., 89 (1996), pp. 507–541.

[7] M. Fukushima, *A finitely convergent algorithm for convex inequalities*, IEEE Trans. Autom. Contr., 27 (1982), pp. 1126–1127.

[8] M. Fukushima, *A successive quadratic programming algorithm with global and superlinear convergence properties*, Math. Programming, 35 (1986), pp. 253–264.

[9] A.J. Hoffman, *On approximate solutions of systems of linear inequalities*, J. Research Nat. Bur. Standards, 49 (1952), pp. 263–265.

[10] S. Kruk and H. Wolkowicz, *SQ2P, sequential quadratic constrained quadratic programming*, in Advances in Nonlinear Programming, Y.X. Yuan, ed., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 177–204.

[11] S. Kruk and H. Wolkowicz, *Sequential, quadratic constrained, quadratic programming for general nonlinear programming*, in Handbook of Semidefinite Programming, H. Wolkowicz, R. Saigal, and L. Vandenberghe, eds., Kluwer Academic Publishers, Boston, MA, 2000, pp. 563–575.

[12] M.S. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret, *Applications of second-order cone programming*, Linear Algebra Appl., 284 (1998), pp. 193–228.

[13] Z.-Q. Luo and S. Zhang, *On extensions of Frank-Wolfe Theorems*, Comput. Optim. Appl., 13 (1999), pp. 87–110.

[14] D.Q. Mayne and E. Polak, *A superlinearly convergent algorithm for constrained optimization problems*, Math. Programming Stud., 16 (1982), pp. 45–61.

[15] R.D.C. Monteiro and T. Tsuchiya, *Polynomial convergence of primal-dual algorithms for the second-order cone programs based on the MZ-family of directions*, Math. Program., 88 (2000), pp. 61–83.

[16] Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM Stud. Appl. Math. 13, SIAM, Philadelphia, PA, 1994.

[17] Y.E. Nesterov, M.J. Todd, and Y. Ye, *Infeasible-start primal-dual methods and infeasibility detectors for nonlinear programming problems*, Math. Program., 84 (1999), pp. 227-267.

[18] J.M. Ortega and W.C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[19] V.M. Panin, *A second-order method for the discrete min-max problem*, U.S.S.R. Comput. Math. and Math. Phys., 19 (1) (1979), pp. 90–100.

[20] V.M. Panin, *Some methods of solving convex programming problems*, U.S.S.R. Comput. Math. and Math. Phys., 21 (2) (1981), pp. 57–72.

[21] E. Polak, D.Q. Mayne, and J.E. Higgins, *On the extension of Newton's method to semi-infinite minimax problems*, SIAM J. Control Optim., 30 (1992), pp. 367–389.

[22] M.J.D. Powell, *Variable metric methods for constrained optimization*, in Mathematical Programming: State of the Art, A. Bachem, M. Grötschel, and B. Korte, eds., Springer-Verlag, Berlin, 1983, pp. 288–311.

[23] D. Ralph and S.J. Wright, *Superlinear convergence of an interior-point method for monotone variational inequalities,* in Complementarity and Variational Problems: State of the Art, M.C. Ferris and J.-S. Pang, eds., SIAM, Philadelphia, 1997, pp. 345–385.

[24] J. Renegar, *Linear programming, complexity theory and elementary functional analysis*, Math. Programming, 70 (1995), pp. 279-351.

[25] R.T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[26] T. Tsuchiya, *A convergence analysis of the scaling-invariant primal-dual path-following algorithms for second-order cone programming*, Optim. Methods Softw., 11/12 (1999), pp. 141–182.

[27] E.J. Wiest and E. Polak, *A generalized quadratic programming-based phase*-I–*phase*-II *method for inequality-constrained optimization*, Appl. Math. Optim., 26 (1992), pp. 223–252.

SIAM J. OPTIM.
Vol. 13, No. 4, pp. 1120–1140

AUGMENTED LAGRANGIANS WITH ADAPTIVE PRECISION CONTROL FOR QUADRATIC PROGRAMMING WITH SIMPLE BOUNDS AND EQUALITY CONSTRAINTS*

Z. DOSTÁL†, A. FRIEDLANDER‡, AND S. A. SANTOS‡

Abstract. In this paper we discuss a specialization of the augmented Lagrangian-type algorithm of Conn, Gould, and Toint to the solution of strictly convex quadratic programming problems with simple bounds and equality constraints. The new feature of the presented algorithm is the adaptive precision control of the solution of auxiliary problems in the inner loop of the basic algorithm which yields a rate of convergence that does not have any term that accounts for inexact solution of auxiliary problems. Moreover, boundedness of the penalty parameter is achieved for the precision control used. Numerical experiments illustrate the efficiency of the presented algorithm and encourage its usage.

Key words. quadratic programming, box and equality constraints, augmented Lagrangians, adaptive precision control

AMS subject classifications. Primary, 65K05; Secondary, 90C20

PII. S1052623499362573

1. Introduction.

We shall be concerned with the problem of finding a minimizer of a quadratic function subject to simple bounds and linear equality constraints, that is,

$$\begin{array}{ll} \text{minimize} & q(x) \\ \text{subject to} & x \in \Omega \end{array} \tag{1.1}$$

with $\Omega = \{x \in \mathbb{R}^n : x \geq 0 \text{ and } Cx = d\}$, $q(x) = \frac{1}{2}x^T Ax - b^T x$, $b \in \mathbb{R}^n$, $d \in \mathbb{R}^m$, $A \in \mathbb{R}^{n\times n}$ symmetric positive definite, and $C \in \mathbb{R}^{m\times n}$ a full rank matrix, with $m < n$.

We are especially interested in problems with m much smaller than n and with the matrix A large and reasonably conditioned (or preconditioned), so that conjugate gradient–based methods are directly applicable. Such problems arise, for example, from the discretization of the variational inequality that describes the equilibrium of a system of elastic bodies in contact in reciprocal formulation whenever such system includes floating bodies [6, 7, 8, 23] or from application of duality-based domain decomposition to the solution of variational inequalities [14, 15].

We restrict our attention to algorithms that reduce problem (1.1) to a sequence of quadratic programming problems with simple bounds. Our approach has been motivated by an effort to exploit recent progress in the solution of the latter problem, namely, effective exploitation of projections [25] that allows drastic changes in the active set from one iteration to another, and results [2, 3, 6, 7, 8, 10, 17, 18, 19] that enable adaptive control of the precision of the solution of auxiliary problems while preserving qualitative properties of the algorithms.

*Received by the editors October 11, 1999; accepted for publication (in revised form) June 5, 2002; published electronically April 30, 2003. This research has been supported by CNPq, FAPESP 95/6574-9, CEZ:J17/98:272400019, and grants GAČR 201/97/0421 and 101/98/0535.
http://www.siam.org/journals/siopt/13-4/36257.html
†On leave from VŠB-Technical University Ostrava, Tř 17 listopadu, CZ-70833 Ostrava, Czech Republic (zdenek.dostal@vsb.cz).
‡Department of Applied Mathematics, IMECC-UNICAMP, University of Campinas, CP 6065, 13083-970 Campinas SP, Brazil (friedlan@ime.unicamp.br, sandra@ime.unicamp.br).

Our starting point is the algorithm presented by Conn, Gould, and Toint [4], who adapted the augmented Lagrangian method of Powell [26] and Hestenes [22] to the solution of problems with a general cost function subject to general equality constraints and simple bounds. When applied to (1.1), their algorithm reduces to a sequence of simple bound constrained problems of the form

$$\begin{array}{ll} \text{minimize} & L(x, \mu^k, \rho_k) \\ \text{subject to} & x \geq 0, \end{array} \tag{1.2}$$

where

$$L(x, \mu^k, \rho_k) = q(x) + (\mu^k)^T(Cx - d) + \frac{\rho_k}{2}\|Cx - d\|^2 \tag{1.3}$$

is known as the augmented Lagrangian function, $\mu^k = (\mu_1^k, \ldots, \mu_m^k)^T$ is the vector of Lagrange multipliers for the equality constraints, $\rho_k > 0$ is the penalty parameter, and $\|\cdot\|$ denotes the Euclidean norm. In [4] the authors developed basic methods of analysis, proved convergence results that also cover the possibility of inexactly solving the auxiliary problems (1.2), and established that a potentially troublesome penalty parameter is bounded away from zero. They implemented their algorithm in the well-known package LANCELOT [5].

The main improvement of the algorithm of Conn, Gould, and Toint that we propose here concerns the precision control of the solution of the auxiliary problems (1.2). In [4] the authors require that problems (1.2) are solved with precision ω_k, where the sequence $\{\omega_k\}$ is defined a priori and converges to zero. Our approach arises from the simple observation that the precision of the solution x^k of the auxiliary problems (1.2) should be related to the feasibility of x^k, i.e., $\|Cx^k - d\|$, since it does not seem reasonable to solve (1.2) with high precision when μ^k is far from the vector of Lagrange multipliers corresponding to the solution of (1.1) (see also [21]). Due to the choice introduced for precision control, an estimate of the rate of convergence is obtained such that it does not have any term accounting for inexact minimization. It is also proved that the penalty parameter generated by our algorithm remains bounded.

After introducing the notation and the notion of extended regularity for points satisfying the nonnegativity constraints, in section 3 we present the algorithm and prove that it is well defined. The sequence of solutions to subproblems is distinguished according to the fulfillment, or not, of extended regularity. The global convergence of the algorithm is proved in section 4, and under the assumption of regularity at the solution x^* of problem (1.1), it is proved that the generated sequence is asymptotically extended regular. These results allow us to extend [12, 13] to include bound constraints. In section 5 we first prove some identification properties of the algorithm. Next we obtain the rate of convergence of the sequence of multipliers and prove that the penalty parameter is bounded. Numerical experiments are described in section 6. In section 7 some comments and conclusions are presented.

2. Notation and preliminaries. Given $J \subseteq N \equiv \{1, \ldots, n\}$, J nonempty, and matrices A and C and vector b, data of problem (1.1), we define the submatrices A_{JJ} and C_J and the subvector b_J that comprise rows and columns determined by the set J. The submatrix C_J is assumed to have the same rows as C. In the same spirit, given a vector $v \in \mathbb{R}^n$, a subvector with components determined by J will be denoted by v_J.

The next equality constrained problem has been discussed by the authors in

[12, 13], and some results obtained there will be useful in this paper:

$$\begin{array}{ll} \text{minimize} & \varphi(y) \\ \text{subject to} & C_J y = d, \end{array} \tag{2.1}$$

where $\varphi(y) = \frac{1}{2} y^T A_{JJ} y - b_J^T y$, $A_{JJ} \in \mathbb{R}^{p \times p}$ is symmetric positive definite, $C_J \in \mathbb{R}^{m \times p}$ is a full row rank matrix, $m \leq p$, $b_J, y \in \mathbb{R}^p$, and $d \in \mathbb{R}^m$ (vector d is the same as in (1.1)). We denote by $\mu^*_{E,J} \in \mathbb{R}^m$ the optimal vector of Lagrange multipliers of problem (2.1). We shall sometimes consider also more general situations including $p \leq m < n$ and C_J with dependent rows. In such cases $\mu^*_{E,J}$ may not exist, and we shall avoid any reference to it.

The following notation will be used throughout the whole paper. The first order updates of the vector of Lagrange multipliers of problems (1.1) and (2.1) will be denoted, respectively, by

$$\widetilde{\mu} = \mu + \rho(Cx - d) \tag{2.2}$$

and

$$\widetilde{\mu}_{E,J} = \mu_{E,J} + \rho(C_J y - d). \tag{2.3}$$

The augmented Lagrangian of (2.1) will be given by

$$L_E(y, \mu_{E,J}, \rho) = \varphi(y) + \mu_{E,J}^T (C_J y - d) + \frac{\rho}{2} \|C_J y - d\|^2. \tag{2.4}$$

The gradients of the augmented Lagrangians (1.3) and (2.4) will be denoted respectively as follows:

$$g(x, \mu, \rho) = \nabla_x L(x, \mu, \rho) = \nabla q(x) + C^T \mu + \rho C^T (Cx - d), \tag{2.5}$$

$$g_E(y, \mu_{E,J}, \rho) = \nabla_y L_E(y, \mu_{E,J}, \rho) = \nabla \varphi(y) + C_J^T \mu_{E,J} + \rho C_J^T (C_J y - d). \tag{2.6}$$

Note that the dimension of vector g_E coincides with the dimension of vector y.

For $B = B^T \in \mathbb{R}^{n \times n}$, we define $\lambda_1(B)$ and $\lambda_n(B)$, respectively, the largest and smallest eigenvalues of matrix B. The smallest eigenvalue of matrix $C_J A_{JJ}^{-1} C_J^T$ will be denoted by γ_J.

The Karush–Kuhn–Tucker (KKT) conditions for problem (1.2) may be conveniently described by the projected gradient g^P that is defined by

$$\begin{array}{lll} g_i^P(x, \mu, \rho) = g_i(x, \mu, \rho) & \text{if} \quad x_i > 0 \quad \text{or} \quad x_i = 0 \quad \text{and} & g_i(x, \mu, \rho) < 0, \\ g_i^P(x, \mu, \rho) = 0 & \text{otherwise, i.e.} \quad x_i = 0 \quad \text{and} & g_i(x, \mu, \rho) \geq 0, \end{array} \tag{2.7}$$

where $g(x, \mu, \rho) = (g_1, \ldots, g_n)^T$. Thus, the KKT conditions for problem (1.2) are satisfied at x if and only if $g^P(x, \mu, \rho) = 0$.

For each vector $x \in \mathbb{R}^n$, we shall denote by $\mathcal{A}(x)$ and $\mathcal{F}(x)$ the *active* and the *free* set of x, respectively, related to the inequality constraints, so that

$$\mathcal{A}(x) = \{i \in N : x_i = 0\} \qquad \text{and} \qquad \mathcal{F}(x) = \{i \in N : x_i \neq 0\}.$$

Note that if $J = \mathcal{F}(x)$, then

$$Cx = C_J x_J, \tag{2.8}$$

$$g_J(x, \mu, \rho) = g_E(x_J, \mu, \rho), \tag{2.9}$$

and if $\mu = \mu_{E,J}$ (and $y = x_J$), then the corresponding update satisfies

$$\widetilde{\mu} = \widetilde{\mu}_{E,J}. \tag{2.10}$$

For the KKT pair (x^*, μ^*) of problem (1.1), we shall also define the binding set

$$\mathcal{B}^* = \{i : x_i^* = 0 \text{ and } [\nabla q(x^*) + C^T\mu^*]_i \geq 0\} \tag{2.11}$$

and its decomposition into the sets

$$\mathcal{B}_0^* = \{i : x_i^* = 0 \text{ and } [\nabla q(x^*) + C^T\mu^*]_i = 0\} \tag{2.12}$$

and

$$\mathcal{B}_1^* = \{i : x_i^* = 0 \text{ and } [\nabla q(x^*) + C^T\mu^*]_i > 0\}. \tag{2.13}$$

DEFINITION 2.1. *Let $x \in \Omega$ (the feasible set of problem (1.1)) denote a given point. If the gradients of all the active constraints (equalities and inequalities) at x are linearly independent, we say that x is regular.*

DEFINITION 2.2. *Given a point $x \in \mathbb{R}^n$ such that $x \geq 0$, let $J = \mathcal{F}(x)$ contain p indices. If the matrix $C_J \in \mathbb{R}^{m\times p}$ is full row rank we say that x is extended regular, or briefly, e-regular.*

For feasible points $x \in \Omega$, the notions of e-regularity and regularity are equivalent. Moreover, for every e-regular point, since the corresponding matrix C_J is full row rank and matrix A_{JJ} is positive definite, it follows that γ_J, the smallest eigenvalue of matrix $C_J A_{JJ}^{-1} C_J^T$, is strictly positive.

3. Algorithm for equality and simple bound constraints. The following algorithm is a modification of the classical augmented Lagrangian method for the solution of strictly convex quadratic functions subject to linear equality and box constraints that enables adaptive precision control of the auxiliary problems.

ALGORITHM 3.1. Given $0 < \alpha < 1$, $\beta > 1$, $M > 0$, $\nu > 0$, $\rho_0 > 0$, $\eta_0 > 0$, $\mu^0 \in \mathbb{R}^m$, and the matrices and vectors that define problem (1.1), set $k = 0$.

Step 1. *Inner iteration with adaptive precision control.*

Find $z \geq 0$ such that

$$\|g^P(z, \mu^k, \rho_k)\| \leq M\|Cz - d\|. \tag{3.1}$$

If z is e-regular or $\|g^P(z, \mu^k, \rho_k)\| \leq \rho_k^{-\nu}$, set $x^k = z$ and go to Step 2. Otherwise, continue with the inner iteration to obtain a point $v \geq 0$ such that

$$\|g^P(v, \mu^k, \rho_k)\| \leq \min\{\rho_k^{-\nu}, M\|Cv - d\|\} \tag{3.2}$$

and set $x^k = v$.

Step 2. *Updating of μ, ρ, η.*

If x^k is e-regular, then

$$\mu^{k+1} = \mu^k + \rho_k(Cx^k - d). \tag{3.3}$$

If $\|Cx^k - d\| \leq \eta_k$, then

$$\rho_{k+1} = \rho_k, \quad \eta_{k+1} = \alpha\eta_k \tag{3.4}$$

else

$$\rho_{k+1} = \beta\rho_k, \quad \eta_{k+1} = \eta_k. \tag{3.5}$$

If x^k is not extended regular, then

$$\mu^{k+1} = \mu^k, \quad \rho_{k+1} = \beta\rho_k, \quad \eta_{k+1} = \eta_k. \tag{3.6}$$

Step 3. Set $k = k + 1$ and return to Step 1.

In Step 1 we can use any algorithm for minimizing a bound constrained quadratic such that the projected gradient converges to zero. Algorithms of this type are presented in [2, 7, 17, 18].

In Algorithm 1 of [4] the vector of Lagrange multipliers is updated after checking if both the projected gradient and the feasibility error are small enough. In Algorithm 3.1 the multipliers are updated by (3.3) every time the current iterate x^k is e-regular. The good behavior of the estimates of the Lagrange multipliers is guaranteed for e-regular points by the adaptive precision control (3.1). Whenever the iterate is not e-regular, we must force this good behavior by (3.6). We prove later that the stopping criterion (3.2) used for points that are not e-regular forces the sequence $\{x^k\}$ produced by Algorithm 3.1 to be asymptotically e-regular. Due to the structure of problem (1.1) it is not necessary to enforce that the feasibility error be bounded by a quantity related to the penalty parameter.

The next theorem shows that Algorithm 3.1 is well defined; that is, any convergent algorithm for the solution of the auxiliary problems will generate either x^k that satisfies the conditions required in Step 1 in a finite number of iterations or a sequence of approximations that converges to the solution of (1.1).

THEOREM 3.2. *Let $M > 0$, $\nu > 0$, $\mu^k \in \mathbb{R}^m$, and $\rho_k > 0$ be given, and let $\{z^j\}$ denote any sequence that converges to the unique solution $\overline{x}$ of the problem*

$$\begin{array}{ll} \textit{minimize} & L(z, \mu^k, \rho_k) \\ \textit{subject to} & z \geq 0. \end{array} \tag{3.7}$$

There exists an index j such that z^j satisfies condition (3.1) *or* (3.2) *stated in Step* 1 *of Algorithm* 3.1 *or $\{z^j\}$ converges to the solution x^* of problem* (1.1).

Proof. Since ρ_k is fixed and $\|g^P(z^j, \mu^k, \rho_k)\|$ tends to zero because $\{z^j\}$ converges to $\overline{x}$, there exists j_0 such that $\|g^P(z^j, \mu^k, \rho_k)\| \leq \rho_k^{-\nu}$ for $j \geq j_0$. Hence, if neither (3.1) nor (3.2) holds for any j, for $j \geq j_0$, $\|g^P(z^j, \mu^k, \rho_k)\| > M\|Cz^j - d\|$ and therefore $\|Cz^j - d\|$ also converges to zero and we must have $C\overline{x} = d$. In this case, if $\overline{r} = g(\overline{x}, \mu^k, \rho_k)$, since $\overline{x}$ is the solution of (3.7), writing the KKT conditions for that problem, it follows that

$$A\overline{x} - b + C^T\mu^k + \rho_k C^T(C\overline{x} - d) = \overline{r}, \tag{3.8}$$

$$\overline{r} \geq 0, \quad \overline{x} \geq 0, \quad \overline{r}^T\overline{x} = 0. \tag{3.9}$$

Substituting $C\overline{x} = d$ into (3.8), we get

$$A\overline{x} - b + C^T\mu^k = \overline{r}. \tag{3.10}$$

However, conditions (3.9), (3.10), and $C\overline{x} = d$ are sufficient for $\overline{x}$ to be the unique solution of (1.1), so that $\overline{x} = x^*$. ∎

The following simple observation will be useful in our proofs.

LEMMA 3.3. *Let* $\{x^k\}$, $\{\mu^k\}$, *and* $\{\rho_k\}$ *be generated by Algorithm* 3.1, $k \geq 1$, $J = \mathcal{F}(x^k)$. *Then*

$$\|g_E(x_J^k, \mu^k, \rho_k)\| \leq M\|C_J x_J^k - d\|. \tag{3.11}$$

Proof. For $\{x^k\}$, $\{\mu^k\}$, and $\{\rho_k\}$ generated by Algorithm 3.1, $k \geq 1$, and $J = \mathcal{F}(x^k)$, by (2.7), (2.8), and (2.9), it follows that

$$\|g_E(x_J^k, \mu^k, \rho_k)\| = \|g_J(x^k, \mu^k, \rho_k)\| \leq \|g^P(x^k, \mu^k, \rho_k)\| \tag{3.12}$$

and

$$Cx^k = C_J x_J^k. \tag{3.13}$$

Therefore, since by the mechanism of Algorithm 3.1 we have

$$\|g^P(x^k, \mu^k, \rho_k)\| \leq M\|Cx^k - d\|, \tag{3.14}$$

then

$$\|g_E(x_J^k, \mu^k, \rho_k)\| \leq M\|C_J x_J^k - d\|, \tag{3.15}$$

where M is given in Algorithm 3.1. □

4. Global convergence. In this section we prove the global convergence of Algorithm 3.1 and, under the assumption of regularity at the solution x^* of problem (1.1), that the sequence $\{x^k\}$ is asymptotically e-regular.

Assumption (AS1). *The solution* x^* *of problem* (1.1) *is regular.*

In [12, 13] the authors addressed the minimization of strictly convex quadratics with equality constraints and proved that the following result holds (Lemma 3.1 of [13]).

LEMMA 4.1. *Given the equality problem* (2.1) *with* $A_{JJ} \in \mathbb{R}^{p\times p}$ *symmetric positive definite and* $C_J \in \mathbb{R}^{m\times p}$ *a full row rank matrix, let* $\gamma_J > 0$ *be the smallest eigenvalue of matrix* $C_J A_{JJ}^{-1} C_J^T$, M *be a positive constant,* $y \in \mathbb{R}^p$, $\mu \in \mathbb{R}^m$, *and*

$$\rho \geq \rho_J \equiv 2\|C_J\|\|A_{JJ}^{-1}\|M/\gamma_J. \tag{4.1}$$

If

$$\|g_E(y, \mu, \rho)\| \leq M\|C_J y - d\|, \tag{4.2}$$

then

$$\|\widetilde{\mu}_{E,J} - \mu_{E,J}^*\| \leq \frac{M_J}{\rho}\|\mu - \mu_{E,J}^*\|, \tag{4.3}$$

where $\widetilde{\mu}_{E,J}$ *is the update defined in* (2.3) *with* $\mu_{E,J} = \mu$, $M_J = \rho_J + 2\gamma_J^{-1}$ *and* $\mu_{E,J}^*$ *is the optimal vector of Lagrange multipliers of problem* (2.1).

In the next lemma we prove that Algorithm 3.1 generates a bounded sequence $\{\mu^k\}$ of approximations to the optimal vector of Lagrange multipliers μ^* for the equality constraints of problem (1.1).

LEMMA 4.2. *Let* $\{\mu^k\}$ *be a sequence generated by Algorithm* 3.1. *Then* $\{\mu^k\}$ *is bounded.*

Proof. Let $\{\mu^k\}$, $\{x^k\}$, and $\{\rho_k\}$ be generated by Algorithm 3.1. In particular, it follows that $\{\rho_k\}$ is nondecreasing.

Assume first that $\{\rho_k\}$ is not bounded. For any $k \geq 1$ and $J = \mathcal{F}(x^k)$, it follows from Lemma 3.3 that

$$\|g_E(x_J^k, \mu^k, \rho_k)\| \leq M\|C_J x_J^k - d\|,$$

where M is given in Algorithm 3.1.

If x^k is e-regular, μ^{k+1} is defined by (3.3) in Step 2 of Algorithm 3.1 and

$$\mu^{k+1} = \mu^k + \rho_k(Cx^k - d) = \mu^k + \rho_k(C_J x_J^k - d), \tag{4.4}$$

so that the vector of multipliers μ^k is updated in the same way as that of the equality problem (2.1) corresponding to $J = \mathcal{F}(x^k)$ at $y = x_J^k$. Moreover, by Definition 2.2, the e-regularity of x^k implies that the smallest eigenvalue γ_J of $C_J A_{JJ}^{-1} C_J^T$ satisfies $\gamma_J > 0$. Applying Lemma 4.1 to this equality problem with $\mu = \mu^k$, $\widetilde{\mu}_{E,J} = \mu^{k+1}$ and using (3.11), we have that, for $\rho_k \geq \rho_J$,

$$\|\mu^{k+1} - \mu_{E,J}^*\| \leq \frac{M_J}{\rho_k}\|\mu^k - \mu_{E,J}^*\|. \tag{4.5}$$

Since there is a finite number of different free sets J corresponding to the extended regular iterates, and by the assumption that ρ_k is not bounded, there exists $\ell > 0$ such that, for $k \geq \ell$, $\rho_k \geq \max \rho_J$ and $0 < \max M_J/\rho_k \leq \delta = \max M_J/\rho_\ell < 1$. It follows that if x^k is e-regular and $k \geq \ell$, then

$$\|\mu^{k+1} - \mu_{E,J}^*\| \leq \delta\|\mu^k - \mu_{E,J}^*\|. \tag{4.6}$$

If x^k is not e-regular, then μ^{k+1} is updated by (3.6) in Step 2 as

$$\mu^{k+1} = \mu^k. \tag{4.7}$$

We have thus proved that if $\{\rho_k\}$ is not bounded, then the subsequence $\{\mu^k\}_{k\geq\ell}$ satisfies either (4.6) or (4.7). However, these are just the assumptions of Corollary A.2 in the appendix, where the vectors $\mu_{E,J}^*$ for the different sets J play the role of $\overline{\mu}^i$ with the different indices i. Thus $\{\mu^k\}$ is bounded whenever $\{\rho_k\}$ is not bounded.

Now, if $\{\rho_k\}$ is bounded, there is k_0 such that, for $k \geq k_0$, the values of ρ_k and η_k are updated by (3.4) in Step 2. It follows that for any $i \geq 0$

$$\|Cx^{k_0+i} - d\| \leq \eta_{k_0+i} = \alpha^i \eta_{k_0}$$

and for $\ell \geq 1$

$$\mu^{k_0+\ell} - \mu^{k_0} = \rho_{k_0} \sum_{i=0}^{\ell-1} (Cx^{k_0+i} - d),$$

so that

$$\begin{aligned}
\|\mu^{k_0+\ell}\| &\leq \|\mu^{k_0}\| + \rho_{k_0} \sum_{i=0}^{\ell-1} \|Cx^{k_0+i} - d\| \\
&\leq \|\mu^{k_0}\| + \rho_{k_0}(1 + \cdots + \alpha^{\ell-1})\eta_{k_0} \\
&\leq \|\mu^{k_0}\| + (\rho_{k_0}/(1-\alpha))\eta_{k_0}.
\end{aligned}$$

Hence $\{\mu^k\}$ is also bounded in this case and the proof is complete. ∎

The next three results are valid for any augmented Lagrangian algorithm that generates bounded sequences of multipliers for problem (1.1). The proofs of these results exploit the particular structure of problem (1.1).

LEMMA 4.3. *Let $\{x^k\}$ and $\{\mu^k\}$ denote sequences in $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively, such that $\{\mu^k\}$ is bounded. If $\{\rho_k\}$ is any sequence of positive numbers and $K \geq 0$ is such that*

$$\|g^P(x^k, \mu^k, \rho_k)\| \leq K, \tag{4.8}$$

then $\{x^k\}$ is bounded.

Proof. For any $k \geq 0$, if $J = \mathcal{F}(x^k)$, by (2.9)

$$\|g_E(x_J^k, \mu^k, \rho_k)\| \leq \|g^P(x^k, \mu^k, \rho_k)\| \leq K. \tag{4.9}$$

Notice that if J is empty, then $x^k = 0$, so that we can assume that J is not empty, without loss of generality. Let

$$r_J^k = g_E(x_J^k, \mu^k, \rho_k) = A_{JJ}x_J^k - b_J + C_J^T\mu^k + \rho_k C_J^T(C_J x_J^k - d). \tag{4.10}$$

Eliminating x_J^k from (4.10) yields

$$x_J^k = (A_{JJ} + \rho_k C_J^T C_J)^{-1}(b_J + r_J^k - C_J^T\mu^k) + \rho^k(A_{JJ} + \rho_k C_J^T C_J)^{-1}C_J^T d. \tag{4.11}$$

We recall the notation of section 2; that is, for matrix $B = B^T \in \mathbb{R}^{n\times n}$, let $\lambda_1(B)$ and $\lambda_n(B)$ be its largest and smallest eigenvalues, respectively. We recast a result on eigenvalues of the sum of symmetric and positive semidefinite matrices (see, e.g., [24, Corollary 4.3.3]), namely

$$\lambda_p(A_{JJ} + \rho_k C_J^T C_J) \geq \lambda_p(A_{JJ}),$$

assuming $A_{JJ} \in \mathbb{R}^{p\times p}$. Therefore,

$$\|(A_{JJ} + \rho_k C_J^T C_J)^{-1}\| = \frac{1}{\lambda_p(A_{JJ} + \rho_k C_J^T C_J)} \leq \frac{1}{\lambda_p(A_{JJ})} = \|A_{JJ}^{-1}\| \tag{4.12}$$

and it follows from (4.11)–(4.12) that

$$\|x_J^k\| \leq \|A_{JJ}^{-1}\|\left(\|b_J\| + \|r_J^k\| + \|C_J^T\|\|\mu^k\|\right) + \rho_k\|(A_{JJ} + \rho_k C_J^T C_J)^{-1}C_J^T d\|. \tag{4.13}$$

To give a bound on the last term in (4.13), notice that it is zero if $C_J = 0$. If $C_J \neq 0$, then by the spectral decomposition theorem [24, Theorem 4.1.5] there is an orthogonal matrix $Q_J = (q_1, \ldots, q_p) \in \mathbb{R}^{p\times p}$ and a nonzero diagonal matrix $\Sigma_J = \operatorname{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0) \in \mathbb{R}^{p\times p}$, $\sigma_1 \geq \cdots \geq \sigma_r > 0$, such that $C_J^T C_J = Q_J \Sigma_J Q_J^T$. Thus, taking $\widehat{\Sigma}_J = \operatorname{diag}(\sigma_1, \ldots, \sigma_r) \in \mathbb{R}^{r\times r}$ and $\widehat{Q}_J = (q_1, \ldots, q_r) \in \mathbb{R}^{p\times r}$, we can define a full rank matrix

$$D_J = \widehat{\Sigma}_J^{\frac{1}{2}}\widehat{Q}_J^T = (\sqrt{\sigma_1}\, q_1, \ldots, \sqrt{\sigma_r}\, q_r)^T \in \mathbb{R}^{r\times p}$$

that satisfies $C_J^T C_J = D_J^T D_J$ and $\|C_J\| = \|D_J\|$. Moreover, as $\operatorname{Im} C_J^T = \operatorname{Im} D_J^T$, there is $\widehat{d} \in \mathbb{R}^r$ such that $C_J^T d = D_J^T \widehat{d}$, so that

$$(A_{JJ} + \rho_k C_J^T C_J)^{-1}C_J^T d = (A_{JJ} + \rho_k D_J^T D_J)^{-1}D_J^T \widehat{d}.$$

Now, combining the last equality with the Sherman–Morrison–Woodbury formula [20, p. 51] yields

$$\begin{aligned}(A_{JJ}+\rho_k C_J^T C_J)^{-1}C_J^T d &= (A_{JJ}+\rho_k D_J^T D_J)^{-1}D_J^T\widehat{d}\\ &= \left(A_{JJ}^{-1}D_J^T - A_{JJ}^{-1}D_J^T(\rho_k^{-1}I + D_J A_{JJ}^{-1}D_J^T)^{-1}D_J A_{JJ}^{-1}D_J^T\right)\widehat{d}\\ &= A_{JJ}^{-1}D_J^T\left(I-(\rho_k^{-1}I+D_JA_{JJ}^{-1}D_J^T)^{-1}D_JA_{JJ}^{-1}D_J^T\right)\widehat{d}\\ &= A_{JJ}^{-1}D_J^T\left(I-(\rho_k^{-1}I+D_JA_{JJ}^{-1}D_J^T)^{-1}\right.\\ &\quad\left.\times\left((D_JA_{JJ}^{-1}D_J^T+\rho_k^{-1}I)-\rho_k^{-1}I\right)\right)\widehat{d}\\ &= \rho_k^{-1}A_{JJ}^{-1}D_J^T(\rho_k^{-1}I+D_JA_{JJ}^{-1}D_J^T)^{-1}\widehat{d}. \qquad (4.14)\end{aligned}$$

As a consequence of the submultiplicativity of the matrix norms, we have from (4.14) that

$$\|(A_{JJ}+\rho_k C_J^T C_J)^{-1}C_J^T d\| \le \frac{\|D_J^T\|\|A_{JJ}^{-1}\|\|\widehat{d}\|}{\rho_k(\rho_k^{-1}+\xi_J)} \le \frac{\|C_J^T\|\|A_{JJ}^{-1}\|\|\widehat{d}\|}{\rho_k\xi_J}, \qquad (4.15)$$

where $\xi_J > 0$ is the smallest eigenvalue of matrix $D_J A_{JJ}^{-1} D_J^T$. Observing that the spectral norm is self-adjoint and applying (4.15), it follows from (4.13) that

$$\|x^k\| = \|x_J^k\| \le M_1^J + M_2^J\|\mu^k\| + M_3^J K, \qquad (4.16)$$

where the constants

$$M_1^J = \|A_{JJ}^{-1}\|\left(\|b_J\| + \frac{\|C_J\|\|\widehat{d}\|}{\xi_J}\right),\ M_2^J = \|A_{JJ}^{-1}\|\|C_J^T\|, \text{ and } M_3^J = \|A_{JJ}^{-1}\|$$

depend only on J. Since the number of all possible free sets is finite and $\{\mu^k\}$ is bounded, it follows that $\{x^k\}$ is bounded. ∎

Remark. The previous result is valid for any iterate x^k, whether or not it is e-regular.

LEMMA 4.4. *Let $\{x^k\}$ be a sequence in $\mathbb{R}^n$ such that $x^k \ge 0$ for every k, $\{\mu^k\}$ be a bounded sequence in $\mathbb{R}^m$, and $\{\rho_k\}$ be a nondecreasing unbounded sequence of positive numbers, and suppose that*

$$g^P(x^k,\mu^k,\rho_k) = 0, \qquad k = 0,1,\ldots. \qquad (4.17)$$

Then $\{x^k\}$ converges to the solution x^* of (1.1).

Proof. By Lemma 4.3, $\{x^k\}$ is bounded. Moreover, (4.17) implies that each x^k is the solution of

$$\begin{array}{ll}\text{minimize} & L(x,\mu^k,\rho_k)\\ \text{subject to} & x \ge 0.\end{array} \qquad (4.18)$$

Thus the solution x^* of (1.1) satisfies

$$L(x^k,\mu^k,\rho_k) \le \min_{\substack{Cx=d\\ x\ge 0}} L(x,\mu^k,\rho_k) = \min_{\substack{Cx=d\\ x\ge 0}} q(x) = q(x^*)$$

so that, for all $k \ge 0$,

$$q(x^k) + (Cx^k-d)^T\mu^k + \frac{1}{2}\rho_k\|Cx^k-d\|^2 \le q(x^*). \qquad (4.19)$$

Let $\overline{x}$ and $\overline{\mu}$ denote limit points of $\{x^k\}$ and $\{\mu^k\}$, respectively. Without loss of generality, we may assume that $\{x^k\}$ converges to $\overline{x}$ and $\{\mu^k\}$ converges to $\overline{\mu}$. By taking the upper limit in (4.19), we get

$$q(\overline{x}) + (C\overline{x} - d)^T\overline{\mu} + \frac{1}{2}\limsup \rho_k \|Cx^k - d\|^2 \leq q(x^*). \tag{4.20}$$

Since $\{\rho_k\}$ is unbounded, it follows that $\|Cx^k - d\|$ converges to zero and $C\overline{x} = d$. Hence, $\overline{x}$ is feasible. However, by (4.20), $q(\overline{x}) \leq q(x^*)$ and thus $\overline{x}$ solves (1.1). Because this argument is valid for any limit point $\overline{x}$ of $\{x^k\}$ and the solution x^* of (1.1) is unique, it follows that $\{x^k\}$ converges to x^*. ∎

Theorem 4.5 establishes that even if the auxiliary problems (1.2) are solved inexactly, with a tolerance that converges to zero, any sequence $\{x^k\}$ generated by an augmented Lagrangian algorithm will converge to the solution x^* of (1.1).

THEOREM 4.5. *Let $\{x^k\}$, $\{\mu^k\}$, and $\{\rho_k\}$ be as in Lemma 4.4, let $\{\varepsilon_k\}$ denote a sequence of nonnegative numbers that converges to zero, and let*

$$\|g^P(x^k, \mu^k, \rho_k)\| \leq \varepsilon_k. \tag{4.21}$$

Then $\{x^k\}$ converges to the solution x^* of (1.1).

Proof. By Lemma 4.3, $\{x^k\}$ is bounded. Let $\overline{x}^k$ be the solution of (4.18) so that

$$\overline{x}^k \geq 0 \qquad \text{and} \qquad g^P(\overline{x}^k, \mu^k, \rho_k) = 0. \tag{4.22}$$

Let $I_0 = \{i \,|\, g_i^P(x^k, \mu^k, \rho_k) = 0\}$ and $I_1 = \{i \,|\, g_i^P(x^k, \mu^k, \rho_k) \neq 0\}$. Thus, by the definition of the projected gradient, if $i \in I_0$, then $g_i(x^k, \mu^k, \rho_k) \geq 0$ and, as $x_i^k \geq 0$, we obtain

$$\begin{aligned}
0 &\geq L(\overline{x}^k, \mu^k, \rho_k) - L(x^k, \mu^k, \rho_k) \\
&= \frac{1}{2}(\overline{x}^k - x^k)^T(A + \rho_k C^T C)(\overline{x}^k - x^k) + (\overline{x}^k - x^k)^T g(x^k, \mu^k, \rho_k) \\
&\geq \frac{\lambda_n(A)}{2}\|\overline{x}^k - x^k\|^2 + (\overline{x}^k_{I_0} - x^k_{I_0})^T g_{I_0}(x^k, \mu^k, \rho_k) + (\overline{x}^k_{I_1} - x^k_{I_1})^T g_{I_1}(x^k, \mu^k, \rho_k) \\
&\geq \frac{\lambda_n(A)}{2}\|\overline{x}^k - x^k\|^2 + (\overline{x}^k_{I_0})^T g_{I_0}(x^k, \mu^k, \rho_k) + (\overline{x}^k_{I_1} - x^k_{I_1})^T g_{I_1}(x^k, \mu^k, \rho_k) \\
&\geq \frac{\lambda_n(A)}{2}\|\overline{x}^k - x^k\|^2 + (\overline{x}^k_{I_1} - x^k_{I_1})^T g_{I_1}(x^k, \mu^k, \rho_k) \\
&\geq \frac{\lambda_n(A)}{2}\|\overline{x}^k - x^k\|^2 - \|\overline{x}^k_{I_1} - x^k_{I_1}\|\|g_{I_1}(x^k, \mu^k, \rho_k)\| \\
&\geq \frac{\lambda_n(A)}{2}\|\overline{x}^k - x^k\|^2 - \varepsilon_k\|\overline{x}^k - x^k\|,
\end{aligned}$$

where the last inequality is due to (4.21). It follows that

$$\|\overline{x}^k - x^k\| \leq 2(\lambda_n(A))^{-1}\varepsilon_k. \tag{4.23}$$

Then

$$\|x^k - x^*\| \leq \|\overline{x}^k - x^k\| + \|\overline{x}^k - x^*\| \leq 2(\lambda_n(A))^{-1}\varepsilon_k + \|\overline{x}^k - x^*\|,$$

and by the assumption on $\{\varepsilon_k\}$ and since Lemma 4.4 ensures that the sequence of solutions $\{\overline{x}^k\}$ converges to x^*, we get the desired result. ∎

Next we show that ultimately the algorithm generates e-regular iterations and finish this section with the global convergence result for Algorithm 3.1.

LEMMA 4.6. *Suppose that Assumption* (AS1) *holds, and let* $\{x^k\}$, $\{\mu^k\}$, *and* $\{\rho_k\}$ *be generated by Algorithm* 3.1. *Then there is* k_0 *such that* x^k *is e-regular for* $k \geq k_0$.

Proof. Since x^* is e-regular, it follows from Definition 2.2 that the matrix C_I where $I = \mathcal{F}(x^*)$ has full row rank. Observing that there is a neighborhood U of x^* such that $\mathcal{F}(x^*) \subseteq \mathcal{F}(x)$ for every $x \in U$, it follows that the matrix C_J, $J = \mathcal{F}(x)$, is full row rank whenever $x \in U$. Thus, there is a neighborhood U of x^* such that every $x \in U$ is e-regular. Suppose that there is an infinite set $S \subseteq \mathbb{N}$ such that for any $k \in S$, x^k is not e-regular. By definition of Step 1, for any $k \in S$

$$\|g^P(x^k, \mu^k, \rho_k)\| \leq \rho_k^{-\nu}$$

and, by (3.6), $\{\rho_k\}_{k\in S}$ is an unbounded nondecreasing sequence. Since $\{\mu^k\}$ is bounded by Lemma 4.2, it follows that if

$$\varepsilon_k = \rho_k^{-\nu}, \quad k \in S,$$

$\{x^k\}_{k\in S}$, $\{\mu^k\}_{k\in S}$, and $\{\rho_k\}_{k\in S}$ satisfy the assumptions of Theorem 4.5 so that $\{x^k\}_{k\in S}$ converges to the solution x^* of (1.1). Hence there is k_0 such that x^k is e-regular for $k \geq k_0$ and $k \in S$. This contradicts our assumption that S is infinite. ∎

THEOREM 4.7. *Let* $\{x^k\}$, $\{\mu^k\}$, *and* $\{\rho_k\}$ *be generated by Algorithm* 3.1, x^* *be the solution of problem* (1.1), *and* μ^* *be the corresponding vector of Lagrange multipliers for the equality constraints, and suppose that Assumption* (AS1) *holds. Then* $\{x^k\}$ *converges to* x^* *and* $\{\mu^k\}$ *converges to* μ^*.

Proof. By Lemma 4.6 there is k_0 such that x^k is extended regular for $k \geq k_0$. Thus, for $k \geq k_0$, all μ^k are updated by (3.3) in Step 2 and

$$\|Cx^k - d\| = \rho_k^{-1}\|\mu^{k+1} - \mu^k\| \leq \rho_k^{-1}(\|\mu^{k+1}\| + \|\mu^k\|). \tag{4.24}$$

If $\{\rho_k\}$ is not bounded, as it is monotone and $\{\mu^k\}$ is bounded by Lemma 4.2, then $\|Cx^k - d\|$ converges to zero.

On the other hand, if $\{\rho_k\}$ is bounded, it follows that there is $k_1 \geq k_0$ such that for $k \geq k_1$, ρ_k and η_k are updated by (3.4) in Step 2 and

$$\|Cx^k - d\| \leq \eta_k = \alpha^{k-k_1}\eta_{k_1}. \tag{4.25}$$

Since $0 < \alpha < 1$ we can conclude that $\|Cx^k - d\|$ converges to zero. However, since at each iteration x^k satisfies (3.1) or (3.2), it follows that $\|g^P(x^k, \mu^k, \rho_k)\|$ also converges to zero. So, by Lemma 4.3, $\{x^k\}$ is bounded.

Since both sequences $\{x^k\}$ and $\{\mu^k\}$ are bounded, they have limit points $\overline{x}$ and $\overline{\mu}$, respectively. As $\|Cx^k - d\|$ converges to zero, $\overline{x}$ is feasible, i.e.,

$$C\overline{x} = d. \tag{4.26}$$

Now, as x^k is e-regular for $k \geq k_1$, then μ^{k+1} is updated by (3.3) in Step 2 and

$$g^P(x^k, \mu^k, \rho_k) = g^P(x^k, \mu^{k+1}, 0). \tag{4.27}$$

As $\|g^P(x^k, \mu^k, \rho_k)\|$ converges to zero, it follows that

$$g^P(\overline{x}, \overline{\mu}, 0) = 0. \tag{4.28}$$

Equations (4.26) and (4.28) are the sufficient conditions for $\overline{x}$ to be the unique solution of problem (1.1), with corresponding vector of Lagrange multipliers $\overline{\mu}$. Therefore, $\overline{x} = x^*$, $\overline{\mu} = \mu^*$, and both sequences $\{x^k\}$, $\{\mu^k\}$ are convergent. ∎

5. Asymptotic convergence analysis. In this section, we first show that the strictly binding set $\mathcal{B}_1^*$ (2.13) of the solution of problem (1.1) is identified in a finite number of steps (i.e., $\mathcal{B}_1^* \subseteq \mathcal{A}(x^k)$ for all k sufficiently large). Next, it is proved that the rate of convergence of the sequence $\{\mu^k\}$ of multipliers is linear. The special structure of the problem under consideration allows us an improvement on the result of Conn, Gould, and Toint [4] in the sense that our estimate does not have any term accounting the errors in the solution of the auxiliary simple bounded problems. Finally, we prove that the penalty parameter is bounded.

LEMMA 5.1. *Suppose that Assumption* (AS1) *holds, and let* $\{x^k\}$ *be a sequence generated by Algorithm* 3.1, x^* *be the solution of* (1.1), *and* μ^* *be the corresponding vector of Lagrange multipliers associated with the equality constraints. Then there is* k_0 *such that for* $k \geq k_0$

$$\mathcal{F}(x^*) \subseteq \mathcal{F}(x^k) \subseteq \mathcal{F}(x^*) \cup \mathcal{B}_0^*, \tag{5.1}$$

and $\mu_{E,J}^* = \mu^*$ *for any* J *that satisfies*

$$\mathcal{F}(x^*) \subseteq J \subseteq \mathcal{F}(x^*) \cup \mathcal{B}_0^*. \tag{5.2}$$

Proof. Let

$$\varepsilon = \min\{x_i^* : i \in \mathcal{F}(x^*)\}, \tag{5.3}$$

$$\delta = \begin{cases} \min\{g_i(x^*, \mu^*, 0) : i \in \mathcal{B}_1^*\} & \text{if} \quad \mathcal{B}_1^* \neq \emptyset \\ 0 & \text{if} \quad \mathcal{B}_1^* = \emptyset \end{cases} \tag{5.4}$$

so that $\varepsilon > 0$ by Assumption (AS1) and $\delta \geq 0$ by (5.4).

Since by Theorem 4.7 $\{x^k\}$ converges to x^*, there is k_1 such that for $k \geq k_1$ and $i \in \mathcal{F}(x^*)$

$$x_i^k \geq \varepsilon/2 > 0$$

so that $i \in \mathcal{F}(x^k)$, proving the first inclusion of (5.1) for $k \geq k_1$. In other words, any free component at the solution is also free for all iterates after k_1.

If $\delta = 0$, by the definition of $\mathcal{B}_1^*$ and (5.4)

$$N = \mathcal{F}(x^*) \cup \mathcal{B}_0^*,$$

and the second inclusion of (5.1) is trivially satisfied in this case for $k \geq k_1$.

If $\delta > 0$, let us assume that $i \in \mathcal{B}_1^*$. As the sequence $\{\mu^k\}$ generated by Algorithm 3.1 converges to μ^* and $\{x^k\}$ converges to x^* by Theorem 4.7, there is $k_2 \geq k_1$ such that, for $k \geq k_2$, x^k is e-regular by Lemma 4.6 and by (3.3) $g_i(x^k, \mu^k, \rho_k) = g_i(x^k, \mu^{k+1}, 0)$. As $g_i(x^k, \mu^{k+1}, 0)$ converges to $g_i(x^*, \mu^*, 0) \geq \delta$, there exists $k_3 \geq k_2$ such that, for $k \geq k_3$,

$$g_i(x^k, \mu^k, \rho_k) > \delta/2. \tag{5.5}$$

Since $\|g^P(x^k, \mu^k, \rho_k)\|$ converges to zero, there is $k_4 \geq k_3$ such that, for $k \geq k_4$,

$$\|g^P(x^k, \mu^k, \rho_k)\| < \delta/2. \tag{5.6}$$

But

$$g_i(x^k, \mu^k, \rho_k) = g_i^P(x^k, \mu^k, \rho_k) \quad \text{for} \quad i \in \mathcal{F}(x^k), \tag{5.7}$$

and so (5.5) contradicts (5.6) for $k \geq k_4$ and $x_i^k > 0$. Thus, $i \in \mathcal{B}_1^*$ implies $i \notin \mathcal{F}(x^k)$ for $k \geq k_0 \equiv \max\{k_1, k_4\}$, which proves the second inclusion in (5.1).

Now, notice that

$$g_J(x^*, \mu^*, 0) = g_J^P(x^*, \mu^*, 0)$$

holds for any J satisfying (5.2), and because (x^*, μ^*) is the KKT pair of (1.1) we obtain by (2.9)

$$g_E(x_J^*, \mu^*, 0) = g_J(x^*, \mu^*, 0) = 0.$$

This shows that the pair (x_J^*, μ^*) satisfies the optimality conditions of problem (2.1), and as this problem has a unique solution, then for each J satisfying (5.2), necessarily $\mu_{E,J}^* = \mu^*$. ∎

THEOREM 5.2. *Suppose that Assumption* (AS1) *holds, and let x^* be the solution of problem* (1.1) *and μ^* be the corresponding vector of optimal Lagrange multipliers for the equality constraints. Then there are positive constants $\widetilde{M}$ and $\overline{\rho}$ such that for any sequences $\{\rho_k\}$ and $\{\mu^k\}$ generated by Algorithm* 3.1 *there is k_0 such that, for $k \geq k_0$ and $\rho_k \geq \overline{\rho}$,*

$$\|\mu^{k+1} - \mu^*\| \leq \frac{\widetilde{M}}{\rho_k} \|\mu^k - \mu^*\|. \tag{5.8}$$

Proof. Let $\{x^k\}$ be a sequence generated by Algorithm 3.1. If $J = \mathcal{F}(x^k)$ and x^k is e-regular, then by Lemma 4.1 applied to problem (2.1) associated with J, there exists M_J such that, if $\rho_k \geq \rho_J$ and

$$\|g_E(x_J^k, \mu^k, \rho_k)\| \leq M \|C_J x_J^k - d\|, \tag{5.9}$$

then

$$\|\widetilde{\mu}_{E,J} - \mu_{E,J}^*\| \leq \rho_k^{-1} M_J \|\mu^k - \mu_{E,J}^*\| \tag{5.10}$$

where M is the value set in the initialization of Algorithm 3.1 and $\widetilde{\mu}_{E,J}$ is the update defined in (2.3) with $\mu_{E,J} = \mu^k$, $\rho = \rho_k$, and $y = x_J^k$. Since, by Lemma 3.3, (5.9) holds for $k \geq 1$, we conclude that (5.10) holds.

By Lemma 4.6 and Lemma 5.1, there exists k_0 such that, for $k \geq k_0$, x^k is e-regular, $J = \mathcal{F}(x^k)$ satisfies (5.2), and $\mu^* = \mu_{E,J}^*$. Denoting $\overline{\rho} = \max\{\rho_J : \mathcal{F}(x^*) \subseteq J \subseteq \mathcal{F}(x^*) \cup \mathcal{B}_0^*\}$ and $\widetilde{M} = \max\{M_J : \mathcal{F}(x^*) \subseteq J \subseteq \mathcal{F}(x^*) \cup \mathcal{B}_0^*\}$, we obtain for $\rho_k \geq \overline{\rho}$, using

$$\widetilde{\mu}_{E,J} = \mu_{E,J} + \rho(C_J y - d) = \mu^k + \rho_k(Cx^k - d) = \mu^{k+1}$$

and (5.10), that

$$\|\mu^{k+1} - \mu^*\| \leq \frac{\widetilde{M}}{\rho_k} \|\mu^k - \mu^*\|. \qquad ∎$$

In section 5 of [4], where the asymptotic convergence analysis of the authors' algorithm is discussed, the typical error bound for the multipliers looks like

$$\|\mu^{k+1} - \mu^*\| \leq c_1 \omega_k + c_2 \rho_k^{-1} \|\mu^k - \mu^*\|,$$

where c_1 and c_2 are positive constants and $\{\omega_k\}$ is the sequence of precision control for the auxiliary problems. This sequence is set a priori and assumed to converge to zero. The particular structure of problem (1.1) makes the bound (5.8) possible, where such a term does not appear.

The next two lemmas are preliminary to the result on the boundedness of the penalty parameter.

LEMMA 5.3. *Suppose that Assumption* (AS1) *holds, and let* x^* *be the solution of problem* (1.1)*,* μ^* *be the corresponding vector of optimal Lagrange multipliers for the equality constraints, and*

$$I = \mathcal{F}(x^*) \cup \mathcal{B}_0^*. \tag{5.11}$$

Then there are positive constants M_1 *and* M_2 *such that, for any* $\mu \in \mathbb{R}^m$*,* $\rho \geq 0$*, and* $x \geq 0$ *such that* $\mathcal{F}(x^*) \subseteq \mathcal{F}(x)$*,*

$$\|g_I(x, \mu, \rho)\| \leq M_1 \|g^P(x, \mu, \rho)\| + M_2 \|\widetilde{\mu} - \mu^*\|, \tag{5.12}$$

where $\widetilde{\mu}$ *is given by* (2.2)*.*

Proof. Let $\mu \in \mathbb{R}^m$, $\rho \geq 0$, and $x \geq 0$ such that $\mathcal{F}(x^*) \subseteq \mathcal{F}(x)$ is given, and let us split the set I into

$$Q = \{i \in I : x_i = 0 \text{ and } g_i(x, \mu, \rho) > 0\} \quad \text{and} \quad R = I \setminus Q.$$

Observe that the indices in I are related to the solution x^* and not to the point x under consideration. Moreover, $\mathcal{F}(x^*) \subseteq \mathcal{F}(x)$ implies that if $x_i = 0$ then $x_i^* = 0$ and if $x_i^* > 0$ then $x_i > 0$. By Assumption (AS1) and $\mathcal{F}(x^*) \subseteq \mathcal{F}(x)$, the set R is not empty. However, it is possible that $Q = \emptyset$. It is easy to verify that

$$\|g_I^P(x, \mu, \rho)\| = \|g_R(x, \mu, \rho)\| \tag{5.13}$$

so that (5.12) holds for $Q = \emptyset$, with any constant $M_1 \geq 1$, because in this case $R = I$ and $\|g_I^P(x, \mu, \rho)\| \leq \|g^P(x, \mu, \rho)\|$.

Let us assume that Q is not empty and notice that, by (5.11),

$$g_I(x^*, \mu^*, 0) = 0 \tag{5.14}$$

and that, for any $x \geq 0$ such that $\mathcal{F}(x^*) \subseteq \mathcal{F}(x)$ and $\widetilde{\mu}$ satisfying (2.2),

$$g_I(x, \widetilde{\mu}, 0) = g_I(x, \widetilde{\mu} - \mu^*, 0) + C_I^T \mu^*. \tag{5.15}$$

Using (2.2), subtracting (5.14) from (5.15), and writing the result in block matrix form, we get

$$\begin{pmatrix} g_R(x, \mu, \rho) \\ g_Q(x, \mu, \rho) \end{pmatrix} = \begin{pmatrix} A_{RR} & A_{RQ} \\ A_{QR} & A_{QQ} \end{pmatrix} \begin{pmatrix} x_R - x_R^* \\ x_Q - x_Q^* \end{pmatrix} + \begin{pmatrix} C_R^T \\ C_Q^T \end{pmatrix} (\widetilde{\mu} - \mu^*), \tag{5.16}$$

so that after using $x_Q = x_Q^* = 0$ and eliminating $x_R - x_R^*$ from (5.16) we obtain

$$x_R - x_R^* = A_{RR}^{-1} \left(g_R(x, \mu, \rho) - C_R^T (\widetilde{\mu} - \mu^*) \right)$$

and

$$g_Q(x, \mu, \rho) = A_{QR} A_{RR}^{-1} g_R(x, \mu, \rho) + (C_Q^T - A_{QR} A_{RR}^{-1} C_R^T)(\widetilde{\mu} - \mu^*),$$

so that

$$\begin{aligned}\|g_I(x,\mu,\rho)\| &\le \|g_R(x,\mu,\rho)\| + \|g_Q(x,\mu,\rho)\| \\ &\le (1+\|A_{QR}A_{RR}^{-1}\|)\|g_R(x,\mu,\rho)\| + \|A_{QR}A_{RR}^{-1}C_R^T - C_Q^T\|\|\widetilde{\mu}-\mu^*\|.\end{aligned}$$

Observe that A_{RR} is a symmetric positive definite matrix, because it is a square submatrix of A. Therefore, $\|A^{-1}\| = 1/\lambda_n(A)$ and $\|A_{RR}^{-1}\| = 1/\lambda_s(A_{RR})$, where $A_{RR} \in \mathbb{R}^{s\times s}$, $s \le n$. By the interlacing property (see [27, pp. 103–104]) it follows that $\lambda_s(A_{RR}) \ge \lambda_n(A)$ and so

$$\|A_{RR}^{-1}\| \le \|A^{-1}\|. \tag{5.17}$$

The inequalities

$$\|C_R^T\| \le \|C\|, \qquad \|C_Q^T\| \le \|C\|, \qquad \|A_{QR}\| \le \|A\|,$$

and (5.17), together with (5.13) and the fact that $\|g_I^P(x,\mu,\rho)\| \le \|g^P(x,\mu,\rho)\|$, complete the proof. ∎

LEMMA 5.4. *Suppose that Assumption* (AS1) *holds, and let* $\{x^k\}$, $\{\mu^k\}$, *and* $\{\rho_k\}$ *be generated by Algorithm* 3.1, *with* $\{\rho_k\}$ *unbounded. Let* $I = \mathcal{F}(x^*)\cup\mathcal{B}_0^*$. *Then there exist* k_0 *and a positive constant* M_1 *such that if* $k \ge k_0$, *then*

$$\|g_I(x^k,\mu^k,\rho_k)\| \le M_1\|Cx^k - d\|. \tag{5.18}$$

Proof. By Lemma 4.6 and Lemma 5.1, there is k_1 such that, for $k \ge k_1$, x^k is e-regular, $\mathcal{F}(x^*) \subseteq \mathcal{F}(x^k) \subseteq \mathcal{F}(x^*)\cup\mathcal{B}_0^*$, and

$$\widetilde{\mu} = \mu^{k+1} = \mu^k + \rho_k(Cx^k - d). \tag{5.19}$$

Moreover, by Lemma 5.3 and the definition of Step 1, there are constants M_1' and M_2' such that, for $I = \mathcal{F}(x^*)\cup\mathcal{B}_0^*$ and $k \ge k_1$,

$$\|g_I(x^k,\mu^k,\rho_k)\| \le M_1'\|Cx^k - d\| + M_2'\|\mu^{k+1} - \mu^*\|. \tag{5.20}$$

Now notice that, if $\mathcal{F}(x^*) \subseteq J \subseteq \mathcal{F}(x^*)\cup\mathcal{B}_0^*$, then by Lemma 5.1, $\mu^* = \mu_{E,J}^*$, and if $\mathcal{F}(x^k) \subseteq J$, then $C_Jx_J^k = Cx^k$. Thus we can estimate the second term on the right-hand side of (5.20) by applying Lemma 4.1 to problem (2.1) corresponding to the set $J = \mathcal{F}(x^k)$. Let M be the value given in Algorithm 3.1, denote by M_1'' the maximum of all M_J over $\mathcal{F}(x^*) \subseteq J \subseteq \mathcal{F}(x^*)\cup\mathcal{B}_0^*$, and let $k_2 \ge k_1$ be such that $k \ge k_2$ implies $\rho_k \ge \rho_J$ for $\mathcal{F}(x^*) \subseteq J \subseteq \mathcal{F}(x^*)\cup\mathcal{B}_0^*$.

By Algorithm 3.1, (2.8) and (2.9), for $k \ge k_2$ and $J = \mathcal{F}(x^k)$, we have

$$\|g_E(x_J^k,\mu^k,\rho_k)\| = \|g_J(x^k,\mu^k,\rho_k)\| \le \|g^P(x^k,\mu^k,\rho_k)\| \le M\|Cx^k - d\| = M\|C_Jx_J^k - d\|$$

so that the assumptions of Lemma 4.1 for problem (2.1) are satisfied with $\mu = \mu^k$ and $y = x_J^k$ and by (2.10) and (5.19) imply that

$$\|\mu^{k+1} - \mu^*\| \le \frac{M_1''}{\rho_k}\|\mu^k - \mu^*\|. \tag{5.21}$$

Now, by (2.2), (5.19), and (5.21) we have

$$\begin{aligned}\rho_k\|C_Jx_J^k - d\| &= \|\mu^{k+1} - \mu^k\| \\ &\ge \|\mu^k - \mu^*\| - \|\mu^{k+1} - \mu^*\| \\ &\ge \left(\frac{\rho_k}{M_1''} - 1\right)\|\mu^{k+1} - \mu^*\|,\end{aligned}$$

and since for sufficiently large k

$$\frac{\rho_k M_1''}{\rho_k - M_1''} = \frac{M_1''}{1 - \frac{M_1''}{\rho_k}} \le 2M_1'',$$

it follows, by (2.8), that there exists $k_0 \ge k_2$ such that, for $k \ge k_0$ and $M_2'' \ge 2M_1''$,

$$\|\mu^{k+1} - \mu^*\| \le M_2''\|C_J x_J^k - d\| = M_2''\|Cx^k - d\|.$$

Hence, from (5.20), for $M_1 = M_1' + M_2' M_2''$ and $k \ge k_0$,

$$\|g_I(x^k, \mu^k, \rho_k)\| \le M_1\|Cx^k - d\|. \qquad \square$$

THEOREM 5.5. *Suppose that Assumption* (AS1) *holds, and let the sequences* $\{x^k\}$, $\{\mu^k\}$, *and* $\{\rho_k\}$ *be generated by Algorithm* 3.1. *Then* $\{\rho_k\}$ *is bounded.*

Proof. Assume that $\{\rho_k\}$ is not bounded. Then, by Lemma 4.6 and (3.5) in Step 2, there exists a subsequence $\mathbb{K}$ such that $\|Cx^k - d\| > \eta_k$ for $k \in \mathbb{K}$. It follows that $\mathbb{N} \setminus \mathbb{K}$ is also infinite by (3.4) and the fact that $\|Cx^k - d\|$ converges to zero. Thus, there exists a subsequence $\mathbb{K}_1$ of $\mathbb{N} \setminus \mathbb{K}$ such that, for $k \in \mathbb{K}_1$, $k+1 \in \mathbb{K}$. For $k \in \mathbb{K}_1$, $\rho_{k+1} = \rho_k$. Let k_0 be large enough so that x^k satisfies (5.1), (5.8), and (5.18). In particular, Assumption (AS1) and (5.1) imply that x^k is e-regular for $k \ge k_0$.

Let $k \in \mathbb{K}_1$, $k \ge k_0$, $J = \mathcal{F}(x^k) \cup \mathcal{F}(x^{k+1})$, $r_J^{k+1} = g_J(x^{k+1}, \mu^{k+1}, \rho_{k+1}) = g_J(x^{k+1}, \mu^{k+1}, \rho_k)$, and $r_J^k = g_J(x^k, \mu^k, \rho_k)$. Eliminating x_J^k and x_J^{k+1} from r_J^k and r_J^{k+1}, respectively, it follows by (3.3) that

$$x_J^k = (A_{JJ} + \rho_k C_J^T C_J)^{-1}(b_J - C_J^T \mu^k + \rho_k C_J^T d + r_J^k)$$

and

$$x_J^{k+1} = (A_{JJ} + \rho_k C_J^T C_J)^{-1}\left(b_J - C_J^T \mu^k + \rho_k C_J^T d - \rho_k C_J^T (C_J x_J^k - d) + r_J^{k+1}\right).$$

Thus,

$$\begin{aligned} C_J x_J^{k+1} - d &= C_J (A_{JJ} + \rho_k C_J^T C_J)^{-1}\left((b_J - C_J^T \mu^k + \rho_k C_J^T d + r_J^k)\right. \\ &\qquad \left. - \rho_k C_J^T (C_J x_J^k - d) + r_J^{k+1} - r_J^k\right) - d \\ &= \left(I - \rho_k C_J (A_{JJ} + \rho_k C_J^T C_J)^{-1} C_J^T\right)(C_J x_J^k - d) \\ &\quad + C_J (A_{JJ} + \rho_k C_J^T C_J)^{-1}(r_J^{k+1} - r_J^k). \end{aligned} \tag{5.22}$$

Applying the Sherman–Morrison–Woodbury formula [20, p. 51]

$$(A_{JJ} + \rho_k C_J^T C_J)^{-1} = A_{JJ}^{-1} - A_{JJ}^{-1} C_J^T (\rho_k^{-1} I + C_J A_{JJ}^{-1} C_J^T)^{-1} C_J A_{JJ}^{-1}$$

and denoting $S_J = C_J A_{JJ}^{-1} C_J^T$ we get

$$\begin{aligned} C_J (A_{JJ} + \rho_k C_J^T C_J)^{-1} C_J^T &= S_J - (S_J + \rho_k^{-1} I - \rho_k^{-1} I)(\rho_k^{-1} I + S_J)^{-1} S_J \\ &= S_J - S_J + \rho_k^{-1}(\rho_k^{-1} I + S_J)^{-1} S_J \\ &= \rho_k^{-1}(\rho_k^{-1} I + S_J)^{-1} S_J \\ &= \rho_k^{-1}(\rho_k^{-1} I + S_J)^{-1}(S_J + \rho_k^{-1} I - \rho_k^{-1} I) \\ &= \rho_k^{-1}\left(I - \rho_k^{-1}(\rho_k^{-1} I + S_J)^{-1}\right) \\ &= \rho_k^{-1}\left(I - \rho_k^{-1}(\rho_k^{-1} I + C_J A_{JJ}^{-1} C_J^T)^{-1}\right). \end{aligned}$$

Thus,

$$I - \rho_k C_J (A_{JJ} + \rho_k C_J^T C_J)^{-1} C_J^T = \rho_k^{-1} (\rho_k^{-1} I + C_J A_{JJ}^{-1} C_J^T)^{-1}, \tag{5.23}$$

and by the analysis of the spectrum of the matrix on the right-hand side of (5.23),

$$\|I - \rho_k C_J (A_{JJ} + \rho_k C_J^T C_J)^{-1} C_J^T\| = \rho_k^{-1} / (\gamma_J + \rho_k^{-1}), \tag{5.24}$$

where γ_J is the smallest eigenvalue of $C_J A_{JJ}^{-1} C_J^T$. With the same reasoning that leads to (4.15) we deduce that

$$\|C_J (A_{JJ} + \rho_k C_J^T C_J)^{-1}\| \leq \frac{\|C_J\| \|A_{JJ}^{-1}\|}{\rho_k (\gamma_J + \rho_k^{-1})},$$

and using (5.24) in (5.22) we obtain that

$$\|C_J x_J^{k+1} - d\| \leq \rho_k^{-1} \left(\frac{\|C_J\| \|A_{JJ}^{-1}\|}{\gamma_J + \rho_k^{-1}} \|r_J^{k+1} - r_J^k\| + \frac{1}{\gamma_J + \rho_k^{-1}} \|C_J x_J^k - d\| \right). \tag{5.25}$$

Now, since x^k is e-regular for $k \geq k_0$, it follows from Definition 2.2 that $\gamma_J > 0$ and $1/(\gamma_J + \rho_k^{-1}) \leq \gamma_J^{-1}$, so, after renaming the constants,

$$\|C_J x_J^{k+1} - d\| \leq \rho_k^{-1} \left(M_J^1 \|r_J^{k+1} - r_J^k\| + M_J^2 \|C_J x_J^k - d\| \right).$$

Observing again that

$$C_J x_J^{k+1} = C x^{k+1} \quad \text{and} \quad C_J x_J^k = C x^k,$$

we get

$$\|C x^{k+1} - d\| \leq \rho_k^{-1} \left(M_J^1 (\|r_J^k\| + \|r_J^{k+1}\|) + M_J^2 \|C x^k - d\| \right). \tag{5.26}$$

After taking maxima of M_J^1, M_J^2 and noting that, for all $k \geq k_0$, $\mathcal{F}(x^k) \subseteq I = \mathcal{F}(x^*) \cup \mathcal{B}_0^*$, we may use Lemma 5.4 to obtain

$$\|C x^{k+1} - d\| \leq \rho_k^{-1} \left(M^1 \|C x^k - d\| + M^2 \|C x^{k+1} - d\| \right), \tag{5.27}$$

where $M^1 = M_1 \max\{M_J^1\} + \max\{M_J^2\}$, $M^2 = M_1 \max\{M_J^1\}$. Thus, for k such that $\rho_k \geq 2M^2$, we have

$$\|C x^{k+1} - d\| \leq \rho_k^{-1} M_3 \|C x^k - d\|,$$

where $M_3 = 2M^1$, and

$$\alpha \eta_k = \eta_{k+1} < \|C x^{k+1} - d\| \leq \rho_k^{-1} M_3 \|C x^k - d\| \leq \rho_k^{-1} M_3 \eta_k$$

for $k \in \mathbb{K}_1$. Thus for arbitrarily large values of ρ_k,

$$\alpha < \rho_k^{-1} M_3,$$

which contradicts our assumption that $\{\rho_k\}$ is unbounded. ∎

6. Numerical experiments. The algorithm has already proved to be useful for numerical solution of contact problems of elasticity discretized by the finite element [11] or boundary element [10] methods. In combination with duality-based domain decomposition methods, the algorithm required as little as 81 conjugate gradient iterations to solve an elliptic variational inequality discretized by 557040 nodes with 513 nodes on the free boundary [14, 15].

The goal of the experiments presented here is just to illustrate the effect of the adaptive precision control of the solution of auxiliary bound constrained problems. We solved a model problem resulting from the finite difference discretization of

$$\begin{aligned} \text{Minimize} \quad & q(u_1,u_2)=\sum_{i=1}^{2}\left(\int_{\Omega_i}|\nabla u_i|^2\,d\Omega-\int_{\Omega_i}Pu_i\,d\Omega\right) \\ \text{subject to} \quad & u_1(0,y)\equiv 0 \text{ and } u_1(1,y)\le u_2(1,y) \text{ for } y\in[0,1], \end{aligned}$$

where $\Omega_1=(0,1)\times(0,1)$, $\Omega_2=(1,2)\times(0,1)$, $P(x,y)=-3$ for $(x,y)\in(0,1)\times[0.75,1)$, $P(x,y)=0$ for $(x,y)\in(0,1)\times(0,0.75)$, $P(x,y)=-1$ for $(x,y)\in(1,2)\times(0,0.25)$, and $P(x,y)=0$ for $(x,y)\in(1,2)\times(0.25,1)$.

This problem is semicoercive due to the lack of Dirichlet data on the boundary of Ω_2. The solution of the model problem may be interpreted as the displacement of two membranes under the traction P. The left membrane is fixed on the left, and the left edge of the right membrane is not allowed to penetrate below the edge of the left membrane. The solution is unique because the right membrane is pressed down. The problem was used as a benchmark in [7, 14, 15].

The discretization scheme consists of a regular grid of 65×65 nodes for each subdomain Ω_i. The duality turns the problem to the minimization problem (1.1) with a strictly convex quadratic function in 65 variables and one equality constraint.

The initial approximation μ^0 for the scalar Lagrange multiplier μ^* for the equality constraint was zero; the other parameters used were defined by $M=100$, $\alpha=0.1$, $\beta=10$, $\nu=1$, $\rho_0=100$, and $\eta_0=1$. The final penalty parameter ρ was 1000 whenever the adaptive precision control was applied. The inner bound constrained quadratic programming problem was solved by the so-called *monotone preconditioning* with the proportioning parameter $\Gamma=1$ (see [8]). The outer stopping rules were defined by $\|g^P\|\le\varepsilon\|b\|$ and $\|Cx-d\|\le 10^{-2}\varepsilon\|b\|$ with $\varepsilon\in\{10^{-5},10^{-7},10^{-9}\}$. To assess the effect of the adaptive precision control, we implemented the algorithm also without such precision control, so that the auxiliary bound constrained problems were solved to the precision defined by $\|g^P\|\le\varepsilon\|b\|$. Moreover, to get an approximate lower bound on the number of the conjugate gradient iterations, we resolved the problem with $\mu^0=\mu^*$, which obviously resulted in one iteration in the outer loop.

The inner stopping rules, that is, the bounds on $\|g^P\|$ used in Step 1 of Algorithm 3.1 and its variants, are on the second row of Table 1. In this table, the total number of inner conjugate gradient iterations is reported in the subcolumns (cg), and the number of the updates of the Lagrange multipliers in the outer loop is presented in the subcolumns (outer). The results of three families of experiments are reported in three major columns. In the first, the auxiliary problems are solved with the adaptive precision control $\|g^P\|\le 100\|Cx-d\|$. In the second and third major columns, the inner stopping criterion coincides with the outer one, although in the second the multipliers are initially zero and in the third they are set as the optimal ones, for reference purposes. Comparing results of the first and second sets, row by row, as the tolerance ε varies, one can see that our adaptive precision control improves the performance of the algorithm as predicted by the theory. Moreover, the first and third

TABLE 1
Performance of Algorithm 3.1.

ε	$\mu^0 = 0$				$\mu^0 = \mu^*$	
	$\|g^P\| \le 100\|Cx - d\|$		$\|g^P\| \le \varepsilon\|b\|$		$\|g^P\| \le \varepsilon\|b\|$	
	cg	outer	cg	outer	cg	outer
10^{-5}	29	6	34	3	29	1
10^{-7}	39	7	50	3	37	1
10^{-9}	54	8	74	4	45	1

sets show that the total number of conjugate gradient iterations necessary to reach the solution with adaptive precision control approaches the optimal performance given by initialization $\mu^0 = \mu^*$.

7. Conclusions. In order to deal with bounded variables, in this paper we extend previous results on the solution of quadratic programming problems with equality constraints [12, 13] and improve, for the quadratic case, results of [4] for the solution of more general problems. The new feature of our algorithm is the adaptive precision control for solution of the auxiliary problems in the context of augmented Lagrangians. A new result for the rate of convergence of the multipliers is obtained, and the boundedness of the penalty parameter is proved. The algorithm was implemented and an experiment illustrates its behavior. Other applications may be found in [10, 11].

The performance of the algorithm may be further improved by using a problem dependent preconditioning [9] or suitable orthogonal projectors that decompose the Hessian of the augmented Lagrangian [14, 15, 16]. These strategies challenge the general wisdom recommending that large penalty parameters should be avoided (see [1]). Indeed, this class of problems, according to [9, 14, 15, 16], yields estimates for the rate of convergence of conjugate gradient independent of both the penalty parameter and the number of equality constraints. We believe that the algorithm may be a powerful tool for the solution of large problems arising in applied sciences and engineering.

A. Appendix.

LEMMA A.1. *Let $\overline{\mu}^1, \ldots, \overline{\mu}^k$ be given m-vectors, $\mathcal{D} \subseteq \mathbb{R}^m$, let a denote a mapping from $\mathcal{D}$ to $\mathbb{R}^m$, and let $0 \le \delta < 1$.*

If for each $\mu \in \mathcal{D}$ there is i such that

$$\|a(\mu) - \overline{\mu}^i\| \le \delta\|\mu - \overline{\mu}^i\|, \tag{A.1}$$

then for each $\mu \in \mathcal{D}$

$$\|a(\mu)\| \le \max\{\|\mu\|, (1+\delta)M/(1-\delta)\}, \tag{A.2}$$

where

$$M = \max\{\|\overline{\mu}^1\|, \ldots, \|\overline{\mu}^k\|\}. \tag{A.3}$$

Proof. For any $\mu \in \mathcal{D}$, there is i such that

$$\begin{aligned} \|a(\mu)\| &\le \|a(\mu) - \overline{\mu}^i\| + \|\overline{\mu}^i\| \le \delta\|\mu - \overline{\mu}^i\| + M \\ &\le \delta\|\mu\| + (1+\delta)M. \end{aligned} \tag{A.4}$$

To finish the proof, it is enough to observe that if

$$\|\mu\| \geq \frac{1+\delta}{1-\delta}M,$$

then substituting into (A.4) we have

$$\|a(\mu)\| \leq \delta\|\mu\| + (1+\delta)M \leq \delta\|\mu\| + (1-\delta)\|\mu\| = \|\mu\|,$$

while if

$$\|\mu\| < \frac{1+\delta}{1-\delta}M,$$

then

$$\|a(\mu)\| \leq \delta\frac{1+\delta}{1-\delta}M + (1+\delta)M = \frac{1+\delta}{1-\delta}M. \qquad \square$$

COROLLARY A.2. *Let $\overline{\mu}^1, \ldots, \overline{\mu}^k$ be given m-vectors, let $\{\mu^\ell\}$ denote any sequence of vectors of $\mathbb{R}^m$, and let $0 \leq \delta < 1$.*

If for each i there is j such that

$$\|\mu^{i+1} - \overline{\mu}^j\| \leq \delta\|\mu^i - \overline{\mu}^j\| \quad \text{or} \quad \|\mu^{i+1}\| \leq \|\mu^i\|, \tag{A.5}$$

then for any ℓ

$$\|\mu^\ell\| \leq \max\{\|\mu^0\|, (1+\delta)M/(1-\delta)\} \tag{A.6}$$

where M is defined by (A.3).

Proof. The inequality (A.6) is trivial for $\ell = 0$. If $\ell > 0$, then by the assumption (A.5) either $\|\mu^\ell\| \leq \|\mu^{\ell-1}\|$, or, applying Lemma A.1 to the mapping a defined on $\mathcal{D} = \{\mu^\ell\}$ by

$$a(\mu^{\ell-1}) = \mu^\ell,$$

we get

$$\|\mu^\ell\| \leq \max\{\|\mu^{\ell-1}\|, (1+\delta)M/(1-\delta)\}. \tag{A.7}$$

Repeating (A.7), we get (A.6). $\square$

Acknowledgments. We thank the anonymous referees for their insightful comments and suggestions that helped us to improve the notation and the reading of this paper.

REFERENCES

[1] D.P. BERTSEKAS, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, London, 1982.

[2] R.H. BIELSCHOWSKY, A. FRIEDLANDER, F.A.M. GOMES, J.M. MARTÍNEZ, AND M. RAYDAN, *An adaptive algorithm for bound constrained quadratic minimization*, Investigación Oper., 7 (1997), pp. 67–102.

[3] J.M. BIRGIN, J.M. MARTÍNEZ, AND M. RAYDAN, *Nonmonotone spectral projected gradient methods on convex sets*, SIAM J. Optim., 10 (2000), pp. 1196–1211.

[4] 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.

[5] A.R. Conn, N.I.M. Gould, and Ph.L. Toint, *LANCELOT: A Fortran Package for Large Scale Nonlinear Optimization*, Springer-Verlag, Berlin, 1992.

[6] Z. Dostál, *Duality based domain decomposition with inexact subproblem solver for contact problems*, in Contact Mechanics II, M.H. Alibiadi and C. Alessandri, eds., Wessex Inst. of Technology, Southampton, 1995, pp. 461–468.

[7] Z. Dostál, *Duality based domain decomposition with proportioning for the solution of free boundary problems*, J. Comput. Appl. Math., 63 (1995), pp. 203–208.

[8] Z. Dostál, *Box constrained quadratic programming with proportioning and projections*, SIAM J. Optim., 7 (1997), pp. 871–887.

[9] Z. Dostál, *On preconditioning and penalized matrices*, Numer. Linear Algebra Appl., 6 (1998), pp. 109–114.

[10] Z. Dostál, A. Friedlander, S.A. Santos, and J. Malík, *Analysis of semicoercive contact problems using symmetric BEM and augmented Lagrangians*, Engineering Analysis with Boundary Elements, 18 (1997), pp. 195–201.

[11] Z. Dostál, A. Friedlander, and S.A. Santos, *Solution of coercive and semicoercive contact problems by FETI domain decomposition*, in Domain Decompostion Methods 10, Contemp. Math. 218, J. Mandel, C. Farhat, and X.-C. Cai, eds., AMS, Providence, RI, 1998, pp. 83–93.

[12] Z. Dostál, A. Friedlander, and S.A. Santos, *Augmented Lagrangians with adaptive precision control for quadratic programming with equality constraints*, Comput. Optim. Appl., 14 (1999), pp. 37–53.

[13] Z. Dostál, A. Friedlander, S.A. Santos, and K. Alesawi, *Augmented Lagrangians with adaptive precision control for quadratic programming with equality constraints: Corrigendum and addendum*, Comput. Optim. Appl., 23 (2000), pp. 127–133.

[14] Z. Dostál, F.A.M. Gomes, and S.A. Santos, *Solution of contact problems by FETI domain decomposition with natural coarse space projection*, Comput. Methods Appl. Mech. Engrg., 190 (2000), pp. 1611–1627.

[15] Z. Dostál, F.A.M. Gomes, and S.A. Santos, *Duality based domain decomposition with natural coarse space for variational inequalities*, J. Comput. Appl. Math., 126 (2000), pp. 397–415.

[16] Z. Dostál, F.A.M. Gomes, and S.A. Santos, *Duality based domain decomposition with adaptive natural coarse grid for contact problems*, in The Mathematics of Finite Elements and Applications X, J.R. Whiteman, ed., Elsevier, Amsterdam, 2000, pp. 259–270.

[17] A. Friedlander, J.M. Martínez, and S.A. Santos, *A new trust region algorithm for bound constrained minimization*, Appl. Math. Optim., 30 (1994), pp. 235–266.

[18] A. Friedlander and J.M. Martínez, *On the maximization of concave quadratic functions with box constraints*, SIAM J. Optim., 4 (1994), pp. 177–192.

[19] A. Friedlander, J.M. Martínez, and M. Raydan, *A new method for large scale box constrained quadratic minimization problems*, Optim. Methods Softw., 5 (1995), pp. 57–74.

[20] G.H. Golub and C.F. Van Loan, *Matrix Computations*, 2nd ed., Johns Hopkins University Press, Baltimore, 1989.

[21] W.W. Hager, *Analysis and implementation of a dual algorithm for constraint optimization*, J. Optim. Theory Appl., 79 (1993), pp. 33–71.

[22] M.R. Hestenes, *Multiplier and gradient methods*, J. Optim. Theory Appl., 4 (1969), pp. 303–320.

[23] I. Hlaváček, J. Haslinger, J. Nečas, and J. Lovíšek, *Solution of Variational Inequalities in Mechanics*, Springer-Verlag, Berlin, 1988.

[24] R.A. Horn and C.R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, 1985.

[25] J.J. Moré and G. Toraldo, *On the solution of large quadratic programming problems with bound constraints*, SIAM J. Optim., 1 (1993), pp. 93–113.

[26] 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.

[27] J.H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, New York, 1988.

SIAM J. OPTIM.
Vol. 13, No. 4, pp. 1141–1158

A NEW EXACT PENALTY FUNCTION*

WALTRAUD HUYER† AND ARNOLD NEUMAIER†

Abstract. For constrained smooth or nonsmooth optimization problems, new continuously differentiable penalty functions are derived. They are proved exact in the sense that under some nondegeneracy assumption, local optimizers of a nonlinear program are precisely the optimizers of the associated penalty function. This is achieved by augmenting the dimension of the program by a variable that controls both the weight of the penalty terms and the regularization of the nonsmooth terms.

Key words. constrained optimization, nonlinear programming, nonsmooth optimization, exact penalty function, Mangasarian–Fromovitz condition, augmented Lagrangian

AMS subject classification. 90C30

PII. S1052623401390537

1. Introduction.

Smooth nonlinear programs are traditionally solved by augmenting the objective function or a corresponding Lagrangian function using penalty or barrier terms to take account of the constraints (see, e.g., the surveys [3, 15]). The resulting merit function is then optimized using either standard unconstrained (or bound constrained) optimization software or sequential quadratic programming (SQP) techniques. Independently of the technique used, the merit function always depends on a small parameter ε (or a large parameter $\rho = \varepsilon^{-1}$); as $\varepsilon \to 0$, minimizers of the merit function converge to the set of minimizers of the original problem. In some SQP approaches, one uses instead so-called exact penalty functions that produce exact optimizers already at sufficiently small positive values of ε. In return, these exact penalty functions have the disadvantage that the evaluation of the merit function either needs Jacobian information (to estimate multipliers) or (for l_1 or l_∞ penalties) is no longer smooth. In addition, both kinds of penalty functions may be unbounded below even when the constrained problem is bounded, which may make it difficult or impossible to locate a minimizer.

For nonsmooth nonlinear programs, solution techniques are much less developed and often restricted to the convex or unconstrained case; in the latter case, constraints are usually handled by an l_∞ exact penalty function (e.g., in SolvOpt [6]). The various approaches are based on combinations of subgradient methods (e.g., [13, 4]), Moreau–Yosida regularization (e.g., [8, 9, 14]), or bundle techniques (e.g., [5, 7, 10]). The regularization again depends on a small smoothing parameter $\varepsilon > 0$ such that for $\varepsilon \to 0$ the original nonsmooth functions are recovered.

In the following, we discuss a new merit function for smooth or nonsmooth optimization problems with equality, inequality, and bound constraints that

- has good smoothness and exactness properties,
- remains bounded below under reasonable conditions,
- combines regularization with penalty techniques, and

*Received by the editors June 6, 2001; accepted for publication (in revised form) August 19, 2002; published electronically April 30, 2003. This research was partially supported by Austrian Fonds zur Förderung der wissenschaftlichen Forschung (FWF) grant P11516-MAT.
http://www.siam.org/journals/siopt/13-4/39053.html
†Institut für Mathematik, Universität Wien, Strudlhofgasse 4, A-1090 Wien, Austria (Waltraud.Huyer@univie.ac.at, Arnold.Neumaier@univie.ac.at).

- is flexible enough to give enough freedom for incorporating available Lagrange multiplier estimates.

The most important new idea is that the merit function is considered as a function of x and ε simultaneously, with the property that under appropriate assumptions, the minimizer (x^*, ε^*) of the merit function satisfies $\varepsilon^* = 0$, so that x^* solves the original problem.

This paper is organized as follows. In section 2, the case of a smooth constrained optimization problem with equality constraints and bound constraints is considered. A penalty function is introduced, and it is proved that under certain assumptions all local minimizers of this penalty function have the form (x, ε), with $\varepsilon = 0$ and x a solution of the original problem. A converse of this can be proved in much greater generality, namely for nonsmooth functions that are suitably regularized. Therefore, in section 3, a nonsmooth objective function and nonsmooth constraints are replaced by regularized functions, and regularization recipes for some common nonsmooth functions are given. As a preparation for the exactness proof, section 4 proves some results involving regular zeros of a not necessarily smooth function. In section 5 the penalty function is generalized to the regularized problem, and it is proved that, for a solution x^* of the constrained optimization problem, $(x^*, 0)$ is a minimizer of the penalty function. In section 6 we illustrate our theory with an example, where the traditional penalty functions are unbounded. Finally, in section 7 our penalty function is generalized to problems involving in addition inequality constraints; results analogous to those for equality constraints are shown to hold by reducing this case to the previous one with the aid of slack variables.

Notation. In the following, the absolute value of a vector is defined componentwise, $|x| := (|x_1|, \ldots, |x_n|)^T$. Similarly, vector inequalities are understood componentwise. The norm used throughout is the Euclidean norm $\|x\| = \sqrt{\sum x_k^2}$, and $B[x_0; r]$ denotes a closed Euclidean ball around x_0 with radius r. The subvector of x indexed by the indices in J is denoted by x_J, and $A_{:J}$ denotes the matrix consisting of the columns of a matrix A indexed by the indices in J. Sets of the form

$$\mathbf{x} = [\underline{x}, \overline{x}] := \{x \in \mathbb{R}^n \mid \underline{x} \le x \le \overline{x}\},$$

where the *lower bound* $\underline{x} \in (\mathbb{R} \cup \{-\infty\})^n$ and the *upper bound* $\overline{x} \in (\mathbb{R} \cup \{\infty\})^n$ are vectors containing proper or infinite bounds on the components of x and $\underline{x} \le \overline{x}$, are referred to as n-dimensional *boxes*.

2. The smooth case. In this section we propose a class of penalty functions for the smooth constrained nonlinear optimization problem

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & x \in [u, v], \quad F(x) = 0, \end{array} \tag{2.1}$$

where $[u, v]$ is a box in $\mathbb{R}^n$ with nonempty interior, $f : D \to \mathbb{R}$ and $F : D \to \mathbb{R}^m$ are continuously differentiable in an open set D containing $[u, v]$. We fix $w \in \mathbb{R}^m$ and consider the equivalent problem

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & F_i(x) = \varepsilon w_i \quad (i = 1, \ldots, m), \\ & x \in [u, v], \quad \varepsilon = 0. \end{array} \tag{2.2}$$

This motivates the definition of the penalty function f_σ on $D \times [0, \bar{\varepsilon}]$ by

$$f_\sigma(x,\varepsilon) = \begin{cases} f(x) & \text{if } \varepsilon = \Delta(x,\varepsilon) = 0, \\ f(x) + \dfrac{1}{2\varepsilon} \cdot \dfrac{\Delta(x,\varepsilon)}{1 - q\Delta(x,\varepsilon)} + \sigma\beta(\varepsilon) & \text{if } \varepsilon > 0,\ \Delta(x,\varepsilon) < q^{-1}, \\ \infty & \text{otherwise,} \end{cases} \tag{2.3}$$

with the *constraint violation measure*

$$\Delta(x,\varepsilon) := \|\varepsilon w - F(x)\|^2, \tag{2.4}$$

where, in addition, $\bar{\varepsilon} > 0$ and $q > 0$ are fixed and $\beta : [0, \bar{\varepsilon}] \to [0, \infty)$ is continuous and continuously differentiable on $(0, \bar{\varepsilon}]$ with $\beta(0) = 0$. The surrogate optimization problem then reads

$$\begin{array}{ll} \min & f_\sigma(x,\varepsilon) \\ \text{s.t.} & (x,\varepsilon) \in [u,v] \times [0,\bar{\varepsilon}]. \end{array} \tag{2.5}$$

Note that $f_\sigma(x,\varepsilon)$ is continuously differentiable in

$$D_\sigma = \{(x,\varepsilon) \in D \times (0,\bar{\varepsilon}) \mid \Delta(x,\varepsilon) < q^{-1}\},$$

with continuous limits at the part of the boundary where the limit values are finite, in particular at $(x, 0)$ with feasible x. Moreover,

$$f_\sigma(x,\varepsilon) = f(x) + \frac{\varepsilon}{2} \cdot \frac{\|w\|^2}{1 - q\|w\|^2\varepsilon^2} + \sigma\beta(\varepsilon) \geq f(x) = f_\sigma(x,0) \quad \text{if } F(x) = 0. \tag{2.6}$$

The shift by εw in the definition of the constraint violation measure allows one to incorporate Lagrange multiplier estimates (that serve to be able to work with better conditioned Hessians; see the remark at the end of this section).

The denominator $1 - q\Delta(x,\varepsilon)$ is included since it forces the level sets of f_σ to remain in the set $\{(x,\varepsilon) \in \mathbb{R}^n \mid \Delta(x,\varepsilon) < q^{-1}\}$ and hence in some sense close to the feasible set of (2.1). In particular, in many cases where the traditional quadratic penalty function (where $w = 0$ and $q = 0$) is unbounded below, moderate positive values for q give a well-behaved penalty problem (cf. the example in section 6). Indeed, f_σ is bounded below on $[u, v] \times [0, \bar{\varepsilon}]$ whenever $f(x)$ is bounded below on the set

$$D' = \{x \in [u,v] \mid \|F(x)\| \leq q^{-1/2} + \bar{\varepsilon}\|w\|\}. \tag{2.7}$$

This is a reasonable condition since it usually holds when f is bounded below on the feasible set, $\bar{\varepsilon}$ is small enough, and q is large enough.

The term $\sigma\beta(\varepsilon)$ is included since it allows us to optimize simultaneously on x and ε, thus automatizing the adaptation of the penalty factor $1/2\varepsilon$. Intuitively, many slices with different, fixed values of ε that are optimized in traditional quadratic penalty methods are arranged consecutively and translated by the term $\sigma\beta(\varepsilon)$ in such a way that the minimizers form a curve with decreasing function values as $\varepsilon \to 0$. Therefore, simultaneous optimization over both x and ε automatically leads to a local minimum at $\varepsilon = 0$. Of course, we need conditions guaranteeing that the term $\sigma\beta(\varepsilon)$ is enough to cause this behavior of the penalty function. As we shall see in section 5, $\beta(\varepsilon) = \sqrt{\varepsilon}$ is an appropriate (but not the only possible) choice.

We say that the *Mangasarian–Fromovitz condition* (see [11]) for (2.1) holds at $x \in [u, v]$ if $F'(x)$ has full rank and there is a $p \in \mathbb{R}^n$ with $F'(x)p = 0$ and

$$p_i \begin{cases} > 0 & \text{if } x_i = u_i, \\ < 0 & \text{if } x_i = v_i. \end{cases}$$

THEOREM 2.1. *In addition to the general assumptions mentioned after* (2.1) *and after* (2.4)*, assume that the set* (2.7) *is bounded, that each* $x \in D'$ *satisfies the Mangasarian–Fromovitz condition, and that*

$$\beta'(\varepsilon) \geq \beta_1 > 0 \quad \textit{for } 0 < \varepsilon < \bar{\varepsilon}. \tag{2.8}$$

If σ *is sufficiently large, there is no Kuhn–Tucker point* (x, ε) *of* (2.5) *with* $\varepsilon > 0$.

In particular, for sufficiently large σ*, every local minimizer* (x^*, ε^*) *of the penalty problem* (2.5) *with finite* $f_\sigma(x^*, \varepsilon^*)$ *has the form* $(x^*, 0)$*, where* x^* *is a local minimizer of the original problem* (2.1).

Proof. If (x, ε) is a Kuhn–Tucker point of (2.5) with $\varepsilon > 0$, then there exist vectors $y, z \in \mathbb{R}^{n+1}$ such that

$$\begin{gathered} \nabla f_\sigma(x, \varepsilon) = y - z, \\ \inf(y_i, x_i - u_i) = \inf(z_i, v_i - x_i) = 0, \qquad i = 1, \ldots, n, \\ y_{n+1} = \inf(z_{n+1}, \bar{\varepsilon} - \varepsilon) = 0, \end{gathered}$$

where $\nabla f_\sigma(x, \varepsilon)$ is the gradient of f_σ with respect to (x, ε). The assertion of the theorem is proved by contradiction. Assume that there exists a sequence $(x^k, \varepsilon_k, \sigma_k)$, $\varepsilon_k \neq 0$ for all k, $\sigma_k \to \infty$ as $k \to \infty$, where (x^k, ε_k) is a Kuhn–Tucker point of f_{σ_k}. We use the abbreviation $\Delta_k := \Delta(x^k, \varepsilon_k)$. The point x^k satisfies

$$\|F(x^k)\| \leq \Delta_k^{1/2} + \varepsilon_k \|w\| \leq q^{-1/2} + \bar{\varepsilon}\|w\|;$$

hence $x^k \in D'$. Since D' is closed and bounded, we may restrict ourselves to a subsequence if necessary and assume that

$$\lim_{k\to\infty} \varepsilon_k = \varepsilon^* \in [0, \bar{\varepsilon}] \quad \text{and} \quad \lim_{k\to\infty} x^k = x^* \in D'. \tag{2.9}$$

The condition $\frac{\partial}{\partial \varepsilon} f_{\sigma_k}(x^k, \varepsilon_k) \leq 0$ yields

$$q\Delta_k^2 + \varepsilon_k^2 \|w\|^2 + 2\varepsilon_k^2 (1 - q\Delta_k)^2 \sigma_k \beta'(\varepsilon_k) \leq \|F(x^k)\|^2, \tag{2.10}$$

with equality in the case $\varepsilon_k \neq \bar{\varepsilon}$. Since the right-hand side is bounded and $\sigma_k \to \infty$, this yields (in view of (2.8) and (2.9))

$$\varepsilon^* = 0 \quad \text{or} \quad \Delta^* = q^{-1}, \tag{2.11}$$

where $\Delta^* := \Delta(x^*, \varepsilon^*)$. The derivatives with respect to x give

$$f_{x_i}(x^k) + \frac{1}{(1 - q\Delta_k)^2 \varepsilon_k} (F'(x^k)^T (F(x^k) - \varepsilon_k w))_i \begin{cases} \geq 0 & \text{if } x_i^k = u_i, \\ = 0 & \text{if } u_i < x_i^k < v_i, \\ \leq 0 & \text{if } x_i^k = v_i \end{cases} \tag{2.12}$$

or

$$(F'(x^k)^T (F(x^k) - \varepsilon_k w))_i + (1 - q\Delta_k)^2 \varepsilon_k f_{x_i}(x^k) \begin{cases} \geq 0 & \text{if } x_i^k = u_i, \\ = 0 & \text{if } u_i < x_i^k < v_i, \\ \leq 0 & \text{if } x_i^k = v_i, \end{cases}$$

where f_{x_i} denotes the partial derivative of f with respect to x_i. By passing to the limit, using (2.9) and (2.11), we obtain

$$(F'(x^*)^T(F(x^*)-\varepsilon^* w))_i \begin{cases} \geq 0 & \text{if } x_i^* = u_i, \\ = 0 & \text{if } u_i < x_i^* < v_i, \\ \leq 0 & \text{if } x_i^* = v_i. \end{cases} \tag{2.13}$$

Since $x^* \in D'$, the Mangasarian–Fromovitz condition holds for $x = x^*$ and some vector $p \in \mathbb{R}^n$. Let $I_1 := \{i \mid x_i^* = u_i\}$, $I_2 := \{i \mid x_i^* = v_i\}$ and $w^* := F(x^*) - \varepsilon^* w$. Then

$$0 = (F'(x^*)p)^T w^* = \sum_{i\in I_1} p_i (F'(x^*)^T w^*)_i + \sum_{i\in I_2} p_i (F'(x^*)^T w^*)_i,$$

and the Mangasarian–Fromovitz condition and (2.13) imply $(F'(x^*)^T w^*)_i = 0$ for $i \in I_1 \cup I_2$ and thus $F'(x^*)^T w^* = 0$. Now the fact that $F'(x^*)$ has full rank yields $w^* = 0$, giving

$$F(x^*) - \varepsilon^* w = 0. \tag{2.14}$$

Hence $\Delta^* = 0$, and by (2.11) we must have $\varepsilon^* = 0$; therefore $F(x^*) = 0$ by (2.14). Now (2.10) and (2.8) yield

$$\frac{q}{\varepsilon_k^2}\Delta_k^2 + \|w\|^2 + 2(1-q\Delta_k)^2\sigma_k\beta_1 \leq \frac{1}{\varepsilon_k^2}\|F(x^k)\|^2.$$

Since $\beta_1 > 0$, the last term on the left-hand side tends to ∞ as $k \to \infty$. Thus the vectors $y^k := \varepsilon_k^{-1} F(x^k)$ satisfy $\|y^k\| \to \infty$, the vectors $z^k := y^k/\|y^k\|$ have norm 1, and (2.12) implies that the numbers μ_i^k $(i = 1, \ldots, n)$, defined by

$$\mu_i^k := \frac{1}{\|y^k\|} f_{x_i}(x^k) + \frac{1}{(1-q\Delta_k)^2}(F'(x^k)^T z^k)_i - \frac{1}{(1-q\Delta_k)^2\|y^k\|}(F'(x^k)^T w)_i,$$

satisfy

$$\mu_i^k \begin{cases} \geq 0 & \text{if } x_i^k = u_i, \\ = 0 & \text{if } u_i < x_i^k < v_i, \\ \leq 0 & \text{if } x_i^k = v_i. \end{cases}$$

If we pick a convergent subsequence z^{k_l} with limit z^* and pass to the limit we obtain

$$(F'(x^*)^T z^*)_i \begin{cases} \geq 0 & \text{if } x_i^* = u_i, \\ = 0 & \text{if } u_i < x_i^* < v_i, \\ \leq 0 & \text{if } x_i^* = v_i. \end{cases}$$

Now similarly as above this yields $z^* = 0$, which is a contradiction to $\|z^*\| = 1$. Thus such a sequence $(x^k, \varepsilon_k, \sigma_k)$ cannot exist, and for sufficiently large σ all Kuhn–Tucker points of f_σ are of the form $(x, 0)$.

Now let (x^*, ε^*) be a local minimizer of f_σ with finite $f_\sigma(x^*, \varepsilon^*)$. If $\varepsilon^* > 0$, then (x^*, ε^*) must be a Kuhn–Tucker point, which is a contradiction. Therefore, $\varepsilon^* = 0$, and since $f_\sigma(x^*, \varepsilon^*)$ is finite, $\Delta(x^*, \varepsilon^*) = 0$. Now (2.4) implies $F(x^*) = 0$, so that x^* is a feasible point of (2.1). Thus (2.6) implies that there is a neighborhood of x^* where $f(x) \geq f(x^*)$ for feasible x. Therefore x^* is a local minimizer of (2.1). □

We conclude that under the stated assumptions, minimizing the penalty function f_σ for sufficiently large σ yields a minimizer of the original problem. Conversely, as we shall prove in section 5 in a more general setting, a minimizer x^* of (2.1) yields a minimizer $(x^*, 0)$ of f_σ for sufficiently large σ and slightly stronger conditions on $\beta(\varepsilon)$.

Remark. If $w_i \neq 0$ for $i = 1, \dots, n$, we can write $w_i = \lambda_i^{-1}$ and rewrite (2.2) as

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & \lambda_i F_i(x) = \varepsilon \quad (i = 1, \dots, m), \\ & x \in [u, v], \quad \varepsilon = 0. \end{array}$$

This is again of the form (2.2) with F_i replaced by $\lambda_i F_i$ and w_i replaced by 1. Therefore, the theorem also holds with $\Delta(x, \varepsilon) = \sum_{i=1}^m (\varepsilon - \lambda_i F_i(x, \varepsilon))^2$ in the penalty function (2.3). Now the penalty function has an augmented Lagrangian interpretation. Indeed, for (x, ε) with small $\Delta(x, \varepsilon)$ we obtain

$$\begin{aligned} f_\sigma(x, \varepsilon) &= f(x) + \frac{1}{2\varepsilon} \cdot \frac{\Delta(x, \varepsilon)}{1 - q\Delta(x, \varepsilon)} + \sigma\beta(\varepsilon) \\ &= f(x) + \frac{1}{2\varepsilon}\Delta(x, \varepsilon) + \sigma\beta(\varepsilon) + O(\Delta(x, \varepsilon)^2) \\ &= f(x) + \frac{1}{2\varepsilon}\sum_{i=1}^m (\lambda_i F_i(x) - \varepsilon)^2 + \sigma\beta(\varepsilon) + O(\Delta(x, \varepsilon)^2) \\ &= f(x) - \sum \lambda_i F_i(x) + \frac{1}{2\varepsilon}\sum \lambda_i^2 F_i(x)^2 + \frac{m\varepsilon}{2} + \sigma\beta(\varepsilon) + O(\Delta(x, \varepsilon)^2). \end{aligned}$$

Thus, for fixed ε and up to constant additive terms and higher order terms, f_σ is an augmented Lagrangian, and the λ_i play the role of (fixed, initial) Lagrange multiplier estimates.

In particular, as in traditional multiplier penalty functions [1], if the λ_i are close to the Lagrange multipliers at the optimizer x^*, then $f_\sigma(x, \varepsilon)$ is nearly stationary at x^* for arbitrary fixed $\varepsilon > 0$. Therefore, numerical schemes for the minimization of $f_\sigma(x, \varepsilon)$ get close to the minimizer already when ε is not very small and hence when the Hessian of f_σ (which gets more and more ill-conditioned as $\varepsilon \to 0$) is still well-conditioned.

3. Regularization of nonsmooth and ill-conditioned problems. In order to regularize the optimization problem (2.1) with not necessarily smooth functions f and F, we assume that we can embed f and F into a family of regularized functions $f(x, \varepsilon)$ and $F(x, \varepsilon)$ that are twice continuously differentiable in (x, ε) when $\varepsilon > 0$ and satisfy

$$f(x) = f(x, 0) = \lim_{\varepsilon \to 0} f(x, \varepsilon), \quad F(x) = F(x, 0) = \lim_{\varepsilon \to 0} F(x, \varepsilon).$$

Of course, the case where the objective function and the constraints are already well-behaved needs no modification, and in this case we simply put $f(x, \varepsilon) = f(x)$ and $F(x, \varepsilon) = F(x)$ for all ε.

Nonsmooth functions arising in practice are often *factorable*, i.e., composed of a finite sequence of elementary operations. Most elementary operations are smooth; the nonsmoothness arises through a small number of nonsmooth elementary functions. A natural regularization approach for factorable functions is to write each nonsmooth

TABLE 1
Some regularization recipes.

$N(x)$	$N(x,\varepsilon)$	Condition
$\max_k x_k$	$\xi + \varepsilon \log \sum \exp((x_k - \xi)/\varepsilon)$	$\xi = \max_k x_k$
$\min_k x_k$	$\xi - \varepsilon \log \sum \exp((\xi - x_k)/\varepsilon)$	$\xi = \min_k x_k$
x^t $(x \geq 0, t < 2)$	$(x+\varepsilon)^t$	$t \neq 0, 1$
$x^t \log x$ $(x \geq 0, t \leq 2)$	$x^t \log(x+\varepsilon)$	$t = 0, 1, 2$
	$(x+\varepsilon)^t \log(x+\varepsilon)$	$t \neq 0, 1, 2$
$\lvert x\rvert^t$ $(t < 2)$	$\lvert x^{t+k}\rvert/(\lvert x\rvert^k + \varepsilon^k)$	$k = \lceil 2 - t \rceil$
$\lvert x\rvert^t \log \lvert x\rvert$ $(0 < t \leq 2)$	$\lvert x^{t+k}\rvert \log \lvert x\rvert/(\lvert x\rvert^k + \varepsilon^k)$	$k = 1 + \lfloor 2 - t \rfloor$
c (huge constant)	$c/(1 + \varepsilon\lvert c\rvert)$	
c (tiny constant)	$c + \varepsilon \operatorname{sign} c$	

elementary function $N(x)$ as a limit of smooth functions $N(x,\varepsilon)$ that are twice continuously differentiable in (x,ε) when $\varepsilon > 0$,

$$N(x) = \lim_{\varepsilon \to 0} N(x,\varepsilon).$$

Assuming that the objective and constraint functions are factorable, we may replace each occurrence $N(r_i)$ of a nonsmooth elementary function in the definition of the objective and constraint functions with $N(r_i, \varepsilon\rho_i)$ depending on an intermediate result r_i and a suitable scaling constant ρ_i. Then we end up with regularized functions $f(x,\varepsilon)$ and $F(x,\varepsilon)$ with the required properties. Possible forms of $N(x,\varepsilon)$ for the most important nonsmooth $N(x)$ are given in Table 1. Note that the first two formulas are independent of ξ; the particular choice indicated is numerically stable and allows us to restrict the sum to those terms where the exponent is $> \log macheps$, where *macheps* is the machine accuracy.

Some smooth nonlinear programs are very difficult to solve since the Hessian matrix of the Lagrangian is severely ill-conditioned everywhere. Often, the reason for this is that the objective function or some constraint contains subexpressions involving some huge or tiny constants. Such constants can be regularized, too, by adapting them according to the last two lines of Table 1.

To approximate elementary functions with step discontinuities, such as

$$\operatorname{pos}(x) = \begin{cases} 1 & \text{if } x > 0, \\ 0 & \text{if } x \leq 0, \end{cases} \qquad \operatorname{nneg}(x) = \begin{cases} 1 & \text{if } x \geq 0, \\ 0 & \text{if } x < 0, \end{cases}$$

$$\operatorname{sign}(x) = \begin{cases} 1 & \text{if } x > 0, \\ 0 & \text{if } x = 0, \\ -1 & \text{if } x < 0, \end{cases}$$

one may use the regularizations given in Table 2 (using $x_+ = \max(x, 0)$). However, no accompanying theory is available since the results presented in section 5 require the Lipschitz continuity of the functions involved.

4. Regular zeros of nonsmooth functions. In this section we derive some technical results that are needed for the successful analysis of nonsmooth equality constraints.

TABLE 2
Regularization of functions with step discontinuities

$N(x)$	$N(x,\varepsilon)$	Condition
$\operatorname{pos}(x)$	$x_+^3/(x_+^3+\varepsilon^3)$	
$\operatorname{nneg}(x)$	$x_+^3/(x_+^3+\varepsilon^3)$	
$\operatorname{sign}(x)$	$x^3/(\lvert x^3\rvert+\varepsilon^3)$	
$\begin{cases} p & \text{if } x \geq 0 (> 0), \\ q & \text{otherwise} \end{cases}$	$\begin{cases} px^3/(x^3+\varepsilon^3) & \text{if } x \geq 0, \\ qx^3/(x^3-\varepsilon^3) & \text{otherwise} \end{cases}$	$pq \leq 0$

DEFINITION 4.1. *A point $x^* \in \mathbb{R}^n$ is called a* regular zero *of a function $H : D \subseteq \mathbb{R}^n \to \mathbb{R}^m$, $m \leq n$, if x^* is in the interior of D and satisfies $H(x^*) = 0$ and there are a closed, convex, and bounded set $\mathcal{A}$ of $m \times n$ matrices and a matrix $B \in \mathbb{R}^{(n-m)\times n}$ such that the augmented matrix $\binom{A}{B}$ is nonsingular for all $A \in \mathcal{A}$ and for every x, y in some neighborhood $N \subseteq D$ of x^* there exists a matrix $A \in \mathcal{A}$ with*

$$H(x) - H(y) = A(x - y). \tag{4.1}$$

The regularity of a zero can be verified under quite general circumstances.

PROPOSITION 4.2. *A point $x^* \in \mathbb{R}^n$ is a regular zero of H if H is continuously differentiable in a neighborhood of x^* and $H'(x^*)$ has full rank.*

Proof. If $H'(x^*)$ has full rank, there is a matrix B such that $\binom{H'(x^*)}{B}$ is square and nonsingular. By continuity, there exists a convex neighborhood N of x^* such that $\binom{H'(x)}{B}$ is nonsingular for $x \in N$ and, as in the proof of [12, Proposition 5.1.4], (4.1) is satisfied if we take for $\mathcal{A}$ the closed convex hull of $\{H'(x) \mid x \in N\}$. □

The preceding result generalizes to certain piecewise differentiable functions if the nonsmoothness is not too severe. For example, we have the following proposition.

PROPOSITION 4.3. *A point $x^* \in \mathbb{R}^n$ is a regular zero of $H(x) = G(x, |x - x^*|)$ if G is continuously differentiable in a neighborhood of $(x^*, 0)$ with $G(x^*, 0) = 0$, and if there exists a matrix B such that, for all diagonal matrices $\Sigma \in \mathbb{R}^{n\times n}$ with $|\Sigma_{ii}| \leq 1$ for $i = 1, \ldots, n$, the matrix $\binom{\partial_1 G(x^*,0)+\partial_2 G(x^*,0)\Sigma}{B}$ is nonsingular.*

Proof. We mimic the proof of Neumaier [12, Proposition 5.1.4]. Without loss of generality, $x^* = 0$. There is a ball $N = B[0; \delta]$ such that $G(x, z)$ is continuously differentiable for x, $z \in N$ and every matrix $\binom{\partial_1 G(x,z)+\partial_2 G(x,z)\Sigma}{B}$ is nonsingular. Let $x, y \in N$. By a version of the mean value theorem we have

$$\begin{aligned} G(x, |x|) - G(y, |y|) &= \int_0^1 \partial_1 G(y + s(x - y), |x|)(x - y)\, ds \\ &\quad + \int_0^1 \partial_2 G(y, |y| + s(|x| - |y|))(|x| - |y|)\, ds \\ &= A(x - y), \end{aligned}$$

where

$$A = \int_0^1 \partial_1 G(y + s(x - y), |x|)\, ds + \int_0^1 \partial_2 G(y, |y| + s(|x| - |y|))\Sigma\, ds$$

and Σ is the diagonal matrix with the diagonal entries

$$\Sigma_{ii} := \begin{cases} (|x_i| - |y_i|)/(x_i - y_i) & \text{if } x_i \neq y_i, \\ 0 & \text{otherwise.} \end{cases}$$

Since N is convex, A is contained in the closed convex hull $\mathcal{A}$ of the set of expressions $\partial_1 G(x,|x'|) + \partial_2 G(y,|y'|)\Sigma$, where $x, x', y, y' \in N$ and Σ is a diagonal matrix with $|\Sigma_{ii}| \le 1$ for $i = 1, \dots, n$. If N is chosen sufficiently small, $\binom{A}{B}$ is nonsingular for all $A \in \mathcal{A}$. □

In particular, this applies to $H(x) = G(x) + \delta G_1(x,|x|)$ if G and G_1 are continuously differentiable near 0, $G'(0)$ has full rank, and δ is sufficiently small.

With a similar argument but now involving generalized derivatives (essentially the convex hull of the limit set of gradients of nicely approximating smooth functions; see, e.g., [2, 12]), the following result can be proved.

PROPOSITION 4.4. *A point $x^* \in \mathbb{R}^n$ is a regular zero of $H(x)$ if H is Lipschitz continuous in a neighborhood of x^* with $H(x^*) = 0$, and if there exists a matrix B such that, for all matrices H' contained in the generalized derivative of H at x^*, the matrix $\binom{H'}{B}$ is nonsingular.*

For x near a regular zero of H and an arbitrary set J of indices, one can find a small perturbation of the order of $O(\|H_J(x)\|)$ such that the perturbed vector y satisfies $H_J(y) = 0$ and $H_i(y) = H_i(x)$ for all $i \notin J$.

THEOREM 4.5. *Let x^* be a regular zero of $H : D \subseteq \mathbb{R}^n \to \mathbb{R}^m$, $m \le n$. Then there are a neighborhood $N_0 \subseteq N$ of x^* and a constant $\gamma_0 > 0$ such that for each $x \in N_0$ and each subset J of $\{1, \dots, m\}$ there exists a vector $y = y(x) \in N$ with $H_i(y) = 0$ for $i \in J$ and $H_i(y) = H_i(x)$ for $i \notin J$ such that*

$$\|x - y\| \le \gamma_0 \|H_J(x)\|.$$

Proof. We define the neighborhood $N_0 := B[x^*; (2\gamma_0(L + \|B\|))^{-1} r] \cap N$, where $r > 0$ is such that the closed ball $N_1 := B[x^*; r]$ is contained in the neighborhood N of x^*, and

$$L := \sup_{A \in \mathcal{A}} \|A\| < \infty, \quad \gamma_0 := \sup_{A \in \mathcal{A}} \left\| \begin{pmatrix} A \\ B \end{pmatrix}^{-1} \right\| < \infty$$

by assumption. By (4.1), the constant L is a Lipschitz constant for H. We fix a vector $x \in N_0$ and a subset J of $\{1, \dots, m\}$, and put $K := \{1, \dots, m\} \setminus J$.

To find y, we want to apply the nonsmooth inverse function theorem given in Neumaier [12, Theorem 5.1.6(iv)] to the mapping $F : N \to \mathbb{R}^n$ defined by

$$F(z) := \begin{pmatrix} H(z) \\ B(z - x^*) \end{pmatrix} \quad \text{for } z \in N,$$

with the right-hand side

$$b^* = \begin{pmatrix} b \\ B(x - x^*) \end{pmatrix}, \tag{4.2}$$

where $b_i = 0$ for $i \in J$ and $b_i = H_i(x)$ for $i \in K$, and $\tilde{x}^0 = x^*$. Since $H(x^*) = 0$ implies $F(x^*) = 0$, this requires that we show

$$F(z) \ne s b^* \quad \text{for all } z \in \partial N_1,\ s \in [0, 1). \tag{4.3}$$

If this holds, [12, Theorem 5.1.6(iv)] implies that there is a unique $y \in N_1$ with $F(y) = b^*$; hence $H_i(y) = 0$ for $i \in J$ and $H_i(y) = H_i(x)$ for $i \in K$. Moreover,

$$F(x) - F(y) = \begin{pmatrix} A \\ B \end{pmatrix} (x - y)$$

for some $A \in \mathcal{A}$ and

$$\|F(x) - F(y)\| = \|H_J(x)\|;$$

hence

$$\|x - y\| \le \left\| \begin{pmatrix} A \\ B \end{pmatrix}^{-1} \right\| \|H_J(x)\| \le \gamma_0 \|H_J(x)\|.$$

Thus y has the required properties.

To show (4.3), we suppose that $F(z) = sb^*$ for some $z \in N_1,\ s \in [0,1)$. Since N_1 is contained in N, there exists an $A \in \mathcal{A}$ such that

$$sb^* = F(z) = F(z) - F(x^*) = \begin{pmatrix} A \\ B \end{pmatrix}(z - x^*),$$

and we obtain

$$\|z - x^*\| = \|s \begin{pmatrix} A \\ B \end{pmatrix}^{-1} b^*\| \le s\gamma_0 \|b^*\|.$$

Since $H(x^*) = 0$, (4.2) implies $\|b^*\| \le \|H_K(x) - H_K(x^*)\| + \|B(x - x^*)\| \le (L + \|B\|)\|x - x^*\|$, and we conclude

$$\|z - x^*\| \le s\gamma_0(L + \|B\|)\|x - x^*\|. \tag{4.4}$$

Hence $z \in B[x^*; r/2]$ cannot lie on the boundary of N_1; in particular, (4.3) holds. This proves the theorem. ∎

5. The local exactness proof. We now consider the optimization problem (2.1), where we now allow f and F to be nonsmooth functions. Clearly, with the embedding of section 3 and a fixed $w \in \mathbb{R}^m$, (2.1) is equivalent to

$$\begin{array}{ll} \min & f(x,\varepsilon) \\ \text{s.t.} & F_i(x,\varepsilon) = \varepsilon w_i \quad (i = 1,\dots,m), \\ & x \in [u,v], \quad \varepsilon = 0. \end{array} \tag{5.1}$$

We assume that f and F are continuous on $D \times [0, \bar{\varepsilon}]$, where D is an open set containing $[u,v]$ and $\bar{\varepsilon} > 0$, and twice continuously differentiable on its interior.

We use again the expression (2.3) for the penalty function but with $f(x,\varepsilon)$ in place of $f(x)$ and the regularized constraint violation measure

$$\Delta(x,\varepsilon) := \|\varepsilon w - F(x,\varepsilon)\|^2. \tag{5.2}$$

If, in addition to the assumptions mentioned after (2.4), β is twice continuously differentiable for $\varepsilon > 0$, the function f_σ is twice continuously differentiable for $(x,\varepsilon) \in (0,\bar{\varepsilon}) \times [u,v]$ with $\Delta(x,\varepsilon) < q^{-1}$.

It is conceivable that for this penalty function, a suitable analogue of Theorem 2.1 holds even in the nonsmooth case, but this requires an extension of the Kuhn–Tucker optimality conditions to nonsmooth problems, and we have not tried to handle the technicalities associated with this. On the other hand, we show here (after some preparation) that a converse of Theorem 2.1 can be proved in the nonsmooth case.

Assumptions. For the formal analysis, we shall suppose that, in addition to the general assumptions made above, the following assumptions (H_f), (H_F), and (H_ε) are satisfied in a neighborhood of some local (or global) minimizer $x = x^*$ of (5.1). Let $I := \{i \mid u_i < x_i^* < v_i\}$; to simplify notation we assume that $I = \{1,\dots,p\}$ with $m \le p \le n$.

(H_f) $f(\cdot,0)$ is Lipschitz continuous with the Lipschitz constant k.

(H_F) x^* is a regular zero of $F(\cdot,0)$ and x_I^* is a regular zero of $G : D_1 \subseteq \mathbb{R}^p \to \mathbb{R}^m$, defined by $G(\tilde{x}) := F(x,0)$, where $x_i = \tilde{x}_i$ for $i = 1,\dots,p$ and $x_i = x_i^*$ for $i = p+1,\dots,n$ and D_1 is an appropriate open set containing $[u_I, v_I]$. N is a neighborhood of x^* according to Definition 4.1 such that, in addition, $F(\cdot,0)$ is Lipschitz continuous in N and $x_I \in [u_I, v_I]$ for $x \in N$.

(H_ε) There are positive constants $\bar{\varepsilon}$ and K such that for all $x \in N$ and all $\varepsilon \in [0, \bar{\varepsilon}]$,

$$\|F(x,0) - F(x,\varepsilon)\|_\infty \le K\varepsilon, \quad |f(x,0) - f(x,\varepsilon)| \le K\varepsilon.$$

(H_β) β satisfies $\liminf_{\varepsilon\to 0} \beta(\varepsilon)/\sqrt{\varepsilon} > 0$.

LEMMA 5.1. *x^* is a regular zero of $H : D \subseteq \mathbb{R}^n \to \mathbb{R}^{m+n-p}$, defined by $H_i(x) := F_i(x,0)$ for $i = 1,\dots,m$ and $H_i(x) := x_{i-m+p} - x^*_{i-m+p}$ for $i = m+1,\dots,m+n-p$, if and only if x^* is a regular zero of $F(\cdot,0)$ and x_I^* is a regular zero of the mapping G defined in* (H_F).

Proof. Let $x, y \in N$, where N is an appropriate neighborhood of x^* according to Definition 4.1. In both cases we have $F(x,0) - F(y,0) = A(x-y)$ for a matrix $A \in \mathcal{A}$, where $\mathcal{A}$ is a closed, convex, and bounded set of $m \times n$ matrices such that the augmented matrix $\binom{A}{B}$ is nonsingular for all $A \in \mathcal{A}$ for some matrix $B \in \mathbb{R}^{(n-m)\times n}$. Then

$$H(x) - H(y) = \begin{pmatrix} A \\ A' \end{pmatrix}(x-y),$$

where $A' := (0 \;\; I_{n-p}) \in \mathbb{R}^{(n-p)\times n}$, and

$$G(x_I) - G(y_I) = A_{:I}(x_I - y_I).$$

Let x^* be a regular zero of H and let $B' \in \mathbb{R}^{(p-m)\times n}$ be such that

$$\begin{pmatrix} A \\ A' \\ B' \end{pmatrix}$$

is nonsingular for all $A \in \mathcal{A}$. Then $\binom{A_{:I}}{B'_{:I}}$ is nonsingular, i.e., x_I^* is a regular zero of G. Conversely, let x_I^* be a regular zero of G and let $B'_{:I}$ be such that $\binom{A_{:I}}{B'_{:I}}$ is nonsingular. Then

$$\begin{pmatrix} A \\ A' \\ B'' \end{pmatrix}$$

is nonsingular with $B'' = (B'_{:I} \;\; 0) \in \mathbb{R}^{(p-m)\times n}$. Therefore x^* is a regular zero of H. ∎

LEMMA 5.2. *If* (H_f) *and* (H_F) *hold, there are a neighborhood $N_0 \subseteq N$ of x^* and a constant $\gamma_1 > 0$ such that*

$$f(x,0) \ge f(x^*,0) - \gamma_1\|F(x,0)\| \quad \textit{for all } x \in N_0 \cap [u,v].$$

Proof. By (H_F), x^* is a regular zero of the mapping H defined in Lemma 5.1. Let $N_0 \subseteq N$ and γ_0 be as in Theorem 4.5, and let $x \in [u, v] \cap N_0$. Then by Theorem 4.5 with $J := \{i \mid F_i(x,0) \ne 0\}$ there exists a $y = y(x)$ with $H_J(y) = 0$ such that

$$\|x - y\| \le \gamma_0\|H_J(x)\| = \gamma_0\|F(x,0)\|.$$

The fact that $y_i = x_i$ for $i = p+1, \dots, n$ and the choice of N in (H_F) imply that $y \in [u, v]$; i.e., y is a feasible point. We therefore have $f(y, 0) \geq f(x^*, 0)$ by assumption, and

$$\begin{aligned} f(x,0) &= f(x,0) - f(y,0) + f(y,0) \geq f(x^*,0) - k\|x-y\| \\ &\geq f(x^*,0) - k\gamma_0\|F(x,0)\|, \end{aligned}$$

which completes the proof. ∎

Now we can prove the main theorem of this section.

THEOREM 5.3. *Under the above assumptions and for sufficiently large σ, there are a neighborhood N' of x^* and an $\varepsilon' \in (0, \bar{\varepsilon}]$ such that*

$$f_\sigma(x,\varepsilon) > f_\sigma(x^*,0) = f(x^*,0) \quad \text{for all } (x,\varepsilon) \in N' \times (0,\varepsilon'].$$

In particular, $(x^*, 0)$ is a local minimizer of f_σ.

Proof. Let the neighborhood $N' \subseteq N_0$ of x^* (N_0 as in Lemma 5.2) be sufficiently small such that

$$\sup_{x \in N'} (f(x^*,0) - f(x,0)) \leq \frac{1}{2}; \tag{5.3}$$

let $\varepsilon' \in (0, \bar{\varepsilon}]$, $\varepsilon' \leq 1$, be sufficiently small such that

$$\beta(\varepsilon) \geq \beta_2 \sqrt{\varepsilon} \tag{5.4}$$

for $0 \leq \varepsilon \leq \varepsilon'$ and a $\beta_2 > 0$; and let

$$\sigma \geq \frac{1}{\beta_2}(K(\gamma_1 + 1) + \gamma_1(\|w\| + 1)). \tag{5.5}$$

For $(x, \varepsilon) \in N' \times (0, \varepsilon']$ we distinguish two cases.

Case 1. Let $\Delta(x, \varepsilon) \geq \varepsilon$. Then

$$\begin{aligned} f_\sigma(x,\varepsilon) &\geq f(x,\varepsilon) + \frac{1}{2} + \sigma\beta(\varepsilon) \\ &\geq f(x^*,0) - \frac{1}{2} - K\varepsilon + \frac{1}{2} + \sigma\beta(\varepsilon) \\ &\geq f(x^*,0) - K\sqrt{\varepsilon} + \sigma\beta_2\sqrt{\varepsilon} > f(x^*,0), \end{aligned}$$

where we have used (H_ε), (5.3), (5.4), (5.5), and the fact that $\varepsilon \leq \varepsilon' \leq 1$.

Case 2. If $\Delta(x, \varepsilon) < \varepsilon$, then $\|F(x,\varepsilon)\| < \varepsilon\|w\| + \|\varepsilon w - F(x,\varepsilon)\| \leq \varepsilon\|w\| + \sqrt{\varepsilon}$; hence by Lemma 5.2 and (H_ε),

$$\begin{aligned} f(x^*,0) &\leq f(x,0) + \gamma_1\|F(x,0)\| \\ &\leq f(x,\varepsilon) + K\varepsilon + \gamma_1(\|F(x,\varepsilon)\| + K\varepsilon) \\ &< f(x,\varepsilon) + K(\gamma_1+1)\varepsilon + \gamma_1(\sqrt{\varepsilon} + \varepsilon\|w\|). \end{aligned}$$

Therefore $f_\sigma(x,\varepsilon) \geq f(x,\varepsilon) + \sigma\beta(\varepsilon) > f(x^*,0)$ by (5.4), (5.5), and $\varepsilon \leq \varepsilon' \leq 1$. ∎

6. An example. To illustrate the theory developed, we consider the simple smooth nonlinear optimization problem

$$\begin{aligned} \min \quad & x_1^3 x_2^3 \\ \text{s.t.} \quad & x_1^2 + x_2^2 = 1. \end{aligned}$$

It has a bounded feasible domain, two global minimizers at $x^* = (\frac{1}{2}\sqrt{2}, -\frac{1}{2}\sqrt{2})^T$ and $x^{**} = (-\frac{1}{2}\sqrt{2}, \frac{1}{2}\sqrt{2})^T$ with $f(x^*) = f(x^{**}) = -\frac{1}{8}$, and no other local minima. The traditional quadratic penalty function for this problem,

$$p(x) = x_1^3 x_2^3 + \frac{1}{2\varepsilon}(x_1^2 + x_2^2 - 1)^2,$$

is unbounded below for all $\varepsilon > 0$ since, e.g., $p(x) \to -\infty$ for $x = (s, -s)^T$, $s \to \infty$. This is also the case for traditional exact penalty functions, including multiplier penalty functions [1] that use an additional term $+\lambda(x_1^2 + x_2^2 - 1)$. On the other hand, our new penalty function is bounded below. For $w = 1$ it reads

$$f_\sigma(x, \varepsilon) = \begin{cases} x_1^3 x_2^3 & \text{for } \varepsilon = \Delta(x, \varepsilon) = 0, \\ x_1^3 x_2^3 + \dfrac{1}{2\varepsilon} \cdot \dfrac{r^2}{1 - qr^2} + \sigma\beta(\varepsilon) & \text{for } \varepsilon > 0,\ |r| < q^{-1/2}, \\ \infty & \text{otherwise,} \end{cases}$$

where $r := 1 + \varepsilon - x_1^2 - x_2^2$. Since $f_\sigma(x, \varepsilon) = \infty$ if $\|x\| \geq \sqrt{q^{-1/2} + 1 + \varepsilon}$, the boundedness of our penalty function below is trivial. The Mangasarian–Fromovitz condition holds for all $x \neq 0$; hence the assumptions of Theorem 2.1 are satisfied if $q^{-1/2} + \bar{\varepsilon} < 1$ and (2.8) holds. In this case, every local minimizer of the penalty function with a sufficiently large σ gives a solution of the original constrained problem.

In this particular example we can give an explicit analysis, and find that in fact weaker assumptions than those given in Theorem 2.1 suffice to get the conclusions, since $\bar{\varepsilon}$ can be chosen arbitrarily large. To show this, let (x, ε) be a Kuhn–Tucker point of f_σ with $\varepsilon > 0$ and $|r| < q^{-1/2}$. Then $\partial f_\sigma / \partial x_1$ and $\partial f_\sigma / \partial x_2$ vanish at (x, ε), and $\partial f_\sigma(x, \varepsilon)/\partial\varepsilon \leq 0$, with equality if $\varepsilon < \bar{\varepsilon}$. We have

$$\begin{aligned} \frac{\partial f_\sigma}{\partial x_1}(x, \varepsilon) &= x_1\Big(3x_1 x_2^3 - \frac{2r}{\varepsilon(1 - qr^2)^2}\Big), \\ \frac{\partial f_\sigma}{\partial x_2}(x, \varepsilon) &= x_2\Big(3x_1^3 x_2 - \frac{2r}{\varepsilon(1 - qr^2)^2}\Big), \\ \frac{\partial f_\sigma}{\partial \varepsilon}(x, \varepsilon) &= \frac{r(2\varepsilon - r + qr^3)}{2\varepsilon^2(1 - qr^2)^2} + \sigma\beta'(\varepsilon). \end{aligned}$$

Since $\sigma\beta'(\varepsilon) > 0$, $\partial f_\sigma / \partial\varepsilon \leq 0$ implies

$$r(2\varepsilon - r + qr^3) < 0; \tag{6.1}$$

in particular $r \neq 0$. Under this condition, the other partial derivatives vanish if and only if either $x_1 = x_2 = 0$ or

$$3x_1 x_2^3 = 3x_1^3 x_2 = \frac{2r}{\varepsilon(1 - qr^2)^2}. \tag{6.2}$$

If (6.2) holds, then $x_1 = \pm x_2$, the definition of r gives $1 + \varepsilon - r = 2x_1^2$, and (6.2) simplifies to

$$8|r| = 3\varepsilon(1 - qr^2)^2(1 + \varepsilon - r)^2. \tag{6.3}$$

If $r > 0$, then $r > 2\varepsilon + qr^3 > 2\varepsilon$ by (6.1), giving $0 \le 1 + \varepsilon - r < 1 - \varepsilon < 1$. Now (6.3) implies $8r \le 3\varepsilon$, contradicting $r > 2\varepsilon$. And if $r < 0$, then (6.3) and $qr^2 < 1$ imply for $q \ge 1$ that

$$\frac{-r}{2\varepsilon(1 - qr^2)^2} = \frac{3}{16}(1 + \varepsilon + |r|)^2 \le \frac{3}{16}(2 + \bar{\varepsilon}) =: t$$

and

$$2\varepsilon - r + qr^3 \le 2\varepsilon - r \le (2 + 2t)\varepsilon.$$

Now $\partial f_\sigma/\partial\varepsilon \le 0$ gives $\sigma\beta'(\varepsilon) \le t(2 + 2t)$. If (2.8) holds and $\sigma > 2t(t + 1)/\beta_1$, this is violated, (6.2) is impossible, and the only Kuhn–Tucker point of f_σ with $\varepsilon > 0$ can be at $x_1 = x_2 = 0$. But then $r = 1 + \varepsilon$ and (6.1) requires $q < (1 - \varepsilon)/(1 + \varepsilon)^3$, again impossible if $q \ge 1$.

Thus, for this example, the conclusion of Theorem 2.1 is valid for arbitrary $\bar{\varepsilon} > 0$ if (2.8) holds and $q \ge 1$, provided that σ is sufficiently large. Note that for $q \ge 1$ we have $\Delta(0, \varepsilon) = (1+\varepsilon)^2 > q^{-1}$, so that the only point violating the Mangasarian–Fromovitz condition does not satisfy $\Delta(x, \varepsilon) < q^{-1}$.

We now look at the converse. Theorem 5.3 gives conditions guaranteeing that every solution of the constrained problem is a local minimizer of f_σ. To verify this explicitly we need to investigate when it is possible that f_σ is below the common value $-\frac{1}{8}$ of $f_\sigma(x^*, 0)$ and $f_\sigma(x^{**}, 0)$. Thus suppose that $f_\sigma(x, \varepsilon) \le -\frac{1}{8}$, $(x, \varepsilon) \ne (x^*, 0), (x^{**}, 0)$. Then $\varepsilon > 0$. Since

$$x_1^3 x_2^3 \ge -\frac{1}{8}(x_1^2 + x_2^2)^3 = -\frac{1}{8}(1 + \varepsilon - r)^3 \ge -\frac{1}{8}(1 + \varepsilon + |r|)^3,$$

we find

$$8\sigma\beta(\varepsilon) \le -1 + (1 + \varepsilon + |r|)^3 - 4r^2/\varepsilon =: p(|r|), \quad |r| < q^{-1/2}. \tag{6.4}$$

Now $p(r)$ is a cubic polynomial. If $\varepsilon < \varepsilon_0 := \frac{1}{6}(-3 + \sqrt{33}) = 0.45742\ldots$, the positive solution of $\varepsilon_0(1 + \varepsilon_0) = \frac{2}{3}$, then $\delta := 2 - 3\varepsilon(1 + \varepsilon) > 0$, and $p(r)$ has a unique local maximum at

$$r_0 = \frac{3\varepsilon(1 + \varepsilon)^2}{2 + \delta + \sqrt{8\delta}} \le \frac{3}{2}\varepsilon(1 + \varepsilon)^2 \le \frac{3}{2}\varepsilon(1 + \varepsilon_0)^2 = \varepsilon(1 + \varepsilon_0^{-1})$$

with function value

$$p(r_0) \le -1 + (1 + \varepsilon + r_0)^3 \le -1 + (1 + \varepsilon(2 + \varepsilon_0^{-1}))^3 \le \frac{-1 + (2 + 2\varepsilon_0)^3}{\varepsilon_0}\varepsilon < 52\varepsilon.$$

Here we used the fact that $(-1 + (1 + \varepsilon(2 + \varepsilon_0^{-1}))^3)/\varepsilon$ is monotone increasing. Now

$$p(|r|) \le \max(p(r_0), p(q^{-1/2})) < 52\varepsilon \quad \text{for } |r| < q^{-1/2}$$

if $p(q^{-1/2}) < 0$, which holds for any fixed q if ε is small enough. Thus (6.4) is violated if $\sigma\beta(\varepsilon) > 6.5\varepsilon$ and ε is small enough. If (H_β) holds, this is the case for arbitrary σ and sufficiently small $\varepsilon > 0$. Thus we have a contradiction, and the conclusion of Theorem 5.3 holds for arbitrary σ. We also see that in this example the conclusion of Theorem 5.3 still holds if the condition $\liminf_{\varepsilon\to 0} \beta(\varepsilon)/\sqrt{\varepsilon} > 0$ in (H_β) is relaxed to $\liminf_{\varepsilon\to 0} \beta(\varepsilon)/\varepsilon > 0$ and σ is chosen sufficiently large.

7. More general constraints. In this section we extend our theory to the more general constrained optimization problem

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & x \in [u, v], \quad F_l \le F(x) \le F_u, \end{array} \tag{7.1}$$

where $F_l \in (\mathbb{R} \cup \{-\infty\})^m$ and $F_u \in (\mathbb{R} \cup \{\infty\})^m$ are vectors containing proper or infinite bounds on the constraint functions and $F_l \le F_u$. This formulation is quite general since equality constraints are allowed by taking equal lower and upper bounds, and one-sided inequalities by taking infinite lower or upper bounds. Thus we treat equality constraints and one- or two-sided inequality constraints on the same footing. Moreover, we again assume that f and F are embedded into families of regularized functions $f(x, \varepsilon)$ and $F(x, \varepsilon)$ with $f(x, 0) = f(x)$ and $F(x, 0) = F(x)$.

For the one-dimensional boxes (which are just the closed intervals) we define the *mignitude* [12]

$$\langle \mathbf{x} \rangle := \min\{|x| \mid x \in \mathbf{x}\} = \begin{cases} \underline{x} & \text{if } \underline{x} > 0, \\ -\overline{x} & \text{if } \overline{x} < 0, \\ 0 & \text{otherwise.} \end{cases}$$

We also need the simple interval operations

$$\alpha + \beta\mathbf{x} = \{\alpha + \beta x \mid x \in \mathbf{x}\} = \begin{cases} [\alpha + \beta\underline{x}, \alpha + \beta\overline{x}] & \text{if } \beta \ge 0, \\ [\alpha + \beta\overline{x}, \alpha + \beta\underline{x}] & \text{if } \beta \le 0. \end{cases}$$

Using this interval notation, we introduce a box-valued function $\mathbf{E}$ on $[u, v] \times [0, \bar{\varepsilon}]$ by

$$\mathbf{E} : (x, \varepsilon) \to \mathbf{E}(x, \varepsilon) := F(x, \varepsilon) - [F_l, F_u], \tag{7.2}$$

where we assume that (H_ε) is satisfied, and consider the following optimization problem in inclusion form:

$$\begin{array}{ll} \min & f(x, \varepsilon) \\ \text{s.t.} & \varepsilon w_i \in \mathbf{E}_i(x, \varepsilon) \quad (i = 1, \ldots, m), \\ & x \in [u, v], \quad \varepsilon = 0, \end{array} \tag{7.3}$$

where again $w \in \mathbb{R}^m$ is fixed. Clearly, this formulation is equivalent to the optimization problem (7.1). The usefulness of this particular formulation will become apparent when we look at the associated penalty function.

The penalty function is again defined by (2.3) with $f(x, \varepsilon)$ in place of $f(x)$, but now the constraint violation measure $\Delta(x, \varepsilon)$ is of the form

$$\Delta(x, \varepsilon) := \sum_{i=1}^{m} \Delta_i(x, \varepsilon) \tag{7.4}$$

with

$$\Delta_i(x,\varepsilon) := \langle \varepsilon w_i - \mathbf{E}_i(x,\varepsilon)\rangle^2, \tag{7.5}$$

the squared distance of εw_i from the interval $\mathbf{E}_i(x,\varepsilon)$. Without loss of generality we assume that the constraints are inequality constraints for $i = 1,\dots,r$ and equality constraints for $i = r+1,\dots,m$, where $1 \le r \le m$. Then

$$\Delta_i(x,\varepsilon) = (\varepsilon w_i - (F_i(x,\varepsilon) - F_{li}))^2 \quad \text{for } i = r+1,\dots,m$$

and

$$\Delta_i(x,\varepsilon) = (\varepsilon w_i - (F_i(x,\varepsilon) - y_i))^2 \quad \text{for } i = 1,\dots,r, \tag{7.6}$$

where

$$y_i = \begin{cases} F_i(x,\varepsilon) - \varepsilon w_i & \text{if } F_i(x,\varepsilon) - \varepsilon w_i \in [F_{li}, F_{ui}], \\ F_{li} & \text{if } F_i(x,\varepsilon) - \varepsilon w_i < F_{li}, \\ F_{ui} & \text{if } F_i(x,\varepsilon) - \varepsilon w_i > F_{ui}. \end{cases} \tag{7.7}$$

Moreover, $\Delta(x,\varepsilon)$ is continuously differentiable for $\varepsilon > 0$, with

$$\frac{\partial}{\partial x}\Delta_i(x,\varepsilon) = -2(\varepsilon w_i - F_i(x,\varepsilon) + y_i)\frac{\partial F_i}{\partial x}(x,\varepsilon)$$

and

$$\frac{\partial}{\partial \varepsilon}\Delta_i(x,\varepsilon) = 2(\varepsilon w_i - F_i(x,\varepsilon) + y_i)\left(w_i - \frac{\partial F_i}{\partial \varepsilon}(x,\varepsilon)\right)$$

for $i = 1,\dots,r$.

We reduce this more general situation to the previous one with the aid of slack variables and use the abbreviations $J := \{1,\dots,r\}$ and $J' := \{r+1,\dots,m\}$. By introducing the slack variables $y_i := F_i(x,\varepsilon)$ for $i = 1,\dots,r$ we obtain the problem

$$\begin{array}{ll} \min & f(x,\varepsilon) \\ \text{s.t.} & x \in [u,v], \quad y \in [F_{lJ}, F_{uJ}], \quad \varepsilon = 0, \\ & F_J(x,\varepsilon) - y = \varepsilon w_J, \quad F_{J'}(x,\varepsilon) - F_{lJ'} (= F_{J'}(x,\varepsilon) - F_{uJ'}) = \varepsilon w_{J'}, \end{array} \tag{7.8}$$

which is of the form (5.1). The penalty function for this problem is given by

$$\tilde f_\sigma(x,y,\varepsilon) = \begin{cases} f(x) & \text{for } \varepsilon = \tilde\Delta(x,y,\varepsilon) = 0, \\ f(x,\varepsilon) + \dfrac{1}{2\varepsilon}\cdot\dfrac{\tilde\Delta(x,y,\varepsilon)}{1 - q\tilde\Delta(x,y,\varepsilon)} + \sigma\beta(\varepsilon) & \text{for } \varepsilon > 0,\ \tilde\Delta(x,y,\varepsilon) < q^{-1}, \\ \infty & \text{otherwise,} \end{cases}$$

where

$$\tilde\Delta(x,y,\varepsilon) = \sum_{i=1}^{r}(\varepsilon w_i - (F_i(x,\varepsilon) - y_i))^2 + \sum_{i=r+1}^{m}(\varepsilon w_i - (F_i(x,\varepsilon) - F_{li}))^2.$$

Let $\tilde E_i(x,y) := F_i(x) - y_i$, $i = 1,\dots,r$, and $\tilde E_i(x,y) := F_i(x) - F_{li}$, $i = r+1,\dots,m$. Then

$$\tilde E'(x,y) = \begin{pmatrix} F_J'(x) & -I_r \\ F_{J'}'(x) & 0 \end{pmatrix},$$

and, for $m \le n + r$, $\tilde{E}'(x, y)$ has full rank if and only if $F'_{J'}(x)$ has full rank.

LEMMA 7.1. *Let* $(x, \varepsilon) \in [u, v] \times [0, \bar{\varepsilon}]$ *with* $\Delta(x, \varepsilon) < q^{-1}$. *Then*

$$\Delta(x, \varepsilon) = \min_{y \in [F_{lJ}, F_{uJ}]} \tilde{\Delta}(x, y, \varepsilon) \tag{7.9}$$

and thus

$$f_\sigma(x, \varepsilon) = \min_{y \in [F_{lJ}, F_{uJ}]} \tilde{f}_\sigma(x, y, \varepsilon).$$

Moreover, if (x, ε) *is a Kuhn–Tucker point of* f_σ, *then there exists a vector* $y \in [F_{lJ}, F_{uJ}]$ *such that* (x, y, ε) *is a Kuhn–Tucker point of* $\tilde{f}_\sigma$.

Proof. Let $(x, \varepsilon) \in [u, v] \times [0, \bar{\varepsilon}]$ with $\Delta(x, \varepsilon) < q^{-1}$, and let y be defined by (7.7). Then, for $i = r + 1, \dots, m$ we have $\Delta_i(x, \varepsilon) = (\varepsilon w_i - (F_i(x, \varepsilon) - F_{li}))^2$, and $\Delta_i(x, \varepsilon)$ is given by (7.6) and (7.7) for $i = 1, \dots, r$, i.e., $\Delta(x, \varepsilon) = \tilde{\Delta}(x, y, \varepsilon)$ for y defined by (7.7) and clearly (7.9) holds. Moreover, $\frac{\partial}{\partial x}\Delta(x, \varepsilon) = \frac{\partial}{\partial x}\tilde{\Delta}(x, y, \varepsilon)$ and $\frac{\partial}{\partial \varepsilon}\Delta(x, \varepsilon) = \tilde{\Delta}(x, y, \varepsilon)$. If (x, ε) is a Kuhn–Tucker point of f_σ, (x, y, ε) thus fulfills the Kuhn–Tucker conditions for x and ε, and due to the definition of y and the fact that

$$\frac{\partial \tilde{f}_\sigma}{\partial y_i}(x, y, \varepsilon) = \frac{1}{\varepsilon(1 - q\tilde{\Delta}(x, y, \varepsilon))^2}(\varepsilon w_i - F_i(x, \varepsilon) + y_i)$$

it also satisfies the Kuhn–Tucker conditions with respect to y. ∎

The Mangasarian–Fromovitz condition for this problem holds for $(x, y) \in [u, v] \times [F_{lJ}, F_{uJ}]$ if

$$\begin{aligned}
&F'_{J'}(x) \text{ has full rank, and there is a } p \in \mathbb{R}^n \text{ with } F'_{J'}(x)p = 0,\\
&p_i \begin{cases} > 0 & \text{if } x_i = u_i,\\ < 0 & \text{if } x_i = v_i,\end{cases}\\
&(F'_J(x)p)_i \begin{cases} > 0 & \text{if } y_i = F_{li},\\ < 0 & \text{if } y_i = F_{ui}.\end{cases}
\end{aligned}$$

Let $I := \{i \mid u_i < x^*_i < v_i\}$ and $J_1 := \{i \in J \mid F_i(x^*, 0) = F_{il} \text{ or } F_i(x^*, 0) = F_{iu}\}$; without loss of generality $I := \{1, \dots, p\}$. Moreover, let $E(x) := F(x, 0) - F(x^*, 0)$, $x \in D$. Then (H_F) is replaced by the following.

(H_E) E is Lipschitz continuous in a neighborhood of x^*, x^* is a regular zero of $E_{J'} : D \subseteq \mathbb{R}^n \to \mathbb{R}^{m-r}$, and x^*_I is a regular zero of $G : D_1 \subseteq \mathbb{R}^p \to \mathbb{R}^{m-r+r_1}$, $r_1 := |J_1|$, defined by $G(\tilde{x}) := E_{J_1 \cup J'}(x)$ with $x_i := \tilde{x}_i$ for $i = 1, \dots, p$ and $x_i := x^*_i$ for $i = p+1, \dots, n$, where D_1 is an appropriate open set containing $[u_I, v_I]$.

THEOREM 7.2. *If the above Mangasarian–Fromovitz condition holds for all* $(x, y) \in [u, v] \times [F_{lJ}, F_{uJ}]$ *with*

$$\sum_{i=1}^{r}(F_i(x) - y_i)^2 + \sum_{i=r+1}^{m}(F_i(x) - F_{li})^2 \le (q^{-1/2} + \bar{\varepsilon}\|w\|)^2,$$

the set of these (x, y) *is bounded, and* (2.8) *is satisfied, then the conclusions of Theorem* 2.1 *hold for the smooth case of the penalty function defined in this section. Moreover, under the assumptions* (H_f), (H_E), (H_ε), *and* (H_β) *the conclusions of Theorem* 5.3 *are satisfied for the penalty function defined by* (2.3) *with* $f(x)$ *replaced by* $f(x, \varepsilon)$ *and* Δ *defined by* (7.4) *and* (7.5).

REFERENCES

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

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

[3] A. R. Conn, N. I. M. Gould, and Ph. L. Toint, *Methods for nonlinear constraints in optimization calculations*, in The State of the Art in Numerical Analysis, I. S. Duff and G. A. Watson, eds., Clarendon Press, Oxford, 1997, pp. 363–390.

[4] Y. M. Ermoliev, A. V. Kryazhimskii, and A. Ruszczyński, *Constraint aggregation principle in convex optimization*, Math. Program., 76 (1997), pp. 353–372.

[5] K. Kiwiel, *Proximity control in bundle methods for convex nondifferentiable minimization*, Math. Program., 46 (1990), pp. 105–122.

[6] A. V. Kuntsevich and F. Kappel, *SolvOpt—The Solver for Local Nonlinear Optimization Problems* (*Version* 1.1, *Matlab, C, FORTRAN*), University of Graz, Graz, Austria, http://www.uni-graz.at/imawww/kuntsevich/solvopt/ (1997).

[7] C. Lemaréchal, A. Nemirovskii, and Y. Nesterov, *New variants of bundle methods*, Math. Program., 69 (1995), pp. 111–147.

[8] C. Lemaréchal and C. Sagastizábal, *Variable metric bundle methods*: *From conceptual to implementable forms*, Math. Program., 76 (1997), pp. 393–410.

[9] C. Lemaréchal and C. Sagastizábal, *Practical aspects of the Moreau–Yosida regularization*: *Theoretical preliminaries*, SIAM J. Optim., 7 (1997), pp. 367–385.

[10] C. Lemaréchal, J. J. Strodiot and A. Bihain, *On a bundle method for nonsmooth optimization*, in Nonlinear Programming 4, O. L. Mangasarian, R. R. Meyer, and S. M. Robinson, eds., Academic Press, New York, 1981, pp. 245–282.

[11] 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.

[12] A. Neumaier, *Interval Methods for Systems of Equations*, Cambridge University Press, Cambridge, UK, 1990.

[13] B. T. Polyak, *A general method of solving extremum problems*, Sov. Math. Dokl., 8 (1967), pp. 593–597 (transl. from Dokl. Akad. Nauk SSSR, 174 (1967), pp. 33–36).

[14] L. Qi and X. Chen, *A preconditioning proximal Newton method for nondifferentiable convex optimization*, Math. Program., 76 (1997), pp. 411–429.

[15] D. F. Shanno and E. M. Simantiraki, *Interior point methods for linear and nonlinear programming*, in The State of the Art in Numerical Analysis, I. S. Duff and G. A. Watson, eds., Clarendon Press, Oxford, 1997, pp. 339–362.

SIAM J. OPTIM.
Vol. 13, No. 4, pp. 1159–1173

ITERATING BREGMAN RETRACTIONS*

HEINZ H. BAUSCHKE† AND PATRICK L. COMBETTES‡

Abstract. The notion of a Bregman retraction of a closed convex set in Euclidean space is introduced. Bregman retractions include backward Bregman projections and forward Bregman projections, as well as their convex combinations, and are thus quite flexible. The main result on iterating Bregman retractions unifies several convergence results on projection methods for solving convex feasibility problems. It is also used to construct new sequential and parallel algorithms.

Key words. backward Bregman projection, Bregman distance, Bregman function, Bregman projection, Bregman retraction, convex feasibility problem, forward Bregman projection, Legendre function, paracontraction, projection algorithm

AMS subject classifications. 90C25, 49M37, 65K05

PII. S1052623402410557

1. Standing assumptions, problem statement, and motivation.

We assume throughout this paper that

$$X \text{ is a Euclidean space with scalar product } \langle \cdot, \cdot \rangle \text{ and induced norm } \| \cdot \| \tag{1.1}$$

and that

$$\begin{aligned} &f\colon X \to \left]-\infty, +\infty\right] \text{ is a proper closed convex Legendre function} \\ &\text{such that } \operatorname{dom} f^* \text{ is open,} \end{aligned} \tag{1.2}$$

where f^* denotes the conjugate of f. Recall that a function is Legendre if it is both essentially smooth and essentially strictly convex (see, e.g., [31] for basic facts and notions from convex analysis). In addition, we assume that

$$\begin{aligned} &(C_i)_{i\in I} \text{ are finitely many closed convex sets in } X \\ &\text{such that } (\operatorname{int}\operatorname{dom} f) \cap \bigcap_{i\in I} C_i \neq \varnothing. \end{aligned} \tag{1.3}$$

Our aim is to study algorithms for solving the fundamental *convex feasibility problem* (see [4], [14], [17], [20], and [27] for further information and references)

$$\text{find } x \in \bigcap_{i\in I} C_i. \tag{1.4}$$

Assumption (1.2) guarantees that we capture a large class of functions (see Example 2.1 below) for which the corresponding *Bregman distance*

$$D_f\colon X \times X \to [0, +\infty] : (x, y) \mapsto \begin{cases} f(x) - f(y) - \langle x - y, \nabla f(y)\rangle & \text{if } y \in \operatorname{int}\operatorname{dom} f; \\ +\infty & \text{otherwise} \end{cases} \tag{1.5}$$

*Received by the editors July 2, 2002; accepted for publication October 31, 2002; published electronically April 30, 2003.
http://www.siam.org/journals/siopt/13-4/41055.html

†Department of Mathematics and Statistics, University of Guelph, Guelph, Ontario N1G 2W1, Canada (hbauschk@uoguelph.ca). The research of this author was supported by the Natural Sciences and Engineering Research Council of Canada.

‡Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie – Paris 6, 75005 Paris, France (plc@math.jussieu.fr).

enjoys useful properties (Proposition 2.2). This type of directed distance was first introduced by Bregman in [8]; see [17] for a historical account. Now fix a closed convex set C in X such that $C \cap \operatorname{int}\operatorname{dom} f \neq \emptyset$ and a point $y \in \operatorname{int}\operatorname{dom} f$. Then there is a unique point in $C \cap \operatorname{int}\operatorname{dom} f$, called the *backward Bregman projection* (or simply the *Bregman projection*) of y onto C and denoted by $\overleftarrow{P}_C y$, which satisfies (Fact 2.3)

$$(\forall c \in C) \quad D_f(\overleftarrow{P}_C y, y) \leq D_f(c, y). \tag{1.6}$$

Moreover, if f *allows forward Bregman projections* (Definition 2.4), then there is analogously a unique point in $C \cap \operatorname{int}\operatorname{dom} f$, called the *forward Bregman projection* of y onto C and denoted by $\overrightarrow{P}_C y$, which satisfies (Fact 2.6)

$$(\forall c \in C) \quad D_f(y, \overrightarrow{P}_C y) \leq D_f(y, c). \tag{1.7}$$

If $f = \frac{1}{2}\|\cdot\|^2$, then both $\overleftarrow{P}_C y$ and $\overrightarrow{P}_C y$ coincide with the *orthogonal projection* of y onto C; however, the backward and forward Bregman projections differ generally, due to the asymmetry of D_f.

With backward and forward Bregman projections in place, we now describe three projection methods for solving (1.4). To this end, fix an index selector map $\mathrm{i}\colon \mathbb{N} = \{0, 1, 2, \dots\} \to I$ that takes on each value in I infinitely often, and a starting point $y_0 \in \operatorname{int}\operatorname{dom} f$. The *method of backward Bregman projections* generates a sequence $(y_n)_{n\in\mathbb{N}}$ by

$$(\forall n \in \mathbb{N}) \quad y_{n+1} = \overleftarrow{P}_{C_{\mathrm{i}(n+1)}} y_n. \tag{1.8}$$

Analogously, if f allows forward Bregman projections, then the update rule for the *method of forward Bregman projections* is

$$(\forall n \in \mathbb{N}) \quad y_{n+1} = \overrightarrow{P}_{C_{\mathrm{i}(n+1)}} y_n. \tag{1.9}$$

Well-known *cyclic* versions arise if $I = \{1, \dots, N\}$ and $\mathrm{i}(n) = n \bmod N$, where the range of the mod function is assumed to be $\{1, \dots, N\}$. The sequence $(y_n)_{n\in\mathbb{N}}$ generated by (1.8) (or by (1.9), if f allows forward Bregman projections) is known to solve (1.4) asymptotically: indeed, $(y_n)_{n\in\mathbb{N}}$ converges to some point in $\bigcap_{i\in I} C_i$; see [5] and [16] (or [7], respectively).

The third algorithm is due to Byrne and Censor [12], who adapted Csiszár and Tusnády's classical alternating minimization procedure [22] to a product space setting (see also section 5). Their algorithm assumes two constraints, $I = \{1, 2\}$, and a sequence $(y_n)_{n\in\mathbb{N}}$ is generated using *alternating backward-forward Bregman projections*:

$$(\forall n \in \mathbb{N}) \quad y_{n+1} = \big(\overleftarrow{P}_{C_2} \circ \overrightarrow{P}_{C_1}\big) y_n. \tag{1.10}$$

They show that, under appropriate conditions, $(y_n)_{n\in\mathbb{N}}$ converges to some point in $C_1 \cap C_2$; see [12, Theorem 1].

The striking resemblance in the update rules of the three preceding algorithms motivates this paper. Our objective is to provide a unified convergence analysis of these algorithms using the notion of a *Bregman retraction*, which encompasses both backward and forward Bregman projections. The main theorem not only recovers known convergence results but also provides a theoretical basis for the application of new sequential and parallel methods.

It is instructive to contrast our Bregman retraction-based framework with Censor and Reich's [16] framework, which is built on *paracontractions* (Definition 3.11).

While backward Bregman projections are both Bregman retractions and paracontractions, the two notions differ in general; actually, Examples 3.12 and 3.13 show that neither framework contains the other.

The key advantage of the Bregman retraction-based framework presented here is its applicability: the conditions on f are mild and easy to check. Moreover, simple constraint qualifications guarantee that Bregman retractions—in the form of backward Bregman projections (and forward Bregman projections, if f allows them)—always exist.

The paper is organized as follows. Background material on Bregman distances and associated projections is included in section 2. In section 3, Bregman retractions are introduced, analyzed, and illustrated by examples. The main result is proved in section 4, and applications are presented in section 5.

2. Preliminary results. Below is a selection of functions satisfying our assumptions (see [5] for additional examples).

EXAMPLE 2.1 (see [5]). *Suppose that $X = \mathbb{R}^J$ and, for every $x \in X$, write $x = (\xi_j)_{j=1}^J$. Then the following functions satisfy* (1.2) *(here and elsewhere, we use the convention $0 \cdot \ln(0) = 0$):*

(i) $f\colon x \mapsto \frac{1}{2}\|x\|^2 = \frac{1}{2}\sum_{j=1}^J |\xi_j|^2$, *with* $\operatorname{dom} f = \mathbb{R}^J$ (energy);

(ii) $f\colon x \mapsto \sum_{j=1}^J \xi_j \ln(\xi_j) - \xi_j$, *with* $\operatorname{dom} f = [0, +\infty[^J$ (negative entropy);

(iii) $f\colon x \mapsto \sum_{j=1}^J \xi_j \ln(\xi_j) + (1-\xi_j)\ln(1-\xi_j)$, *with* $\operatorname{dom} f = [0,1]^J$ (Fermi–Dirac entropy);

(iv) $f\colon x \mapsto -\sum_{j=1}^J \ln(\xi_j)$, *with* $\operatorname{dom} f = \,]0, +\infty[^J$ (Burg entropy);

(v) $f\colon x \mapsto -\sum_{j=1}^J \sqrt{\xi_j}$, *with* $\operatorname{dom} f = [0, +\infty[^J$.

The assumptions imposed on f in (1.2) guarantee the following very useful properties of D_f.

PROPOSITION 2.2. *Let D_f be defined as in* (1.5). *Then we have the following:*

(i) *D_f is continuous on* $(\operatorname{int}\operatorname{dom} f)^2$.

(ii) *If $x \in \operatorname{dom} f$ and $y \in \operatorname{int}\operatorname{dom} f$, then $D_f(x,y) \geq 0$, and $D_f(x,y) = 0 \Leftrightarrow x = y$.*

(iii) *If $(x_n)_{n\in\mathbb{N}}$ and $(y_n)_{n\in\mathbb{N}}$ are two sequences in* $\operatorname{int}\operatorname{dom} f$ *converging to $x \in \operatorname{int}\operatorname{dom} f$ and $y \in \operatorname{int}\operatorname{dom} f$, respectively, then $D_f(x_n, y_n) \to 0 \Leftrightarrow x = y$.*

(iv) *If $x \in \operatorname{int}\operatorname{dom} f$ and $(y_n)_{n\in\mathbb{N}}$ is a sequence in* $\operatorname{int}\operatorname{dom} f$ *such that the sequence $(D_f(x, y_n))_{n\in\mathbb{N}}$ is bounded, then $(y_n)_{n\in\mathbb{N}}$ is bounded and all its cluster points belong to* $\operatorname{int}\operatorname{dom} f$.

(v) *If $x \in \operatorname{int}\operatorname{dom} f$ and $(y_n)_{n\in\mathbb{N}}$ is a sequence in* $\operatorname{int}\operatorname{dom} f$ *such that $D_f(x, y_n) \to 0$, then $y_n \to x$.*

Proof. (i) This follows from the definition of D_f and the continuity of f (respectively, ∇f) on $\operatorname{int}\operatorname{dom} f$; see [31, Theorem 10.1] (respectively, [31, Theorem 25.5]). (ii) [5, Theorem 3.7.(iv)]. (iii) This is a consequence of (i) and (ii). (iv) [5, Theorem 3.7.(vi) and Theorem 3.8.(ii)]. (v) (See also [7, Fact 2.18].) By (iv), $(y_n)_{n\in\mathbb{N}}$ is bounded and has all its cluster points in $\operatorname{int}\operatorname{dom} f$. Pick an arbitrary cluster point of $(y_n)_{n\in\mathbb{N}}$, say $y_{k_n} \to y \in \operatorname{int}\operatorname{dom} f$. Then $D_f(x, y_{k_n}) \to 0$ and thus $x = y$ by (iii). ∎

We now turn to backward and forward Bregman projections.

FACT 2.3 (backward Bregman projection). *Suppose that C is a closed convex set in X such that $C \cap \operatorname{int}\operatorname{dom} f \neq \emptyset$. Then, for every $y \in \operatorname{int}\operatorname{dom} f$, there exists a unique point $\overleftarrow{P}_C y \in C \cap \operatorname{dom} f$ such that $D_f(\overleftarrow{P}_C y, y) \leq D_f(c, y)$ for all $c \in C$. The point $\overleftarrow{P}_C y$ is called the* backward Bregman projection *(or simply the* Bregman

projection*) of y onto C, and it is characterized by*

$$(2.1)\quad \overleftarrow{P}_C \in C \cap \operatorname{int}\operatorname{dom} f \quad \text{and} \quad (\forall c \in C)\ \big\langle c - \overleftarrow{P}_C y, \nabla f(y) - \nabla f(\overleftarrow{P}_C y)\big\rangle \leq 0;$$

equivalently, it is characterized by

$$(2.2)\quad \overleftarrow{P}_C \in C \cap \operatorname{int}\operatorname{dom} f \quad \text{and} \quad (\forall c \in C)\ D_f(c,y) \geq D_f(c, \overleftarrow{P}_C y) + D_f(\overleftarrow{P}_C y, y).$$

Finally, the operator $\overleftarrow{P}_C$ is continuous on $\operatorname{int}\operatorname{dom} f$.

Proof. Under the present assumptions on f, the claims follow from [5, Theorem 3.14 and Proposition 3.16], except for the continuity of $\overleftarrow{P}_C$, which we derive now. Suppose that $(x_n)_{n\in\mathbb{N}}$ is a sequence in $\operatorname{int}\operatorname{dom} f$ converging to $\bar{x} \in \operatorname{int}\operatorname{dom} f$. Set $(c_n)_{n\in\mathbb{N}} = (\overleftarrow{P}_C x_n)_{n\in\mathbb{N}}$ and $\bar{c} = \overleftarrow{P}_C \bar{x}$. We must show that $(c_n)_{n\in\mathbb{N}}$ converges to $\bar{c}$. Using Proposition 2.2(i) and (2.2), we have

$$(2.3)\qquad D_f(\bar{c},\bar{x}) \leftarrow D_f(\bar{c},x_n) \geq D_f(\bar{c},c_n) + D_f(c_n,x_n) \geq D_f(\bar{c},c_n).$$

Hence $(D_f(\bar{c},c_n))_{n\in\mathbb{N}}$ is bounded. By Proposition 2.2(iv), $(c_n)_{n\in\mathbb{N}}$ is bounded and all its cluster points belong to $C \cap \operatorname{int}\operatorname{dom} f$. Let $\hat{c}$ be such a cluster point, say $c_{k_n} \to \hat{c} \in \operatorname{int}\operatorname{dom} f$. Using the definition of $\bar{c}$, Proposition 2.2(i), and (2.2), we deduce $D_f(\hat{c},\bar{x}) \geq D_f(\bar{c},\bar{x}) \leftarrow D_f(\bar{c},x_{k_n}) \geq D_f(\bar{c},c_{k_n}) + D_f(c_{k_n},x_{k_n}) \to D_f(\bar{c},\hat{c}) + D_f(\hat{c},\bar{x}) \geq D_f(\hat{c},\bar{x})$; thus $D_f(\bar{c},\hat{c}) = 0$, and hence, by Proposition 2.2(ii), $\bar{c} = \hat{c}$. ∎

DEFINITION 2.4. *The function f* allows forward Bregman projections *if it satisfies the following additional properties:*

(i) $\nabla^2 f$ *exists and is continuous on* $\operatorname{int}\operatorname{dom} f$;
(ii) D_f *is convex on* $(\operatorname{int}\operatorname{dom} f)^2$;
(iii) *for every* $x \in \operatorname{int}\operatorname{dom} f$, $D_f(x,\cdot)$ *is strictly convex on* $\operatorname{int}\operatorname{dom} f$.

REMARK 2.5. *The function f allows forward Bregman projections if and only if it satisfies the standing assumptions of* [7], *which allows us to apply the results of* [7]. *This equivalence follows from* [7, *Remark* 2.1] *and*

$$(2.4)\qquad D_f \text{ is convex on } (\operatorname{int}\operatorname{dom} f)^2 \ \Leftrightarrow\ D_f \text{ is convex on } X^2.$$

We now verify (2.4). *The implication "$\Leftarrow$" is clear. To establish "$\Rightarrow$", let us fix* $(y_1,y_2) \in (\operatorname{int}\operatorname{dom} f)^2$, $(x_1,x_2) \in (\operatorname{dom} f)^2$, *and* $(\lambda_1,\lambda_2) \in \left]0,1\right[^2$ *such that* $\lambda_1+\lambda_2 = 1$. *For* $\varepsilon \in \left]0,1\right[$ *and* $i \in \{1,2\}$, *set* $x_{i,\varepsilon} = (1-\varepsilon)x_i + \varepsilon y_i \in \operatorname{int}\operatorname{dom} f$. *Then* $D_f(\lambda_1 x_{1,\varepsilon} + \lambda_2 x_{2,\varepsilon}, \lambda_1 y_1 + \lambda_2 y_2) \leq \lambda_1 D_f(x_{1,\varepsilon},y_1) + \lambda_2 D_f(x_{2,\varepsilon},y_2)$. *Now take* $y \in \operatorname{int}\operatorname{dom} f$. *Since f is closed and convex, so is $D_f(\cdot,y)$. Hence, as* $\varepsilon \downarrow 0^+$, *the line segment continuity property of $D_f(\cdot,y)$* [31, *Corollary* 7.5.1] *results in* $D_f(\lambda_1 x_1 + \lambda_2 x_2, \lambda_1 y_1 + \lambda_2 y_2) \leq \lambda_1 D_f(x_1,y_1) + \lambda_2 D_f(x_2,y_2)$. *Thus D_f is convex on* $\operatorname{dom} f \times \operatorname{int}\operatorname{dom} f = \operatorname{dom} D_f$ *and, thereby, on* X^2.

FACT 2.6 (forward Bregman projection). *Suppose that f allows forward Bregman projections and C is a closed convex set in X such that* $C \cap \operatorname{int}\operatorname{dom} f \neq \emptyset$. *Then, for every* $y \in \operatorname{int}\operatorname{dom} f$, *there exists a unique point* $\overrightarrow{P}_C y \in C \cap \operatorname{dom} f$ *such that* $D_f(y, \overrightarrow{P}_C y) \leq D_f(y,c)$ *for all* $c \in C$. *The point $\overrightarrow{P}_C y$ is called the* forward Bregman projection *of y onto C and it is characterized by*

$$(2.5)\quad \overleftarrow{P}_C y \in C \cap \operatorname{int}\operatorname{dom} f \quad \text{and} \quad (\forall c \in C)\ \big\langle c - \overrightarrow{P}_C y, \nabla^2 f(\overrightarrow{P}_C y)(y - \overrightarrow{P}_C y)\big\rangle \leq 0;$$

equivalently, it is characterized by

$$(2.6)\quad \overleftarrow{P}_C y \in C \cap \operatorname{int}\operatorname{dom} f \ \text{ and } \ (\forall c \in C)\ D_f(c,y) \geq D_f(c, \overrightarrow{P}_C y) + D_{D_f}\big((c,c),(y,\overrightarrow{P}_C y)\big).$$

Finally, the operator $\overrightarrow{P}_C$ is continuous on $\operatorname{int}\operatorname{dom} f$.

Proof. This follows from [7, Lemma 2.9, Lemma 3.5, Lemma 3.6, and Corollary 3.7]. ∎

The key requirement in Definition 2.4 is the convexity of D_f, which is studied separately in [6]. Not every Legendre function allows forward Bregman projections, but the most important ones from Example 2.1 do.

EXAMPLE 2.7 (functions allowing forward Bregman projections; see [7, Example 2.16]). *Let $X = \mathbb{R}^J$. Then the energy, the negative entropy, and the Fermi–Dirac entropy allow forward Bregman projections.*

The following example shows that backward and forward Bregman projections are different notions.

EXAMPLE 2.8 (entropic averaging in $\mathbb{R}^2$). *Let $f\colon \mathbb{R}^2 \to \left]-\infty,+\infty\right] : (\xi_1,\xi_2) \mapsto \sum_{i=1}^{2} \xi_i \ln(\xi_i) - \xi_i$ be the negative entropy on $\mathbb{R}^2$, and let $\Delta = \{(\xi_1,\xi_2) \in \mathbb{R}^2 \colon \xi_1 = \xi_2\}$. Then $\operatorname{dom} f = [0,+\infty[^2$ and clearly $\Delta \cap \operatorname{int}\operatorname{dom} f \neq \emptyset$. Using (2.1) and (2.5), it is straightforward to verify that, for every $(\xi_1,\xi_2) \in \operatorname{int}\operatorname{dom} f = \left]0,+\infty\right[^2$,*

$$\overleftarrow{P}_{\Delta}(\xi_1,\xi_2) = \big(\sqrt{\xi_1\xi_2},\sqrt{\xi_1\xi_2}\big) \quad \text{and} \quad \overrightarrow{P}_{\Delta}(\xi_1,\xi_2) = \big(\tfrac{1}{2}(\xi_1+\xi_2),\tfrac{1}{2}(\xi_1+\xi_2)\big). \tag{2.7}$$

These formulae can also be deduced from Example 3.16 *below.*

We close this section with a characterization of convergence for Bregman monotone sequences. Note that when $f = \frac{1}{2}\|\cdot\|^2$, Bregman monotonicity reverts to the standard notion of Fejér monotonicity, which is discussed in detail in [4] and [21].

PROPOSITION 2.9 (Bregman monotonicity). *Suppose that C is a closed convex set in X such that $C \cap \operatorname{int}\operatorname{dom} f \neq \emptyset$. Suppose further that $(y_n)_{n\in\mathbb{N}}$ is a sequence which is* Bregman monotone *with respect to $C \cap \operatorname{int}\operatorname{dom} f$; i.e., it lies in* $\operatorname{int}\operatorname{dom} f$ *and*

$$(\forall c \in C \cap \operatorname{int}\operatorname{dom} f)(\forall n \in \mathbb{N}) \quad D_f(c,y_{n+1}) \leq D_f(c,y_n). \tag{2.8}$$

Then $(y_n)_{n\in\mathbb{N}}$ converges to some point in $C \cap \operatorname{int}\operatorname{dom} f$ $\Leftrightarrow$ all cluster points of $(y_n)_{n\in\mathbb{N}}$ are in C.

Proof. The implication "$\Rightarrow$" is clear. "$\Leftarrow$": pick $c \in C \cap \operatorname{int}\operatorname{dom} f$. Then the sequence $(D_f(c,y_n))_{n\in\mathbb{N}}$ is decreasing and nonnegative, and hence bounded. By Proposition 2.2(iv), $(y_n)_{n\in\mathbb{N}}$ is bounded and all its cluster points lie in $\operatorname{int}\operatorname{dom} f$. Let $\{c,\hat{c}\} \subset C \cap \operatorname{int}\operatorname{dom} f$ be two cluster points of $(y_n)_{n\in\mathbb{N}}$, say $y_{k_n} \to c$ and $y_{l_n} \to \hat{c}$. By Proposition 2.2(iii), $D_f(c,y_{k_n}) \to 0$. Since $(y_n)_{n\in\mathbb{N}}$ is Bregman monotone, we have $D_f(c,y_n) \to 0$ and, in particular, $D_f(c,y_{l_n}) \to 0$. Using Proposition 2.2(v), we conclude that $c = \hat{c}$. ∎

3. Bregman retractions.

3.1. Properties and examples.

DEFINITION 3.1 (Bregman retraction). *Suppose that C is a closed convex set in X such that $C \cap \operatorname{int}\operatorname{dom} f \neq \emptyset$ and μ is a function from $\operatorname{dom}\mu = (C \cap \operatorname{int}\operatorname{dom} f) \times \operatorname{int}\operatorname{dom} f$ to $[0,+\infty[$. Then $R\colon \operatorname{dom} R = \operatorname{int}\operatorname{dom} f \to C \cap \operatorname{int}\operatorname{dom} f$ is a* Bregman retraction of C with modulus μ *if the following two properties hold for every $c \in C \cap \operatorname{int}\operatorname{dom} f$ and every $x \in \operatorname{int}\operatorname{dom} f$:*

(i) $D_f(c,x) \geq D_f(c,Rx) + \mu(c,x)$.

(ii) *If $(x_n)_{n\in\mathbb{N}}$ is a sequence in $\operatorname{int}\operatorname{dom} f$ and y is a point in $\operatorname{int}\operatorname{dom} f$ such that $x_n \to x$, $Rx_n \to y$, and $\mu(c,x_n) \to 0$, then $x = y$.*

PROPOSITION 3.2 (basic properties of Bregman retractions). *Suppose that C is a closed convex set in X such that $C \cap \operatorname{int}\operatorname{dom} f \neq \emptyset$. Suppose further that R is*

a Bregman retraction of C with modulus μ. Then the following holds true for every $c \in C \cap \operatorname{int}\operatorname{dom} f$ and every $x \in \operatorname{int}\operatorname{dom} f$:

(i) *Suppose that $(x_n)_{n\in\mathbb{N}}$ is a sequence in $\operatorname{int}\operatorname{dom} f$ and y is a point in $\operatorname{int}\operatorname{dom} f$ such that $x_n \to x$ and $Rx_n \to y$. Then $\mu(c, x_n) \to 0 \Leftrightarrow x = y$.*

(ii) *$\mu(c, x) = 0 \Leftrightarrow x = Rx \Leftrightarrow x \in C$.*

Proof. (i) The implication "$\Rightarrow$" is clear by Definition 3.1(ii). If $x = y$, then (using Proposition 2.2(i) and Definition 3.1(i)) $0 = D_f(c, x) - D_f(c, x) \leftarrow D_f(c, x_n) - D_f(c, Rx_n) \geq \mu(c, x_n) \geq 0$. Hence $\mu(c, x_n) \to 0$ and "$\Leftarrow$" is verified. (ii) The first equivalence is a special case of (i), while the implication $x = Rx \Rightarrow x \in C$ is clear. Now assume $x \in C$. Then Definition 3.1(i) with $c = x$ yields $0 = D_f(x, x) \geq D_f(x, Rx) + \mu(x, x) \geq D_f(x, Rx) \geq 0$. Hence $D_f(x, Rx) = 0$ and thus $x = Rx$ by Proposition 2.2(ii). ∎

Every nonempty closed convex set in X possesses a Bregman retraction with respect to the energy.

EXAMPLE 3.3 (orthogonal projection). *Suppose that $f = \frac{1}{2}\|\cdot\|^2$ and C is a nonempty closed convex set in X. Then its orthogonal projection P_C is a Bregman retraction with modulus $\mu\colon (c, x) \mapsto \frac{1}{2}\|x - P_C x\|^2$.*

Proof. This will turn out to be a special case of Example 3.6 or Example 3.7. ∎

However, the next example shows that there exist Bregman retractions that are not projections.

EXAMPLE 3.4. *Let $f = \frac{1}{2}\|\cdot\|^2$ and $C = \{x \in X\colon \|x\| \leq 1\}$. Fix $\varepsilon \in]0, 1[$, define $\lambda\colon X \to [0, +\infty[\,: x \mapsto 1 + \min\{\varepsilon, \|x - P_C x\|\}$, and let $R\colon X \to C\colon x \mapsto (1 - \lambda(x))x + \lambda(x)P_C x$, where P_C is the orthogonal projection onto C. Then R is a Bregman retraction of C with modulus $\mu\colon (c, x) \mapsto \frac{1}{2}(2 - \lambda(x))\lambda(x)\|x - P_C x\|^2$.*

Proof. Fix $x \in X$ and $c \in C$. It follows from standard properties of orthogonal projections (see, e.g., [4, Corollary 2.5]) that

$$\|x - c\|^2 - \|x - Rx\|^2 \geq \big(2 - \lambda(x)\big)\lambda(x)\|x - P_C x\|^2, \tag{3.1}$$

which corresponds to Definition 3.1(i). Now assume $(x_n)_{n\in\mathbb{N}}$ converges to x. Since P_C, and hence λ, is continuous, we have $P_C x_n \to P_C x$ and $Rx_n \to Rx$. Assume further that $\mu(c, x_n) \to 0$. Then $x_n - P_C x_n \to 0$ and thus $Rx_n = x_n + \lambda(x_n)(P_C x_n - x_n) \to x$. Hence $Rx = x$ and therefore R is a Bregman retraction. ∎

REMARK 3.5. *In passing, note that if C is a closed convex set in X such that $(\operatorname{int} C) \cap \operatorname{int}\operatorname{dom} f \neq \varnothing$ and $y \in \operatorname{int}\operatorname{dom} f \smallsetminus C$, then both points $\overleftarrow{P}_C y$ and $\overrightarrow{P}_C y$ belong to $(\operatorname{bdry} C) \cap \operatorname{int}\operatorname{dom} f$. Now let R and C be as in Example 3.4. Since R maps points outside C to the interior of C, there is no function f such that R is the backward or forward Bregman projection onto C with respect to D_f.*

The following two examples contain Example 3.3 if we let $f = \frac{1}{2}\|\cdot\|^2$.

EXAMPLE 3.6 (backward Bregman projection). *Suppose that C is a closed convex set in X such that $C \cap \operatorname{int}\operatorname{dom} f \neq \varnothing$. Then the backward Bregman projection $\overleftarrow{P}_C$ is a continuous Bregman retraction with modulus $\mu\colon (c, x) \mapsto D_f(\overleftarrow{P}_C x, x)$.*

Proof. This follows from Fact 2.3 and Proposition 2.2(iii). ∎

EXAMPLE 3.7 (forward Bregman projection). *Suppose that f allows forward Bregman projections and C is a closed convex set in X such that $C \cap \operatorname{int}\operatorname{dom} f \neq \varnothing$. Then the forward Bregman projection $\overrightarrow{P}_C$ is a continuous Bregman retraction with*

modulus

$$\mu\colon (c,x)\mapsto D_{D_f}\big((c,c),(x,\overrightarrow{P}_C x)\big) = D_f(c,x)-D_f(c,\overrightarrow{P}_C x)+\big\langle c-\overrightarrow{P}_C x,\nabla^2 f(\overrightarrow{P}_C x)(x-\overrightarrow{P}_C x)\big\rangle. \tag{3.2}$$

Proof. See [7, Lemma 2.9] for the nonnegativity of D_{D_f} and for the expression of D_{D_f}. Fact 2.6 states that $\overrightarrow{P}_C$ is continuous and (2.6) verifies Definition 3.1(i). It remains to establish condition (ii) of Definition 3.1. So pick $c\in C\cap \operatorname{int}\operatorname{dom} f$ and $(x_n)_{n\in\mathbb{N}}$ in $\operatorname{int}\operatorname{dom} f$ such that $x_n\to x\in \operatorname{int}\operatorname{dom} f$, $\overrightarrow{P}_C x_n\to y\in \operatorname{int}\operatorname{dom} f$, and $\mu(c,x_n)\to 0$. By [7, Lemma 2.9], D_{D_f} is continuous on $(\operatorname{int}\operatorname{dom} f)^4$ and therefore $\mu(c,x_n)\to D_{D_f}((c,c),(x,y))$. Altogether, $D_{D_f}((c,c),(x,y))=0$ and [7, Lemma 2.10] implies $x=y$. ∎

The following example is motivated by [30, section 4.7].

EXAMPLE 3.8. *Let $X=\mathbb{R}^J$, f be the negative entropy, and let*

$$C=\big\{x\in X\colon x\geq 0 \text{ and } \langle x,\mathbf{1}\rangle\leq 1\big\},\quad \text{where}\quad \mathbf{1}=(1,\dots,1)\in X. \tag{3.3}$$

Let

$$R\colon \operatorname{int}\operatorname{dom} f\to C\cap\operatorname{int}\operatorname{dom} f\colon x\mapsto \begin{cases} x & \text{if } x\in C;\\ x/\langle x,\mathbf{1}\rangle & \text{otherwise.}\end{cases} \tag{3.4}$$

Then $R=\overleftarrow{P}_C=\overrightarrow{P}_C$. Consequently, R is a continuous Bregman retraction of C.

Proof. Fix $c\in C$ and $x\in\operatorname{int}\operatorname{dom} f\smallsetminus C$. Then, abusing notation slightly,

$$\begin{aligned}\langle c-Rx,\nabla f(x)-\nabla f(Rx)\rangle &= \big\langle c-x/\langle x,\mathbf{1}\rangle,\ln(x)-\ln\big(x/\langle x,\mathbf{1}\rangle\big)\big\rangle\\ &= \big\langle c-x/\langle x,\mathbf{1}\rangle,\ln\big(\langle x,\mathbf{1}\rangle\big)\mathbf{1}\big\rangle\\ &= \ln\big(\langle x,\mathbf{1}\rangle\big)\big(\langle c,\mathbf{1}\rangle-1\big)\\ &\leq 0.\end{aligned}$$

By (2.1), we see that $Rx=\overleftarrow{P}_C x$. Similarly,

$$\begin{aligned}\big\langle c-Rx,\nabla^2 f(Rx)(x-Rx)\big\rangle &= \big\langle c-x/\langle x,\mathbf{1}\rangle,\big(\langle x,\mathbf{1}\rangle-1\big)\cdot\mathbf{1}\big\rangle\\ &= \big(\langle x,\mathbf{1}\rangle-1\big)\big(\langle c,\mathbf{1}\rangle-1\big)\\ &\leq 0.\end{aligned}$$

Thus, using (2.5), $Rx=\overrightarrow{P}_C x$. ∎

REMARK 3.9. *In* [30, *Section* 4.7], *it is observed that the* orthogonal *projection of an arbitrary point in $\mathbb{R}^J$ onto C is hard to compute explicitly, and hence the use of the following extension $\widetilde{R}$ of R is suggested. Denoting the nonnegative part of a vector $x\in\mathbb{R}^J$ by x^+ (i.e., x^+ is the orthogonal projection of x onto the nonnegative orthant), the extension $\widetilde{R}$ is defined by*

$$\widetilde{R}\colon X\to C\colon x\mapsto \begin{cases} x^+/\langle x^+,\mathbf{1}\rangle & \text{if } \langle x^+,\mathbf{1}\rangle>1;\\ x^+ & \text{otherwise.}\end{cases} \tag{3.5}$$

It is important to note that for certain points $x\in\operatorname{dom} f\smallsetminus C$ and $c\in C$, the inequality $\|\widetilde{R}x-c\|=\|Rx-c\|\leq\|x-c\|$ does not *hold. Indeed, take $X=\mathbb{R}^2$, let $c=(1,0)$, consider the ray emanating from 0 that makes an angle of $\pi/6$ with $[0,+\infty[\cdot c$, and let x be the orthogonal projection of c onto this ray. Then $\|\widetilde{R}x-c\|=\|Rx-c\|>\|x-c\|$ (and this example can be lifted to $\mathbb{R}^J$, where $J\geq 3$). Therefore, Example* 3.8 *shows that an operator which is not a Bregman retraction with respect to the energy may turn out to be a Bregman retraction with respect to some other function.*

3.2. Comparison with Censor and Reich's paracontractions. Let us first recall the concept of a *Bregman function*, as defined in [15] or [17] (see also [5], [9], [25], and [32] for more concise definitions).

DEFINITION 3.10. *Let S be a nonempty open convex subset of $\mathbb{R}^J$, let $g\colon \overline{S} \to \mathbb{R}$ be a continuous and strictly convex function, and let D_g be the corresponding Bregman distance. Then g is a* Bregman function *with* zone S *if the following conditions hold:*

(i) *g is continuously differentiable on S;*
(ii) *for every $x \in \overline{S}$ the sets $(\{y \in S\colon D_g(x,y) \leq \eta\})_{\eta\in\mathbb{R}}$ are bounded;*
(iii) *for every $y \in S$ the sets $(\{x \in \overline{S}\colon D_g(x,y) \leq \eta\})_{\eta\in\mathbb{R}}$ are bounded;*
(iv) *if $(y_n)_{n\in\mathbb{N}}$ lies in S and $y_n \to y$, then $D_g(y, y_n) \to 0$;*
(v) *if $(x_n)_{n\in\mathbb{N}}$ is a bounded sequence in $\overline{S}$, $(y_n)_{n\in\mathbb{N}}$ lies in S, $y_n \to y$, and $D_g(x_n, y_n) \to 0$, then $x_n \to y$.*

The following notion is due to Censor and Reich.

DEFINITION 3.11 (see [16, Definition 3.2]). *Suppose that g is a Bregman function with zone $S \subset \mathbb{R}^J$, and let $T\colon S \to \mathbb{R}^J$ be an operator with domain S. A point $\bar{x} \in \mathbb{R}^J$ is called an* asymptotic fixed point *of T if there exists a sequence $(x_n)_{n\in\mathbb{N}}$ in S such that $x_n \to \bar{x}$ and $Tx_n \to \bar{x}$. The set of asymptotic fixed points is denoted by $\widehat{F}(T)$. The operator T is a* paracontraction *if $\widehat{F}(T) \neq \varnothing$ and the following two conditions hold:*

(i) $(\forall c \in \widehat{F}(T))(\forall x \in S) \quad D_g(c, Tx) \leq D_g(c, x)$.
(ii) *If $(x_n)_{n\in\mathbb{N}}$ is a bounded sequence in S and $c \in \widehat{F}(T)$ satisfies $D_g(c, x_n) - D_g(c, Tx_n) \to 0$, then $D_g(Tx_n, x_n) \to 0$.*

EXAMPLE 3.12 (Bregman retraction $\not\Rightarrow$ paracontraction). *Let f and Δ be as in Example* 2.8, *and set $T = \overrightarrow{P}_{\Delta}$. Then T is a continuous Bregman retraction but not a paracontraction.*

Proof. The first claim follows from Examples 2.7 and 3.7. We now show that T is not a paracontraction. First, f is a Bregman function with zone $S = \operatorname{int}\operatorname{dom} f = \left]0, +\infty\right[^2$. In addition,

$$D_f(x,y) = \xi_1 \ln(\xi_1/\eta_1) - \xi_1 + \eta_1 + \xi_2 \ln(\xi_2/\eta_2) - \xi_2 + \eta_2 \tag{3.6}$$

for $x = (\xi_1, \xi_2) \in \operatorname{dom} f = [0, +\infty[^2$ and $y = (\eta_1, \eta_2) \in \operatorname{int}\operatorname{dom} f = \left]0, +\infty\right[^2$. The set of asymptotic fixed points of T is seen to be

$$\widehat{F}(T) = \Delta \cap \operatorname{dom} f = \big\{(\xi_1, \xi_2) \in \mathbb{R}^2 : \xi_1 = \xi_2 \geq 0\big\} \neq \varnothing. \tag{3.7}$$

Fix $c = (0,0) \in \widehat{F}(T)$ and pick an arbitrary $x = (\xi_1, \xi_2) \in \operatorname{int}\operatorname{dom} f \smallsetminus \Delta$. By Example 2.8, $Tx = \overrightarrow{P}_{\Delta} x = \frac{1}{2}(\xi_1 + \xi_2, \xi_1 + \xi_2)$. Hence,

$$\begin{aligned} &D_f(c,x) - D_f(c,Tx) \\ &\quad = D_f(0,x) - D_f(0,Tx) = \big(\xi_1 + \xi_2\big) - \big(\tfrac{1}{2}(\xi_1 + \xi_2) + \tfrac{1}{2}(\xi_1 + \xi_2)\big) = 0. \end{aligned} \tag{3.8}$$

However, since $x \notin \Delta$, we have $Tx = \overrightarrow{P}_{\Delta} x \neq x$ and so, by Proposition 2.2(ii), $D_f(Tx, x) > 0$. Therefore Definition 3.11(ii) fails, and it follows that T is not a paracontraction. ∎

EXAMPLE 3.13 (paracontraction $\not\Rightarrow$ Bregman retraction). *Let $X = \mathbb{R}$ and $f = \frac{1}{2}|\cdot|^2$. Then f is a Bregman function with zone $S = X$ and $T\colon X \to X\colon x \mapsto \frac{1}{2}x$ is a paracontraction with $\widehat{F}(T) = \{0\}$. Now suppose that T is a Bregman retraction. Then, by Proposition* 3.2(ii), *the underlying set must be $C = \{0\}$. However, by Definition* 3.1,

this is absurd since the range of T is not a subset of C. Therefore T is not a Bregman retraction.

3.3. New Bregman retractions via averages and products.

PROPOSITION 3.14 (averaged Bregman retractions). *Suppose that f allows forward Bregman projections and C is a closed convex set in X such that $C \cap \operatorname{int}\operatorname{dom} f \neq \emptyset$. Suppose further that R_1 and R_2 are two continuous Bregman retractions of C with moduli μ_1 and μ_2. Fix $\lambda_1 > 0$ and $\lambda_2 > 0$ such that $\lambda_1 + \lambda_2 = 1$, and set $R = \lambda_1 R_1 + \lambda_2 R_2$. Then R is a Bregman retraction of C with modulus $\mu = \lambda_1\mu_1 + \lambda_2\mu_2$.*

Proof. It is clear that the range of R is contained in $C \cap \operatorname{int}\operatorname{dom} f$ and that $\operatorname{dom} R = \operatorname{int}\operatorname{dom} f$. Fix $c \in C \cap \operatorname{int}\operatorname{dom} f$ and $x \in \operatorname{int}\operatorname{dom} f$. Since both R_1 and R_2 are Bregman retractions of C and since $D_f(c, \cdot)$ is convex on $\operatorname{int}\operatorname{dom} f$, we have

$$\begin{aligned} D_f(c,x) &= \lambda_1 D_f(c,x) + \lambda_2 D_f(c,x) \\ &\geq \lambda_1\big(D_f(c,R_1x) + \mu_1(c,x)\big) + \lambda_2\big(D_f(c,R_2x) + \mu_2(c,x)\big) \\ &= \big(\lambda_1 D_f(c,R_1x) + \lambda_2 D_f(c,R_2x)\big) + \mu(c,x) \\ &\geq D_f(c,Rx) + \mu(c,x). \end{aligned}$$

Hence condition (i) of Definition 3.1 holds. Next, assume $(x_n)_{n\in\mathbb{N}}$ is a sequence in $\operatorname{int}\operatorname{dom} f$ converging to x such that $Rx_n \to y \in \operatorname{int}\operatorname{dom} f$ and $\mu(c,x_n) \to 0$. Then $\mu_1(c,x_n) \to 0$ and $\mu_2(c,x_n) \to 0$. On the other hand, since R_1 and R_2 are continuous on $\operatorname{int}\operatorname{dom} f$, $(R_1x_n, R_2x_n) \to (R_1x, R_2x)$, and hence $Rx_n \to Rx$. Thus $y = Rx$. Using condition (ii) of Definition 3.1 on each R_i, we also have $x = R_1x = R_2x$ and thus $x = Rx$. Altogether, $x = y$ and condition (ii) of Definition 3.1 is verified as well. ∎

EXAMPLE 3.15 (averaged backward-forward Bregman projections). *Suppose that f allows forward Bregman projections and C is a closed convex set in X such that $C \cap \operatorname{int}\operatorname{dom} f \neq \emptyset$. Denote the Bregman retraction and its modulus from Example* 3.6 *(respectively, Example* 3.7*) by R_1 and μ_1 (respectively, R_2 and μ_2). Fix $\lambda_1 > 0$ and $\lambda_2 > 0$ such that $\lambda_1 + \lambda_2 = 1$, and set $R = \lambda_1 R_1 + \lambda_2 R_2$. Then R is a Bregman retraction of C with modulus $\mu = \lambda_1\mu_1 + \lambda_2\mu_2$.*

We conclude this section with a product space construction first introduced by Pierra in [28] (see also [29]). The extension to a Bregman distance setting is due to Censor and Elfving [13]. The product space technique will be extremely useful for analyzing the parallel projection methods presented in section 5.

EXAMPLE 3.16 (product space setup). *For convenience, let $I = \{1, \ldots, N\}$ in* (1.3)*. Denote the standard Euclidean product space X^N by $\mathbf{X}$, and write $\mathbf{x} = (x_i)_{i\in I}$ for $\mathbf{x} \in \mathbf{X}$. Let*

$$\boldsymbol{\Delta} = \big\{(x, \ldots, x) \in \mathbf{X} : x \in X\big\} \quad \text{and} \quad \mathbf{C} = C_1 \times \cdots \times C_N. \tag{3.9}$$

Fix $(\lambda_i)_{i\in I}$ in $]0,1]$ such that $\sum_{i\in I} \lambda_i = 1$, and set

$$\mathbf{f} \colon \mathbf{X} \to \,]-\infty, +\infty] : \mathbf{x} \mapsto \sum_{i\in I} \lambda_i f(x_i). \tag{3.10}$$

Then $\mathbf{f}$ is Legendre, $\operatorname{dom}\mathbf{f}^*$ is open, and $\boldsymbol{\Delta} \cap \mathbf{C} \cap \operatorname{int}\operatorname{dom}\mathbf{f} \neq \emptyset$. In addition, if $\mathbf{x} \in \operatorname{dom}\mathbf{f}$ and $\mathbf{y} \in \operatorname{int}\operatorname{dom}\mathbf{f}$, then $D_{\mathbf{f}}(\mathbf{x},\mathbf{y}) = \sum_{i\in I} \lambda_i D_f(x_i, y_i)$. Moreover, we have the following:

(i) *The operators* $\overleftarrow{P}_{\boldsymbol{\Delta}}$ *and* $\overleftarrow{P}_{\mathbf{C}}$ *are continuous Bregman retractions of* $\boldsymbol{\Delta}$ *and* $\mathbf{C}$*, respectively, and*

$$\begin{aligned}\overleftarrow{P}_{\boldsymbol{\Delta}}\mathbf{y} &= (z,\ldots,z), \quad \text{where} \quad z = \nabla f^*\left(\sum_{i\in I}\lambda_i\nabla f(y_i)\right),\\ \overleftarrow{P}_{\mathbf{C}}\mathbf{y} &= \big(\overleftarrow{P}_{C_i}y_i\big)_{i\in I}.\end{aligned} \tag{3.11}$$

(ii) *Suppose that* f *allows forward Bregman projections. Then so does* $\mathbf{f}$*. The operators* $\overrightarrow{P}_{\boldsymbol{\Delta}}$ *and* $\overrightarrow{P}_{\mathbf{C}}$ *are continuous Bregman retractions of* $\boldsymbol{\Delta}$ *and* $\mathbf{C}$*, respectively, and*

$$\begin{aligned}\overrightarrow{P}_{\boldsymbol{\Delta}}\mathbf{y} &= (z,\ldots,z), \quad \text{where} \quad z = \sum_{i\in I}\lambda_i y_i,\\ \overrightarrow{P}_{\mathbf{C}}\mathbf{y} &= \big(\overrightarrow{P}_{C_i}y_i\big)_{i\in I}.\end{aligned} \tag{3.12}$$

Proof. The fact that the operators $\overleftarrow{P}_{\boldsymbol{\Delta}}$, $\overleftarrow{P}_{\mathbf{C}}$ (and $\overrightarrow{P}_{\boldsymbol{\Delta}}$, $\overrightarrow{P}_{\mathbf{C}}$, provided they exist) are continuous Bregman retractions follows from Example 3.6 (and Example 3.7, respectively). (i) See [5, Corollary 7.2] or [13, Lemmata 4.1 and 4.2]. (ii) Using Definition 2.4, it is straightforward to check that $\mathbf{f}$ allows forward Bregman projections. Next, let $z = \sum_{i\in I}\lambda_i y_i$ and $\mathbf{z} = (z,\ldots,z) \in \mathbf{X}$. Then $\mathbf{z} \in \boldsymbol{\Delta}$. Observe that $\nabla^2\mathbf{f}(\mathbf{z})\mathbf{y} = (\lambda_i\nabla^2 f(z)y_i)_{i\in I}$ and $\nabla^2\mathbf{f}(\mathbf{z})\mathbf{z} = (\lambda_i\nabla^2 f(z)z)_{i\in I}$. Hence $\nabla^2\mathbf{f}(\mathbf{z})(\mathbf{y}-\mathbf{z}) \in \boldsymbol{\Delta}^\perp = \{\mathbf{x}\in\mathbf{X} : \sum_{i\in I}x_i = 0\}$, because $\sum_{i\in I}\lambda_i\nabla^2 f(z)y_i = \nabla^2 f(z)(\sum_{i\in I}\lambda_i y_i) = \nabla^2 f(z)z = \sum_{i\in I}\lambda_i\nabla^2 f(z)z$. Thus, it follows from (2.5) that $\mathbf{z} = \overrightarrow{P}_{\boldsymbol{\Delta}}\mathbf{y}$. In view of the separability of $D_{\mathbf{f}}$ and $\mathbf{C}$, the formula for $\overrightarrow{P}_{\mathbf{C}}$ is clear. ∎

REMARK 3.17. *The case when* $\mathbf{f}$ *is replaced by* $\sum_{i\in I}\lambda_i g_i(x_i)$*, where* $(g_i)_{i\in I}$ *is a family of possibly different Bregman functions, was considered in* [12] *and* [13]*. This setup is too general to permit closed forms for* $\overleftarrow{P}_{\boldsymbol{\Delta}}$ *or* $\overrightarrow{P}_{\boldsymbol{\Delta}}$*. Furthermore, since Bregman functions are not necessarily Legendre, the existence of Bregman projections is not guaranteed and must therefore be imposed.*

4. Main result. Going back to (1.4), we henceforth set

$$C = \bigcap_{i\in I} C_i \tag{4.1}$$

and assume that (the existence of the Bregman retractions is guaranteed by (1.3) and Example 3.6)

$$(\forall i \in I) \quad R_i \text{ is a Bregman retraction of } C_i \text{ with modulus } \mu_i. \tag{4.2}$$

We now formulate our main result.

THEOREM 4.1 (method of Bregman retractions). *Given an arbitrary starting point* $y_0 \in \operatorname{int}\operatorname{dom} f$*, generate a sequence by*

$$(\forall n \in \mathbb{N}) \quad y_{n+1} = R_{\mathrm{i}(n+1)}y_n, \tag{4.3}$$

where $\mathrm{i}\colon \mathbb{N} \to I$ *takes on each value in* I *infinitely often. Then the sequence* $(y_n)_{n\in\mathbb{N}}$ *converges to a point in* $C \cap \operatorname{int}\operatorname{dom} f$*.*

Proof. We proceed in several steps.

Step 1. We have

$$(\forall n \in \mathbb{N})(\forall c \in C_{\mathrm{i}(n+1)} \cap \operatorname{int}\operatorname{dom} f) \quad D_f(c, y_n) \geq D_f(c, y_{n+1}) + \mu_{\mathrm{i}(n+1)}(c, y_n).$$

Indeed, $y_{n+1} = R_{\mathrm{i}(n+1)} y_n$ and $R_{\mathrm{i}(n+1)}$ is a Bregman retraction of $C_{\mathrm{i}(n+1)}$.

Step 2. $(y_n)_{n\in\mathbb{N}}$ is Bregman monotone with respect to $C \cap \operatorname{int}\operatorname{dom} f$.

This is clear from Step 1.

Step 3. $(y_n)_{n\in\mathbb{N}}$ is bounded and all its cluster points belong to $\operatorname{int}\operatorname{dom} f$.

Fix $c \in C \cap \operatorname{int}\operatorname{dom} f$. In view of Step 2, the sequence $\big(D_f(c, y_n)\big)_{n\in\mathbb{N}}$ is decreasing, and hence bounded. Now apply Proposition 2.2(iv).

Next, let us consider an arbitrary cluster point of $(y_n)_{n\in\mathbb{N}}$, say $y_{k_n} \to y$.

Step 4. $y \in \operatorname{int}\operatorname{dom} f$ and $D_f(y, y_{k_n}) \to 0$.

This follows from Step 3 and Proposition 2.2(i).

Because I is finite, after passing to a further subsequence and relabelling if necessary, we assume that $\mathrm{i}(k_n) \equiv i_{\mathrm{in}}$. Since $C_{i_{\mathrm{in}}}$ is closed, we have $y \in C_{i_{\mathrm{in}}}$.

We now define $I_{\mathrm{in}} = \{i \in I : y \in C_i\}$ and $I_{\mathrm{out}} = \{i \in I : y \notin C_i\}$.

Step 5. $I_{\mathrm{out}} = \emptyset$.

Suppose to the contrary that $I_{\mathrm{out}} \neq \emptyset$. After passing to a further subsequence and relabelling if necessary, we assume that $\{\mathrm{i}(k_n), \mathrm{i}(k_n+1), \ldots, \mathrm{i}(k_{n+1}-1)\} = I$—this is possible by our assumptions on the index selector i. For each $n \in \mathbb{N}$, let

$$m_n = \min\big\{k_n \leq k \leq k_{n+1} - 1 : \mathrm{i}(k) \in I_{\mathrm{out}}\big\} - 1. \tag{4.4}$$

The current assumptions imply that each m_n is a well-defined integer in $[k_n, k_{n+1}-2]$ satisfying $y \in \bigcap_{k_n \leq k \leq m_n} C_{\mathrm{i}(k)}$. Repeated use of Step 1 thus yields

$$(\forall n \in \mathbb{N}) \quad D_f(y, y_{m_n}) \leq D_f(y, y_{k_n}). \tag{4.5}$$

Using Step 4 and Proposition 2.2(v), we deduce $y_{m_n} \to y$. After passing to a further subsequence and relabelling if necessary, we assume that $\mathrm{i}(m_n+1) \equiv i_{\mathrm{out}}$ and that $y_{m_n+1} = R_{i_{\mathrm{out}}} y_{m_n} \to z \in C_{i_{\mathrm{out}}} \cap \operatorname{int}\operatorname{dom} f$ (using Step 3 again). Now fix $c \in C \cap \operatorname{int}\operatorname{dom} f$. Step 1 implies that $(\mu_{\mathrm{i}(n+1)}(c, y_n))_{n\in\mathbb{N}}$ is summable; in particular,

$$\mu_{i_{\mathrm{out}}}(c, y_{m_n}) = \mu_{\mathrm{i}(m_n+1)}(c, y_{m_n}) \to 0. \tag{4.6}$$

Since $R_{i_{\mathrm{out}}}$ is a Bregman retraction, we obtain $y = z \in C_{i_{\mathrm{out}}}$. But this in turn implies $i_{\mathrm{out}} \in I_{\mathrm{in}}$, which is the desired contradiction.

Last step. We have shown that $(y_n)_{n\in\mathbb{N}}$ is Bregman monotone with respect to $C \cap \operatorname{int}\operatorname{dom} f$ (Step 2) and that all its cluster points lie in $C \cap \operatorname{int}\operatorname{dom} f$ (Step 4 and Step 5). Therefore, by Proposition 2.9, the entire sequence $(y_n)_{n\in\mathbb{N}}$ converges to some point in $C \cap \operatorname{int}\operatorname{dom} f$. ∎

REMARK 4.2. *The proof of Theorem* 4.1 *is guided by the proof of* [7, *Theorem* 4.1] *and similar convergence results on iterating operators under such general control; see* [5], [16], *and* [26]. *The present proof clearly shows when properties of the Bregman distance are used, as opposed to those of the modulus. This distinction is blurred in other proofs, because the implicit surrogates for the modulus depend on* D_f*: see the roles of* $D_f(\overleftarrow{P}_{C_{r(n+1)}} y_n, y_n)$, $D_{D_f}((c,c),(y_n, \overrightarrow{P}_{C_{r(n+1)}} y_n))$, $D_f(T_s z(t), z(t))$, *and* $D_h^k(x^{k+1}, x^k)$ *in the proofs of* [5, *Theorem* 8.1], [7, *Theorem* 4.1], *and* [16, *Theorem* 3.1], [26, *Theorem* 4.1], *respectively.*

REMARK 4.3 (Bregman retractions must correspond to the same Bregman distance). *It is natural to ask whether it is possible to use iterates of Bregman retractions*

coming from possibly different *underlying Bregman distances to solve convex feasibility problems. Unfortunately, this approach is not successful in general. To see this, let* $X = \mathbb{R}^2$, *and set* $R_C = \overleftarrow{P}_C = \overrightarrow{P}_C$, *where* f *is the* negative entropy *and* C *is as in Example* 3.8 *(with* $J = 2$*). Further, let* L *be the straight line through the points* $(0, \frac{53}{56})$ *and* $(\frac{1}{4}, \frac{7}{8})$, *and let* R_L *be the orthogonal projection, i.e., the backward or forward Bregman projection with respect to the* energy. *Then, although* $L \cap \operatorname{int} C \neq \emptyset$, *iterating the map* $T = R_L \circ R_C$ *may not lead to a point in* $C \cap L$: *indeed,* $(\frac{1}{4}, \frac{7}{8})$ *is a fixed point of* T outside C.

5. Applications. Continuing to work under assumptions (4.1) and (4.2), we now discuss various sequential and parallel algorithms derived from Theorem 4.1.

5.1. Sequential algorithms.

Application 5.1 (sequential Bregman projections). *For each* $i \in I$, *let* $R_i = \overleftarrow{P}_{C_i}$. *Then Theorem* 4.1 *coincides with* [5, *Theorem* 8.1.(ii)]; *see also* [16, *Theorem* 3.2]. *For cyclic Bregman projections, see Bregman's classical paper* [8].

Application 5.2 (new method of mixed backward-forward Bregman projections). *Suppose that* f *allows forward Bregman projections. For each* $i \in I$, *let either* $R_i = \overleftarrow{P}_{C_i}$ *or* $R_i = \overrightarrow{P}_{C_i}$. *Then Theorem* 4.1 *yields a convergence result on iterating a mixture of backward and forward Bregman projections.* Note: *If desired, it is possible to use both* $\overleftarrow{P}_{C_i}$ *and* $\overrightarrow{P}_{C_i}$ *for a given set* C_i *infinitely often, by counting this set twice.*

The following three algorithms are special instances of Application 5.2.

Application 5.3 (sequential forward Bregman projections). *Suppose that* f *allows forward Bregman projections and let* $R_i = \overrightarrow{P}_{C_i}$ *for every* $i \in I$. *Then Theorem* 4.1 *reduces to* [7, *Theorem* 4.1].

Application 5.4 (sequential orthogonal projections). *Suppose that* $f = \frac{1}{2}\|\cdot\|^2$, *and let each* R_i *be the orthogonal projection* P_{C_i}. *Then Theorem* 4.1 *turns into a convergence result on (chaotic or random) iterations of orthogonal projections; see also* [2], [19], *and references therein.*

Application 5.5 (alternating backward-forward Bregman projections). *Suppose that* f *allows forward Bregman projections, and let* $I = \{1, 2\}$, $R_1 = \overrightarrow{P}_{C_1}$, *and* $R_2 = \overleftarrow{P}_{C_2}$. *Then the method of Bregman retractions* (4.3) *corresponds to an alternating backward-forward Bregman projection method, which can be viewed as Csiszár and Tusnády's* alternating minimization procedure [22] *applied to* D_f *(this covers the* Expectation-Maximization method *for a specific Poisson model; see* [22] *and* [24]*).*

5.2. Parallel algorithms. Various parallel algorithms arise by specializing Application 5.2 to the product space setting of Example 3.16. Using Example 3.16 and its notation, we deduce that the sequence $(\mathbf{T}^n\mathbf{x}_0)_{n\in\mathbb{N}}$, where $\mathbf{x}_0 \in \boldsymbol{\Delta}$ and $\mathbf{T} = \overleftarrow{P}_{\boldsymbol{\Delta}} \circ \overleftarrow{P}_{\mathbf{C}}$, converges to some point in $\boldsymbol{\Delta} \cap \mathbf{C} \cap \operatorname{int}\operatorname{dom}\mathbf{f}$. The same holds true when $\mathbf{T} \in \{\overrightarrow{P}_{\boldsymbol{\Delta}} \circ \overleftarrow{P}_{\mathbf{C}}, \overleftarrow{P}_{\boldsymbol{\Delta}} \circ \overrightarrow{P}_{\mathbf{C}}, \overrightarrow{P}_{\boldsymbol{\Delta}} \circ \overrightarrow{P}_{\mathbf{C}}\}$, provided that f allows forward Bregman projections.

Translating back to the original space X, we obtain the following four parallel algorithms.

Application 5.6 (parallel projections à la Censor and Elfving). *Given* $x_0 \in \operatorname{int}\operatorname{dom} f$, *the sequence generated by*

$$(\forall n \in \mathbb{N}) \quad x_{n+1} = \nabla f^* \left(\sum_{i\in I} \lambda_i \nabla f(\overleftarrow{P}_{C_i} x_n) \right) \tag{5.1}$$

converges to a point in $C \cap \operatorname{int}\operatorname{dom} f$. *This method, which amounts to iterating* $\overleftarrow{P}_{\boldsymbol{\Delta}} \circ \overleftarrow{P}_{\mathbf{C}}$ *in* $\mathbf{X}$, *was first suggested implicitly in* [13]; *see also* [5] *and Remark* 3.17.

APPLICATION 5.7 (parallel projections à la Byrne and Censor I). *Suppose that f allows forward Bregman projections. Given $x_0 \in \operatorname{int}\operatorname{dom} f$, the sequence generated by*

$$(\forall n \in \mathbb{N}) \quad x_{n+1} = \sum_{i\in I} \lambda_i \overleftarrow{P}_{C_i} x_n \tag{5.2}$$

converges to a point in $C \cap \operatorname{int}\operatorname{dom} f$. This method, which amounts to iterating $\overrightarrow{P}_{\mathbf{\Delta}} \circ \overleftarrow{P}_{\mathbf{C}}$ in $\mathbf{X}$, can be found implicitly in [11, *section* 4.1] *(see also Remark* 3.17*).*

APPLICATION 5.8 (parallel projections à la Byrne and Censor II). *Suppose that f allows forward Bregman projections. Given $x_0 \in \operatorname{int}\operatorname{dom} f$, the sequence generated by*

$$(\forall n \in \mathbb{N}) \quad x_{n+1} = \nabla f^* \left(\sum_{i\in I} \lambda_i \nabla f(\overrightarrow{P}_{C_i} x_n) \right) \tag{5.3}$$

converges to a point in $C \cap \operatorname{int}\operatorname{dom} f$. This method, which amounts to iterating $\overleftarrow{P}_{\mathbf{\Delta}} \circ \overrightarrow{P}_{\mathbf{C}}$ in $\mathbf{X}$, can be found implicitly in [11, *section* 4.2] *(see also Remark* 3.17*).*

APPLICATION 5.9 (new parallel method). *Suppose that f allows forward Bregman projections. Given $x_0 \in \operatorname{int}\operatorname{dom} f$, the sequence generated by*

$$(\forall n \in \mathbb{N}) \quad x_{n+1} = \sum_{i\in I} \lambda_i \overrightarrow{P}_{C_i} x_n \tag{5.4}$$

converges to a point in $C \cap \operatorname{int}\operatorname{dom} f$. This corresponds to iterating $\overrightarrow{P}_{\mathbf{\Delta}} \circ \overrightarrow{P}_{\mathbf{C}}$ in $\mathbf{X}$.

The negative entropy and the energy lead to concrete examples.

APPLICATION 5.10 (averaged entropic projections à la Butnariu, Censor, and Reich). *Let f be the negative entropy. Given $x_0 \in \operatorname{int}\operatorname{dom} f$, the sequence generated by*

$$(\forall n \in \mathbb{N}) \quad x_{n+1} = \sum_{i\in I} \lambda_i \overleftarrow{P}_{C_i} x_n \tag{5.5}$$

converges to a point in $C \cap \operatorname{int}\operatorname{dom} f$. Convergence is guaranteed by [10, *Theorem* 3.3], *which holds true in more general settings, or by Application* 5.7.

We conclude with a classical method which can be obtained from Application 5.6, 5.7, 5.8, or 5.9 by setting $f = \frac{1}{2}\|\cdot\|^2$.

APPLICATION 5.11 (parallel orthogonal projections à la Auslender). *For each $i \in I$, let P_{C_i} be the orthogonal projection onto C_i. Given $x_0 \in X$, the sequence generated by*

$$(\forall n \in \mathbb{N}) \quad x_{n+1} = \sum_{i\in I} \lambda_i P_{C_i} x_n \tag{5.6}$$

converges to some point in C [1] *(see also* [3], [18], *and* [23] *for the case when $C = \emptyset$).*

Acknowledgments. We wish to thank Charlie Byrne and Yair Censor for sending us [11] and an anonymous referee for careful reading and helpful comments.

REFERENCES

[1] A. AUSLENDER, *Optimisation – Méthodes Numériques*, Paris, Masson, 1976.
[2] H. H. BAUSCHKE, *A norm convergence result on random products of relaxed projections in Hilbert space*, Trans. Amer. Math. Soc., 347 (1995), pp. 1365–1374.

[3] H. H. Bauschke and J. M. Borwein, *On the convergence of von Neumann's alternating projection algorithm for two sets*, Set-Valued Anal., 1 (1993), pp. 185–212.
[4] H. H. Bauschke and J. M. Borwein, *On projection algorithms for solving convex feasibility problems*, SIAM Rev., 38 (1996), pp. 367–426.
[5] H. H. Bauschke and J. M. Borwein, *Legendre functions and the method of random Bregman projections*, J. Convex Anal., 4 (1997), pp. 27–67.
[6] H. H. Bauschke and J. M. Borwein, *Joint and separate convexity of the Bregman distance*, in Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications, D. Butnariu, Y. Censor, and S. Reich, eds., Elsevier, Amsterdam, The Netherlands, 2001, pp. 23–36.
[7] H. H. Bauschke and D. Noll, *The method of forward projections*, J. Nonlinear Convex Anal., 3 (2002), pp. 191–205.
[8] L. M. Bregman, *The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming*, U.S.S.R. Comput. Math. and Math. Phys., 7 (1967), pp. 200–217.
[9] D. Butnariu, C. Byrne, and Y. Censor, *Redundant axioms in the definition of Bregman functions,* J. Convex Anal., 10 (2003), to appear.
[10] D. Butnariu, Y. Censor, and S. Reich, *Iterative averaging of entropic projections for solving stochastic convex feasibility problems*, Comput. Optim. Appl., 8 (1997), pp. 21–39.
[11] C. Byrne and Y. Censor, *Proximity Function Minimization Using Multiple Bregman Projections, with Applications to Entropy Optimization and Kullback-Leibler Distance Minimization*, unpublished research report, http://math.haifa.ac.il/yair/bc29b070699.ps (June 7, 1999).
[12] C. Byrne and Y. Censor, *Proximity function minimization using multiple Bregman projections, with applications to split feasibility and Kullback-Leibler distance minimization*, Ann. Oper. Res., 105 (2001), pp. 77–98.
[13] Y. Censor and T. Elfving, *A multiprojection algorithm using Bregman projections in a product space*, Numer. Algorithms, 8 (1994), pp. 221–239.
[14] Y. Censor and G. T. Herman, *Block-iterative algorithms with underrelaxed Bregman projections*, SIAM J. Optim., 13 (2002), pp. 283–297.
[15] Y. Censor and A. Lent, *An iterative row-action method for interval convex programming*, J. Optim. Theory Appl., 34 (1981), pp. 321–353.
[16] Y. Censor and S. Reich, *Iterations of paracontractions and firmly nonexpansive operators with applications to feasibility and optimization*, Optimization, 37 (1996), pp. 323–339.
[17] Y. Censor and S. A. Zenios, *Parallel Optimization: Theory, Algorithms, and Applications,* Oxford University Press, New York, 1997.
[18] P. L. Combettes, *Inconsistent signal feasibility problems: Least-squares solutions in a product space*, IEEE Trans. Signal Process., 42 (1994), pp. 2955–2966.
[19] P. L. Combettes, *Construction d'un point fixe commun à une famille de contractions fermes*, C. R. Acad. Sci. Paris Sér. I Math., 320 (1995), pp. 1385–1390.
[20] P. L. Combettes, *Hilbertian convex feasibility problem: Convergence of projection methods*, Appl. Math. Optim., 35 (1997), pp. 311–330.
[21] P. L. Combettes, *Fejér monotonicity in convex optimization*, in Encyclopedia of Optimization, Vol. 2, C. A. Floudas and P. M. Pardalos, eds., Kluwer Academic Publishers, Boston, 2001, pp. 106–114.
[22] I. Csiszár and G. Tusnády, *Information geometry and alternating minimization procedures*, Statist. Decisions, (Supplement 1) (1984), pp. 205–237.
[23] A. R. De Pierro and A. N. Iusem, *A parallel projection method for finding a common point of a family of convex sets*, Pesqui. Oper., 5 (1985), pp. 1–20.
[24] A. N. Iusem, *A short convergence proof of the EM algorithm for a specific Poisson model*, Rev. Brasil. Prob. Estatística, 6 (1992), pp. 57–67.
[25] K. C. Kiwiel, *Free-steering relaxation methods for problems with strictly convex costs and linear constraints*, Math. Oper. Res., 22 (1997), pp. 326–349.
[26] K. C. Kiwiel, *Generalized Bregman projections in convex feasibility problems*, J. Optim. Theory Appl., 96 (1998), pp. 139–157.
[27] K. C. Kiwiel and B. Łopuch, *Surrogate projection methods for finding fixed points of firmly nonexpansive mappings*, SIAM J. Optim., 7 (1997), pp. 1084–1102.
[28] G. Pierra, *Eclatement de contraintes en parallèle pour la minimisation d'une forme quadratique*, Lecture Notes in Comput. Sci. 41, Springer-Verlag, New York, 1976, pp. 200–218.
[29] G. Pierra, *Decomposition through formalization in a product space*, Math. Programming, 28 (1984), pp. 96–115.
[30] B. T. Polyak, *Random algorithms for solving convex inequalities*, in Inherently Parallel Al-

gorithms in Feasibility and Optimization and Their Applications, D. Butnariu, Y. Censor, and S. Reich, eds., Elsevier, Amsterdam, The Netherlands, 2001, pp. 409–422.
[31] R. T. Rockafellar, *Convex Analysis,* Princeton University Press, Princeton, NJ, 1970.
[32] M. V. Solodov and B. F. Svaiter, *An inexact hybrid generalized proximal point algorithm and some new results on the theory of Bregman functions,* Math. Oper. Res., 25 (2000), pp. 214–230.

SIAM J. OPTIM.
Vol. 13, No. 4, pp. 1174–1194

PRIMAL-DUAL GRADIENT STRUCTURED FUNCTIONS: SECOND-ORDER RESULTS; LINKS TO EPI-DERIVATIVES AND PARTLY SMOOTH FUNCTIONS*

ROBERT MIFFLIN† AND CLAUDIA SAGASTIZÁBAL‡

Abstract. We give second-order expansions for quite general nonsmooth functions from the $\mathcal{VU}$-space decomposition point of view. The results depend on primal-dual gradient structure, which we relate to general concepts of second-order epi-derivatives and partly smooth functions. Expressions for the associated second-order objects are given in terms of $\mathcal{U}$-subspace Hessians.

Key words. second-order derivatives, $\mathcal{VU}$-decomposition, nonsmooth analysis, subdifferential

AMS subject classifications. Primary, 90C31, 49J52; Secondary, 65K10, 58C20

PII. S1052623402412441

1. Introduction and motivation. Nonsmooth analysis is essential for the understanding of optimization problems. In applications, nonsmoothness of functions typically appears, not in a general way, but in a structured manner. The problem of defining "good" structures for nonsmooth functions has been addressed by many authors in various ways. A good structure should be special enough to result in smooth behavior on manifolds but also general enough to apply to a broad class of functions. Ideally, the structure should characterize the manifolds and associated Hessians.

In this paper, we extend primal-dual gradient (pdg) structure, first defined in [8], to a class of not necessarily convex functions. This subclass of lower semicontinuous functions encompasses many examples, such as extreme value functions that have an underlying C^2-structure. It is similar to the amenable class [15, p. 442] in the sense of having desirable smooth substructure and containing functions that are pointwise maxima of finite collections of smooth functions [6]. It is different from the amenable class and the fully amenable subclass [15, p. 443] due to containing functions that are not regular [15, p. 260] and containing regular ones that are not fully amenable (for example, maximum eigenvalue functions and other functions arising from infinite collections of smooth functions [7]). A related class that includes maximum eigenvalue functions, as in Example 3.2 below, is introduced in [16]. However, it should be noted that this new class does not include the first and third examples in section 3.

pdg structure is tied closely to $\mathcal{VU}$-space decomposition, where $\mathcal{V}$ is the subspace parallel to a function's Clarke subdifferential [2] at a point and $\mathcal{U}$ is orthogonal to $\mathcal{V}$. The structure provides a finite set of vectors that span $\mathcal{V}$, even though the subdifferential may have a continuum of extreme points. The $\mathcal{U}$-component of the subdifferential

*Received by the editors July 31, 2002; accepted for publication December 2, 2002; published electronically April 30, 2003.
http://www.siam.org/journals/siopt/13-4/41244.html

†Department of Mathematics, Washington State University, Pullman, WA 99164-3113 (mifflin@math.wsu.edu). The research of this author was supported by National Science Foundation grant DMS-0071459.

‡IMPA, Estrada Dona Castorina 110, Jardim Botânico, 22460-320 Rio de Janeiro-RJ, Brazil (sagastiz@impa.br). On leave from INRIA, B.P. 105, 78153 Le Chesnay, France (Claudia.Sagastizabal@inria.fr). The research of this author was supported by FAPERJ (Brazil) grant E26/150.205/98.

is a singleton, and hence the function appears to be differentiable relative to $\mathcal{U}$ at the point of definition. This $\mathcal{VU}$-terminology is meant to describe the shape of a function's graph near the point in question. However, since in our (possibly) nonconvex setting the corresponding V and U shapes could be inverted, it is perhaps best to call them "sharp" and "smooth," respectively.

For a pdg-structured function we develop second-order results depending on basic index sets. These correspond to "active manifolds" such as those defined by active constraints in nonlinear programming or those defined by constant eigenvalue multiplicity in eigenvalue optimization; see [5]. The structure provides us with one or more of these sets whose associated vectors span a subspace of $\mathcal{V}$ and generate an implicit function therein from which a smooth trajectory tangent to $\mathcal{U}$ can be defined. This information also generates multipliers that are smooth functions of $u \in \mathbb{R}^{dim\ \mathcal{U}}$ and depend on the $\mathcal{V}$-subspace component of a parameter vector $y \in \mathbb{R}^n$. Combining these elements and making a primal feasibility assumption leads to the definition of a Lagrangian-like "primal-dual" function that is C^2 in u for each y. This function then provides expansions of second order in u for its associated pdg-structured function. Because y need not be a subgradient at the point in question, the final expansion contained in Theorem 5.2 below is new even for a convex function. In the nonconvex case it is useful for obtaining the result that certain primal-dual $\mathcal{U}$-Hessians associated with a local minimizer are positive semidefinite.

Under a $\mathcal{V}$-minimality assumption for a certain subgradient, the corresponding primal-dual function becomes a $\mathcal{U}$-Lagrangian, along the lines of [4], even for the nonconvex case. In addition, we are able to give expressions for second-order epi-derivatives [14] in $\mathcal{U}$-space directions in terms of the related $\mathcal{U}$-Hessians. For an example from [5] we demonstrate below that the second-order behavior of the function can be captured well either by using epi-derivatives, defined via complicated epi-limits, or by ordinary second derivatives from pdg structure.

We also connect regular members of the pdg-structured family to partly smooth functions [5] and hence identifiable surfaces [18]. An important feature of the partly smooth class is the existence of a sensitivity theory akin to that of nonlinear programming and of eigenvalue optimization. A partly smooth function is smooth on a manifold in the sense that on the manifold it equals a representative function that is C^2. Our pdg structure provides expressions for these objects and for the corresponding representative Hessian, in terms of a primal-dual Hessian associated with a basic index set that satisfies both primal and dual feasibility conditions.

The paper is organized as follows. We define the pdg-structured class in section 2 and relate the structure to $\mathcal{VU}$-space decomposition. Section 3 provides three example functions. The last one, from [5], is used in subsequent sections to illustrate the principal results. Smooth trajectories, manifolds, and associated multiplier functions are studied in section 4. Section 5 gives second-order expansions for pdg-structured functions. The connection between $\mathcal{U}$-Hessians and second-order epi-derivatives is made in section 6. Section 7 relates the previous results to functions that are partly smooth and gives expressions for manifold restricted Hessians.

We use notation from [8] and the results therein which depend on the subdifferential being a convex set rather than the function being convex. Much of the additional notation comes from [15]. For algebraic purposes we consider (sub-) gradients to be column vectors. For a vector function $v(\cdot)$, its Jacobian $Jv(\cdot)$ is a matrix, each row of which is the transposed gradient of the corresponding component of $v(\cdot)$. For a given set Y, we denote its convex hull by con$Y$ and its linear hull by lin$Y$.

2. Function structure and space decomposition. For variational analysis we use the Clarke normal cone and subgradients, as defined in [15, eqs. 6(19) and 8(32)] depending on the basic normal cone [15, Def. 6.3] introduced in [11].

More precisely, for a set $C \subset \mathbb{R}^n$ and a point $x \in C$, a vector v is *normal* to C at x if there are sequences $x^\nu \to_C x$ and $v^\nu \to v$ such that $\langle v^\nu, z - x^\nu \rangle \leq o(|z - x^\nu|)$ for all $z \in C$. The set of normal vectors is denoted by $N_C(x)$. The closure of its convex hull yields the Clarke normal cone, denoted by $\bar{N}_C(x)$.

Let $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ be a lower semicontinuous (lsc) function so that its epigraph, denoted and defined by $\operatorname{epi} f := \{(x, \beta) \in \mathbb{R}^n \times \mathbb{R} : \beta \geq f(x)\}$, is a closed set in $\mathbb{R}^{n+1}$. Take $x \in \mathbb{R}^n$, where f is finite-valued, and consider the Clarke cone normal to the epigraph of f at $(x, f(x))$. The set of Clarke subgradients of f at x is denoted and defined by

$$\bar{\partial} f(x) := \left\{ g : (g, -1) \in \bar{N}_{\operatorname{epi} f}(x, f(x)) \right\}.$$

When f is Lipschitz around x, from [2, Thm. 2.5.1], $\bar{\partial} f(x)$ is the convex hull of all possible limits of gradients at points of differentiability of f in sequences converging to x. When f is subdifferentially regular at x, then $\bar{\partial} f(x)$ equals the Mordukhovich subdifferential [11] denoted by $\partial f(x)$ in [15]; see page 337 and Definition 8.3 therein. If f is convex (and hence regular) this common subdifferential is the forerunner from convex analysis.

In order to give second-order expansions for f, it is essential to properly describe all of the subgradients in the Clarke subdifferential $\bar{\partial} f$, at least on a subset of $\mathbb{R}^n$. This is the purpose of the structure depending on a set $\mathcal{P}$ introduced next. In particular, the presence of the functions φ_ℓ and corresponding multipliers allows the subdifferential to be nonpolyhedral, and hence permits the class to contain functions that are not fully amenable [15, pp. 443–444], such as the two convex example functions considered throughout [8]. In addition, the way in which these functions enter the structure allows for the possibility of $\bar{\partial} f$ being unbounded at some points.

2.1. `pdg` structure. In the following, $\mathcal{P}$ is a subset of $\mathbb{R}^n$ containing $\bar{x}$. We say that an lsc function $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ has `pdg` *structure at* $\bar{x}$ *relative to* $\mathcal{P}$ if there exists $m_1 + 1 + m_2$ *primal functions*

$$\{f_i(x)\}_{i=0}^{m_1} \quad \text{and} \quad \{\varphi_\ell(x)\}_{\ell=1}^{m_2}$$

that are C^2 on a ball about $\bar{x}$, denoted by $B(\bar{x})$, and a *dual multiplier set* $\Delta \subset \mathbb{R}^{m_1+1+m_2}$ satisfying the following conditions:

(i)

$$\bar{x} \in \left\{ x \in B(\bar{x}) : \left\{ \begin{array}{ll} f_i(x) = f(x) & \text{for } i = 0, 1, \ldots, m_1 \\ \varphi_\ell(x) = 0 & \text{for } \ell = 1, \ldots, m_2 \end{array} \right\} \subseteq \mathcal{P} \subseteq B(\bar{x});$$

(ii) Δ is a closed convex set such that

(a) if $\alpha := (\alpha_0, \ldots, \alpha_{m_1}, \alpha_{m_1+1}, \ldots, \alpha_{m_1+m_2}) \in \Delta$, then $(\alpha_0, \ldots, \alpha_{m_1})$ is an element of the canonical simplex

$$\Delta_1 := \left\{ (\alpha_0, \alpha_1, \ldots, \alpha_{m_1}) : \sum_{i=0}^{m_1} \alpha_i = 1, \alpha_i \geq 0, i = 0, 1, \ldots, m_1 \right\};$$

(b) for each $i = 0, 1, \ldots, m_1$, $\mathbf{1}_{i+1} \in \Delta$, where $\mathbf{1}_j$ is the jth unit vector in $\mathbb{R}^{m_1+1+m_2}$;

and

(c) for each $\ell = 1, 2, \ldots, m_2$, there exists $\alpha^\ell \in \Delta$ such that $\alpha^\ell_{m_1+\ell} \neq 0$ and $\alpha^\ell_{m_1+i} = 0$ for $i \in \{1, 2, \ldots, m_2\} \backslash \{\ell\}$;

(iii) for each $x \in \mathcal{P}$,

(a) $f(x) \in \{f_i(x) : i = 0, 1, \ldots, m_1\}$;

(b) $g \in \bar{\partial} f(x)$ if and only if

$$g = \sum_{i=0}^{m_1} \alpha_i \nabla f_i(x) + \sum_{i=m_1+1}^{m_1+m_2} \alpha_i \nabla \varphi_{i-m_1}(x),$$

where the multipliers $\alpha_0, \alpha_1, \ldots, \alpha_{m_1+m_2}$ satisfy

$$\textit{complementary slackness: } \alpha_i = 0 \text{ if } \begin{cases} f_i(x) \neq f(x) & \text{for } i \leq m_1, \\ \varphi_{i-m_1}(x) \neq 0 & \text{for } i > m_1, \end{cases}$$

and

$$\textit{dual feasibility: } \alpha = (\alpha_0, \alpha_1, \ldots, \alpha_{m_1+m_2}) \in \Delta. \qquad \square$$

As a consequence of item (iii), for all $x \in \mathcal{P}$ f is finite-valued at x and $\bar{\partial} f(x)$ is a nonempty set depending on a multiplier set Δ that is independent of x and is the part of the structure which reflects the shape of the subdifferential.

This `pdg` definition is broader than its convex forerunner in [8], because the set $\mathcal{P}$ considered here can be larger than the one defined there; see the first two examples in section 3 below.

It is not difficult to show, as in [8], that this definition covers a function that is the pointwise maximum (or minimum) value of a finite collection of C^2-functions by taking $m_1 + 1$ to be the number of "active" functions at $\bar{x}$, m_2 to be 0, Δ to be Δ_1, and $\mathcal{P}$ to be a ball about $\bar{x}$ chosen small enough to exclude the "inactive" functions from the local structure. Such a function can be nonconvex and for the minimum value case can be nonregular [15, p. 260] at points of nondifferentiability and, as a consequence, not amenable [15, pp. 442–444].

Before giving additional examples, we show the connection between `pdg` structure and $\mathcal{VU}$-space decomposition.

2.2. Relation to $\mathcal{VU}$-decomposition. The $\mathcal{VU}$-theory introduced in [4] and further studied in [6], [8], [7], [10], is based on decomposing $\mathbb{R}^n$ into two orthogonal subspaces $\mathcal{V}$ and $\mathcal{U}$ depending on a point such that $\mathcal{V}$ being nontrivial indicates nonsmooth behavior of the function at the point. As a result of (2.2) below, from the $\mathcal{U}$-subspace point of view, the function appears to be differentiable.

More precisely, given an lsc function f and a point $\bar{x} \in \mathbb{R}^n$ such that $f(\bar{x})$ finite and $\bar{\partial} f(\bar{x})$ is nonempty, let $g \in \bar{\partial} f(\bar{x})$ be an arbitrary subgradient and define the orthogonal subspaces

$$\mathcal{V} := \operatorname{lin}(\bar{\partial} f(\bar{x}) - g) \quad \text{and} \quad \mathcal{U} := \mathcal{V}^\perp. \tag{2.1}$$

Note that $\mathbb{R}^n = \mathcal{U} \oplus \mathcal{V}$. Thus, letting $\bar{U}$ be an orthonormal basis matrix for $\mathcal{U}$, the $\mathcal{U}$-component of $x \in \mathbb{R}^n$ is given by $x_\mathcal{U} := \bar{U}^\top x$. In particular, from (2.1), the $\mathcal{U}$-component of a subgradient $g \in \bar{\partial} f(\bar{x})$ is the same as that of any other subgradient at $\bar{x}$; we call it the *$\mathcal{U}$-gradient* of f at $\bar{x}$ and denote it by $\bar{g}_\mathcal{U}$:

$$\bar{g}_\mathcal{U} := \bar{U}^\top g \quad \text{for any } g \in \bar{\partial} f(\bar{x}). \tag{2.2}$$

In addition, note that $\mathcal{V}$, depending on $\bar{x}$, is independent of the particular choice of $g \in \bar{\partial} f(\bar{x})$. Furthermore, when f has pdg structure it is possible to completely characterize $\mathcal{V}$ in terms of the gradients of the primal functions. More precisely, relations (4.3) and (4.5) and Lemma 4.1 in [8] imply the following.

LEMMA 2.1. *If f has* pdg *structure at $\bar{x}$ relative to $\mathcal{P}$, then*

$$\nabla f_i(\bar{x}) \in \bar{\partial} f(\bar{x}) \text{ for } i = 0, 1, \ldots, m_1, \nabla \varphi_\ell(\bar{x}) \in \mathcal{V} \text{ for } \ell = 1, \ldots, m_2, \tag{2.3}$$

and the subspace $\mathcal{V}$ from (2.1) *can be written as*

$$\mathcal{V} = \operatorname{lin}\left(\{\nabla f_i(\bar{x}) - \nabla f_0(\bar{x})\}_{i=0}^{m_1} \cup \{\nabla \varphi_\ell(\bar{x})\}_{\ell=1}^{m_2}\right).$$

3. Specific functions with pdg structure. In general, pdg structure is not unique, and it is desirable to have $\mathcal{P}$ as large as possible in order to exhibit interesting structure of a function. The following examples illustrate three different ways to define $\mathcal{P}$. The initial example shows that it is possible for $\mathcal{P}$ to be larger than $\{x \in B(\bar{x}) : \varphi_\ell(x) = 0 \text{ for } \ell = 1, \ldots, m_2\}$, whereas the second example has $\mathcal{P}$ equal to this set. In addition, the first example provides a function for which $\bar{\partial} f(\bar{x})$ is unbounded.

Example 3.1. For $x \in \mathbb{R}$ let

$$f(x) := \begin{cases} -x & \text{if } |x| \le 1, \\ -x + \sqrt{x^2 - 1} & \text{if } |x| \ge 1, \end{cases}$$

and let $\bar{x} := 1$. This function from [1] is the *spectral abscissa* (the largest real part of all eigenvalues) of the matrix

$$\begin{bmatrix} 0 & 1 \\ -1 & -2x \end{bmatrix},$$

and $\bar{x} = 1$ is a minimizing point at which f is nondifferentiable.

Take $B(\bar{x}) := (0, 2)$, $m_1 := 0$, $f_0(x) := -x$, $m_2 := 1$, $\varphi_1(x) := x^2 - 1$, $\mathcal{P} := \{x \in B(\bar{x}) : \varphi_1(x) \le 0\} = (0, 1]$, and $\Delta := \{(\alpha_0, \alpha_1) : \alpha_0 = 1\,, \alpha_1 \ge 0\}$. Then $f_0(1) = -1 = f(1)$, $\varphi_1(1) = 0$, and

$$\begin{aligned} \bar{\partial} f(1) &= \{\gamma : -1 \le \gamma\} \\ &= \{\alpha_0(-1) + \alpha_1(2) : \alpha_0 = 1\,, \alpha_1 \ge 0\} \\ &= \{\alpha_0 f_0'(1) + \alpha_1 \varphi_1'(1) : (\alpha_0, \alpha_1) \in \Delta\}, \end{aligned}$$

$\mathcal{V} = \mathbb{R}$, and $\mathcal{U} = \{0\}$. Finally, for $x \in \mathcal{P} \backslash \bar{x} = (0, 1)$, $f_0(x) = -x = f(x)$, $\varphi_1(x) = x^2 - 1 \ne 0$, and

$$\bar{\partial} f(x) = \{-1\} = \{\alpha_0 f_0'(x) + \alpha_1 \varphi_1'(x) : (\alpha_0, \alpha_1) \in \Delta\,, \alpha_1 = 0\}\,,$$

so f has the desired pdg structure.

Example 3.2. For $x \in \mathbb{R}^n$ let $f(x)$ be the maximum eigenvalue of a symmetric matrix whose elements are C^2-functions of x. Then from the analysis in [8, sect. 3.2] f has pdg structure at any $\bar{x} \in \mathbb{R}^n$, because the subdifferential used there, from [12], is the Clarke subdifferential for this locally Lipschitz function. At each $\bar{x}$, $m_1 + 1$ is the multiplicity of the maximum eigenvalue and $m_2 = m_1(m_1 + 1)/2$. The f_i's and φ_ℓ's depend on some C^2 vector functions that form a basis for a subspace spanned by certain eigenvectors; see [17]. The set Δ corresponds to a set of positive semidefinite matrices having unit traces and $\mathcal{P} = \{x \in B(\bar{x}) : \varphi_\ell(x) = 0 \text{ for } \ell = 1, \ldots, m_2\}$.

Finally, we give a pdg structure for a locally Lipschitz function $f : \mathbb{R}^2 \to \mathbb{R}$ from [5, sect. 7].

Example 3.3. For $x = (x_1, x_2)^\top$ consider the following function, defined on a partition of $\mathbb{R}^2$:

$$f(x) := \begin{cases} f_0(x) & \text{on} \quad S_0 := \{(x_1, x_2) \in \mathbb{R}^2 : x_2 \le 0\}, \\ h_1(x) & \text{on} \quad S_1 := \{(x_1, x_2) \in \mathbb{R}^2 : 0 < x_2 < 2x_1^2\}, \\ f_2(x) & \text{on} \quad S_2 := \{(x_1, x_2) \in \mathbb{R}^2 : 0 < 2x_1^2 \le x_2 \le 4x_1^2\}, \\ f_3(x) & \text{on} \quad S_3 := \{(x_1, x_2) \in \mathbb{R}^2 : 4x_1^2 < x_2\}, \end{cases}$$

where

$$\begin{aligned} f_0(x) &:= x_1^2 - x_2, \\ h_1(x) &:= \sqrt{x_1^4 + 2x_1^2 x_2 - x_2^2}, \\ f_2(x) &:= 3x_1^2 - x_2, \\ f_3(x) &:= -5x_1^2 + x_2, \end{aligned}$$

and the sets S_i defined here correspond to the sets S_{i+1} in [5].

Take $\bar{x} = (0, 0)^\top$, and note that $f(\bar{x}) = 0$. As shown in [5],

$$\bar{\partial} f(\bar{x}) = \operatorname{con}\{(0, -1)^\top, (0, 1)^\top\}.$$

Thus

$$\mathcal{V} = \operatorname{lin}\{(0, 1)^\top\} \quad \text{and} \quad \mathcal{U} = \operatorname{lin}\{(1, 0)^\top\},$$

and letting $\bar{U} = \left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right]$ gives in (2.2) $\bar{g}_{\mathcal{U}} = 0$. Furthermore, since $(0, 0)^\top \in \bar{\partial} f(\bar{x})$, $\bar{x}$ is a stationary point for f.

The difficulty in determining a pdg structure for f at $\bar{x}$ comes from the function h_1 that is not C^1 on any ball about $\bar{x}$ and yet defines f on S_1. Accordingly, we omit S_1 by taking $\mathcal{P} := S_0 \cup S_2 \cup S_3$ and consider f's behavior on the regions where the closure of S_1 intersects S_0 or S_2, i.e., on

$$\mathcal{M}_1 := \{(x_1, x_2) : x_2 = 0\} \tag{3.1}$$

and on $\mathcal{R} := \{(x_1, x_2) : x_2 = 2x_1^2, x_1 \neq 0\}$. The gradient of h_1 is given by

$$\nabla h_1(x) = \left(1 + 2\frac{x_2}{x_1^2} - \left(\frac{x_2}{x_1^2}\right)^2\right)^{-1/2} \left(2x_1\left(1 + \frac{x_2}{x_1^2}\right), 1 - \frac{x_2}{x_1^2}\right)^\top.$$

Since on $\mathcal{R}$, $\nabla h_1(x) = (6x_1, -1)^\top = \nabla f_2(x)$, f has a continuous gradient and there is no loss with respect to $\mathcal{R}$ from omitting h_1. Thus, we replace h_1 by a C^2-function f_1 whose value and gradient agree with those of h_1 on $\mathcal{M}_1 \backslash \bar{x}$. Since

$$h_1(x_1, 0) = x_1^2 \quad \text{and} \quad \nabla h_1(x_1, 0) = (2x_1, 1)^\top,$$

the desired function f_1 is given by

$$f_1(x) := x_1^2 + x_2.$$

In order to show that f has a pdg structure at $\bar{x}$ relative to $\mathcal{P}$, we let the ball $B(\bar{x})$ be $\mathbb{R}^2$ and consider the primal functions f_0, f_1, f_2, f_3 as defined above (so $m_1 = 3$ and $m_2 = 0$). As for the dual multiplier set Δ, it is the canonical simplex Δ_1 in $\mathbb{R}^4$.

We now check satisfaction of the three items defining pdg structure in section 2.1. Items (i), (ii), and (iii)(a) follow immediately from the definition of f and those of the structure objects f_i, $\mathcal{P}$, and Δ.

We conclude by demonstrating satisfaction of (iii)(b). The gradients of the primal functions are given by

$$\nabla f_0(x)=\begin{pmatrix}2x_1\\-1\end{pmatrix},\quad \nabla f_1(x)=\begin{pmatrix}2x_1\\1\end{pmatrix},\quad \nabla f_2(x)=\begin{pmatrix}6x_1\\-1\end{pmatrix},\quad \nabla f_3(x)=\begin{pmatrix}-10x_1\\1\end{pmatrix}. \tag{3.2}$$

First consider the case $x \in S_0$, with $x \neq \bar{x}$ (for $x = \bar{x}$, this item is easily verified). Because $\nabla h_1(x_1, 0) = (2x_1, 1)^\top$ we have that

$$\bar{\partial} f(x) = \begin{cases} \operatorname{con}\left((2x_1,1)^\top, (2x_1,-1)^\top\right) & \text{for } x_2 = 0, \\ \{(2x_1,-1)^\top\} & \text{for } x_2 < 0. \end{cases} \tag{3.3}$$

To write the subdifferential with pdg structure, the multipliers α_i must satisfy complementary slackness conditions, as given in the following table:

$x \in S_0 \backslash \bar{x}$	$f_i = f$	$f_i \neq f$	$\alpha_i = 0$
$x_2 = 0, x_1 \neq 0$	f_0, f_1	f_2, f_3	α_2, α_3
$x_2 < 0, x_1 \neq 0$	f_0	f_1, f_2, f_3	$\alpha_1, \alpha_2, \alpha_3$
$x_2 < 0, x_1 = 0$	f_0, f_2	f_1, f_3	α_1, α_3

As a result, the respective convex combinations of gradients are given by

$x \in S_0 \backslash \bar{x}$	$\sum_i \alpha_i \nabla f_i$
$x_2 = 0, x_1 \neq 0$	$\alpha_0 \nabla f_0(x) + (1-\alpha_0)\nabla f_1(x)$
$x_2 < 0, x_1 \neq 0$	$\nabla f_0(x)$
$x_2 < 0, x_1 = 0$	$\alpha_0 \nabla f_0(x) + (1-\alpha_0)\nabla f_2(x) = \nabla f_0(x)$

where the last equality follows from the fact that f_0 and f_2 have the same gradient if $x_1 = 0$. Comparing these results to (3.3) via (3.2) shows that item (iii)(b) holds for all $x \in S_0$.

Similar calculations can be carried out for S_2 and S_3, whose closures intersect at

$$\mathcal{M}_2 := \{(x_1, x_2) : x_2 = 4x_1^2\}. \tag{3.4}$$

Table 3.1 summarizes the main results from such calculations.

TABLE 3.1
Results for S_2 and S_3.

$x(\neq \bar{x})$	$\partial f(x)$	$f_i = f$	$f_i \neq f$	$\alpha_i = 0$
$x \in S_2 \backslash \mathcal{M}_2$	$\nabla f_2(x)$	f_2	f_0, f_1, f_3	$\alpha_0, \alpha_1, \alpha_3$
$x \in \mathcal{M}_2$	$\operatorname{con}(\nabla f_2(x), \nabla f_3(x))$	f_2, f_3	f_1, f_2	α_1, α_2
$x \in S_3$	$\nabla f_3(x)$	f_3	f_0, f_1, f_2	$\alpha_0, \alpha_1, \alpha_2$

4. Basic trajectories and multipliers. Throughout the remainder of the paper we assume that f has pdg structure at $\bar{x}$ relative to $\mathcal{P}$. Lemma 2.1 gives a spanning set of generators for $\mathcal{V}$ based on the structure. Next we consider linearly independent subsets of these generating vectors which lead to defining associated smooth trajectories and multiplier functions.

4.1. Basic index sets. An index set K of the form $K = K_f \cup K_\varphi \subseteq \{0, 1, \ldots, m_1\} \cup \{m_1 + 1, \ldots, m_1 + m_2\}$ with $0 \in K_f$ is called a *basic index set* if the $(n+1)$-dimensional vectors

$$\left\{ \begin{bmatrix} \nabla f_i(\bar{x}) \\ 1 \end{bmatrix} \right\}_{i \in K_f} \bigcup \left\{ \begin{bmatrix} \nabla \varphi_{i-m_1}(\bar{x}) \\ 0 \end{bmatrix} \right\}_{i \in K_\varphi}$$

are linearly independent.

In the theory that follows, we always suppose that $K = K_f \cup K_\varphi$ is a basic index set and, if necessary (with the exception of Example 3.3), the f_i's are reindexed so that the nonempty set K_f contains $i = 0$.

Associated with a basic index set K we define a full column rank $n \times (|K_f| - 1 + |K_\varphi|)$ matrix

$$\bar{V} := [\{\nabla f_i(\bar{x}) - \nabla f_0(\bar{x})\}_{0 \neq i \in K_f} \cup \{\nabla \varphi_{i-m_1}(\bar{x})\}_{i \in K_\varphi}].$$

This is a basis matrix for the $|K_f| - 1 + |K_\varphi|$ dimensional subspace of $\mathcal{V}$ defined by

$$\mathcal{V}_K := \operatorname{lin}\left(\{\nabla f_i(\bar{x}) - \nabla f_0(\bar{x})\}_{i \in K_f} \cup \{\nabla \varphi_{i-m_1}(\bar{x})\}_{i \in K_\varphi}\right) \tag{4.1}$$

with the interpretation that if K is a singleton, then $\bar{V}$ is vacuous and $\mathcal{V}_K = \{0\}$.

When K is such that $\mathcal{V}_K = \mathcal{V}$, we say that K is a *transversal* basic index set.

REMARK 4.1. *Recall that $\bar{U}$ is an orthogonal basis matrix for $\mathcal{U}$. Corresponding to a basic index set K and its associated subspace $\mathcal{V}_K \subseteq \mathcal{V}$ from* (4.1)*, if K is transversal, then $\bar{V}$ is a basis matrix for $\mathcal{V}$ which is not necessarily orthonormal. These basis matrices can be used to express the identity matrix in $\mathbb{R}^n$ as the sum of the projections onto the subspaces $\mathcal{U}$ and $\mathcal{V} = \mathcal{V}_K$:*

$$I = \bar{U}\bar{U}^\top + \bar{V}[\bar{V}^\top \bar{V}]^{-1}\bar{V}^\top. \tag{4.2}$$

Accordingly, for any $x \in \mathbb{R}^n$, $\bar{U}^\top x$ and $[\bar{V}^\top \bar{V}]^{-1}\bar{V}^\top x$ are, respectively, the $\mathcal{U}$ and $\mathcal{V}$ components of x.

Example (continuation of Example 3.3*).* For convenience we do not reindex the f_i's so that index $i = 0$ is always in a basic index set K. Since for this function the vectors $\left[\begin{smallmatrix} \nabla f_i(\bar{x}) \\ 1 \end{smallmatrix}\right]$ are equal to

$$\begin{bmatrix} 0 \\ (-1)^{i+1} \\ 1 \end{bmatrix}$$

for $i = 0, \ldots, 3$, there are no basic index sets with three or four elements. All of the singleton index sets are basic, but none of them is transversal, because $dim\, \mathcal{V} = 1$. Finally, the only two-element sets that are basic are

$$\{0, 1\}, \{0, 3\}, \{1, 2\}, \text{ and } \{2, 3\}.$$

These four sets are all transversal and for each of them we take $\bar{V} = \left[\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right]$ to be the corresponding basis matrix for $\mathcal{V}$.

4.2. Smooth trajectories. The purpose of introducing basic index sets is to identify trajectories along which f behaves in a smooth manner. Our next result shows how to find a smooth trajectory that is tangent to $\mathcal{U}$ using the `pdg` structure of f and an implicit function theorem to parameterize the trajectory in terms of $u \in \mathbb{R}^{dim\, \mathcal{U}}$.

From here on, we assume that $\mathcal{U} \neq \{0\}$ so that $dim\, \mathcal{U} \geq 1$ and $\mathcal{V} \neq \mathbb{R}^n$.

THEOREM 4.2. *Let f have* `pdg` *structure at $\bar{x}$ relative to $\mathcal{P}$, and suppose $K = K_f \cup K_\varphi$ is a nonsingleton basic index set. For all u small enough,*

(i) *the nonlinear system with variables* (u, v)

$$\begin{cases} f_i(\bar{x} + \bar{U}u + \bar{V}v) - f_0(\bar{x} + \bar{U}u + \bar{V}v) = 0, & 0 \neq i \in K_f, \\ \varphi_{i-m_1}(\bar{x} + \bar{U}u + \bar{V}v) = 0, & i \in K_\varphi, \end{cases} \tag{4.3}$$

has a unique solution $v = v_K(u)$*, where* $v_K : \mathbb{R}^{dim\, \mathcal{U}} \to \mathbb{R}^{dim\, \mathcal{V}_K}$ *is a* C^2*-function;*

(ii) *the trajectory*

$$\chi(u) := \bar{x} + \bar{U}u + \bar{V}v_K(u), \tag{4.4}$$

has a C^1 *Jacobian* $J\chi(u) = \bar{U} - \bar{V}(V(u)^\top \bar{V})^{-1} V(u)^\top \bar{U}$*, where*

$$V(u) := \left[\{\nabla f_i(\chi(u)) - \nabla f_0(\chi(u))\}_{0 \neq i \in K_f} \bigcup \{\nabla \varphi_{i-m_1}(\chi(u))\}_{i \in K_\varphi} \right];$$

(iii) *in particular,* $v_K(0) = 0$*,* $\chi(0) = \bar{x}$*,* $V(0) = \bar{V}$*,* $Jv_K(0) = 0$*,* $J\chi(0) = \bar{U}$*;*

(iv) $v_K(u) = O(|u|^2)$ *and the trajectory* $\chi(u)$ *is tangent to* $\mathcal{U}$ *at* $\chi(0) = \bar{x}$*.*

Proof. These results follow from the asssumption that the structure functions are C^2 along the lines of [8, Thm. 5.1], by applying a second-order implicit function theorem; see, for example, [3, p. 364]. ∎

When K is a singleton, $\bar{V}$ and $V(u)$ are vacuous. In this case, it is useful to define $\bar{V}v_K(u) := 0$ so that $\chi(u) = \bar{x} + \bar{U}u$ with Jacobian $\bar{U}$.

REMARK 4.3. *When* f *is a convex function (not necessarily* pdg*-structured) the* $\mathcal{U}$*-Lagrangian introduced in* [4] *provides an alternative way to obtain smooth trajectories. More specifically, when* $\bar{V}$ *is a basis matrix for* $\mathcal{V}$*, trajectories* $\chi(u) = \bar{x} + \bar{U}u + \bar{V}v(u)$ *are given by means of the following minimization problem:*

$$v(u) \in \operatorname*{Argmin}_{v \in \mathbb{R}^{dim\, \mathcal{V}}} \{f(\bar{x} + \bar{U}u + \bar{V}v) - y^\top \bar{V}v\} = \operatorname*{Argmin}_{v \in \mathbb{R}^{dim\, \mathcal{V}}} \{f(\bar{x} + \bar{U}u + \bar{V}v) - z^\top v\},$$

where $y \in \partial f(\bar{x})$ *and* $z = \bar{V}^\top y$*. Such smooth trajectories give a convex function, called the* $\mathcal{U}$*-Lagrangian, which is defined by*

$$L_{\mathcal{U}}(u; z) := f(\chi(u)) - z^\top v(u). \tag{4.5}$$

When a $\mathcal{U}$*-Lagrangian depending on* $z = \bar{V}^\top y$ *has a Hessian in* u*, it is possible to obtain an expansion for* f *that is superlinear in* u*, since* $v(u) = o(|u|)$*.*

When f *is* pdg*-structured (not necessarily convex) the solution of system* (4.3)*, depending on a basic index set* K*, gives* $v_K(u)$ *and a corresponding trajectory. In section* 5 *below we give an associated "*$\mathcal{U}$*-Lagrangian-related" function whose definition is analogous to* (4.5)*, except that it depends on* K*. This function is denoted by* $L_K(u; z)$ *and is called a "primal-dual" function. From* (4.5)*, since both* $L_K(u; z)$ *and* $v_K(u)$ *are* C^2*-functions, we are able to obtain an expansion for* f *that is of second order in* u *(rather than only superlinear as in the case without special structure); see Theorems* 5.1 *and* 5.2 *below. When the* $\mathcal{V}$*-minimality condition of Theorem* 6.1 *below is satisfied, the function* $v_K(u)$ *becomes an above* $v(u)$*, even for a nonconvex* f*; see* [8, *sect.* 6] *for sufficient conditions for* $\mathcal{V}$*-minimality when* f *is convex.*

We will show in section 7 below that a pdg-structured function f is partly smooth relative to certain manifolds $\mathcal{M}_K$. More precisely, depending on a basic index set K we consider the following smooth manifold:

$$\mathcal{M}_K := \left\{ x \in B(\bar{x}) : \left\{ \begin{array}{rcll} f_i(x) - f_0(x) & = & 0, & 0 \neq i \in K_f \\ \varphi_{i-m_1}(x) & = & 0, & i \in K_\varphi \end{array} \right\} \right. . \tag{4.6}$$

In order to establish a relation between a manifold $\mathcal{M}_K$ and the function f, the set K is required to satisfy a feasibility assumption depending on the following definition.

A basic index set K with corresponding trajectory $\chi(u)$ is *primal feasible* if for all $u \in \mathbb{R}^{dim\,\mathcal{U}}$ small enough $\chi(u) \in \mathcal{P}$ and $f(\chi(u)) = f_i(\chi(u))$ for some (and hence all) $i \in K_f$.

LEMMA 4.4. *Let f have* pdg *structure at $\bar{x}$ relative to $\mathcal{P}$, and suppose K is a basic index set. Then the following hold:*

(i) $N_{\mathcal{M}_K}(\bar{x}) = \mathcal{V}_K$, *the subspace defined in* (4.1).

(ii) *If K is transversal, then for all x close enough to $\bar{x}$*

$$x \in \mathcal{M}_K \iff x = \bar{x} + \bar{U}u(x) + \bar{V}v_K(u(x)), \text{ where } u(x) := \bar{U}^\top(x - \bar{x}).$$

Furthermore, if K is also primal feasible, then for $x \in \mathcal{M}_K$

(a) $x \in \mathcal{P}$ *and*

(b) *if $i \in \{0, 1, \ldots, m_1 + m_2\}$ is such that $f_i(x) \neq f(x)$ or $\varphi_{i-m_1}(x) \neq 0$, then $i \notin K$.*

Proof. From [15, p. 203], the subspace $\mathcal{V}_K$ as defined in (4.1) is the normal cone to $\mathcal{M}_K$ at $\bar{x}$, so (i) follows. Now we show that (ii) holds.

By the transversality assumption, $N_{\mathcal{M}_K}(\bar{x}) = \mathcal{V}$ and $\bar{V}$ is a basis matrix for $\mathcal{V}$. Take any $x \in \mathcal{M}_K$, and define $v(x) := [\bar{V}^\top\bar{V}]^{-1}\bar{V}^\top(x - \bar{x})$. From (4.2),

$$x - \bar{x} = \bar{U}\bar{U}^\top(x - \bar{x}) + \bar{V}[\bar{V}^\top\bar{V}]^{-1}\bar{V}^\top(x - \bar{x}), \text{ so } x = \bar{x} + \bar{U}u(x) + \bar{V}v(x).$$

With this notation, by the definition of $\mathcal{M}_K$ in (4.6), $(u, v) = (u(x), v(x))$ satisfies nonlinear system (4.3). Since the solution $v_K(u)$ is unique for a given u, $v(x) = v_K(u(x))$, so any $x \in \mathcal{M}_K$ has the form $x = \bar{x} + \bar{U}u(x) + \bar{V}v_K(u(x))$. The converse is immediate from (4.6) and Theorem 4.2(i) with $u = u(x)$. The remaining results follow from the assumption that K is primal feasible. ∎

Example (continuation of Example 3.3*).* Although singleton basic sets could be considered, we focus on the two-element sets that are basic and transversal, i.e., on $\{0, 1\}$, $\{0, 3\}$, $\{1, 2\}$, and $\{2, 3\}$.

Next we perform some calculations for $K = \{2, 3\}$ only and give a summary for all four sets in Table 4.1.

Since for this example

$$\bar{x} + \bar{U}u + \bar{V}v = \begin{pmatrix} 0 \\ 0 \end{pmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u + \begin{bmatrix} 0 \\ 1 \end{bmatrix} v = \begin{pmatrix} u \\ v \end{pmatrix},$$

when $K = \{2, 3\}$, system (4.3) corresponds to (recall that index $0 \notin K$ here)

$$f_2(u, v) - f_3(u, v) = 0 \quad \iff \quad 3u^2 - v - (-5u^2 + v) = 0,$$

TABLE 4.1
Trajectories for Example 3.3.

K	$v_K(u)$	$\chi(u) \in S_i$	$f(\chi(u))$	$f_0(\chi(u))$	$f_1(\chi(u))$	$f_2(\chi(u))$	$f_3(\chi(u))$
$0, 1$	0	$\binom{u}{0} \in S_0$	u^2	u^2	u^2	$3u^2$	$-5u^2$
$0, 3$	$3u^2$	$\binom{u}{3u^2} \in S_2$	0	$-2u^2$	$4u^2$	0	$-3u^2$
$1, 2$	u^2	$\binom{u}{u^2} \in S_1$	$\sqrt{2}u^2$	0	$2u^2$	$2u^2$	$-4u^2$
$2, 3$	$4u^2$	$\binom{u}{4u^2} \in S_2$	$-u^2$	$-3u^2$	$5u^2$	$-u^2$	$-u^2$

which is solved for v by $v_K(u) = 4u^2$. Therefore, $\chi(u) = (u, 4u^2)^\top$. For $u \neq 0$, $\chi(u) \in S_2 \subset \mathcal{P}$, so $f(\chi(u)) = f_2(\chi(u)) = f_3(\chi(u))$. Thus, $K = \{2, 3\}$ is primal feasible.

Note that for none of the trajectories $\chi(u)$ from Table 4.1 is $f(\chi(u))$ equal to the maximum or the minimum over all four functions $f_i(\chi(u))$. Also, we see that only

$$K_1 := \{0, 1\} \quad \text{and} \quad K_2 := \{2, 3\} \tag{4.7}$$

are primal feasible basic index sets. Since they are also transversal, letting $u = \bar{U}x = x_1$, the manifolds corresponding to (4.6) given via Lemma 4.4 and Table 4.1 are

$$\{x = (x_1, x_2) \in \mathbb{R}^2 : f_0(x) = f_1(x)\} = \{(u, 0) : u \in \mathbb{R}\} = \mathcal{M}_1$$

and

$$\{x = (x_1, x_2) \in \mathbb{R}^2 : f_2(x) = f_3(x)\} = \{(u, 4u^2) : u \in \mathbb{R}\} = \mathcal{M}_2,$$

i.e., the smooth manifolds from (3.1) and (3.4), respectively, containing the points of nondifferentiability of f.

4.3. Multiplier functions. So far we have developed only a primal object $\chi(u)$ depending on the pdg structure primal functions and on a basic index set K. Now we turn our attention to an associated dual object that is also a smooth function of $u \in \mathbb{R}^{dim\, \mathcal{U}}$. It is a multiplier vector $\alpha(u)$ which depends on structure function gradients, as well as on $\chi(u)$, and on an arbitrary parameter vector y that need not be a subgradient at $\bar{x}$. The next result follows along the lines of Theorem 5.2 in [8].

THEOREM 4.5. *Suppose f has* pdg *structure at $\bar{x}$ with respect to $\mathcal{P}$. Corresponding to a basic index set $K = K_f \cup K_\varphi$ with trajectory $\chi(u) = \bar{x} + \bar{U}u + \bar{V}v_K(u)$ and to a parameter vector $y \in \mathbb{R}^n$, for each u small enough, the linear system with variables α_i*

$$\begin{cases} \bar{V}^\top \left(\displaystyle\sum_{i \in K_f} \alpha_i \nabla f_i(\chi(u)) + \sum_{i \in K_\varphi} \alpha_i \nabla \varphi_{i-m_1}(\chi(u)) \right) = \bar{V}^\top y \in \mathbb{R}^{|K|-1}, \\ \displaystyle\sum_{i \in K_f} \alpha_i = 1 \end{cases}$$

has a unique solution $\alpha = \alpha(u)$, given by

$$\{\alpha_i(u)\}_{0 \neq i \in K} = (\bar{V}^\top V(u))^{-1} \bar{V}^\top (y - \nabla f_0(\chi(u))),$$

$$\alpha_0(u) = 1 - \sum_{0 \neq i \in K_f} \alpha_i(u),$$

where $V(u)$ is defined in Theorem 4.2(ii) *if K is not a singleton and $\alpha_0(u) = 1$ otherwise.*

Note that multipliers $\alpha_i(u)$ depend on K and $\bar{V}^\top y$, so we should think of the abbreviated vector notation $\alpha(u)$ as standing for something like $\alpha_K(u; \bar{V}^\top y)$. An expression for the Jacobian of $\alpha(u)$ is given in [8].

Example (continuation of Example 3.3*).* Table 4.2 shows the results for obtaining multiplier functions for the basic index sets K_1 and K_2 defined in (4.7). The results follow from Table 4.1, (3.2), and the choice of $\bar{V} = \left[\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right]$ which implies that $\bar{V}^\top y = y_2$.

TABLE 4.2
Multipliers for Example 3.3 with $\bar{V}^\top = [0 \;\; 1]$.

	$K_1 = \{0,1\}$	$K_2 = \{2,3\}$
$\chi(u)$	$\begin{pmatrix} u \\ 0 \end{pmatrix}$	$\begin{pmatrix} u \\ 4u^2 \end{pmatrix}$
$\sum_{i\in K} \alpha_i \nabla f_i(\chi(u))$	$\alpha_0 \begin{pmatrix} 2u \\ -1 \end{pmatrix} + \alpha_1 \begin{pmatrix} 2u \\ 1 \end{pmatrix}$	$\alpha_2 \begin{pmatrix} 6u \\ -1 \end{pmatrix} + \alpha_3 \begin{pmatrix} -10u \\ 1 \end{pmatrix}$
Linear system in Theorem 4.5	$\begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix} = \begin{pmatrix} y_2 \\ 1 \end{pmatrix}$	same, except the variables are α_2, α_3
$\alpha(u)$	$\alpha_0(u) = (1 - y_2)/2$ $\alpha_1(u) = (1 + y_2)/2$	$\alpha_2(u) = (1 - y_2)/2$ $\alpha_3(u) = (1 + y_2)/2$

5. Primal-dual functions; second-order order expansions. Consider a primal feasible basic index set K with associated quantities $\bar{V}$, $v_K(u)$, and $\chi(u)$, and let $z = \bar{V}^\top y$, where y is a parameter vector in $\mathbb{R}^n$. In view of Remark 4.3, in particular definition (4.5), we define for $u \in \mathbb{R}^{dim\, \mathcal{U}}$ small enough the following primal-dual function:

$$L_K(u; z) := f(\chi(u)) - z^\top v_K(u). \tag{5.1}$$

Next we show that this function is C^2 and give explicit expressions for its first and second derivatives with respect to u.

THEOREM 5.1. *Let f have* pdg *structure at $\bar{x}$ with respect to $\mathcal{P}$. Suppose $K = K_f \cup K_\varphi$ is a primal feasible basic index set with corresponding trajectory $\chi(u)$ and basis matrix $\bar{V}$. Let $y \in \mathbb{R}^n$ be a parameter vector, and consider the primal-dual function defined in* (5.1) *and the multiplier functions $\alpha_i(u)$ from Theorem* 4.5 *corresponding to K and to $z = \bar{V}^\top y \in \mathbb{R}^{dim\, \mathcal{V}_K}$.*

Then for all u small enough we have the following:

(i) *$L_K(u; z)$ is a C^2-function of u satisfying the particular Lagrangian-like result that*

$$L_K(u; 0) = \left(\sum_{i\in K_f} \alpha_i(u) f_i(\chi(u)) + \sum_{i\in K_\varphi} \alpha_i(u) \varphi_{i-m_1}(\chi(u)) \right).$$

(ii) *The gradient of L_K is given by*

$$\nabla L_K(u; z) = \bar{U}^\top g_K(u; z),$$

where $g_K(u; z)$ is defined by

$$g_K(u; z) := \left(\sum_{i\in K_f} \alpha_i(u) \nabla f_i(\chi(u)) + \sum_{i\in K_\varphi} \alpha_i(u) \nabla \varphi_{i-m_1}(\chi(u)) \right). \tag{5.2}$$

(iii) *The Hessian of L_K is given by*

$$\nabla^2 L_K(u; z) = J\chi(u)^\top M_K(u; z) J\chi(u),$$

where $M_K(u; z)$ is the $n \times n$ matrix function defined by

$$M_K(u; z) := \sum_{i\in K_f} \alpha_i(u) \nabla^2 f_i(\chi(u)) + \sum_{i\in K_\varphi} \alpha_i(u) \nabla^2 \varphi_{i-m_1}(\chi(u)), \tag{5.3}$$

which depends on z via $\alpha(u)$ even when $u = 0$.

(iv) *In particular, for all* z, $L_K(0;z) = f(\bar{x})$, $\nabla L_K(0;z) = \bar{g}_{\mathcal{U}}$, *the* $\mathcal{U}$*-gradient of* f *at* $\bar{x}$ *defined in* (2.2), *and* $\nabla^2 L_K(0;z) = \bar{U}^\top M_K(0;z)\bar{U}$.

Proof. Since K is primal feasible, using (5.1) with $z = 0$ gives

$$L_K(u;0) = f_i(\chi(u)) \qquad \text{for all } i \in K_f.$$

In addition, since

$$0 = \varphi_{i-m_1}(\chi(u)) \qquad \text{for all } i \in K_\varphi,$$

multiplying each of the above equations by the appropriate multiplier $\alpha_i(u)$ and then summing and recalling that $\sum_{i\in K_f} \alpha_i(u) = 1$ gives the Lagrangian-like expression in item (i).

The result that L_K is C^2 in u and the next two corresponding items follow along the lines of [8, Thm. 6.3(iii)–(iv)], with $L_{\mathcal{U}}$ and $\bar{g}$ therein replaced by L_K and y, respectively.

Item (iv) follows because $v_K(0) = 0$, $\chi(0) = \bar{x}$, $J\chi(0) = \bar{U}$, $\sum_{i\in K_f} \alpha_i(0) = 1$, and for each $i \in K$, by (2.3),

$$\begin{array}{llllll} \text{either} & \nabla f_i(\bar{x}) \in \bar{\partial} f(\bar{x}) & \text{so} & \bar{U}^\top \nabla f_i(\bar{x}) & = & \bar{g}_{\mathcal{U}} \\ \text{or} & \nabla \varphi_{i-m_1}(\bar{x}) \in \mathcal{V} & \text{so} & \bar{U}^\top \nabla \varphi_{i-m_1}(\bar{x}) & = & 0. \end{array} \qquad \square$$

If K is primal feasible and y is such that $z = \bar{V}^\top y = 0$, we call the corresponding Hessian of L_K at $u = 0$ a *basic* $\mathcal{U}$*-Hessian for* f *at* $\bar{x}$ and denote it by $\bar{H}_K := \nabla^2 L_K(0;0)$. Second-order $\mathcal{U}$-derivatives are useful for specifying second-order expansions for f and giving related necessary conditions for optimizers, as is shown next.

THEOREM 5.2. *Let* f *have* pdg *structure at* $\bar{x}$ *with respect to* $\mathcal{P}$. *Suppose* K *is a primal feasible basic index set with corresponding trajectory* $\chi(u) = \bar{x} + \bar{U}u + \bar{V}v_K(u)$. *Then for all* u *small enough and all* $y \in \mathbb{R}^n$

$$f(\chi(u)) = f(\bar{x}) + \bar{g}_{\mathcal{U}}^\top u + y^\top \bar{V} v_K(u) + \tfrac{1}{2} u^\top \nabla^2 L_K(0; \bar{V}^\top y) u + o(|u|^2),$$

where $\bar{g}_{\mathcal{U}}$ *is the* $\mathcal{U}$*-gradient defined in* (2.2).

In particular, when $y = \bar{g} \in \bar{\partial} f(\bar{x})$,

$$f(\chi(u)) = f(\bar{x}) + \bar{g}^\top(\chi(u) - \bar{x}) + \tfrac{1}{2}(\chi(u) - \bar{x}))^\top \bar{U} \nabla^2 L_K(0; \bar{V}^\top \bar{g}) \bar{U}^\top (\chi(u) - \bar{x}) + o(|\chi(u) - \bar{x}|^2),$$

or, when $\bar{V}^\top y = 0$,

$$f(\chi(u)) = f(\bar{x}) + \bar{g}_{\mathcal{U}}^\top u + \tfrac{1}{2} u^\top \bar{H}_K u + o(|u|^2).$$

Proof. The first expansion follows from expanding $L_K(u;z)$ with $z = \bar{V}^\top y$ about $u = 0$ and using (5.1) and Theorem 5.1(ii), (iii), and (iv). The second then follows from (2.2) and the result from Theorem 4.2 that $v_K(u) = O(|u|^2)$. The third expansion is a consequence of the definition of $\bar{H}_K$. $\square$

COROLLARY 5.3. *Let* f *have* pdg *structure at* $\bar{x}$ *with respect to* $\mathcal{P}$, *and suppose* $\bar{x}$ *is a local minimizer of* f. *Then* $0 \in \bar{\partial} f(\bar{x})$ *and for any primal feasible basic index set* K *the associated basic* $\mathcal{U}$*-Hessian* $\bar{H}_K$ *is positive semidefinite.*

Example (continuation of Example 3.3*).* To compute the primal-dual functions and their derivatives given in Table 5.1, we substitute into (5.1) the relevant expressions from Table 4.1 and differentiate with respect to u. Equivalently, we could use the multipliers $\alpha(u)$ given in Table 4.2 and the formulas from Theorem 5.1. Recall that $z = \bar{V}^\top y = y_2$ for this example.

TABLE 5.1
Primal-dual information for Example 3.3.

	$K_1 = \{0,1\}$	$K_2 = \{2,3\}$
$L_K(u; y_2)$	u^2	$-u^2 - 4y_2u^2$
$\nabla L_K(u; y_2)$	$2u$	$-2(1+4y_2)u$
$\nabla^2 L_K(u; y_2)$	2	$-2(1+4y_2)$
$\bar{H}_K$	2	-2

6. Link to second-order epi-derivatives. We now show that some $\mathcal{U}$-Hessians give second-order epi-derivatives for $\mathcal{U}$-subspace directions. Thus, primal-dual functions L_K, having ordinary second derivatives, can capture much of the second-order epi-differential behavior of a `pdg`-structured function.

We start by recalling the formal definition of epigraphical convergence.

6.1. Characterization of epi-limits and epi-derivatives. The epigraphical convergence theory developed in [15, Chap. 7] includes the following useful characterization of epi-limits.

Proposition 7.2 and equation 7(3) in [15]. Let $\{q^\nu\}$ be a sequence of functions on $\mathbb{R}^n$, and let w be any point in $\mathbb{R}^n$. The value $q(w)$ is the *epi-limit* of the sequence q^ν at w if and only if

$$\begin{cases} \liminf\limits_{\nu} \; q^\nu(w^\nu) \geq q(w) & \text{for every sequence } w^\nu \to w\,, \\ \limsup\limits_{\nu} \; q^\nu(w^\nu) \leq q(w) & \text{for some sequence } w^\nu \to w\,. \end{cases} \qquad \square$$

Note that in the expression above the limit $q(w)$ can be infinity.

For a function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ and point $\bar{x} \in dom\ h$ with $h(\bar{x})$ finite we consider various limits of the following first-order and second-order difference quotients:

$$\frac{h(\bar{x}+\tau\cdot) - h(\bar{x})}{\tau} \quad \text{and} \quad \frac{h(\bar{x}+\tau\cdot) - h(\bar{x}) - \tau y^\top \cdot}{\frac{1}{2}\tau^2}$$

for $\tau > 0$ and $y \in \mathbb{R}^n$.

The *(first) subderivative* of h at $\bar{x}$ in the direction w is denoted and defined by

$$\mathrm{d}h(\bar{x})(w) := \liminf_{\tau \searrow 0\,, w^\tau \to w} \frac{h(\bar{x}+\tau w^\tau) - h(\bar{x})}{\tau}\,.$$

When the first-order difference quotient has an epi-limit at w as $\tau \searrow 0$, then $\mathrm{d}h(\bar{x})(w)$ is this limit and it is then called the *first epi-derivative* of h at $\bar{x}$ in the direction w and h is said to be *epi-differentiable* at $\bar{x}$ for w.

Similarly, but for each $y \in \mathbb{R}^n$, when the second-order difference quotient has an epi-limit at w as $\tau \searrow 0$, it is denoted by $\mathrm{d}^2h(\bar{x}|y)(w)$ and is called the *second epi-derivative* of h at $\bar{x}$ relative to y in the direction w. In this case

$$\mathrm{d}^2h(\bar{x}|y)(w) = \liminf_{\tau \searrow 0\,, w^\tau \to w} \frac{h(\bar{x}+\tau w^\tau) - h(\bar{x}) - \tau y^\top w^\tau}{\frac{1}{2}\tau^2}\,,$$

generally called the *second subderivative* of h at $\bar{x}$ relative to y in the direction w.

6.2. Connection with $\mathcal{U}$-Hessians. It is important to note that, in the epigraphical setting, the second-order epi-derivative provides a second-order approximation $f(\bar{x}+\tau w) \approx f(\bar{x})+\tau y^\top w+\frac{1}{2}\tau^2 \mathrm{d}^2 f(\bar{x}|y)(w)$, not in the usual sense of local uniform convergence, but of closeness of the epigraphs of the second-order difference quotient function and $\mathrm{d}^2 f(\bar{x}|y)(\cdot)$; see [13]. In contrast, primal-dual functions $L_K(u;\bar{V}^\top y)$ provide second-order approximations in u in the classical sense. We now establish a relation between these second-order objects when a basic index set K satisfies a property relative to certain subgradients which for a convex function follows from the $\mathcal{V}$-optimality conditions given in [7, Section 6.1].

THEOREM 6.1. *Let f have* pdg *structure at $\bar{x}$ with respect to $\mathcal{P}$. Suppose K is a transversal and primal feasible basic index set with corresponding trajectory $\chi(u) = \bar{x} + \bar{U}u + \bar{V}v_K(u)$ and primal-dual Hessian $\nabla^2 L_K(u;z)$ from Theorem* 5.1(iii). *Let*

$$G_K := \left\{ \bar{g} \in \bar{\partial} f(\bar{x}) : \begin{array}{l} v_K(u) \in \operatorname*{Argmin}_{v \in \mathbb{R}^{dim\,\mathcal{V}}} \{f(\bar{x} + \bar{U}u + \bar{V}v) - \bar{g}^\top \bar{V}v\} \\ \text{for all small enough } u \in \mathbb{R}^{dim\,\mathcal{U}} \end{array} \right\}. \tag{6.1}$$

Then for all $w \in \mathbb{R}^n$ the corresponding first and second subderivatives, respectively, satisfy

$$\mathrm{d} f(\bar{x})(w) \geq \sup\{\bar{g}^\top w : \bar{g} \in G_K\}$$

and

$$\mathrm{d}^2 f(\bar{x}|\bar{g})(w) \geq w^\top \bar{U} \nabla^2 L_K(0; \bar{V}^\top \bar{g}) \bar{U}^\top w \quad \text{for all } \bar{g} \in G_K.$$

Furthermore, for all $w \in \mathcal{U}$ the corresponding first and second epi-derivatives, respectively, are given by

$$\mathrm{d} f(\bar{x})(w) = \bar{g}_{\mathcal{U}}^\top \bar{U}^\top w$$

and

$$\mathrm{d}^2 f(\bar{x}|\bar{g})(w) = w^\top \bar{U} \nabla^2 L_K(0; \bar{V}^\top \bar{g}) \bar{U}^\top w \quad \text{for all } \bar{g} \in G_K.$$

Proof. We start by showing the second-order results. Suppose $\bar{g} \in G_K$. Then, for all $v \in \mathbb{R}^{dim\,\mathcal{V}}$, $f(\chi(u)) - \bar{g}^\top \bar{V} v_K(u) \leq f(\bar{x} + \bar{U}u + \bar{V}v) - \bar{g}^\top \bar{V}v$. Subtracting $f(\bar{x}) + \bar{g}^\top \bar{U}u$ from both sides of the inequality gives

$$f(\chi(u)) - f(\bar{x}) - \bar{g}^\top(\chi(u) - \bar{x}) \leq f(\bar{x} + \bar{U}u + \bar{V}v) - f(\bar{x}) - \bar{g}^\top(\bar{U}u + \bar{V}v).$$

Then, since K is primal feasible, Theorem 5.2 written with $y = \bar{g}$ implies that for all $u \in \mathbb{R}^{dim\,\mathcal{U}}$ small enough and all $v \in \mathbb{R}^{dim\,\mathcal{V}}$

$$\frac{1}{2} u^\top \nabla^2 L_K(0; \bar{V}^\top \bar{g}) u + o(|u|^2) \leq f(\bar{x} + \bar{U}u + \bar{V}v) - f(\bar{x}) - \bar{g}^\top(\bar{U}u + \bar{V}v). \tag{6.2}$$

Take a sequence $w^\tau \to w$ as $\tau \searrow 0$, and let $u^\tau := \bar{U}^\top w^\tau$. Then, using (4.2) and the transversality of K,

$$w^\tau = \bar{U}\bar{U}^\top w^\tau + \bar{V}[\bar{V}^\top \bar{V}]^{-1} \bar{V}^\top w^\tau = \bar{U}u^\tau + \bar{V}[\bar{V}^\top \bar{V}]^{-1} \bar{V}^\top w^\tau.$$

With this notation

$$\bar{x} + \tau w^\tau = \bar{x} + \bar{U}(\tau u^\tau) + \bar{V}[\bar{V}^\top \bar{V}]^{-1} \bar{V}^\top (\tau w^\tau).$$

From (6.2) with $u = \tau u^\tau \in \mathbb{R}^{dim\,\mathcal{U}}$, τ small enough, and $v = [\bar{V}^\top \bar{V}]^{-1}\bar{V}^\top(\tau w^\tau) \in \mathbb{R}^{dim\,\mathcal{V}}$ we obtain

$$\frac{1}{2}(\tau u^\tau)^\top \nabla^2 L_K(0;\bar{V}^\top\bar{g})\tau u^\tau + o(|\tau u^\tau|^2) \le f(\bar{x}+\tau w^\tau) - f(\bar{x}) - \tau\bar{g}^\top w^\tau. \tag{6.3}$$

Dividing both sides of this inequality by $\frac{1}{2}\tau^2$ yields

$$(u^\tau)^\top \nabla^2 L_K(0;\bar{V}^\top\bar{g})u^\tau + \frac{o(\tau^2|u^\tau|^2)}{\frac{1}{2}\tau^2} \le \frac{f(\bar{x}+\tau w^\tau) - f(\bar{x}) - \tau\bar{g}^\top w^\tau}{\frac{1}{2}\tau^2}.$$

Note that since $w^\tau \to w$, the definition of u^τ implies that $u^\tau \to \bar{U}^\top w$. Hence, passing to the limit gives the following inequality involving the second subderivative:

$$w^\top \bar{U}\nabla^2 L_K(0;\bar{V}^\top\bar{g})\bar{U}^\top w \le \mathrm{d}^2 f(\bar{x}|\bar{g})(w).$$

To show that the left-hand side is an epi-limit for $w \in \mathcal{U}$ we exhibit below a sequence w_K^τ converging to w with the property that $\bar{x} + \tau w_K^\tau = \chi(\tau u^\tau)$. We have just shown, using Theorem 5.2, that the second-order difference quotient function of such a direction vector w_K^τ converges to $w^\top\bar{U}\nabla^2 L_K(0;\bar{V}^\top\bar{g})\bar{U}^\top w$, which therefore equals the second-order epi-derivative (by the above definition of epi-limit, since the liminf is greater than or equal to this limit).

Consider the sequence

$$w_K^\tau := \bar{U}u^\tau + \bar{V}\frac{1}{\tau}v_K(\tau u^\tau), \text{ which implies } \bar{x} + \tau w_K^\tau = \chi(\tau u^\tau).$$

By Theorem 4.2(iv), the term $\frac{1}{\tau}v_K(\tau u^\tau) = \frac{1}{\tau}O(\tau^2|u^\tau|^2)$ converges to 0 together with τ. As a result, $\lim_\tau w_K^\tau = \lim_\tau \bar{U}u^\tau = \lim_\tau \bar{U}\bar{U}^\top w^\tau = \bar{U}\bar{U}^\top w$. The result follows by noting from (4.2) that $w \in \mathcal{U} \iff w = \bar{U}\bar{U}^\top w$.

The first-order results are obtained in a similar manner, by dividing (6.3) by τ instead of $\frac{1}{2}\tau^2$ and noting that if $w \in \mathcal{U}$, then $\bar{g}^\top w = \bar{g}^\top\bar{U}\bar{U}^\top w = \bar{g}_{\mathcal{U}}^\top\bar{U}^\top w$ from (2.2). □

Remark 6.2. *When $\bar{g} \in G_K$, then from* (6.1) *and the definition of the minimizer $v(u)$ in Remark* 4.3, *whether or not f is convex, $L_K(u;\bar{V}^\top\bar{g})$ can be called the $\mathcal{U}$-Lagrangian relative to $\bar{V}^\top\bar{g}$, and hence $\nabla^2 L_K(0;\bar{V}^\top\bar{g})$ is the related $\mathcal{U}$-Hessian of f at $\bar{x}$. A similar generalization from the convex case to the nonconvex can be made for a fast track* [10], *i.e., if $G_K = \bar{\partial}f(\bar{x})$, then the corresponding $\chi(u)$ can be called a fast track. We note in passing that the dependence of the $\mathcal{U}$-Lagrangian on the pair $(u;z) \in \mathbb{R}^{dim\mathcal{U}} \times \mathbb{R}^{dim\mathcal{V}} = \mathbb{R}^n$ is exploited in* [10] *in order to apply an implicit function theorem to show that for a convex f proximal points are on the fast track.*

Example (continuation of Example 3.3*).* Recall that $\bar{\partial}f(\bar{x}) = \{\bar{g} = (0,\bar{g}_2) : -1 \le \bar{g}_2 \le 1\}$. We will see below that for the transversal and primal feasible sets K_1 and K_2 from (4.7) the associated sets G_{K_1} and G_{K_2} from (6.1) satisfy $\bar{\partial}f(\bar{x}) = G_{K_1} \cup G_{K_2}$.

In view of Table 4.1, we first find values of $\bar{g}_2 \in [-1,1]$ which correspond to $v_{K_1}(u) = 0$ or to $v_{K_2}(u) = 4u^2$ being in

$$\operatorname*{Argmin}_{v\in\mathbb{R}}\{f(u,v) - \bar{g}_2 v\}. \tag{6.4}$$

Table 6.1 below summarizes the main results for computing the various minimizers in (6.4). To obtain these results from the definition of f we partition $\mathbb{R}$ into four v-subintervals. For all $u \in \mathbb{R}$ and $\bar{g}_2 \in [-1,1]$, on each subinterval the objective in

TABLE 6.1
Subinterval v-minimizers for Example 3.3.

Subinterval	Objective in (6.4)	Argmin value	Objective min value
$v \le 0$	$u^2 - (1+\bar{g}_2)v$	$v = 0$ for all $\bar{g}_2$	u^2
$0 \le v \le 2u^2$	$h_1(u,v) - \bar{g}_2 v$	$\left[\begin{array}{l} v = 0 \text{ or } v = 2u^2, \\ \text{depending on } \bar{g}_2 \end{array}\right.$	$\min\{u^2, (1-2\bar{g}_2)u^2\}$
$2u^2 \le v \le 4u^2$	$3u^2 - (1+\bar{g}_2)v$	$v = 4u^2$ for all $\bar{g}_2$	$-(1+4\bar{g}_2)u^2$
$4u^2 \le v$	$-5u^2 + (1-\bar{g}_2)v$	$v = 4u^2$ for all $\bar{g}_2$	$-(1+4\bar{g}_2)u^2$

(6.4) is concave in v (actually, it is affine on three subintervals and strictly concave on $[0, 2u^2]$, since $\frac{\partial^2 h_1}{\partial v^2} < 0$ there), and hence corresponding minimizers are at subinterval endpoints.

From the last column in Table 6.1 the minimum objective value in (6.4) is

$$\min\{u^2, (1-2\bar{g}_2)u^2, -(1+4\bar{g}_2)u^2\} = \left\{\begin{array}{ll} u^2 & \text{for } \bar{g}_2 \in [-1, -1/2], \\ -(1+4\bar{g}_2)u^2 & \text{for } \bar{g}_2 \in [-1/2, 1]. \end{array}\right.$$

Combined with the third column, this implies that

$$G_{K_1} = \left\{\bar{g} = (0, \bar{g}_2)^\top : \bar{g}_2 \in [-1, -\tfrac{1}{2}]\right\} \text{ and } G_{K_2} = \left\{\bar{g} = (0, \bar{g}_2)^\top : \bar{g}_2 \in [-\tfrac{1}{2}, 1]\right\}.$$

Thus, from Theorem 6.1 with $\bar{U}^\top w = w_1$ and Table 5.1 with $y_2 = \bar{g}_2$ we find that for all $w = (w_1, 0)^\top \in \mathcal{U}$ and $\bar{g} = (0, \bar{g}_2)^\top \in \bar{\partial} f(\bar{x})$ the second-order epi-derivatives are given by

$$\mathrm{d}^2 f(\bar{x}|\bar{g})(w) = \left\{\begin{array}{ll} 2w_1^2 & \text{if } \bar{g}_2 \in [-1, -\frac{1}{2}], \\ -2(1+4\bar{g}_2)w_1^2 & \text{if } \bar{g}_2 \in [-\frac{1}{2}, 1]. \end{array}\right. \tag{6.5}$$

7. Link to partly smooth functions. The concept of partial smoothness, introduced in [5], is closely related to $\mathcal{VU}$-space decomposition and `pdg` structure. The smooth and sharp (generalized U- and V-shaped, respectively) behaviors of a function are decoupled via an active manifold as, for example, in (4.6).

More precisely (see [5, Def. 2.7]), given a set $\mathcal{M} \subset X$ that contains $\bar{x}$, the function $h : X \to \bar{\mathbb{R}}$ is *partly smooth at $\bar{x}$ relative to $\mathcal{M}$* if $\mathcal{M}$ is a manifold around $\bar{x}$ and the following four properties hold:

(i) (restricted smoothness). The restriction $h|_{\mathcal{M}}$ equals a representative function that is C^2 around $\bar{x}$.
(ii) (regularity). At every point close to $\bar{x}$ in $\mathcal{M}$, the function h is (subdifferentially) regular and has a subgradient.
(iii) (V-type sharpness). For all nonzero directions $w \in N_{\mathcal{M}}(\bar{x})$ the first subderivative satisfies $\mathrm{d}h(\bar{x})(w) > -\mathrm{d}h(\bar{x})(-w)$.
(iv) (subgradient continuity). The subdifferential map ∂h is continuous at $\bar{x}$ relative to $\mathcal{M}$.

We will show that regular `pdg`-structured functions are partly smooth relative to certain structure-defined manifolds. We start by showing the properties of V-type sharpness and restricted smoothness.

LEMMA 7.1. *Let f have* `pdg` *structure at $\bar{x}$ relative to $\mathcal{P}$, and suppose K is any basic index set with associated manifold $\mathcal{M}_K$ from* (4.6). *If f is regular at $\bar{x}$, then*

$$\mathrm{d}f(\bar{x})(w) > -\mathrm{d}f(\bar{x})(-w) \quad \textit{for all nonzero } w \in N_{\mathcal{M}_K}(\bar{x}).$$

Proof. Suppose w is a nonzero vector in $\mathcal{V}$. Let $\bar{g}$ be a subgradient in the relative interior of $\bar{\partial} f(\bar{x})$, which, from (2.1), is the interior of $\bar{\partial} f(\bar{x})$ relative to the affine set $g + \mathcal{V}$ for any subgradient g. Then,

$$\bar{g} + \tau w \in \bar{\partial} f(\bar{x}) \quad \text{for some } \tau > 0.$$

The regularity assumption implies, by [15, p. 337 and Thm. 8.30], that $\mathrm{d}f(\bar{x})(w) = \sup\{g^\top w : g \in \bar{\partial} f(\bar{x})\}$. Thus, for all nonzero $w \in \mathcal{V}$

$$\mathrm{d}f(\bar{x})(w) \geq \bar{g}^\top w + \tau |w|^2 > \bar{g}^\top w = -\bar{g}^\top(-w) \geq -\mathrm{d}f(\bar{x})(-w).$$

The proof is ended by noting, from Lemma 4.4(i), that $N_{\mathcal{M}_K}(\bar{x}) = \mathcal{V}_K \subseteq \mathcal{V}$. ∎

Next we show that the richness of pdg structure allows us to exhibit a *smooth representative of* $f|_{\mathcal{M}_K}$ (in terms of a primal-dual function with $z = 0$) and to give its first and second derivatives.

THEOREM 7.2. *Let f have* pdg *structure at $\bar{x}$ with respect to $\mathcal{P}$. Suppose K is a basic index set that is primal feasible and transversal. Then the following hold:*

(i) *$L_K(u(x); 0)$, with $u(x) := \bar{U}^\top(x - \bar{x})$, is C^2 around $\bar{x} \in \mathbb{R}^n$ and equals f on $\mathcal{M}_K$.*

(ii) *The gradient and Hessian of $L_K(u(x); 0)$ are given, respectively, by*

$$\bar{U}\bar{U}^\top g_K(u(x); 0) \quad \textit{and} \quad \bar{U}\nabla^2 L_K(u(x); 0)\bar{U}^\top.$$

These derivatives depend on expressions in Theorem 5.1(ii) *and* (iii) *and Theorem* 4.2(ii) *with u and $\chi(u)$ therein replaced by $u(x)$ and x, respectively.*

Proof. Take $x \in \mathcal{M}_K$ close enough to $\bar{x}$ so that $u(x)$ is small enough for $v_K(u(x))$ to be defined. For such $x \in \mathcal{M}_K$ Lemma 4.4(ii) gives $x = \chi(u(x))$, where $\chi(u)$ corresponds to K. Then (5.1) with $(u; z) = (u(x); 0)$ gives $L_K(u(x); 0) = f(x)$ for $x \in \mathcal{M}_K$ close enough to $\bar{x}$. Since the Jacobian of $u(x)$ is $\bar{U}^\top$, the results follow from Theorem 5.1 and the chain rule. ∎

In order to show continuity of the subdifferential on $\mathcal{M}_K$, we require an additional assumption on the basic index set K.

A basic index set $K = K_f \cup K_\varphi$ is called *dual feasible with respect to* $\bar{g} \in \bar{\partial} f(\bar{x})$ if there exists an $\bar{\alpha} \in \Delta$ such that

$$\bar{g} = \sum_{i \in K_f} \bar{\alpha}_i \nabla f_i(\bar{x}) + \sum_{i \in K_\varphi} \bar{\alpha}_i \nabla \varphi_{i-m_1}(\bar{x}) \text{ and } \bar{\alpha}_i = 0 \quad \text{for all } i \notin K. \tag{7.1}$$

THEOREM 7.3. *Let f have* pdg *structure at $\bar{x}$ with respect to $\mathcal{P}$. Suppose K is a basic index set that is dual feasible with respect to all $\bar{g} \in \bar{\partial} f(\bar{x})$. Then the following hold:*

(i) *K is transversal.*

(ii) *If, in addition, K is primal feasible, then $\bar{\partial} f$ is inner semicontinous at $\bar{x}$ relative to $\mathcal{M}_K$.*

Proof. To show (i), i.e., $\mathcal{V}_K = \mathcal{V}$, we need only to prove the inclusion $\mathcal{V} \subseteq \mathcal{V}_K$. Since the subgradient g used in defining $\mathcal{V}$ in (2.1) is arbitrary, take $g = \nabla f_0(\bar{x})$, which is in $\bar{\partial} f(\bar{x})$ by (2.3). Suppose $v \in \mathcal{V}$. Then there exist scalars β_j and subgradients $g^j \in \bar{\partial} f(\bar{x})$ such that

$$v = \sum_j \beta_j (g^j - \nabla f_0(\bar{x})). \tag{7.2}$$

By the assumption that K is dual feasible for all subgradients, there exist vectors $\alpha^j \in \Delta$ such that

$$g^j = \sum_{i\in K_f} \alpha_i^j \nabla f_i(\bar{x}) + \sum_{i\in K_\varphi} \alpha_i^j \nabla \varphi_{i-m_1}(\bar{x}) \text{ and } \alpha_i^j = 0 \quad \text{for all } i \notin K.$$

Then, from the definition of Δ, $\sum_{i\in K_f} \alpha_i^j = 1$ for each j, so

$$g^j - 1\nabla f_0(\bar{x}) = \sum_{i\in K_f} \alpha_i^j (\nabla f_i(\bar{x}) - \nabla f_0(\bar{x})) + \sum_{i\in K_\varphi} \alpha_i^j \nabla \varphi_{i-m_1}(\bar{x}).$$

Multiplying each of these equalities by β_j, and then summing over j, and using (7.2) gives

$$v = \sum_{i\in K_f} \gamma_i (\nabla f_i(\bar{x}) - \nabla f_0(\bar{x})) + \sum_{i\in K_\varphi} \gamma_i \nabla \varphi_{i-m_1}(\bar{x}),$$

where γ_i is defined by $\gamma_i := \sum_j \beta_j \alpha_i^j$ for $i \in K$. Thus, from (4.1), $v \in \mathcal{V}_K$.

To show (ii), let $\bar{g} \in \bar{\partial} f(\bar{x})$, and let $\bar{\alpha} \in \Delta$ be a multiplier vector satisfying (7.1) in the definition of dual feasibility. Let $\{x^r\} \subset \mathcal{M}_K$ be a sequence converging to $\bar{x}$. Then the sequence $\{g^r\}$ defined by

$$g^r := \sum_{i\in K_f} \bar{\alpha}_i \nabla f_i(x^r) + \sum_{i\in K_\varphi} \bar{\alpha}_i \nabla \varphi_{i-m_1}(x^r)$$

converges to $\bar{g}$ by (7.1) and the continuity of the primal function gradients. To complete the proof, we need to show that $g^r \in \bar{\partial} f(x^r)$ using condition (iii)(b) of the `pdg` structure definition in section 2.1. By item (i), K is transversal, so, by Lemma 4.4(ii), $x^r \in \mathcal{P}$ and if i is such that $f_i(x^r) \neq f(x^r)$ or $\varphi_{i-m_1}(x^r) \neq 0$, then $i \notin K$. From (7.1) $\bar{\alpha}_i = 0$ for $i \notin K$, so we have the complementary slackness between $x^r \in \mathcal{P}$ and $\bar{\alpha} \in \Delta$ needed to conclude from (7.1) and section 2.1(iii)(b) with $x = x^r$ and $\alpha = \bar{\alpha}$ that $g^r \in \bar{\partial} f(x^r)$. □

We are now in a position to give the main result of this section.

THEOREM 7.4. *Let f have* `pdg` *structure at $\bar{x}$ with respect to $\mathcal{P}$. Suppose K is a primal feasible basic index set that is dual feasible with respect to all $\bar{g} \in \bar{\partial} f(\bar{x})$. If f is regular at all $x \in \mathcal{M}_K$ close to $\bar{x}$, then f is partly smooth at $\bar{x}$ relative to $\mathcal{M}_K$.*

Proof. Lemma 7.1 and Theorems 7.2 and 7.3 show all of the conditions for f to be partly smooth, except for outer semicontinuity of $\bar{\partial} f$ at $\bar{x}$ relative to $\mathcal{M}_K$. By the regularity assumption, on $\mathcal{M}_K$ $\bar{\partial} f(\cdot)$ equals the outer semicontinuous map $\partial f(\cdot)$; see [15, Prop. 8.7]. □

Example (conclusion of Example 3.3*).* It is shown in [5] that this example f is subdifferentially regular everywhere. From (3.2) and (3.3) with $(x_1, x_2) = (0, 0)$ and the fact that Δ is the canonical simplex in $\mathbb{R}^4$, it is easy to show that both of the primal feasible basic index sets $K_1 = \{0, 1\}$ and $K_2 = \{2, 3\}$ are dual feasible for all $\bar{g} \in \bar{\partial} f(\bar{x})$. Hence, this function is partly smooth at $\bar{x}$ relative to both $\mathcal{M}_1$ and $\mathcal{M}_2$. From Theorem 7.2 with $\bar{U}^\top = [1 \;\; 0]$ and Table 5.1 with $y_2 = 0$ the respective corresponding manifold restricted Hessians are

$$\begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix} \text{ and } \begin{bmatrix} -2 & 0 \\ 0 & 0 \end{bmatrix}.$$

8. Conclusion. In this paper we have related $\mathcal{U}$-subspace second-order smoothness results for the broad class of pdg-structured functions to the general concepts of second-order epi-derivatives and partly smooth functions. The principal results given here are the second-order expansions in u depending on a parameter y that need not be a subgradient. They are contained in Theorem 5.2 and based on a primal-dual function corresponding to a basic index set that is assumed only to be primal feasible. Without having to develop second-order epi-derivatives (as in Theorem 6.1) or manifold restricted Hessians (as in Theorem 7.2) for Example 3.3 we can conclude from Corollary 5.3 and Table 5.1 (having two basic $\mathcal{U}$-Hessians of opposite sign) that $\bar{x} = (0, 0)^\top$ is neither a local maximizer nor a local minimizer.

Theorem 6.1 provides somewhat more second-order information, since from (6.5) the second-order epi-derivatives for Example 3.3 vary from 2 to -10 as $\bar{g}_2$, the $\mathcal{V}$-component of the subgradients, varies from -1 to 1. This is at the expense of the additional assumption of $\mathcal{V}$-minimality, which provides the important link between primal-dual functions given here and $\mathcal{U}$-Lagrangians defined earlier for convex functions.

The connection to partly smooth functions in Theorem 7.4 requires a regularity assumption as well as dual feasibility for all subgradients. This latter assumption is important for obtaining inner semicontinuity of the subdifferential. This Theorem 7.3 result and its proof may have future application for rate of convergence analysis for minimization algorithms based on approximating fast tracks as in [9]. Such algorithms would try to exploit pdg structure implicitly without having to know it explicitly.

REFERENCES

[1] J. Burke, A. Lewis, and M. Overton, *Two numerical methods for optimizing matrix stability*, Linear Algebra Appl., 351/352 (2002), pp. 117–145.

[2] F. Clarke, *Optimization and Nonsmooth Analysis*, Wiley, 1983; reprinted, SIAM, Philadelphia, 1990.

[3] S. Lang, *Real and Functional Analysis*, 3rd ed., Springer-Verlag, New York, 1993.

[4] C. Lemaréchal, F. Oustry, and C. Sagastizábal, *The $\mathcal{U}$-Lagrangian of a convex function*, Trans. Amer. Math. Soc., 352 (2000), pp. 711–729.

[5] A. S. Lewis, *Active sets, nonsmoothness, and sensitivity*, SIAM J. Optim., 13 (2002), pp. 702–725.

[6] R. Mifflin and C. Sagastizábal, *$\mathcal{VU}$-decomposition derivatives for convex max-functions*, in Ill-posed Variational Problems and Regularization Techniques, Lecture Notes in Econom. and Math. Systems 477, R. Tichatschke and M. Théra, eds., Springer-Verlag, Berlin, Heidelberg, 1999, pp. 167–186.

[7] R. Mifflin and C. Sagastizábal, *Functions with primal-dual gradient structure and $\mathcal{U}$-Hessians*, in Nonlinear Optimization and Related Topics, Applied Optim. 36, G. D. Pillo and F. Giannessi, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000, pp. 219–233.

[8] R. Mifflin and C. Sagastizábal, *On $\mathcal{VU}$-theory for functions with primal-dual gradient structure*, SIAM J. Optim., 11 (2000), pp. 547–571.

[9] R. Mifflin and C. Sagastizábal, *A $\mathcal{VU}$-Proximal Point Algorithm for Minimization,* working paper, 2001, http://www.math.wsu.edu/math/faculty/mifflin/.

[10] R. Mifflin and C. Sagastizábal, *Proximal points are on the fast track*, J. Convex Anal., 9 (2002), pp. 563–579.

[11] B. Mordukhovich, *Maximum principle in the problem of time optimal response with nonsmooth constraints*, J. Appl. Math. Mech., 40 (1976), pp. 960–969.

[12] M. L. Overton, *Large-scale optimization of eigenvalues*, SIAM J. Optim., 2 (1992), pp. 88–120.

[13] R. Poliquin and R. T. Rockafellar, *Second-order nonsmooth analysis in nonlinear programming*, in Recent Advances in Nonsmooth Optimization, World Scientific, River Edge, NJ, 1995, pp. 322–349.

[14] R. T. Rockafellar, *First and second-order epi-differentiability in nonlinear programming*, Trans. Amer. Math. Soc., 307 (1988), pp. 75–108.
[15] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Grundlehren Math. Wiss. 317, Springer-Verlag, Berlin, 1998.
[16] A. Shapiro, *On a class of nonsmooth composite functions*, Math. Oper. Res., to appear.
[17] A. Shapiro and M. K. H. Fan, *On eigenvalue optimization*, SIAM J. Optim., 5 (1995), pp. 552–569.
[18] S. J. Wright, *Identifiable surfaces in constrained optimization*, SIAM J. Control Optim., 31 (1993), pp. 1063–1079.

SIAM J. OPTIM.
Vol. 13, No. 4, pp. 1195–1221

A GLOBALLY AND LOCALLY SUPERLINEARLY CONVERGENT NON–INTERIOR-POINT ALGORITHM FOR P_0 LCPs*

YUN-BIN ZHAO† AND DUAN LI‡

Abstract. Based on the concept of the regularized central path, a new non–interior-point path-following algorithm is proposed for solving the P_0 linear complementarity problem (P_0 LCP). The condition ensuring the global convergence of the algorithm for P_0 LCPs is weaker than most conditions previously used in the literature. This condition can be satisfied even when the strict feasibility condition, which has often been assumed in most existing non–interior-point algorithms, fails to hold. When the algorithm is applied to P_* and monotone LCPs, the global convergence of this method requires no assumption other than the solvability of the problem. The local superlinear convergence of the algorithm can be achieved under a nondegeneracy assumption. The effectiveness of the algorithm is demonstrated by our numerical experiments.

Key words. linear complementarity problem, non–interior-point algorithm, Tikhonov regularization, P_0 matrix, regularized central path

AMS subject classifications. 90C30, 90C33, 90C51, 65K05

PII. S1052623401384151

1. Introduction.

We consider a new path-following algorithm for the linear complementarity problem (LCP):

$$x \geq 0,\ Mx + d \geq 0,\ x^T(Mx + d) = 0,$$

where M is an n by n matrix and d is a vector in R^n. This problem is said to be a P_0 LCP if M is a P_0 matrix, and a P_* LCP if M is a P_* matrix. We recall that M is said to be a P_0 matrix (see [13]) if

$$\max_{1\leq i\leq n} x_i(Mx)_i \geq 0 \ \text{ for any } \ 0 \neq x \in R^n.$$

M is said to be a P_* matrix (see [26]) if there exists a nonnegative constant $\tau \geq 0$ such that

$$(1+\tau)\sum_{i\in I_+} x_i(Mx)_i + \sum_{i\in I_-} x_i(Mx)_i \geq 0 \ \text{ for any } \ 0 \neq x \in R^n,$$

where $I_+ = \{i : x_i(Mx)_i > 0\}$ and $I_- = \{i : x_i(Mx)_i < 0\}$.

We first give a synopsis of non–interior-point methods and related results for complementarity problems. The first non–interior-point path-following method for LCPs was proposed by Chen and Harker [6]. This method was improved by Kanzow [24], who also studied other closely related methods, later called the Chen–Harker–Kanzow–Smale (CHKS) smoothing function method (see [20]). The CHKS function

*Received by the editors January 29, 2001; accepted for publication (in revised form) August 6, 2002; published electronically May 2, 2003. This work was partially supported by grant CUHK4392/99E, Research Grants Council, Hong Kong.
http://www.siam.org/journals/siopt/13-4/38415.html

†Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, NT, Hong Kong, and Institute of Applied Mathematics, AMSS, Chinese Academy of Sciences, Beijing, China (ybzhao@amss.ac.cn).

‡Corresponding author. Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, NT, Hong Kong (dli@se.cuhk.edu.hk).

$\phi : R^3 \to R$ is defined by

$$\phi(t_1, t_2, \mu) = t_1 + t_2 - \sqrt{(t_1 - t_2)^2 + 4\mu}.$$

Based on such a function, Hotta and Yoshise [20] studied the structural properties of a non–interior-point trajectory and proposed a globally convergent path-following algorithm for a class of P_0 LCPs. However, no rate of convergence was reported in these papers. The first global linear convergence result for the LCP with a P_0 and R_0 matrix was obtained by Burke and Xu [3], who also proposed in [4] a non–interior-point predictor-corrector algorithm for monotone LCPs which was both globally linearly and locally quadratically convergent under certain assumptions. Further development of non–interior-point methods can be found in [35, 5, 40, 33, 8, 7, 21]. It is worth mentioning that Chen and Xiu [8] and Chen and Chen [7] proposed a class of non–interior-point methods using the Chen–Mangasarian smoothing function family [9] that includes the CHKS smoothing function as a special case.

Since most existing non–interior-point path-following algorithms are based on the CHKS function, these methods actually follow the central path to locate a solution of the LCP. The central path is the set of solutions of the following system as the parameter $\mu > 0$ varies:

$$x > 0,\ Mx + d > 0,\ X(Mx + d) = \mu e,$$

where $X = \mathrm{diag}(x)$ and $e = (1, \ldots, 1)^T$. For P_0 LCPs, it is shown (see [42, 43]) that most assumptions used for non–interior-point algorithms (for instance, Condition 1.5 in [25], Condition 1.2 in Hotta and Yoshise [20], and the $P_0 + R_0$ assumption in Burke and Xu [3], and Chen and Chen [7]) imply that the solution set of the problem is bounded. As shown by Ravindran and Gowda in [34], the P_0 complementarity problem with a bounded solution set must have a strictly feasible point, i.e., there exists an x^0 such that $Mx^0 + d > 0$. (This implies that a P_0 LCP with no strictly feasible point either has no solution or has an unbounded solution set.) We conclude that the above-mentioned conditions all imply that the problem has a strictly feasible point. Thus, for a solvable P_0 LCP without a strictly feasible point (in this case, the central path does not exist), it is unknown whether most existing non–interior-point algorithms are globally convergent or not. An interesting problem is how to improve these algorithms so that they are able to handle those P_0 problems with unbounded solution sets or without strictly feasible points.

Recently, Zhao and Li [42] proposed a new continuation trajectory for complementarity problems, which is defined as follows:

$$x > 0, \quad Mx + d + \theta^p x > 0, \quad x_i[(Mx + d)_i + \theta^p x_i] = \theta^q a_i, \quad i = 1, \ldots, n,$$

where θ is a parameter in $(0, 1]$; $p \in (0, \infty)$ and $q \in [1, \infty)$ are two fixed scalars; and $a = (a_1, \ldots, a_n)^T \in R^n_{++}$ is a fixed vector, for example, $a = e$. For a P_0 matrix M, it turns out (see [42]) that the above system has a unique solution for each given parameter θ, and this solution, denoted by $x(\theta)$, is continuously differentiable on (0,1). Thus, the set $\{x(\theta) : \theta \in (0, 1]\}$ forms a smooth path approaching the solution set of the P_0 LCP as θ tends to zero. Notice that, for a given θ, the term $Mx + d + \theta^p x$ is the Tikhonov regularization of $Mx + d$, which has been used by several authors, such as Isac [22], Venkateswaran [36], Facchinei and Kanzow [14], Ravindran and Gowda [34], and Zhao and Li [42], to study complementarity problems. We may refer to the above smooth path as the *regularized central path*. A good feature of the regularized

central path is that its existence and boundedness can be guaranteed under a very weak assumption. In particular, the boundedness of the solution set and the strict feasibility condition are not needed for the existence of this path. Combining the CHKS function and the Tikhonov regularization method, Zhao and Li [43] extended the results in [42] to non–interior-point methods and studied the existence as well as the limiting behavior of a new non–interior-point smooth path.

The theoretical results established in [43] motivate us to construct a new non–interior-point path-following algorithm for P_0 LCPs. The purpose of this paper is to provide such a practical numerical algorithm. It is worth stressing the differences between the proposed method in this paper and previous algorithms in the literature. (i) The proposed algorithm follows the regularized central path instead of the central path. (ii) The condition ensuring the global convergence of the algorithm for P_0 LCPs is strictly weaker than those used in most existing non–interior-point methods. The local superlinear convergence of the algorithm can be achieved under a standard nondegeneracy assumption that was used in many works such as [38, 39, 33]. In particular, we also study the important special case of P_* LCPs and derive some stronger results than that of the P_0 case.

The paper is organized as follows. In section 2, we introduce some basic results and describe the algorithm. In section 3, we prove the global convergence of the algorithm for a class of P_0 LCPs. The local convergence analysis of the algorithm is given in section 4. The special case of P_* LCPs is discussed in section 5, and some numerical results are reported in section 6.

Notation: R^n denotes the n-dimensional Euclidean space. R^n_+ and R^n_{++} denote the nonnegative orthant and positive orthant, respectively. A vector $x \geq 0$ ($x > 0$) means $x \in R^n_+$ ($x \in R^n_{++}$). All the vectors, unless otherwise stated, are column vectors. T denotes the transpose of a vector. For any vector x, the capital X denotes the diagonal matrix $\mathrm{diag}(x)$, and for any index set $I \subseteq \{1, \ldots, n\}$, x_I denotes the subvector made up of the components x_i for $i \in I$. The symbol e denotes the vector in R^n with all of its components equal to one. For given vectors u, w, v in R^n, the triplet (u, w, v) (the pair (x, y)) denotes the column vector $(u^T, w^T, v^T)^T$ ($(x^T, y^T)^T$). For any $u \in R^n$, the symbol u^p denotes the pth power of the vector u, i.e., the vector $(u_1^p, \ldots, u_n^p)^T$ where $p > 0$ is a positive scalar, and U^p denotes the diagonal matrix $\mathrm{diag}(u^p)$. For any vector $x \leq y$, we denote by $[x, y]$ the rectangular box $[x_1, y_1] \times \cdots \times [x_n, y_n]$.

2. A non–interior-point path-following algorithm. Let p and q be two given positive scalars. Define the map $\mathcal{H} : R^n_+ \times R^{2n} \to R^{3n}$ as follows:

$$\mathcal{H}(u, x, y) = \begin{pmatrix} u \\ x + y - \sqrt{(x-y)^2 + 4u^q} \\ y - (Mx + d + U^p x) \end{pmatrix}, \qquad (u, x, y) \in R^n_+ \times R^{2n}, \tag{2.1}$$

where $U^p = \mathrm{diag}(u^p)$ and all the algebraic operations are performed componentwise. The above homotopy map first appeared in [43]. Clearly, if $\mathcal{H}(u, x, y) = 0$, then (x, y) is a solution to the LCP; conversely, if (x, y) is a solution to the LCP, then $(0, x, y)$ is a solution to the equation $\mathcal{H}(u, x, y) = 0$. Thus, an LCP can be solved by locating a solution of the nonlinear equation $\mathcal{H}(u, x, y) = 0$. From the discussion in [43], we can conclude that it is a judicious choice to use the above version of the homotopy formulation in order to deal with the LCP with an unbounded solution set.

Before embarking on stating the algorithm, we first introduce some results established in [43]. Let $(a, b, c) \in R^n_{++} \times R^{2n}$ be given. Consider the following system with

the parameter θ:

$$\mathcal{H}(u,x,y)=\theta(a,b,c), \tag{2.2}$$

where $\theta\in(0,1]$. For P_0 LCPs, it is shown in [43] that for each given $\theta\in(0,1]$ the above system has a unique solution denoted by $(u(\theta),x(\theta),y(\theta))$, which is also continuously differentiable with respect to θ. Therefore, the set

$$\{(u(\theta),x(\theta),y(\theta)):\mathcal{H}(u,x,y)=\theta(a,b,c),\ \theta\in(0,1]\} \tag{2.3}$$

forms a smooth path. Also, in this paper, we refer to this path as the *regularized central path*. The existence of such a smooth path for P_0 LCPs needs no assumption (see Theorem 2.1 below). An additional condition is assumed to guarantee the boundedness of this path. We now introduce such a condition proposed in [43].

For given $(a,b,c)\in R^n_{++}\times R^{2n}$ and $\theta\in(0,1]$, we define a mapping $\mathcal{F}_{(a,b,c,\theta)}: R^{2n}\to R^{2n}$ as follows:

$$\mathcal{F}_{(a,b,c,\theta)}(x,y)=\begin{pmatrix} Xy \\ y-M(x+\frac{1}{2}\theta b)-d-\theta^p A^p x-\theta c\end{pmatrix},$$

where $X=\operatorname{diag}(x)$ and $A^p=\operatorname{diag}(a^p)$.

Condition 2.1. *For any given $(a,b,c)\in R^n_{++}\times R^{2n}$ and scalar $\hat{t}$, there exists a scalar $\theta^*\in(0,1]$ such that*

$$\bigcup_{\theta\in(0,\theta^*]}\mathcal{F}^{-1}_{(a,b,c,\theta)}(D_\theta)$$

is bounded, where

$$\mathcal{F}^{-1}_{(a,b,c,\theta)}(D_\theta):=\{(x,y)\in R^{2n}_{++}:\mathcal{F}_{(a,b,c,\theta)}(x,y)\in D_\theta\}$$

and $D_\theta:=[0,\theta a^q]\times[-\theta\hat{t}e,\theta\hat{t}e]\subseteq R^n_+\times R^n$.

The following result states that Condition 2.1 is weaker than most known assumptions used for non–interior-point methods. An example was given in [43] to show that Condition 2.1 can be satisfied even when the strict feasibility condition fails to hold.

Proposition 2.1 (see [43]). *Let $f=Mx+d$, where M is a P_0 matrix. If one of the following conditions holds, then Condition* 2.1 *is satisfied:*

(a) *Condition* 1.5 *of Kojima, Megiddo, and Noma* [25];

(b) *Condition* 2.2 *of Hotta and Yoshise* [20];

(c) *Assumption* 2.2 *of Chen and Chen* [7];

(d) *f is a P_0 and R_0 function* [3, 7];

(e) *f is a P_* function (i.e., M is a P_* matrix) and there is a strictly feasible point* [26];

(f) *f is a uniform P function, i.e., M is a P matrix* [27];

(g) *The solution set of the LCP is nonempty and bounded.*

The converse, however, is not true; i.e., Condition 2.1 *cannot imply any one of the above conditions.*

Restricted to LCPs, the main result established in [43] is summarized as follows.

Theorem 2.1 (see [43]). *Let M be a P_0 matrix.*

(i) *For each $\theta\in(0,1]$, the system* (2.2) *has a unique solution denoted by* $(u(\theta),$ $x(\theta),$ $y(\theta))$, *which is also continuously differentiable in θ.*

(ii) *If Condition* 2.1 *is satisfied, then the regularized central path* (2.3) *is bounded. Hence, there exists a subsequence* $(u(\theta^k), x(\theta^k), y(\theta^k))$ *that converges, as* $\theta^k \to 0$, *to* $(0, x^*, y^*)$ *where* x^* *is a solution to the LCP.*

For P_* LCPs, the only condition for the result (ii) above is the solvability of the problem.

THEOREM 2.2 (see [43]). *Let* M *be a* P_* *matrix. Assume that the solution set of the LCP is nonempty.*

(i) *If* $p \le 1$ *and* $q \in [1, \infty)$, *then the regularized central path* (2.3) *is bounded.*

(ii) *If* $p > 1$, $q \in [1, \infty)$ *and* $c \in R^n_{++}$, *then the regularized central path* (2.3) *is bounded.*

The boundedness of the path (2.3) implies that the problem has a solution. Combining this fact and the above result, we may conclude that the solvability of a P_* LCP, roughly speaking, is a necessary and sufficient condition for the boundedness of the regularized central path. For monotone LCPs, we have a much stronger result than the above; i.e., the entire path (2.3) is convergent as $\theta \to 0$. The property of the limiting point of this path, as $\theta \to 0$, depends on the choice of the scalars p and q.

THEOREM 2.3 (see [43]). *Let* M *be a positive semidefinite matrix. Assume that the solution set of the LCP is nonempty.*

(i) *If* $p \le 1$ *and* $q \in [1, \infty)$, *then the regularized central path* (2.3) *converges, as* $\theta \to 0$, *to the unique least N-norm solution of the LCP, where* $N = A^{p/2}$.

(ii) *If* $p > 1$, $q \in [1, \infty)$, *and* $c \in R^n_{++}$, *then the regularized central path* (2.3) *converges, as* $\theta \to 0$, *to a maximally complementary solution of the LCP.*

We now introduce the algorithm. We choose the following neighborhood around the regularized central path $\{(u(\theta), x(\theta), y(\theta)) : \theta \in (0, 1]\}$:

$$\mathcal{N}(\beta) = \{(u, x, y) : \|u - \theta a\| = 0,\ \|\mathcal{H}(u, x, y) - \theta(a, b, c)\| \le \beta\theta,\ \theta \in (0, 1]\}.$$

Denote

$$G_\theta(x, y) = \begin{pmatrix} x + y - \sqrt{(x - y)^2 + 4(\theta a)^q} \\ y - (Mx + d + \theta^p A^p x) \end{pmatrix}. \tag{2.4}$$

Then the above neighborhood reduces to

$$\mathcal{N}(\beta) = \{(x, y) : \|G_\theta(x, y) - \theta(b, c)\| \le \beta\theta,\ \theta \in (0, 1]\},$$

where G_θ is given by (2.4). For a given $\theta \in (0, 1]$, we denote

$$\mathcal{N}(\beta, \theta) = \{(x, y) : \|G_\theta(x, y) - \theta(b, c)\| \le \beta\theta\}.$$

Throughout the paper, $\nabla G_\theta(x, y)$ denotes the Jacobian of $G_\theta(x, y)$ with respect to (x, y). Let $\varepsilon > 0$ be a given tolerance. We now describe the algorithm as follows.

ALGORITHM 2.1. Let $p \in (0, \infty)$, $q \in [1, \infty)$, $\sigma \in (0, 1)$, and $\alpha \in (0, 1)$ be given.

Step 1. Select $(a, b, c) \in R^n_{++} \times R^{2n}$, $(x^0, y^0) \in R^{2n}$, $\theta^0 \in (0, 1)$, and $\beta > 0$ such that $(x^0, y^0) \in \mathcal{N}(\beta, \theta^0)$.

Step 2 (approximate Newton step). If $\|G_0(x^k, y^k)\| \le \varepsilon$, stop; otherwise, let $(d\hat{x}^k, d\hat{y}^k)$ solve the equation

$$G_0(x^k, y^k) + \nabla G_{\theta^k}(x^k, y^k)(dx, dy) = 0. \tag{2.5}$$

Let

$$(\hat{x}^{k+1}, \hat{y}^{k+1}) = (x^k, y^k) + (d\hat{x}^k, d\hat{y}^k).$$

If $\|G_0(\hat{x}^{k+1}, \hat{y}^{k+1})\| \le \varepsilon$, stop; otherwise, if

$$(\hat{x}^{k+1}, \hat{y}^{k+1}) \in \mathcal{N}(\beta, (\theta^k)^2),$$

then set

$$\theta^{k+1} = (\theta^k)^2, \ (x^{k+1}, y^{k+1}) = (\hat{x}^{k+1}, \hat{y}^{k+1}).$$

Set $k := k + 1$, and repeat step 2. Otherwise, go to step 3.

Step 3 (centering step). If $G_{\theta^k}(x^k, y^k) = \theta^k(b, c)$, set $(x^{k+1}, y^{k+1}) = (x^k, y^k)$, and go to step 4. Otherwise, let (dx^k, dy^k) be the solution to the equation

$$G_{\theta^k}(x^k, y^k) - \theta^k(b, c) + \nabla G_{\theta^k}(x^k, y^k)(dx, dy) = 0. \tag{2.6}$$

Let λ_k be the maximum among the values of $1, \alpha, \alpha^2, \ldots$ such that

$$\|G_{\theta^k}(x^k + \lambda_k dx^k, y^k + \lambda_k dy^k) - \theta^k(b, c)\| \le (1 - \sigma\lambda_k)\|G_{\theta^k}(x^k, y^k) - \theta^k(b, c)\|.$$

Set

$$(x^{k+1}, y^{k+1}) = (x^k, y^k) + \lambda_k(dx^k, dy^k).$$

Step 4 (reduce θ^k). Let γ_k be the maximum among the values $1, \alpha, \alpha^2, \ldots$ such that

$$(x^{k+1}, y^{k+1}) \in \mathcal{N}(\beta, (1 - \gamma_k)\theta^k),$$

i.e.,

$$\|G_{(1-\gamma_k)\theta^k}(x^{k+1}, y^{k+1}) - (1 - \gamma_k)\theta^k(b, c)\| \le \beta(1 - \gamma_k)\theta^k.$$

Set $\theta^{k+1} = (1 - \gamma_k)\theta^k$. Set $k := k + 1$ and go to step 2.

Remark 2.1. (i) To start the algorithm, we need an initial point within the neighborhood of the regularized central path. Such an initial point can be found at no cost for the above algorithm. For example, let (a, b, c) be an arbitrary triplet in $R^n_{++} \times R^{2n}$, (x^0, y^0) be an arbitrary vector in R^{2n}, and θ^0 be an arbitrary scalar in (0,1). Then choose β such that

$$\beta \ge \frac{\|G_{\theta^0}(x^0, y^0) - \theta^0(b, c)\|}{\theta^0}.$$

Clearly, this initial point satisfies $(x^0, y^0) \in \mathcal{N}(\beta, \theta^0)$.

(ii) The step 3 of the algorithm is a centering step in the sense that it forces the iterate close to the regularized central path such that the iterate is always confined in the neighborhood of the path. In the next section, we show that step 3 together with step 4 guarantees the global convergence of the algorithm. Step 2 is an approximate Newton step which was shown to have good local convergence properties (see, for example, [10, 11]). This step is used to accelerate the iteration such that a local rapid convergence can be achieved. Similar strategies were used in several works such as [38, 39, 28, 7, 8]. We also note that linear systems (2.5) and (2.6) have the same coefficient matrix, and thus only one matrix factorization is needed at each iteration.

We now show that the algorithm is well-defined.

PROPOSITION 2.2. *Algorithm* 2.1 *is well-defined and satisfies the following properties:* (i) θ^k *is monotonically decreasing, and* (ii) $\|G_{\theta^k}(x^k, y^k) - \theta^k(b, c)\| \le \beta\theta^k$ *for all* $k \ge 0$, *i.e.,* $(x^k, y^k) \in \mathcal{N}(\beta, \theta^k)$ *for all* $k \ge 0$.

Proof. We verify that each step of the algorithm is well-defined. As we pointed out in Remark 2.1, step 1 of the algorithm is well-defined. Consider the following $2n \times 2n$ matrix:

$$\nabla G_{\theta^k}(x^k, y^k) = \begin{pmatrix} I - (X^k - Y^k)D^k & I + (X^k - Y^k)D^k \\ -(M + (\theta^k)^p A^p) & I \end{pmatrix}, \tag{2.7}$$

where $X^k = \operatorname{diag}(x^k)$, $Y^k = \operatorname{diag}(y^k)$, and $D^k = \operatorname{diag}(d^k)$ with $d^k = (d_1^k, \ldots, d_n^k)^T$, where

$$d_i^k = \frac{1}{\sqrt{(x_i^k - y_i^k)^2 + 4(\theta^k)^q a_i^q}}, \qquad i = 1,\ 2,\ \ldots,\ n.$$

Since $a \in R^n_{++}$, for each given $\theta^k \in (0, 1)$ it is easy to see that $I-(X^k-Y^k)D^k$ and $I+(X^k-Y^k)D^k$ are positive diagonal matrices for every $(x^k, y^k) \in R^{2n}$. Thus, by Lemma 5.4 in Kojima, Megiddo, and Noma [25], the matrix $\nabla G_{\theta^k}(x^k, y^k)$ is nonsingular when M is a P_0 matrix. Thus, step 2 is well-defined.

Since (dx^k, dy^k) is a descent direction for the function at (x^k, y^k)

$$\tilde{f}(x, y) = \frac{1}{2}\|G_{\theta^k}(x, y) - \theta^k(b, c)\|_2^2,$$

the line search in step 3 is well-defined, and thus the whole step 3 is well-defined.

We finally prove that the step 4 is well-defined. For any scalar $\mu_1 > \mu_2 \ge 0$, we have

$$\begin{aligned}
&\|G_{\mu_1}(x, y) - G_{\mu_2}(x, y)\| \\
&= \left\| \begin{pmatrix} x + y - \sqrt{(x-y)^2 + 4\mu_1^q a^q} \\ y - (Mx + d + \mu_1^p A^p x) \end{pmatrix} - \begin{pmatrix} x + y - \sqrt{(x-y)^2 + 4\mu_2^q a^q} \\ y - (Mx + d + \mu_2^p A^p x) \end{pmatrix} \right\| \\
&= \left\| \begin{pmatrix} \sqrt{(x-y)^2 + 4\mu_1^q a^q} - \sqrt{(x-y)^2 + 4\mu_2^q a^q} \\ (\mu_1^p - \mu_2^p) A^p x \end{pmatrix} \right\| \\
&= \left\| \begin{pmatrix} \frac{4(\mu_1^q - \mu_2^q)a^q}{\sqrt{(x-y)^2 + 4\mu_1^q a^q} + \sqrt{(x-y)^2 + 4\mu_2^q a^q}} \\ (\mu_1^p - \mu_2^p) A^p x \end{pmatrix} \right\| \\
&\le \left\| \begin{pmatrix} \frac{4(\mu_1^q - \mu_2^q)a^q}{\sqrt{4\mu_1^q a^q}} \\ (\mu_1^p - \mu_2^p) A^p x \end{pmatrix} \right\| \\
&\le \max\{\mu_1^{q/2}(1 - (\mu_2/\mu_1)^q), \mu_1^p(1 - (\mu_2/\mu_1)^p)\}\|(2a^{q/2}, A^p x)\|. \qquad (2.8)
\end{aligned}$$

In particular, setting $(x, y) = (x^k, y^k)$, $\mu_1 = \theta^k > 0$, and $\mu_2 = (1 - \gamma)\theta^k$ with $\gamma \in (0, 1)$, we then have

$$\begin{aligned}
&\|G_{\theta^k}(x^k, y^k) - G_{(1-\gamma)\theta^k}(x^k, y^k)\| \\
&\le \max\{(\theta^k)^{q/2}(1 - (1 - \gamma)^q),\ (\theta^k)^p(1 - (1 - \gamma)^p)\}\|(2a^{q/2}, A^p x^k)\|. \qquad (2.9)
\end{aligned}$$

There are two cases to be considered.

Case (i): $G_{\theta^k}(x^k, y^k) = \theta^k(b, c)$ in step 3. Then, $(x^{k+1}, y^{k+1}) = (x^k, y^k)$. By (2.9), for all sufficiently small γ we have

$$\begin{aligned}
&\|G_{(1-\gamma)\theta^k}(x^{k+1}, y^{k+1}) - (1-\gamma)\theta^k(b, c)\| \\
&= \|G_{(1-\gamma)\theta^k}(x^k, y^k) - G_{\theta^k}(x^k, y^k) + \theta^k(b, c) - (1-\gamma)\theta^k(b, c)\| \\
&\le \|G_{(1-\gamma)\theta^k}(x^k, y^k) - G_{\theta^k}(x^k, y^k)\| + \gamma\theta^k\|(b, c)\| \\
&\le \max\{(\theta^k)^{q/2}(1-(1-\gamma)^q), (\theta^k)^p(1-(1-\gamma)^p)\}\|(2a^{q/2}, A^p x^k)\| + \gamma\theta^k\|(b, c)\| \\
&\le \beta(1-\gamma)\theta^k.
\end{aligned}$$

Case (ii): $G_{\theta^k}(x^k, y^k) \neq \theta^k(b, c)$ in step 3. For this case, according to step 3 we have

$$\begin{aligned}
\|G_{\theta^k}(x^{k+1}, y^{k+1}) - \theta^k(b, c)\| &\le (1-\sigma\lambda_k)\|G_{\theta^k}(x^k, y^k) - \theta^k(b, c)\| \\
&\le (1-\sigma\lambda_k)\beta\theta^k.
\end{aligned}$$

The second inequality follows from the fact that $\|G_{\theta^k}(x^k, y^k) - \theta^k(b, c)\| \le \beta\theta^k$, which is evident from the construction of the algorithm. Notice that $1 - \sigma\lambda_k < 1$. By (2.9) and the above inequality, for all sufficiently small γ we have

$$\begin{aligned}
&\|G_{(1-\gamma)\theta^k}(x^{k+1}, y^{k+1}) - (1-\gamma)\theta^k(b, c)\| \\
&\le \|G_{(1-\gamma)\theta^k}(x^{k+1}, y^{k+1}) - G_{\theta_k}(x^{k+1}, y^{k+1})\| + \|G_{\theta^k}(x^{k+1}, y^{k+1}) - \theta^k(b, c)\| \\
&+\gamma\theta^k\|(b, c)\| \\
&\le \max\{(\theta^k)^{q/2}(1-(1-\gamma)^q), (\theta^k)^p(1-(1-\gamma)^p)\}\|(2a^{q/2}, A^p x^{k+1})\| \\
&+(1-\sigma\lambda_k)\beta\theta^k + \gamma\theta^k\|(b, c)\| \\
&\le (1-\gamma)\beta\theta^k.
\end{aligned}$$

Thus, the step 4 is well-defined.

We now show that all the iterates are in the neighborhood defined by the algorithm. By the construction of the algorithm, it is evident that either $\theta^{k+1} = (\theta^k)^2$ or $\theta^{k+1} = (1-\gamma_k)\theta^k$. Thus, θ^k is monotonically decreasing. When $k = 0$, it follows from step 1 that $(x^0, y^0) \in \mathcal{N}(\beta, \theta^0)$. Assume that this property holds for k, i.e., $(x^k, y^k) \in \mathcal{N}(\beta, \theta^k)$. We show that it holds for $k+1$. Indeed, if step 2 is accepted, then the criterion $(x^{k+1}, y^{k+1}) \in \mathcal{N}(\beta, \theta^{k+1})$ is satisfied, where $\theta^{k+1} = (\theta^k)^2$. If step 2 is rejected, then (x^{k+1}, y^{k+1}) is created by step 3 together with step 4. It follows from step 4 that $(x^{k+1}, y^{k+1}) \in \mathcal{N}(\beta, \theta^{k+1})$, where $\theta^{k+1} = (1-\gamma_k)\theta^k$. Thus, for all $k \ge 0$, we have that $(x^k, y^k) \in \mathcal{N}(\beta, \theta^k)$, i.e., $\|G_{\theta^k}(x^k, y^k) - \theta^k(b, c)\| \le \beta\theta^k$. □

3. Global convergence for $\mathbf{P_0}$ LCPs. We now show that the proposed algorithm is globally convergent for P_0 LCPs provided that Condition 2.1 is satisfied. By Proposition 2.2, for every $k \ge 1$, the iterate (x^k, y^k) satisfies the following:

$$\|G_{\theta^k}(x^k, y^k) - \theta^k(b, c)\| \le \beta\theta_k,\ \theta^k = (1-\gamma_{k-1})\theta^{k-1} \ \text{ or } \ \theta^k = (\theta^{k-1})^2. \tag{3.1}$$

Let $(b^k, c^k) \in R^{2n}$ be two auxiliary vectors determined by

$$(b^k, c^k) = \frac{G_{\theta^k}(x^k, y^k) - \theta^k(b, c)}{\theta^k} \ \text{ for all } \ k. \tag{3.2}$$

Then, $\{(b^k, c^k)\}$ is uniformly bounded. In fact, by (3.1), we have that $\|(b^k, c^k)\| \le \beta$, and hence

$$-\beta e \le b^k \le \beta e,\ -\beta e \le c^k \le \beta e. \tag{3.3}$$

By the definition of (2.4), we can write (3.2) as

$$x^k + y^k - \sqrt{(x^k - y^k)^2 + 4(\theta^k)^q a^q} = \theta^k(b + b^k),$$
$$y^k - (Mx^k + d + (\theta^k)^p A^p x^k) = \theta^k(c + c^k).$$

By the property of the CHKS function (see Lemma 1 in [20]), the system above is equivalent to

$$x^k - \frac{1}{2}\theta^k(b + b^k) > 0, \quad y^k - \frac{1}{2}\theta^k(b + b^k) > 0, \tag{3.4}$$

$$\left[X^k - \frac{1}{2}\theta^k(B + B^k)\right]\left(y^k - \frac{1}{2}\theta^k(b + b^k)\right) = (\theta^k)^q a^q, \tag{3.5}$$

$$y^k = Mx^k + d + (\theta^k)^p A^p x^k + \theta^k(c + c^k), \tag{3.6}$$

where X^k, B, and B^k are diagonal matrices corresponding to x^k, b, and b^k, respectively.

Remark 3.1. The fact that all iterates generated by Algorithm 2.1 satisfy the system (3.4)–(3.6) plays a key role in the analysis throughout the paper. By continuity, from (3.6) it follows that $\{y^k\}$ is bounded if $\{x^k\}$ is. Thus, if the sequence (x^k, y^k) is unbounded, then $\{x^k\}$ must be unbounded.

The following result is a minor revision of Lemma 1 in [34].

LEMMA 3.1 (see [42, 44]). *Let M be a P_0 matrix. Let $\{z^k\}$ be an arbitrary sequence with $\|z^k\| \to \infty$ and $z^k \geq \bar{z}$ for all k, where $\bar{z} \in R^n$ is a fixed vector. Then there exist a subsequence of $\{z^k\}$, denoted by $\{z^{k_j}\}$, and a fixed index i_0 such that $z_{i_0}^{k_j} \to \infty$ and $(Mz^{k_j} + d)_{i_0}$ is bounded from below.*

The next result shows that the iterative sequence $\{(x^k, y^k)\}$ generated by Algorithm 2.1 is bounded under Condition 2.1.

THEOREM 3.1. *Let M be a P_0 matrix. If Condition* 2.1 *is satisfied, then the iterative sequence $\{(x^k, y^k)\}$ generated by Algorithm* 2.1 *is bounded.*

Proof. We prove this result by contradiction. Assume that $\{(x^k, y^k)\}$ is unbounded. Then $\{x^k\}$ is unbounded (see Remark 3.1). Without loss of generality, we may assume that $\|x^k\| \to \infty$. Notice that $\theta^k < 1$ and $\|b^k\| \leq \beta$. It follows from (3.4) that

$$x^k \geq \frac{1}{2}\theta^k(b + b^k) \geq -\frac{1}{2}(\|b\| + \beta)e \quad \text{for all} \quad k.$$

Thus, by Lemma 3.1, there exist a subsequence of $\{x^k\}$, denoted also by $\{x^k\}$, and an index m such that $x_m^k \to \infty$ and $(Mx^k + d)_m$ is bounded from below. By (3.5), for each i we have

$$\left(x_i^k - \frac{1}{2}\theta^k(b_i + b_i^k)\right)\left(y_i^k - \frac{1}{2}\theta^k(b_i + b_i^k)\right) = (\theta^k)^q a_i^q,$$

and thus,

$$y_m^k - \frac{1}{2}\theta^k(b_m + b_m^k) = \frac{(\theta^k)^q a_m^q}{x_m^k - \theta^k(b_m + b_m^k)/2}.$$

By using (3.6), the above equation can be further written as

$$(Mx^k + d)_m + \theta^k(c_m + c_m^k) - \frac{1}{2}\theta^k(b_m + b_m^k) - \frac{(\theta^k)^q a_m^q}{x_m^k - \theta^k(b_m + b_m^k)/2}$$
$$= -(\theta^k)^p a_m^p x_m^k.$$

Since b^k and c^k are bounded, $x_m^k \to \infty$, and $(Mx^k + d)_m$ is bounded from below, we conclude that the left-hand side of the above equation is bounded from below. This implies that $\theta^k \to 0$ (since otherwise the right-hand side tends to $-\infty$).

In what follows, we denote

$$\bar{x}^k = x^k - \frac{1}{2}\theta^k(b + b^k), \qquad \bar{y}^k = y^k - \frac{1}{2}\theta^k(b + b^k). \tag{3.7}$$

From (3.4) and (3.5), we see that $(\bar{x}^k, \bar{y}^k) > 0$ for all k, and

$$\bar{X}^k\bar{y}^k = (\theta^k)^q a^q. \tag{3.8}$$

Since $\|b^k\| \le \beta$, it follows that

$$\left\| \frac{M(x^k - \bar{x}^k - \theta^k b/2)}{\theta^k} \right\| \le \frac{1}{2}\beta\|M\|. \tag{3.9}$$

By using (3.6) and (3.7), we have

$$\begin{aligned}
&\bar{y}^k - M\left(\bar{x}^k + \frac{1}{2}\theta^k b\right) - d - (\theta^k)^p A^p \bar{x}^k - \theta^k c \\
&\quad = Mx^k + d + (\theta^k)^p A^p x^k + \theta^k(c + c^k) - \frac{1}{2}\theta^k(b + b^k) \\
&\qquad -M\left(\bar{x}^k + \frac{1}{2}\theta^k b\right) - d - (\theta^k)^p A^p \bar{x}^k - \theta^k c \\
&\quad = \theta^k\left(\frac{M(x^k - \bar{x}^k - \theta^k b/2)}{\theta^k} - \frac{1}{2}(b + b^k) + c^k + \frac{1}{2}(\theta^k)^p A^p (b + b^k)\right).
\end{aligned}$$

By (3.9) and the boundedness of θ^k, b^k, and c^k, there exists a scalar $\hat{t} > 0$ such that

$$-\hat{t}e \le \frac{M(x^k - \bar{x}^k - \theta^k b/2)}{\theta^k} - \frac{1}{2}(b + b^k) - c^k + \frac{1}{2}(\theta^k)^p A^p (b + b^k) \le \hat{t}e$$

for all k. Therefore,

$$\bar{y}^k - M(\bar{x}^k + \frac{1}{2}\theta^k b) - d - (\theta^k)^p A^p \bar{x}^k - \theta^k c \ \in \ \theta^k[-\hat{t}e, \hat{t}e]$$

for all k. Notice that $q \in [1, \infty)$. Combining (3.8) and the above leads to

$$\begin{aligned}
\mathcal{F}_{(a,b,c,\theta^k)}(\bar{x}^k, \bar{y}^k) \ &= \ \begin{pmatrix} \bar{X}^k\bar{y}^k \\ \bar{y}^k - M(\bar{x}^k + \frac{1}{2}\theta^k b) - d - (\theta^k)^p A^p \bar{x}^k - \theta^k c \end{pmatrix} \\
&\in \ \theta^k[0, a^q] \times \theta^k[-\hat{t}e, \hat{t}e] \\
&=: D_{\theta^k}
\end{aligned}$$

for all k. Thus,

$$(\bar{x}^k, \bar{y}^k) \in \mathcal{F}^{-1}_{(a,b,c,\theta^k)}(D_{\theta^k}) \ \text{ for all } \ k.$$

By Condition 2.1, there exists a θ^* such that

$$\bigcup_{\theta \in (0,\theta^*]} \mathcal{F}^{-1}_{(a,b,c,\theta)}(D_\theta)$$

is bounded. Since $\theta^k \to 0$, there exists some k_0 such that for all $k \geq k_0$ we have $\theta^k \leq \theta^*$. Thus,

$$\{(\bar{x}^k, \bar{y}^k)\}_{k \geq k_0} \subseteq \bigcup_{\theta^k \leq \theta^*} \mathcal{F}^{-1}_{(a,b,c,\theta^k)}(D_{\theta^k}) \subseteq \bigcup_{\theta \in (0,\theta^*]} \mathcal{F}^{-1}_{(a,b,c,\theta)}(D_\theta).$$

The right-hand side of the above is bounded. This contradicts the left-hand side, which (by assumption) is an unbounded sequence. ∎

We are ready to prove the global convergence of Algorithm 2.1 for P_0 LCPs.

THEOREM 3.2. *Let M be a P_0 matrix. Assume that Condition 2.1 is satisfied. If (x^k, y^k, θ^k) is generated by Algorithm 2.1, then $\{(x^k, y^k)\}$ has at least one accumulation point, and*

$$\lim_{k \to \infty} \theta^k \to 0, \qquad \lim_{k \to \infty} \|G_{\theta^k}(x^k, y^k)\| \to 0. \tag{3.10}$$

Thus, every accumulation point of (x^k, y^k) is a solution to the LCP.

Proof. By Theorem 3.1, the iterative sequence $\{(x^k, y^k)\}$ generated by the algorithm is bounded, and hence it has at least one accumulation point. By Proposition 2.2, we have

$$\begin{aligned} \|G_{\theta^k}(x^k, y^k)\| &\leq \|G_{\theta^k}(x^k, y^k) - \theta^k(b, c)\| + \theta^k \|(b, c)\| \\ &\leq \theta^k[\beta + \|(b, c)\|]. \end{aligned}$$

Thus, to show the second limiting property in (3.10) it is sufficient to show that $\theta^k \to 0$. By the construction of the algorithm, we have either $\theta^{k+1} = (1 - \gamma_k)\theta^k$ or $\theta^{k+1} = (\theta^k)^2$. Thus θ^k is monotonically decreasing, and thus there exists a scalar $1 > \bar{\theta} \geq 0$ such that $\theta^k \to \bar{\theta}$. If $\bar{\theta} = 0$, the desired result follows.

Assume the contrary, that $\bar{\theta} > 0$. We now derive a contradiction. Since $\bar{\theta} > 0$, the algorithm eventually phases out the approximate Newton step and takes only step 3 and step 4. In fact, if step 2 is accepted infinitely many times, then there exists a subsequence $\{k_j\}$ such that $\theta^{k_j+1} = (\theta^{k_j})^2$ which implies that $\bar{\theta} = \bar{\theta}^2$. This is impossible since $0 < \bar{\theta} < 1$. Thus, there exists a k_0 such that for all $k \geq k_0$, the iterates $\{(x^k, y^k)\}_{k \geq k_0}$ are generated only by step 3, and hence $\theta^{k+1} = (1 - \gamma_k)\theta^k$ for all $k \geq k_0$. Since $\theta^k \to \bar{\theta} > 0$, it follows that $\gamma_k \to 0$. Thus, for all sufficiently large k, we have $(x^{k+1}, y^{k+1}) \notin \mathcal{N}(\beta, (1 - \frac{1}{\alpha}\gamma_k)\theta^k$, that is,

$$\left\| G_{(1-\frac{1}{\alpha}\gamma_k)\theta^k}(x^{k+1}, y^{k+1}) - \left(1 - \frac{1}{\alpha}\gamma_k\right)\theta^k(b, c) \right\| > \beta \left(1 - \frac{1}{\alpha}\gamma_k\right)\theta^k.$$

Since the iterate (x^{k+1}, y^{k+1}) is bounded, taking a subsequence if necessary we may assume that this sequence converges to some $(\hat{x}, \hat{y})$. Notice that $\gamma_k \to 0$. Taking the limit in the above inequality, we have

$$\|G_{\bar{\theta}}(\hat{x}, \hat{y}) - \bar{\theta}(b, c)\| \geq \beta\bar{\theta} > 0.$$

Since $\bar{\theta} > 0$, the matrix $\nabla G_{\bar{\theta}}(\hat{x}, \hat{y})$ is nonsingular. Let $(d\hat{x}, d\hat{y})$ be the solution to

$$G_{\bar{\theta}}(\hat{x}, \hat{y}) - \bar{\theta}(b, c) + \nabla G_{\bar{\theta}}(\hat{x}, \hat{y})(dx, dy) = 0.$$

Then $(d\hat{x}, d\hat{y})$ is a strictly descent direction for $\|G_{\bar{\theta}}(x, y) - \bar{\theta}(b, c)\|$ at $(\hat{x}, \hat{y})$. As a result, the line search steplengths, $\hat{\lambda}$ (in step 3) and $\hat{\gamma}$ (in step 4), are both positive

constants. Since G and ∇G are continuous in the neighborhood of $(\hat{x}, \hat{y})$, it follows that $(dx^k, dy^k, \lambda_k, \gamma_k) \to (d\hat{x}, d\hat{y}, \hat{\lambda}, \hat{\gamma})$, and therefore λ_k, γ_k must be uniformly bounded from below by some positive constant for all sufficiently large k. This contradicts the fact $\gamma_k \to 0$. Therefore, $\theta^k \to 0$ must hold. Assume that $(\hat{x}, \hat{y})$ is an arbitrary accumulation point of (x^k, y^k); then by (3.10),

$$0 = \lim_{k\to\infty} \|G_{\theta^k}(x^k, y^k)\| = \|G_0(\hat{x}, \hat{y})\|,$$

which implies that $(\hat{x}, \hat{y})$ is a solution to the LCP. □

Remark 3.2. We have pointed out that the global convergence of most existing non–interior-point methods for P_0 LCPs actually requires the boundedness assumption of the solution set, in which case the P_0 problem must have a strictly feasible point. In order to relax this requirement, Chen and Ye [12] designed a big-M smoothing method for P_0 LCPs. They proved that if the P_0 LCP has a solution and if certain conditions such as "$\bar{x}_{n+2} - \bar{y}_{n+2} \neq -2\bar{\varepsilon}$" are satisfied at the accumulation point of their iterative sequence, then their algorithm is globally convergent. We note that Condition 2.1 in this paper is quite different from Chen and Ye's. However, it is not clear what relation is between the two conditions.

While the global convergence for P_0 LCPs is proved under Condition 2.1, it should be pointed out that this condition is not necessary for the global convergence of P_* problems. We can prove that Algorithm 2.1 is globally convergent provided that the P_* LCP has a solution. Since this result cannot follow from Theorem 3.2, and since its proof is not straightforward, we postpone the discussion for this special case until the local convergence analysis for P_0 LCPs is complete.

4. Local behavior of the algorithm. Under a nondegeneracy assumption, we show in this section the local superlinear convergence of the algorithm when $p = 2 \leq q$. Let (x^*, y^*) be an accumulation point of the iterative sequence (x^k, y^k) generated by Algorithm 2.1. We make use of the following assumption that can also be found in [38, 39, 33].

CONDITION 4.1. *Assume that (x^*, y^*) is strictly complementary, i.e., $x^* + y^* > 0$, where $y^* = Mx^* + d$, and the matrix M_{II} is nonsingular, where $I = \{i : x_i^* > 0\}$.*

While this condition for local convergence has been used by several authors, it is stronger than some existing non–interior-point algorithms. Let M be a P_0 matrix. Under the above condition, it is easy to verify the nonsingularity of the matrix:

$$\nabla G_0(x^*, y^*) = \begin{pmatrix} I - W & I + W \\ -M & I \end{pmatrix}, \tag{4.1}$$

where $W = \mathrm{diag}(w)$ is a diagonal matrix with $w_i = 1$ if $x_i^* > 0$ and $w_i = -1$ otherwise. If Condition 4.1 is satisfied, it follows easily from Proposition 2.5 of Qi [32] that the solution (x^*, y^*) is a locally isolated solution. On the other hand, it is well known that a P_0 complementarity problem has a unique solution when it has a locally isolated solution (Jones and Gowda [23] and Gowda and Sznajder [17]). Thus, Condition 4.1 implies the uniqueness of the solution for a P_0 LCP, and hence it implies Condition 2.1. By Theorem 3.2, we conclude that under Conditions 4.1 the entire sequence (x^k, y^k), generated by Algorithm 2.1, converges to the unique solution of the P_0 LCP, i.e., $(x^k, y^k) \to (x^*, y^*)$. By continuity of ∇G_θ and nonsingularity of $\nabla G_0(x^*, y^*)$, there exists a local neighborhood of (x^*, y^*), denoted by $N(x^*, y^*)$, such that for all $(x, y) \in N(x^*, y^*)$ and all sufficiently small θ the matrix $\nabla G_\theta(x, y)$ is nonsingular,

and there exists a constant C and $\hat{\theta} \in (0,1)$ such that

$$\|\nabla G_\theta(x,y)^{-1}\| \le C \text{ for all } (x,y) \in N(x^*,y^*) \text{ and } \theta \in (0,\hat{\theta}].$$

The following result is very useful for our local convergence analysis.

LEMMA 4.1. *Let M be a P_0 matrix. Under Condition* 4.1, *there exists a neighborhood $N(x^*,y^*)$ of (x^*,y^*) such that for all $(x^k,y^k) \in N(x^*,y^*)$ we have*

(i) $\|\nabla G_{\theta^k}(x^k,y^k) - \nabla G_0(x^k,y^k)\| \le \kappa \max\{(\theta^k)^q,(\theta^k)^p\}$, *where κ is a constant;*

(ii) $G_0(x^k,y^k) - G_0(x^*,y^*) - \nabla G_0(x^k,y^k)[(x^k,y^k) - (x^*,y^*)] = 0$.

Proof. Let

$$I = \{i : x_i^* > 0\}, \qquad J = \{j : y_j^* > 0\}.$$

Then, by strict complementarity, $I \cap J$ is empty and $I \cup J = \{1,2,\ldots,n\}$. Denote

$$\eta = \frac{1}{2}\min\{\|x_I^*\|_\infty,\ \|y_J^*\|_\infty\}.$$

We show first the following inequality:

$$\|W - (X^k - Y^k)D^k\| \le \frac{2}{\eta^2}(\theta^k)^q\|a^q\|_\infty \text{ for all } (x^k,y^k) \in N(x^*,y^*), \tag{4.2}$$

where W is given as in (4.1) and D^k is defined as in (2.7). As we have pointed out, under Condition 4.1 the sequence $\{(x^k,y^k)\}$ converges to (x^*,y^*). For all $(x^k,y^k) \in N(x^*,y^*)$, without loss of generality, we may assume that

$$x_i^k - y_i^k \ge \eta > 0 \ \text{ for } i \in I; \qquad -(x_i^k - y_i^k) \ge \eta > 0 \ \text{ for } i \in J.$$

Hence, when k is sufficiently large, for each $i \in I$ we have

$$\begin{aligned}
&|W_i - (x_i^k - y_i^k)d_i^k| \\
&= |1 - (x_i^k - y_i^k)d_i^k| \\
&= \frac{|\sqrt{(x_i^k - y_i^k)^2 + 4(\theta^k)^q a_i^q} - (x_i^k - y_i^k)|}{\sqrt{(x_i^k - y_i^k)^2 + 4(\theta^k)^q a_i^q}} \\
&= \frac{4(\theta^k)^q a_i^q}{\sqrt{(x_i^k - y_i^k)^2 + 4(\theta^k)^q a_i^q}\left(\sqrt{(x_i^k - y_i^k)^2 + 4(\theta^k)^q a_i^q} + x_i^k - y_i^k\right)} \\
&\le \frac{4(\theta^k)^q a_i^q}{\sqrt{\eta^2 + 4(\theta^k)^q a_i^q}\left(\sqrt{\eta^2 + 4(\theta^k)^q a_i^q} + \eta\right)} \\
&\le \frac{2}{\eta^2}(\theta^k)^q a_i^q. \qquad (4.3)
\end{aligned}$$

Similarly, for $j \in J$ we have

$$|W_j - (x_j^k - y_j^k)d_j^k| \le \frac{2}{\eta^2}(\theta^k)^q a_j^q,$$

which together with (4.3) yields the desired inequality (4.2). On the other hand, by strict complementarity, for every sufficiently large k it is evident that $(X^k - Y^k)\overline{D}^k =$

W, where $\overline{D}^k = \mathrm{diag}(\bar{d}^k)$, where $(\bar{d}^k)_i = 1/\sqrt{(x_i^k - y_i^k)^2}$ $(i = 1, \ldots, n)$. Thus, for every sufficiently large k we have

$$\begin{aligned}
&\nabla G_0(x^k, y^k) - \nabla G_0(x^*, y^*) \\
&= \begin{pmatrix} I - (X^k - Y^k)\overline{D}^k & I + (X^k - Y^k)\overline{D}^k \\ -M & I \end{pmatrix} - \begin{pmatrix} I - W & I + W \\ -M & I \end{pmatrix} \\
&= \begin{pmatrix} W - (X^k - Y^k)\overline{D}^k & (X^k - Y^k)\overline{D}^k - W \\ 0 & 0 \end{pmatrix} \\
&= 0. \qquad (4.4)
\end{aligned}$$

By using (2.7), (4.2), and (4.4), for every sufficiently large k we have

$$\begin{aligned}
\|\nabla G_{\theta^k}(x^k, y^k) - \nabla G_0(x^k, y^k)\| &= \|\nabla G_{\theta^k}(x^k, y^k) - \nabla G_0(x^*, y^*)\| \\
&= \left\| \begin{pmatrix} W - (X^k - Y^k)D^k & (X^k - Y^k)D^k - W \\ -(\theta^k)^p A^p & O \end{pmatrix} \right\| \\
&\le 2\|W - (X^k - Y^k)D^k\| + \|(\theta^k)^p A^p\| \\
&\le \frac{4}{\eta^2}(\theta^k)^q \|a^q\|_\infty + (\theta^k)^p \|a^p\|_\infty \\
&\le \kappa \max\{(\theta^k)^q, (\theta^k)^p\},
\end{aligned}$$

where $\kappa = (4\|a^q\|_\infty)/\eta^2 + \|a^p\|_\infty$ is a constant independent of k. Result (i) is proved.

We now prove result (ii). By the strict complementarity and the definition of W, it is easy to see that for every sufficiently large k the following holds:

$$\begin{aligned}
x^k + y^k - \sqrt{(x^k - y^k)^2} &= (I - W)(x^k - x^*) + (I + W)(y^k - y^*), \\
y^k - Mx^k - d &= -M(x^k - x^*) + y^k - y^*.
\end{aligned}$$

Therefore, by using (4.4) and the above two equations, we have

$$\begin{aligned}
&G_0(x^k, y^k) - G_0(x^*, y^*) - \nabla G_0(x^k, y^k)((x^k, y^k) - (x^*, y^*)) \\
&= \begin{pmatrix} x^k + y^k - \sqrt{(x^k - y^k)^2} \\ y^k - Mx^k - d \end{pmatrix} - \begin{pmatrix} I - W & I + W \\ -M & I \end{pmatrix} \begin{pmatrix} x^k - x^* \\ y^k - y^* \end{pmatrix} \\
&= 0,
\end{aligned}$$

as desired. □

In the next result, we show that under Condition 4.1 the algorithm is at least locally superlinear. The key to the proof is to show that the algorithm eventually rejects the centering step and finally switches to step 2 when the iterate approaches the solution set.

THEOREM 4.1. *Let M be a P_0 matrix. Let $p = 2 \le q$ and $\beta > 2\|a^{q/2}\| + \|(b, c)\|$. Assume that Condition 4.1 is satisfied. Then there exists a k_0 such that $\theta^{k+1} = (\theta^k)^2$ for all $k \ge k_0$, and*

$$\lim_{k \to \infty} \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} = 0,$$

which implies that the algorithm is locally superlinearly convergent.

Proof. Let $N(x^*, y^*)$ be a neighborhood of (x^*, y^*) defined as in Lemma 4.1. We first show that for all $(x^k, y^k) \in N(x^*, y^*)$, there exists a constant $\delta > 0$ such that

$$\|(\hat{x}^{k+1}, \hat{y}^{k+1}) - (x^*, y^*)\| \le \delta \max\{(\theta^k)^q, (\theta^k)^p\}\|(x^k, y^k) - (x^*, y^*)\|.$$

As we have pointed out, Condition 4.1 implies that $(x^k, y^k) \to (x^*, y^*)$, and there exist constants C and $\hat{\theta}$ such that

$$\|\nabla G_{\theta^k}(x^k, y^k)^{-1}\| \le C$$

for all $(x^k, y^k) \in N(x^*, y^*)$ and $\theta^k \in (0, \hat{\theta}]$. Therefore, for all sufficiently large k, by Lemma 4.1 we have

$$\begin{aligned}
&\|(\hat{x}^{k+1}, \hat{y}^{k+1}) - (x^*, y^*)\| \\
&= \|(x^k, y^k) - (x^*, y^*) - \nabla G_{\theta^k}(x^k, y^k)^{-1} G_0(x^k, y^k)\| \\
&= \|\nabla G_{\theta^k}(x^k, y^k)^{-1}\{[\nabla G_{\theta^k}(x^k, y^k) + \nabla G_0(x^k, y^k)]((x^k, y^k) - (x^*, y^*)) \\
&\quad - G_0(x^k, y^k) + G_0(x^*, y^*) - \nabla G_0(x^k, y^k)((x^k, y^k) - (x^*, y^*))\}\| \\
&\le \|\nabla G_{\theta^k}(x^k, y^k)^{-1}[\nabla G_{\theta^k}(x^k, y^k) - \nabla G_0(x^k, y^k)]((x^k, y^k) - (x^*, y^*))\| \\
&\quad + \|\nabla G_{\theta^k}(x^k, y^k)^{-1}[G_0(x^k, y^k) - G_0(x^*, y^*) - \nabla G_0(x^k, y^k)((x^k, y^k) - (x^*, y^*))]\| \\
&\le C\|\nabla G_{\theta^k}(x^k, y^k) - \nabla G_0(x^k, y^k)\|\|(x^k, y^k) - (x^*, y^*)\| \\
&\le C\kappa \max\{(\theta^k)^q, (\theta^k)^p\}\|(x^k, y^k) - (x^*, y^*)\|.
\end{aligned}$$

Set $\delta = C\kappa$. The desired inequality follows. The above inequality implies that the sequence $(\hat{x}^{k+1}, \hat{y}^{k+1})$ also converges to (x^*, y^*). Notice that $\theta^k \to 0$ (by Theorem 3.2). To show the local superlinear convergence of Algorithm 2.1, the above inequality implies that it is sufficient to show that the algorithm eventually takes the approximate Newton step alone.

Since (x^*, y^*) is a strictly complementary solution, $G_0(x, y)$ is continuously differentiable in the neighborhood of (x^*, y^*), and thus it must be Lipschitzian in the neighborhood of (x^*, y^*). Hence, there exists a constant $L > 0$ such that for all sufficiently large k

$$\begin{aligned}
\|G_0(\hat{x}^{k+1}, \hat{y}^{k+1}) - G_0(x^*, y^*)\| &\le L\|(\hat{x}^{k+1}, \hat{y}^{k+1}) - (x^*, y^*)\| \\
&\le L\delta \max\{(\theta^k)^q, (\theta^k)^p\}\|(x^k, y^k) - (x^*, y^*)\| \\
&= \tau_k \max\{(\theta^k)^q, (\theta^k)^p\},
\end{aligned}$$

where $\tau_k = L\delta\|(x^k, y^k) - (x^*, y^*)\| \to 0$ as $k \to \infty$. That is,

$$\|G_0(\hat{x}^{k+1}, \hat{y}^{k+1})\| \le \tau_k \max\{(\theta^k)^q, (\theta^k)^p\} \tag{4.5}$$

for all sufficiently large k. Setting $(\mu_1, \mu_2) = (\mu, 0)$ in (2.8), where $\mu \in (0, 1)$, we see from the first inequality in (2.8) that

$$\|G_\mu(x, y) - G_0(x, y)\| \le \mu\|(2\mu^{q/2-1}a^{q/2}, \mu^{p-1}A^p x)\| \text{ for all } (x, y) \in R^{2n}. \tag{4.6}$$

Thus, by using (4.5) and (4.6), for all sufficiently large k we have

$$\begin{aligned}
&\|G_{(\theta^k)^2}(\hat{x}^{k+1}, \hat{y}^{k+1}) - (\theta^k)^2(b, c)\| \\
&\le \|G_{(\theta^k)^2}(\hat{x}^{k+1}, \hat{y}^{k+1}) - G_0(\hat{x}^{k+1}, \hat{y}^{k+1})\| + \|G_0(\hat{x}^{k+1}, \hat{y}^{k+1})\|
\end{aligned}$$

$$
\begin{aligned}
&+(\theta^k)^2\|(b,c)\| \\
&\le (\theta^k)^2\|(2[(\theta^k)^2]^{q/2-1}a^{q/2}, [(\theta^k)^2]^{p-1}A^p\hat{x}^{k+1}\|) + \tau_k \max\{(\theta^k)^q, (\theta^k)^p\} \\
&+(\theta^k)^2\|(b,c)\| \\
&\le (\theta^k)^2(2\|a^{q/2}\| + \|(b,c)\| + [(\theta^k)^2]^{p-1}\|A^p\hat{x}^{k+1}\|) + \tau_k \max\{(\theta^k)^q, (\theta^k)^p\} \\
&= \beta(\theta^k)^2\left[\frac{2\|a^{q/2}\| + \|(b,c)\|}{\beta} + \frac{[(\theta^k)^2]^{p-1}\|A^p\hat{x}^{k+1}\|}{\beta}\right. \\
&\qquad\qquad \left. + \frac{\tau_k \max\{(\theta^k)^{q-2}, (\theta^k)^{p-2}\}}{\beta}\right] \\
&\le \beta(\theta^k)^2. \qquad (4.7)
\end{aligned}
$$

The third inequality follows from that $q \ge 2$ and $[(\theta^k)^2]^{q/2-1} \le 1$. The last inequality follows from the fact that $p = 2 \le q$, $\beta > 2\|a^{q/2}\| + \|(b,c)\|$, $\tau_k \to 0$, and

$$\lim_{k\to 0} \frac{[(\theta^k)^2]^{p-1}\|A^p\hat{x}^{k+1}\|}{\beta} = 0.$$

Thus, from (4.7), the approximate Newton step is accepted at the kth step provided that k is a large number. Therefore, the next iterate $(x^{k+1}, y^{k+1}) = (\hat{x}^{k+1}, \hat{y}^{k+1})$. Repeating the above proof, we can see that at the $(k+1)$th step $(x^{k+2}, y^{k+2}) = (\hat{x}^{k+2}, \hat{y}^{k+2})$, i.e., the approximate Newton step is still accepted at the $(k+1)$th step. By induction, we conclude that the algorithm eventually takes only the approximate Newton step. Hence, for some k_0, we have $\theta^{k+1} = (\theta^k)^2$ for all $k \ge k_0$, and $\lim_{k\to 0} \|x^{k+1} - x^*\|/\|x^k - x^*\| = 0$. □

The proof above shows that if an iterate (x^k, y^k) lies in a sufficiently small neighborhood of (x^*, y^*), then the next iterate still falls in this neighborhood and much closer to the solution (x^*, y^*) than (x^k, y^k). Since the centering step is gradually phased out and only approximate Newton steps are executed at the end of iteration, the superlinear convergence of the algorithm can be achieved.

5. Special cases. In this section, we show some much deeper global convergence results than Theorem 3.2 when the algorithm is applied to P_* LCPs. For the special case, the only assumption to ensure the global convergence is the nonemptiness of the solution set. In other words, this algorithm is able to solve any P_* LCP provided that a solution exists. For a given LCP, we denote

$$I = \{i : x_i^* > 0 \text{ for some solution } x^*\}, \qquad (5.1)$$

$$J = \{j : (Mx^* + d)_j > 0 \text{ for some solution } x^*\}, \qquad (5.2)$$

$$K = \{k : x_k^* = (Mx^* + d)_k = 0 \text{ for all solutions } x^*\}. \qquad (5.3)$$

The above partition of the set $\{1, 2, \ldots, n\}$ is unique for a given P_* LCP. Consider the affine set:

$$\overline{S} = \{(x,y) \in R^{2n} : \ x_{J\cup K} = 0, \ y_{I\cup K} = 0, \ y = Mx + d\}.$$

In fact, $\overline{S}$ is the affine hull of the solution set of the LCP, i.e., the smallest affine set containing the solution set. For any $(\tilde{x}, \tilde{y}) \in \overline{S}$, it is easy to see that $\tilde{x}_i\tilde{y}_i = 0$ for all $i = 1, \ldots, n$. We now prove a very useful result.

LEMMA 5.1. *Let $(\tilde{x}, \tilde{y})$ be an arbitrary vector in $\overline{S}$. Let M be a P_* matrix. Let $\{(x^k, y^k, \theta^k)\}$ be generated by Algorithm* 2.1 *and $(\bar{x}^k, \bar{y}^k)$ be defined by* (3.7). *Then*

$$\begin{aligned}
&\tilde{x}^T\bar{y}^k + \tilde{y}^T\bar{x}^k \\
&\quad \le (\theta^k)^q(1+\tau n)e^T a^q - \tau n\left(\min_{1\le i\le n}\rho_i^k\right) \\
&\quad\quad - (\bar{x}^k - \tilde{x})^T\left[(\theta^k)^p A^p\bar{x}^k + \theta^k(c+c^k) + \frac{1}{2}\theta^k(M - I + (\theta^k)^p A^p)(b+b^k)\right],
\end{aligned} \tag{5.4}$$

where

$$\begin{aligned}
\rho_i^k &= \bar{x}_i^k\tilde{y}_i + \tilde{x}_i\bar{y}_i^k \\
&\quad + (\bar{x}_i^k - \tilde{x}_i)\left\{(\theta^k)^p a_i^p\bar{x}_i^k + \theta^k(c_i + c_i^k) + \frac{1}{2}\theta^k[(M - I + (\theta^k)^p A^p)(b+b^k)]_i\right\}.
\end{aligned}$$

Proof. Since $(\bar{x}^k, \bar{y}^k) > 0$ and $\tilde{x}_i\tilde{y}_i = 0$ for all $i = 1, \ldots, n$, by (3.8) we have

$$\begin{aligned}
(\bar{x}_i^k - \tilde{x}_i)(\bar{y}_i^k - \tilde{y}_i) &= \bar{x}_i^k\bar{y}_i^k - \bar{x}_i^k\tilde{y}_i - \tilde{x}_i\bar{y}_i^k + \tilde{x}_i\tilde{y}_i \\
&= (\theta^k)^q a_i^q - \bar{x}_i^k\tilde{y}_i - \tilde{x}_i\bar{y}_i^k.
\end{aligned}$$

It is easy to verify that

$$\bar{y}^k = M\bar{x}^k + d + (\theta^k)^p A^p\bar{x}^k + \theta^k(c+c^k) + \frac{1}{2}\theta^k(M - I + (\theta^k)^p A^p)(b+b^k).$$

Thus, we have

$$\begin{aligned}
(\bar{x}_i^k - \tilde{x}_i)[M(\bar{x}^k - \tilde{x})]_i &= (\bar{x}_i^k - \tilde{x}_i)[(M\bar{x}^k + d)_i - \tilde{y}_i] \\
&= (\bar{x}_i^k - \tilde{x}_i)\Bigg\{\bar{y}_i^k - (\theta^k)^p a_i^p\bar{x}_i^k - \theta^k(c_i + c_i^k) \\
&\quad - \frac{1}{2}\theta^k[(M - I + (\theta^k)^p A^p)(b+b^k)]_i - \tilde{y}_i\Bigg\} \\
&= (\bar{x}_i^k - \tilde{x}_i)(\bar{y}_i^k - \tilde{y}_i) - (\bar{x}_i^k - \tilde{x}_i)\Bigg\{(\theta^k)^p a_i^p\bar{x}_i^k \\
&\quad + \theta^k(c_i + c_i^k) + \frac{1}{2}\theta^k[(M - I + (\theta^k)^p A^p)(b+b^k)]_i\Bigg\} \\
&\le (\theta^k)^q a_i^q - \bar{x}_i^k\tilde{y}_i - \tilde{x}_i\bar{y}_i^k - (\bar{x}_i^k - \tilde{x}_i)\Bigg\{(\theta^k)^p a_i^p\bar{x}_i^k \\
&\quad + \theta^k(c_i + c_i^k) + \frac{1}{2}\theta^k[(M - I + (\theta^k)^p A^p)(b+b^k)]_i\Bigg\} \\
&\le (\theta^k)^q e^T a^q - \min_{1\le i\le n}\rho_i^k,
\end{aligned} \tag{5.5}$$

where

$$\begin{aligned}
\rho_i^k &= \bar{x}_i^k\tilde{y}_i + \tilde{x}_i\bar{y}_i^k + (\bar{x}_i^k - \tilde{x}_i)\{(\theta^k)^p a_i^p\bar{x}_i^k + \theta^k(c_i + c_i^k) \\
&\quad + \frac{1}{2}\theta^k[(M - I + (\theta^k)^p A^p)(b+b^k)]_i\}.
\end{aligned}$$

Therefore, by (3.8), (5.5), and the definition of the P_* matrix, we have

$$\begin{aligned}
\tilde{x}^T\bar{y}^k+\tilde{y}^T\bar{x}^k &= -(\bar{x}^k-\tilde{x})^T(\bar{y}^k-\tilde{y})+(\bar{x}^k)^T\bar{y}^k\\
&= -(\bar{x}^k-\tilde{x})^T\Bigg[M\bar{x}^k+d+(\theta^k)^pA^p\bar{x}^k+\theta^k(c+c^k)\\
&\quad+\frac{1}{2}\theta^k(M-I+(\theta^k)^pA^p)(b+b^k)-\tilde{y}\Bigg]+(\theta^k)^qe^Ta^q\\
&= -(\bar{x}^k-\tilde{x})^T(M\bar{x}^k+d-\tilde{y})-(\bar{x}^k-\tilde{x})^T\Bigg[(\theta^k)^pA^p\bar{x}^k+\theta^k(c+c^k)\\
&\quad+\frac{1}{2}\theta^k(M-I+(\theta^k)^pA^p)(b+b^k)\Bigg]+(\theta^k)^qe^Ta^q\\
&= -(\bar{x}^k-\tilde{x})^TM(\bar{x}^k-\tilde{x})-(\bar{x}^k-\tilde{x})^T\Bigg[(\theta^k)^pA^p\bar{x}^k+\theta^k(c+c^k)\\
&\quad+\frac{1}{2}\theta^k(M-I+(\theta^k)^pA^p)(b+b^k)\Bigg]+(\theta^k)^qe^Ta^q\\
&\le \tau\sum_{i\in I_+}(\bar{x}_i^k-\tilde{x}_i)[M(\bar{x}^k-\tilde{x})]_i-(\bar{x}^k-\tilde{x})^T\Bigg[(\theta^k)^pA^p\bar{x}^k\\
&\quad+\theta^k(c+c^k)+\frac{1}{2}\theta^k(M-I+(\theta^k)^pA^p)(b+b^k)\Bigg]+(\theta^k)^qe^Ta^q\\
&\le \tau n\left((\theta^k)^qe^Ta^q-\min_{1\le i\le n}\rho_i^k\right)-(\bar{x}^k-\tilde{x})^T\Bigg[(\theta^k)^pA^p\bar{x}^k\\
&\quad+\theta^k(c+c^k)+\frac{1}{2}\theta^k(M-I+(\theta^k)^pA^p)(b+b^k)\Bigg]+(\theta^k)^qe^Ta^q\\
&= (\theta^k)^q(1+\tau n)e^Ta^q-\tau n\left(\min_{1\le i\le n}\rho_i^k\right)-(\bar{x}^k-\tilde{x})^T\Bigg[(\theta^k)^pA^p\bar{x}^k\\
&\quad+\theta^k(c+c^k)+\frac{1}{2}\theta^k(M-I+(\theta^k)^pA^p)(b+b^k)\Bigg].
\end{aligned}$$

The proof is complete. □

The following result shows that under a suitable choice of parameters our algorithm can locate a solution of the P_* LCP as long as a solution exists.

THEOREM 5.1. *Let M be a P_* matrix. Assume that the solution set of the P_* LCP is nonempty. If one of the following holds,*

(i) $p\le 1$,

(ii) $p>1, c>\frac{1}{2}\|(M-I)b\|e$, *and* $0<\beta<\min_{1\le i\le n}\frac{c_i-(1/2)\|(M-I)b\|}{1+\|M-I\|/2}$,

then the sequence $\{(x^k,y^k,\theta^k)\}$, generated by Algorithm 2.1, is bounded, and

$$\lim_{k\to\infty}\theta^k\to 0,\qquad \lim_{k\to\infty}\|G_{\theta^k}(x^k,y^k)\|=0.$$

Therefore, any accumulation point of (x^k,y^k) is a solution to the LCP.

Proof. We focus on the proof of the boundedness of $\{(x^k, y^k)\}$. Let (x^*, y^*) be an arbitrary solution to the LCP. Set $(\tilde{x}, \tilde{y}) = (x^*, y^*)$ in Lemma 5.1. Since for this case $\bar{y}_i^k x_i^* + \bar{x}_i^k y_i^* \geq 0$, we have that

$$\rho_i^k \geq \eta_i^k := (\bar{x}_i^k - x_i^*) \left\{ (\theta^k)^p a_i^p \bar{x}_i^k + \theta^k (c_i + c_i^k) + \frac{1}{2}\theta^k [(M - I + (\theta^k)^p A^p)(b + b^k)]_i \right\}.$$

This, together with (5.4), implies that

$$(x^*)^T \bar{y}^k + (y^*)^T \bar{x}^k \leq (\theta^k)^q (1 + \tau n) e^T a^q - \tau n \left(\min_{1 \leq i \leq n} \eta_i^k \right) - (\bar{x}^k - x^*)^T \left\{ (\theta^k)^p A^p \bar{x}^k \right.$$

$$\left. + \theta^k (c + c^k) + \frac{1}{2}\theta^k (M - I + (\theta^k)^p A^p)(b + b^k) \right\}. \tag{5.6}$$

Dividing both sides of the above by $(\theta^k)^p$ and noting that the left-hand side is nonnegative, we have

$$\begin{aligned} &(\bar{x}^k - x^*)^T A^p \bar{x}^k + \tau n \left(\min_{1 \leq i \leq n} \frac{\eta_i^k}{(\theta^k)^p} \right) \\ &+ (\theta^k)^{1-p} (\bar{x}^k - x^*)^T \left[c + c^k + \frac{1}{2}(M - I + (\theta^k)^p A^p)(b + b^k) \right] \\ &\leq (\theta^k)^{q-p} (1 + \tau n) e^T a^q. \end{aligned} \tag{5.7}$$

If $p \leq 1$, the right-hand side of the above inequality is bounded since $q \geq 1$ and $\theta_k < 1$. This implies that the sequence $\{\bar{x}^k\}$ is bounded (otherwise the left-hand side is unbounded from above), and thus $\{x^k\}$ is bounded, as is $\{y^k\}$ by (3.6). The boundedness of $\{(x^k, y^k)\}$ under (i) is proved.

We now prove the boundedness of (x^k, y^k) in the case (ii). Consider two subcases.

Subcase 1: $\theta_k \not\to 0$. In this case, there exists a constant $\hat{\theta} > 0$ such that $1 > \theta^k \geq \hat{\theta}$. It is easy to see from (5.7) that the sequence $\{\bar{x}^k\}$ is bounded, and thus (x^k, y^k) is bounded.

Subcase 2: $\theta_k \to 0$. In this case, by the choice of p, β, and c, it is easy to see that

$$\begin{aligned} c_i + c_i^k + \frac{1}{2}[(M - I)(b + b^k)]_i &> c_i - \beta - \frac{1}{2}\|(M - I)(b + b^k)\| \\ &\geq c_i - \beta - \frac{1}{2}\|(M - I)b\| - \frac{1}{2}\|(M - I)b^k\| \\ &\geq c_i - \frac{1}{2}\|(M - I)b\| - \beta \left(1 + \frac{1}{2}\|M - I\| \right) \\ &> 0. \end{aligned} \tag{5.8}$$

Since $\theta_k \to 0$, for all sufficiently large k it follows that

$$c_i + c_i^k + \frac{1}{2}[(M - I + (\theta_k)^p A^p)(b + b^k)]_i > 0.$$

Thus, for all sufficiently large k we have

$$\begin{aligned} \frac{\eta_i^k}{\theta^k} &\geq -x_i^* \left\{ (\theta^k)^{p-1} a_i^p \bar{x}_i^k + c_i + c_i^k + \frac{1}{2}[(M - I + (\theta^k)^p A^p)(b + b^k)]_i \right\} \\ &\geq -(\theta^k)^{p-1} (x^*)^T A^p \bar{x}^k \\ &\quad - \max_{1 \leq i \leq n} x_i^* \left\{ c_i + c_i^k + \frac{1}{2}[(M - I + (\theta^k)^p A^p)(b + b^k)]_i \right\}. \end{aligned} \tag{5.9}$$

Since the left-hand side of (5.6) is nonnegative, dividing both sides of (5.6) by θ^k and using (5.9), we have

$$\begin{aligned}
0 &\le (\theta^k)^{q-1}(1+\tau n)e^T a^q - \tau n\left(\min_{1\le i\le n}\frac{\eta_i^k}{\theta^k}\right) + (\theta^k)^{p-1}(x^*)^T A^p \bar{x}^k \\
&\quad -(\bar{x}^k - x^*)^T(c+c^k) - \frac{1}{2}(\bar{x}^k - x^*)^T(M - I + (\theta^k)^p A^p)(b+b^k) \\
&\le (\theta^k)^{q-1}(1+\tau n)e^T a^q + \tau n \max_{1\le i\le n} x_i^*\left\{c_i + c_i^k + \frac{1}{2}[(M - I + (\theta^k)^p A^p)(b+b^k)]_i\right\} \\
&\quad +(\theta^k)^{p-1}(1+\tau n)(x^*)^T A^p \bar{x}^k - (\bar{x}^k - x^*)^T(c+c^k) \\
&\quad -\frac{1}{2}(\bar{x}^k - x^*)^T(M - I + (\theta^k)^p A^p)(b+b^k).
\end{aligned}$$

It follows that

$$\begin{aligned}
&(\bar{x}^k)^T\left[c + c^k - (\theta^k)^{p-1}(1+\tau n)A^p x^* + \frac{1}{2}(M - I + (\theta^k)^p A^p)(b+b^k)\right] \\
&\le (\theta^k)^{q-1}(1+\tau n)e^T a^q + \tau n \max_{1\le i\le n} x_i^*\Big\{c_i + c_i^k + \frac{1}{2}[(M - I \\
&\quad +(\theta^k)^p A^p)(b+b^k)]_i\Big\} + (x^*)^T\left[c + c^k + \frac{1}{2}(M - I + (\theta^k)^p A^p)(b+b^k)\right]. \qquad (5.10)
\end{aligned}$$

Since $p > 1$ and $\theta^k \to 0$, by a proof similar to (5.8), for all sufficiently large k we have

$$\begin{aligned}
&c + c^k - (\theta^k)^{p-1}(1+\tau n)A^p x^* + \frac{1}{2}(M - I + (\theta^k)^p A^p)(b+b^k) \\
&\ge \frac{1}{2}\left\{c - \frac{1}{2}\|(M-I)b\|e - \beta\left(1 + \frac{1}{2}\|M - I\|\right)e\right\} \\
&> 0.
\end{aligned}$$

Since the right-hand side of (5.10) is bounded and $\bar{x}^k > 0$, from the above inequality and (5.10) it follows that $\{\bar{x}^k\}$ is bounded, and hence (x^k, y^k) is bounded.

Based on the boundedness of $\{(x^k, y^k)\}$, repeating the proof of Theorem 3.2 we can prove that $\theta^k \to 0$. □

Remark 5.1. It is worth mentioning the difference between (i) and (ii) of the above theorem. In the case (i), there is no restriction on the parameter $\beta > 0$. Thus, β can be assigned a large number so that the neighborhood is wide enough to ensure a large steplength at each iteration. For the case (ii), however, the parameter β is required to be relatively small. To satisfy this requirement, the initial point of Algorithm 2.1 can be also obtained easily. For example, set $x^0 = 0, a \in R^n_{++}, \theta^0 \in (0,1)$, and choose $y^0 \in R^n_{++}$ to be large enough such that $c > \frac{1}{2}\|(M-I)b\|e$, where

$$b = \frac{x^0 + y^0 - \sqrt{(x^0 - y^0)^2 + 4(\theta^0)^q a^q}}{\theta^0} = \frac{4(\theta^0)^q a^q}{y^0 + \sqrt{(y^0)^2 + 4(\theta^0)^q a^q}},$$

$$c = \frac{y^0 - (f(x^0) + (\theta^0)^p A^p x^0)}{\theta^0} = \frac{y^0 - f(0)}{\theta^0}.$$

The above choice implies that $\|G_{\theta^0}(x^0, y^0)\| = 0$. Thus, $(x^0, y^0) \in \mathcal{N}(\beta, \theta^0)$ for any $\beta > 0$. In particular, β can be taken such that

$$0 < \beta < \min_{1\le i\le n} \frac{c_i - (1/2)\|(M-I)b\|}{1 + \|M - I\|/2}.$$

In the rest of this section, we characterize the accumulation point of the sequence $\{(x^k, y^k)\}$. We first recall some concepts. Let S denote the solution set of the LCP. An element x^* of S is said to be the *N-norm least solution*, where N is a positive, definite, symmetric matrix, if $\|N^{1/2}x^*\| \le \|N^{1/2}u\|$ for all $u \in S$. In particular, if $N = I$, the solution x^* is called the *least 2-norm solution* of S. An element x^* of S is said to be the *least element* of S if $x^* \le u$ for all $u \in S$ (see, for example, [30, 13]). The solution x^* is called a maximally complementary solution if $x_i^* > 0$ for all $i \in I$, $(Mx^* + d)_i > 0$ for all $i \in J$, and $x_i^* = (Mx^* + d)_i = 0$ for all $i \in K$. Clearly, a strictly complementary solution is a maximally complementary solution with $K = \emptyset$.

THEOREM 5.2. *Let M be a P_* matrix. Assume that the solution set of the LCP is nonempty.*

(i) *If $p < 1$, then every accumulation point $(\hat{x}, \hat{y})$ of the sequence (x^k, y^k) satisfies the following property: For any solution x^*, there exists a corresponding index i_0 such that*

$$(\hat{x})^T A^p(\hat{x} - x^*) + \tau n a_{i_0}^p \hat{x}_{i_0}(\hat{x}_{i_0} - x_{i0}^*) \le 0. \tag{5.11}$$

Moreover, if the least element solution exists, then the entire sequence (x^k, y^k) is convergent, and its accumulation point coincides with the least element solution.

(ii) *If $p > 1, c > \frac{1}{2}\|(M - I)b\|e$, $0 < \beta < \lim_{1\le i\le n} \frac{c_i - (1/2)\|(M-I)b\|}{1+\|M-I\|/2}$, and $q = 1$, then each accumulation point is a maximally complementary solution of the LCP.*

Proof. For $p < 1$, by the result (i) of Theorem 5.1, $\{x^k\}$ is bounded and $\theta^k \to 0$. Let $(\hat{x}, \hat{y})$ be an arbitrary accumulation point of $\{(x^k, y^k)\}$. Taking the limit in (5.7) where x^* is an arbitrary solution of the LCP, we see that there exists an index i_0 such that

$$(\hat{x})^T A^p(\hat{x} - x^*) + \tau n a_{i_0}^p \hat{x}_{i_0}(\hat{x}_{i_0} - x_{i0}^*) \le 0.$$

Moreover, if the least element solution exists, setting x^* to be the least element, we conclude from the above inequality that $\hat{x}$ is equal to the least element. Since such an element is unique, the sequence $\{x^k\}$ is convergent.

We now consider the case (ii). By result (ii) of Theorem 5.1, the sequence (x^k, y^k) is bounded, $\theta^k \to 0$, and each accumulation point of (x^k, y^k) is a solution to the LCP. Let (x^*, y^*) be a maximally complementary solution and I, J, K be defined by (5.1)–(5.3). Then we have

$$\begin{aligned}(x^*)^T\bar{y}^k + (y^*)^T\bar{x}^k &= (x_I^*)^T\bar{y}_I^k + (y_J^*)^T\bar{x}_J^k \\ &= (x_I^*)^T(\bar{X}_I^k)^{-1}\bar{X}_I^k\bar{y}_I^k + (y_J^*)^T(\bar{Y}_J^k)^{-1}\bar{Y}_J^k\bar{x}_J^k \\ &= (\theta^k)^q\left[(x_I^*)^T(\bar{X}_I^k)^{-1}a_I^q + (y_J^*)^T(\bar{Y}_J^k)^{-1}a_J^q\right].\end{aligned}$$

By (5.6) and the above inequality, we have

$$\begin{aligned}&(x_I^*)^T(\bar{X}_I^k)^{-1}a_I^q + (y_J^*)^T(\bar{Y}_J^k)^{-1}a_J^q \\ &\le (1 + \tau n)e^T a^q - \tau n\left(\min_{1\le i\le n}\frac{\eta_i^k}{(\theta^k)^q}\right) - (\bar{x}^k - x^*)^T\Bigg[(\theta^k)^{p-q}A^p\bar{x}^k \\ &\quad + (\theta^k)^{1-q}(c + c^k) + \frac{1}{2}(\theta^k)^{1-q}(M - I + (\theta^k)^p A^p)(b + b^k)\Bigg].\end{aligned}$$

Let $(\hat{x}, \hat{y})$ be an arbitrary accumulation point of the iterates. Since $\theta^k \to 0$ and $p > 1 = q$, we can see that $\eta_i^k/(\theta^k)^q$ is bounded. The right-hand side of the above

inequality is bounded. Since $(x_I^*, y_J^*) > 0$, we conclude that $\bar{x}_I^k \to \hat{x}_I > 0$; otherwise, if $\hat{x}_i = 0$ for some $i \in I$, then $x_i^*/\bar{x}_i \to \infty$, and hence the left-hand side tends to infinity, contradicting the boundedness of the right-hand side. In a similar way, we have that $\hat{y}_I > 0$. Thus, $(\hat{x}, \hat{y})$ is a maximally complementary solution. □

Since every positive semidefinite matrix is a P_* matrix with $\tau = 0$, the result (i) above can be further improved for monotone LCPs. In fact, from Theorem 2.3, the following result is natural since the algorithm follows the regularized central path approximately.

THEOREM 5.3. *Let M be a positive semidefinite matrix. Assume that the solution set of the LCP is nonempty. For $p < 1$, the entire sequence (x^k, y^k), generated by Algorithm* 2.1*, converges to $(\hat{x}, \hat{y})$, where $\hat{x}$ is the least N-norm solution with $N = A^{p/2}$. In particular, if $a = e$ is taken, the sequence converges to the (unique) least* 2*-norm solution.*

Proof. For the case of $p < 1$, setting $\tau = 0$ in (5.11) we have

$$(\hat{x})^T A^p (\hat{x} - x^*) \le 0,$$

which implies that $\|A^{p/2}\hat{x}\| \le \|A^{p/2}x^*\|$. Since x^* is an arbitrary solution, it follows that the solution $\hat{x}$ is the least N-norm solution where $N = A^{p/2}$. It is also easy to see from the above inequality that the solution $\hat{x}$ is unique, and thus the entire sequence is convergent. □

Remark 5.2. For P_* LCPs, the boundedness assumption of the solution set (or the strict feasibility condition) is not required for the global convergence of our algorithm. Further, all results in this section can be easily extended to nonlinear P_* complementarity problems. We notice that Ye's homogeneous model [41] for monotone LCPs, which was later generalized to nonlinear monotone complementarity problems by Andersen and Ye [2], also does not require the boundedness of the solution set (or the strict feasibility) of the original problem. However, it is unknown whether Ye's algorithm can be generalized to the nonlinear P_* problems.

6. Numerical examples. Algorithm 2.1 was tested on some LCPs, nonlinear complementarity problems (NCPs), and nonlinear programming problems (NLPs) which can be written as complementarity problems by KKT optimality conditions. For all test examples, common parameters and initial points were used in our algorithm. From the analysis of section 4 and our experiments, the value of parameters p and q should be relatively large for the sake of rapid convergence. The constant σ should be taken relatively small such that a possible large steplength λ_k can be taken. The vector $(a, b, c) \in R^n_{++} \times R^{2n}$ and the initial point $(x^0, y^0) \in R^{2n}$ can be chosen freely. In general cases, the value of β should be taken relatively large to ensure that the neighborhood is wide enough to permit a large iterative steplength. Thus, the parameters used in our code were set as $p = 2,\ q = 3,\ \sigma = 0.001$ and $\alpha = 0.9$. The vectors $a,\ b,\ c$ were set as $a = b = c = e$. The initial values of (x^0, y^0, θ^0) were set as $\theta^0 = 0.9$ and $x^0 = y^0 = e$. The parameter β was given by

$$\beta = \frac{\|G_{\theta^0}(x^0, y^0) - \theta^0(b, c)\|}{\theta^0} + 100.$$

Since $G_0(x^*, y^*) = 0$ if and only if (x^*, y^*) is a solution to the complementarity problem, we use $\|G_0(x^k, y^k)\| < \varepsilon$ as the stopping criterion, where $\varepsilon > 0$ is a given tolerance. In our experiments, $\varepsilon = 10^{-14}$ was taken for all numerical examples. All results were undertaken on a DEC Alpha V4.0 workstation by Fortran 90, and all the

arithmetic operations were performed in double precision to avoid round-off errors. We recorded the following aspects to examine the effectiveness of the algorithm: the dimension of problems, the total number of iterations, the total number of functions called, the CPU time used, the final value of θ^k, and the residual, i.e., the final value of $\|G_0(x^k, y^k)\|$. All CPU times reported here include time for input and output. We now introduce test examples and provide the numerical results for them.

Linear complementarity problems.

LCP1. This is Watson's first problem [37].

LCP2. This is Watson's second problem [37].

LCP3. The matrix M_1 is a P_* matrix given in (6.1), and $d = -e$. The solution set is unbounded. There is no strictly feasible point for this LCP. The central path does not exist for this problem. However, Algorithm 2.1 deals with this problem very efficiently.

LCP4. This is a P_0 LCP given by Chen and Ye [12]. The matrix M_2 is given in (6.1), and $d = (0, 0, 1)$. The solution set is unbounded.

LCP5. This is a P_0 LCP with the matrix M_3 given in (6.1), and $d = (0, 0, 1)$. This problem has no strictly feasible point, and its solution set is unbounded:

$$M_1 = \begin{pmatrix} 0 & 0 & 2 & 1 \\ 0 & 0 & 1 & 2 \\ -2 & -1 & 0 & 0 \\ 4 & 8 & 0 & 0 \end{pmatrix}, \; M_2 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 1 \end{pmatrix}, \; M_3 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & -2 \\ 0 & 2 & 1 \end{pmatrix}. \tag{6.1}$$

LCP6. This example was given by Fathi [15]. The matrix M_4 is given in (6.2) and the vector $d = -e$.

LCP7. This example was given by Ahn [1]. The vector $d = -e$, and the matrix M_5 is given in (6.2):

$$M_4 = \begin{pmatrix} 1 & 2 & 2 & \cdots & 2 \\ 2 & 5 & 6 & \cdots & 6 \\ 2 & 6 & 9 & \cdots & 10 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 2 & 6 & 9 & \cdots & 4n-3 \end{pmatrix}, \; M_5 = \begin{pmatrix} 4 & -2 & 0 & \cdots & 0 \\ 1 & 4 & -2 & \cdots & 0 \\ 0 & 1 & 4 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 4 \end{pmatrix}. \tag{6.2}$$

LCP8. This example was used by Geiger and Kanzow [16], where $d = -e$ and the matrix M_7 is given as in (6.3).

LCP9. This LCP was given in [29]. The matrix M_8 is given in (6.3) and $d = -e$:

$$M_7 = \begin{pmatrix} 4 & -1 & 0 & 0 & 0 & \cdots & 0 \\ -1 & 4 & -1 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 4 & -1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & \cdots & 0 & -1 & 4 \end{pmatrix}, \; M_8 = \begin{pmatrix} 1 & 2 & 2 & \cdots & 2 \\ 0 & 1 & 2 & \cdots & 2 \\ 0 & 0 & 1 & \cdots & 2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}. \tag{6.3}$$

LCP10. This example can be found in [16], where $M = \mathrm{diag}(1/n,\ 2/n,\ \ldots,\ 1)$ and $d = -e$.

LCP11. The matrix is obtained from M_5 by replacing the first diagonal entry by -4, and the vector $d = (0, 1, \ldots, 1)$. This LCP has no strictly feasible point.

LCP12. The matrix is obtained from M_7 by replacing the first diagonal entry by -4, and the vector $d = (0, 0, 1, \ldots, 1)$. This LCP has no strictly feasible point.

TABLE 6.1
LCPs, $\varepsilon = 1e-14$.

Problems	Dim.	No. of Iter.	No. of fun.	θ^k	Residual	CPU (sec.)
LCP1	10	8	9	1.4e−06	2.4e−15	0.00
LCP2	5	9	10	1.9e−12	1.6e−15	0.00
LCP3	4	8	9	1.4e−06	2.1e−16	0.00
LCP4	3	8	9	1.4e−06	5.5e−17	0.00
LCP5	3	8	9	1.4e−06	7.8e−17	0.00
LCP6	300	12	19	1.9e−16	9.9e−16	6.53
LCP6	500	12	19	1.9e−16	8.3e−16	42.75
LCP7	300	8	9	1.4e−06	2.3e−15	4.41
LCP7	500	8	9	1.4e−06	3.0e−15	29.16
LCP8	300	8	9	1.4e−06	2.7e−15	4.54
LCP8	500	8	9	1.4e−06	3.6e−15	28.52
LCP9	300	10	13	1.9e−16	0.0	5.66
LCP9	500	10	13	1.9e−16	0.0	40.12
LCP10	300	10	13	3.3e−13	5.6e−15	5.50
LCP10	500	11	16	2.2e−16	8.1e−15	40.57
LCP11	300	9	10	1.9e−12	4.4e−18	6.08
LCP11	500	9	10	1.9e−12	3.7e−18	35.83
LCP12	300	9	10	2.1e−12	2.4e−16	5.61
LCP12	500	9	10	2.0e−12	5.5e−16	34.08
LCP13	300	10	13	1.9e−16	1.4e−17	6.38
LCP13	500	10	13	1.8e−16	1.3e−17	39.50

LCP13. The matrix is obtained from M_8 by replacing the last diagonal entry by -1, and the vector $d = (-1, \ldots, -1, 0)$. This LCP has no strictly feasible point.

Nonlinear complementarity problems.

NCP1. (Kojima–Shindo [31]) This is an NCP which is difficult to solve by the conventional Newton-type methods.

NCP2. (Watson's fourth problem [37]) This is an NCP representing the KKT conditions for a convex programming problem.

NCP3. (Mathiesen's Walrasian equilibrium model [31]) This is a 4-variable equilibrium problem depending on three parameters (α, b_2, b_3). We use two sets of constants: $(\alpha, b_2, b_3) = (0.75, 1, 0.5)$ and $(0.75, 1, 2)$. In Table 6.2, NCP3a and NCP3b denote, respectively, the problems corresponding to the above two cases.

NCP4. (invariant capital stock model [31]) This is an NCP (see [31]) formulated from an invariant capital stock model described by Hansen and Koopmans.

NCP5. (Nash–Cournot production problem [18]) We solve this NCP problem with $\gamma = 1.1$, and the data α_i, L_i, β_i can be found in [18]. The 5- and 10-variable problems were solved in our experiments. We use NCP5a and NCP5b in Table 6.2 to denote the 5- and 10-variable problems, respectively.

For NCPs, the number of evaluations of the Jacobian $\nabla f(x)$ should be recorded. However, by the construction of the algorithm, the total number of evaluations of the Jacobian $\nabla f(x)$ equals to the total number of iterations, and hence it is omitted here.

Nonlinear programming problems. We also test the algorithm for some NLPs. These examples can be found in Hock and Schittkowski [19]. We solve these examples via the KKT conditions for these problems, which can be formulated as complementarity problems.

The computational results for LCPs are summarized in Table 6.1, those for NCPs are reported in Table 6.2, and those for NLPs are summarized in Table 6.3, in which the "Dim" stands for the dimension of the corresponding complementarity problems.

TABLE 6.2
NCPs, $\varepsilon = 1e-14$.

Problems	Dim.	No. of Iter.	No. of fun.	θ^k	Residual	CPU (sec.)
NCP1	4	9	12	3.2e−09	1.3e−15	0.00
NCP2	5	16	35	4.0e−15	3.8e−16	0.00
NCP3a	4	8	9	1.4e−06	2.7e−16	0.00
NCP3b	4	8	9	1.4e−06	1.3e−16	0.00
NCP4	14	9	10	1.9e−12	5.6e−16	0.00
NCP5a	5	8	9	1.3e−06	6.9e−15	0.00
NCP5b	10	14	33	1.0e−17	9.7e−15	0.00

TABLE 6.3
NLPs, $\varepsilon = 1e-14$.

Problems	Dim.	No. of Iter.	No. of fun.	θ^k	Residual	CPU (sec.)
HS18	7	17	76	5.2e−11	6.5e−16	0.00
HS24	4	7	8	1.5e−06	2.3e−16	0.00
HS33	6	12	19	2.1e−10	9.6e−16	0.00
HS34	8	10	33	7.3e−11	1.8e−15	0.00
HS35	4	8	9	1.3e−06	1.5e−15	0.00
HS36	7	14	90	1.0e−13	8.9e−16	0.00
HS44	10	8	9	1.4e−06	1.5e−15	0.00
HS63	7	9	82	8.1e−08	6.1e−15	0.00
HS66	8	14	56	3.8e−10	1.0e−15	0.00

From the experiments, we found that the algorithm can solve all these examples effectively. It should be pointed out that the NCP1 is difficult to solve by conventional Newton-type methods, and as pointed out in [37] none of the standard algebraic techniques can solve the LCP2 easily. However, the proposed algorithm deals with the two problems very easily, and a quick convergence is observed. We also note that the value of β has a close relation to the convergence speed of the algorithm. The convergence speed will be slow if β is too small. In fact, a big value of β enables a large iterative steplength to be taken such that a rapid convergence can be achieved. This is indeed shown from our experiments.

7. Final remarks. A new non–interior-point algorithm is presented for P_0 LCPs. The global convergence of the algorithm is proved under a new condition which is different from ones previously used in the literature. A good feature of this condition is that it does not imply the boundedness of the solution set of the problem. Especially for P_* LCPs, the algorithm is globally convergent, provided that a solution exists. The superlinear convergence of the algorithm is also proved under a standard nondegeneracy assumption and a suitable choice of some parameters. The effectiveness of the algorithm was verified by our numerical experiments.

The essence of our algorithm is to follow a newly introduced regularized central path whose existence and theoretical properties were proved in [43]. Although the discussion in this paper was limited to LCPs, all the analysis in this paper can be extended to nonlinear P_0 complementarity problems as long as the function f is assumed to be continuously differentiable and Lipschitzian.

Acknowledgments. The authors would like to thank the anonymous referees for their helpful comments and suggestions that helped improve the paper.

REFERENCES

[1] B. H. Ahn, *Iterative methods for linear complementarity problem with upperbounds and lowerbounds,* Math. Program., 26 (1983), pp. 265–315.

[2] E. D. Andersen and Y. Ye, *On a homogeneous algorithm for the monotone complementarity problem,* Math. Program., 84 (1999), pp. 375–399.

[3] J. V. Burke and S. Xu, *The global linear convergence of a non-interior path-following algorithm for linear complementarity problems,* Math. Oper. Res., 23 (1998), pp. 719–734.

[4] J. V. Burke and S. Xu, *A non-interior predictor-corrector path following algorithm for the monotone linear complementarity problem,* Math. Program., 87 (2000), pp. 113–130.

[5] J. V. Burke and S. Xu, *The complexity of a non-interior-path following method for the linear complementarity problem*, J. Optim. Theory Appl., 112 (2002), pp. 53–76.

[6] 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.

[7] B. Chen and X. Chen, *A global and local superlinear continuation-smoothing method for P_0 and R_0 or monotone NCP,* SIAM J. Optim., 9 (1999), pp. 624–645.

[8] B. Chen and N. Xiu, *A global linear and local quadratic noninterior continuation method for nonlinear complementarity problems based on Chen–Mangasarian smoothing functions,* SIAM J. Optim., 9 (1999), pp. 605–623.

[9] C. Chen and O. L. Mangasarian, *A class of smoothing functions for nonlinear and mixed complementarity problems*, Comput. Optim. Appl., 5 (1996), 1168–1190.

[10] X. Chen, L. Qi, and D. Sun, *Global and superlinear convergence of the smoothing Newton method and its application to general box constrained variational inequalities,* Math. Comp., 67 (1998), pp. 519–540.

[11] X. Chen and Y. Ye, *On homotopy-smoothing methods for box-constrained variational inequalities,* SIAM J. Control Optim., 37 (1999), pp. 589–616.

[12] X. Chen and Y. Ye, *On smoothing methods for the P_0 matrix linear complementarity problem,* SIAM J. Optim., 11 (2000), pp. 341–363.

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

[14] F. Facchinei and C. Kanzow, *Beyond monotonicity in regularization methods for nonlinear complementarity problems,* SIAM J. Control Optim., 37 (1999), pp. 1150–1161.

[15] Y. Fathi, *Computational complexity of LCPs associated with positive definite matrices,* Math. Program., 17 (1979), pp. 355–344.

[16] C. Geiger and C. Kanzow, *On the resolution of monotone complementarity problems,* Comput. Optim. Appl., 5 (1996), pp. 155–173.

[17] M. S. Gowda and R. Sznajder, *Weak univalence and connectedness of inverse images of continuous functions*, Math. Oper. Res., 24 (1999), pp. 255–261.

[18] P. T. Harker, *Accelerating the convergence of the diagonalization and projection algorithms for finite-dimensional variational inequalities,* Math. Program., 41 (1988), pp. 25–59.

[19] W. Hock and K. Schittkowski, *Test Examples for Nonlinear Programming Codes,* Lecture Notes in Econ. and Math. Systems 187, Springer-Verlag, Berlin, 1981.

[20] K. Hotta and A. Yoshise, *Global convergence of a class of non-interior-point algorithms using Chen-Harker-Kanzow-Smale functions for nonlinear complementarity problems,* Math. Program., 86 (1999), pp. 105–133.

[21] K. Hotta, M. Inaba, and A. Yoshise, *A complexity analysis of a smoothing method using CHKS-function for monotone linear complementarity problems,* Comput. Optim. Appl., 17 (2000), pp. 183–201.

[22] G. Isac, *Tikhonov's regularization and the complementarity problem in Hilbert space,* J. Math. Anal. Appl., 174 (1993), pp. 53–66.

[23] C. Jones and M. S. Gowda, *On the connectedness of solution sets in linear complementarity problems,* Linear Algebra Appl., 272 (1998), pp. 33–44.

[24] C. Kanzow, *Some nonlinear continuation methods for linear complementarity problems,* SIAM J. Matrix Anal. Appl., 17 (1996), pp. 851–865.

[25] M. Kojima, N. Megiddo, and T. Noma, *Homotopy continuation methods for nonlinear complementarity problems,* Math. Oper. Res., 16 (1991), pp. 754–774.

[26] M. Kojima, N. Megiddo, T. Noma, and A. Yoshise, *A Unified Approach to Interior Point Algorithms for Linear Complementarity Problems,* Lecture Notes in Comput. Sci., 538, Springer-Verlag, New York, 1991.

[27] M. Kojima, M. Mizuno, and T. Noma, *A new continuation method for complementarity problems with uniformed P-functions,* Math. Oper. Res., 43 (1989), pp. 107–113.

[28] Z. Q. Luo and Y. Ye, *A genuine quadratically convergent polynomial interior-point algorithm*

for linear programming, in Advances in Optimization and Approximation, Nonconvex Optim. Appl. 1, D. Du and J. Sun, eds., Kluwer, Dordrecht, 1994, pp. 235–246.

[29] K. G. Murty, *Linear Complementarity, Linear and Nonlinear Programming,* Sigma. Ser. Appl. Math. 3, Heldermann-Verlag, Berlin, 1988.

[30] J. S. Pang, *On a class of least-element complementarity problems,* Math. Program., 16 (1979), pp. 111–126.

[31] J. S. Pang and S. A. Gabriel, *NE/SQP: A robust algorithm for the nonlinear complementarity problem,* Math. Program., 60 (1993), pp. 295–338.

[32] L. Qi, *Convergence analysis of some algorithms for solving nonsmooth equations*, Math. Oper. Res., 18 (1993), pp. 227–243.

[33] L. Qi and D. Sun, *Improving the convergence of non-interior point algorithm for nonlinear complementarity problems,* Math. Comp., 69 (2000), pp. 282–304.

[34] G. Ravindran and M. S. Gowda, *Regularization of P_0-functions in box variational inequality problems,* SIAM J. Optim., 11 (2000), pp. 748–760.

[35] P. Tseng, *Analysis of a non-interior continuation method based on Chen-Mangasarian smoothing functions for complementarity problems,* in Reformulation: Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods, Appl. Optim. 22, M. Fukushima and L. Qi, eds., Kluwer, Dordrecht, 1998, pp. 381–404.

[36] V. Venkateswaran, *An algorithm for the linear complementarity problem with a P_0-matrix,* SIAM J. Matrix Anal. Appl., 14 (1993), pp. 967–977.

[37] L. T. Watson, *Solving the nonlinear complementarity problem by a homotopy method,* SIAM J. Control Optim., 17 (1979), pp. 36–46.

[38] S. Wright, *An infeasible-interior-point algorithm for linear complementarity problems,* Math. Program., 67 (1994), pp. 29–51.

[39] S. Wright and D. Ralph, *Superlinear infeasible-interior-point algorithm for monotone complementarity problems*, Math. Oper. Res., 21 (1996), pp. 815–838.

[40] S. Xu and J. V. Burke, *A polynomial time interior-point path-following algorithm for LCP based on Chen-Harker-Kanzow smoothing techniques,* Math. Program., 86 (1999), pp. 91–103.

[41] Y. Ye, *On homogeneous and self-dual algorithms for LCP*, Math. Program., 76 (1997), pp. 211–222.

[42] Y. B. Zhao and D. Li, *On a new homotopy continuation trajectory for complementarity problems,* Math. Oper. Res., 26 (2001), pp. 119–146.

[43] Y.-B. Zhao and D. Li, *Existence and limiting behavior of a non–interior-point trajectory for complementarity problems without strict feasible condition,* SIAM J. Control Optim., 40 (2001), pp. 898–924.

[44] Y.-B. Zhao and G. Isac, *Properties of a multivalued mapping associated with some nonmonotone complementarity problems,* SIAM J. Control Optim., 39 (2000), pp. 571–593.

SIAM J. OPTIM.
Vol. 13, No. 4, pp. 1222–1244

A FEASIBLE SEQUENTIAL LINEAR EQUATION METHOD FOR INEQUALITY CONSTRAINED OPTIMIZATION*

YU-FEI YANG†, DONG-HUI LI†, AND LIQUN QI‡

Abstract. In this paper, by means of the concept of the working set, which is an estimate of the active set, we propose a feasible sequential linear equation algorithm for solving inequality constrained optimization problems. At each iteration of the proposed algorithm, we first solve one system of linear equations with a coefficient matrix of size $m \times m$ (where m is the number of constraints) to compute the working set; we then solve a subproblem which consists of four reduced systems of linear equations with a common coefficient matrix. Unlike existing QP-free algorithms, the subproblem is concerned with only the constraints corresponding to the working set. The constraints not in the working set are neglected. Consequently, the dimension of each subproblem is not of full dimension. Without assuming the isolatedness of the stationary points, we prove that every accumulation point of the sequence generated by the proposed algorithm is a KKT point of the problem. Moreover, after finitely many iterations, the working set becomes independent of the iterates and is essentially the same as the active set of the KKT point. In other words, after finitely many steps, only those constraints which are active at the solution will be involved in the subproblem. Under some additional conditions, we show that the convergence rate is two-step superlinear or even Q-superlinear. We also report some preliminary numerical experiments to show that the proposed algorithm is practicable and effective for the test problems.

Key words. sequential linear equation algorithm, optimization, active set strategy, global convergence, superlinear convergence

AMS subject classifications. 90C30, 65K10

PII. S1052623401383881

1. Introduction.

We consider the nonlinear inequality constrained optimization problem

$$\text{(P)} \qquad \begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & g(x) \le 0, \end{aligned}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^m$ are assumed to be twice continuously differentiable. We denote by

$$\mathcal{F} = \{x \in \mathbb{R}^n \mid g(x) \le 0\}$$

the feasible set of problem (P).

The Lagrangian function associated with problem (P) is defined by

$$L(x, \lambda) = f(x) + \lambda^T g(x).$$

A pair $(x^*, \lambda^*) \in \mathbb{R}^{n \times m}$ is called a KKT point or a KKT pair of problem (P) if it satisfies the following KKT conditions:

$$\begin{aligned} & \nabla_x L(x^*, \lambda^*) = 0, \quad g(x^*) \le 0, \quad \lambda^* \ge 0, \\ & g_i(x^*)\lambda_i^* = 0 \quad \forall\, i \in I, \end{aligned} \tag{1.1}$$

*Received by the editors January 22, 2001; accepted for publication (in revised form) November 6, 2002; published electronically May 2, 2003. This work was supported by the NSF (grant 10171030) of China, the RGC (grant PolyU/5314/01p) of Hong Kong, and the Australian Research Council.
http://www.siam.org/journals/siopt/13-4/38388.html

†Institute of Applied Mathematics, Hunan University, Changsha 410082, China. Current address: Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong (mayfyang@polyu.edu.hk, madhli@polyu.edu.hk).

‡Corresponding author. Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong (maqilq@polyu.edu.hk).

where $I := \{1, \ldots, m\}$ and

$$\nabla_x L(x, \lambda) := \nabla f(x) + \sum_{i=1}^{m} \lambda_i \nabla g_i(x). \tag{1.2}$$

Sometimes, we also call the point x^* satisfying (1.1) a KKT point of problem (P). If (x^*, λ^*) satisfies all conditions in (1.1) except for the inequality $\lambda^* \geq 0$, we call the point x^* a stationary point of problem (P).

Throughout the paper, we assume that the following blanket hypotheses hold.

Assumption A1. The set $\mathcal{F}$ is bounded.

Assumption A2. At every $x \in \mathcal{F}$, the vectors $\nabla g_i(x)$, $i \in I_0(x)$, are linearly independent, where $I_0(x) := \{i \in I \,|\, g_i(x) = 0\}$.

Note that Assumption A1 is often substituted by the assumption that the level sets of the objective function of some unconstrained optimization problem are compact or the sequence of points generated by the algorithm is bounded, while Assumption A2 is a common assumption in dealing with the global convergence of most algorithms for solving problem (P).

The sequential quadratic programming (SQP) methods are a class of efficient methods for solving nonlinearly constrained optimization problems. They have received much attention in recent decades. We refer to a review paper [2] for a good survey on SQP methods.

The iterative process of a typical SQP method is as follows. Let the current iterate be x^k. Compute a search direction d^k by solving the following quadratic program (QP):

$$\begin{array}{ll} \min\limits_{d} & \dfrac{1}{2}\langle d, H_k d\rangle + \langle \nabla f(x^k), d\rangle, \\ \text{s.t.} & g_i(x^k) + \langle \nabla g_i(x^k), d\rangle \leq 0 \quad \forall\, i \in I, \end{array} \tag{1.3}$$

where $H_k \in \mathbb{R}^{n\times n}$ is symmetric positive definite. Perform a line search to determine a steplength t_k and let the next iterate be $x^{k+1} = x^k + t_k d^k$.

SQP methods possess global and superlinear convergence properties under certain conditions. However, in a traditional SQP method, the QP subproblem (1.3) may be inconsistent; that is, the feasible set of (1.3) may be empty. To overcome this shortcoming, various techniques have been proposed; see, e.g., [6, 18, 21, 24, 25, 29, 31]. In particular, Panier and Tits [21] presented a feasible SQP (FSQP) algorithm in which the generated iterates lie in the feasible region $\mathcal{F}$. Under certain conditions, this FSQP algorithm is globally convergent and locally two-step superlinearly convergent. Further study on FSQP algorithms can be found in [17, 22, 27, 28].

FSQP methods are particularly useful for solving those problems arising from engineering design where the objective function f might be undefined outside the feasible region $\mathcal{F}$. Another advantage of FSQP methods is that the objective function f can be used as a merit function to avoid the use of a penalty function. However, FSQP algorithms still require solving QP subproblems at each iteration, which is computationally expensive. In [23], Panier, Tits, and Herskovits proposed a feasible QP-free algorithm in which, at every iteration, only three systems of linear equations need to be solved. Specifically, the iterative process of the QP-free algorithm is as follows. Let (x^k, λ^k) be the current iterate. To guarantee the feasibility of the next iterate, they first solve two systems of linear equations of the form

$$\begin{pmatrix} H_k & \nabla g(x^k) \\ \operatorname{diag}(\mu^k)\nabla g(x^k)^T & \operatorname{diag}(g(x^k)) \end{pmatrix} \begin{pmatrix} d \\ \lambda \end{pmatrix} = \begin{pmatrix} -\nabla f(x^k) \\ c \end{pmatrix} \tag{1.4}$$

by choosing a different vector c, where $H_k \in \mathbb{R}^{n\times n}$ is positive definite, $\mu^k \in \mathbb{R}^m$, $c \in \mathbb{R}^m$, and $\mathrm{diag}(\mu^k)$ denotes the $m \times m$ diagonal matrix whose ith diagonal element is μ_i^k. Then they further "bend" the primal search direction by solving a least squares subproblem to avoid the Maratos effect. It has been shown in [23] that under appropriate conditions, this QP-free method possesses global convergence as well as a locally two-step superlinear convergence rate. However, the QP-free algorithm proposed in [23] may have instability problems. The linear system (1.4) may become very ill-conditioned if some multiplier μ_i corresponding to a nearly active constraint g_i becomes very small. In addition, in the global convergence theorem, there is a restrictive condition which requires that the number of stationary points is finite. The idea of this QP-free algorithm has been further used by Urban, Tits, and Lawrence [34] to develop a primal-dual logarithmic barrier interior-point method; see also [1]. Under similar conditions, the method possesses global and fast local convergence properties.

Recently, by means of the Fischer–Burmeister function, Qi and Qi [26] presented a new feasible QP-free algorithm for solving problem (P). At each iteration, the subproblem of the new QP-free method consists of three systems of linear equations of the form

$$\begin{pmatrix} H_k & \nabla g(x^k) \\ \mathrm{diag}(\eta^k)\nabla g(x^k)^T & -\sqrt{2}\mathrm{diag}(\theta^k) \end{pmatrix} \begin{pmatrix} d \\ \lambda \end{pmatrix} = \begin{pmatrix} -\nabla f(x^k) \\ c, \end{pmatrix}, \tag{1.5}$$

where c is a suitable vector and for each $i \in I$

$$\eta_i^k := \frac{g_i(x^k)}{\sqrt{g_i^2(x^k) + (\mu_i^k)^2}} + 1 \quad \text{and} \quad \theta_i^k := \left(1 - \frac{\mu_i^k}{\sqrt{g_i^2(x^k) + (\mu_i^k)^2}}\right)^{1/2}.$$

To avoid the Maratos effect, they also solve a least squares subproblem. Their algorithm shares some advantages of the method in [23]. Moreover, the matrix in (1.5) is nonsingular even if the strict complementarity does not hold. The method achieves global convergence without requiring the isolatedness of the stationary points. The local one-step superlinear convergence rate of the method has also been established.

In this paper, we propose a feasible sequential linear equation (FSLE) algorithm for solving problem (P). At each step, we first solve three reduced systems of linear equations with the following form:

$$\begin{pmatrix} H_k & \nabla g_{A^k}(x^k) \\ \nabla g_{A^k}(x^k)^T & 0 \end{pmatrix} \begin{pmatrix} d \\ \lambda_{A^k} \end{pmatrix} = \begin{pmatrix} -\nabla f(x^k) \\ c_{A^k} \end{pmatrix}, \tag{1.6}$$

where $A^k \subset I$ is called a *working set* which is an estimate of the active set $I_0(x^k)$. The calculation of the working set depends on some multiplier function which is the solution of a system of linear equations. If x^k is sufficiently close to a KKT point x^*, then A^k is an identification of the active set $I_0(x^*)$. The working set and the identification of the active set have been studied by some authors [9, 10, 12, 13, 31, 32]. They are also very important in our algorithm. It is clear that the dimension of system (1.6) is no more than the dimension of system (1.5). Moreover, as we shall show in section 4 (see Lemma 4.1), under appropriate conditions we have $A^k = I_0(x^*)$ for all k sufficiently large. This means that after finitely many iterations, the inactive constraints at x^* will be neglected.

Like other existing feasible QP-free methods, the method proposed in this paper also generates a sequence of iterates that are interior points in the feasible region.

However, feasible QP-free methods are different from interior-point methods. An interior-point method follows a central path, while a feasible QP-free method does not.

In order to achieve a superlinear convergence rate, we solve another system of linear equations. This system is equivalent to a least squares problem. Unlike algorithms proposed in [23, 26], the coefficient matrix of the last linear system is the same as the previous reduced ones. Furthermore, our algorithm provides a special technique to update the working set and makes it possible to remove multiple inactive constraints in one iteration. This technique for updating the working set has also been used recently in [32].

The main advantage of the proposed algorithm lies in that it has the potential of saving computational cost. Moreover, it reserves all the advantages of algorithms proposed in [23, 26].

Interesting features of the proposed algorithm include the following:

- All iterates are feasible and the sequence of objective functions is decreasing.
- At each iteration, we need to solve only one $m \times m$ system of linear equations and four reduced systems of linear equations with a common coefficient matrix.
- Under appropriate conditions, the generated direction sequences are uniformly bounded.
- The iterative matrices are nonsingular without the requirement of strict complementarity.
- Every accumulation point of the sequence generated by the proposed algorithm is a KKT point of problem (P) without assuming that the stationary points are isolated.
- Locally two-step superlinear or Q-superlinear convergence rate is achieved.

Recently, Facchinei and Lazzari [11] presented a local feasible QP-free algorithm for solving problem (P) with an SC^1 objective function. Their algorithm possesses some favorable properties, such as fast local convergence and feasibility of all iterates. In addition, at each iteration, only systems of linear equations need to be solved. Their algorithm produces a sequence $\{x^k\}$ according to the following formula:

$$x^{k+1} = x^k + d^k + \hat{d}^k.$$

The local structure of our algorithm is similar to theirs. In some sense, our algorithm can be regarded as a globalization of their algorithm. However, compared with their algorithm, we used quasi-Newton algorithms. Moreover, the computation of the directions d^k and $\hat{d}^k$ is different from that in [11].

The paper is organized as follows. In the next section we introduce a multiplier function to define the working set. We then describe the algorithm and show that it is well defined. In section 3, we establish a global convergence theorem for the algorithm. In section 4, we prove that under appropriate conditions the sequence $\{x^k\}$ generated by the proposed algorithm is locally two-step superlinearly or Q-superlinearly convergent. We report some preliminary numerical results in section 5. In the last section, we give some remarks to conclude the paper.

A few words for the notation. The symbol $\|\cdot\|$ always stands for the Euclidean vector norm or its associated matrix norm. Given $h : \mathbb{R}^n \to \mathbb{R}^m$ and a subset A of I, we denote by $h_A(x)$ the subvector of $h(x)$ with components $h_i(x), i \in A$, and by $\nabla h_A(x)$ the transpose of the Jacobian of $h_A(x)$. We use $e \in \mathbb{R}^m$ to denote the vector of all ones, and $E \in \mathbb{R}^{m\times m}$ is the unit matrix.

2. Algorithm. In this section we first define the working set based on a multiplier function; then we present an FSLE algorithm for solving problem (P) and show that it is well defined.

The following proposition comes from [14] and [19].

PROPOSITION 2.1. *The following statements hold.*

(i) *For every $x \in \mathcal{F}$, there exists a unique minimizer $\lambda(x)$ of the quadratic function in λ,*

$$\|\nabla_x L(x,\lambda)\|^2 + \|G(x)\lambda\|^2$$

over $\mathbb{R}^m$, given by

$$\lambda(x) = -M^{-1}(x)\nabla g(x)^T \nabla f(x), \tag{2.1}$$

where

$$G(x) := \operatorname{diag}(g_i(x)) \quad \textit{and} \quad M(x) := \nabla g(x)^T \nabla g(x) + G^2(x).$$

(ii) *The multiplier function $\lambda(x)$ is continuously differentiable in $\mathcal{F}$.*

(iii) *If $(x^*, \lambda^*) \in \mathbb{R}^n \times \mathbb{R}^m$ is a KKT pair for problem* (P)*, we have $\lambda(x^*) = \lambda^*$.*

For $x \in \mathcal{F}$, we now make the following "guess" for the active set $I_0(x)$:

$$A(x;\epsilon) := \{i \mid g_i(x) + \epsilon\rho(x,\lambda(x)) \geq 0\},$$

where ϵ is a nonnegative parameter and $\rho(x,\lambda) := \sqrt{\|\Phi(x,\lambda)\|}$ with

$$\Phi(x,\lambda) := \begin{pmatrix} \nabla_x L(x,\lambda) \\ \min\{-g(x), \lambda\} \end{pmatrix}.$$

It is obvious that (x^*, λ^*) is a KKT pair of problem (P) if and only if $\Phi(x^*, \lambda^*) = 0$ or $\rho(x^*, \lambda^*) = 0$. Facchinei, Fischer, and Kanzow [9] showed that if the second order sufficient condition and the Mangasarian–Fromovotz constraint qualification hold, then for any $\epsilon > 0$, when x is sufficiently close to x^*, the working set $A(x;\epsilon)$ is an exact identification of $I_0(x^*)$. It is not difficult to see from Assumption A1 and Proposition 2.1(ii) that $\rho(x,\lambda(x))$ is bounded on $\mathcal{F}$. This property will enable us to keep the parameter ϵ fixed after a finite number of iterations in our algorithm. Details will be given subsequently.

Let

$$V(x,H;A) = \begin{pmatrix} H & \nabla g_A(x) \\ \nabla g_A(x)^T & 0 \end{pmatrix},$$

where H is an $n \times n$ positive definite matrix and A is a subset of I. We now state the steps of our algorithm for solving problem (P).

ALGORITHM 2.1.

Parameters. $\beta \in (0,1)$, $\mu \in (0,1/2)$, $\nu > 2$, $\tau \in (2,3)$, $\vartheta \in (0,1)$, and $\sigma \in (0,1)$.

Data. x^1, a strictly feasible point in $\mathcal{F}$; $H_1 \in \mathbb{R}^{n\times n}$, a symmetric positive definite matrix; and $\epsilon^0 > 0$, an initial parameter.

Set $k := 1$.

Step 1. *Set $\epsilon := \epsilon^{k-1}$.*

Step 2. *Set $A^k(\epsilon) := A(x^k;\epsilon)$.*

If $\nabla g_{A^k(\epsilon)}(x^k)$ is not of full rank, then set $\epsilon := \sigma\epsilon$ and go to Step 2.

Step 3. *Set* $\epsilon^k := \epsilon$, $A^k := A^k(\epsilon^k)$, *and* $V_k := V(x^k, H_k; A^k)$.

Step 4. *Computation of a search direction.*

(i) *Compute* $(d^{k0}, z^{k0}_{A^k})$ *by solving the system of linear equations in* (d, z_{A^k}),

$$V_k \begin{pmatrix} d \\ z_{A^k} \end{pmatrix} = \begin{pmatrix} -\nabla f(x^k) \\ 0 \end{pmatrix}. \tag{2.2}$$

(ii) *Compute* $(d^{k1}, z^{k1}_{A^k})$ *by solving the system of linear equations in* (d, z_{A^k}),

$$V_k \begin{pmatrix} d \\ z_{A^k} \end{pmatrix} = \begin{pmatrix} -\nabla f(x^k) \\ \varphi^k \end{pmatrix}, \tag{2.3}$$

where $\varphi^k \in \mathbb{R}^{|A^k|}$ *is defined by*

$$\varphi^k_i := \begin{cases} z^{k0}_i & \text{if } \ z^{k0}_i < 0, \\ -g_i(x^k) & \text{if } \ z^{k0}_i > 0, \\ 0 & \text{otherwise.} \end{cases}$$

If $d^{k1} = 0$, *stop.*

(iii) *Compute* $(d^{k2}, z^{k2}_{A^k})$ *by solving the system of linear equations in* (d, z_{A^k}),

$$V_k \begin{pmatrix} d \\ z_{A^k} \end{pmatrix} = \begin{pmatrix} -\nabla f(x^k) \\ \varphi^k - \|d^{k1}\|^\nu e_{A^k} \end{pmatrix}. \tag{2.4}$$

(iv) *Compute the search direction* d^k *and the approximate multiplier vector* $z^k_{A^k}$ *according to*

$$\begin{pmatrix} d^k \\ z^k_{A^k} \end{pmatrix} := (1 - \phi^k) \begin{pmatrix} d^{k1} \\ z^{k1}_{A^k} \end{pmatrix} + \phi^k \begin{pmatrix} d^{k2} \\ z^{k2}_{A^k} \end{pmatrix},$$

where

$$\phi^k := (\vartheta - 1) \frac{\langle \nabla f(x^k), d^{k1} \rangle}{1 + \|d^{k1}\|^\nu |\sum_{i \in A^k} z^{k0}_i|}.$$

Step 5. *Compute a correction* $\hat{d}^k$ *by solving the system of linear equations in* (d, z_{A^k}),

$$V_k \begin{pmatrix} d \\ z_{A^k} \end{pmatrix} = \begin{pmatrix} 0 \\ -\|d^k\|^\tau e_{A^k} - g_{A^k}(x^k + d^k) \end{pmatrix}. \tag{2.5}$$

If $\|\hat{d}^k\| > \|d^k\|$, *set* $\hat{d}^k := 0$.

Step 6. *Line search. Compute* t_k, *the first number* t *in the sequence* $\{1, \beta, \beta^2, \ldots\}$ *satisfying*

$$f(x^k + td^k + t^2\hat{d}^k) \le f(x^k) + \mu t \langle \nabla f(x^k), d^k \rangle \tag{2.6}$$

and

$$g_i(x^k + td^k + t^2\hat{d}^k) < 0 \qquad \forall\, i \in I. \tag{2.7}$$

Step 7. *Set* $x^{k+1} := x^k + t_k d^k + t_k^2 \hat{d}^k$ *and generate a new symmetric definite positive matrix* H_{k+1}. *Set* $k := k + 1$ *and go to Step* 1.

Remarks.

(i) It follows from Assumption A2 that there exists some $\delta_0 > 0$ such that $\nabla g_{I(x;\delta)}(x)$ is of full rank, where $I(x;\delta) := \{i \in I : g_i(x) \geq -\delta\}$ and $0 \leq \delta \leq \delta_0$. By the continuity of $\lambda(x)$ and Assumption A1, there exists some $\bar{\epsilon}_0 > 0$ such that the inequality $\epsilon\rho(x, \lambda(x)) \leq \delta_0$ holds for all $\epsilon \leq \bar{\epsilon}_0$ and $x \in \mathcal{F}$, and hence $A(x;\epsilon) \subseteq I(x;\delta_0)$. This implies that $\nabla g_{A(x;\epsilon)}(x)$ is of full rank. Therefore, for symmetric positive definite matrix $H \in \mathbb{R}^{n\times n}$, the matrix $V(x, H; A(x;\epsilon))$ is nonsingular. Consequently, V_k is nonsingular for each k. This shows that $(d^{k0}, z^{k0}_{A^k})$, $(d^{k1}, z^{k1}_{A^k})$, $(d^{k2}, z^{k2}_{A^k})$, and $\hat{d}^k$ are well defined.

On the other hand, the above analysis also indicates that at Step 2 of Algorithm 2.1 the parameter ϵ is reduced only finitely many times. In other words, ϵ_k will remain fixed after finitely many iterations. Without loss of generality, we assume that $\epsilon^k = \tilde{\epsilon}$ for all k.

(ii) In order to guarantee the feasibility of all iterates and the decrease of the objective function at each iteration, we solve three linear systems with the same coefficient matrix but different right vectors. This technique is similar to that in [26]. Notice that the choice of φ^k at Step 4(ii) ensures that x^k is a trivial KKT point of problem (P) whenever $d^{k1} = 0$ (see Lemma 2.2).

(iii) The role of Step 5 is to avoid the Maratos effect. It is not difficult to see that $\hat{d}^k$ is also the unique solution of the least squares problem in d,

$$\begin{aligned} \min \quad & \frac{1}{2}\|d\|^2_{H_k} \\ \text{s.t.} \quad & g_i(x^k + d^k) + \langle \nabla g_i(x^k), d\rangle = -\|d^k\|^\tau \qquad \forall\, i \in A^k. \end{aligned} \tag{2.8}$$

An important difference between our algorithm and those in [23, 26] lies in the fact that the coefficient matrix in (2.5) is the same as that in Step 4. Hence, our algorithm needs fewer computational efforts. If H_k is taken to be the unit matrix for every k, $A^k = I_0(x^*)$, and $\tau = 2$, then problem (2.8) reduces to the subproblem of computing the correction direction $\hat{d}^k$ in [11].

(iv) It is not difficult to deduce that the direction $(d^k, z^k_{A^k})$ is the unique solution of the following system of linear equations:

$$V_k \begin{pmatrix} d \\ z_{A^k} \end{pmatrix} = \begin{pmatrix} -\nabla f(x^k) \\ \varphi^k - \phi^k \|d^{k1}\|^\nu e_{A^k} \end{pmatrix}. \tag{2.9}$$

We now analyze the updating technique for the working set. For $i \in A^k$, we obtain from (2.5) and (2.9)

$$\begin{aligned} & g_i(x^k + td^k + t^2\hat{d}^k) \\ &= g_i(x^k) + t\nabla g_i(x^k)^T d^k + t^2 \nabla g_i(x^k)^T \hat{d}^k + O((t\|d^k\|)^2) \\ &= g_i(x^k) + t\nabla g_i(x^k)^T d^k - t^2 g_i(x^k + d^k) + O((t\|d^k\|)^2) \\ &= (1-t^2) g_i(x^k) + (t - t^2)\nabla g_i(x^k)^T d^k + O((t\|d^k\|)^2) \\ &= O((t\|d^k\|)^2) - (t - t^2)\phi^k \|d^{k1}\|^\nu \\ &\quad + \begin{cases} (1-t^2) g_i(x^k) + (t-t^2) z^{k0}_i & \text{if } z^{k0}_i < 0, \\ (1-t) g_i(x^k) & \text{if } z^{k0}_i > 0, \\ (1-t^2) g_i(x^k) & \text{otherwise.} \end{cases} \end{aligned}$$

Hence, if $z^{k0}_i < 0$ is not small and $t_k - t_k^2$ is not very small, it is likely that $i \notin A^{k+1}$ because g_i becomes strongly negative now. Thus, it is reasonable to exclude these i from A^{k+1}. This technique was also used by Spellucci [32].

For the sake of convenience, we let for each k

$$z_i^{k0} = z_i^{k1} = z_i^{k2} = z_i^k = 0 \qquad \forall\, i \notin A^k.$$

To analyze the well-definedness and convergence of the above algorithm, we make the following hypothesis on the choice of matrix H_k.

Assumption A3. There exist positive constants C_1 and C_2 such that, for all k and $d \in \mathbb{R}^n$,

$$C_1\|d\|^2 \le d^T H_k d \le C_2\|d\|^2.$$

It is not difficult to see from the discussion of Remark (i) that every limit of the sequence $\{\nabla g_{A^k}(x^k)\}$ is also of full rank. Therefore, Assumption A3 shows that every limit of the sequence $\{V_k\}$ is nonsingular, which implies that $\{\|V_k^{-1}\|\}$ is bounded. We assume that $\|V_k^{-1}\| \le \tilde{M}$ for all k.

Let $N_{A^k} := \nabla g_{A^k}(x^k)$. Then, by Step 2, N_{A^k} is of full rank. Since V_k is nonsingular, it is clear that matrix $D_k := N_{A^k}^T H_k^{-1} N_{A^k}$ is also nonsingular. Let

$$B_k := H_k^{-1} N_{A^k} D_k^{-1} \quad \text{and} \quad Q_k := H_k^{-1}(E - N_{A^k} B_k^T).$$

By Step 4 of the algorithm, it is not difficult to deduce the following relations:

$$\left\{\begin{array}{ll} d^{k0} = -Q_k \nabla f(x^k), & z_{A^k}^{k0} = -B_k^T \nabla f(x^k), \\ d^{k1} = d^{k0} + B_k \varphi^k, & z_{A^k}^{k1} = z_{A^k}^{k0} - D_k^{-1}\varphi^k, \\ d^{k2} = d^{k1} - \|d^{k1}\|^\nu B_k e_{A^k}, & z_{A^k}^{k2} = z_{A^k}^{k1} + \|d^{k1}\|^\nu D_k^{-1} e_{A^k}, \\ d^k = d^{k1} - \phi^k \|d^{k1}\|^\nu B_k e_{A^k}, & z_{A^k}^{k} = z_{A^k}^{k1} + \phi^k \|d^{k1}\|^\nu D_k^{-1} e_{A^k}. \end{array}\right. \tag{2.10}$$

LEMMA 2.2. *If the algorithm stops at Step* 4(ii)*, i.e.,* $d^{k1} = 0$*, then* $\nabla f(x^k) = 0$.

Proof. If $d^{k1} = 0$, then it follows from (2.3) that

$$\left\{\begin{array}{l} \nabla f(x^k) + \nabla g_{A^k}(x^k) z_{A^k}^{k1} = 0, \\ \varphi^k = 0. \end{array}\right. \tag{2.11}$$

By the construction of φ^k, we have $z_{A^k}^{k0} = 0$, and hence by (2.10) $z_{A^k}^{k1} = 0$. The assertion then follows from the first equation of (2.11). ∎

The above lemma shows that if the algorithm stops at Step 4(ii), then x^k is an unconstrained stationary point of f. Since we always have $x^k \in \mathcal{F}$, this means that x^k is actually a KKT point of problem (P). In what follows, we assume that the algorithm never stops at Step 4(ii). Therefore, the algorithm generates an infinite sequence $\{x^k\}$.

LEMMA 2.3. (i) $\langle \nabla f(x^k), d^{k0}\rangle = -\langle d^{k0}, H_k d^{k0}\rangle$.

(ii) $\langle \nabla f(x^k), d^{k1}\rangle = \langle \nabla f(x^k), d^{k0}\rangle - \langle \varphi^k, z_{A^k}^{k0}\rangle \le \langle \nabla f(x^k), d^{k0}\rangle$.

(iii) $\langle \nabla f(x^k), d^k\rangle \le \vartheta \langle \nabla f(x^k), d^{k1}\rangle$.

Proof. By (2.2), we deduce

$$\begin{aligned} \langle \nabla f(x^k), d^{k0}\rangle &= -\langle d^{k0}, H_k d^{k0}\rangle - \langle d^{k0}, \nabla g_{A^k}(x^k) z_{A^k}^{k0}\rangle \\ &= -\langle d^{k0}, H_k d^{k0}\rangle - \langle \nabla g_{A^k}(x^k)^T d^{k0}, z_{A^k}^{k0}\rangle \\ &= -\langle d^{k0}, H_k d^{k0}\rangle. \end{aligned}$$

This establishes (i). From (2.10), we have

$$\begin{aligned}\langle \nabla f(x^k), d^{k1}\rangle &= \langle \nabla f(x^k), d^{k0}\rangle + \langle \nabla f(x^k), B_k\varphi^k\rangle \\ &= \langle \nabla f(x^k), d^{k0}\rangle + \langle B_k^T \nabla f(x^k), \varphi^k\rangle \\ &= \langle \nabla f(x^k), d^{k0}\rangle - \langle z^{k0}_{A^k}, \varphi^k\rangle \\ &\le \langle \nabla f(x^k), d^{k0}\rangle,\end{aligned}$$

where the last inequality holds because by the definition of φ^k we have $\langle z^{k0}_{A^k}, \varphi^k\rangle \ge 0$. This establishes (ii). We now turn to verify (iii).

It follows from (2.10) and the definition of ϕ^k that

$$\begin{aligned}\langle \nabla f(x^k), d^k\rangle &= \langle \nabla f(x^k), d^{k1}\rangle - \phi^k \|d^{k1}\|^\nu \langle \nabla f(x^k), B_k e_{A^k}\rangle \\ &= \langle \nabla f(x^k), d^{k1}\rangle - \phi^k \|d^{k1}\|^\nu \langle B_k^T \nabla f(x^k), e_{A^k}\rangle \\ &= \langle \nabla f(x^k), d^{k1}\rangle + \phi^k \|d^{k1}\|^\nu \sum_{i\in A^k} z_i^{k0} \\ &\le \vartheta \langle \nabla f(x^k), d^{k1}\rangle.\end{aligned}$$

This establishes (iii). □

Lemma 2.3 shows that the direction d^k is a descent direction of the merit function f. Similar to the proof of Proposition 3.3 in [23], we can deduce that for each k there is a nonnegative integer $j(k)$ such that inequalities (2.6) and (2.7) are satisfied with $t_k = \beta^{j(k)}$.

The above discussion has shown that Algorithm 2.1 is well defined.

3. Global convergence. In this section we will show that Algorithm 2.1 is globally convergent. First, we see from Step 2 of Algorithm 2.1 and the discussion after Assumption A3 that $\|V_k^{-1}\| \le \tilde{M}$ for all k. The following lemma is then obvious.

LEMMA 3.1. *The sequences* $\{(d^{k0}, z^{k0})\}$, $\{(d^{k1}, z^{k1})\}$, *and* $\{(d^{k2}, z^{k2})\}$ *are all bounded.*

Proof. By (2.2), Assumption A1, and the boundedness of $\{\|V_k^{-1}\|\}$, we deduce that $\{(d^{k0}, z^{k0})\}$ is bounded, which implies that $\{(d^{k1}, z^{k1})\}$ is also bounded by (2.3). The boundedness of $\{(d^{k2}, z^{k2})\}$ directly follows from (2.4) and the boundedness of $\{d^{k1}\}$ and $\{z^{k0}\}$. □

LEMMA 3.2. *There exists a constant* $\kappa > 0$ *such that, for all* $k = 1, 2, \ldots,$

$$\|d^k - d^{k1}\| \le \kappa \|d^{k1}\|^\nu.$$

Proof. Assumption A1 and Lemma 3.1 imply that $\{\phi^k\}$ is bounded. It follows from Step 4 of Algorithm 2.1 that

$$\begin{pmatrix} d^k - d^{k1} \\ z^k_{A^k} - z^{k1}_{A^k} \end{pmatrix} = V_k^{-1} \begin{pmatrix} 0 \\ -\phi^k \|d^{k1}\|^\nu e_{A^k} \end{pmatrix},$$

which shows that the assertion holds with $\kappa := \tilde{M} \sup\{\phi^k\}$. □

The following proposition gives a sufficient condition for the global convergence of Algorithm 2.1.

PROPOSITION 3.3. *Let* x^* *be an accumulation point of the sequence* $\{x^k\}$ *generated by Algorithm* 2.1 *and suppose that* $\{x^k\}_{K_0} \to x^*$. *If*

$$\{\langle \nabla f(x^k), d^{k1}\rangle\}_{K_0} \to 0, \tag{3.1}$$

then x^* is a KKT point of problem (P) *and $\{z^{k0}\}_{K_0}$ converges to the unique multiplier vector λ^* associated with x^*.*

Proof. It follows from Assumption A3, Lemma 2.3, and (3.1) that

$$\{d^{k0}\}_{K_0} \to 0 \quad \text{and} \quad \{\langle \varphi^k, z^{k0}_{A^k}\rangle\}_{K_0} \to 0. \tag{3.2}$$

Let z^* be an arbitrary accumulation point of $\{z^{k0}\}_{K_0}$, and let $\{z^{k0}\}_{K_1}$ be a subsequence of $\{z^{k0}\}_{K_0}$ such that $\{z^{k0}\}_{K_1} \to z^*$. The boundedness of $\{z^{k0}\}$ implies that z^* exists. From (2.2), (3.2), and the definition of φ^k, we deduce

$$\begin{cases} \nabla f(x^*) + \nabla g(x^*) z^* = 0, \\ z_i^* \geq 0, \quad z_i^* g_i(x^*) = 0 \ \ \forall\, i \in I. \end{cases}$$

It is also obvious that $g(x^*) \leq 0$. Thus, x^* is a KKT point of problem (P) and z^* is its associated multiplier vector (i.e., $z^* = \lambda^*$). The uniqueness of the multiplier vector implies that $\{z^{k0}\}_{K_0} \to \lambda^*$. ∎

By Proposition 3.3, we establish a global convergence theorem for Algorithm 2.1.

THEOREM 3.4. *If (x^*, λ^*) is an accumulation point of the sequence $\{(x^k, z^{k0})\}$ generated by Algorithm* 2.1*, then (x^*, λ^*) is a KKT pair of problem* (P).

Proof. We prove the theorem by contradiction. Suppose that there is a subsequence $\{(x^k, z^{k0})\}_K$ converging to (x^*, λ^*), but (x^*, λ^*) is not a KKT pair of problem (P). We first prove that there must be a subset K_0 of K such that (3.1) holds. Otherwise, there exist $\gamma > 0$ and $\underline{d} > 0$ such that

$$\langle \nabla f(x^k), d^{k1}\rangle \leq -\gamma \ \ \forall\, k \in K \qquad \text{and} \qquad \liminf_{k \in K} \|d^{k1}\| \geq \underline{d}. \tag{3.3}$$

By the definition of ϕ^k, Lemma 3.1, and (3.3), it follows that there exists $\tilde{\phi} > 0$ such that

$$\phi^k \geq \tilde{\phi} \qquad \forall \ k \in K.$$

In a way similar to the proof of Lemma 3.9 in [23], we deduce

$$\begin{aligned} & f(x^k + td^k + t^2\hat{d}^k) - f(x^k) - \mu t \langle \nabla f(x^k), d^k\rangle \\ & \leq t\Big\{ \sup_{\xi \in [0,1]} \|\nabla f(x^k + t\xi d^k + t^2\xi \hat{d}^k) - \nabla f(x^k)\| \|d^k\| \\ & \qquad + 2t \sup_{\xi \in [0,1]} \|\nabla f(x^k + t\xi d^k + t^2\xi \hat{d}^k)\| \|\hat{d}^k\| - (1-\mu)\vartheta C_1 \underline{d}^2 \Big\}, \end{aligned} \tag{3.4}$$

where C_1 and $\underline{d}$ are specified by Assumption A3 and (3.3), respectively. We also have for each $i \in I$

$$g_i(x^k + td^k + t^2\hat{d}^k) \leq g_i(x^k) + t\{u_i^k(t) + \langle \nabla g_i(x^k), d^k\rangle\} \tag{3.5}$$

with

$$\begin{aligned} u_i^k(t) := & \sup_{\xi \in [0,1]} \|\nabla g_i(x^k + t\xi d^k + t^2\xi \hat{d}^k) - \nabla g_i(x^k)\| \|d^k\| \\ & + 2t \sup_{\xi \in [0,1]} \|\nabla g_i(x^k + t\xi d^k + t^2\xi \hat{d}^k)\| \|\hat{d}^k\|. \end{aligned}$$

Hence, by (2.9), (3.3), (3.5), and the definition of φ^k, we have, for $i \in A^k$,

$$
\begin{aligned}
& g_i(x^k + td^k + t^2\hat{d}^k) \\
& \le g_i(x^k) + t\{u_i^k(t) + \varphi_i^k - \phi^k \|d^{k1}\|^\nu\} \\
& \le g_i(x^k) + t\{u_i^k(t) + \varphi_i^k - \tilde{\phi}\underline{d}^\nu\} \\
& = \begin{cases} g_i(x^k) + tz_i^{k0} + t\{u_i^k(t) - \tilde{\phi}\underline{d}^\nu\} & \text{if } z_i^{k0} < 0, \\ (1-t)g_i(x^k) + t\{u_i^k(t) - \tilde{\phi}\underline{d}^\nu\} & \text{if } z_i^{k0} > 0, \\ g_i(x^k) + t\{u_i^k(t) - \tilde{\phi}\underline{d}^\nu\} & \text{otherwise} \end{cases} \\
& \le t\{u_i^k(t) - \tilde{\phi}\underline{d}^\nu\}.
\end{aligned} \tag{3.6}
$$

On the other hand, for $i \notin A^k$, $g_i(x^k) < -\tilde{\epsilon}\rho(x^k, \lambda(x^k))$, and hence by (3.5), we get

$$
g_i(x^k + td^k + t^2\hat{d}^k) < -\tilde{\epsilon}\rho(x^k, \lambda(x^k)) + t\{u_i^k(t) + \langle \nabla g_i(x^k), d^k \rangle\}. \tag{3.7}
$$

Since $\rho(x^*, \lambda(x^*)) > 0$, $\{x^k\}_K \to x^*$, $\|\hat{d}^k\| \le \|d^k\|$, and $\{d^k\}$ is bounded, it follows from (3.6) and (3.7) that for all $i \in I$ there exists $\bar{t}_i > 0$, independent of k, such that, for all $t \in [0, \bar{t}_i]$ and $k \in K$ sufficiently large,

$$
g_i(x^k + td^k + t^2\hat{d}^k) < 0.
$$

Moreover, (3.4) implies that there exists $\bar{t}_f > 0$, independent of k, such that, for all $t \in [0, \bar{t}_f]$ and $k \in K$ sufficiently large,

$$
f(x^k + td^k + t^2\hat{d}^k) - f(x^k) - \mu t \langle \nabla f(x^k), d^k \rangle \le 0. \tag{3.8}
$$

Let

$$
\bar{t} := \min\{\bar{t}_f, \bar{t}_1, \ldots, \bar{t}_m\}.
$$

The line search rules (2.6) and (2.7) show that $t_k \ge \beta\bar{t}$ for all $k \in K$ sufficiently large, and hence by Lemma 2.3, (3.3), and (3.8) we deduce

$$
f(x^k + t_k d^k + t_k^2\hat{d}^k) - f(x^k) \le -\mu\beta\bar{t}\vartheta\gamma, \quad k \in K. \tag{3.9}
$$

Since $\{f(x^k)\}$ is monotonically decreasing and bounded below, it converges. Taking limits in (3.9) as $k \to \infty$ with $k \in K$ yields a contradiction. The contradiction shows that (3.1) holds for some $K_0 \subseteq K$. It then follows from Proposition 3.3 that (x^*, λ^*) is a KKT pair of problem (P). The proof is complete. □

4. Superlinear convergence. In this section we analyze the rate of convergence of Algorithm 2.1. Let (x^*, λ^*) be an accumulation point of the sequence $\{(x^k, z^{k0})\}$. Then it follows from Theorem 3.4 that (x^*, λ^*) is a KKT pair of problem (P). For simplicity, we let $I_0 = I_0(x^*)$.

Assumption A4. The strict complementarity condition holds at (x^*, λ^*), i.e., $\lambda^* - g(x^*) > 0$.

Assumption A5. The second order sufficiency condition holds at (x^*, λ^*); i.e., the Hessian $\nabla^2_{xx} L(x^*, \lambda^*)$ is positive definite on the space $\{u \,|\, \langle \nabla g_i(x^*), u \rangle = 0 \text{ for all } i \in I_0\}$.

We first show that under the conditions of Assumptions A1–A3 and A5, the whole sequence $\{x^k\}$ converges to x^* and the sequence $\{z^k\}$ converges to λ^*. Then we prove

that under Assumptions A1–A5, together with Assumption A6′, which will be introduced later in this section, the unit steplength is accepted for all k sufficiently large, and hence the Maratos effect does not occur. Finally, we show that the convergence rate is two-step superlinear or even Q-superlinear.

The following lemma follows from Theorems 2.3 and 3.7 in [9] directly.

LEMMA 4.1. *Let x^* be a KKT point of problem* (P) *and assume that Assumption* A5 *holds. Then there exists a neighborhood of x^* such that, for each x in this neighborhood,*

$$A(x;\tilde{\epsilon}) = I_0.$$

The above lemma indicates that the active constraints can be accurately identified close to a KKT point even if the strict complementarity condition does not hold at that point. To prove that the whole sequence $\{x^k\}$ converges to x^*, we cite another useful result from Proposition 7 in [16]. The original version of this result is due to Moré and Sorensen [20], which is slightly different from this version.

LEMMA 4.2. *Assume that $\omega^* \in \mathbb{R}^t$ is an isolated accumulation point of a sequence $\{\omega^k\} \subset \mathbb{R}^t$ such that, for every subsequence $\{\omega^k\}_K$ converging to ω^*, there is an infinite subset $K' \subseteq K$ such that $\{\|\omega^{k+1} - \omega^k\|\}_{K'} \to 0$. Then the whole sequence $\{\omega^k\}$ converges to ω^*.*

The next proposition claims the convergence of the whole sequence $\{x^k\}$.

PROPOSITION 4.3. *Under Assumptions* A1–A3 *and* A5*, the whole sequence $\{x^k\}$ converges to x^* and the sequence $\{z^{k0}\}$ converges to λ^*.*

Proof. Assumptions A2 and A5 imply that x^* is an isolated accumulation point of $\{x^k\}$ (see [30]). Let $\{x^k\}_K$ be a subsequence converging to x^*. It is clear from Lemma 4.1 that $A^k = I_0$ holds for $k \in K$ sufficiently large. We first prove that there must exist an infinite subset $K' \subseteq K$ such that

$$\{\|d^k\|\}_{K'} \to 0. \tag{4.1}$$

Suppose by contradiction that (4.1) does not hold for any infinite subset of K. Then

$$\liminf_{k\in K} \|d^k\| > 0,$$

which implies by Lemma 3.2 that

$$\liminf_{k\in K} \|d^{k1}\| > 0. \tag{4.2}$$

Without loss of generality, by Lemma 3.1 we assume that

$$\{(d^{k0}, z^{k0})\}_K \to (d^{*0}, z^{*0}) \quad \text{and} \quad \{(d^{k1}, z^{k1})\}_K \to (d^{*1}, z^{*1}).$$

Furthermore, we assume that $\{H_k\}_K \to H_*$. Taking limit in both sides of (2.2) as $k \to \infty$ with $k \in K$, we deduce that $(d^{*0}, z^{*0}_{I_0})$ solves the following system of linear equations:

$$V_* \begin{pmatrix} d \\ z_{I_0} \end{pmatrix} = \begin{pmatrix} -\nabla f(x^*) \\ 0 \end{pmatrix}, \tag{4.3}$$

where $V_* := V(x^*, H_*; I_0)$ is nonsingular. On the other hand, it is easy to see from the KKT system (1.1) that $(0, \lambda^*_{I_0})$ is the solution of system (4.3). So, we have $z^{*0}_{I_0} = \lambda^*_{I_0}$.

It then follows from the definition of φ^k that $\{\varphi^k\}_K \to 0$. Taking limit in (2.3) as $k \to \infty$ with $k \in K$, we see that $(d^{*1}, z^{*1}_{I_0})$ also satisfies (4.3), and hence $d^{*1} = 0$, which contradicts (4.2). This contradiction shows that (4.1) holds for some infinite subset $K' \subseteq K$. Therefore, we get from (4.1)

$$\|x^{k+1} - x^k\| \le \|d^k\| + \|\hat{d}^k\| \le 2\|d^k\| \to 0, \quad \text{as } k \to \infty \text{ with } k \in K'.$$

By means of Lemma 4.2, we claim that the whole sequence $\{x^k\}$ converges to x^*. Moreover, the uniqueness of the multiplier vector λ^* implies that $\{z^{k0}\}$ converges to λ^*. □

The following results are a direct corollary of Proposition 4.3 and will play an important role in the analysis of the convergence rate.

COROLLARY 4.4. *Let Assumptions* A1–A3 *and* A5 *hold. Then the equality* $A^k = I_0$ *holds for all* k *sufficiently large. Furthermore, we have*

(i) $d^k \to 0$, $d^{k0} \to 0$, $d^{k1} \to 0$, *as* $k \to \infty$.

(ii) $z^k \to \lambda^*$, $z^{k1} \to \lambda^*$, *as* $k \to \infty$.

(iii) *If, in addition, Assumption* A4 *holds, then for* k *sufficiently large it holds that* $\varphi^k = -g_{I_0}(x^k)$.

Proof. We have by Lemma 4.1 and Proposition 4.3 that $\{z^{k0}\} \to \lambda^*$ and that $A^k = I_0$ holds for all k sufficiently large. It is also not difficult to see from Proposition 4.3 and the uniqueness of the solution of system (4.3) that the sequences $\{(d^{k0}, z^{k0}_{I_0})\}$, $\{(d^{k1}, z^{k1}_{I_0})\}$, and $\{(d^{k2}, z^{k2}_{I_0})$ converge to the unique solution of system (4.3). This shows (i) and (ii).

If Assumption A4 holds, then we have $z^{k0}_{I_0} > 0$ for all k sufficiently large. This implies (iii). □

System (2.3) and Corollary 4.4(iii) show that for k sufficiently large $(d^{k1}, z^{k1}_{I_0})$ is the unique solution of the following system of linear equations:

$$\begin{cases} H_k d + \nabla g_{I_0}(x^k) z_{I_0} = -\nabla f(x^k), \\ \nabla g_{I_0}(x^k)^T d = -g_{I_0}(x^k). \end{cases} \tag{4.4}$$

This means that d^{k1} produced by (2.3) can be regarded as a quasi-Newton direction for the equality constrained optimization problem

$$\begin{array}{ll} \min & f(x), \\ \text{s.t.} & g_{I_0}(x) = 0. \end{array} \tag{4.5}$$

It is interesting to note that the local algorithm proposed by Facchinei and Lazzari [11] generates a direction d^k which is a Newton direction of (4.5). In other words, d^k generated by the algorithm in [11] is the solution of (4.4) with H_k taken from the generalized Hessian $\partial^2_{xx} L(x^k, \lambda(x^k))$. Our method is slightly different from the method in [11] in that d^k in our method is only an approximate solution of (4.4) because we have $d^k = d^{k1} + O(\|d^{k1}\|^\nu)$ with $\nu > 2$ by Lemma 3.2.

We are going to prove the superlinear convergence of the proposed method. It is well known that the Dennis and Moré condition [7] is necessary and sufficient for superlinear convergence of a quasi-Newton method for solving nonlinear equations or unconstrained optimization problems. Boggs, Tolle, and Wang [3] extended this result to the quasi-Newton method for solving equality constrained optimization problems (see also [33]). We will extend this result to our algorithm.

Assumption A6′. The sequence of matrices $\{H_k\}$ satisfies

$$\frac{\|P_k(H_k - \nabla^2_{xx}L(x^*,\lambda^*))P_k d^k\|}{\|d^k\|} \to 0,$$

where

$$P_k := E - N_k(N_k^T N_k)^{-1}N_k^T \quad \text{and} \quad N_k := \nabla g_{I_0}(x^k).$$

We will show that Assumption A6′ is a sufficient condition for our algorithm to be two-step superlinearly convergent. To this end, we first prove two lemmas.

LEMMA 4.5. *When k is sufficiently large, the direction d^k can be decomposed into*

$$d^k = P_k d^k + \tilde{d}^k$$

with

$$\|\tilde{d}^k\| = O(\|g_{I_0}(x^k)\|) + o(\|d^{k1}\|^2).$$

Proof. It follows from (2.9) and Corollary 4.4(iii) that for k sufficiently large

$$\begin{aligned}\langle \nabla g_i(x^k), d^k\rangle &= \varphi_i^k - \phi^k\|d^{k1}\|^\nu \\ &= -g_i(x^k) - \phi^k\|d^{k1}\|^\nu \qquad \forall\, i \in I_0.\end{aligned}$$

This implies that

$$N_k^T d^k = h^k,$$

where

$$h^k := -g_{I_0}(x^k) - \phi^k\|d^{k1}\|^\nu e_{I_0}.$$

Thus, we have

$$\begin{aligned}d^k &= P_k d^k + N_k(N_k^T N_k)^{-1}N_k^T d^k \\ &= P_k d^k + N_k(N_k^T N_k)^{-1}h^k \\ &= P_k d^k + \tilde{d}^k,\end{aligned}$$

where

$$\tilde{d}^k := N_k(N_k^T N_k)^{-1}h^k$$

satisfies

$$\|\tilde{d}^k\| = O(\|h^k\|) = O(\|g_{I_0}(x^k)\|) + o(\|d^{k1}\|^2). \qquad \square$$

LEMMA 4.6. *When k is sufficiently large, the direction $\hat{d}^k$ is determined by solving system* (2.5)*, and it satisfies*

$$\|\hat{d}^k\| = O(\|d^k\|^2).$$

Proof. It follows from (2.5) and Corollary 4.4 that when k is sufficiently large the direction $\hat{d}^k$ is first computed by solving the following system of linear equations:

$$V_k \begin{pmatrix} d \\ z_{I_0} \end{pmatrix} = \begin{pmatrix} 0 \\ -\|d^k\|^\tau e_{I_0} - g_{I_0}(x^k + d^k) \end{pmatrix} \tag{4.6}$$

with $V_k = V(x^k, H_k; I_0)$.

By Taylor's expansion, we get for each $i \in I_0$

$$\begin{aligned} & -\|d^k\|^\tau - g_i(x^k + d^k) \\ &= -\|d^k\|^\tau - [g_i(x^k) + \langle \nabla g_i(x^k), d^k\rangle + O(\|d^k\|^2)] \\ &= -\|d^k\|^\tau + \phi^k \|d^{k1}\|^\nu + O(\|d^k\|^2) \\ &= O(\|d^k\|^2), \end{aligned}$$

where the second equality follows from (2.9) and Corollary 4.4(iii), and the last equality follows from Lemma 3.2, respectively. The assertion then follows from (4.6) and the fact that $\|V_k^{-1}\| \le \tilde{M}$ for all k. □

We are now in a position to prove that a unit step is eventually accepted by Algorithm 2.1.

PROPOSITION 4.7. *Let Assumptions* A1–A5 *and* A6′ *hold. Then when k is sufficiently large the step $t_k = 1$ is accepted.*

Proof. By the line search rules (2.6) and (2.7), we need only to show that for k sufficiently large the following two conditions hold:

(i) The sufficient decrease condition (2.6) on f holds for $t = 1$.

(ii) The strict feasibility condition (2.7) on g holds for $t = 1$.

It follows from Lemma 4.6 that

$$\begin{aligned} f(x^k + d^k + \hat{d}^k) = f(x^k) + \langle \nabla f(x^k), d^k + \hat{d}^k\rangle \\ + \frac{1}{2}\langle d^k, \nabla^2_{xx} f(x^k) d^k\rangle + o(\|d^k\|^2). \end{aligned} \tag{4.7}$$

In view of (2.5), (2.9), and Corollary 4.4, for k sufficiently large

$$H_k d^k + \nabla f(x^k) + \sum_{i \in I_0} z_i^k \nabla g_i(x^k) = 0, \tag{4.8}$$

$$\langle \nabla g_i(x^k), d^k\rangle = -g_i(x^k) - \phi^k \|d^{k1}\|^\nu \quad \forall\, i \in I_0, \tag{4.9}$$

and

$$g_i(x^k + d^k) + \langle \nabla g_i(x^k), \hat{d}^k\rangle = -\|d^k\|^\tau \quad \forall\, i \in I_0. \tag{4.10}$$

By (4.8) and Lemma 4.6, we have

$$\langle \nabla f(x^k), d^k\rangle = -\langle d^k, H_k d^k\rangle - \sum_{i \in I_0} z_i^k \langle \nabla g_i(x^k), d^k\rangle \tag{4.11}$$

and

$$\langle \nabla f(x^k), \hat{d}^k\rangle = -\sum_{i \in I_0} z_i^k \langle \nabla g_i(x^k), \hat{d}^k\rangle + o(\|d^k\|^2). \tag{4.12}$$

Thus, from (4.7), (4.11), and (4.12) we deduce

$$\begin{aligned} & f(x^k + d^k + \hat{d}^k) \\ &= f(x^k) + \frac{1}{2}\langle \nabla f(x^k), d^k\rangle + \frac{1}{2}\langle d^k, (\nabla^2_{xx} f(x^k) - H_k) d^k\rangle \\ & \quad -\frac{1}{2}\sum_{i \in I_0} z_i^k \langle \nabla g_i(x^k), d^k\rangle - \sum_{i \in I_0} z_i^k \langle \nabla g_i(x^k), \hat{d}^k\rangle + o(\|d^k\|^2). \end{aligned} \tag{4.13}$$

Furthermore, for all $i \in I_0$ it follows from (4.10) that

$$g_i(x^k) + \langle \nabla g_i(x^k), d^k + \hat{d}^k \rangle + \frac{1}{2}\langle d^k, \nabla^2_{xx} g_i(x^k) d^k \rangle = o(\|d^k\|^2). \tag{4.14}$$

Using (4.9), (4.14), and Lemma 3.2, we obtain

$$\begin{aligned}
&-\frac{1}{2}\sum_{i\in I_0} z_i^k \langle \nabla g_i(x^k), d^k \rangle - \sum_{i\in I_0} z_i^k \langle \nabla g_i(x^k), \hat{d}^k \rangle \\
&= \frac{1}{2}\sum_{i\in I_0} z_i^k \langle \nabla g_i(x^k), d^k \rangle - \sum_{i\in I_0} z_i^k \langle \nabla g_i(x^k), d^k + \hat{d}^k \rangle \\
&= \frac{1}{2}\sum_{i\in I_0} z_i^k g_i(x^k) - \frac{1}{2}\sum_{i\in I_0} \phi^k \|d^{k1}\|^\nu z_i^k \\
&\quad + \frac{1}{2}\sum_{i\in I_0} z_i^k \langle d^k, \nabla^2_{xx} g_i(x^k) d^k \rangle + o(\|d^k\|^2) \\
&= \frac{1}{2}\sum_{i\in I_0} z_i^k g_i(x^k) + \frac{1}{2}\sum_{i\in I_0} z_i^k \langle d^k, \nabla^2_{xx} g_i(x^k) d^k \rangle + o(\|d^k\|^2). \qquad (4.15)
\end{aligned}$$

Clearly, Assumption A4 and Corollary 4.4 imply that, for each $i \in I_0$ and any k sufficiently large, $z_i^k \geq 0.5\lambda_i^* > 0$; hence we get for k sufficiently large

$$\frac{1}{2}\sum_{i\in I_0} z_i^k g_i(x^k) + o(\|g_{I_0}(x^k)\|) < 0. \tag{4.16}$$

In view of (4.15)–(4.16) and Assumption A6′, we obtain from (4.13)

$$\begin{aligned}
&f(x^k + d^k + \hat{d}^k) \\
&= f(x^k) + \frac{1}{2}\langle \nabla f(x^k), d^k \rangle + \frac{1}{2}\sum_{i\in I_0} z_i^k g_i(x^k) \\
&\quad + \frac{1}{2}\langle d^k, (\nabla^2_{xx} f(x^k) + \sum_{i\in I_0} z_i^k \nabla^2_{xx} g_i(x^k) - H_k) d^k \rangle + o(\|d^k\|^2) \\
&= f(x^k) + \frac{1}{2}\langle \nabla f(x^k), d^k \rangle + \frac{1}{2}\sum_{i\in I_0} z_i^k g_i(x^k) + o(\|g_{I_0}(x^k)\|) \\
&\quad + \frac{1}{2}\langle d^k, P_k(\nabla^2_{xx} f(x^k) + \sum_{i\in I_0} z_i^k \nabla^2_{xx} g_i(x^k) - H_k) P_k d^k \rangle + o(\|d^k\|^2) \\
&\leq f(x^k) + \frac{1}{2}\langle \nabla f(x^k), d^k \rangle \\
&\quad + \frac{1}{2}\|d^k\| \|P_k\Big(\nabla^2_{xx} f(x^k) + \sum_{i\in I_0} z_i^k \nabla^2_{xx} g_i(x^k) - H_k\Big) P_k d^k\| + o(\|d^k\|^2) \\
&= f(x^k) + \frac{1}{2}\langle \nabla f(x^k), d^k \rangle + o(\|d^k\|^2), \qquad (4.17)
\end{aligned}$$

where the second equality follows from Lemmas 4.5 and 3.2. We also have from (4.4)

$$\begin{aligned}
\langle \nabla f(x^k), d^{k1} \rangle &= -\langle d^{k1}, H_k d^{k1} \rangle - \langle d^{k1}, \nabla g_{I_0}(x^k) z_{I_0}^{k1} \rangle \\
&= -\langle d^{k1}, H_k d^{k1} \rangle + \langle g_{I_0}(x^k), z_{I_0}^{k1} \rangle \\
&< -\langle d^{k1}, H_k d^{k1} \rangle, \qquad (4.18)
\end{aligned}$$

where the last inequality is due to $g_{I_0}(x^k) < 0$ and for k sufficiently large $z^{k1}_{I_0} > 0$. This, together with Lemma 2.3(iii), Assumption A3, and Lemma 3.2, implies that

$$\begin{aligned}\langle \nabla f(x^k), d^k\rangle &\le \vartheta \langle \nabla f(x^k), d^{k1}\rangle \\ &\le -\vartheta \langle d^{k1}, H_k d^{k1}\rangle \\ &\le -\vartheta C_1 \|d^{k1}\|^2 \\ &= -\vartheta C_1 \|d^k\|^2 + o(\|d^k\|^2). \end{aligned} \tag{4.19}$$

Due to $\mu < \frac{1}{2}$, inequalities (4.17) and (4.19) show that for k sufficiently large $t = 1$ satisfies inequality (2.6), i.e.,

$$f(x^k + d^k + \hat{d}^k) \le f(x^k) + \mu \langle \nabla f(x^k), d^k\rangle.$$

This proves (i). We now turn to prove (ii).

It is clear from Corollary 4.4 and Lemma 4.6 that $d^k \to 0$ and $\hat{d}^k \to 0$. For $i \notin I_0$, $g_i(x^*) < 0$ implies that for k sufficiently large

$$g_i(x^k + d^k + \hat{d}^k) < 0. \tag{4.20}$$

For $i \in I_0$, we have from (2.5) and Lemma 4.6 that for k sufficiently large

$$\begin{aligned} g_i(x^k + d^k + \hat{d}^k) &= g_i(x^k + d^k) + \langle \nabla g_i(x^k + d^k), \hat{d}^k\rangle + O(\|\hat{d}^k\|^2) \\ &= g_i(x^k + d^k) + \langle \nabla g_i(x^k), \hat{d}^k\rangle + O(\|d^k\|\|\hat{d}^k\|) \\ &= -\|d^k\|^\tau + O(\|d^k\|^3) \\ &= -\|d^k\|^\tau + o(\|d^k\|^\tau) \\ &\le -\frac{1}{2}\|d^k\|^\tau < 0. \end{aligned}$$

This, together with (4.20), shows (ii). This completes the proof. □

Proposition 4.7 shows that the use of $\hat{d}^k$ on the search direction makes the unit step accepted for all k sufficiently large. Consequently, the Maratos effect does not appear. The next theorem indicates that Algorithm 2.1 is two-step superlinearly convergent.

THEOREM 4.8. *Let Assumptions* A1–A5 *and* A6′ *hold. Then the sequence* $\{x^k\}$ *generated by Algorithm* 2.1 *converges two-step superlinearly, i.e.,*

$$\lim_{k\to\infty} \frac{\|x^{k+2} - x^*\|}{\|x^k - x^*\|} = 0.$$

The proof of the above theorem follows step by step, with minor modifications, that of Theorem 4.6 in [23]. The details are omitted.

Furthermore, in the following, we show the Q-superlinear convergence of Algorithm 2.1 if Assumption A6′ is replaced by a stronger assumption.

Assumption A6. The sequence of matrices $\{H_k\}$ satisfies

$$\frac{\|P_k(H_k - \nabla^2_{xx} L(x^*, \lambda^*))d^k\|}{\|d^k\|} \to 0.$$

THEOREM 4.9. *Let Assumptions* A1–A6 *hold. Then the sequence* $\{x^k\}$ *generated by Algorithm* 2.1 *converges Q-superlinearly, i.e.,*

$$\|x^{k+1} - x^*\| = o(\|x^k - x^*\|).$$

If, in addition, supposing that $\nabla^2 f$ and $\nabla^2 g_i$, for all $i \in I$, are Lipschitz continuous and $H_k = \nabla^2_{xx} L(x^k, \lambda(x^k))$, then the convergence rate is Q-quadratic, i.e.,

$$\|x^{k+1} - x^*\| = O(\|x^k - x^*\|^2).$$

Proof. In view of Proposition 4.7 and Lemmas 3.2 and 4.6, we have for k sufficiently large

$$x^{k+1} - x^k = d^k + \hat{d}^k = d^{k1} + O(\|d^{k1}\|^2) = d^{k1} + o(\|d^{k1}\|). \tag{4.21}$$

For k sufficiently large, d^{k1} can be viewed as a quasi-Newton direction for the equality constrained optimization problem (4.5). It follows from Lemma 3.2 that $d^k = d^{k1} + o(\|d^{k1}\|)$. By the boundedness of P_k, H_k, and $\nabla^2_{xx} L(x^*, \lambda^*)$, Assumption A6 is equivalent to

$$\frac{\|P_k(H_k - \nabla^2_{xx} L(x^*, \lambda^*))d^{k1}\|}{\|d^{k1}\|}.$$

Combining this expression and the results in [33], we have

$$\|x^k + d^{k1} - x^*\| = o(\|x^k - x^*\|). \tag{4.22}$$

Furthermore, by the use of Lemma 3.1 in [8], we get

$$\lim_{k\to\infty} \frac{\|d^{k1}\|}{\|x^k - x^*\|} = 1. \tag{4.23}$$

So, by (4.21)–(4.23), it holds that

$$\|x^{k+1} - x^*\| = \|x^k + d^{k1} + o(\|d^{k1}\|) - x^*\| = o(\|x^k - x^*\|),$$

which shows that $\{x^k\}$ converges to x^* Q-superlinearly.

If $H_k = \nabla^2_{xx} L(x^k, \lambda(x^k))$ for k sufficiently large, we get from Theorem 3.1 in [12] that

$$\|x^k + d^{k1} - x^*\| = O(\|x^k - x^*\|^2).$$

This, together with (4.21) and (4.23), yields

$$\|x^{k+1} - x^*\| = \|x^k + d^{k1} + O(\|d^{k1}\|^2) - x^*\| = O(\|x^k - x^*\|^2),$$

which shows that the convergence rate is Q-quadratic. ∎

5. Numerical experiments. In this section we report the numerical results on a test set that includes some of Hock and Schittkowski's problems [15] as well as several other large-scale real-world problems from the CUTE [4] and the COPS [5] collections. The algorithm was implemented by a Matlab code. For each test problem, we chose $H_1 = E$ as the initial guess of the Lagrangian Hessian. At each step, the matrix H_k was updated by the damped BFGS formula from Powell [24] as in [17, 26]. Specifically, we set

$$H_{k+1} = H_k - \frac{H_k s_k s_k^T H_k}{s_k^T H_k s_k} + \frac{y_k y_k^T}{s_k^T y_k},$$

where

$$y_k = \begin{cases} \hat{y}_k & \text{if } \hat{y}_k^T s_k \geq 0.2 s_k^T H_k s_k, \\ \theta_k \hat{y}_k + (1-\theta_k) H_k s_k & \text{otherwise} \end{cases}$$

and

$$\begin{cases} s_k = x^{k+1} - x^k, \\ \hat{y}_k = \nabla f(x^{k+1}) - \nabla f(x^k) + (\nabla g(x^{k+1}) - \nabla g(x^k))\lambda^{k0}, \\ \theta_k = 0.8 s_k^T H_k s_k / (s_k^T H_k s_k - s_k^T \hat{y}_k). \end{cases}$$

We set the parameters as follows:

$$\beta = 0.5, \quad \mu = 0.1, \quad \nu = 3.0, \quad \tau = 2.5, \quad \vartheta = 0.5, \quad \sigma = 0.1, \text{ and } \epsilon^0 = 3.0.$$

The algorithm stops if one of the following termination criteria is satisfied:

(a) $\|\Phi(x^k, \lambda(x^k))\| \leq 10^{-5}$.
(b) $\|\Phi(x^k, z^{k0})\| \leq 10^{-5}$.
(c) $\|d^{k1}\|/(1 + \|x^k\|) \leq 10^{-5}$.

The first and second criteria state the KKT conditions for problem (P). At Step 2 of the algorithm $\Phi(x^k, \lambda(x^k))$ has to be computed so as to estimate the working set and to update the parameter ϵ. Hence the first criterion is used here. Moreover, Lemma 2.2 implies that x^k is only a trivial KKT point of problem (P) if $d^{k1} = 0$. Hence in our implementation the second or third criterion is used at Step 4(ii) as the termination criterion.

The check of full rankness in Step 2 is done by using the Matlab command "rank."

We first tested some problems taken from [15]. For these test problems, we used the initial point given in [15] if it was strictly feasible. For some problems whose initial points given in [15] were not strictly feasible, we chose other initial points which were strictly feasible. These initial points are listed in Table 1.

TABLE 1
Starting points for some HS problems.

Problem	Starting point
HS25	(25, 5, 1)
HS30	(3, 2, 1)
HS31	(4, 3, -2)
HS33	(1, 3, 4)
HS34	(0.1, 1.15, 3.2)
HS65	(0, 0, 0)
HS66	(0.5, 2, 8)

The computational results are shown in Table 2, where the columns have the following meanings:

TABLE 2
Numerical results on the HS problems.

Problem	n	m	Iter	Nf	Ng	Fv	Term	Prec	Final-ϵ	Aset
HS1	2	1	36	57	59	6.662078e-14	(a)	3.862116e-06	0.3000	0
HS3	2	1	10	17	26	2.293930e-08	(a)	2.388519e-08	0.3000	1
HS4	2	2	3	4	7	2.666667	(a)	4.049708e-11	0.3000	2
HS5	2	4	5	8	8	-1.913223	(c)	6.007380e-06	0.3000	0
HS12	2	1	7	24	28	-30.000000	(c)	2.902889e-06	3.0000	1
HS24	2	5	9	13	22	-1.000000	(a)	7.192276e-08	0.3000	2
HS25	3	6	14	55	62	3.318784e-06	(c)	1.268899e-06	0.3000	0
HS29	3	1	9	28	34	-22.627417	(a)	6.168718e-06	3.0000	1
HS30	3	7	10	27	34	1.000000	(a)	6.986504e-07	0.3000	2
HS31	3	7	11	32	40	6.000000	(c)	1.925959e-06	0.3000	1
HS33	3	6	15	74	87	-4.585782	(a)	7.512854e-07	0.3000	3
HS34	3	8	17	76	92	-0.834024	(c)	1.611863e-06	0.3000	3
HS35	3	4	7	13	19	0.111111	(c)	8.006110e-06	0.3000	1
HS36	3	7	11	33	44	-3.300000e+03	(a)	6.296636e-07	0.3000	3
HS37	3	8	15	45	58	-3.456000e+03	(c)	6.947600e-06	0.3000	1
HS38	4	8	49	91	91	5.128073e-11	(c)	1.890126e-06	0.0300	0
HS43	4	3	12	36	45	-44.000000	(c)	6.631011e-06	0.3000	2
HS44	4	10	17	60	73	-14.999860	(c)	3.117109e-06	0.3000	4
HS65	3	7	8	19	22	0.953529	(a)	6.193657e-08	0.3000	1
HS66	3	8	11	24	35	0.518164	(a)	7.452867e-06	3.0000	2
HS76	4	7	9	15	23	-4.681818	(a)	8.369451e-10	0.3000	2
HS93	6	8	18	51	69	135.075964	(c)	2.708803e-06	0.3000	2
HS100	7	4	14	44	58	680.630057	(c)	6.183027e-06	3.0000	2
HS113	10	8	21	58	79	24.306209	(c)	1.424558e-06	3.0000	6

Problem: the problem number given in [15],
n: the number of variables,
m: the number of constraints (including bound constraints),
Iter: the number of iterations,
Nf: the number of function evaluations for f,
Ng: the number of function evaluations for g,
Fv: the objective function value at the final iterate,
Term: the label of the termination criterion,
Prec: the final value of the norm function used in the termination criteria,
Final-ϵ: the value of the parameter ϵ at the final iterate,
Aset: the number of indices in the final working set.

We succeeded in solving all test problems chosen in Table 2, and for most of these problems the number of iterations was small. The computational results illustrate that our algorithm is competitive with those in [26, 34].

All of the problems in the Hock and Schittkowski set [15] are very small. To see more clearly the effectiveness of our algorithm, we tested several problems from the CUTE collection [4] and two problems from the COPS collection [5] that contained no equality constraints. Some of these problems are larger and therefore more interesting. Table 3 lists starting points of these problems, except for the last problem, whose initial points vary with its dimension. We also succeeded in solving all these test problems. The computational results are listed in Tables 4 and 5, where the termination criterion (c) is changed to $\|d^{k1}\| \leq 10^{-5}$.

The results reported in Tables 4 and 5 are encouraging. First, we note that here the number of iterations and hence the number of objective function evaluations are

TABLE 3
Starting points for problems in Tables 4 and 5.

Problem	Starting point
Expfit	$x = (6, 1, 6, 0, 0)^T$
Ngone-k	$x_i = 0.8 * i/k, i = 1, \cdots, k; y_i = 0.6, i = 1, \cdots, k-1$
Obstclae-k	$x_{i,j} = 1, i, j = 2, \cdots, k-1$
Svanberg-k	$x_i = 0, i = 1, \cdots, k$
Polygon-k	$r_i = 0.5, \theta_i = \pi * i/k, i = 1, \cdots, k-1$

TABLE 4
Numerical results on the CUTE problems.

Problem	n	m	Iter	Nf	Ng	Fv	Term	Prec	Final-ϵ	Aset
Expfita	5	22	228	1758	1897	0.00113661	(a)	1.866081e-08	0.0300	4
Expfitb	5	102	157	1107	1204	0.00501937	(a)	1.159079e-08	0.0300	4
Expfitc	5	502	273	2824	2964	0.02330257	(a)	2.493369e-06	0.0003	3
Ngone-3	8	9	5	11	15	-0.500000	(c)	8.261035e-06	0.3000	2
Ngone-5	12	20	14	26	39	-0.620366	(a)	9.202883e-07	0.3000	7
Ngone-24	50	324	241	95	1199	-0.643097	(b)	8.010034e-06	0.0300	26
Ngone-49	100	1274	1414	10876	12290	-0.643421	(c)	8.811006e-06	0.0300	51
Obstclae-4	16	32	4	5	9	0.753660	(a)	5.825214e-07	3.0000	4
Obstclae-10	100	200	165	979	1144	1.397898	(a)	8.353902e-06	3.0000	29
Obstclae-23	529	1058	908	7179	8087	1.678027	(a)	7.187001e-06	3.0000	221
Obstclae-32	1024	2048	2438	22024	24462	1.748270	(a)	9.658341e-06	3.0000	472
Svanberg-10	10	30	36	227	258	15.731517	(c)	5.365582e-06	0.0300	6
Svanberg-30	30	90	101	777	864	49.142526	(c)	9.130506e-06	0.0300	22
Svanberg-50	50	150	108	881	968	82.581912	(c)	9.472167e-06	0.0300	38
Svanberg-80	80	240	190	1666	1835	132.749819	(c)	4.663239e-06	0.0300	61
Svanberg-100	100	300	178	1628	1782	166.197171	(c)	7.281111e-06	0.0300	77
Svanberg-500	500	1500	402	4020	4407	835.186918	(c)	5.299494e-06	0.0300	398

TABLE 5
Numerical results on the COPS problems.

Problem	n	m	Iter	Nf	Ng	Fv	Term	Prec	Final-ϵ	Aset
Polygon-4	6	17	6	17	21	-0.500000	(a)	9.911252e-09	0.3000	2
Polygon-6	10	34	13	26	38	-0.674981	(a)	1.813151e-07	0.3000	6
Polygon-10	18	80	18	31	49	-0.749137	(a)	3.669190e-06	0.3000	10
Polygon-15	28	160	43	159	199	-0.768622	(a)	7.178912e-06	0.0300	15
Polygon-20	38	265	78	348	422	-0.776859	(a)	9.648048e-06	0.0300	20
Polygon-25	48	395	96	403	494	-0.780232	(a)	9.167416e-06	0.0300	25
Polygon-30	58	550	137	739	872	-0.781674	(a)	9.128916e-06	0.0300	30
Polygon-40	78	935	416	3113	3509	-0.783069	(a)	7.798918e-06	0.0030	40
Polygon-50	98	1420	1416	14855	16192	-0.783799	(b)	8.082852e-06	0.0003	50
Cam-10	10	43	15	155	170	-43.85994	(a)	1.336130e-07	3.0e-4	10
Cam-20	20	83	14	110	124	-86.55864	(a)	3.498434e-06	3.0e-4	20
Cam-50	50	203	14	166	180	-214.6961	(b)	6.369979e-06	3.0e-6	50
Cam-100	100	403	17	244	261	-427.8899	(b)	9.198429e-06	3.0e-6	100
Cam-200	200	803	55	552	607	-855.7000	(c)	7.568989e-08	3.0e-7	200
Cam-400	400	1603	98	1207	1305	-1710.275	(c)	9.040695e-08	3.0e-8	400

generally larger than those reported in [17] for a feasible SQP method. This is understandable because the subproblems of Algorithm 2.1 are low dimensional, which use only partial information of the problems. The number of constraint function evaluations here is competitive with that of a feasible SQP method. On the other

TABLE 6
Number of indices in the working set on the problem "Obstclae-10."

Iteration	1	120	123	131	134	135	140	143	144	147
Working set	64	62	64	62	60	58	56	55	56	55
Iteration	148	149	150	154	156	158	161	162	163	
Working set	53	39	36	35	34	33	32	31	29	

hand, Tables 4 and 5 also show that the cardinality of the final working set "Aset" is generally much smaller than the number of constraints. This means that the subproblems of Algorithm 2.1 are generally much smaller than that of the full dimensional feasible SQP methods. Moreover, as the number of constraints in problem (P) increases, this benefit becomes extremely apparent. This shows the potential advantage of our algorithm when applied to solving problems with large numbers of constraints. Table 6 positively supports this possibility. Table 6 lists the numbers of indices in the working set corresponding to iterations when Algorithm 2.1 is applied to solving problem "Obstclae-10." The results show that as iteration increases, the number of corresponding indices in the working set exhibits the decreasing tendency.

6. Conclusion. In this paper an FSLE algorithm for inequality constrained optimization is proposed. The proposed algorithm is based on an efficient identification technique of the active constraints and has some nice properties. We have proved that every accumulation point of the sequence generated by the proposed algorithm is a KKT point of problem (P) without requiring the isolatedness of the stationary points. We have also established locally two-step superlinear or Q-superlinear or Q-quadratic convergence for the proposed algorithm under mild assumptions. The preliminary numerical experiments show that the proposed method is effective for the test problems. However, to achieve superlinear convergence of the algorithm we still need the strict complementarity condition. Recently, Facchinei, Lucidi, and Palagi [13] proposed a globally and superlinearly convergent truncated Newton method for solving the box constrained optimization. In particular, they established superlinear convergence without requiring the strict complementarity condition. How to remove this condition for the general constrained optimization is an important topic for further research.

REFERENCES

[1] S. BAKHTIARI AND A. TITS, *A simple primal-dual feasible interior-point method for nonlinear programming with monotone descent*, Comput. Optim. Appl., 25 (2003), pp. 17–38.

[2] P. T. BOGGS AND J. W. TOLLE, *Sequential Quadratic Programming*, Acta Numer. 4, Cambridge University Press, Cambridge, UK, 1995, pp. 1–51.

[3] P. T. BOGGS, J. W. TOLLE, AND P. WANG, *On the local convergence of quasi-Newton methods for constrained optimization*, SIAM J. Control Optim., 20 (1982), pp. 161–171.

[4] I. BONGARTZ, A. R. CONN, N. I. M. GOULD, AND PH. L. TOINT, *CUTE: Constrained and unconstrained testing environment*, ACM Trans. Math. Software, 21 (1995), pp. 123–160.

[5] A. S. BONDARENKO, D. M. BORTZ, AND J. J. MORÉ, *COPS: Large-Scale Nonlinearly Constrained Optimization Problems*, Technical report ANL/MCS-TM-237, Argonne National Laboratory, Argonne, IL, 1998.

[6] J. V. BURKE AND S. P. HAN, *A robust sequential quadratic programming method*, Math. Programming, 43 (1989), pp. 277–303.

[7] J. E. DENNIS AND J. J. MORÉ, *A characterization of superlinear convergence and its application to quasi-Newton methods*, Math. Comp., 28 (1974), pp. 549–560.

[8] F. FACCHINEI, *Minimization of SC^1-functions and the Maratos effect*, Oper. Res. Lett., 17 (1995), pp. 131–137.

[9] F. Facchinei, A. Fischer, and C. Kanzow, *On the accurate identification of active constraints*, SIAM J. Optim., 9 (1998), pp. 14–32.
[10] F. Facchinei, J. Júdice, and J. Soares, *An active set Newton algorithm for large-scale nonlinear programs with box constraints*, SIAM J. Optim., 8 (1998), pp. 158–186.
[11] F. Facchinei and C. Lazzari, *Local feasible QP-free algorithm for the constrained minimization of SC^1 functions*, J. Optim. Theory Appl., to appear.
[12] F. Facchinei and S. Lucidi, *Quadratically and superlinearly convergent algorithms for the solution of inequality constrained minimization problems*, J. Optim. Theory Appl., 85 (1995), pp. 265–289.
[13] F. Facchinei, S. Lucidi, and L. Palagi, *A truncated Newton algorithm for large scale box constrained optimization*, SIAM J. Optim., 12 (2002), pp. 1100–1125.
[14] T. Glad and E. Polak, *A multiplier method with automatic limitation of penalty growth*, Math. Programming, 17 (1979), pp. 140–155.
[15] W. Hock and K. Schittkowski, *Test Examples for Nonlinear Programming Codes*, Lecture Notes in Econom. and Math. Systems 187, Springer-Verlag, Berlin, New York, 1981.
[16] C. Kanzow and H. D. Qi, *A QP-free constrained Newton-type method for variational inequality problems*, Math. Program., 85 (1999), pp. 81–106.
[17] C. T. Lawrence and A. L. Tits, *A computationally efficient feasible sequential quadratic programming algorithm*, SIAM J. Optim., 11 (2001), pp. 1092–1118.
[18] X.-W. Liu and Y.-X. Yuan, *A robust algorithm for optimization with general equality and inequality constraints*, SIAM J. Sci. Comput., 22 (2000), pp. 517–534.
[19] S. Lucidi, *New results on a continuously differentiable exact penalty function*, SIAM J. Optim., 2 (1992), pp. 558–574.
[20] J. J. Moré and D. C. Sorensen, *Computing a trust region step*, SIAM J. Sci. Statist. Comput., 4 (1983), pp. 553–572.
[21] E. R. Panier and A. L. Tits, *A superlinearly convergent feasible method for the solution of inequality constrained optimization problems*, SIAM J. Control Optim., 25 (1987), pp. 934–950.
[22] E. Panier and A. L. Tits, *On combining feasibility, descent and superlinear convergence in inequality constrained optimization*, Math. Programming, 59 (1993), pp. 261–276.
[23] E. R. Panier, A. L. Tits, and J. N. Herskovits, *A QP-free globally convergent, locally superlinearly convergent algorithm for inequality constrained optimization*, SIAM J. Control Optim., 26 (1988), pp. 788–811.
[24] M. J. D. Powell, *A fast algorithm for nonlinearly constrained optimization calculations*, in Numerical Analysis, Lecture Notes in Math. 630, Springer, Berlin, 1978, pp. 144–157.
[25] M. J. D. Powell, *The convergence of variable metric methods for nonlinearly constrained optimization calculations*, in Nonlinear Programming 3, O. L. Mangasarian, R. R. Meyer, and S. M. Robinson, eds., Academic Press, New York, 1978, pp. 27–63.
[26] H.-D. Qi and L. Qi, *A new QP-free globally convergent, locally superlinearly convergent algorithm for inequality constrained optimization*, SIAM J. Optim., 11 (2000), pp. 113–132.
[27] 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.
[28] L. Qi and Z. Wei, *Corrigendum: On the constant positive linear dependence condition and its application to SQP methods*, SIAM J. Optim., 11 (2001), pp. 1145–1146.
[29] L. Qi and Y. F. Yang, *A globally and superlinearly convergent SQP algorithm for nonlinear constrained optimization*, J. Global Optim., 21 (2001), pp. 157–184.
[30] S. M. Robinson, *Strongly regular generalized equations*, Math. Oper. Res., 5 (1980), pp. 43–62.
[31] P. Spellucci, *A new technique for inconsistent QP problems in the SQP methods*, Math. Methods Oper. Res., 47 (1998), pp. 355–400.
[32] P. Spellucci, *An SQP method for general nonlinear programs using only equality constrained subproblems*, Math. Programming, 82 (1998), pp. 413–448.
[33] J. Stoer and A. Tapia, *On the characterization of q-superlinear convergence of quasi-Newton methods for constrained optimization*, Math. Comp., 49 (1987), pp. 581–584.
[34] T. Urban, A. L. Tits, and C. T. Lawrence, *A Primal-Dual Interior-Point Method for Nonconvex Optimization with Multiple Logarithmic Barrier Parameters and with Strong Convergence Properties*, Institute for Systems Research Technical report TR 98-27, University of Maryland, College Park, MD, 1998.