SIAM J. OPTIM.
Vol. 17, No. 1, pp. 1–36
© 2006 Society for Industrial and Applied Mathematics

SOLVING MATRIX INEQUALITIES WHOSE UNKNOWNS ARE MATRICES*

JUAN F. CAMINO†, J. WILLIAM HELTON‡, AND ROBERT E. SKELTON§

Abstract. This paper provides algorithms for numerical solution of convex matrix inequalities in which the variables naturally appear as matrices. This includes, for instance, many systems and control problems. To use these algorithms, no knowledge of linear matrix inequalities is required. However, as tools, they preserve many advantages of the linear matrix inequality framework. Our method has two components: (1) a numerical algorithm that solves a large class of matrix optimization problems and (2) a symbolic "convexity checker" that automatically provides a region which, if convex, guarantees that the solution from (1) is a global optimum on that region. The algorithms are partly numerical and partly symbolic and since they aim at exploiting the matrix structure of the unknowns, the symbolic part requires the development of new computer techniques for treating noncommutative algebra.

Key words. matrix inequalities, convex optimization, semidefinite programming, noncommutative algebra, computer algebra

AMS subject classifications. 90C25, 90C22, 15A42, 15A45, 93A99

DOI. 10.1137/040613718

1. The basic idea.

Since the early 1990s, matrix inequalities (MIs) have become very important in engineering, particularly in control theory. If one has the ability to convert the MIs arising in a particular problem to a linear matrix inequality (LMI), then the problem can be solved up to substantial size. The wide acceptance of LMIs stems from the following advantages:

1. If a control problem is posed as an LMI, then any local solution is a global optimum.
2. Efficient numerical LMI solvers are readily available.
3. Once a control problem is posed as an LMI, adding constraints in the form of LMIs results in a LMI problem.

On the other hand, the LMI framework has the following disadvantages:

1. There is no systematic way to produce LMIs for general classes of problems.
2. There is no way of knowing whether it is possible to reduce a system problem to an LMI without actually doing it.
3. The user must possess the knowledge of manipulating LMIs, which takes considerable training. Indeed, if one does not have the ability to deal with LMIs, then it is not clear what one should do.
4. Transformations via Schur complements can lead to a large LMI representation.

*Received by the editors August 20, 2004; accepted for publication (in revised form) September 19, 2005; published electronically April 21, 2006.
http://www.siam.org/journals/siopt/17-1/61371.html

†Department of Computational Mechanics, School of Mechanical Engineering, State University of Campinas, 13083-970, Campinas, SP, Brazil (camino@fem.unicamp.br). This author was partly supported by CAPES/Brazil and by NSF grants DMS 0100576 and 0400794.

‡Department of Mathematics, University of California at San Diego, La Jolla, CA 92093-0112 (helton@math.ucsd.edu). This author was partly supported by NSF grants DMS 0100576 and 0400794 and the Ford Motor Company.

§Department of Mechanical and Aerospace Engineering, University of California at San Diego, La Jolla, CA 92093-0411 (bobskelton@ucsd.edu). This author was partly supported by DARPA.

1.1. Our method. The main objective for this paper is to provide a method for solving MIs that possesses similar advantages to the LMI framework but without its main disadvantages. Our method has two components:

1. a numerical algorithm, called NCSDP, that solves a large class of matrix optimization problems;
2. a symbolic "convexity checker" that automatically provides a region $\mathcal{G}$. If $\mathcal{G}$ is convex, then the solution from (1) is a global optimum on $\mathcal{G}$. Also, convexity ensures good numerical behavior of NCSDP on $\mathcal{G}$.

1.2. The convexity region algorithm. The symbolic convexity region algorithm receives as input a function $F(x)$ and gives as output a family of inequalities that determine a region $\mathcal{G}$ of x on which $F(x)$ is "matrix convex." Often, we just refer to matrix convexity as convexity and it is defined precisely in section 4.5. This algorithm produces sufficient conditions, which with some very weak hypotheses are necessary conditions for convexity. A concern is that the output might produce a "region of convexity $\mathcal{G}$" with several connected components, in which case the user must select one of them. (See section 3.1 for an example.)

1.3. The numerical solver for matrix inequalities. Our NCSDP solver can be used to solve optimization problems involving matrix inequalities. It is designed for situations where there are only a few unknown matrices and it attempts with symbolic manipulation (as well as numerics) to use the matrix structure to advantage. The solver has very reliable behavior in convex situations. The novel features of our algorithm that allow us to view the matrices as unknowns, rather than the entries of these matrices as unknowns, are discussed in sections 4.4, 7.2, and 8.

1.4. Combining the tools. Putting together the convexity checker and the NCSDP solver, we have a set of tools to solve many engineering problems that can be posed as matrix inequalities with matrix unknowns. Section 3 gives an example of these tools. Our method is effective on problems with few unknowns, but we reiterate that we can take each unknown to be a matrix. This is not a serious restriction for many system problems (e.g., most of the classics [23]).

1.5. LMI analogues. In some sense, there is a parallel between the conventional LMI approach and our approach. In the former, one needs to be able to convert the optimization problem over matrix functions into an equivalent LMI problem, so that some available LMI solver can be used. In our approach, the convexity checker provides a region $\mathcal{G}$ which, if convex, guarantees reliable behavior of our NCSDP solver and that a solution is a global optimum on $\mathcal{G}$.

1.6. Matrix unknowns. The advantage of dealing with matrices as single letters is that one letter z can stand for a matrix Z with n^2 commuting variables. In typical engineering situations, most problems have few matrix unknowns, often two or three, and few (not exceedingly complicated) constraints (usually fewer than 10). This contrasts with treating matrices in terms of their entries where one often has several thousand variables. A disadvantage is that matrix multiplication is not commutative and so we must develop computer tools for performing algebraic operations on noncommuting variables. The major focus of this research is how to use the matrix structure of the unknowns to advantage, and this will come out as the article unfolds. Our algorithms, including the NCSDP solver, combine symbolic and numerical manipulations and lead to several very natural open questions.

1.7. Software availability. The user interface of our NCSDP code is not polished and we do not yet distribute NCSDP. However, the convexity checker algorithm is well documented and available through NCAlgebra, a noncommutative algebra package that runs under Mathematica. This package provides a large number of useful commands and functions for symbolic computation. It can be downloaded from http://math.ucsd.edu/∼ncalg.

2. Nomenclature. We use uppercase letters (e.g., X) for matrices and lowercase letters (e.g., x) for symbolic variables. The notation $\mathcal{Q}, \mathcal{H}$ stands for the symbolic gradient and Hessian maps, and the notation $\mathbb{Q}, \mathbb{H}$ is used to indicate we have substituted matrices of compatible size for the symbolic variables in $\mathcal{Q}, \mathcal{H}$. The n-dimensional Euclidean space is denoted by $\mathbb{R}^n$. The space of $n \times m$ real matrices is denoted by $\mathbb{R}^{n\times m}$. The space of $n \times n$ symmetric matrices with real entries is denoted by $\mathbb{S}^n$. Let $(\mathbb{S}^n)^g$ stand for the direct product $\mathbb{S}^n \times \mathbb{S}^n \times \cdots \times \mathbb{S}^n$ of order g. The expression $A \geq B$ ($A > B$) means that $A - B$ is a positive semidefinite (positive definite) matrix. The associated spaces are respectively denoted by $\mathbb{S}_+$ and $\mathbb{S}_{++}$. The usual Kronecker product of two matrices A and B is denoted by $A \otimes B$ and the trace of A is $\mathrm{Tr}\,\{A\}$. To define the vec operation, let us associate the vector $\mathrm{vec}(X) \in \mathbb{R}^{nm}$ with each matrix $X \in \mathbb{R}^{n\times m}$ by listing the entries of the columns, column by column, that is, $\mathrm{vec}(X) = [X_{11}, X_{21}, \ldots, X_{n1}, X_{12}, \ldots, X_{n2}, \ldots, X_{1m}, \ldots, X_{nm}]^T$.

3. Introducing our approach by an example. Suppose one is given two matrices[1] A and S, where S is symmetric, and one needs to solve the following problem:

$$\max\{\mathrm{Tr}\,\{X\} \,:\, (X, Y, A, S) \in \mathrm{closure}(\mathcal{S})\}, \tag{P1}$$

where the domain $\mathcal{S}$ is given by

$$\mathcal{S} = \left\{(X, Y, A, S) \in \mathcal{V} \,:\, F(X, Y, A, S) < 0, \quad X^2 < I, \quad Y > 0, \quad Y^2 < I\right\}$$

with $\mathcal{V} = \mathbb{S}^n \times \mathbb{S}^n \times \mathbb{R}^{n\times n} \times \mathbb{S}^n$ and

$$\begin{aligned} F(X, Y, A, S) := &-AX(XA^TY^{-1}AX - Y)^{-1}XA^T \\ &- (Y^{-1}(XA^TY^{-1}AXY^{-1} - Y)Y^{-1})^{-1} - AX(Y^{-1}(XA^TY^{-1}AX - Y))^{-1} \\ &- ((XA^TY^{-1}AX - Y)Y^{-1})^{-1}XA^T + XA^TY^{-1}AX - S. \end{aligned}$$

To solve this problem, we apply our two-step methodology:

1. Determine a domain $\mathcal{G}$ on which the above problem is convex.
2. Solve numerically the optimization problem on $\mathcal{G}$ using NCSDP.

3.1. Step 1. Determining a region of convexity in problem (P1). This step is purely symbolic and we do not use the particular numerical values of A and S given in (3.1). We describe this step using standard TEX notation rather than displaying actual computer runs.

The problem (P1) is to maximize $\mathrm{Tr}\,\{X\}$ over the domain

$$\mathcal{S} := \mathcal{S}_1 \cap \mathcal{S}_2$$

[1] Ultimately, in our example we shall take the matrices A and S to be

$$A = \begin{bmatrix} 1 & -1 \\ 0 & 2 \end{bmatrix}, \qquad S = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}. \tag{3.1}$$

The matrices are chosen small to save space.

with

$$\mathcal{S}_1 := \{(X,Y,A,S) \in \mathcal{V} : Y^2 < I, \quad X^2 < I\}$$

and

$$\mathcal{S}_2 := \{(X,Y,A,S) \in \mathcal{V} : F(X,Y,A,S) < 0, \quad Y > 0\}.$$

This optimization problem will be convex whenever $\mathcal{S}$ is a convex domain, since the objective function $\mathrm{Tr}\{X\}$ is linear in X. It is clear that $\mathcal{S}_1$ is convex; we wish to show that $\mathcal{S}_2$ is convex so that we can conclude that $\mathcal{S}$ is convex. For this purpose, we use our symbolic package to find the region where $F(X,Y,A,S)$ is convex with respect to X, Y in the domain $\mathcal{S}_2$.

Since matrix multiplication is not commutative, we must treat the matrices X, Y, A, and S symbolically as noncommutative variables. Thus, we load the Mathematica package NCAlgebra, which contains our convexity checker software. We type in the function F just as we see it in the definition of F and apply the convexity checker algorithm **NCConvexityRegion[]** (see section 5) using its default set of permutations. One of the outputs is the list

$$\{2y^{-1},\ -2(xa^Ty^{-1}ax - y)^{-1},\ 2y^{-1},\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0\}.$$

The interpretation of the output is that $F(X,Y,A,S)$ will be a convex function on the region consisting of all matrices that make each nonzero entry in the output list a positive definite expression, which, in this case, is the region given by

$$2Y^{-1} > 0 \qquad \text{and} \qquad -2(XA^TY^{-1}AX - Y)^{-1} > 0.$$

Thus, we conclude from this output that $F(X,Y,A,S)$ is simultaneously convex in X and Y whenever A, S, X, and Y are matrices of compatible dimension in the region $\mathcal{G}_{\mathcal{S}_2}$ given by

$$\mathcal{G}_{\mathcal{S}_2} := \{(X,Y,A,S) \in \mathcal{V} : Y > 0, \quad XA^TY^{-1}AX < Y\}. \tag{3.2}$$

To find if the above region $\mathcal{G}_{\mathcal{S}_2}$ is itself simultaneously convex in X and Y, we run the convexity checker once more on the function $G(X,Y,A) := XA^TY^{-1}AX - Y$,

NCConvexityRegion$[xa^Ty^{-1}ax - y, \{x,\ y\}]$.

This command outputs the list $\{2y^{-1},\ 0\}$. Thus, the region $\mathcal{G}$ is convex on matrices Y satisfying $Y > 0$. Thus, the region where the function G is convex consists of matrices Y satisfying $Y > 0$; consequently the region $\mathcal{G}_{\mathcal{S}_2}$ in (3.2) is convex. Thus, we can conclude that the optimization problem (P1) is convex inside the convex region

$$\mathcal{G} := \{(X,Y,A,S) \in \mathcal{V} : (X,Y,A,S) \in \mathcal{S} \cap \mathcal{G}_{\mathcal{S}_2}\}.$$

Equivalently

$$\mathcal{G} := \{(X,Y,A,S) \in \mathcal{V} : F(X,Y,A,S) < 0, \quad XA^TY^{-1}AX < Y, \quad Y > 0, \quad Y^2 < I, \quad X^2 < I\}.$$

Note that the region of convexity $\mathcal{G}$ for the optimization problem (P1) was determined without considering any specific numerical values for A and S. Thus, the set

of inequalities $\mathcal{G}$ characterizes a convex region for any arbitrary choice of two $n \times n$ matrices A and symmetric S no matter what value n is. Whether $\mathcal{G}$ is the biggest such region we have not said. In fact, the algorithm addresses this, requiring an interpretation, for which we refer to [5], or see section 5 for an abbreviated account. For the example above this gives that the largest subregion of matrix tuples (of large enough size) on which the Hessian is positive is the closure of $\mathcal{G}$.

3.2. Step 2. Invoking the NCSDP solver. Until this point, all calculations were symbolic. Now, we make the particular numerical choice for the matrices A, S given in (3.1). The optimization problem (P1) can now be solved with the NCSDP solver reliably and globally on the convex region of 2×2 matrices satisfying the constraints $\mathcal{G}$. We emphasize that this amounts to adding the following convex constraint

$$XA^TY^{-1}AX < Y \tag{3.3}$$

to the constraints defining $\mathcal{S}$. *Thus, we are not solving exactly the original problem, and the user must decide if this constraint meets his or her engineering needs.* Beware that declining to add the constraint (3.3) subjects one to the difficulties found in nonconvex situations, but one can still run numerical optimization routines.

To use NCSDP, we define the objective for this optimization problem as

$$\text{obj} := -\operatorname{Tr}\{X\},$$

subject to the constraint $G_i < 0$, where

$$\begin{aligned} G_1 &:= F(X, Y, A, S), \qquad G_2 := XA^TY^{-1}AX - Y, \\ G_3 &:= -Y, \qquad G_4 := YY - I, \qquad G_5 := XX - I, \end{aligned}$$

with

$$A = \begin{bmatrix} 1 & -1 \\ 0 & 2 \end{bmatrix}, \qquad S = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}.$$

Using this input, namely, (obj, $\{G_1, G_2, G_3, G_4, G_5\}$, $\{X, Y\}$), we run NCSDP. The solver finds the global optimizers over $\mathcal{G}$ to be

$$X^* = \begin{bmatrix} 0.3421 & 0.0263 \\ 0.0263 & 0.0788 \end{bmatrix}, \qquad Y^* = \begin{bmatrix} 0.8107 & 0.0016 \\ 0.0016 & 0.4255 \end{bmatrix}.$$

The optimal cost is therefore $\operatorname{Tr}\{-X^*\} = -0.4208$. We repeat that X^*, Y^* is a global optimum over the region

$$\mathcal{S} \cap \{(X, Y, A, S) \in \mathbb{S}^{2\times 2} \,:\, XA^TY^{-1}AX < Y, \quad Y > 0\}$$

with the matrices A and S as given above.

We emphasize there is no need to know much Mathematica to use NCSDP, unless one desires to use the convexity checker, and related symbolic algorithms, as we did in this example.

3.3. Scope of our methods. The method we shall describe here applies to problems of the form

$$\min_{X_i} \{\operatorname{Tr}\{X_1\} \,:\, X_i \in \text{closure}(\mathcal{G})\},$$

where the feasibility region $\mathcal{G}$ is given by

$$\mathcal{G} = \Big\{ (X_i, A_j) \,:\, F_1(X_i, A_j) > 0, \ldots, F_\ell(X_i, A_j) > 0 \Big\}$$

with $F_1(X_i, A_j), \ldots, F_\ell(X_i, A_j)$ rational expressions of noncommutative variables A_j, X_i, X_i^T. We assume the closure of the set $\mathcal{G}$ is compact. We can take some of the variables to be formally symmetric, like $X_7 = X_7^T$. The methods also apply to the feasibility problem, namely, determining if $\mathcal{G}$ is empty. We expect that (once refined) such methods will have advantages when the F_k are not highly complicated expressions.

An example of a problem we do not treat here is

$$\min \mathrm{Tr}\,\{X\} \qquad \text{subject to} \quad \mathrm{Tr}\,\big\{X^2\big\} \le 1.$$

However, we think our methods extend to such situations.

Space and expository considerations forced us to consider a single function $F(X)$ of a single symmetric variable $X = X^T$. The extension to the multivariate case stated above, found in [4], follows similar ideas, but it is too long to present here.

3.4. Comparing to the LMI approach. The optimization problem (P1) was actually selected to correspond to an LMI problem, so that we could compare approaches. There is not enough space to describe this in detail. (The corresponding LMI system has dimension 4×4.) We found that our approach produced exactly what was obtained using the LMI. Indeed our "extra condition" $XA^TY^{-1}AX < Y$ was a necessary condition for the LMI to be positive definite. Thus, from the LMI point of view, it is an essential constraint.

Since transformations via Schur complements can lead to an LMI representation with large constraint block matrices, the NCSDP solver has the potential to reduce the optimization time significantly compared to primal-dual solvers (see section 9).

3.5. Comparing to optimization over functions of commutative variables. If one has a complicated polynomial or rational function F, then there are typically many isolated regions on which the Hessian of F is positive definite. In our terminology, there are many "regions of convexity" for F. Thus, our technique requires selecting those convexity regions of interest and finding optima on them.

To those whose experience is with classical rational optimization, this seems odd, because there are many regions of convexity for F. However, our motivation comes from systems engineering problems, where we reemphasize that the number of matrix unknowns is small and that the rational functions are not terribly complicated; consequently F has a few connected regions of convexity. Moreover, the inequality constraints in a problem (e.g., $Y > 0$, $X^2 < I$) often select one convexity region.

4. Background on NC rational functions and convexity.

4.1. NC polynomials. We work with noncommutative (NC) polynomials with real numbers as coefficients in variables $x = \{x_1, \ldots x_g\}$. They cause little confusion, so a few examples suffice for an introduction:

$$p(x) = x_1x_2x_1 + x_1x_2 + x_2x_1, \qquad x_1^T = x_1, \quad x_2^T = x_2,$$

where the variables x_j are formally symmetric. In this next expression

$$p(x) = x_1^Tx_2x_1 + x_1^Tx_2 + x_2x_1, \qquad x_2^T = x_2,$$

the variable x_2 is formally symmetric but x_1 is not. Often, the term indeterminant is used instead of the term variable.

An NC polynomial p is symmetric provided that it is formally symmetric with respect to the involution ^{T}. Often, we shall substitute $n \times n$ matrices $X_1, \ldots, X_g$ into p for the variables $x_1, \ldots, x_g$. For a symmetric p, if the x_j are designated as symmetric variables, then the matrices X_j must be taken to be symmetric, and the resulting matrix $p(X_1, \ldots, X_g)$ is symmetric. The variables x_j which are not declared symmetric, if substituted by the matrix X_j, also result in the variables x_j^T being substituted by X_j^T.

4.2. NC rational functions. We shall discuss the notion of an NC rational function in terms of rational expressions. There is a technicality, "analytic at 0," which we include, since it makes formal definitions simpler. Casual readers can ignore it, since assuming analyticity elsewhere suffices.

An NC rational expression analytic at 0 is defined recursively. NC polynomials are NC rational expressions as are all sums and products of NC rational expressions. If r is a NC rational expression and $r(0) \neq 0$, then the inverse of r is a rational expression.

The notion of the formal domain of a rational expression r, denoted $\mathcal{F}_{r,\text{formal}}$, very roughly speaking, is

$$\mathcal{F}_{r,\text{formal}} := \{X \,:\, r(X) \text{ is defined } (\text{is not infinite})\}\,.$$

More precisely, the formal domain and the evaluation $r(X)$ of the rational expression at a tuple $X \in (\mathbb{S}^n)^g \bigcap \mathcal{F}_{r,\text{formal}}$ are both defined recursively.[2]

The following example illustrates it conveniently.

Example 4.1. Let the symmetric NC rational expressions $r(x)$ be given by

$$r(x_1, x_2) = (1 + x_1 - (3 + x_2)^{-1})^{-1}$$

with $x_1 = x_1^T$ and $x_2 = x_2^T$. The domain $\mathcal{F}_{r,\text{formal}}$ is

$$\bigcup_{n>0} \left\{ X_1, X_2 \in \mathbb{S}^n \,:\, I + X_1 - (3I + X_2)^{-1} \text{ and } 3I + X_2 \text{ are invertible} \right\}.$$

A difficulty is two different expressions, such as

$$r_1 = x_1(1 - x_2x_1)^{-1} \quad \text{and} \quad r_2 = (1 - x_1x_2)^{-1}x_1,$$

that can be converted into each other with algebraic manipulation. Thus they represent the same function and one needs to specify an equivalence relation on rational expressions to arrive at what are typically called NC rational functions. (This is standard and simple for commutative (ordinary) rational functions.) There are many alternate ways to describe NC rational functions and they go back 50 years or so in the algebra literature; cf. [17]. For engineering purposes, one need not be too concerned, since what happens is that two expressions r_1 and r_2 are equivalent whenever the usual manipulations one is accustomed to with matrix expressions convert r_1 to r_2. We say more on this in section 4.3.

[2]The formal domain of a polynomial p is all of $(\mathbb{S}^n)^g$ and $p(X)$ is defined as before. The formal domain of sums and products of rational expressions is the intersection of their respective formal domains. If r is an invertible rational expression analytic at 0 and $r(X)$ is invertible, then X is in the formal domain of r^{-1}.

For $\mathfrak{r}$ a rational function, that is, an "equivalence class of rational expressions r," we define its domain by

$$\mathcal{F}_{\mathfrak{r}} := \bigcup_{\{r \text{ represents } \mathfrak{r}\}} \mathcal{F}_{r,\text{formal}}.$$

Henceforth we do not distinguish between rational functions $\mathfrak{r}$ and rational expressions r, since this causes no confusion.

4.3. Partial fraction expansion of an NC rational. A computer algebra package must have a way to put functions into a canonical form. For example, if two rational expressions r_1 and r_2 represent the same rational function, then the canonical form of $r_1 - r_2$ would be 0. In NCAlgebra we have the command NCSimplifyRational, which in principle, when applied to a rational expression r, outputs what one might think of as a noncommutative partial fraction expansion of r; in practice, our command gives the true canonical form on a broad class of NC rational expressions but not all, since doing all of them is an infinite process. The theory behind producing this kind of canonical form is found in [11] and [24]. The idea is to generate what is called a Gröbner basis (GB) from the defining equations for inverses and store key elements of the GB as replacement rules in NCSimplifyRational. This is well suited to systems whose input operators B are left invertible, output operators C are right invertible, and state operators are generically invertible. Indeed they naturally lie in what is called a path algebra. It is not hard (for a GB expert) to prove that GB production respects the path algebra structure; thus, for example, the right inverse of B will never occur. See [10] for an extensive treatment of GBs in a path algebra.

4.4. Symbolic differentiation of noncommutative functions. Since our goal is to use symbolic computation to determine the gradient and the Hessian of functions in our optimization problems and to preserve the matrix structure of the unknowns, we need the notion of derivatives of function of variables which are symbolic noncommutative elements.

Noncommutative rational functions of x are polynomials in x and in inverses of polynomials in x. An example of a symmetric noncommutative function is

$$F(a, b, x) = ax + xa^T - \frac{3}{4}xbb^Tx, \quad x = x^T. \tag{4.1}$$

It is also assumed there is an involution on these rational functions which is denoted by the superscript ^{T} and which will play the role of transpose later when we substitute matrices for the indeterminates.

The *first directional derivative* of a noncommutative rational function $F(x)$ with respect to x in the direction δ_x is defined in the usual way

$$DF(x)[\delta_x] := \lim_{t\to 0} \frac{1}{t}\left(F(x + t\delta_x) - F(x)\right) = \frac{d}{dt}\, F(x + t\delta_x)\Big|_{t=0}.$$

For example, with F in (4.1),

$$DF(x)[\delta_x] = a\delta_x + \delta_x a^T - \frac{3}{4}\delta_x bb^T x - \frac{3}{4}xbb^T\delta_x,$$

and if $p(x) = x^4$,

$$Dp(x)[\delta_x] = \delta_x xxx + x\delta_x xx + xx\delta_x x + xxx\delta_x.$$

It is easy to check that derivatives of symmetric noncommutative rational functions always have the form

$$DF(x)[\delta_x] = \operatorname{sym}\left\{\sum_{i=1}^{k} a_i \delta_x b_i\right\}. \tag{4.2}$$

The sym operator is defined as $\operatorname{sym}\{M\} = M + M^T$.

The *second directional derivative* of a noncommutative rational function $F(x)$ with respect to x in the direction δ_x is defined by

$$D^2F(x)\left[\delta_x, \delta_x\right] = \frac{d^2}{dt^2} F(x + t\delta_x)\bigg|_{t=0}.$$

For example, if $p(x) = x^4$, then

$$D^2p(x)\left[\delta_x, \delta_x\right] = 2(\delta_x\delta_x xx + \delta_x x\delta_x x + \delta_x xx\delta_x + x\delta_x\delta_x x + x\delta_x x\delta_x + xx\delta_x\delta_x).$$

One can easily show that the second directional derivative of a symmetric noncommutative rational functions has the form

$$\begin{aligned} D^2F(x)\left[\delta_x, \delta_x\right] = \operatorname{sym}\Bigg\{ & \sum_{j=1}^{w_1} m_j \delta_x n_j \delta_x t_j \\ & + \sum_{j=1+w_1}^{w_2} m_j {\delta_x}^T n_j \delta_x t_j + \sum_{j=1+w_2}^{w_3} m_j \delta_x n_j {\delta_x}^T t_j \Bigg\}. \end{aligned} \tag{4.3}$$

For $r(x)$ given by $r(x_1, x_2) = (1 + x_1 - (3 + x_2)^{-1})^{-1}$, we have

$$Dr(x)[\delta_{x_1}] = -(1 + x_1 - (3 + x_2)^{-1})^{-1}\delta_{x_1}(1 + x_1 - (3 + x_2)^{-1})^{-1}$$

and

$$\begin{aligned} D^2r(x)\left[\delta_{x_1}, \delta_{x_1}\right] = 2(1 + x_1 - (3 + x_2)^{-1})^{-1} \cdots \\ \delta_{x_1}(1 + x_1 - (3 + x_2)^{-1})^{-1}\delta_{x_1}(1 + x_1 - (3 + x_2)^{-1})^{-1}. \end{aligned}$$

4.4.1. Symbolic NC differentiator algorithm. Derivatives of rational expressions can be defined recursively from the following rules:

1. If $r(x)$ is a polynomial, use the standard formula.
2. The product rule is, if $r(x) = r_1(x)r_2(x)$, then $Dr(x)[\delta_x] = Dr_1(x)[\delta_x]r_2(x) + r_1(x)Dr_2(x)[\delta_x]$.
3. The sum rule is, if $r(x) = r_1(x) + r_2(x)$, then $Dr(x)[\delta_x] = Dr_1(x)[\delta_x] + Dr_2(x)[\delta_x]$.
4. If $r(x)$ is the inverse $r(x) = f^{-1}(x)$ of an NC rational expression satisfying $f(0) \neq 0$, then $Dr(x)[\delta_x] := -f^{-1}(x)Df(x)[\delta_x]f^{-1}(x)$.

Our differentiation algorithm applies these rules (in a natural order) to an NC rational expression $r(x)$ and gives a new NC rational expression $Dr(x)[\delta_x]$, the directional derivative of the rational expression $r(x)$ in direction δ_x. Similarly, there are the natural formulas for the second directional derivative $D^2r(x)\left[\delta_x, \delta_x\right]$, for sums products and inverses. Our algorithm uses these recursively to compute our symbolic Hessian of $r(x)$.

4.5. Matrix convex functions. It will be shown that the definition just presented for the Hessian of a symmetric noncommutative rational function F is the key to determine the region of convexity for F. Therefore, it is the main ingredient of our NCConvexityRegion algorithm. There are several (almost equivalent) notions of noncommutative convexity; thus we define matrix convex functions as it is the definition used throughout the paper. For formal definitions, a detailed presentation, and an substantial theory behind the algorithm, see [5].

Let us suppose that F is the symmetric noncommutative rational function to be analyzed. Say F is a function of the noncommutative variables $x_1, \ldots, x_k$. Then, the function F is said to be matrix convex with respect to the variables $x_1, \ldots, x_k$ on a certain domain $\mathcal{G}$ provided its Hessian, denoted by $D^2F(X_1, \ldots, X_k)[\delta_{X_1}^2, \ldots, \delta_{X_k}^2]$, is a positive semidefinite matrix for all $X_1, \ldots, X_k$ in[3] $\mathcal{G}$ and all $\delta_{X_1}, \ldots, \delta_{X_k}$.

It is known (cf. [5]) that if $\mathcal{G}$ is a convex set, then this definition is equivalent to the usual notion of convexity, the geometrically matrix convex functions, which states that

$$F(\alpha X + (1-\alpha)Y) \leq \alpha F(X) + (1-\alpha)F(Y)$$

with $X := \{X_1, \ldots, X_k\}$ and $Y := \{Y_1, \ldots, Y_k\}$ tuples of matrices of compatible dimensions, and $0 \leq \alpha \leq 1$ a scalar. Of course, $\mathcal{G}$ might have separate components; then one often can focus on the component of primary interest with convexity on that component alone.

5. How the convexity checker algorithm works. With these notions of convexity, we now briefly introduce the algorithm underlying the command

NCConvexityRegion$[F, \{x\}]$

that provides a region $\mathcal{G}$ on which $F(X)$ is matrix convex. The main steps of the algorithm are as follows:

1. The second directional derivative with respect to $x_1, \ldots, x_k$, called the Hessian $D^2F[\delta_x, \delta_x]$ of the function F, is computed.
2. As the Hessian is always a quadratic function of the δ_x directions, it can be associated with a symmetric matrix $M_{D^2F[\delta_x, \delta_x]}$ with entries which are NC rational functions of x but not δ_x.
3. The noncommutative LDL$^{\mathrm{T}}$ factorization is applied to the coefficient matrix $M_{D^2F[\delta_x, \delta_x]}$.
4. And finally specifying positive definiteness of the resulting diagonal matrix $D(x_1, \ldots, x_k)$ gives inequalities describing a region $\mathcal{G}$ of variables on which F is matrix convex.
5. If a linear independence condition which is usually true holds and if the region $\mathcal{G}$ is nonempty for matrices of large enough size, then the closure of $\mathcal{G}$ is the largest domain on which $D^2F[\delta_x, \delta_x]$ is positive semidefinite. (See [5] for details on this rather complicated fact.)

Our implementation assumes that each pivot in the LDL$^{\mathrm{T}}$ decomposition is invertible. Possibly the most informative thing that could be done briefly is to give a simple example, presented in [5].

Example 5.1. Define the function $F(x)$ by

$$F(x) = g^T x^T a x g + x^T b x + g^T x^T c x + x^T c^T x g,$$

where $b = b^T$ and $a = a^T$.

[3] More precisely, for all X in the set $\mathcal{G}$ that intersect the domain of the rational function F.

1. The Hessian of $F(x)$ is given by
$$D^2F(X)\,[\delta_x,\delta_x] = 2({\delta_x}^T b\delta_x + {\delta_x}^T c^T\delta_x G + g^T{\delta_x}^T a\delta_x g + g^T{\delta_x}^T c\delta_x).$$
2. Equivalently, this quadratic expression takes the form
$$D^2F(X)\,[\delta_x,\delta_x] = V[\delta_x]^T M_{D^2F} V[\delta_x] = 2({\delta_x}^T, g^T{\delta_x}^T)\begin{pmatrix} b & c^T \\ c & a \end{pmatrix}\begin{pmatrix}\delta_x \\ \delta_x g\end{pmatrix}.$$
3. The LDL$^{\mathrm{T}}$ decomposition with no permutation applied to M_{D^2F} is
$$\begin{pmatrix} 1 & 0 \\ cb^{-1} & 1 \end{pmatrix}\begin{pmatrix} b & 0 \\ 0 & a - cb^{-1}c^T \end{pmatrix}\begin{pmatrix} 1 & b^{-1}c^T \\ 0 & 1 \end{pmatrix},$$
provided that b is invertible.[4]
4. Therefore, when b is invertible, sufficient conditions for the Hessian to be positive semidefinite are
$$b > 0 \quad \text{and} \quad a - cb^{-1}c^T > 0.$$
5. If the Hessian is "positive," then for large enough dimension, a, b, c are in the closure of the set described in step 4.

A finer property of our algorithm (and implementation) is that it includes the possibility of permutations. Thus, if we know that a (instead of b) in the example above is invertible, then a permutation can be applied before applying the LDL$^{\mathrm{T}}$ decomposition. This makes **NCConvexityRegion[]** somewhat delicate. In practice, the runs will finish with some choices of permutation and not with others. Also, the expressions that appear in the outputs from successful runs using different permutations can be different; however, the theory behind NCConvexityRegion tells us that the sets they describe are all the same.

Example 5.2. Let $p(x)$ be given by $p(x) = a^Tx^2b + b^Tx^2a$ with $x = x^T$. Its Hessian is $a^T{\delta_x}^2b + b^T{\delta_x}^2a$. Represent this as v^TMv with
$$v^T := \begin{pmatrix} a^T\delta_x & b^T\delta_x \end{pmatrix} \quad \text{and} \quad M = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$
whose eigenvalues are 1 and -1. Thus, we conclude from steps 4 and 5 that (for large enough n) there is no set $\mathcal{G}$ of matrices a and b, open in the set of pairs of $n \times n$ matrices, on which the Hessian is positive definite.

A related example is $p(x) = a^Tx^2a + a^Tx^2a$. Its Hessian is $a^T{\delta_x}^2a + a^T{\delta_x}^2a$; the representation above still works with $v^T := (a^T\delta_x, a^T\delta_x)$. However, the linear independence condition of step 4 on $v^T := (a^T\delta_x, a^T\delta_x)$ does not hold, that is, the condition $a^T\delta_x$ and $a^T\delta_x$ are linearly independent fails, and thus no definitive conclusion is possible. Another (more natural) representation is $a^T{\delta_x}^2a + a^T{\delta_x}^2a = 2a^T\delta_x 1\delta_x a$, that is, M is the 1×1 matrix 1. Thus step 4 implies, the Hessian of $p(x)$ is everywhere positive; note that in this case, linear independence holds. NCConvexityRegion does the later not the former calculation.

Our convexity region algorithm was new with [5], where it is described in detail, together with the proofs of its properties. The guarantee it produces for the nonnegativity of the NC Hessian (as in step 4) is straightforward to prove, while the converse (as in step 5) is quite surprising and not at all easy to prove.

[4]The list returned by **NCConvexityRegion** is $\{b, \quad a - cb^{-1}c^T\}$.

5.1. Noncommutativity is essential. A great advantage of our framework is that treating matrices as single letters is likely the only practical necessary and sufficient approach available for checking convexity of rational functions on matrices of large dimension. For a commutative rational function F, one might imagine a Parrilo type of sum of squares algorithm [22], which could affirm positivity of the Hessian of F, and thus convexity of F. Unfortunately, it would be practical with only a few dozen variables. If the unknown X and Y were symmetric matrices on even a 10-dimensional state-space, the entrywise representation would give about 100 commuting variables. This is prohibitive. On the other hand, our convexity checking method is insensitive to the dimension of the state-space.

6. Convex optimization over matrix functions. In this paper, the presentation of the numerical NCSDP optimization solver for matrix functions is limited to the single variable case. The extension to the multivariate case, found in [4], follows similar ideas, but it is too long to present here. This solver is based on an implementation of the method of centers.

There are in the literature few papers on solving nonlinear matrix inequalities. In [15], the authors presented and analyzed a numerical interior point trust region algorithm that can be used for solving a class of nonlinear (nonconvex) semidefinite programming problem. For MI problems that concern the minimization of the largest eigenvalue of a matrix (this is a convex but highly nonsmooth problem), the work by [21, 16] is a good source. See [1, 26, 6, 27, 28, 29] for general SDP problems and [3] for a comprehensive introduction to convex optimization.

The main distinctions between these approaches and ours is that our method focus on unknowns which themselves are treated as matrices. In our research, we deal with the entire matrix structure instead of dealing with the individual entries of the matrix unknowns. Also, the user does not need to calculate first and second derivatives by hand, since this is done automatically in our method.

The outline of our method is as follows. We compute the first and second derivatives of a potential function noncommutatively (symbolically) in a way that keeps the matrix structure and does not split up the matrices (see section 4.4). This step provides the Hessian map $\mathcal{H}(\delta_x)$ and the gradient map $\mathcal{Q}$. It is this step whose efficiency is improved by our MinimumSylvesterIndex algorithm for symbolically obtaining an efficient form for $\mathcal{H}(\delta_x)$ (this is described in section 8). After this, our algorithm turns numerical (by substituting matrices for the indeterminate that appears in the expressions for $\mathcal{H}(\delta_x)$ and $\mathcal{Q}$) and the code aims to solve the respective numerical linear system of equations $\mathbb{H}(\delta_X) = \mathbb{Q}$ in the direction δ_X. We find that the numerical linear subproblem has an elegant form, and it is an interesting open question how to fully exploit this form. We numerically solve this linear system for δ_X in a conventional way. The method successively iterates, at the numerical level, until the algorithm converges to an optimal solution.

As described above, one needs to compute derivatives of an auxiliary potential function. The formula for this potential function depends on which member of the family of penalty/barrier methods one wishes to adhere to. The approach we have selected is known as the *analytic method of centers*. Before describing this method in section 6.2, we characterize the optimization problem we are interested in section 6.1.

6.1. Constrained optimization problem. Throughout the paper, we shall be primarily concerned with the convex optimization problem

$$\text{(COP)} \qquad f^{\text{opt}} = \min \{\text{Tr}\,\{X\} \,:\, X \in \text{closure}(\mathcal{G})\}$$

with the feasibility domain $\mathcal{G}$ given by

$$\mathcal{G} = \Big\{ X \in \mathcal{V} \,:\, F(X) > 0 \Big\},$$

where we assume the closure of $\mathcal{G}$ is compact, the set $\mathcal{V}$ is a subspace of $\mathbb{R}^{p\times q}$, and $F : \mathcal{V} \to \mathbb{S}^n$ is a concave function. This type of problem incorporates the eigenvalue minimization problem as a particular case.

6.2. Review of the method of centers. The idea behind the method of centers ([2, 19] and references therein) is to replace the above constrained problem by a sequence of unconstrained minimization problems whose solutions eventually tend to the set of optimal solutions of (COP). This occurs in the context of *interior penalty methods*. It follows therefore, that under certain hypotheses, the original problem (COP) can be approximated by a sequence of unconstrained convex optimization problems of the form

$$X^*(\gamma) = \operatorname{argmin} \{ \phi_\gamma(X) \,:\, X \in \mathcal{G}_\gamma \} \tag{UOP}$$

with the auxiliary potential function $\phi_\gamma : \mathcal{G}_\gamma \to \mathbb{R}$ given by

$$\phi_\gamma(X) = \zeta \log \left(1/(\gamma - \operatorname{Tr}\{X\})\right) - \log\det F(X),$$

where ζ is a scalar satisfying $\zeta \geq 1$ and $\mathcal{G}_\gamma$ is the domain given by

$$\mathcal{G}_\gamma = \Big\{ X \in \mathcal{G} \,:\, \operatorname{Tr}\{X\} < \gamma \Big\}.$$

The decrease of the parameter γ has to be done in such a way that the method maintains feasibility at each iteration and that the sequence $\{\gamma^k\}$ is guaranteed to converge to f^{opt} (the minimum values of the objective function). The formula for updating γ at some iteration k is given by

$$\gamma^{k+1} = (1-\theta)\operatorname{Tr}\left\{X^k\right\} + \theta\gamma^k, \quad 0 < \theta < 1, \tag{6.1}$$

where X^k denotes $X^*(\gamma^k)$. Under mild conditions, the solution $X^*(\gamma)$ of (UOP) approaches the set of optimal solutions of (COP) for an appropriate sequence of decreasing centralization parameter γ (see [2, 7, 19]).

Using these facts, one possible algorithm based on the method of centers can be described by the following algorithm.

Algorithm 6.1. Method of centers.
Fix θ such that $0 < \theta < 1$;
Choose X^0 and γ^0 such that $X^0 \in \mathcal{G}_{\gamma^0}$;
$k \leftarrow 0$;
while not converged **do**
  $\gamma^{k+1} \leftarrow (1-\theta)\operatorname{Tr}\left\{X^k\right\} + \theta\gamma^k$;
  $X^{k+1} \leftarrow \operatorname{argmin}\left\{\phi_{\gamma^{k+1}}(X^k) \,:\, X^k \in \mathcal{G}_{\gamma^{k+1}}\right\}$;
  $k \leftarrow k+1$;
end while

There are some important comments concerning this algorithm:

1. The bound γ^{k+1} from (6.1), used in the determination of the analytic center of the potential $\phi_{\gamma^{k+1}}(X^k)$, never produces infeasible starting points X^k, γ^{k+1}.

2. It will be necessary to find feasible starting points X^0 and γ^0 to be used in Algorithm 6.1. This is a *feasibility problem* that can be solved by the same method of centers.
3. Evidently, the expensive part of the algorithm is the *inner loop*, the part that computes the *analytic center*.

6.3. An algorithm to solve the inner loop. This section sketches briefly a standard algorithm to solve the inner loop. The algorithm implemented in the NCSDP code to find the analytic center

$$X^{k+1} = \operatorname{argmin} \left\{ \phi_\gamma(X^k) \,:\, X^k \in \mathcal{G}_\gamma \right\}$$

for fixed scalar γ is based on a conventional modified Newton's method [19, 3, 20], as follows:

while not converged **do**
$X^{k+1} \leftarrow X^k + \sigma {\delta_X}^k$;
end while

where ${\delta_X}^k$ is the Newton direction (see section 7.1). The step length used in this algorithm [19] is given by

$$\sigma = \begin{cases} 1/(1+\tau) & \text{if } \tau > 1/4, \\ 1 & \text{otherwise} \end{cases}$$

for $\tau = \sqrt{g^T H^{-1} g}$ with g and H, respectively, the gradient and the Hessian of $\phi_\gamma(X^k)$. The stopping criteria used in our experiments was, stop as soon as $\sigma = 1$. In practice Newton's method works better with a line search instead of the above fixed step length. Certainly, we shall consider implementing a line search in a more elaborate version of the code.

7. Solving for the analytic center. In sections 7.1, 7.2, and 7.3 we discuss in depth the linear subproblem that provides the update direction δ_X, which is the core of the modified Newton's algorithm presented in section 6.3. Therefore, in these sections we show how to exploit the matrix structure of the unknowns to find an elegant formula for the linear subproblem.

7.1. Describing the main steps. The original convex optimization problem (COP) has now been replaced by a sequence of unconstrained convex minimization problems of the form (UOP) for a decreasing sequence of scalars $\{\gamma^k\}$ provided by formula (6.1). To find the update directions which lead toward the central path for fixed values of γ, Newton's method is applied by minimizing an approximation, the second-order Taylor series expansion, of the potential function $\phi_\gamma(X)$. In a vague sense, these procedures can be summarized as follows:

1. Compute symbolically the second-order Taylor expansion of the potential function $\phi_\gamma(x + \delta_x)$ in some direction δ_x

$$\phi_\gamma(x) + D\phi_\gamma(x)\,[\delta_x] + \frac{1}{2} D^2\phi_\gamma(x)\,[\delta_x, \delta_x]\,.$$

2. The Newton step ${\delta_x}^*$ must satisfy the necessary optimality conditions for the following quadratic minimization problem:

$$\min_{\delta_x} \; D\phi_\gamma(x)\,[\delta_x] + \frac{1}{2} D^2\phi_\gamma(x)\,[\delta_x, \delta_x]\,.$$

3. This first-order necessary optimality condition is algebraically[5] given by

$$0 = D\left[D\phi_\gamma(x)\,[\delta_x] + \frac{1}{2}D^2\phi_\gamma(x)\,[\delta_x,\delta_x]\right]\Big[\delta_v\Big] \quad \text{for all symmetric } \delta_v. \tag{7.1}$$

Which will be shown (in Theorem 7.1) to be equivalently[6] written as

$$\mathrm{Tr}\left\{\delta_v(\mathcal{H}(\delta_x)-\mathcal{Q})^T + (\mathcal{H}(\delta_x)-\mathcal{Q}){\delta_v}^T\right\} = 0 \quad \text{for all symmetric } \delta_v. \tag{7.2}$$

4. Finally, find a Newton update ${\delta_X}^*$ satisfying (7.1) or (7.2) for all δ_v.

Section 7.2 concerns steps 1 through 3, which are performed symbolically. On the other hand, step 4, presented in section 7.3, is completely numerical.

7.2. Obtaining the formulas for the linear subproblem. The main ingredient of our approach is how we use symbolic computation to determine the algebraic linear system of equations that provides the update direction δ_X toward the central paths. At the outset of this work, it was not obvious that we could find a clean symbolic formula for the linear subproblem which treated both known and unknown matrices as a whole and did not break them into entries. Fortunately, this is possible, as the next theorem shows.

THEOREM 7.1. *Let $\mathcal{V}$ be a subspace of $\mathbb{R}^{p\times q}$, and let the map $F : \mathcal{V} \to \mathbb{S}$ be concave. Consider the unconstrained auxiliary potential function $\phi_\gamma : \mathcal{G}_\gamma \to \mathbb{R}$ given by*

$$\phi_\gamma(X) = \zeta \log\Big(1/(\gamma - \mathrm{Tr}\,\{X\})\Big) - \log\det F(X),$$

where ζ is a scalar satisfying $\zeta \geq 1$ and the feasibility domains $\mathcal{G}$ and $\mathcal{G}_\gamma$ are respectively given by

$$\mathcal{G} = \Big\{X \in \mathcal{V} \,:\, F(X) > 0\Big\} \quad \textit{and} \quad \mathcal{G}_\gamma = \Big\{X \in \mathcal{G} \,:\, \mathrm{Tr}\,\{X\} < \gamma\Big\},$$

where we assume the closure of $\mathcal{G}$ is compact. Then, the update direction ${\delta_X}^*$ toward the central path for the above potential is the solution of the following symbolically[7] computable algebraic linear equation:

$$\mathrm{Tr}\left\{\delta_v(\mathcal{H}(\delta_x)-\mathcal{Q})^T + (\mathcal{H}(\delta_x)-\mathcal{Q}){\delta_v}^T\right\} = 0 \quad \textit{for all } \delta_v \in \mathcal{V}, \tag{7.3}$$

where $\mathcal{H}(\delta_x)$ is linear as regarded as a function of δ_x. Moreover, $\mathcal{Q}$ and $\mathcal{H}(\delta_x)$ are given by

$$\mathcal{Q} = \sum_{i=1}^{k} a_i^T F(x)^{-1} b_i^T - \frac{1}{2}\zeta\,(\gamma - \mathrm{Tr}\,\{x\})^{-1}\,\mathfrak{I}_d$$

[5] Assuming that all x belong to some space $\mathcal{V}$, then $\delta_x, \delta_v \in \mathcal{V}$. Cf. footnotes 6, 7.

[6] When we say $\mathrm{Tr}\,\{x\}$, we mean that the operation $\mathrm{Tr}\,\{\}$ is to be performed after x has been replaced by a matrix. Cf. footnote 7.

[7] When we say $\delta_v \in \mathcal{V}$, we mean that we will substitute matrices in $\mathcal{V}$ for the indeterminate δ_v. The same is true for $\mathrm{Tr}\,\{x\}$. Cf. footnote 6.

and

$$\begin{aligned}\mathcal{H}(\delta_x) = &\sum_{i=1}^{k}\sum_{j=1}^{k} a_i^T F(x)^{-1} a_j \delta_x b_j F(x)^{-1} b_i^T + \sum_{i=1}^{k}\sum_{j=1}^{k} a_i^T F(x)^{-1} b_j^T \delta_x^T a_j^T F(x)^{-1} b_i^T \\ &- \frac{1}{2}\sum_{j=1}^{w_1} n_j^T \delta_x^T m_j^T F(x)^{-1} t_j^T + m_j^T F(x)^{-1} t_j^T \delta_x^T n_j^T \\ &- \frac{1}{2}\sum_{j=1+w_1}^{w_2} n_j \delta_x t_j F(x)^{-1} m_j + n_j^T \delta_x m_j^T F(x)^{-1} t_j^T \\ &- \frac{1}{2}\sum_{j=1+w_2}^{w_3} t_j F(x)^{-1} m_j \delta_x n_j + m_j^T F(x)^{-1} t_j^T \delta_x n_j^T \\ &+ \frac{1}{2}\zeta \left(\gamma - \operatorname{Tr}\{x\}\right)^{-2} \operatorname{Tr}\{\delta_x\}\, \mathfrak{I}_d,\end{aligned}$$

where the terms a_i, b_i, m_i, n_i, t_i are obtained from the first and second directional derivatives of $F(x)$ as given by (4.2) *and* (4.3). *The term $\mathfrak{I}_d$ stands for the symbolic analogue of the identity matrix.*

Proof. The theorem follows from manipulations of (7.1). For a detailed presentation see [4]. □

The result of Theorem 7.1, the algebraic linear equation (7.3), can be further specialized depending upon the structure of the underlying subspace $\mathcal{V}$; in other words, if there is or is not some restriction imposed on X. Specifying various structures for the underlying subspace $\mathcal{V}$ is the subject of Corollary 7.2 which is the main result of this section.

COROLLARY 7.2. *Let $\mathcal{V}$ be a subspace of $\mathbb{R}^{p\times q}$ and $\mathcal{C}$ be a convex domain in $\mathcal{V}$. Let the map $F : \mathcal{C} \to \mathbb{S}$ be concave. Consider the unconstrained auxiliary potential function $\phi_\gamma : \mathcal{G}_\gamma \to \mathbb{R}$ given by*

$$\phi_\gamma(X) = \zeta \log\Big(1/(\gamma - \operatorname{Tr}\{X\})\Big) - \log\det F(X),$$

where ζ is a scalar satisfying $\zeta \geq 1$ and the feasibility domains $\mathcal{G}$ and $\mathcal{G}_\gamma$ are respectively given by

$$\mathcal{G} = \Big\{X \in \mathcal{C} \subset \mathcal{V} \,:\, F(X) > 0\Big\} \quad \textit{and} \quad \mathcal{G}_\gamma = \Big\{X \in \mathcal{G} \,:\, \operatorname{Tr}\{X\} < \gamma\Big\}.$$

Then, depending upon the structure of the underlying subspace $\mathcal{V}$, the update direction $\delta_x{}^*$ toward the central path for the above potential is the solution of one of the following symbolically computable algebraic linear equations:

1. *The subspace $\mathcal{V}$ equals $\mathbb{R}^{p\times q}$ so that the unknown X can be any matrix in $\mathbb{R}^{p\times q}$:*

$$\sum_{i=1}^{c_1} \mathfrak{a}_i \delta_x \mathfrak{b}_i + \sum_{j=c_1+1}^{c_2} \mathfrak{a}_j \delta_x^T \mathfrak{b}_j + \varrho \operatorname{Tr}\{\delta_x\} = \mathfrak{Q}.$$

2. *The subspace $\mathcal{V}$ equals $\mathbb{S}^p$ so that the unknown X is restricted to being symmetric:*

$$\sum_{i=1}^{c_2} \mathfrak{b}_i^T \delta_x \mathfrak{a}_i^T + \mathfrak{a}_i \delta_x \mathfrak{b}_i + \varrho \operatorname{Tr}\{\delta_x\} = \mathfrak{Q} + \mathfrak{Q}^T.$$

3. *The unknown X is restricted to being a scalar multiple of the identity, that is, $X = \sigma I$, for some scalar σ:*

$$\mathrm{Tr}\left\{\sum_{i=1}^{c_2} \mathfrak{a}_i \mathfrak{b}_i + \varrho \mathrm{Tr}\left\{\mathfrak{I}_d\right\}\right\} \delta_\sigma = \mathrm{Tr}\left\{\mathfrak{Q}\right\}, \qquad \delta_x = \delta_\sigma \mathfrak{I}_d, \quad \delta_\sigma \in \mathbb{R}.$$

For these expressions, $\mathfrak{Q}$ is the gradient term given by

$$\mathfrak{Q} = \sum_{i=1}^{k} a_i^T F(x)^{-1} b_i^T - \frac{1}{2}\zeta \left(\gamma - \mathrm{Tr}\left\{x\right\}\right)^{-1} \mathfrak{I}_d.$$

The term ϱ is the cost term given by $\varrho = \frac{1}{2}\zeta(\gamma - \mathrm{Tr}\left\{x\right\})^{-2}\mathfrak{I}_d$. And, by an appropriate relabeling, the terms $\mathfrak{a}_i$ and $\mathfrak{b}_i$ are obtained from the Hessian map $\mathcal{H}(\delta_x)$ presented in Theorem 7.1*.*

Proof. The corollary follows from Theorem 7.1 by expressing the linear system of equations (7.3) considering the structure of the underlying subspace $\mathcal{V}$. (See [4].) ∎

The above results provide the necessary conditions that the update δ_X must satisfy in order to be a Newton direction toward the central path of the unconstrained auxiliary potential function $\phi_\gamma(X)$.

We provide a tutorial example in appendix A to illustrate Theorem 7.1 and Corollary 7.2 and to give an idea of how they are proved.

7.3. Solving the linear subproblem. The algebraic linear subproblem[8] provided in Corollary 7.2 always has the form

$$\sum_{i}^{N} \mathfrak{a}_i \delta_x \mathfrak{b}_i + \varrho\, \mathrm{Tr}\left\{\delta_x\right\} = \bar{\mathfrak{Q}}, \tag{7.4}$$

where the $\mathfrak{a}_i$, $\mathfrak{b}_i$, and $\bar{\mathfrak{Q}}$ are rational functions of the known noncommutative variables given in the problem formulation and δ_x is the unknown variable (the update direction). The notation $\mathfrak{a}_i, \mathfrak{b}_i$ stands for symbolic Sylvester terms, and the notation $\mathcal{A}_i, \mathcal{B}_i$ will indicate we have substituted matrices of compatible size for the symbolic variables in $\mathfrak{a}_i, \mathfrak{b}_i$.

The integer N has been called the Sylvester index in [14]. A key point is that the same linear system can have several representations of the form (7.4), that is, the representation of the Hessian map $\mathcal{H}(\delta_x)$ in Theorem 7.1 is not unique. We will see later in section 8 that there is a substantial advantage to obtaining a representation with a small Sylvester index.

There are two main costs in treating the linear subproblem (7.4):

FE: evaluating the matrices $\mathfrak{a}_i$, $\mathfrak{b}_i$, and $\bar{\mathfrak{Q}}$ at each iteration, that is, converting them from symbols to numeric matrices whose entries are numbers is time-consuming (see section 8);

NLS: solving numerically the resulting linear system for δ_X.

[8] (a) Depending upon the structure of the underlying subspace $\mathcal{V}$, the term $\bar{\mathfrak{Q}}$ will be either $\bar{\mathfrak{Q}} = \mathfrak{Q}$ or $\bar{\mathfrak{Q}} = \mathfrak{Q} + \mathfrak{Q}^T$. (b) The third case in Corollary 7.2 behaves in a similar way, so we do not go through it.

After the evaluation step FE has been performed, we rewrite the linear subproblem (7.4) as

$$\sum_{i}^{N} \mathcal{A}_i \delta_X \mathcal{B}_i + \varrho \operatorname{Tr}\{\delta_X\} = \mathbb{Q}, \tag{7.5}$$

indicating that the indeterminate have already been substituted by matrices of compatible dimension. Then, using the vec operation (see [12]), the matrix system (7.5) can be transformed into the equivalent vector form

$$Hv = g, \tag{7.6}$$

where H is the Hessian matrix given by

$$H = \sum_{i}^{N} \mathcal{B}_i^T \otimes \mathcal{A}_i + \operatorname{vec}(\varrho)\operatorname{vec}(I)^T,$$

where the vector g is given by $g = \operatorname{vec}(\mathbb{Q})$, and v is the vector of unknowns given by $v = \operatorname{vec}(\delta_X)$. The symbol $\otimes$ denotes the Kronecker product.

Therefore, the cost of numerically solving the linear subproblem can be split into two distinct costs:

KP: applying Kronecker products to build the Hessian matrix H;

LS: numerically solving $Hv = g$ for the unknown vector v.

The above "brute force" procedure does not take advantage of the particular structure of $\mathcal{H}(\delta_x)$. Of course, Lyapunov equations are very special cases for which there are extremely fast algorithms (see [9, 14]). Naturally, an open question highly motivated by this research is how one uses this special Sylvester structure to solve efficiently (7.5).

Iterative methods are attractive for solving Sylvester-type linear equations. Related to this is [13] and references therein. However, in our paper, we do not investigate numerical linear solvers special to Sylvester forms. It is a separate topic and our focus was on our new noncommutative symbolic methodology. Consequently, we just used our brute force Kronecker product approach since it is reliable. However, in order to speed up the implementation of our linear solver, we plan a careful study of iterative methods like conjugate gradient in a separate project.

8. Improving the evaluation time for the linear subproblem. In this section, we illustrate by examples that for a system of linear equations, the Sylvester form (7.4) is not unique. Moreover, we show that the Sylvester index has a great influence on the evaluation cost given in Step FE of section 7.3.

Consider an expression in the Sylvester form

$$\mathcal{H}(\delta_x) = a\delta_x a^T + x^T\delta_x x + b\delta_x b^T - a\delta_x x - x^T\delta_x a^T + b\delta_x a^T + a\delta_x b^T. \tag{8.1}$$

The Sylvester index in this case is seven. This expression can be written in at least two different ways, having the same number of terms. One possibility is

$$\mathcal{H}(\delta_x) = (a - x^T)\delta_x(a - x^T)^T + (a + b)\delta_x(a + b)^T - a\delta_x a^T = \sum_{i=1}^{N=3} \mathfrak{a}_i \delta_x \mathfrak{b}_i$$

for $\mathfrak{a}_i$ and $\mathfrak{b}_i$ given by

$$\begin{aligned} \mathfrak{a}_1 &= (a - x^T), \qquad \mathfrak{a}_2 = (a + b), \qquad \mathfrak{a}_3 = -a, \\ \mathfrak{b}_1 &= (a - x^T)^T, \quad \mathfrak{b}_2 = (a + b)^T, \quad \mathfrak{b}_3 = a^T. \end{aligned}$$

Another one is

$$\mathcal{H}(\delta_x) = (a + b - x^T)\delta_x(a + b - x^T)^T + b\delta_x x + x^T\delta_x b^T = \sum_{i=1}^{N=3} \mathfrak{a}_i\delta_x\mathfrak{b}_i$$

for $\mathfrak{a}_i$ and $\mathfrak{b}_i$ given by

$$\begin{aligned} \mathfrak{a}_1 &= (a + b - x^T), \qquad \mathfrak{a}_2 = b, \qquad \mathfrak{a}_3 = x^T, \\ \mathfrak{b}_1 &= (a + b - x^T)^T, \quad \mathfrak{b}_2 = x, \qquad \mathfrak{b}_3 = b^T. \end{aligned}$$

In both cases, the Sylvester index is now three, going down by over one half. Thus, for a given Hessian map $\mathcal{H}(\delta_x)$, the Sylvester index is not unique. Moreover, the $\mathcal{H}(\delta_x)$ may have different representations for a specific Sylvester index (as illustrated above). It is also easy to see that a significant reduction in the Sylvester index might happen for an expression which contains a large number of Sylvester terms. Based on those ideas, a few natural questions can be formulated:

1. Given an expression for the Hessian map $\mathcal{H}(\delta_x)$, what is the minimum Sylvester index associated with this expression?
2. Is there a symbolic algorithm to compute a minimum Sylvester index representation?
3. How many different expressions which achieve this minimal Sylvester index are possible?
4. Does the evaluation time in Step FE vary substantially for small versus large Sylvester index N?

This section addresses the first two questions. We describe preliminarily a symbolic algorithm which is fast and which often reduces the Sylvester index N dramatically. Later, we describe a more powerful (but slower) symbolic algorithm which gives the minimal Sylvester index when the coefficients $\mathfrak{a}_j$ and $\mathfrak{b}_j$ are polynomials. For our problems, the coefficients are not polynomials, but this algorithm applies with no restrictions. However, we can no longer guarantee that we obtain the minimal Sylvester index.

As to question 4, we have found through examples (see section 8.3.1) that the overall computational time spent on numerically solving an optimization problem using our NCSDP code dramatically reduces when the Sylvester index of the Hessian map $\mathcal{H}(\delta_x)$ is reduced by one of these two algorithms. However, we should consider the time consumed at the symbolic level by the algorithm itself. We found that the first algorithm to be presented is faster than the second algorithm. (The second provides the minimal Sylvester index).

8.1. A Sylvester index reducing algorithm. We now describe our first Sylvester index reducing algorithm, which is denoted by

NCCollectSylvester[exp, var].

The implementation used in our NCSDP optimization code is a command that sequentially applies two commands, called NCRightSylvester[] and NCLeftSylvester[], to the expression. These two "sided" commands have analogous implementation,

which uses a pattern match that collects similar terms on the right (respectively, on the left) side of the expression. We now present the idea behind these commands.

ALGORITHM 8.1 (NCRIGHTSYLVESTER ALGORITHM).

1. IDENTIFY THE TERMS IN WHICH THE EXPRESSION SHOULD BE COLLECTED. In the example given by expression (8.1), this term is δ_x.
2. BUILD A RIGHT LIST. This list contains the terms that multiplies δ_x from the right side (including δ_x itself). For the expression (8.1), we would obtain
$$\text{RightList} = \{\delta_x a^T,\ \delta_x x,\ \delta_x b^T\}.$$
3. BUILD A COLLECTLIST. For each element inside RightList, we add together all the terms that multiply this element from the left side. For our example we obtain
$$\text{CollectList} = \{(a + b - x^T),\ (x^T - a),\ (a + b)\}.$$
4. COMBINE THE COLLECTLIST AND THE RIGHTLIST. This gives the answer. For our example it is
$$\mathcal{H}(\delta_x) = (a + b - x^T)\delta_x a^T + (x^T - a)\delta_x x + (a + b)\delta_x b^T.$$

The above right-sided implementation of the collecting algorithm begins by building a list of multipliers from the right side of δ_x. Clearly, a similar implementation can also be done by obtaining a left list of terms that multiplies δ_x from the left side, instead of the right side. In this way, we can implement two collect commands that differ only by the side in which the process of collecting begins; thus, we can have an **NCRightSylvester[]** command (described above) and an **NCLeftSylvester[]** command. As already mentioned, our implementation encompasses these two commands into a single command

NCCollectSylvester[exp, var]
:= NCRightSylvester[NCLeftSylvester[exp, vars], vars].

These algorithms, when applied to an expression in the Sylvester form, in practice provide a large reduction on the Sylvester index. However, these algorithms do not guarantee that one can obtain the lowest possible Sylvester index. On the other hand, in the next section, we provide an algorithm which under some hypothesis provides the lowest possible Sylvester index.

8.2. The minimum Sylvester index algorithm. Consider a function $L(\delta)$ of a noncommutative variable δ in the Sylvester form

$$L(\delta) := \sum_{j=1}^{N} \mathfrak{a}_j \, \delta \, \mathfrak{b}_j,$$

with the terms $\mathfrak{a}_j$ and $\mathfrak{b}_j$ polynomials in noncommutative variables. The algorithm proposed in this section has property that if the $\mathfrak{a}_j$ and $\mathfrak{b}_j$ are restricted to be polynomials, then it always gives the lowest possible Sylvester index. This fact is presented in Theorem 8.3. We now present the steps of the algorithm and give an example. For

this purpose, consider the following expression:

$$\begin{aligned}(8.2)\quad L(\delta) &= (xb+axb)\,\delta\,(-2axb+bxb)+(xb+axb)\,\delta\,(xb-axb+bxb)\\ &-(xb+axb)\,\delta\,(xb-axb+2xax)+(xb+axb)\,\delta\,(xb-2axb+bxb+xax)\\ &+(c-xb-bxb+xax)\,\delta\,(-axb+xax)+(c+axb-bxb+xax)\,\delta\,(xb+xax)\\ &+(c+xb+2axb-bxb+xax)\,\delta\,(2xb+bxb).\end{aligned}$$

The Sylvester index associated with this expression is $N = 7$. Using our algorithm, we will see that the minimum Sylvester index is $N^* = 2$.

Algorithm 8.2 (MinimumSylvesterIndex Algorithm).

1. Identify the terms in which the expression should be collected. In the example given by expression (8.2), this term is δ.
2. Build a right list. This list contains the terms that multiplies δ from the right side in $L(\delta)$. Denote this list by $\mathfrak{b}$. For our example, this list is

$$\begin{aligned}\mathfrak{b} = \{&(-2axb+bxb),\ (xb-axb+bxb),\ (xb-axb+2xax),\\ &(xb-2axb+bxb+xax),\ (-axb+xax),\ (xb+xax),\ (2xb+bxb)\}.\end{aligned}$$

3. Build a monomial list. This list contains the terms that appears in $\mathfrak{b}$. This list, denoted by $\mathfrak{m}$, contains only monomial that are linearly independents:

$$\mathfrak{m} = \{xb,\ axb,\ bxb,\ xax\}.$$

4. Find a matrix G such that $\mathfrak{b} = G\mathfrak{m}$. For our example, G is given by

$$G^T = \begin{bmatrix} 1 & 0 & 1 & 1 & 0 & 1 & 2\\ -1 & -2 & -1 & -2 & -1 & 0 & 0\\ 0 & 1 & 1 & 1 & 0 & 0 & 1\\ 2 & 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix}.$$

5. Build a left list. This list contains the terms that multiplies δ from the left side in $L(\delta)$. Denote this list by $\mathfrak{a}$:

$$\begin{aligned}\mathfrak{a} = \{&(xb+axb),\ (xb+axb),\ (xb+axb),\ -(xb+axb),\\ &(c-xb-bxb+xax),\ (c+axb-bxb+xax),\ (c+xb+2axb-bxb+xax)\}.\end{aligned}$$

6. The expression $L(\delta)$ can now be rewritten as

$$L(\delta) = \sum_{j=1}^{4} \mathfrak{c}_j\,\delta\,\mathfrak{m}_j$$

with $\mathfrak{c} = G^T\mathfrak{a}$ given by

$$\begin{aligned}\mathfrak{c} = \{&3(c+xb+2axb-bxb+xax),\ (-c-3xb-4axb+bxb-xax),\\ &(c+4xb+5axb-bxb+xax),\ 2(c-xb-bxb+xax)\}.\end{aligned}$$

7. Build a monomial list from $\mathfrak{c}$. For the above example:

$$\bar{\mathfrak{m}} = \{c,\ xb,\ axb,\ bxb,\ xax\}.$$

8. Find a transformation matrix $\bar{G}$ such that $\mathfrak{c} = \bar{G}\bar{\mathfrak{m}}$:

$$\bar{G} = \begin{bmatrix} 3 & 3 & 6 & -3 & 3 \\ -1 & -3 & -4 & 1 & -1 \\ 1 & 4 & 5 & -1 & 1 \\ 2 & -2 & 0 & -2 & 2 \end{bmatrix}.$$

9. Decompose matrix $\bar{G}$ as $\bar{G} = QR$, with Q and R full rank matrices:

$$Q^T = \frac{1}{5}\begin{bmatrix} 15 & -5 & 5 & 10 \\ 3 & -11 & 16 & -18 \end{bmatrix}, \quad R = \frac{1}{5}\begin{bmatrix} 5 & 4 & 9 & -5 & 5 \\ 0 & 5 & 5 & 0 & 0 \end{bmatrix}.$$

10. The minimal Sylvester index N^* is the rank of $\bar{G}$. Thus, the final expression is

$$L(\delta) = \sum_j^{N^*} [R\bar{\mathfrak{m}}]_j \, \delta \, [Q^T \mathfrak{m}]_j \,. \tag{8.3}$$

For our example (8.2), the result is

$$L(\delta) = \frac{1}{5}(5\,c + 4\,xb + 9\,axb - 5\,bxb + 5\,xax)\,\delta\,(3\,xb - axb + bxb + 2\,xax) + \frac{1}{5}(xb + axb)\,\delta\,(3\,xb - 11\,axb + 16\,bxb - 18\,xax)$$

with the minimum Sylvester index guaranteed to be $N^* = 2$.

The implementation of our **MinimumSylvesterIndex**$[L(\delta)$**,** $\delta]$ command is described by these steps. When the original expression $L(\delta)$ contains a large number of Sylvester terms, the time spent on generating the matrix G in step 4 might be long. However, we emphasize that the expression $L(\delta)$ provided in step 6 can alternatively be provided by **NCRightSylvester[],** which is significantly faster than steps 1 through 6. In fact, this is how the MinimumSylvesterIndex command was implemented.

Theorem 8.3. *Provided that the $\mathfrak{a}_j$ and $\mathfrak{b}_j$ are polynomials, the lowest Sylvester index for $L(\delta)$ is given by N^*, which is the dimension of the span of $\mathfrak{c}_j$ for $j = 1, \ldots, d_{\mathfrak{b}}$, i.e., the rank of $\bar{G}$.*

Proof. This theorem follows immediately from Lemmas 8.4 and 8.6. □

Lemma 8.4. *The representation* (8.3) *produced by the MinimumSylvesterIndex algorithm has the property that the polynomials $[R\bar{\mathfrak{m}}]_1, \ldots, [R\bar{\mathfrak{m}}]_{N^*}$ are linearly independent and that the polynomials $[Q^T\mathfrak{m}]_1, \ldots, [Q^T\mathfrak{m}]_{N^*}$ are also linearly independent.*

Proof. Since the vectors $\mathfrak{m}_j$ are linearly independent and Q has full rank, the vectors $[Q^T\mathfrak{m}]_j$ for $j = 1, \ldots, N^*$ are linearly independent. Similarly, since the vectors $\bar{\mathfrak{m}}_j$ are linearly independent and R has full rank, the vectors $[R\bar{\mathfrak{m}}]_j$ for $j = 1, \ldots, N^*$ are linearly independent. □

Definition 8.5. *We call a dependence free Sylvester representation any Sylvester expression $L(\delta) = \sum \mathfrak{a}_j \, \delta \, \mathfrak{b}_j$ with $\mathfrak{a}_j$ linearly independent and $\mathfrak{b}_j$ also linearly independent.*

Lemma 8.6. *Let $L(\delta)$ and $\widetilde{L}(\delta)$ be Sylvester representations such that*

$$L(\delta) := \sum_{j=1}^{N} \mathfrak{a}_j \, \delta \, \mathfrak{b}_j = \sum_{k=1}^{\widetilde{N}} \widetilde{\mathfrak{a}}_k \, \delta \, \widetilde{\mathfrak{b}}_k =: \widetilde{L}(\delta).$$

If the polynomials $\mathfrak{a}_j$ are linearly independent, and if the polynomials $\mathfrak{b}_j$ are linearly independent, then for each $k = 1, \ldots, \widetilde{N}$ we have

$$\mathfrak{a}_j \in \operatorname{span}\{\widetilde{\mathfrak{a}}_k\}_1^{\widetilde{N}} \qquad \textit{and} \qquad \mathfrak{b}_j \in \operatorname{span}\left\{\widetilde{\mathfrak{b}}_k\right\}_1^{\widetilde{N}}.$$

Consequently $N \leq \widetilde{N}$, and if $\widetilde{L}(\delta)$ is also a dependence free Sylvester representation, then their Sylvester indexes are the same, $\widetilde{N} = N$.

Proof. Let β denote the maximum of the degrees of all of the polynomials $\mathfrak{a}_j, \mathfrak{b}_j, \widetilde{\mathfrak{a}}_k, \widetilde{\mathfrak{b}}_k$ for $j = 1, \ldots, N$ and $k = 1, \ldots, \widetilde{N}$. Let $\mathcal{P}(y)$ denote the space of all polynomials of degree less than or equal to β in $y = \{y_1, \ldots, y_g\}$. Let $\mathcal{P}(y)\delta$ denote all polynomials in the variables $\{y_1, y_2, \ldots, y_g, \delta\}$ of the form $p(y)\delta$ for $p \in \mathcal{P}(y)$. Since $\{\mathfrak{b}_j\}$ for $j = 1, \ldots, N$ is a linearly independent subset of the finite dimensional vector space $\mathcal{P}(y)$, there is an inner product $(\ ,\ )$ defined on $\mathcal{P}(y)$ satisfying

$$(\mathfrak{b}_i, \mathfrak{b}_j) = \begin{cases} 0, & i \neq j, \\ 1, & i = j\,. \end{cases} \tag{8.4}$$

For each $p \in \mathcal{P}(y)$, let us define a map $E : \mathcal{P}(y) \to \mathcal{P}(y)\delta$ for any Sylvester form by

$$E(L(\delta), p) := \sum_{j=1}^{N} \mathfrak{a}_j\, \delta\, (\mathfrak{b}_j, p) = \sum_{j=1}^{N} (\mathfrak{b}_j, p)\mathfrak{a}_j\, \delta\,.$$

With this notation, for each $\ell \leq N$ we obtain

$$E(L(\delta), \mathfrak{b}_\ell) = \sum_{j=1}^{N} \mathfrak{a}_j\, \delta\, (\mathfrak{b}_j, \mathfrak{b}_\ell) = \mathfrak{a}_\ell\, \delta\,.$$

Since $L(\delta) = \widetilde{L}(\delta)$ we have

$$E(L(\delta), \mathfrak{b}_\ell) = E(\widetilde{L}(\delta), \mathfrak{b}_\ell) \longrightarrow \mathfrak{a}_\ell\, \delta = \sum_{k=1}^{\widetilde{N}} (\widetilde{\mathfrak{b}}_k, \mathfrak{b}_\ell)\widetilde{\mathfrak{a}}_k\, \delta\,.$$

Thus, the polynomial $\mathfrak{a}_\ell$ is a linear combination of the polynomials $\widetilde{\mathfrak{a}}_k$, i.e., $\mathfrak{a}_\ell \in \operatorname{span}\{\widetilde{\mathfrak{a}}_k\}$. In a similar way, we can define an inner product $(\mathfrak{a}_i, \mathfrak{a}_j)$ satisfying property (8.4) and apply it to $L(\delta)$ to obtain that $\mathfrak{b}_\ell \in \operatorname{span}\{\widetilde{\mathfrak{b}}_k\}$. ∎

8.2.1. Rational coefficients. We have presented an algorithm which has the property that if the $\mathfrak{a}_j$ and $\mathfrak{b}_j$ are polynomials, then it always gives the lowest possible Sylvester index. However, in our optimization application the $\mathfrak{a}_j$ and $\mathfrak{b}_j$ may be rational functions. Thus, we shall describe how one can extend the algorithm to rational functions rather than polynomials.

The conceptual idea is to think of inverses of expressions as new variables, say, w_j. Then any rational expression is a polynomial in the original variables together with the new letters w_j. In this way, one can apply directly the Sylvester index minimizing algorithm. As an example, suppose that $L(\delta)$ is given by

$$L(\delta) = x(1-x)^{-1}\,\delta\, x - (1-x)^{-1}\,\delta\, x + \delta\, x. \tag{8.5}$$

Using the change of variable

$$w = (1-x)^{-1} \tag{8.6}$$

this expression can be written as

$$L(\delta) = xw\delta x - w\delta x + \delta x.$$

Now, one can apply the MinimumSylvesterIndex command to obtain

$$L(\delta) = (xw - w + 1)\delta x.$$

In this way, the Sylvester index for $L(\delta)$ was reduced to $N = 1$.

That is how the MinimumSylvesterIndex command is used in NCSDP. Unpleasantly, the algorithm did not take into account the "side relationships" that w_j and the other variables might satisfy, which for the above example is

$$(1 - x)^{-1} \equiv x(1 - x)^{-1} + 1$$

or in terms of w

$$xw - w + 1 = 0.$$

Consequently $L(\delta)$ is identically zero. Thus, our algorithm when applied to rational functions fails to produce a minimal Sylvester representation.

To some extent, we are not optimistic about finding a practical exact algorithm for $L(\delta)$ having rational coefficients, because noncommutative Gröbner basis algorithms are very time-consuming. However, we are looking into more empirical methods. One effective test for linear dependence is as follows. Suppose that

$$L(\delta) = \sum_{j=1}^{N} \mathfrak{a}_j \delta \mathfrak{b}_j$$

has already been reduced with the command MinimumSylvesterIndex. Then we replace the symbols appearing in the expressions for $\mathfrak{a}_j$ and for $\mathfrak{b}_j$ by matrices of large dimensions generated randomly. In this way, we obtain random large matrices $\mathcal{A}_j$ and $\mathcal{B}_j$. After, we build numerically the matrices A and B as follows:

$$A = \begin{bmatrix} \operatorname{vec}(\mathcal{A}_1) & \operatorname{vec}(\mathcal{A}_2) & \cdots & \operatorname{vec}(\mathcal{A}_N) \end{bmatrix},$$
$$B = \begin{bmatrix} \operatorname{vec}(\mathcal{B}_1) & \operatorname{vec}(\mathcal{B}_2) & \cdots & \operatorname{vec}(\mathcal{B}_N) \end{bmatrix}.$$

Naturally, N is the number of columns of A and B. Denoting by r_A the rank of the matrix A (respectively, r_B for the rank of B), then the minimum Sylvester index for $L(\delta)$ will be $\min(r_A, r_B)$. If the $\mathcal{A}_j$ and $\mathcal{B}_j$ are polynomials, then we know that r_A and r_B remains N. However, when the $\mathcal{A}_j$ and $\mathcal{B}_j$ are rational functions rather than polynomial, this might not be the case, as described by the example (8.5).

An optimization problem will be presented in section 8.3.1 in which the Sylvester index of the Hessian map is $N = 1043$. After applying the MinimumSylvesterIndex command, the Sylvester index was reduced to $N = 26$. However, $N = 26$ is not the lowest possible index for this Hessian map. When we apply the empirical procedure just described, we found that $r_A = 22$ and $r_B = 22$. Therefore, we know that the minimum Sylvester index is less than or equal to $N = 22$. An empirical algorithm along these lines for actually computing the dependences is under investigation.

Remark 8.7. Another step is taken in order to improve the overall timing, and it is not related to the idea of simplifying expressions by collecting terms, but it is valuable. At the symbolic level we look for inverses of matrices which appear repeatedly inside the symbolic expressions for the Hessian map and we replace each occurrence of an inverse by a new variable. In this way, all numerical inverses are evaluated only once at the beginning of the linear subproblem. This can considerably improve the overall run times.

8.3. Experiments with MinimumSylvesterIndex. The previous examples were presented to illustrate methods for reducing the Sylvester index. Now, we present numerical evidence validating the usefulness of these ideas. For this purpose, let us consider the following eigenvalue minimization problem, whose numerical behavior is to be presented in section 9:

$$\text{(P2)} \qquad \begin{array}{l} \inf \lambda_{\max}(CXC^T) \\ \text{subject to} \end{array}$$
$$0 < X,$$
$$0 < G(X) := A_3X + XA_3^T - XR_3^{-1}X + S_3,$$
$$0 < F(X) := A_1X + XA_1^T - XR_1^{-1}X + S_1 - (A_2^TX + XA_2)G(X)^{-1}(A_2^TX + XA_2)$$

with all the matrices having dimension $n \times n$.

As already described, we need to compute symbolically the Hessian and the gradient of an auxiliary potential function. For the above example, this potential function is given by the symbolic formula

$$\phi_\gamma(x) = -\log\det x - \log\det F(x) - \log\det G(x) - \log\det(\gamma\mathfrak{I}_d - cxc^T),$$

where $\mathfrak{I}_d$ stand for the symbolic analogue of the identity matrix and γ is a scalar which is not relevant here. The expression $\phi_\gamma(x)$ is a function of the unknown x. If the update direction is taken to be δ_x, the Hessian map $\mathcal{H}(\delta_x)$ as a function of δ_x will have a structure of the form

$$\mathcal{H}(\delta_x) = \sum_i^N \mathfrak{a}_i\delta_x\mathfrak{b}_i,$$

where the $\mathfrak{a}_i$ and $\mathfrak{b}_i$ are noncommutative rational functions of the variables c, a_1, a_2, a_3, r_1, r_3, s_1, s_3, x. At this stage, one can apply the MinimumSylvesterIndex command to reduce the Sylvester index N. The gradient map $\mathcal{Q}$ is obtained from the first directional derivative of $\phi_\gamma(x)$ along the direction δ_x. Thus, for this symmetric case, one obtains a "symbolic" system given by

$$\mathcal{H}(\delta_x) = \mathcal{Q} + \mathcal{Q}^T. \tag{8.7}$$

The next step is to substitute for matrices of compatible dimensions the symbols appearing in $\mathcal{H}(\delta_x)$ and $\mathcal{Q}$. Thus, the code becomes numerical, and to find numerically the update direction δ_X, we must be able to solve the linear system of equations given by

$$\mathbb{H}(\delta_X) = \mathbb{Q} + \mathbb{Q}^T.$$

Using the vec operation, the above system can be equivalently written as

$$Hv = g, \tag{8.8}$$

where $H = \sum_i^N \mathcal{B}_i^T \otimes \mathcal{A}_i$, $g = \mathrm{vec}(\mathbb{Q} + \mathbb{Q}^T)$, and $v = \mathrm{vec}(\delta_X)$.

Therefore, in order to solve numerically the linear system given in (8.7), one needs

FE: to substitute matrices for the symbols appearing in $\mathcal{H}(\delta_x)$ and $\mathcal{Q}$;
KP: to evaluate the Hessian matrix H by applying N Kronecker products;
LS: to solve the system $Hv = g$ for the update direction v.

TABLE 8.1
Timing (seconds): formulas evaluation, Kronecker products, and linear solver.

SIZE	16			32			64		
	MSI	CS	UNT	MSI	CS	UNT	MSI	CS	UNT
FE	0.071	0.096	0.409	0.258	0.287	1.09	1.47	1.74	6.1
KP	0.039	0.061	1.341	0.603	0.974	21.47	9.74	15.71	344.3
LS	0.029	0.028	0.029	0.397	0.410	0.41	9.65	9.87	9.8
TOT	0.139	0.185	1.779	1.258	1.671	22.96	20.85	27.31	360.1

Ratio	UNT / MSI	UNT / MSI	UNT / MSI
FE	5.8	4.2	4.1
KP	34.4	35.6	35.3
TOT	12.8	18.3	17.3

The first two steps, namely, FE (formula evaluations) and KP (Kronecker products), are the two main steps where reducing the Sylvester index N of the expression for $\mathcal{H}(\delta_x)$ can significantly affect the evaluation time. We do not show the formulas for $\mathcal{H}(\delta_x)$ and $\mathcal{Q}$, since these expressions are quite large and would consume several pages. What is important is the fact that the formula for $\mathcal{H}(\delta_x)$ as computed originally, before applying any simplification rule, has $N = 1014$ Sylvester terms. However, after applying the NCCollectSylvester command to the original expression, the Sylvester index decreases to $N = 43$, and after applying our MinimumSylvesterIndex command to the original expression, we obtain the index $N = 26$ for this Hessian.

8.3.1. Time saved by applying MinimumSylvesterIndex. To find out how much time is actually saved at the numerical level, the NCSDP code is executed using the collected formulas for $\mathcal{H}(\delta_x)$ with $N = 26$ (MinimumSylvesterIndex command) and with $N = 43$ (NCCollectSylvester command) and the not collected formula for $\mathcal{H}(\delta_x)$ with $N = 1014$. For this set of experiments, the size n of the matrices involved assume the values $n = 16, 32, 64$. For each case, we execute the inner loop where the linear system (8.8) is numerically solved 20 times. Thus, we measure the CPU time per call (average over 20 iterations) spent on the above items, FE, KP, and LS (linear solver). In this way, we can analyze how the time spent on formula evaluations behaves as a function of the size of the matrices involved in the expressions, as well as the Sylvester index N.

The results are presented in Table 8.1, where MSI stands for the Hessian simplified by MinimumSylvesterIndex (the Sylvester index is $N = 26$), CS stands for the Hessian simplified by NCCollectSylvester (the Sylvester index is $N = 43$), and UNT stands for the untreated Hessian (the Sylvester index is $N = 1014$). In this table, the row labeled Ratio is the ratio between the untreated column and the MSI column. The time spent on solving the linear system, presented in the row labeled LS, is not affected by the expression being or not being collected. SIZE stands for matrix size, and TOT for the total time FE+KP+LS.

The results provided in Table 8.1 show that collecting terms in the expression for the Hessian map $\mathcal{H}(\delta_x)$ represents a huge saving, since the average time spent on substituting matrices for the symbols that appear in the expressions for the $\mathfrak{a}_i$ and $\mathfrak{b}_i$ when the expressions are not collected (UNT case) is approximately four to five times longer than the time for the MSI collected case (row FE in Ratio). Moreover, collecting the expressions significantly improved the time spent on evaluating Kronecker products, as seen from row KP under Ratio, where this timing improved by a factor ranging from 34.4 to 35.6.

For matrices of dimension 16 and 32, the time per call spent (over 20 iterations) on numerically solving the equation $Hv = g$ for the unknown v was relatively small, as seen from row LS. On the other hand, for matrices of dimension 64, the (LS) cost becomes significantly large. To understand this fact better, suppose the dimension of the matrices involved is chosen to be $n = 64$. Thus, the symmetric unknown matrix X having size 64×64 implies that the unknown vector v and the system to be solved will have size approximately $64^2/2 = 2048$.

Kronecker products are also extremely expensive, as seen from row KP for size 64. In fact, if we did not have a theory for decreasing the Sylvester index, our approach using Kronecker products would be intractable for matrices of large dimensions. If one could solve the linear system of equations for δ_X in its original structured form $\mathbb{H}(\delta_X) = \mathbb{Q}$, without applying Kronecker products and keeping the dimension of the linear system low, a huge saving on the numerical linear solver would probably be attained. This is an open area which we hope members of the community will pursue.

8.4. Some more experiments using MinimumSylvesterIndex. Another interesting experiment is to analyze how the Sylvester index behaves by applying our MinimumSylvesterIndex command to a variety of matrix inequalities which appear in control design. The example just presented, taken from section 9, has shown a great improvement since the Sylvester index reduced from $N = 1014$ to $N = 26$. Now, we present two more examples.

Example 8.8. For the standard Riccati inequality

$$AX + XA^T - XRX + S > 0$$

the Hessian map $\mathcal{H}(\delta_x)$ for the untreated case has a Sylvester index of $N = 20$, while our MinimumSylvesterIndex algorithm applied to it provides a Sylvester index of $N = 4$.

Example 8.9. Now, a more realistic example is used: a mixed H_2/H_∞ control problem

$$\text{(P3)} \qquad \begin{array}{l} \inf \operatorname{Tr}\{W\} \\ \text{subject to} \\ \qquad\qquad 0 < X, \\ \qquad 0 < W - (C_2X + D_{2u}F)X^{-1}(C_2X + D_{2u}F)^T, \\ 0 > AX + XA^T + B_uF + F^TB_u^T + B_wB_w^T \\ \qquad + \left[XC_1^T + F^TD_{1u}^T + B_wD_{1w}^T\right] R^{-1} \left[XC_1^T + F^TD_{1u}^T + B_wD_{1w}^T\right]^T \end{array}$$

with $R = \eta^2 I - D_{1w}D_{1w}^T > 0$.

For the above control problem, there are three unknowns denoted by $W = W^T$, $X = X^T$, and F (not symmetric). Thus, the linear subproblem to be solved will have dimension 3×3, and consequently each entry on this system will contain a Sylvester operator. For instance, the (1,1) entry will be an expression of the form

$$\sum_i^{N_{11}} \mathfrak{a}_i^{11} \delta_w \mathfrak{b}_i^{11} + \sum_i^{\hat{N}_{11}} \hat{\mathfrak{a}}_i^{11} {\delta_w}^T \hat{\mathfrak{b}}_i^{11}.$$

The (1,2) entry will have the form $\sum_i^{N_{12}} \mathfrak{a}_i^{12} \delta_x \mathfrak{b}_i^{12} + \sum_i^{\hat{N}_{12}} \hat{\mathfrak{a}}_i^{12} \delta_x^T \hat{\mathfrak{b}}_i^{12}$. The (1,3) entry will have the form $\sum_i^{N_{13}} \mathfrak{a}_i^{13} \delta_f \mathfrak{b}_i^{13} + \sum_i^{\hat{N}_{13}} \hat{\mathfrak{a}}_i^{13} \delta_f^T \hat{\mathfrak{b}}_i^{13}$. The (2,1) entry is the adjoint case

TABLE 8.2
Sylvester index N and $\hat{N}$ for the MSI Hessian $\mathcal{H}(\delta_w, \delta_x, \delta_f)$.

	Sylv. index N_{ij}			Sylv. index $\hat{N}_{ij}$		
i,j	1	2	3	1	2	3
1	1	2	1	0	0	1
2	2	10	4	0	0	4
3	1	4	2	1	4	4

TABLE 8.3
Sylvester index N and $\hat{N}$ for the UNT Hessian $\mathcal{H}(\delta_w, \delta_x, \delta_f)$.

	Sylv. index N_{ij}			Sylv. index $\hat{N}_{ij}$		
i,j	1	2	3	1	2	3
1	1	2	2	0	0	2
2	2	73	38	0	0	38
3	2	38	42	2	38	38

of the (1,2) entry. The (2,2) entry will have the form $\sum_i^{N_{22}} \mathfrak{a}_i^{22} \delta_x \mathfrak{b}_i^{22} + \sum_i^{\hat{N}_{22}} \hat{\mathfrak{a}}_i^{22} \delta_x^T \hat{\mathfrak{b}}_i^{22}$ and so forth. It should be noticed that the Sylvester index $\hat{N}_{11}$, $\hat{N}_{12}$, $\hat{N}_{21}$, and $\hat{N}_{22}$ are zero, since the corresponding variables w and x are symmetric.

For the matrix inequalities given in problem (P3), the set of Sylvester indexes N and $\hat{N}$ for the Hessian map $\mathcal{H}(\delta_w, \delta_x, \delta_f)$ simplified by MinimumSylvesterIndex (MSI case) and for the untreated Hessian (UNT case) are respectively presented in Table 8.2 and Table 8.3.

In Tables 8.2 and 8.3, the variables x and f are associated with the entries

$$\text{(i,j)} \in \{(2,2) \quad (2,3) \quad (3,2) \quad (3,3)\}$$

for each one of the subtables. If we only pay attention to the Sylvester index N, we see that the submatrix associated with x and f for the

UNT case $\begin{array}{|c|c|}\hline 73 & 38 \\ \hline 38 & 42 \\ \hline \end{array}$ reduces to only $\begin{array}{|c|c|}\hline 10 & 4 \\ \hline 4 & 2 \\ \hline \end{array}$ in the MSI case.

Similarly, a large reduction is also obtained for the Sylvester index $\hat{N}$. Thus, for the variables x and f, we found that a large reduction on the Sylvester index N and $\hat{N}$ are obtained after applying our MinimumSylvesterIndex command. Naturally, this will represent a considerable saving on the evaluation time for the numerical linear solver.

It is also true that the process of simplifying rational functions, at the symbolic level of Mathematica, can consume a considerable amount of time. However, this computation is performed only once at the beginning of the run. This is in contrast with the numerical part, where solving the linear system to provide the update direction takes place at each inner iteration (which occurs several times). Therefore, the ability to collect factors in an expression (decreasing the Sylvester index) plays a very important role.

9. Numerical experiment: Timing of NCSDP. In this section, our NCSDP code is numerically compared to some available semidefinite programming solvers.

TABLE 9.1
Total CPU time in seconds.

	Matrix Size			
	8	16	32	64
SDPT3	2.59	12.94	163.77	3132.68
LMI-Lab	0.43	2.58	66.03	2124.54
SeDuMi	0.79	2.20	33.63	1254.30
NCSDP	7.73	12.57	81.40	1224.49

TABLE 9.2
CPU time per iteration in seconds.

	Matrix Size					
	16		32		64	
	CPI	IT	CPI	IT	CPI	IT
SDPT3	1.62	8	18.20	9	348.08	9
SeDuMi	0.20	11	2.59	13	89.59	14
NCSDP	0.60	21	4.28	19	72.03	17

For this purpose, the optimization problem to be used is the following eigenvalue minimization problem (stated earlier in section 8.3.1):

$$\text{(P2)} \qquad \begin{array}{l} \inf \lambda_{\max}(CXC^T) \\ \text{subject to} \\ \qquad 10^{-1}I < X, \\ \qquad 0 < G(X) := A_3X + XA_3^T - XR_3^{-1}X + S_3, \\ 0 < F(X) := A_1X + XA_1^T - XR_1^{-1}X + S_1 - (A_2^TX + XA_2)G(X)^{-1}(A_2^TX + XA_2). \end{array}$$

The matrices C, A_1, A_2, and A_3 belong to $\mathbb{R}^{n\times n}$, the invertible matrices R_1, R_3 belong to $\mathbb{S}^n_{++}$ and the matrices S_1, S_3, and X belong to $\mathbb{S}^n$. We do not present the numerical values of those matrices since it would take considerable space. Note that by Schur complement techniques problem (P2) can be equivalently restated as an LMI problem.

The results of this experiment (shown in Table 9.1) show the overall CPU time spent by the solvers SDPT3, LMI-Lab, SeDuMi, and NCSDP to solve the above optimization problem (P2) within the required accuracy of 10^{-4} for the objective value. The LMI-Lab toolbox (Version 1.0.8) is based on the projective method of [8]. The SeDuMi solver (Version 1.02) from [25] implements the self-dual embedding technique for optimization over self-dual homogeneous cones. The SDPT3 solver (Version 3.0) from [28] implements an infeasible path-following algorithm that employs a predictor-corrector method. The starting feasible points were the same for all the solvers.

From Table 9.1, one sees that for matrices of size 64, the solvers SeDuMi and NCSDP were the fastest code for the eigenvalue minimization problem (P2). The CPU times per iteration (CPI) and number of (outer) iterations (IT) are presented in Table 9.2. We believe that NCSDP might be significantly faster than SeDuMi for matrices of dimensions larger than 64×64. However, we did not run this experiment

TABLE 9.3
Numerical behavior of NCSDP.

SIZE	IT/NeNe	FE	KP	LS	g	$\lambda_{\min}(H)$	$\lambda_{\max}(H)$
8	25/94	2.32	0.07	0.14	$1.14E+04$	$6.74E+03$	$2.59E+10$
16	21/75	3.05	2.88	2.16	$8.94E+03$	$1.62E+03$	$2.41E+10$
32	19/69	8.24	41.9	26.7	$7.38E+03$	$1.14E+03$	$2.24E+10$
64	17/61	43.4	595	580	$5.84E+03$	$9.81E+02$	$2.76E+10$

since the overall elapsed time would be extremely long and because of the requirement of large RAM memory availability. The computer used for our experiments was an Intel Celeron at 2800 MHz CPU clock, 512MB of RAM, 1GB of swap, running Linux (kernel 2.4.20-31.9), MATLAB Version 6.5.0 R13, Mathematica 4.0, and NCAlgebra Version 3.7.

We believe our NCSDP code, even in its raw stage, has been competitive mainly because it allows nonlinear matrix inequalities, so it avoids the increase in dimensions when converting to LMIs using Schur complements. Also, we think that the techniques in the paper allow the numerical Newton equations to be derived more efficiently. However, we did not take advantage of the special structure of the linear subproblem when solving it.

For these experiments, we installed the codes listed above using their default installer. However, since these codes are for general SDP problems, where the input data should be expressed in a "standard" SDP form, which is not the standard LMI form (like the input from LMILab), we make use of the package LMILab Translator (LMITrans) that translates from LMILab form to the SeDuMi and SDPT3 form. The timing presented in Table 9.1 did not incorporate the (modest amount of) time consumed by this interface.

We do not know if LMITrans does nearly the "optimal" conversion for each solver; this adds uncertainty to the experiments. It might be the case that there exists an optimal conversion for a particular solver, in this case, the solver would perform better than for the default options used in LMITrans. However, we also used YALMIP as a front-end for SeDuMi and SDPT3 at the suggestion of a referee after the paper was complete. In a few tests, we found that it did not affect the timings significantly: the timing was approximately the same as the timing obtained using LMITrans.

9.1. Numerical behavior of NCSDP. We now provide the numerical details of the results from NCSDP presented in Table 9.1. The trade-off between the number of inner iterations, NeNe, and the number of outer iterations, Iter (as seen from Table 9.3), is a characteristic of Barrier methods, in particular, the method of centers [2], and depends mainly on the centralization parameter θ, given in (6.1) from section 6.2. In practice, a smaller θ induces a higher number of inner iterations, NeNe. For all the experiments, θ was set to $\theta = 0.2$.

For matrices of small size, the most expensive part is the time spent on evaluating the Sylvester terms $\mathfrak{a}_i$, $\mathfrak{b}_i$, and $\mathcal{Q}$, presented in column FE. However, when the size of the matrices increases above 8, the time spent on Kronecker products, column KP, and the time spent on solving the linear system, column LS, begins to dominate.

In Table 9.3, the column g stands for the gradient, and columns $\lambda_{\min}(H)$ and $\lambda_{\max}(H)$ stand, respectively, for the minimum and the maximum eigenvalues of the

Hessian matrix H at the optimum. Those values show that the condition number of the Hessian is large at the optimal solution. This ill-conditioning in the Hessian is a well-known fact for classical barrier methods [30, 18], where it has been shown that this behavior is highly influenced by the set of constraints that are or are not active (binding) at the solution. However, it is not an immediate task to determine the set of active constraints in the semidefinite programming framework.

9.2. Implementation speed-ups. At this stage, the numeric part of our NCSDP code is "completely" implemented using MATLAB functions (not compiled). The only compiled part of our code is the Kronecker product, since the MATLAB function *kron.m* was extremely slow for our needs. On the other hand, most of the other solvers have their core subroutines written in either Fortran or C, which significantly improve their overall performance.

To make the experiment transparent, the stopping criteria for the inner loop in NCSDP is kept constant throughout all the iterations. We stop as soon as $\sigma = 1$ (see section 6.3). Changing dynamically this stopping criteria might also improve the timing of the solver.

Although in this paper we have focused on convex optimization problems over matrix inequalities, the extension of our numerical ideas to finding local solutions in nonconvex situations is immediate; however, reliability has not been tested [4].

We reiterate that fast methods for solving numerical linear equations of Sylvester form have not been investigated and are the main open question motivated by this paper. A big advance here would translate directly into a big reduction in run times.

Appendix A. Illustrating our methodology by an example. In this section we explain the ideas behind Theorem 7.1 and Corollary 7.2 through a simple optimization problem. The extrapolation to a more general case is straightforward, but it gives messy formulas. Let us consider the optimization problem

$$\min \{\mathrm{Tr}\,\{X\} \,:\, X \in \mathrm{closure}(\mathcal{G})\} \tag{A.1}$$

with $F(X) := AX + XA^T - XRX + Q$, and the domain $\mathcal{G}$ given by

$$\mathcal{G} = \{X \in \mathbb{S}^n \,:\, F(X) > 0\}\,.$$

We assume (1) all matrices have dimension $n \times n$; (2) the matrices X, R, Q are symmetric; (3) the closure of the set $\mathcal{G}$ is compact.

A.1. Describing the central path. Let us define the unconstrained auxiliary potential function $\phi_\gamma(X)$ as described in Theorem 7.1 as

$$\phi_\gamma(X) = \zeta \log\Big(1/(\gamma - \mathrm{Tr}\,\{X\})\Big) - \log\det F(X). \tag{A.2}$$

The analytic center for the potential $\phi_\gamma(X)$ is the path given by

$$X^*(\gamma^k) = \mathrm{argmin}\,\phi_{\gamma^k}(X). \tag{A.3}$$

A.2. Solving for the analytic center. The above optimization problem (A.1) has now been replaced by a sequence of unconstrained minimization problems in the form (A.3). In this way, we are interested in finding update directions which lead toward the central path of (A.3). To find those directions, Newton's method is applied to minimize the second-order Taylor series expansion of the potential function $\phi(x)$.

These procedures are summarized in section 7.1. Here, we go through each step precisely. For clarity of notation, we omit the subscript γ in $\phi_\gamma(x)$. To compute the quadratic approximation of $\phi(x)$, we take δ_x to be the update direction for x. Thus, assuming $x^* = x + \delta_x$, the series expansion of $\phi(x)$ up to the second term is given by

$$\tilde{\phi}(\delta_x) := \phi(x^*) - \phi(x) = D\phi(x)\,[\delta_x] + \frac{1}{2}D^2\phi(x)\,[\delta_x, \delta_x]\,. \tag{A.4}$$

A.3. Directional eerivatives of $F(x)$. In order to compute the derivatives in (A.4), we need to have at hand the first and second directional derivatives of $F(x)$. Recalling that x is symmetric, and therefore so is the update direction δ_x, the first directional derivative of $F(x)$ in the direction δ_x is given by

$$DF(x)\,[\delta_x] = (a - xr)\delta_x + \delta_x(a^T - rx) = \operatorname{sym}\{(a - xr)\delta_x\}$$

and the second directional derivative is

$$D^2F(x)\,[\delta_x, \delta_x] = -\operatorname{sym}\{\delta_x r \delta_x\}$$

A.4. Connection with Theorem 7.1. Comparing the above derivatives of $F(x)$ with the formulas (4.2) and (4.3), one readily verifies that $k = 1$, $a_1 = (a - xr)$, and $b_1 = 1$, for the first directional derivative, and that $w_1 = 1, w_2 = 0, w_3 = 0$, $m_1 = 1$, $n_1 = -r$, and $t_1 = 1$, for the second directional derivative. With this notation, we can directly apply Theorem 7.1 to obtain the algebraic linear system of equations. For the gradient term $\mathcal{Q}$ we have

$$\begin{aligned}\mathcal{Q} &= \sum_{i=1}^{k} a_i^T F(x)^{-1} b_i^T - \frac{1}{2}\zeta\,(\gamma - \operatorname{Tr}\{x\})^{-1}\,\mathfrak{I}_d \\ &= (a - xr)^T F(x)^{-1} - \frac{1}{2}\zeta(\gamma - \operatorname{Tr}\{x\})^{-1}\mathfrak{I}_d.\end{aligned}$$

For the Hessian $\mathcal{H}(\delta_x)$ we calculate

1. $\sum_{i=1}^{k}\sum_{j=1}^{k} a_i^T F(x)^{-1} a_j \delta_x b_j F(x)^{-1} b_i^T = (a - xr)^T F(x)^{-1}(a - xr)\delta_x F(x)^{-1}$;
2. $\sum_{i=1}^{k}\sum_{j=1}^{k} a_i^T F(x)^{-1} b_j^T \delta_x^T a_j^T F(x)^{-1} b_i^T = (a - xr)^T F(x)^{-1}\delta_x(a - xr)F(x)^{-1}$;
3. $-\sum_{j=1}^{w_1} n_j^T \delta_x^T m_j^T F(x)^{-1} t_j^T + m_j^T F(x)^{-1} t_j^T \delta_x^T n_j^T = r\delta_x F(x)^{-1} + F(x)^{-1}\delta_x r$.

Thus $\mathcal{H}(\delta_x)$ becomes

$$\begin{aligned}\mathcal{H}(\delta_x) = {} & \frac{1}{2}F(x)^{-1}\delta_x r + \frac{1}{2}r\delta_x F(x)^{-1} + (a - xr)^T F(x)^{-1}(a - xr)\delta_x F(x)^{-1} \\ & + (a - xr)^T F(x)^{-1}\delta_x(a - xr)^T F(x)^{-1} + \frac{1}{2}\zeta(\gamma - \operatorname{Tr}\{x\})^{-2}\operatorname{Tr}\{\delta_x\}\,\mathfrak{I}_d.\end{aligned}$$

Consequently, the algebraic linear system of equations is described by

$$\operatorname{Tr}\left\{\delta_V(\mathcal{H}(\delta_x) - \mathcal{Q})^T + (\mathcal{H}(\delta_x) - \mathcal{Q})\delta_V\right\} = 0 \tag{A.5}$$

with $\mathcal{H}(\delta_X)$ and $\mathcal{Q}$ as given above

These are the steps someone would need in order to apply Theorem 7.1 directly. However, we shall go through the details of the manipulation that leads to this main result.

A.5. Directional derivatives of the barrier function. Having the above directional derivatives of $F(x)$ available, we are ready to take the directional derivatives needed in (A.4). However, to clarify the exposition, we split the potential function into two parts:

$$\phi_1(x) = -\log\det F(x) \quad \text{and} \quad \phi_2(x) = \zeta \log\Big(1/(\gamma - \operatorname{Tr}\{x\})\Big).$$

A.6. Symbolic directional derivatives of $\phi_1(x) = -\log\det F(x)$. The first and second directional derivative of $\phi_1(x)$ in the direction δ_x are given by

$$\begin{aligned} D\phi_1(x)[\delta_x] &= -\operatorname{Tr}\left\{F(x)^{-1} DF(x)[\delta_x]\right\} \\ &= -\operatorname{Tr}\left\{F(x)^{-1}\operatorname{sym}\{(a - xr)\delta_x\}\right\}, \end{aligned}$$

$$\begin{aligned} D^2\phi_1(x)[\delta_x,\delta_x] &= \operatorname{Tr}\left\{\left(F(x)^{-1}DF(x)[\delta_x]\right)^2\right\} - \operatorname{Tr}\left\{F(x)^{-1}D^2F(x)[\delta_x,\delta_x]\right\} \\ &= \operatorname{Tr}\left\{\left(F(x)^{-1}\operatorname{sym}\{(a-xr)\delta_x\}\right)^2\right\} \\ &\quad + \operatorname{Tr}\left\{F(x)^{-1}\operatorname{sym}\{\delta_x r \delta_x\}\right\}. \end{aligned}$$

A.7. Symbolic directional derivatives of $\phi_2(x) = \zeta\log(1/(\gamma - \operatorname{Tr}\{x\}))$. The first derivative is given by

$$D\phi_2(x)[\delta_x] = \zeta\left(\gamma - \operatorname{Tr}\{x\}\right)^{-1}\operatorname{Tr}\{\delta_x\}$$

and the second by

$$D^2\phi_2(x)[\delta_x,\delta_x] = \zeta\left(\left(\gamma - \operatorname{Tr}\{x\}\right)^{-1}\operatorname{Tr}\{\delta_x\}\right)^2.$$

A.8. Optimality conditions. Now we are ready to write the optimality conditions which will provide the update direction. These conditions are the first-order necessary optimality conditions for problem (A.3), obtained by taking directional derivatives of the Taylor expansion $\phi(x+\delta_x)$, given by (A.4), as a function of δ_x in the direction δ_V. To accomplish this step, we should compute $D\tilde{\phi}(\delta_x)[\delta_V]$ or equivalently

$$D\left(D\phi(x)[\delta_x] + \frac{1}{2}D^2\phi(x)[\delta_x,\delta_x]\right)[\delta_V] = 0. \tag{A.6}$$

Using the directional derivatives just computed in the previous sections, the expression for the second-order approximation $\tilde{\phi}(\delta_x)$ is given

$$\begin{aligned} \tilde{\phi}(\delta_x) &= -\operatorname{Tr}\left\{F(x)^{-1}\operatorname{sym}\{(a-xr)\delta_x\}\right\} + \frac{1}{2}\operatorname{Tr}\left\{\left(F(x)^{-1}\operatorname{sym}\{(a-xr)\delta_x\}\right)^2\right\} \\ &\quad + \frac{1}{2}\operatorname{Tr}\left\{F(x)^{-1}\operatorname{sym}\{\delta_x r\delta_x\}\right\} + \zeta\left(\gamma - \operatorname{Tr}\{x\}\right)^{-1}\operatorname{Tr}\{\delta_x\} \\ &\quad + \frac{1}{2}\zeta\left(\left(\gamma - \operatorname{Tr}\{x\}\right)^{-1}\operatorname{Tr}\{\delta_x\}\right)^2. \end{aligned} \tag{A.7}$$

To proceed, we now set to zero the directional derivative of $\tilde{\phi}(\delta_x)$ as a function of δ_x in the direction δ_V. After a few manipulations, the term $D\tilde{\phi}(\delta_x)\,[\delta_V]$ is given by

$$\begin{aligned}
D\tilde{\phi}(\delta_x)\,[\delta_V] = \operatorname{Tr}\Big\{ \delta_V \Big(& F(x)^{-1}\delta_x(a-xr)^T F(x)^{-1}(a-xr) - F(x)^{-1}(a-xr) \\
& + \frac{1}{2} r\delta_x F(x)^{-1} + \frac{1}{2} F(x)^{-1}\delta_x r \\
& + F(x)^{-1}(a-xr)\delta_x F(x)^{-1}(a-xr) \\
& + \frac{1}{2}\zeta(\gamma - \operatorname{Tr}\{x\})^{-1}\mathfrak{I}_d + \frac{1}{2}\zeta(\gamma - \operatorname{Tr}\{x\})^{-2}\operatorname{Tr}\{\delta_x\}\,\mathfrak{I}_d \Big) \\
+ \Big(& (a-xr)^T F(x)^{-1}(a-xr)\delta_x F(x)^{-1} - (a-xr)^T F(x)^{-1} \\
& + \frac{1}{2} F(x)^{-1}\delta_x r + \frac{1}{2} r\delta_x F(x)^{-1} \\
& + (a-xr)^T F(x)^{-1}\delta_x (a-xr)^T F(x)^{-1} \\
& + \frac{1}{2}\zeta(\gamma - \operatorname{Tr}\{x\})^{-1}\mathfrak{I}_d + \frac{1}{2}\zeta(\gamma - \operatorname{Tr}\{x\})^{-2}\operatorname{Tr}\{\delta_x\}\,\mathfrak{I}_d \Big)\delta_V \Big\}.
\end{aligned}$$

Therefore, the algebraic linear system of equations is described by

$$\operatorname{Tr}\left\{\delta_V(\mathcal{H}(\delta_x) - \mathcal{Q})^T + (\mathcal{H}(\delta_x) - \mathcal{Q})\delta_V\right\} = 0 \tag{A.8}$$

with $\mathcal{H}(\delta_X)$ and $\mathcal{Q}$ respectively given by

$$\mathcal{Q} = (a-xr)^T F(x)^{-1} - \frac{1}{2}\zeta(\gamma - \operatorname{Tr}\{x\})^{-1}\mathfrak{I}_d,$$

$$\begin{aligned}
\mathcal{H}(\delta_x) = & \frac{1}{2} F(x)^{-1}\delta_x r + \frac{1}{2} r\delta_x F(x)^{-1} + (a-xr)^T F(x)^{-1}(a-xr)\delta_x F(x)^{-1} \\
& + (a-xr)^T F(x)^{-1}\delta_x (a-xr)^T F(x)^{-1} + \frac{1}{2}\zeta(\gamma - \operatorname{Tr}\{x\})^{-2}\operatorname{Tr}\{\delta_x\}\,\mathfrak{I}_d.
\end{aligned}$$

A.9. Connection with Theorem 7.1. We shall emphasize that this illustrative example gives a reasonable idea of how the proof of Theorem 7.1 was constructed, since it follows very similar steps.

A.10. The algebraic linear system of equations. Since the unknown x is restricted to being symmetric (so is δ_x) the subspace $\mathcal{V}$ equals the space of symmetric matrices. Consequently, its orthogonal complement $\mathcal{V}^\perp$ is the set of all skew symmetric matrices. Therefore, we obtain the following linear system in δ_x:

$$\mathcal{H}(\delta_x) + \mathcal{H}(\delta_x)^T = \mathcal{Q} + \mathcal{Q}^T.$$

We can rewrite this equation using a suitable choice of variables $\mathfrak{a}_i$ and $\mathfrak{b}_i$ as follows:

$$\operatorname{sym}\left\{\sum_{i=1}^{4}\left(\mathfrak{a}_i\delta_x\mathfrak{b}_i\right)\right\} + \varrho\operatorname{Tr}\{\delta_x\} = \mathcal{Q} + \mathcal{Q}^T \tag{A.9}$$

with

$$\begin{aligned}
\mathfrak{a}_1 &= F(x)^{-1}(a - xr), & \mathfrak{b}_1 &= F(x)^{-1}(a - xr),\\
\mathfrak{a}_2 &= F(x)^{-1}, & \mathfrak{b}_2 &= (a - xr)^T F(x)^{-1}(a - xr),\\
\mathfrak{a}_3 &= \frac{1}{2}F(x)^{-1}, & \mathfrak{b}_3 &= r,\\
\mathfrak{a}_4 &= r, & \mathfrak{b}_4 &= \frac{1}{2}F(x)^{-1},
\end{aligned}$$

and

$$\mathcal{Q} = (a - xr)^T F(x)^{-1} - \frac{1}{2}\zeta(\gamma - \operatorname{Tr}\{x\})^{-1}\mathfrak{I}_d, \qquad \varrho = \frac{1}{2}\zeta(\gamma - \operatorname{Tr}\{x\})^{-2}\mathfrak{I}_d.$$

A.11. Connection with Corollary 7.2. These are the $\mathfrak{a}_i$ and $\mathfrak{b}_i$ described in the corollary for the specific case where the subspace $\mathcal{V}$ equals the space of symmetric matrices $\mathbb{S}^n$. The proof of Corollary 7.2 is illustrated by our example, since the proof mainly consists in determining the orthogonal complement of the subspace $\mathcal{V}$.

Acknowledgments. Thanks are due to M. C. de Oliveira for many ideas and suggestions. We also thank the referees for their very valuable comments and suggestions.

REFERENCES

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

[2] S. Boyd and L. El Ghaoui, *Method of centers for minimizing generalized eigenvalues*, Linear Algebra Appl., 188 (1993), pp. 63–111.

[3] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, New York, 2004.

[4] J. F. Camino, *Optimization over Convex Matrix Inequalities*, Ph.D. thesis, University of California, San Diego, 2003.

[5] J. F. Camino, J. W. Helton, R. E. Skelton, and J. Ye, *Matrix inequalities: A symbolic procedure to determine convexity automatically*, Integral Equations Operator Theory, 46 (2003), pp. 399–454.

[6] L. El Ghaoui and S. Niculescu, *Advances in Linear Matrix Inequality Methods in Control*, Adv. Des. Control, SIAM, Philadelphia, 1999.

[7] A. V. Fiacco and G. P. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, Classics Appl. Math., SIAM, Philadelphia, 1990.

[8] P. Gahinet, A. Nemirovskii, A. J. Laub, and M. Chilali, *LMI Control Toolbox*, The Math Works, Natick, MA, 1995.

[9] G. Golub and C. V. Loan, *Matrix Computation*, Johns Hopkins University Press, Baltimore, 1983.

[10] E. L. Green, *Multiplicative bases, Gröbner bases, and right Gröbner bases*, J. Symbolic Comput., 29 (2000), pp. 601–623.

[11] J. W. Helton and J. J. Wavrik, *Rules for computer simplification of the formulas in operator model theory and linear systems*, in Nonselfadjoint Operators and Related Topics Oper. Theory Adv. Appl. 73, Birkhäuser, Basel, 1994, pp. 325–354.

[12] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, New York, 1999.

[13] K. Jbilou and A. Messaoudi, *Matrix recursive interpolation algorithm for block linear systems direct methods*, Linear Algebra Appl., (1999), pp. 137–154.

[14] M. Konstantinov, V. Mehrmann, and P. Petkov, *On properties of Sylvester and Lyapunov operators*, Linear Algebra Appl., 312 (2000), pp. 35–71.

[15] F. Leibfritz and E. M. Mostafa, *An interior point constrained trust region method for a special class of nonlinear semidefinite programming problems*, SIAM J. Optim., 12 (2002), pp. 1048–1074.

[16] A. S. Lewis and M. L. Overton, *Eigenvalue optimization*, Acta Numer., 5 (1996), pp. 149–190.
[17] P. A. Linnell, *Noncommutative Localization in Group Rings*, arXiv: Math. RA/0311071.
[18] W. Murray, *Analytic expression for the eigenvalues and eigenvectors of the Hessian matrices of barrier and penalty functions*, J. Optim. Theory Appl., 7 (1971), pp. 189–196.
[19] Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*, Stud. Appl. Math., SIAM, Philadelphia, 1994.
[20] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Classics in Appl. Math., SIAM, Philadelphia, 2000.
[21] M. L. Overton, *On minimizing the maximum eigenvalue of a symmetric matrix*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 256–268.
[22] P. A. Parrilo, *Structured Semidefinite Programs and Semialgebraic Geometry Methods in Robustness and Optimization*, Ph.D. thesis, California Institute of Technology, 2000.
[23] R. E. Skelton, T. Iwasaki, and K. M. Grigoriadis, *A Unified Algebraic Approach to Linear Control Design*, Taylor & Francis, London, 1998.
[24] M. Stankus, J. W. Helton, and J. Wavrik, *Computer simplification of formulas in linear systems theory*, IEEE Trans. Automat. Control, 43 (1998), pp. 302–314.
[25] J. F. Sturm, *Using SeDuMi* 1.02*, a MATLAB toolbox for optimization over symmetric cones*, Optim. Methods Softw., 11/12 (1999), pp. 625–653.
[26] J. F. Sturm, *Implementation of interior point methods for mixed semidefinite and second order cone optimization problems*, Optim. Methods Softw., 17 (2002), pp. 1105–1154.
[27] M. J. Todd, *Semidefinite optimization*, Acta Numer., 10 (2001), pp. 515–560.
[28] K. C. Toh, M. J. Todd, and R. H. Tütüncü, *SDPT*3*—A MATLAB software package for semidefinite programming, version* 2.1, Optim. Methods Softw., 11 (1999), pp. 545–581.
[29] H. Wolkowicz, R. Saigal, and L. Vandenberghe, eds., *Handbook of Semidefinite Programming*, Internat. Ser. Oper. Res. Management Sci. 27, Kluwer Academic Publishers, Boston, 2000.
[30] M. H. Wright, *Interior methods for constrained optimization*, Acta Numer., 1 (1992), pp. 341–407.

SIAM J. OPTIM.
Vol. 17, No. 1, pp. 37–51
© 2006 Society for Industrial and Applied Mathematics

CONSTRUCTING GENERALIZED MEAN FUNCTIONS USING CONVEX FUNCTIONS WITH REGULARITY CONDITIONS*

YUN-BIN ZHAO†, SHU-CHERNG FANG‡, AND DUAN LI§

Abstract. The generalized mean function has been widely used in convex analysis and mathematical programming. This paper studies a further generalization of such a function. A necessary and sufficient condition is obtained for the convexity of a generalized function. Additional sufficient conditions that can be easily checked are derived for the purpose of identifying some classes of functions which guarantee the convexity of the generalized functions. We show that some new classes of convex functions with certain regularity (such as S^*-regularity) can be used as building blocks to construct such generalized functions.

Key words. convexity, mathematical programming, generalized mean function, self-concordant functions, S^*-regular functions

AMS subject classifications. 90C30, 90C25, 52A41, 49J52

DOI. 10.1137/040603838

1. Introduction.

In this paper, we denote the n-dimensional Euclidean space by R^n, its nonnegative orthant by R^n_+, and its positive orthant by R^n_{++}.

In 1934, Hardy, Littlewood, and Pólya [13] considered the following function under the name of generalized mean:

$$\Upsilon_w(x) = \phi^{-1}\left(\sum_{i=1}^{n} w_i \phi(x_i)\right), \tag{1.1}$$

where $\phi(\cdot)$ is a real, strictly increasing, convex function defined on a subset of R and $w = (w_1, w_2, \ldots, w_n)^T$ is a given vector in R^n_+. Assuming that $\phi > 0$, $\phi' > 0$, and $\phi'' > 0$, they showed an equivalent condition for the convexity of Υ_w. When ϕ is three times differentiable, Ben-Tal and Teboulle [2] established another equivalent condition for Υ_w being convex (see section 2 for details).

The generalized mean function (1.1) has many applications in optimization. Ben-Tal and Teboulle [2] demonstrated an interesting application of (1.1) (in a continuous form) on penalty functions and duality formulation of stochastic nonlinear programming problems. However, the most widely used generalized means are the logorithmic-exponentional and p-norm functions:

$$f_w(x) = \log\left(\sum_{i=1}^{n} w_i e^{x_i}\right), \quad p_w(x) = \left(\sum_{i=1}^{n} w_i x_i^p\right)^{1/p} \quad \text{for } x = (x_1, \ldots, x_n)^T \in R^n.$$

*Received by the editors February 4, 2004; accepted for publication (in revised form) November 11, 2005; published electronically April 21, 2006.
http://www.siam.org/journals/siopt/17-1/60383.html

†Institute of Applied Mathematics, AMSS, Chinese Academy of Sciences, Beijing 100080, China (ybzhao@amss.ac.cn). This author's work was partially supported by the National Natural Science Foundation of China under grants 10201032 and 70221001.

‡Industrial Engineering and Operations Research, North Carolina State University, Raleigh, NC 26695-7906 (fang@eos.ncsu.edu). Also with the Departments of Mathematical Sciences and Industrial Engineering, Tsinghua University, Beijing, China. This author's work was partially supported by the U.S. Army Research Office grant W911NF-04-D-0003-0002.

§Department of Systems Engineering and Engineering Management, Chinese University of Hong Kong, Shatin, NT, Hong Kong (dli@se.cuhk.edu.hk). This author's work was partially supported by grant CUHK4180/03E, Research Grant Council, Hong Kong.

They correspond to the special cases of Υ_w with $\phi(t) = e^t$ and $\phi(t) = t^p$, respectively.

Needless to say, the log-exp function has been widely used in convex analysis and mathematical programming. For example, a geometric program (see Duffin, Peterson, and Zener [8] and Boyd and Vandenberghe [6]) can be converted into a convex programming problem by using the log-exp function so that the interior-point algorithms can be developed to solve geometric programs with great efficiency (see Kortanek, Xu, and Ye [14]). Another example is concerned with the nondifferentiable minimax problem

$$\min_{y\in D} \max_{1\le i\le n} g_i(y),$$

where $g_i(\cdot)$, $i = 1, \ldots, n$, are real functions defined on a convex set D in R^m. Since the recession function of the log-exp function is the "max-function"(see Rockafellar [20]), i.e., $\max_{1\le i\le n} x_i = \lim_{\varepsilon\to 0_+} \varepsilon f(\frac{x}{\varepsilon})$ where $f(\cdot) = f_w(\cdot)$ and $w = (1, 1, \ldots, 1)$, the above nondifferential optimization problem can be approximated by solving the following optimization problem:

$$\min_{y\in D} \varepsilon \log\left(\sum_{i=1}^{n} e^{\frac{g_i(y)}{\varepsilon}}\right).$$

The objective function is differentiable and convex if every $g_i(y)$ is. Other applications of the log-exp function in optimization can be found in Ben-Tal [1], Ben-Tal and Teboulle [3], Zang [25], Bertsekas [4], Polyak [19], Fang [9], Fang and Tsao [10], Li and Fang [15], Peng and Lin [17], Birbil et al. [5], and Sun and Li [22, 23, 24].

It is worth mentioning that the conjugate function of the log-exp function happens to be the well-known Shannon entropy function [21] which plays a vital role in so many fields ranging from image enhancement to economics and from statistical mechanics to nuclear physics (see Buck and Macaulay [7] and Fang, Rajasekera, and Tsao [11]).

We consider in this paper a further generalization of (1.1) in the form

$$\Gamma_w(x) = \Psi^{-1}\left(\sum_{i=1}^{n} w_i\phi_i(x_i)\right), \tag{1.2}$$

where $\phi_i : \Omega \to R$, $i = 1, \ldots, n$, are convex, twice differentiable (but not necessarily strictly increasing) functions defined on an open convex set $\Omega \subset R$, $\Psi : \Omega \to R$ is convex, twice differentiable, and strictly increasing, and $w \in R^n_+$ is a given vector. Clearly, $\Upsilon_w(\cdot)$ is a special case of $\Gamma_w(\cdot)$ with $\phi_1 = \phi_2 = \cdots = \phi_n = \Psi = \phi$. For convenience, in this paper, we still call Γ_w given by (1.2) a generalized mean function, and we call ϕ_i the inner function and Ψ the outer function of Γ_w.

To ensure the well-definedness of Γ_w, we naturally require that $\sum_{i=1}^{n} Cone[\phi_i(\Omega)] \subseteq \Psi(\Omega)$, where $Cone[\phi_i(\Omega)]$ denotes the cone generated by the set $\phi_i(\Omega)$.

As in the case of Υ_w, we would like to derive certain sufficient and necessary conditions for the function Γ_w to be convex. Moreover, we hope to find a systematic way to explicitly construct some classes of convex Γ_w.

It is interesting to point out that Γ_w is by no means a new research subject. In fact, it was essentially studied by Fenchel in his lecture notes on *Convex Cones, Sets and Functions* in 1953 [12]. Based on the properties of level sets and characteristic roots of Hessian matrices of functions involved, Fenchel derived some sufficient and necessary conditions for the convexity of the generalized mean function Γ_w. The conditions he derived, however, are rather complicated, and there is no simple test to decide what

kind of functions may admit these complicated properties. Unlike Fenchel's approach, our analysis in this paper depends only on the function value, its first derivative, and its second derivative to provide a sufficient and necessary condition for Γ_w being convex. The necessary and sufficient condition we derive in this paper can be viewed as a generalization of that in [13] concerning the function (1.1). We can also use related sufficient conditions to explicitly construct concrete examples of convex Γ_w. Moreover, we show how the so-called S^*-regular functions (to be defined in this paper) can be used to construct convex generalized mean functions.

The rest of the paper is organized as follows. In section 2, we investigate the conditions that ensure the convexity of the generalized mean function Γ_w. In section 3, we identify some classes of functions that satisfy the conditions derived in section 2 and illustrate how the generalized mean function Γ_w can be explicitly constructed. Conclusions are given in the last section.

2. Necessary and sufficient conditions for the convexity of Γ_w. Let us start with a simple lemma (proof omitted) that shows that the inverse of an increasing convex function is concave and increasing.

LEMMA 2.1. *Let Ω be an open convex subset of R and $\Psi : \Omega \to R$ be a real function defined on Ω. Then Ψ is (strictly) convex and strictly increasing if and only if its inverse $\Psi^{-1} : R \to \Omega$ is (strictly) concave and strictly increasing.*

Notice that if $w_i = 0$ for some i, then the term $w_i\phi_i(x)$ can be removed from the expression of $\Gamma_w(x)$, and it suffices to consider Γ_w defined on R^{n-1}. Thus, without loss of generality, we may assume that the vector $w \in R^n_{++}$ throughout the rest of the paper.

To study the convexity of Γ_w, when assuming that ϕ_i, $i = 1, \ldots, n$, and Ψ^{-1} are twice differentiable, we need to check the properties of its Hessian matrix. Let

$$x_w = \sum_{i=1}^{n} w_i\phi_i(x_i).$$

Since $\frac{\partial x_w}{\partial x_i} = w_i\phi_i'(x_i)$, we have

$$\frac{\partial \Gamma_w}{\partial x_i} = (\Psi^{-1})'(x_w)w_i\phi_i'(x_i).$$

Moreover,

$$\frac{\partial^2 \Gamma_w}{\partial x_i^2} = (\Psi^{-1})''(x_w)(w_i\phi_i'(x_i))^2 + (\Psi^{-1})'(x_w)w_i\phi_i''(x_i),$$
$$\frac{\partial^2 \Gamma_w}{\partial x_i \partial x_j} = (\Psi^{-1})''(x_w)w_iw_j\phi_i'(x_i)\phi_j'(x_j) \quad \text{for} \quad i \neq j.$$

Consequently, the Hessian matrix of Γ_w becomes

$$\frac{\partial^2 \Gamma_w}{\partial x^2} = (\Psi^{-1})'(x_w) \begin{bmatrix} w_1\phi_1''(x_1) & 0 & \ldots & 0 \\ 0 & w_2\phi_2''(x_2) & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & w_n\phi_n''(x_n) \end{bmatrix}$$
$$+ (\Psi^{-1})''(x_w) \begin{bmatrix} w_1\phi_1'(x_1) \\ w_2\phi_2'(x_2) \\ \ldots \\ w_n\phi_n'(x_n) \end{bmatrix} [w_1\phi_1'(x_1), w_2\phi_2'(x_2), \ldots, w_n\phi_n'(x_n)]. \tag{2.1}$$

Note that when ϕ_i, $i = 1, \ldots, n$, is convex and Ψ is convex and increasing, by Lemma 2.1, we see that the first term on the right-hand side of (2.1) is a positive semidefinite matrix multiplied by a positive coefficient $(\Psi^{-1})'(x_w)$, while the second is a rank one matrix multiplied by a negative coefficient $(\Psi^{-1})''(x_w)$.

Some conditions for convexity of the function $\Upsilon_w(x)$ have already been studied in [13] and [2]. We summarize their results here.

THEOREM 2.2 (see [13]). *Under the conditions of $\phi > 0$, $\phi' > 0$, and $\phi'' > 0$, the function $\Upsilon_w(x)$ defined by (1.1) is convex if and only if the following condition holds:*

$$\sum_{i=1}^{n} w_i \frac{[\phi'(x_i)]^2}{\phi''(x_i)} \le \frac{[\phi'(y)]^2}{\phi''(y)} \quad \textit{for} \quad y = \Upsilon_w(x).$$

Ben-Tal and Teboulle [2] also provided a different sufficient and necessary condition, under certain assumptions, for the convexity of the function $\Upsilon_w(x)$.

THEOREM 2.3 (see [2]). *Let $\phi(t) \in C^3$. $\Upsilon_w(x)$ is convex if and only if $1/\rho(t)$ is convex, where $\rho(t) = -\phi''/\phi'$.*

It is possible to extend the analysis in [2] for deriving sufficient conditions for the convexity of $\Gamma_w(x)$ defined by (1.2). For example, the following result is actually implied in [2] and can be proved along the line of the proof of "Lemma 1" and "Theorem 1" therein.

THEOREM 2.4 (see [2]). *Let $\Psi(t) \in C^3$ and $\phi_i(t) \in C^3$ be strictly increasing and $\rho(t) = -\Psi''(t)/\Psi'(t)$. If $1/\rho(t)$ is convex and $\Psi^{-1}(\phi_i(t))$ is convex for $i = 1, \ldots, n$, then $\Gamma_w(x)$ given by (1.2) is convex.*

Note that if ϕ is sufficiently smooth, $1/\rho(t)$ is convex, where $\rho(t) = -\phi''(t)/\phi'(t)$, if and only if its second derivative is nonnegative; i.e.,

$$\left(\frac{1}{\rho}\right)'' = \frac{(\phi'')^3\phi''' - 2\phi'\phi''(\phi''')^2 + \phi'(\phi'')^2\phi''''}{(\phi'')^2} \ge 0.$$

Thus, to check the convexity of $1/\rho(t)$, it is usually needed to check the above inequality involving the third and forth derivative of the function ϕ. Theorem 2.2 does not require the third or fourth differentiability of the function ϕ.

In what follows, we generalize the above Theorem 2.2 to the function $\Gamma_w(x)$. Although the basic idea of our proof is essentially tied to that of [13], the proof is not straightforward. For completeness, we give a detailed proof for the result.

THEOREM 2.5. *Let $\Omega \subset R$ be open and convex, $\Psi : \Omega \to R$ be convex, twice differentiable, and strictly increasing, $\phi_i : \Omega \to R$, $i = 1, \ldots, n$, be strictly convex and twice differentiable, and $w \in R^n_{++}$ be a given vector. Then the generalized mean function*

$$\Gamma_w(x) = \Psi^{-1}\left(\sum_{i=1}^{n} w_i \phi_i(x_i)\right)$$

is convex on $\Omega^n := \overbrace{\Omega \times \cdots \times \Omega}^{n}$ if and only if

$$\Psi''(y)\left(\sum_{i=1}^{n} w_i \frac{[\phi_i'(x_i)]^2}{\phi_i''(x_i)}\right) \le [\Psi'(y)]^2 \quad \textit{for} \quad x \in \Omega^n \ \textit{and} \ \ y = \Gamma_w(x). \tag{2.2}$$

Moreover, $\Gamma_w(x)$ is strictly convex if and only if the inequality in (2.2) holds strictly.

Proof. Let $y = \Gamma_w(x) = \Psi^{-1}(x_w)$. Then $x_w = \Psi(y)$ and

$$(\Psi^{-1})'(x_w)\Psi'(y) = 1. \tag{2.3}$$

Differentiating both sides with respect to y and using the above relations, we have

$$\begin{aligned} 0 &= (\Psi^{-1})''(x_w)[\Psi'(y)]^2 + (\Psi^{-1})'(x_w)\Psi''(y) \\ &= (\Psi^{-1})''(x_w)[\Psi'(y)]^2 + \frac{\Psi''(y)}{\Psi'(y)}. \end{aligned}$$

Therefore,

$$(\Psi^{-1})''(x_w) = -\frac{\Psi''(y)}{[\Psi'(y)]^3}. \tag{2.4}$$

Combining (2.3) and (2.4) yields

$$(\Psi^{-1})'(x_w) + (\Psi^{-1})''(x_w)\sum_{i=1}^{n} w_i \frac{[\phi_i'(x_i)]^2}{\phi_i''(x_i)} = \frac{[\Psi'(y)]^2 - \left(\sum_{i=1}^{n} w_i \frac{[\phi_i'(x_i)]^2}{\phi_i''(x_i)}\right)\Psi''(y)}{[\Psi'(y)]^3}. \tag{2.5}$$

First we prove that $\Gamma_w(x)$ is convex if (2.2) holds. It suffices to show that the Hessian matrix of $\Gamma_w(x)$ is positive semidefinite.

For any $d \in R^n$ and $x \in \Omega^n$, the Cauchy–Schwarz inequality implies that

$$\begin{aligned} \left(\sum_{i=1}^{n} w_i \phi_i'(x_i) d_i\right)^2 &= \left(\sum_{i=1}^{n} \left[\sqrt{w_i \phi_i''(x_i)}\, d_i\right] \cdot \sqrt{\frac{w_i}{\phi_i''(x_i)}}\, \phi_i'(x_i)\right)^2 \\ &\le \left(\sum_{i=1}^{n} w_i \phi_i''(x_i) d_i^2\right)\left(\sum_{i=1}^{n} w_i \frac{[\phi_i'(x_i)]^2}{\phi_i''(x_i)}\right). \end{aligned}$$

By Lemma 2.1, we know Ψ^{-1} is concave and hence $(\Psi^{-1})''(x_w) \le 0$ for all x_w. Combining this fact with the above inequality, we see that, for any $d \in R^n$,

$$\begin{aligned} & d^T \frac{\partial^2 \Gamma_w}{\partial x^2} d \\ &= (\Psi^{-1})'(x_w)\left(\sum_{i=1}^{n} w_i \phi_i''(x_i) d_i^2\right) + (\Psi^{-1})''(x_w)\left(\sum_{i=1}^{n} w_i \phi_i'(x_i) d_i\right)^2 \\ &\ge (\Psi^{-1})'(x_w)\left(\sum_{i=1}^{n} w_i \phi_i''(x_i) d_i^2\right) + (\Psi^{-1})''(x_w)\left(\sum_{i=1}^{n} w_i \phi_i''(x_i) d_i^2\right)\left(\sum_{i=1}^{n} w_i \frac{[\phi_i'(x_i)]^2}{\phi_i''(x_i)}\right) \\ &= \left(\sum_{i=1}^{n} w_i \phi_i''(x_i) d_i^2\right)\left[(\Psi^{-1})'(x_w) + (\Psi^{-1})''(x_w)\left(\sum_{i=1}^{n} \frac{w_i [\phi_i'(x_i)]^2}{\phi_i''(x_i)}\right)\right] \\ &= \left(\sum_{i=1}^{n} w_i \phi_i''(x_i) d_i^2\right) \frac{[\Psi'(y)]^2 - \left(\sum_{i=1}^{n} w_i \frac{[\phi_i'(x_i)]^2}{\phi_i''(x_i)}\right)\Psi''(y)}{[\Psi'(y)]^3} \\ &\ge 0. \end{aligned}$$

The last equality follows from (2.5) and the last inequality follows from the fact that the first quantity on the right-hand side, i.e., $\sum_{i=1}^{n} w_i \phi_i''(x_i) d_i^2$, is nonnegative, and

the second quantity is also nonnegative due to our assumption. Consequently, we have proven that the Hessian matrix $\frac{\partial^2 \Gamma_w}{\partial x^2}$ is positive semidefinite, as desired.

Conversely, we would like to show that inequality (2.2) holds if $\Gamma_w(x)$ is convex. For any vector $0 \neq d \in R^n$, knowing (2.3), (2.4), and the convexity of $\Gamma_w(x)$, we have

$$\begin{aligned} 0 \leq d^T \frac{\partial^2 \Gamma_w}{\partial x^2} d &= (\Psi^{-1})'(x_w) \left(\sum_{i=1}^n w_i \phi_i''(x_i) d_i^2 \right) + (\Psi^{-1})''(x_w) \left(\sum_{i=1}^n w_i \phi_i'(x_i) d_i \right)^2 \\ &= \frac{1}{\Psi'(y)} \left(\sum_{i=1}^n w_i \phi_i''(x_i) d_i^2 \right) - \frac{\Psi''(y)}{\Psi'(y)^3} \left(\sum_{i=1}^n w_i \phi_i'(x_i) d_i \right)^2 \\ &= \left(\sum_{i=1}^n w_i \phi_i''(x_i) d_i^2 \right) \left[\frac{1}{\Psi'(y)} - \frac{\Psi''(y)}{\Psi'(y)^3} \frac{[\sum_{i=1}^n w_i \phi_i'(x_i) d_i]^2}{\sum_{i=1}^n w_i \phi_i''(x_i) d_i^2} \right]. \end{aligned} \tag{2.6}$$

Notice that the above inequality holds for any vector $d \in R^n$. In particular, let

$$d_i = \frac{\phi_i'(x_i)}{\phi_i''(x_i) \sum_{k=1}^n w_k \frac{[\phi_k'(x_k)]^2}{\phi_k''(x_k)}}, \ i = 1, \ldots, n.$$

Then, we have

$$\sum_{i=1}^n w_i \phi_i'(x_i) d_i = 1, \ \sum_{i=1}^n w_i \phi_i''(x_i) d_i^2 = \frac{1}{\sum_{i=1}^n w_i \frac{[\phi_i'(x_i)]^2}{\phi_i''(x_i)}}.$$

As a result, the inequality (2.6) reduces to

$$0 \leq \left(\frac{1}{\sum_{i=1}^n w_i \frac{[\phi_i'(x_i)]^2}{\phi_i''(x_i)}} \right) \left[\frac{1}{\Psi'(y)} - \frac{\Psi''(y)}{\Psi'(y)^3} \left(\sum_{i=1}^n w_i \frac{[\phi_i'(x_i)]^2}{\phi_i''(x_i)} \right) \right].$$

We see that inequality (2.2) indeed holds. The result about strict convexity can be easily checked out. □

Theorem 2.5 generalizes the result of Theorem 2.2 (concerning $\Upsilon_w(x)$) to the more general function $\Gamma_w(x)$, while Theorem 2.4 generalizes the sufficient condition of Theorem 2.3 (concerning $\Upsilon_w(x)$) to the function $\Gamma_w(x)$. Except for some very simple cases, like e^t or x^p, these results do not give us the concrete class of functions which can be used to construct of the examples of generalized mean functions. The purpose of the remainder of this paper is to provide a way to identify the desired class of functions. Our analysis here is based only on the result of Theorem 2.5 instead of Theorem 2.4. We believe that there are also some systematic ways to identify the desired class of function based on Theorem 2.4.

To this end, two related sufficiency results of Theorem 2.5 are derived below for convenient usage in constructing convex Γ_w (see section 3).

THEOREM 2.6. *Let Ω be an open convex subset of R, $\Psi : \Omega \to R$ be strictly increasing, twice differentiable, and convex, $\phi_i : \Omega \to R$, $i = 1, \ldots, n$, be strictly convex and twice differentiable, and $w \in R^n_{++}$ be a given vector. Assume that there exists a scalar $\alpha \in R$ such that*

$$\alpha \Psi(t) \Psi''(t) \leq [\Psi'(t)]^2 \quad \textit{for } t \in \Omega. \tag{2.7}$$

Then the function Γ_w *is convex on* Ω^n *if*

$$\sum_{i=1}^{n} \frac{w_i[\phi_i'(x_i)]^2}{\phi_i''(x_i)} \le \alpha\Psi(y) \quad \text{for } x \in \Omega^n, \tag{2.8}$$

where $y = \Gamma_w(x)$.

Proof. Multiplying both sides of (2.8) by $\Psi''(y)$ and applying (2.7), we see that condition (2.2) holds. The result follows from Theorem 2.5 immediately. □

THEOREM 2.7. *Let* Ω *be an open convex subset of* R, $\Psi : \Omega \to R$ *be strictly increasing, twice differentiable, and convex,* $\phi_i : \Omega \to R$, $i = 1, \ldots, n$, *be strictly convex and twice differentiable, and* $w \in R^n_{++}$ *be a given vector. Assume that there exist* $0 \ne \alpha_i \in R$, $i = 1, \ldots, n$, *holding the same sign such that*

$$\alpha_i\phi_i(t)\phi_i''(t) \ge [\phi_i'(t)]^2 \quad \text{for } t \in \Omega, \tag{2.9}$$

and there exists an $\alpha \in R$ *such that the inequality* (2.7) *holds. Then the function* Γ_w *is convex if*

$$\alpha \ge \max_{1\le i\le n} \alpha_i \quad (\text{when } \alpha_i > 0 \text{ for all } i), \tag{2.10}$$

or

$$\alpha \le \min_{1\le i\le n} \alpha_i \quad (\text{when } \alpha_i < 0 \text{ for all } i). \tag{2.11}$$

Proof. Taking $y = \Gamma_w(x)$, we see two cases.

Case 1. $\alpha_i > 0$ for $i = 1, \ldots, n$. In this case, (2.9) implies that $\phi_i(t) \ge 0$ for $t \in \Omega$ and (2.10) implies that

$$\sum_{i=1}^{n} w_i \frac{[\phi_i'(t)]^2}{\phi_i''(x_i)} \le \sum_{i=1}^{n} w_i\alpha_i\phi_i(x_i) \le \left(\max_{1\le i\le n} \alpha_i\right) \sum_{i=1}^{n} w_i\phi_i(x_i) \le \alpha\Psi(y).$$

Case 2. $\alpha_i < 0$ for $i = 1, \ldots, n$. In this case, (2.9) implies that $\phi_i(t) \le 0$ for $t \in \Omega$ and (2.11) implies that

$$\sum_{i=1}^{n} w_i \frac{[\phi_i'(t)]^2}{\phi_i''(x_i)} \le \sum_{i=1}^{n} w_i\alpha_i\phi_i(x_i) \le \left(\min_{1\le i\le n} \alpha_i\right) \sum_{i=1}^{n} w_i\phi_i(x_i) \le \alpha\Psi(y).$$

Both cases yield (2.8) and the desired result follows from Theorem 2.2. □

A special case of $\phi_1(t) = \phi_2(t) = \cdots = \Psi(t)$ immediately leads to the next result.

COROLLARY 2.8. *Let* Ω *be an open convex set in* R, $\phi : \Omega \to R$ *be a convex, twice differentiable, and strictly increasing function, and* $w \in R^n_{++}$ *be a given vector. If there exists an* $\alpha \ne 0$ *such that*

$$[\phi'(t)]^2 = \alpha\phi(t)\phi''(t) \quad \text{for } t \in \Omega, \tag{2.12}$$

then the function $\Upsilon_w(x) = \phi^{-1}(\sum_{i=1}^{n} w_i\phi(x_i))$ *is convex on* Ω^n.

This result can also follow directly from the aforementioned Theorem 2.3 (due to Ben-Tal and Teboulle [2]). In fact, it is easy to verify that the relation (2.12) implies that the second derivative of ϕ'/ϕ'' is equal to zero, and thus by Theorem 2.3 the function $\Upsilon_w(x)$ is convex.

Remark 2.1. The functions satisfying a differential inequality such as (2.7) are related to the so-called self-concordant barrier function introduced by Nesterov and Nemirovskii [16]. Recall that a C^3 function $\xi : (0,\infty) \to R$ is said to be self-concordant if ξ is convex and there exists a constant $\mu_1 > 0$ such that

$$|\xi'''(t)| \le \mu_1(\xi''(t))^{\frac{3}{2}} \quad \text{for } t \in (0,\infty). \tag{2.13}$$

Moreover, the self-concordant function ξ is called a self-concordant barrier function if there exists a constant $\mu_2 > 0$ such that

$$|\xi'(t)| \le \mu_2[\xi''(t)]^{\frac{1}{2}} \quad \text{for } t \in (0,\infty). \tag{2.14}$$

Combining (2.13) and (2.14) yields

$$\xi'(t)\xi'''(t) \le \mu[\xi''(t)]^2.$$

This indicates that the first-order derivative function of a self-concordant barrier function, i.e., $g(t) := \xi'(t)$, satisfies the inequality (2.7). A self-concordant function $\xi(\cdot)$ itself may also satisfy an inequality like (2.7) or (2.9).

Remark 2.2. The functions satisfying a differential inequality such as (2.7) also appear in convexity theory. Given a twice differentiable function $\phi(t) > 0$ on its domain Ω, we consider the convexity of the function $h(t) := \frac{1}{\phi(t)}$ on Ω. Notice that

$$h''(t) = \frac{2[\phi'(t)]^2 - \phi(t)\phi''(t)}{[\phi(t)]^3} \quad \text{for} \quad t \in \Omega.$$

Hence the function $h(t) = \frac{1}{\phi(t)}$ is convex if and only if the inequality $\phi(t)\phi''(t) \le 2[\phi'(t)]^2$ holds on Ω. Moreover, if $\phi(t)\phi''(t) \le [\phi'(t)]^2$, the convex function $h(t)$ satisfies a reverse inequality, i.e., $h(t)h''(t) \ge [h'(t)]^2$ on Ω.

From this observation, a related question arises. Given a function $\phi(t) > 0$ on Ω and a constant $r > 0$, when will the function $h(t) := \frac{1}{\phi(t)^r}$ become convex and satisfy an inequality such as (2.9)? A straightforward analysis leads to the next result.

THEOREM 2.9. (i) *Let Ω be a convex subset of R and $\phi : \Omega \to (0,\infty)$ be a function. If $\phi(t)\phi''(t) \le [\phi'(t)]^2$ for $t \in \Omega$, then, for any $r > 0$, the function $h(t) := \frac{1}{\phi(t)^r}$ is convex and $h(t)h''(t) \ge [h'(t)]^2$ for $t \in \Omega$. Conversely, if there exists an $r > 0$ such that $h(t) := \frac{1}{\phi(t)^r}$ is convex and $h(t)h''(t) \ge [h'(t)]^2$ for $t \in \Omega$, then $\phi(t)\phi''(t) \le [\phi'(t)]^2$ for $t \in \Omega$.*

(ii) *Let Ω be a convex subset of R, $\tau > 0$, and $\phi : \Omega \to (\tau,\infty)$ be a function. If $\phi(t)\phi''(t) \le [\phi'(t)]^2$ for $t \in \Omega$, then, for any scalar $r > 0$ and $T > 0$, the function $h_T(t) := T + \frac{1}{\phi(t)^r}$ is convex and $\alpha h_T(t)h_T''(t) \ge [h_T'(t)]^2$ for $t \in \Omega$, where $\alpha = \frac{1}{T\tau^r+1}$.*

Proof. For case (i), it is sufficient to see that

$$h''(t) = \frac{r^2(\phi'(t))^2 + r[(\phi'(t))^2 - \phi(t)\phi''(t)]}{\phi(t)^{r+2}},$$

and

$$h(t)h''(t) - [h'(t)]^2 = \frac{r[(\phi'(t))^2 - \phi(t)\phi''(t)]}{\phi(t)^{2(r+1)}}.$$

For case (ii), it is easy to verify that $h_T''(t) = h''(t)$ and

$$\left(\frac{1}{T\phi(t)^r + 1}\right) h_T(t)h_T''(t) - [h_T'(t)]^2 = \frac{r[(\phi'(t))^2 - \phi(t)\phi''(t)]}{\phi(t)^{2(r+1)}}.$$

Then the desired result follows. ∎

The above results indicate that if we have a function ϕ satisfying the inequality (2.7) with $\alpha = 1$, then we may construct a function h from ϕ such that h satisfies the converse differentiable inequality $\alpha h(t)h''(t) \geq [h'(t)]^2$ for some constant α. Moreover, if we take a T-translation of the value of the function h, then the resulting function satisfies the converse differentiable inequality with an α that can be reduced to be smaller than any threshold given in (0,1) provided a suitable choice of $T > 0$. This fact will be used near the end of section 3.

3. Constructing convex generalized mean functions Γ_w. In this section, we try to identify some classes of functions that satisfy inequality (2.7) and/or inequality (2.9) so that we have building blocks for constructing the concrete convex function $\Gamma_w(x)$. First, we give a result that identifies functions satisfying (2.12). Obviously, this class of functions satisfies both inequalities (2.7) and (2.9).

THEOREM 3.1. *Let Ω be an open set in R and $\phi : \Omega \to R$ be a convex, twice differentiable, and strictly increasing function satisfying* (2.12) *with a constant $\alpha \neq 0$. Then, the following hold:*

(i) *When $\alpha = 1$, ϕ is in the form of $\phi(t) = \gamma e^{\frac{t}{\beta}}$ for some $\gamma > 0$ and $\beta > 0$.*

(ii) *When $0 < \alpha \neq 1$ with $v^* := \sup_{t\in\Omega} \frac{1-\alpha}{\alpha} t$ being finite, ϕ is in the form of*

$$\phi(t) = \gamma \left(\frac{\alpha - 1}{\alpha} t + \beta \right)^{\frac{\alpha}{\alpha-1}}$$

for some $\gamma > 0$ and $\beta \geq v^*$.

(iii) *When $\alpha < 0$ with $u^* := \sup_{t\in\Omega} \frac{\alpha-1}{\alpha} t$ being finite, ϕ is in the form of*

$$\phi(t) = -\gamma \left(\beta - \frac{\alpha - 1}{\alpha} t \right)^{\frac{\alpha}{\alpha-1}}$$

for some $\gamma > 0$ and $\beta \geq u^*$.

Note that results (i) and (ii) were pointed out in [2] and [13] and result (iii) can be easily derived. The above result leads to the following consequence related to Υ_w.

COROLLARY 3.2. *The following functions can be used to explicitly construct a convex generalized mean function $\Upsilon_w(x) = \phi^{-1}(\sum_{i=1}^n w_i \phi(x_i))$ over Ω^n:*

(i) *$\phi(t) = \gamma e^{\frac{t}{\beta}}$ over $\Omega = R$ with $\gamma > 0$ and $\beta > 0$.*

(ii) *$\phi(t) = \gamma\left(\frac{1}{p}t + \beta\right)^p$ over $\Omega = (\eta, \infty)$ with $p > 1$, $\gamma > 0$, and $\beta \geq -\frac{\eta}{p}$.*

(iii) *$\phi(t) = \frac{\gamma}{\left(\beta - \frac{1}{p}t\right)^p}$ over $\Omega = (-\infty, \eta)$ with $p > 0$, $\gamma > 0$, and $\beta \geq -\frac{\eta}{p}$.*

(iv) *$\phi(t) = -\gamma(\beta - \frac{1}{p}t)^p$ over $\Omega = (-\infty, \eta)$ with $0 < p < 1$, $\gamma > 0$, and $\beta \geq \frac{\eta}{p}$.*

Again, results (i) and (ii) were given in [2] and [13] and results (iii) and (iv) can be easily derived. The functions listed in Corollary 3.2 actually form a complete basis in the sense that the function ϕ in case (i) satisfies condition (2.12) with $\alpha = 1$; the function ϕ in case (ii) satisfies condition (2.12) with $\alpha = \frac{p}{p-1} > 1$; the function ϕ in case (iii) satisfies condition (2.12) with $\alpha = \frac{p}{p+1} \in (0, 1)$; and the function ϕ in (iv) satisfies condition (2.12) with $\alpha = \frac{p}{p-1} < 0$.

We now try to identify some class of functions that satisfies inequalities (2.7) and/or (2.9). For simplicity, we consider only convex, twice differentiable, strictly increasing functions ϑ on $\Omega = (0, \infty)$. Let us first define the following four categories of such functions:

$$\mathcal{U}_1 = \{\vartheta : \text{There exists } \alpha \in R \text{ such that } \alpha\vartheta(t)\vartheta''(t) \geq [\vartheta'(t)]^2 \text{ for } t \in \Omega\};$$

$\mathcal{U}_2 = \{\vartheta : \text{There exists } \alpha \in R \text{ such that } \alpha\vartheta(t)\vartheta''(t) \le [\vartheta'(t)]^2 \text{ for } t \in \Omega\}$;
$\mathcal{U}_3 = \{\vartheta : \text{There exist } \alpha_1 \le \alpha_2 \text{ such that } \alpha_1\vartheta(t)\vartheta''(t) \le [\vartheta'(t)]^2 \le \alpha_2\vartheta(t)\vartheta''(t)$ for $t \in \Omega\}$;
$\mathcal{U}_4 = \{\vartheta : \text{There exists } \alpha \in R \text{ such that } \alpha\vartheta(t)\vartheta''(t) = [\vartheta'(t)]^2 \text{ for all } t \in \Omega\}$.

It is evident that

$$\mathcal{U}_4 \subset \mathcal{U}_3 \subset (\mathcal{U}_2 \cap \mathcal{U}_1).$$

As pointed out in Theorem 3.1, the class $\mathcal{U}_4$ can be given explicitly. By allowing $\alpha_1 \neq \alpha_2$, we show that $\mathcal{U}_3$ is much broader than $\mathcal{U}_4$. In fact, many convex functions with certain regularities fall into the category $\mathcal{U}_3$. To start, we introduce a new class of functions with certain regularity properties.

DEFINITION 3.3. *A convex, twice differentiable, strictly increasing function* $\delta(t) : (0,\infty) \to R$ *is called an* S^*-regular function *if* (i) $\delta(t)$ *vanishes at* $t = 0$ *in the sense of*

$$\lim_{t\to 0_+} \delta(t) = \lim_{t\to 0_+} \delta'(t) = \lim_{t\to 0_+} \delta''(t) = 0;$$

and (ii) *there exist positive constants* $0 < \beta_1 \le \beta_2$, $p \ge 1$, *and* $q \ge 1$ *such that*

$$\beta_1[(t+1)^{p-1} - (t+1)^{-1-q}] \le \delta''(t) \le \beta_2[(t+1)^{p-1} - (t+1)^{-1-q}],\ t > 0. \tag{3.1}$$

Note that condition (3.1) actually implies the strict convexity of an S^*-regular function on $(0,\infty)$. In particular, setting $\beta_1 = \beta_2$, condition (3.1) reduces to an equation

$$\delta''(t) = (t+1)^{p-1} - (t+1)^{-1-q}. \tag{3.2}$$

Taking integration twice and noting that $\lim_{t\to 0_+} \delta(t) = \lim_{t\to 0_+} \delta'(t) = 0$, the unique solution to (3.2) is

$$\Delta_{p,q}(t) = \frac{(t+1)^{p+1} - 1}{p(p+1)} - \frac{(t+1)^{1-q} - 1}{q(q-1)} - \frac{p+q}{pq}t \quad \text{for } p \ge 1 \text{ and } \ q > 1. \tag{3.3}$$

In addition, since $\lim_{q\to 1_+}[1 - (t+1)^{1-q}]/(q-1) = \ln(t+1)$, we have

$$\Delta_{p,1}(t) = \frac{(t+1)^{p+1} - 1}{p(p+1)} + \ln(t+1) - \frac{p+1}{p}t \quad \text{for } p \ge 1. \tag{3.4}$$

Taking $p = 1$ in (3.4), we have

$$\Delta_{1,1}(t) = \frac{(t+1)^2 - 1}{2} + \ln(t+1) - 2t = \frac{1}{2}t^2 - t + \ln(t+1). \tag{3.5}$$

Moreover, taking $p = 1$ and $q = 2$ in (3.3) yields

$$\Delta_{1,2}(t) = \frac{1}{2}\left[(t+1)^2 - (t+1)^{-1} - 3t\right]. \tag{3.6}$$

In terms of this particular solution $\Delta_{p,q}(t)$, condition (3.1) can be written as

$$\beta_1\Delta''_{p,q}(t) \le \delta''(t) \le \beta_2\Delta''_{p,q}(t). \tag{3.7}$$

By integrating and noting that $\lim_{t\to 0_+}\delta'(t)=\lim_{t\to 0_+}\delta(t)=0$, we further have

$$\beta_1\Delta'_{p,q}(t)\le\delta'(t)\le\beta_2\Delta'_{p,q}(t) \tag{3.8}$$

and

$$\beta_1\Delta_{p,q}(t)\le\delta(t)\le\beta_2\Delta_{p,q}(t). \tag{3.9}$$

Therefore, we can see that the class of S^*-regular functions is quite broad. Later, by using (3.7)–(3.9), we show that S^*-regular functions fall into the category $\mathcal{U}_3$.

It is worth mentioning that for any $p\ge 1, q>1$ (including the case of $q\to 1_+$) the S^*-regular function $\Delta_{p,q}(t)$ is not self-concordant. In fact, the function $\Delta_{p,q}(t)$ does not satisfy the inequality (2.13) since $\delta''(t)\to 0$ and $\delta'''(t)\to p+q$ as $t\to 0_+$.

S^*-regular functions are somewhat analogous to (but different from) the self-regular functions defined in [18]. As we have mentioned above, S^*-regular functions are not self-concordant; however, the class of self-regular functions has a large overlap with self-concordant functions. In what follows, we display the relation among the first and second derivatives of S^*-regular functions, which shows that any S^*-regular function belongs to the category $\mathcal{U}_3$. It should be mentioned that the relations among the first and second derivative for the self-regular function have been studied in [18].

THEOREM 3.4. *Let $\delta(t):(0,\infty)\to R$ be S^*-regular on $(0,\infty)$. Then there exist $c_2\ge c_1>0$ such that*

$$c_1\le\frac{\delta(t)\delta''(t)}{[\delta'(t)]^2}\le c_2\quad \textit{for all } t\in(0,\infty), \tag{3.10}$$

i.e., the function $\delta(t)\in\mathcal{U}_3$.

Proof. We show only that an S^*-regular function $\Delta_{p,q}(t)$ satisfies the property (3.10). Actually, we have

$$\frac{\Delta_{p,q}(t)\Delta''_{p,q}(t)}{[\Delta'_{p,q}(t)]^2}=\frac{\left(\frac{(t+1)^{p+1}-1}{p(p+1)}-\frac{(t+1)^{1-q}-1}{q(q-1)}-\frac{p+q}{pq}t\right)\left[(t+1)^{p-1}-(t+1)^{-1-q}\right]}{\left(\frac{(t+1)^p}{p}+\frac{(t+1)^{-q}}{q}-\frac{p+q}{pq}\right)^2}.$$

Dividing the numerator and denominator of the right-hand side of the above equation by $(t+1)^{2p}=(t+1)^{p+1}(t+1)^{p-1}$, we have

$$\frac{\Delta_{p,q}(t)\Delta''_{p,q}(t)}{[\Delta'_{p,q}(t)]^2}=\frac{\left(\frac{1-(t+1)^{-(p+1)}}{p(p+1)}+\frac{(t+1)^{-(p+1)}-(t+1)^{-(p+q)}}{q(q-1)}-\frac{(p+q)t}{pq(t+1)^{(p+1)}}\right)\left(1-\frac{1}{(t+1)^{(p+q)}}\right)}{\left(\frac{1}{p}+\frac{1}{q(t+1)^{(p+q)}}-\frac{p+q}{pq(t+1)^p}\right)^2}.$$

Therefore,

$$\lim_{t\to\infty}\frac{\Delta_{p,q}(t)\Delta''_{p,q}(t)}{[\Delta'_{p,q}(t)]^2}=\frac{p}{p+1}. \tag{3.11}$$

Since $\Delta''_{p,q}(t)=(t+1)^{p-1}-(t+1)^{-1-q}$, we have $\lim_{t\to 0_+}\Delta'''_{p,q}(t)=p+q$. Since $\Delta''_{p,q}(t)\to 0$, $\Delta'_{p,q}(t)\to 0$, and $\Delta_{p,q}(t)\to 0$ as $t\to 0_+$, we have

$$\lim_{t\to 0_+}\frac{(\Delta''_{p,q}(t))^2}{\Delta'_{p,q}(t)}=\lim_{t\to 0_+}\frac{[(\Delta''_{p,q}(t))^2]'}{[\Delta'_{p,q}(t)]'}=\lim_{t\to 0_+}\frac{2\Delta''_{p,q}(t)\Delta'''_{p,q}(t)}{\Delta''_{p,q}(t)}=2(p+q).$$

Hence, we have

$$\lim_{t\to 0_+}\frac{\Delta_{p,q}(t)}{2\Delta'_{p,q}(t)\Delta''_{p,q}(t)}=\lim_{t\to 0_+}\frac{\Delta'_{p,q}(t)}{2[\Delta''_{p,q}(t)]^2+2\Delta'_{p,q}(t)\Delta'''_{p,q}(t)}=\frac{1}{6(p+q)}.$$

Using the above relations, we further have

$$\begin{aligned}\lim_{t\to 0_+}\frac{\Delta_{p,q}(t)\Delta''_{p,q}(t)}{[\Delta'_{p,q}(t)]^2}&=\lim_{t\to 0_+}\frac{\Delta'_{p,q}(t)\Delta''_{p,q}(t)+\Delta_{p,q}(t)\Delta'''_{p,q}(t)}{2\Delta'_{p,q}(t)\Delta''_{p,q}(t)}\\&=\frac{1}{2}+\lim_{t\to 0_+}\frac{\Delta_{p,q}(t)}{2\Delta'_{p,q}(t)\Delta''_{p,q}(t)}\lim_{t\to 0_+}\Delta'''_{p,q}(t)=\frac{2}{3}.\end{aligned}\tag{3.12}$$

Notice that $\Delta_{p,q}(t)>0, \Delta''_{p,q}(t)>0$, and $\Delta'_{p,q}(t)>0$ in $(0,\infty)$. From (3.11) and (3.12), we can see by continuity that there exist two constants $\mu_2\ge\mu_1>0$ such that

$$\mu_1\le\frac{\Delta_{p,q}(t)\Delta''_{p,q}(t)}{[\Delta'_{p,q}(t)]^2}\le\mu_2\quad\text{for } t\in(0,\infty).$$

Together with (3.7) through (3.9), this implies that an S^*-regular function $\delta(t)$ satisfies the following inequality:

$$0<\mu_1\beta_1\le\frac{\delta(t)\delta''(t)}{[\delta'(t)]^2}\le\beta_2\mu_2.$$

Therefore, (3.10) holds with $c_1:=\mu_1\beta_1$ and $c_2:=\mu_2\beta_2$. □

A fact that should be pointed out here is that new functions in $\mathcal{U}_1$ or $\mathcal{U}_2$ can be constructed by using the basic operations (addition, multiplication, division, and composition) on known functions. The proof of the following fact is omitted.

LEMMA 3.5. (i) *If $\phi:(0,\infty)\to(0,\infty)$, $\phi\in\mathcal{U}_1$ with $\alpha=\alpha_1$, and $\varphi:(0,\infty)\to(0,\infty)$, $\varphi\in\mathcal{U}_1$ with $\alpha=\alpha_2$, then $\phi+\varphi\in\mathcal{U}_1$ with $\alpha=2\max\{\alpha_1,\alpha_2\}$.*

(ii) *If $\phi:(0,\infty)\to(0,\infty)$, $\phi\in\mathcal{U}_1$ with $\alpha_1\in(0,1]$, and $\varphi:(0,\infty)\to(0,\infty)$, $\varphi\in\mathcal{U}_1$ with $\alpha_2\in(0,1]$, then the multiplicative function $\phi(t)\cdot\varphi(t)\in\mathcal{U}_1$ with $\alpha=1$. Similarly, if $\phi\in\mathcal{U}_2$ with $\alpha_1\ge 1$ and $\varphi\in\mathcal{U}_2$ with $\alpha_2\ge 1$, then $\phi(t)\cdot\varphi(t)\in\mathcal{U}_2$ with $\alpha=1$.*

(iii) *If $\phi:(0,\infty)\to(0,\infty)$, $\phi\in\mathcal{U}_2$ with $\alpha_1\ge 1$, and $\varphi:(0,\infty)\to(0,\infty)$, $\varphi\in\mathcal{U}_1$ with $\alpha_2\in(0,1]$, then the function $\frac{\phi}{\varphi}\in\mathcal{U}_2$ with $\alpha=1$. Similarly, if $\phi\in\mathcal{U}_1$ with $\alpha_1\in(0,1]$ and $\varphi\in\mathcal{U}_2$ with $\alpha_2\ge 1$, then $\frac{\phi}{\varphi}\in\mathcal{U}_1$ with $\alpha=1$.*

(iv) *Let $\varphi:(0,\infty)\to\Omega_1\subset R$ and $\phi:\Omega_1\to(0,\infty)$ be two convex functions. If $\phi\in\mathcal{U}_1$ with $\alpha>0$, then the composite function $(\phi\circ\varphi)(t)=\phi(\varphi(t))\in\mathcal{U}_1$ with the same constant α.*

The next result shows that the composite functions of e^t belong to $\mathcal{U}_3$.

LEMMA 3.6. *Denote the exponential function e^t by $\exp(t)$ and the composition of m ($m\ge 1$) exponential functions by*

$$\theta_m(t):=\overbrace{(\exp\circ\exp\circ\cdots\circ\exp)}^{m}(t).$$

Then

$$\frac{1}{m}\theta_m(t)\theta''_m(t)\le[\theta'_m(t)]^2\le\theta_m(t)\theta''_m(t)\quad\text{for } t\in R.\tag{3.13}$$

Proof. Let $\alpha_m(t) := [\theta'_m(t)]^2/(\theta_m(t)\theta''_m(t))$ for $t \in R$. Since $\alpha_1(t) \equiv 1$, we can prove the right-hand side inequality of (3.13) using (iv) of Lemma 3.5 and mathematical induction. For the left-hand side inequality, notice that

$$\theta'_m(t) = \theta_m(t)\theta'_{m-1}(t),\ \theta''_m(t) = \theta_m(t)(\theta'_{m-1}(t))^2 + \theta_m(t)\theta''_{m-1}(t) \ \text{ for } t \in R.$$

This indicates that

$$\alpha_m(t) = \frac{1}{1 + \frac{1}{\alpha_{m-1}(t)\theta_{m-1}(t)}} > \frac{1}{1 + \frac{1}{\alpha_{m-1}(t)}} \ \text{ for } t \in R.$$

It is easy to check that $\alpha_2(t) \in (\frac{1}{2}, 1)$. The desired result follows by induction. ∎

To construct examples of the convex function Γ_w, Theorem 2.7 tells us that it suffices to find functions satisfying the inequalities (2.7) and (2.9) and compare their α values. The next result is to estimate the α values, or, equivalently, to estimate the values of c_1 and c_2 in (3.10). For simplicity, we use the S^*-regular functions with $p = 1$ and $q = 1, 2$ to estimate the required c_1 and c_2. In fact, we have the following result. Its proof was omitted here.

LEMMA 3.7. *The S^*-regular functions $\Delta_{1,1}(t)$ and $\Delta_{1,2}(t)$ given by* (3.5) *and* (3.6)*, respectively, satisfy condition* (3.10) *with $c_1 = \frac{1}{2}$ and $c_2 = \frac{2}{3}$; that is,*

$$\frac{3}{2}\Delta_{1,1}(t)\Delta''_{1,1}(t) \le [\Delta'_{1,1}(t)]^2 \le 2\Delta_{1,1}(t)\Delta''_{1,1}(t), \tag{3.14}$$

$$\frac{3}{2}\Delta_{1,2}(t)\Delta''_{1,2}(t) \le [\Delta'_{1,2}(t)]^2 \le 2\Delta_{1,2}(t)\Delta''_{1,2}(t) \tag{3.15}$$

for $t \in (0, \infty)$.

We now give the last result on how to construct some convex functions Γ_w.

THEOREM 3.8. *Let Ω be an open convex subset of R.*

(i) *Let $\phi : \Omega \to (0, \infty)$ be a convex, twice differentiable, strictly increasing function on Ω. If $\phi(t)\phi''(t) \le [\phi'(t)]^2$ for $t \in \Omega$, then the generalized mean function*

$$\Gamma_w^{(1)}(x) := \phi^{-1}\left(\sum_{i=1}^n \frac{w_i}{\phi(x_i)^r}\right)$$

is convex on Ω^n for any given $w \in R^n_{++}$ and $r > 0$.

(ii) *Let $\kappa > 0$ be a constant and $\phi : \Omega \to (\kappa, \infty)$ be a convex, twice differentiable, strictly increasing function satisfying the inequality $\phi(t)\phi''(t) \le [\phi'(t)]^2$ for $t \in \Omega$. Then, for any given $w \in R^n_{++}$ and $T > 0, r > 0$, the function*

$$\Gamma_w^{(2)}(x) := \overbrace{\ln \circ \ln \circ \cdots \circ \ln}^{\ell}\left(\sum_{i=1}^n w_i\left(T + \frac{1}{\phi(x_i)^r}\right)\right)$$

is convex on Ω^n for any positive integer $\ell \le T\kappa^r + 1$.

In fact, result (i) comes from part (i) of Theorem 2.9 and Theorem 2.7. Result (ii) follows from Lemma 3.6, Theorem 2.7, and part (ii) of Theorem 2.9. In fact, it suffices to take the inner function $h_T(t) = T + \frac{1}{\phi(t)^r}$ and outer function $\theta_m(t)$, as defined in Lemma 3.6, whose inverse function is given by $\overbrace{\ln \circ \ln \cdots \circ \ln}^{m}(t)$.

The above result partially answers the following interesting question: *Given a convex function, how many times can* log*-transformations be applied while retaining the convexity?*

Using Theorems 2.7, 2.9, and 3.8 and Lemma 3.7, we have the following examples of convex Γ_w.

Example 3.1.

(i) $\Delta_{1,j}^{-1}\left[\sum_{i=1}^{n}\frac{1}{\Delta_{1,j}(x_i)^r}\right]$,

(ii) $\ln\left(\sum_{i=1}^{n}\frac{1}{\Delta_{1,j}(x_i)^r}\right)$,

(iii) $\overbrace{\ln\circ\ln\circ\cdots\circ\ln}^{\ell\le m+1}\left(\sum_{i=1}^{n}\left(m+e^{-rx_i}\right)\right),\ x\in(0,\infty)^n.$

(iv) $\overbrace{\ln\circ\ln\circ\cdots\circ\ln}^{\ell}\left[\sum_{i=1}^{n}\left(m+\frac{1}{\Delta_{1,1}(x_i)^r}\right)\right],\ x\in(\tau,\infty)^n, \ell\le m\Delta_{1,1}(\tau)^r+1, \tau>0.$

It follows from Corollary 3.2 that the function x^p over $(0,\infty)$ satisfies (2.12) with $\alpha=\frac{p}{p-1}$. Hence, when $1<p\le 2$, we have $\alpha\ge 2$, and when $1<p\le\frac{29}{17}$, we have $\alpha\ge\frac{29}{12}\ge\frac{9}{4}$. By Lemma 3.7, both $\Delta_{1,2}(t)$ and $\Delta_{1,1}(t)$ satisfy condition (2.9) with $\alpha=2$. From Theorem 2.7, we see that the functions below are examples of convex Γ_w.

Example 3.2. Let $1<p\le 2$ and $\delta_i(t)=\Delta_{1,2}(t)$ or$\Delta_{1,1}(t)$ for $t\in(0,\infty)$ and $i=1,\ldots,n$. Then $\Gamma_w(x)=(\sum_{i=1}^{n}w_i\delta_i(x_i))^{\frac{1}{p}}$ is convex on $(0,\infty)^n$.

Before closing this section, we briefly illustrate a possible application of involving function Γ_w in the regularization method for solving a nonlinear programming problem:

$$\min\{f_0(x): x\in C\}.$$

For simplicity, we assume that C is a convex set and f_0 is a convex function. Let $\mu>0$ be a positive parameter. Given a strictly convex function Γ_w, we consider the following problem:

$$\min\{f_0(x)+\mu\Gamma_w(x): x\in C\}.$$

This problem becomes a strictly convex programming problem with a unique solution, denoted by $x(\mu)$, which comprises a continuation trajectory $\{x(\mu):\mu>0\}$. Under suitable conditions of f_0,Φ, and ϕ, this trajectory becomes bounded. In this case, by setting $\mu\to 0$, any accumulation point of $x(\mu)$, as $\mu\to 0$, is a solution to the original problem. Thus, a path-following algorithm can be designed to follow this trajectory to achieve the solution of the original problem. The performance of such a path-following algorithm certainly depends on the choice of the function Γ_w with regularity conditions.

4. Conclusions. In this paper, we have further extended the theoretical foundation for the generalized mean function. We have established a necessary and sufficient condition for such a generalization to be convex. Moreover, a systematic way to explicitly construct convex Γ_w has been illustrated. To this end, the concept of S^*-regular functions has been introduced. It should be noted that any S^*-regular function is not self-concordant [16].

Acknowledgments. We would like to thank the two anonymous referees for their insightful comments which helped improve significantly the results and presentation of the paper. We are especially grateful to one of the referees who brought the references [2] and [13] to our attention and provided Theorem 2.4 in this paper.

REFERENCES

[1] A. Ben-Tal, *The entropic penalty approach to stochastic programming*, Math. Oper. Res., 10 (1985), pp. 263–279.
[2] A. Ben-Tal and M. Teboulle, *Expected utility, penalty functions, and duality in stochastic nonlinear programming*, Management Sci., 32 (1986), pp. 1445–1466.
[3] A. Ben-Tal and M. Teboulle, *A smoothing technique for nondifferentiable optimization problems*, in Optimization, Lecture Notes in Math. 1405, Springer-Verlag, Berlin, 1989, pp. 1–11.
[4] D. P. Bertsekas, *Constrained Optimization and Lagrangian Multiplier Methods*, Academic Press, New York, 1982.
[5] S. I. Birbil, S.-C. Fang, J. Frenk, and S. Zhang, *Recursive approximate of the high dimensional MAX function*, Oper. Res. Lett., 33 (2005), pp. 450–458.
[6] S. Boyd and L. Vandenberghe, *Introduction to Convex Optimization with Engineering Applications*, Stanford University, Stanford, CA, 1997.
[7] B. Buck and V. A. Macaulay, *Maximum Entropy in Action: A Collection of Expository Essays*, Oxford University Press, Oxford, UK, 1991.
[8] R. J. Duffin, E. L. Peterson, and C. Zener, *Geometric Programming—Theory and Applications*, Wiley, New York, 1967.
[9] S.-C. Fang, *An unconstrained convex programming view of linear programming*, Math. Methods Oper. Res., 36 (1992), pp. 149–161.
[10] S.-C. Fang and H. S. J. Tsao, *On the entropic perturbation and exponential penalty methods for linear programming*, J. Optim. Theory Appl., 89 (1996), pp. 461–466.
[11] S.-C. Fang, J. R. Rajasekera, and H. Tsao, *Entropy Optimization and Mathematical Programming*, Kluwer Academic Publishers, Boston, 1997.
[12] W. Fenchel, *Convex Sets, Cones, and Functions*, Lectures in Princeton University, Princeton, NJ, 1953.
[13] G. Hardy, J. L. Littlewood, and G. Pólya, *Inequalities*, Cambridge University Press, Cambridge, UK, 1934.
[14] K. O. Kortanek, X. Xu, and Y. Ye, *An infeasible interior-point algorithm for solving primal and dual geometric programs*, Math. Programming, 76 (1996), pp. 155–181.
[15] X.-S. Li and S.-C. Fang, *On the regularization method for solving min-max problems with applications*, Math. Methods Oper. Res., 46 (1997), pp. 119–130.
[16] Y. Nesterov and A. Nemirovskii, *Interior Point Polynomial Algorithms in Convex Programming*, SIAM Studies in Applied Mathematics 13, SIAM, Philadelphia, 1994.
[17] J. Peng and Z. Lin, *A non-interior continuation method for generalized linear complementarity problems*, Math. Program., 86 (1999), pp. 533–563.
[18] J. Peng, C. Roos, and T. Terlaky, *Self-Regularity: A New Paradigm for Primal-Dual Interior-Point Algorithms*, Princeton University Press, Princeton, NJ, 2002.
[19] R. A. Polyak, *Smooth optimization methods for minimax problems*, SIAM J. Control Optim., 26 (1988), pp. 1274–1286.
[20] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[21] C. E. Shannon, *A Mathematical Theory of Communication*, Bell System Technical Journal, 27 (1948), pp. 379–423, 623–656.
[22] X. L. Sun and D. Li, *Value-estimation function method for constrained global optimization*, J. Optim. Theory Appl., 102 (1999), pp. 385–409.
[23] X. L. Sun and D. Li, *Logarithmic-exponential penalty formulation for integer programming*, Appl. Math. Lett., 12 (1999), pp. 73–77.
[24] X. L. Sun and D. Li, *Asymptotic strong duality for bounded integer programming: A logarithmic-exponential dual formulation*, Math. Oper. Res., 25 (2000), pp. 625–644.
[25] I. Zang, *A smoothing technique for min-max optimization*, Math. Programming, 19 (1980), pp. 61–77.

SIAM J. OPTIM.
Vol. 17, No. 1, pp. 52–77
© 2006 Society for Industrial and Applied Mathematics

INTERIOR METHODS FOR MATHEMATICAL PROGRAMS WITH COMPLEMENTARITY CONSTRAINTS*

SVEN LEYFFER†, GABRIEL LÓPEZ-CALVA‡, AND JORGE NOCEDAL§

Abstract. This paper studies theoretical and practical properties of interior-penalty methods for mathematical programs with complementarity constraints. A framework for implementing these methods is presented, and the need for adaptive penalty update strategies is motivated with examples. The algorithm is shown to be globally convergent to strongly stationary points, under standard assumptions. These results are then extended to an interior-relaxation approach. Superlinear convergence to strongly stationary points is also established. Two strategies for updating the penalty parameter are proposed, and their efficiency and robustness are studied on an extensive collection of test problems.

Key words. mathematical programs with complementarity constraints, nonlinear programming, interior-point methods, exact penalty, equilibrium constraints, complementarity constraints

AMS subject classifications. 90C30, 90C33, 90C51, 49M37, 65K10

DOI. 10.1137/040621065

1. Introduction.

In this paper we study the numerical solution of mathematical programs with complementarity constraints (MPCCs) of the form

$$\begin{aligned}
&\text{minimize} && f(x) && \text{(1.1a)}\\
&\text{subject to} && c_i(x) = 0, \quad i \in \mathcal{E}, && \text{(1.1b)}\\
& && c_i(x) \geq 0, \quad i \in \mathcal{I}, && \text{(1.1c)}\\
& && 0 \leq x_1 \perp x_2 \geq 0. && \text{(1.1d)}
\end{aligned}$$

The variables have been divided as $x = (x_0, x_1, x_2)$, with $x_0 \in \mathbb{R}^n$, $x_1, x_2 \in \mathbb{R}^p$. The complementarity condition (1.1d) stands for

$$x_1 \geq 0,\ x_2 \geq 0 \ \text{ and either } \ x_{1i} = 0 \ \text{ or } \ x_{2i} = 0 \ \text{ for } \ i = 1, \ldots, p, \tag{1.2}$$

where x_{1i}, x_{2i} are the ith components of vectors x_1 and x_2, respectively.

Complementarity (1.2) represents a logical condition (a disjunction) and must be expressed in analytic form if we wish to solve the MPCC using nonlinear programming

*Received by the editors December 17, 2004; accepted for publication (in revised form) November 14, 2005; published electronically April 21, 2006. This work was supported in part by National Science Foundation grant CCR-0219438 and Department of Energy grant DE-FG02-87ER25047-A004. Support was also provided by the Mathematical, Information, and Computational Sciences Division subprogram of the Office of Advanced Scientific Computing Research, Office of Science, U.S. Department of Energy, under Contract W-31-109-ENG-38. The U.S. Government retains for itself, and others acting on its behalf, a paid-up, nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government. Copyright is owned by SIAM to the extent not limited by these rights.

http://www.siam.org/journals/siopt/17-1/62106.html

†Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439 (leyffer@mcs.anl.gov).

‡Industrial Engineering and Management Sciences Department, Northwestern University, Evanston, IL 60208 (g-lopez-calva@northwestern.edu).

§Electrical Engineering and Computer Science Department, Northwestern University, Evanston, IL 60208 (nocedal@ece.northwestern.edu).

methods. A popular reformulation of the MPCC is

$$\begin{aligned}
&\text{minimize} && f(x) && \text{(1.3a)}\\
&\text{subject to} && c_i(x) = 0, \quad i \in \mathcal{E}, && \text{(1.3b)}\\
& && c_i(x) \geq 0, \quad i \in \mathcal{I}, && \text{(1.3c)}\\
& && x_1 \geq 0, \quad x_2 \geq 0, && \text{(1.3d)}\\
& && x_{1i}x_{2i} \leq 0, \quad i = 1, \ldots, p. && \text{(1.3e)}
\end{aligned}$$

This formulation preserves the solution set of the MPCC but is not totally adequate because it violates the Mangasarian–Fromowitz constraint qualification (MFCQ) at any feasible point. This lack of regularity can create problems when applying *classical* nonlinear programming algorithms. For example, sequential quadratic programming (SQP) methods can give rise to inconsistent constraint linearizations. Interior methods exhibit inefficiencies caused by the conflicting goals of enforcing complementarity while keeping the variables x_1, x_2 away from their bounds.

Modern nonlinear programming algorithms include, however, regularization techniques and other safeguards to deal with degeneracy, and one cannot rule out the possibility that they can cope with the difficulties created by the formulation (1.3) without having to exploit the special structure of MPCCs. If this level of robustness could be attained (and this is a laudable goal) there might be no need to develop algorithms specifically for MPCCs.

Numerical experiments by Fletcher and Leyffer [12] suggest that this goal is almost achieved by modern active-set SQP methods. In [12], FILTERSQP [11] was used to solve the problems in the MacMPEC collection [18], which contains more than a hundred MPCCs, and fast convergence was almost always observed. The reason for this practical success is that, even though the formulation (1.3) fails to satisfy MFCQ, it is locally equivalent to a nonlinear program that satisfies MFCQ, and a robust SQP solver is able to identify the right set of active constraints in the well-behaved program and converge to a solution. Failures, however, are still possible for the SQP approach. Fletcher et al. [13] give several examples that illustrate ways in which an SQP method may fail to converge.

Interior methods are less successful when applied directly to the nonlinear programming formulation (1.3). Fletcher and Leyffer [12] tested LOQO [25] and KNITRO [4] and observed that they were slower and less reliable than the SQP solvers FILTERSQP and SNOPT [15] (all codes as of 2002). This result contrasts starkly with the experience in nonlinear programming, where interior methods compete well with SQP methods. These studies have stimulated considerable interest in developing interior methods for MPCCs that guarantee both global convergence and efficient practical performance. The approaches can be broadly grouped into two categories.

The first category comprises relaxation approaches, where (1.3) is replaced by a family of problems in which (1.3e) is changed to

$$x_{1i}x_{2i} \leq \theta, \quad i = 1, \ldots, p, \tag{1.4}$$

and the relaxation parameter $\theta > 0$ is driven to zero. This type of approach has been studied from a theoretical perspective by Scholtes [24] and Ralph and Wright [22]. Interior methods based on the relaxation (1.4) have been proposed by Liu and Sun [19] and Raghunathan and Biegler [21]. In both studies, the parameter θ is proportional to the barrier parameter μ and is updated only at the end of each barrier problem. Raghunathan and Biegler focus on local analysis and report very good numerical

results on the MacMPEC collection. Liu and Sun analyze global convergence of their algorithm and report limited numerical results. Numerical difficulties may arise when the relaxation parameter gets small, since the interior of the regularized problem shrinks toward the empty set.

DeMiguel et al. [8] address this problem by proposing a different relaxation scheme where, in addition to (1.4), the nonnegativity bounds on the variables are relaxed to

$$x_{1i} \geq -\delta, \quad x_{2i} \geq -\delta. \tag{1.5}$$

Under fairly general assumptions, their algorithm drives either θ or δ, but not both, to zero. This provides the resulting family of problems with a strict interior, even when the appropriate relaxation parameters are approaching zero, which is a practical advantage over the previous relaxation approach. The drawback is that the algorithm has to correctly identify the parameters that must be driven to zero, a requirement that can be difficult to meet in some cases.

The second category involves a regularization technique based on an exact-penalty reformulation of the MPCC. Here, (1.3e) is moved to the objective function in the form of an ℓ_1-penalty term, so that the objective becomes

$$f(x) + \pi x_1^T x_2, \tag{1.6}$$

where $\pi > 0$ is a penalty parameter. If π is chosen large enough, the solution of the MPCC can be recast as the minimization of a single penalty function. The appropriate value of π is, however, unknown in advance and must be estimated during the course of the minimization.

This approach was first studied by Anitescu [1] in the context of active-set SQP methods, although it had been used before to solve engineering problems (see, e.g., [10]). It has been adopted as a heuristic to solve MPCCs with interior methods in LOQO by Benson et al. [3], who present very good numerical results on the MacMPEC set. A more general class of exact penalty functions was analyzed by Hu and Ralph [17], who derive global convergence results for a sequence of penalty problems that are solved exactly. Anitescu [2] derives similar global results in the context of inexact subproblem solves.

In this paper, we focus on the penalization approach, because we view it as a general tool for handling degenerate nonlinear programs. Our goal is to study global and local convergence properties of interior-penalty methods for MPCCs and to propose efficient and robust practical implementations.

In section 2 we present the interior-penalty framework; some examples motivate the need for proper updating strategies for the penalty parameter. Section 3 shows that the proposed interior-penalty method converges globally to strongly stationary points, under standard assumptions. These results are then extended to the interior-relaxation approaches considered in [19] and [21]. In section 4 we show that, near a solution that satisfies some standard regularity properties, the penalty parameter is not updated and the iterates converge superlinearly to the solution. Section 5 presents two practical implementations of the interior-penalty method with different updating strategies for the penalty parameter. Our numerical experiments, reported in the same section, favor a dynamic strategy that assesses the magnitude of the penalty parameter at every iteration.

2. An interior-penalty method for MPCCs. To circumvent the difficulties caused by the complementarity constraints, we replace (1.3) by the ℓ_1-penalty problem

$$\begin{aligned} &\text{minimize} && f(x) + \pi x_1^T x_2, \\ &\text{subject to} && c_i(x) = 0, \quad i \in \mathcal{E}, \\ &&& c_i(x) \geq 0, \quad i \in \mathcal{I}, \\ &&& x_1 \geq 0,\ x_2 \geq 0, \end{aligned} \tag{2.1}$$

where $\pi > 0$ is a penalty parameter. In principle, the ℓ_1-penalty term should have the form $\sum_i \max\{0, x_{1i}x_{2i}\}$, but we can write it as $x_1^T x_2$ if we enforce the nonnegativity of x_1, x_2. This exact penalty reformulation of MPCCs has been studied in [1, 2, 3, 17, 22, 23]. Since problem (2.1) is smooth, we can safely apply standard nonlinear programming algorithms, such as interior methods, to solve it. The barrier problem associated to (2.1) is

$$\begin{aligned} &\text{minimize} && f(x) + \pi x_1^T x_2 - \mu \sum_{i \in \mathcal{I}} \log s_i - \mu \sum_{i=1}^{p} \log x_{1i} - \mu \sum_{i=1}^{p} \log x_{2i} \\ &\text{subject to} && c_i(x) = 0, \quad i \in \mathcal{E}, \\ &&& c_i(x) - s_i = 0, \quad i \in \mathcal{I}, \end{aligned} \tag{2.2}$$

where $\mu > 0$ is the barrier parameter and $s_i > 0$, $i \in \mathcal{I}$, are slack variables. The Lagrangian of this barrier problem is given by

$$\begin{aligned} \mathcal{L}_{\mu,\pi}(x, s, \lambda) = f(x) + \pi x_1^T x_2 - \mu \sum_{i \in \mathcal{I}} \log s_i - \mu \sum_{i=1}^{p} \log x_{1i} - \mu \sum_{i=1}^{p} \log x_{2i} \\ - \sum_{i \in \mathcal{E}} \lambda_i c_i(x) - \sum_{i \in \mathcal{I}} \lambda_i (c_i(x) - s_i), \end{aligned} \tag{2.3}$$

and the first-order Karush–Kuhn–Tucker (KKT) conditions of (2.2) can be written as

$$\begin{aligned} \nabla f(x) - \nabla c_{\mathcal{E}}(x)^T \lambda_{\mathcal{E}} - \nabla c_{\mathcal{I}}(x)^T \lambda_{\mathcal{I}} - \begin{pmatrix} 0 \\ \mu X_1^{-1} e - \pi x_2 \\ \mu X_2^{-1} e - \pi x_1 \end{pmatrix} = 0, \\ s_i \lambda_i - \mu = 0, \quad i \in \mathcal{I}, \\ c_i(x) = 0, \quad i \in \mathcal{E}, \\ c_i(x) - s_i = 0, \quad i \in \mathcal{I}, \end{aligned} \tag{2.4}$$

where we have grouped the components $c_i(x), i \in \mathcal{E}$, into the vector $c_{\mathcal{E}}(x)$, and similarly for $c_{\mathcal{I}}(x), \lambda_{\mathcal{E}}, \lambda_{\mathcal{I}}$. We also define $\lambda = (\lambda_{\mathcal{E}}, \lambda_{\mathcal{I}})$. X_1 denotes the diagonal matrix containing the elements of x_1 on the diagonal (the same convention is used for X_2 and S), and e is a vector of ones of appropriate dimension.

The KKT conditions (2.4) can be expressed more compactly as

$$\nabla_x \mathcal{L}_{\mu,\pi}(x, s, \lambda) = 0, \tag{2.5a}$$

$$S\lambda_{\mathcal{I}} - \mu e = 0, \tag{2.5b}$$

$$c(x, s) = 0, \tag{2.5c}$$

where we define

$$c(x,s) = \begin{pmatrix} c_{\mathcal{E}}(x) \\ c_{\mathcal{I}}(x) - s \end{pmatrix}. \tag{2.6}$$

In Figure 1, we describe an interior method for MPCCs based on the ℓ_1-penalty formulation. Here, and in the rest of the paper, $\|\cdot\|$ denotes the infinity norm. This is consistent with our implementation; it also simplifies the exposition, without compromising the generality of our results.

In addition to requiring that the optimality conditions (2.7) of the barrier problem are satisfied approximately, we impose a reduction in the complementarity term by means of (2.8). For now, the only requirement on the sequence of barrier parameters $\{\mu^k\}$ and the stopping tolerances $\{\epsilon^k_{pen}\}, \{\epsilon^k_{comp}\}$ is that they all converge to 0 as $k \to \infty$. Later, in the local analysis of section 4, we impose further conditions on the relative rate of convergence of these sequences.

Algorithm I: Interior-Penalty Method for MPCCs

Initialization: Let x^0, s^0, λ^0 be the initial primal and dual variables. Set $k = 1$.

repeat

1. Choose a barrier parameter μ^k, stopping tolerances ϵ^k_{pen} and ϵ^k_{comp}
2. Find π^k and an approximate solution (x^k, s^k, λ^k) of problem (2.2) with parameters μ^k and π^k that satisfy $x^k_1 > 0, x^k_2 > 0, s^k > 0, \lambda^k_{\mathcal{I}} > 0$ and the following conditions:

$$\|\nabla_x \mathcal{L}_{\mu^k,\pi^k}(x^k, s^k, \lambda^k)\| \le \epsilon^k_{pen}, \tag{2.7a}$$

$$\|S^k \lambda^k_{\mathcal{I}} - \mu^k e\| \le \epsilon^k_{pen}, \tag{2.7b}$$

$$\|c(x^k, s^k)\| \le \epsilon^k_{pen}, \tag{2.7c}$$

and

$$\|\min\{x^k_1, x^k_2\}\| \le \epsilon^k_{comp} \tag{2.8}$$

3. Let $k \leftarrow k + 1$

until a stopping test for the MPCC is satisfied.

FIG. 1. *An interior-penalty method for MPCCs.*

We use $\|\min\{x^k_1, x^k_2\}\|$ in (2.8) as a measure of complementarity, rather than $x^{kT}_1 x^k_2$, because it is less sensitive to the scaling of the problem and is independent of the number of variables. Moreover, this measure is accurate even when both x^k_{1i} and x^k_{2i} converge to zero.

Our formulation of Algorithm I is sufficiently general to permit various updating strategies for the penalty parameter in step 2. One option is to choose μ^k and solve (2.2) with $\pi^k = \pi^{k-1}$, until conditions (2.7) are satisfied. If condition (2.8) also holds, then we proceed to step 3. Otherwise, we increase π^k and solve (2.2) again using the same barrier parameter μ^k. The process is repeated, if necessary, until (2.8) is satisfied. We show in section 5 that Algorithm I with this penalty update strategy is much more robust and efficient than the direct application of an interior method to (1.3). Nevertheless, there are some flaws in a strategy that holds the penalty

parameter fixed throughout the minimization of a barrier problem, as illustrated by the following examples.

The results reported next were obtained with an implementation of Algorithm I that uses the penalty update strategy described in the previous paragraph. The initial parameters are $\pi^1 = 1, \mu^1 = 0.1$, and we set $\epsilon^k_{comp} = (\mu^k)^{0.4}$ for all k. When the penalty parameter is increased, it is multiplied by 10. The other details of the implementation are discussed in section 5 and are not relevant to the discussion that follows.

Example 1 (`ralph2`). Consider the MPCC

$$\begin{aligned} &\text{minimize} && x_1^2 + x_2^2 - 4x_1x_2 \\ &\text{subject to} && 0 \le x_1 \perp x_2 \ge 0, \end{aligned} \tag{2.9}$$

whose solution is $(0, 0)$. The associated penalty problem is

$$\begin{aligned} &\text{minimize} && (x_1 - x_2)^2 + (\pi - 2)x_1x_2 \\ &\text{subject to} && x_1 \ge 0,\ x_2 \ge 0, \end{aligned} \tag{2.10}$$

which is unbounded for any $\pi < 2$. Starting with $\pi^1 = 1$, the first barrier problem is never solved. The iterates increase monotonically because, by doing so, the objective function is reduced and feasibility is maintained for problem (2.10). Eventually, the iterates diverge. Table 1 shows the values of x_1x_2 during the first eight iterations of the inner algorithm in step 2.

TABLE 1
Complementarity values for problem `ralph2`.

Iterate	1	2	3	4	5	6	7	8
Complementarity	0.0264	0.0916	0.1480	51.70	63.90	79.00	97.50	120.0

The upward trend in complementarity should be taken as a warning sign that the penalty parameter is not large enough, since no progress is made toward satisfaction of (2.8). This suggests that we should be prepared to increase the penalty parameter dynamically. How to do so, in a robust manner, is not a simple question because complementarity can oscillate. We return to this issue in section 5, where we describe a dynamic strategy for updating the penalty parameter. □

Example 2 (`scale1`). Even if the penalty problem is bounded, there are cases where efficiency can be improved with a more flexible strategy for updating π^k. For example, consider the MPCC

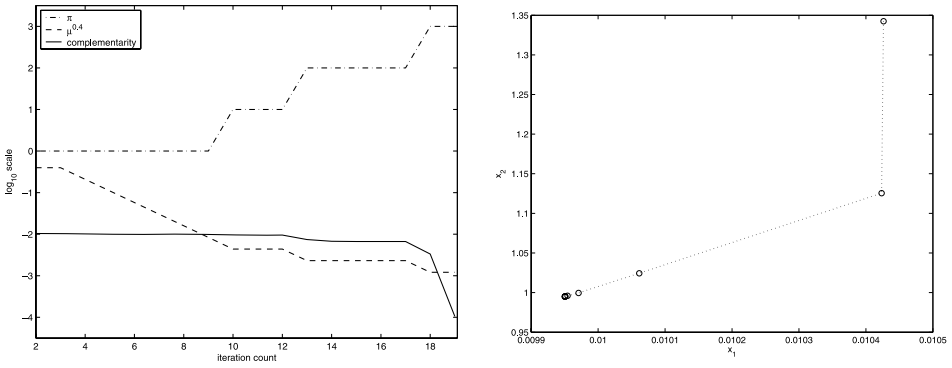
$$\begin{aligned} &\text{minimize} && (100x_1 - 1)^2 + (x_2 - 1)^2 \\ &\text{subject to} && 0 \le x_1 \perp x_2 \ge 0, \end{aligned} \tag{2.11}$$

which has two local solutions: $(0.01, 0)$ and $(0, 1)$. Table 2 shows the first seven values of x^k satisfying (2.7) and (2.8), and the corresponding values of μ^k.

We observe that complementarity, as measured by $\min\{x_1^k, x_2^k\}$, stagnates. This result is not surprising because the minimum penalty parameter required to recover the solution $(0, 1)$ is $\pi^* = 200$ and we have used the value $\pi^1 = 1$. In fact, for any $\pi < 200$, there is a saddle point close to $(0, 1)$, and the iterates approach that saddle point. Seven barrier problems must be solved before the test (2.8) is violated for the first time, triggering the first update of π^k.

TABLE 2
Solutions of 7 *consecutive barrier-penalty problems for* `scale1`.

k	1	2	3	4	5	6	7
μ^k	0.1	0.02	0.004	0.0008	0.00016	0.000032	0.0000064
x_1^k	0.010423	0.010061	0.009971	0.009954	0.009951	0.009950	0.009950
x_2^k	1.125463	1.024466	0.999634	0.995841	0.995186	0.995057	0.995031
ϵ_{comp}^k	0.398107	0.209128	0.109856	0.057708	0.030314	0.015924	0.008365



(a) Complementarity gap. (b) Solution path.

FIG. 2. *A numerical solution of problem* `scale1`.

The behavior of the algorithm is illustrated in Figure 2(a), which plots three quantities as a function of the inner iterations. Complementarity (continuous line) stalls at a nonzero value during the first 10 iterations, while μ^k (dashed line) decreases monotonically. The penalty parameter (dashed-dotted line) is increased for the first time at iteration 9. It must be increased three times to surpass the threshold value $\pi^* = 200$, which finally forces complementarity down to zero. Figure 2(b) shows the path of the iterates up to the solution of the seventh barrier problem. There is a clear pattern of convergence to the stationary point where none of the variables is zero. If this convergence pattern can be identified early, the penalty parameter can be increased sooner, saving some iterations in the solution of the MPCC. □

One could ask whether the penalty parameter needs to be updated at all, or whether choosing a very large value of π and holding it fixed during the execution of Algorithm I could prove to be an effective strategy. In section 5 we show that excessively large penalty parameters can result in substantial loss of efficiency. More important, no matter how large π is, for some problems the penalty function is unbounded outside a small neighborhood of the solution, and a bad initial point makes the algorithm diverge if π is kept fixed (see [20] for an example).

In section 5, we describe a dynamic strategy for updating the penalty parameter. We show that it is able to promptly identify the undesirable behavior described in these examples and to react accordingly.

3. Global convergence analysis. In this section, we present the global convergence analysis of an interior-penalty method. We start by reviewing an MPCC constraint qualification that suffices to derive first-order optimality conditions for MPCCs. We then review stationarity concepts.

DEFINITION 3.1. *We say that the* MPCC linear independence constraint qualification *(MPCC-LICQ) holds at a feasible point x for the MPCC* (1.1) *if and only if the standard LICQ holds at x for the set of constraints*

$$\begin{aligned} c_i(x) &= 0, \quad i \in \mathcal{E}, \\ c_i(x) &\geq 0, \quad i \in \mathcal{I}, \\ x_1 \geq 0, &\quad x_2 \geq 0. \end{aligned} \tag{3.1}$$

We denote indices of the active constraints at a feasible point x by

$$\begin{aligned} \mathcal{A}_c(x) &= \{i \in \mathcal{I} : c_i(x) = 0\}, \\ \mathcal{A}_1(x) &= \{i \in \{1, \ldots, p\} : x_{1i} = 0\}, \\ \mathcal{A}_2(x) &= \{i \in \{1, \ldots, p\} : x_{2i} = 0\}. \end{aligned} \tag{3.2}$$

For ease of notation, we use $i \notin \mathcal{A}_1(x)$ as shorthand for $i \in \{1, \ldots, p\} \setminus \mathcal{A}_1(x)$ (likewise for $\mathcal{A}_2, \mathcal{A}_c$). We sometimes refer to variables satisfying $x_{1i} + x_{2i} > 0$ as *branch variables*; those for which $x_{1i} + x_{2i} = 0$, that is, variables indexed by $\mathcal{A}_1(x) \cap \mathcal{A}_2(x)$, are called *corner variables*.

The next theorem establishes the existence of multipliers for minimizers that satisfy MPCC-LICQ. It can be viewed as a counterpart for MPCCs of the first-order KKT theorem for nonlinear programs.

THEOREM 3.2. *Let x^* be a minimizer of the MPCC* (1.1), *and suppose MPCC-LICQ holds at x^*. Then, there exist multipliers $\lambda^*, \sigma_1^*, \sigma_2^*$ that, together with x^*, satisfy the system*

$$\nabla f(x) - \nabla c_{\mathcal{E}}(x)^T \lambda_{\mathcal{E}} - \nabla c_{\mathcal{I}}(x)^T \lambda_{\mathcal{I}} - \begin{pmatrix} 0 \\ \sigma_1 \\ \sigma_2 \end{pmatrix} = 0, \tag{3.3a}$$

$$c_i(x) = 0, \quad i \in \mathcal{E}, \tag{3.3b}$$

$$c_i(x) \geq 0, \quad i \in \mathcal{I}, \tag{3.3c}$$

$$x_1 \geq 0, x_2 \geq 0, \tag{3.3d}$$

$$x_{1i} = 0 \text{ or } x_{2i} = 0, \quad i = 1, \ldots, p, \tag{3.3e}$$

$$c_i(x)\lambda_i = 0, \quad i \in \mathcal{I}, \tag{3.3f}$$

$$\lambda_i \geq 0, \quad i \in \mathcal{I}, \tag{3.3g}$$

$$x_{1i}\sigma_{1i} = 0 \quad \text{and} \quad x_{2i}\sigma_{2i} = 0, \quad i = 1, \ldots, p, \tag{3.3h}$$

$$\sigma_{1i} \geq 0, \sigma_{2i} \geq 0, \quad i \in \mathcal{A}_1(x) \cap \mathcal{A}_2(x). \tag{3.3i}$$

For a proof of this theorem, see [23] or an alternative proof in [20].

We note that the multipliers σ_1, σ_2 are required to be nonnegative only for corner variables. This requirement reflects the geometry of the feasible set: If $x_{1i} > 0$, then $x_{2i} = 0$ acts like an equality constraint, and the corresponding multiplier can be positive or negative. Theorem 3.2 motivates the following definition.

DEFINITION 3.3. (a) *A point x^* is called a* strongly stationary point *of the MPCC* (1.1) *if there exist multipliers $\lambda^*, \sigma_1^*, \sigma_2^*$ such that* (3.3) *is satisfied.* (b) *A point x^* is called a* C-stationary point *of the MPCC* (1.1) *if there exist multipliers $\lambda^*, \sigma_1^*, \sigma_2^*$ such that conditions* (3.3a)–(3.3h) *hold and*

$$\sigma_{1i}^* \sigma_{2i}^* \geq 0, \quad i \in \mathcal{A}_1(x^*) \cap \mathcal{A}_2(x^*). \tag{3.4}$$

Strong stationarity implies the absence of first-order feasible descent directions. These are the points that the algorithms should aim for. Although C-stationarity does not characterize the solutions of an MPCC, since it allows descent directions if $\sigma_{1i} < 0$ or $\sigma_{2i} < 0$, we consider C-stationary points because they are attractors of iterates generated by Algorithm I. One can find examples in which a sequence of stationary points of the penalty problem converge to a C-stationary point where descent directions exist, and this phenomenon can actually be observed in practice (see case 1 of Example 3.1 in [17] and the comments on problem `scale4` in section 5). The reader further interested in stationarity for MPCCs is referred to [23].

3.1. Global convergence of the interior-penalty algorithm. Many algorithms have been proposed to solve the barrier problem in step 2; see, for example, [7, 14] and the references therein. As is well known, these inner algorithms may fail to satisfy (2.7), and therefore Algorithm I can fail to complete step 2. The analysis of the inner algorithm is beyond the scope of this paper, and we concentrate only on the analysis of the outer iterations in Algorithm I. We assume that the inner algorithm is always successful and that Algorithm I generates an infinite sequence of iterates $\{x^k, s^k, \lambda^k\}$ that satisfies conditions (2.7) and (2.8).

We present the following result in the slightly more general setting in which a vector of penalties $\pi = (\pi_1, \ldots, \pi_p)$ is used, with the objective function as

$$f(x) + \pi^T X_1 x_2, \tag{3.5}$$

and with minor changes in the Lagrangian of the problem. This allows us to extend the global convergence result to the relaxation approach in the next subsection. For the implementation, however, we use a uniform (i.e., scalar-valued) penalty.

THEOREM 3.4. *Suppose that Algorithm* I *generates an infinite sequence of iterates* $\{x^k, s^k, \lambda^k\}$ *and parameters* $\{\pi^k, \mu^k\}$ *that satisfies conditions* (2.7) *and* (2.8), *for sequences* $\{\epsilon^k_{pen}\}, \{\epsilon^k_{comp}\}, \{\mu^k\}$ *converging to zero. If* x^* *is a limit point of the sequence* $\{x^k\}$, *and* f *and* c *are continuously differentiable in an open neighborhood* $\mathcal{N}(x^*)$ *of* x^*, *then* x^* *is feasible for the MPCC* (1.1). *If, in addition, MPCC-LICQ holds at* x^*, *then* x^* *is a C-stationary point of* (1.1). *Moreover, if* $\pi^k_i x^k_{ji} \to 0$ *for* $j = 1, 2$ *and* $i \in \mathcal{A}_1(x^*) \cap \mathcal{A}_2(x^*)$, *then* x^* *is a strongly stationary point of* (1.1).

Proof. Let x^* be a limit point of the sequence $\{x^k\}$ generated by Algorithm I, and let $\mathcal{K}$ be an infinite index set such that $\{x^k\}_{k\in\mathcal{K}} \to x^*$. Then, $x^k \in \mathcal{N}(x^*)$ for all k sufficiently large; from the assumption of continuous differentiability on $\mathcal{N}(x^*)$, and $\{x^k\}_{k\in\mathcal{K}} \to x^*$, we conclude that the sequences $\{f(x^k)\}$, $\{c(x^k)\}$, $\{\nabla f(x^k)\}$, $\{\nabla c_{\mathcal{E}}(x^k)\}$, $\{\nabla c_{\mathcal{I}}(x^k)\}$, $k \in \mathcal{K}$, have limit points and are therefore bounded.

Since the inner algorithm used in Step 2 enforces positivity of the slacks s^k, by continuity of c and the condition $\epsilon^k_{pen} \to 0$ we have

$$c_i(x^*) = 0, \quad i \in \mathcal{E},$$

$$c_i(x^*) = s^*_i \geq 0, \quad i \in \mathcal{I},$$

where $s^*_i = \lim_{k\in\mathcal{K}} s^k_i$. Therefore x^* satisfies (3.3b) and (3.3c), and it also satisfies (3.3d) because the inner algorithm enforces the positivity of x^k. The complementarity condition (3.3e) follows directly from (2.8) and $\epsilon^k_{comp} \to 0$. Therefore, x^* is feasible for the MPCC (1.1).

Existence of multipliers. Let us define

$$\sigma_{1i}^k = \frac{\mu^k}{x_{1i}^k} - \pi_i^k x_{2i}^k, \qquad \sigma_{2i}^k = \frac{\mu^k}{x_{2i}^k} - \pi_i^k x_{1i}^k \tag{3.6}$$

and

$$\alpha^k = \|(\lambda^k, \sigma_1^k, \sigma_2^k)\|. \tag{3.7}$$

(Recall $\|\cdot\|$ denotes the infinity norm.) We first show that $\{\alpha^k\}_{k\in\mathcal{K}}$ is bounded, a result that implies that the sequence of multipliers $(\lambda^k, \sigma_1^k, \sigma_2^k)$ has a limit point. Then we show that any limit point satisfies C-stationarity at x^*.

We can assume, without loss of generality, that $\alpha_k \geq \tau > 0$ for all $k \in \mathcal{K}$. Indeed, if there were a further subsequence $\{\alpha_k\}_{k\in\mathcal{K}'}$ converging to 0, this subsequence would be trivially bounded, and we would apply the analysis below to $\{\alpha_k\}_{k\in\mathcal{K}\setminus\mathcal{K}'}$, which is bounded away from 0, to prove the boundedness of the entire sequence $\{\alpha_k\}_{k\in\mathcal{K}}$.

Let us define the "normalized multipliers"

$$\hat{\lambda}^k = \frac{\lambda^k}{\alpha^k}, \quad \hat{\sigma}_1^k = \frac{\sigma_1^k}{\alpha^k}, \quad \hat{\sigma}_2^k = \frac{\sigma_2^k}{\alpha^k}. \tag{3.8}$$

We now show that the normalized multipliers corresponding to inactive constraints converge to 0 for $k \in \mathcal{K}$. Consider an index $i \notin \mathcal{A}_c(x^*)$, where $\mathcal{A}_c$ is defined by (3.2). Since $s_i^k \to c_i(x^*) > 0$ and $s_i^k \lambda_i^k \to 0$ by (2.7b), we have that λ_i^k converges to 0, and so does $\hat{\lambda}_i^k$.

Next consider an index $i \notin \mathcal{A}_1(x^*)$. We want to show that $\hat{\sigma}_{1i}^k \to 0$. If $i \notin \mathcal{A}_1(x^*)$, then $x_{1i}^k \to x_{1i}^* > 0$, which implies that $x_{2i}^k \to 0$, by (2.8) and $\epsilon_{comp}^k \to 0$. We also have, from (3.6), that for any $k \in \mathcal{K}$,

$$\sigma_{1i}^k \neq 0 \quad \Rightarrow \quad \frac{\mu^k}{x_{1i}^k} - \pi_i^k x_{2i}^k \neq 0 \quad \Rightarrow \quad \frac{\mu^k}{x_{2i}^k} - \pi_i^k x_{1i}^k \neq 0 \quad \Rightarrow \quad \sigma_{2i}^k \neq 0. \tag{3.9}$$

Using this and the fact that $|\sigma_{2i}^k| \leq \alpha^k$, we have that, if there is any subsequence of indices k for which $\sigma_{1i}^k \neq 0$, then

$$\begin{aligned} |\hat{\sigma}_{1i}^k| = \frac{|\sigma_{1i}^k|}{\alpha^k} \leq \frac{|\sigma_{1i}^k|}{|\sigma_{2i}^k|} &= \frac{\left|\frac{\mu^k}{x_{1i}^k} - \pi_i^k x_{2i}^k\right|}{\left|\frac{\mu^k}{x_{2i}^k} - \pi_i^k x_{1i}^k\right|} \\ &= \frac{\left|\frac{\mu^k - \pi_i^k x_{1i}^k x_{2i}^k}{x_{1i}^k}\right|}{\left|\frac{\mu^k - \pi_i^k x_{1i}^k x_{2i}^k}{x_{2i}^k}\right|} = \frac{x_{2i}^k}{x_{1i}^k} \to 0. \end{aligned}$$

Since clearly $\hat{\sigma}_{1i}^k \to 0$ for those indices with $\sigma_{1i}^k = 0$, we have that the whole sequence $\hat{\sigma}_{1i}^k$ converges to zero for $i \notin \mathcal{A}_1(x^*)$. The same argument can be applied to show that $\hat{\sigma}_{2i}^k \to 0$ for $i \notin \mathcal{A}_2(x^*)$. Therefore we have shown that the normalized multipliers (3.8) corresponding to the inactive constraints converge to zero for $k \in \mathcal{K}$.

To prove that $\{\alpha^k\}_{k\in\mathcal{K}}$ is bounded, we proceed by contradiction and assume that there exists $\mathcal{K}' \subseteq \mathcal{K}$ such that $\{\alpha^k\}_{k\in\mathcal{K}'} \to \infty$. By definition, the sequences of normalized multipliers (3.8) are bounded, so we restrict $\mathcal{K}'$ further, if necessary, so that the sequences of normalized multipliers are convergent within $\mathcal{K}'$. Given that

$\mathcal{K}' \subseteq \mathcal{K}$, all the sequences of gradients $\{\nabla f(x^k)\}$, $\{\nabla c_{\mathcal{E}}(x^k)\}$, $\{\nabla c_{\mathcal{I}}(x^k)\}$, $k \in \mathcal{K}'$, are convergent. We can then divide both sides of (2.7a) by α^k and take limits to get

$$\lim_{k\to\infty, k\in\mathcal{K}'} \left\| \frac{1}{\alpha^k} \nabla_x \mathcal{L}_{\mu^k,\pi^k}(x^k, s^k, \lambda^k) \right\| \leq \lim_{k\to\infty, k\in\mathcal{K}'} \frac{\epsilon^k_{pen}}{\alpha^k} = 0$$

or

$$\lim_{k\to\infty, k\in\mathcal{K}'} \left[\frac{1}{\alpha^k} \nabla f^k - \sum_{i\in\mathcal{E}\cup\mathcal{I}} \hat{\lambda}^k_i \nabla c_i(x^k) - \begin{pmatrix} 0 \\ \hat{\sigma}^k_1 \\ \hat{\sigma}^k_2 \end{pmatrix} \right] = 0. \tag{3.10}$$

It is immediate that the first term of (3.10) converges to 0. We showed that the coefficients (the normalized multipliers (3.8)) of the inactive constraints also converge to zero. Since the corresponding sequences of gradients have limits (hence are bounded), all the terms corresponding to inactive constraints get cancelled in the limit, and we have

$$\sum_{i\in\mathcal{E}\cup\mathcal{A}_c(x^*)} \hat{\lambda}^*_i \nabla c_i(x^*) + \sum_{i\in\mathcal{A}_1(x^*)} \hat{\sigma}^*_{1i} \begin{pmatrix} 0 \\ e_i \\ 0 \end{pmatrix} + \sum_{i\in\mathcal{A}_2(x^*)} \hat{\sigma}^*_{2i} \begin{pmatrix} 0 \\ 0 \\ e_i \end{pmatrix} = 0.$$

If the limit point x^* satisfies MPCC-LICQ, then the constraint gradients involved in this expression are linearly independent, and we get

$$\hat{\lambda}^* = 0, \;\; \hat{\sigma}^*_1 = 0, \;\; \hat{\sigma}^*_2 = 0.$$

This result, however, contradicts the fact that $\|(\hat{\lambda}^k, \hat{\sigma}^k_1, \hat{\sigma}^k_2)\| = 1$ for all $k \in \mathcal{K}'$, which follows from (3.7), (3.8), and the assumption that $\lim_{k\to\infty, k\in\mathcal{K}'} \alpha^k \to \infty$. Therefore, we conclude that no such unbounded subsequence exists, and hence all the sequences $\{\lambda^k\}, \{\sigma^k_1\}, \{\sigma^k_2\}$, with $k \in \mathcal{K}$, are bounded and have limit points.

C-stationarity. Choose any such limit point $(\lambda^*, \sigma^*_1, \sigma^*_2)$ and restrict $\mathcal{K}$, if necessary, so that

$$(x^k, s^k, \lambda^k, \sigma^k_1, \sigma^k_2) \to (x^*, s^*, \lambda^*, \sigma^*_1, \sigma^*_2).$$

By (2.7a) and (2.4) and by continuity of f and c, we have that

$$\nabla f(x^*) - \nabla c_{\mathcal{E}}(x^*)^T \lambda^*_{\mathcal{E}} - \nabla c_{\mathcal{I}}(x^*)^T \lambda^*_{\mathcal{I}} - \begin{pmatrix} 0 \\ \sigma^*_1 \\ \sigma^*_2 \end{pmatrix} = 0,$$

which proves (3.3a). We have already shown that the limit point x^* satisfies conditions (3.3b) through (3.3e). The nonnegativity of $\lambda^*_{\mathcal{I}}$, condition (3.3g), follows from the fact that the inner algorithm maintains $\lambda^k_i > 0$ for $i \in \mathcal{I}$. Condition (3.3f) holds because, for any $i \in \mathcal{I}$, if $c_i(x^*) = s^*_i > 0$, then since $s^k_i \lambda^k_i \to 0$, we must have $\lambda^*_i = 0$.

We now establish that conditions (3.3h) hold at the limit point $(x^*, s^*, \lambda^*, \sigma^*_1, \sigma^*_2)$. They are clearly satisfied when $i \in \mathcal{A}_1(x^*)$ and $i \in \mathcal{A}_2(x^*)$. Consider an index $i \notin \mathcal{A}_1(x^*)$. If there is any infinite subset $\mathcal{K}'' \subseteq \mathcal{K}$ with $\sigma^k_{1i} \neq 0$ for all $k \in \mathcal{K}''$, then, as argued in (3.9), $\sigma^k_{1i} \neq 0 \Rightarrow \sigma^k_{2i} \neq 0$ for all $k \in \mathcal{K}''$ and

$$\lim_{k\to\infty, k\in\mathcal{K}''} \frac{|\sigma^k_{1i}|}{|\sigma^k_{2i}|} = \lim_{k\to\infty, k\in\mathcal{K}''} \frac{\left| \frac{\mu^k}{\pi^k_i x^k_{1i}} - x^k_{2i} \right|}{\left| \frac{\mu^k}{\pi^k_i x^k_{2i}} - x^k_{1i} \right|} = \lim_{k\to\infty, k\in\mathcal{K}''} \frac{x^k_{2i}}{x^k_{1i}} = 0, \tag{3.11}$$

where the limit follows from the fact that $x_{1i}^* > 0$, which implies that $x_{2i}^k \to 0$. $\{\sigma_{2i}^k\}$ has a limit and is therefore bounded. Hence, (3.11) can hold only if $\lim_{k\to\infty, k\in\mathcal{K}''} \sigma_{1i}^k = 0$ and, by definition, $\sigma_{1i}^k = 0$ for all $k \in \mathcal{K} \setminus \mathcal{K}''$. We conclude that $\sigma_{1i}^* = 0$ for $i \notin \mathcal{A}_1(x^*)$. A similar argument can be used to get $\sigma_{2i}^* = 0$ if $i \notin \mathcal{A}_2(x^*)$.

To prove (3.4), we consider an index $i \in \mathcal{A}_1(x^*) \cap \mathcal{A}_2(x^*)$. If $\sigma_{1i}^* = 0$, we immediately have $\sigma_{1i}^*\sigma_{2i}^* = 0$. If $\sigma_{1i}^* > 0$, then for all $k \in \mathcal{K}$ large enough, $\sigma_{1i}^k > 0$. Then

$$\frac{\mu^k}{x_{1i}^k} > \pi_i^k x_{2i}^k \quad \Rightarrow \quad \frac{\mu^k}{x_{2i}^k} > \pi_i^k x_{1i}^k,$$

or $\sigma_{2i}^k > 0$. Hence, $\sigma_{1i}^*\sigma_{2i}^* \geq 0$, as desired. The same argument can be used to show that if $\sigma_{1i}^* < 0$, then $\sigma_{2i}^* < 0$, and hence $\sigma_{1i}^*\sigma_{2i}^* \geq 0$. Therefore, condition (3.4) holds, and x^* is a C-stationary point of the MPCC.

Strong stationarity. Let $i \in \mathcal{A}_1(x^*) \cap \mathcal{A}_2(x^*)$. If $\pi_i^k x_{2i}^k \to 0$, then

$$\sigma_{1i}^* = \lim_{k\in\mathcal{K}} \sigma_{1i}^k = \lim_{k\in\mathcal{K}} \left(\frac{\mu^k}{x_{1i}^k} - \pi_i^k x_{2i}^k \right) = \lim_{k\in\mathcal{K}} \frac{\mu^k}{x_{1i}^k} \geq 0. \tag{3.12}$$

A similar argument shows that $\sigma_{2i}^* \geq 0$. Therefore, condition (3.3i) holds, and x^* is a strongly stationary point for the MPCC (1.1). □

The proof of Theorem 3.4 builds on a similar proof in [17], where an analogous result is derived for *exact* subproblem solves. Our result is related to the analysis in [2] (derived independently), except that we explicitly work within an interior-method framework and we do not analyze the convergence of the inner algorithm. In [2], stronger assumptions are required (e.g., that the lower-level problem satisfies a mixed-P property) to guarantee that the inner iteration always terminates.

For strong stationarity, we require a condition on the behavior of the penalty parameter, relative to the sequences converging to the corners. This is the same condition that Scholtes required for strong stationarity in [24]. A simpler, though stronger, assumption on the penalties is a boundedness condition, which we use for the following corollary that corresponds to the particular case of our implementations.

COROLLARY 3.5. *Suppose Algorithm* I *is applied with a uniform (i.e., scalar-valued) penalty parameter, and let the assumptions of Theorem* 3.4 *hold. Then, if the sequence of penalty parameters* $\{\pi^k\}$ *is bounded,* x^* *is a strongly stationary point for* (1.1). □

In our algorithmic framework, the sequence of penalty parameters does not have to be monotone, although practical algorithms usually generate nondecreasing sequences. Monotonicity is required neither in the description of the algorithm nor in the proof. This flexibility could be exploited to correct unnecessarily large penalty parameters in practice. For theoretical purposes, on the other hand, this nonmonotonicity property is important for the derivation of Theorem 3.6 in the next subsection.

3.2. Relationship to interior-relaxation methods. An alternative to exact penalization for regularizing the complementarity constraints of an MPCC is to relax the complementarity constraints. This approach has been combined with interior methods in [19, 21]; we refer to it as the "interior-relaxation" method. The objective of this subsection is to show that there is a correspondence between interior-penalty and interior-relaxation approaches and that this correspondence can be exploited to give an alternative global convergence proof for an interior-relaxation method, based on Theorem 3.4.

Interior-relaxation methods solve a sequence of barrier subproblems associated with (1.3) with one modification; namely, the complementarity constraints (1.3e) are relaxed by introducing a parameter $\theta^k > 0$ that goes to 0 as the barrier parameter μ^k approaches 0. Effectively, a sequence of problems

$$\begin{aligned} \text{minimize} \quad & f(x) - \mu^k \sum_{i\in\mathcal{I}} \log s_i - \mu^k \sum_{i=1}^{p} \log s_{ci} - \mu^k \sum_{i=1}^{p} \log x_{1i} - \mu^k \sum_{i=1}^{p} \log x_{2i} \\ \text{subject to} \quad & c_i(x) = 0, \qquad i \in \mathcal{E}, \\ & c_i(x) - s_i = 0, \qquad i \in \mathcal{I}, \\ & \theta^k - x_{1i}x_{2i} - s_{ci} = 0, \qquad i = 1,\dots,p, \end{aligned} \tag{3.13}$$

has to be solved, where s_c are the slacks for the relaxed complementarity constraints, the multipliers of which are denoted by ξ. Let $\mathcal{L}_{\mu^k,\theta^k}$ denote the Lagrangian of (3.13).

An approximate solution of (3.13), for some μ^k and θ^k, is given by variables $x^k, s^k, s_c^k, \lambda^k, \xi^k$, with $x_1^k > 0, x_2^k > 0, s^k > 0, s_c^k > 0, \lambda_{\mathcal{I}}^k > 0, \xi^k > 0$, satisfying the following inexact KKT system, where $\epsilon_{rel}^k > 0$ is some tolerance

$$\|\nabla_x \mathcal{L}_{\mu^k,\theta^k}(x^k, s^k, \lambda^k, \xi^k)\| \le \epsilon_{rel}^k, \tag{3.14a}$$

$$\|S^k \lambda_{\mathcal{I}}^k - \mu^k e\| \le \epsilon_{rel}^k, \tag{3.14b}$$

$$\|S_c^k \xi^k - \mu^k e\| \le \epsilon_{rel}^k, \tag{3.14c}$$

$$\|c(x^k, s^k)\| \le \epsilon_{rel}^k, \tag{3.14d}$$

$$\|\theta^k e - X_1^k x_2^k - s_c^k\| \le \epsilon_{rel}^k. \tag{3.14e}$$

THEOREM 3.6. *Suppose an interior-relaxation method generates an infinite sequence of solutions* $\{x^k, s^k, s_c^k, \lambda^k, \xi^k\}$ *and parameters* $\{\mu^k, \theta^k\}$ *that satisfies conditions* (3.14) *for sequences* $\{\mu^k\}, \{\theta^k\}$, *and* $\{\epsilon_{rel}^k\}$, *all converging to 0. If* x^* *is a limit point of the sequence* $\{x^k\}$, *and* f *and* c *are continuously differentiable in an open neighborhood* $\mathcal{N}(x^*)$ *of* x^*, *then* x^* *is feasible for the MPCC* (1.1). *If, in addition, MPCC-LICQ holds at* x^*, *then* x^* *is a C-stationary point of* (1.1). *Moreover, if* $\xi_i^k x_{ji}^k \to 0$ *for* $j = 1, 2$ *and* $i \in \mathcal{A}_1(x^*) \cap \mathcal{A}_2(x^*)$, *then* x^* *is a strongly stationary point of* (1.1).

Proof. We provide an indirect proof. Given sequences of variables $\{x^k,\ s^k,\ s_c^k,\ \lambda^k,\ \xi^k\}$, parameters $\{\mu^k, \theta^k\}$, and tolerances $\{\epsilon_{rel}^k\}$ satisfying the assumptions, we define sequences of parameters $\{\mu^k, \pi^k := \xi^k\}$ and tolerances $\{\epsilon_{pen}^k := \epsilon_{rel}^k, \epsilon_{comp}^k := (\theta^k + \epsilon_{rel}^k)^{1/2}\}$; for the variables, we keep $\{x^k, s^k, \lambda^k\}$ only. Note that we have not changed the sequence of decision variables $\{x^k\}$, so the limit points are unchanged. We show that the sequences that we just defined satisfy the assumptions of Theorem 3.4. Observe that there is no reason why the sequence of multipliers $\{\xi^k\}$ should be monotone. This is not a problem, however, because there is no monotonicity requirement for the sequence $\{\pi^k\}$ in Theorem 3.4, as noted earlier.

First, $\{\mu^k\}, \{\epsilon_{pen}^k\}, \{\epsilon_{comp}^k\}$ all converge to 0, by construction. Next, it is easy to see that

$$\nabla_x \mathcal{L}_{\mu^k,\pi^k}(x^k, s^k, \lambda^k) = \nabla_x \mathcal{L}_{\mu^k,\theta^k}(x^k, s^k, \lambda^k, \xi^k).$$

This, together with conditions (3.14a), (3.14b), and (3.14d), yields (2.7).

Recall that the infinity norm is used for (2.8) (without loss of generality). Combining (3.14e) with $\min\{x_1^k, x_2^k\} \le x_1^k$ and $\min\{x_1^k, x_2^k\} \le x_2^k$, we get

$$\begin{aligned} 0 &\le \min\{x_{1i}^k, x_{2i}^k\} \le (x_{1i}^k x_{2i}^k)^{1/2} \\ &\le (x_{1i}^k x_{2i}^k + s_{ci}^k)^{1/2} \le (\theta^k + \epsilon_{rel}^k)^{1/2} = \epsilon_{comp}^k. \end{aligned}$$

Therefore, the sequence $\{x^k, s^k, \lambda^k\}$, with corresponding parameters $\{\mu^k, \pi^k\}$, satisfies conditions (2.7) and (2.8) for all k. The conclusions follow from a direct application of Theorem 3.4. ∎

A similar global result is proved directly in [19], under somewhat different assumptions. The key for the proof presented here is that there exists a one-to-one correspondence between KKT points of problems (2.2) and (3.13), which is easily seen by comparing the corresponding first-order conditions. In fact, this one-to-one relation between KKT points of relaxation and penalization schemes can be extended to general nonlinear programs. Such an extension is useful because some convergence results can be derived directly for one approach only and then extended to the alternative regularization scheme in a simple way.

4. Local convergence analysis. In this section, we show that if the iterates generated by Algorithm I approach a solution x^* of the MPCC that satisfies certain regularity conditions and if the penalty parameter is sufficiently large, then this parameter is never updated and the iterates converge to x^* at a superlinear rate.

We start by defining a second-order sufficient condition (SOSC) for MPCCs (see [23]). For this purpose, we define the Lagrangian

$$\mathcal{L}(x, \lambda, \sigma_1, \sigma_2) = f(x) - \lambda_{\mathcal{E}}^T c_{\mathcal{E}}(x) - \lambda_{\mathcal{I}}^T c_{\mathcal{I}}(x) - \sigma_1^T x_1 - \sigma_2^T x_2. \tag{4.1}$$

DEFINITION 4.1. *The* MPCC second-order sufficient condition *(MPCC-SOSC) holds at x^* if x^* is a strongly stationary point of* (1.1) *with multipliers $\lambda^*, \sigma_1^*, \sigma_2^*$ and*

$$d^T \nabla_{xx}^2 \mathcal{L}(x^*, \lambda^*, \sigma_1^*, \sigma_2^*)\, d > 0 \tag{4.2}$$

for all critical directions d, with $\|d\| = 1$, satisfying

$$\nabla f(x)^T d = 0, \tag{4.3a}$$

$$\nabla c_i(x)^T d = 0 \ \ \textit{for all} \ \ i \in \mathcal{E}, \tag{4.3b}$$

$$\nabla c_i(x)^T d \geq 0 \ \ \textit{for all} \ \ i \in \mathcal{A}_c(x), \tag{4.3c}$$

$$\min_{\{j : x_{ji} = 0\}} \{d_{ji}\} = 0 \ \ \textit{for all} \ \ i = 1, \ldots, p. \tag{4.3d}$$

Notice that (4.3d) is a convenient way to summarize the following conditions, which characterize the set of feasible directions with respect to the complementarity constraints: If $x_{1i} = 0, x_{2i} > 0$, then $d_{1i} = 0$ and d_{2i} is free; if $x_{2i} = 0, x_{1i} > 0$, then $d_{2i} = 0$ and d_{1i} is free; and if $x_{1i} = x_{2i} = 0$, then $0 \leq d_{1i} \perp d_{2i} \geq 0$.

For the local analysis, we make the following assumptions.

Assumptions 4.2. There exists a strongly stationary point x^* of the MPCC (1.1), with multipliers $\lambda^*, \sigma_1^*, \sigma_2^*$, satisfying the following conditions:

1. f and c are twice Lipschitz continuously differentiable in an open neighborhood of x^*.
2. MPCC-LICQ holds at x^*.
3. The following primal-dual strict complementarity holds at x^*: $\lambda_i^* \neq 0$ for all $i \in \mathcal{E} \cup \mathcal{A}_c(x^*)$, and $\sigma_{ji}^* > 0$ for all $i \in \mathcal{A}_1(x^*) \cap \mathcal{A}_2(x^*)$, for $j = 1, 2$.
4. MPCC-SOSC holds at x^*.

The following lemma shows that the penalty formulation inherits the desirable properties of the MPCC for a sufficiently large penalty parameter. The multipliers for the bound constraints $x_1 \geq 0, x_2 \geq 0$ of the penalty problem (2.1) are denoted by $\nu_1 \geq 0, \nu_2 \geq 0$, respectively.

LEMMA 4.3. *If Assumptions* 4.2 *hold at* x^* *and* $\pi > \pi^*$, *where*

$$\pi^* = \pi^*(x^*, \sigma_1^*, \sigma_2^*) = \max\left\{0, \max_{\{i:x_{1i}^*>0\}} \frac{-\sigma_{2i}^*}{x_{1i}^*}, \max_{\{i:x_{2i}^*>0\}} \frac{-\sigma_{1i}^*}{x_{2i}^*}\right\}, \tag{4.4}$$

then it follows that

1. *LICQ holds at* x^* *for* (2.1);
2. x^* *is a KKT point of* (2.1);
3. *primal-dual strict complementarity holds at* x^* *for* (2.1); *that is,* $\lambda_i^* \neq 0$ *for all* $i \in \mathcal{E} \cup \mathcal{A}_c(x^*)$ *and* $\nu_{ji}^* > 0$ *for all* $i \in \mathcal{A}_j(x^*)$, *for* $j = 1, 2$;
4. *SOSC holds at* x^* *for* (2.1).

Proof. LICQ at x^* for (2.1) follows from the definition of MPCC-LICQ.

The proof of part 2 is similar to the proof of Proposition 4.1 in [13]. The key for the proof is the relationship between the multipliers σ_1^*, σ_2^* of (1.1) and $\nu_1^* \geq 0, \nu_2^* \geq 0$ of (2.1), given by

$$\nu_1^* = \sigma_1^* + \pi x_2^* \quad \text{and} \quad \nu_2^* = \sigma_2^* + \pi x_1^*. \tag{4.5}$$

The result is evident when the strong stationarity conditions (3.3) and the first-order KKT conditions of (2.1) are compared, except for the nonnegativity of ν_1^* and ν_2^*. To see that $\nu_1^*, \nu_2^* \geq 0$, suppose first that $i \in \mathcal{A}_1(x^*) \cap \mathcal{A}_2(x^*)$. In that case, from (4.5), we have $\nu_{ji} = \sigma_{ji}, j = 1, 2$, and the nonnegativity follows directly from (3.3i). If, on the other hand, $i \notin \mathcal{A}_2(x^*)$, then (4.5) and $\pi > \pi^*$ imply

$$\nu_{1i}^* = \sigma_{1i}^* + \pi x_{2i}^* > \sigma_{1i}^* + \frac{-\sigma_{1i}^*}{x_{2i}^*} x_{2i}^* = 0. \tag{4.6}$$

The same argument applies for $i \notin \mathcal{A}_1(x^*)$, which completes the proof of part 2.

Note that $\pi \geq \pi^*$ suffices for the nonnegativity of ν_1, ν_2. The strict inequality $\pi > \pi^*$ is required for part 3; that is, we need it for primal-dual strict complementarity at x^* for (2.1). In fact, (4.6) yields primal-dual strict complementarity for $i \notin \mathcal{A}_2(x^*)$ (and a similar argument works for $i \notin \mathcal{A}_1(x^*)$). For $i \in \mathcal{E} \cup \mathcal{A}_c(x^*)$, strict complementarity comes directly from the assumptions. For $i \in \mathcal{A}_2(x^*) \cap \mathcal{A}_1(x^*)$, relation (4.5) shows that $\nu_{ji}^* = \sigma_{ji}^*, j = 1, 2$, which is positive by Assumption 4.2(3).

For part 4, Assumption 4.2(3) implies that the multipliers of the complementarity variables satisfy $\nu_{1i}^* + \nu_{2i}^* > 0$ for all $i \in \mathcal{A}_1(x^*) \cap \mathcal{A}_2(x^*)$, which, together with $\pi > \pi^*$, constitutes a sufficient condition for SOSC of the penalty problem (2.1); see [20] for details. Therefore, SOSC holds at x^* for (2.1). ∎

We note that Assumption 4.2(3) can be weakened and still get SOSC for the penalized problem (2.1). In [20], two alternative sufficient conditions for SOSC of (2.1) are given. The first involves $\nu_{1i}^* + \nu_{2i}^* > 0$ for all $i \in \mathcal{A}_1(x^*) \cap \mathcal{A}_2(x^*)$ (which is called partial strict complementarity in [22]) and $\pi > \pi^*$. The second condition involves a possibly larger penalty parameter and shows how the curvature term of the complementarity constraint $x_1^T x_2$ can be exploited to ensure the penalized problem satisfies a second-order condition. We state the result here for completeness (the proof can be found in [20]).

LEMMA 4.4. *Let MPCC-SOSC hold at* x^*. *If either*

1. $\pi > \pi^*$ *and* $\nu_{1i}^* + \nu_{2i}^* > 0$ *for all* $i \in \mathcal{A}_1(x^*) \cap \mathcal{A}_2(x^*)$, *or*
2. $\pi > \max\{\pi^*, \pi_{SO}\}$, *for a (possibly higher) value* π_{SO} *defined in* [20],

then SOSC holds at x^* *for* (2.1). ∎

We now show that an adequate penalty parameter stabilizes near a regular solution and superlinear convergence takes place.

We group primal and dual variables in a single vector $z = (x, s, \lambda)$. Given a strongly stationary point x^* with multipliers $\lambda^*, \sigma_1^*, \sigma_2^*$, we associate to it the triplet $z^* = (x^*, s^*, \lambda^*)$, where $s^* = c_{\mathcal{I}}(x^*)$. We also group the left-hand side of (2.5) in the function

$$F_\mu(z;\pi) = \begin{pmatrix} \nabla_x \mathcal{L}_{\mu,\pi}(x,\lambda) \\ S\lambda_{\mathcal{I}} - \mu e \\ c(x,s) \end{pmatrix}. \tag{4.7}$$

At every inner iteration in step 2 of Algorithm I, a step d is computed by solving a system of the form

$$\nabla F_\mu(z;\pi) d = -F_\mu(z;\pi). \tag{4.8}$$

Note that (2.7) is equivalent to $\|F_\mu(z;\pi)\| \le \epsilon_{pen}$.

The following theorem shows that there are practical implementations of Algorithm I that, near a regular solution x^* of the MPCC and for a sufficiently large penalty parameter, satisfy the stopping tests (2.7) and (2.8) at every iteration, with no backtracking and no updating of the penalty parameter. Using this fact, one can easily show that the iterates converge to x^* superlinearly. To state this result, we introduce the following notation. Let z be an iterate satisfying $\|F_\mu(z;\pi)\| \le \epsilon_{pen}$ and $\|\min\{x_1, x_2\}\| \le \epsilon_{comp}$. We define z^+ to be the new iterate computed using a barrier parameter $\mu^+ < \mu$, namely,

$$z^+ = z + d, \qquad \text{with} \quad F_{\mu^+}(z;\pi) d = -F_{\mu^+}(z;\pi). \tag{4.9}$$

THEOREM 4.5. *Suppose that Assumptions* 4.2 *hold at a strongly stationary point* x^*. *Assume that* $\pi > \pi^*$, *with* π^* *given by* (4.4), *and that the tolerances* $\epsilon_{pen}, \epsilon_{comp}$ *in Algorithm* I *are functions of* μ *that converge to* 0 *as* $\mu \to 0$. *Furthermore, assume that the barrier parameter and these tolerances are updated so that the following limits hold as* $\mu \to 0$*:*

$$\frac{(\epsilon_{pen} + \mu)^2}{\epsilon_{pen}^+} \to 0, \tag{4.10a}$$

$$\frac{(\epsilon_{pen} + \mu)^2}{\mu^+} \to 0, \tag{4.10b}$$

$$\frac{\mu^+}{\epsilon_{comp}^+} \to 0. \tag{4.10c}$$

Assume also that

$$\frac{\mu^+}{\|F_0(z;\pi)\|} \to 0 \ \ as \ \ \|F_0(z;\pi)\| \to 0. \tag{4.11}$$

Then, if μ *is sufficiently small and* z *is sufficiently close to* z^*, *the following conditions hold:*

1. *The stopping criteria* (2.7) *and* (2.8), *with parameters* $\mu^+, \epsilon_{pen}^+, \epsilon_{comp}^+$ *and* π, *are satisfied at* z^+.
2. $\|z^+ - z^*\| = o(\|z - z^*\|)$.

Proof. By the implicit function theorem, Assumptions 4.2, the condition $\pi > \pi^*$, and Lemma 4.3, it follows that, for all sufficiently small μ, there exists a solution

$z^*(\mu)$ of problem (2.2); see, for example, [14]. If, in addition, z is close to z^*, then

$$L_1\mu \leq \|z^* - z^*(\mu)\| \leq U_1\mu, \tag{4.12}$$

$$L_2\|F_\mu(z;\pi)\| \leq \|z - z^*(\mu)\| \leq U_2\|F_\mu(z;\pi)\|. \tag{4.13}$$

(Condition (4.12) is Corollary 3.14 in [14], and (4.13) is Lemma 2.4 in [5].) Here and in the rest of the proof L_i and U_i denote positive constants; recall that $\|\cdot\|$ denotes the infinity norm (without loss of generality). By standard Newton analysis (see, e.g., Theorem 2.3 in [5]) we have that

$$\|z^+ - z^*(\mu^+)\| \leq U_3\|z - z^*(\mu^+)\|^2. \tag{4.14}$$

We also use the inequality

$$\|z^+ - z^*(\mu^+)\| \leq U_4(\epsilon_{pen} + \mu)^2, \tag{4.15}$$

which is proved as follows:

$$\begin{aligned}
\|z^+ - z^*(\mu^+)\| &\leq U_3\|z - z^*(\mu^+)\|^2 \quad \text{(from (4.14))} \\
&\leq U_3\left(\|z - z^*(\mu)\| + \|z^*(\mu) - z^*\| + \|z^* - z^*(\mu^+)\|\right)^2 \\
&\leq U_3\left(U_2\|F_\mu(z;\pi)\| + U_1\mu + U_1\mu^+\right)^2 \quad \text{(from (4.13) and (4.12))} \\
&\leq U_4\left(\epsilon_{pen} + \mu\right)^2,
\end{aligned}$$

where the last inequality holds because z satisfies (2.7) with μ, ϵ_{pen}, π and because $\mu^+ < \mu$.

We now show that (2.7) holds at z^+, with parameters $\mu^+, \epsilon^+_{pen}, \pi$, as follows:

$$\begin{aligned}
\|F_{\mu^+}(z^+;\pi)\| &\leq L_2^{-1}\|z^+ - z^*(\mu^+)\| \quad \text{(from (4.13))} \\
&\leq L_2^{-1}U_4(\epsilon_{pen} + \mu)^2 \quad \text{(from (4.15))} \\
&= L_2^{-1}U_4\frac{(\epsilon_{pen} + \mu)^2}{\epsilon^+_{pen}}\epsilon^+_{pen} \\
&\leq \epsilon^+_{pen} \quad \text{(from (4.10a)).}
\end{aligned}$$

To see that $x_1^+ > 0$, we can apply (4.15) componentwise to get

$$|x^+_{1i} - x^*_{1i}(\mu^+)| \leq U_4(\epsilon_{pen} + \mu)^2,$$

from which we have that

$$x^+_{1i} \geq x^*_{1i}(\mu^+) - U_4(\epsilon_{pen} + \mu)^2. \tag{4.16}$$

If $x^*_{1i} = 0$, we have by (4.12) and the positivity of $x^*_{1i}(\mu^+)$ that $x^*_{1i}(\mu^+) \geq L_1\mu^+$. Therefore

$$\begin{aligned}
x^+_{1i} &\geq L_1\mu^+ - U_4\frac{(\epsilon_{pen} + \mu)^2}{\mu^+}\mu^+ \quad \text{(from (4.12))} \\
&\geq L_5\mu^+ \quad \text{(from (4.10b)).}
\end{aligned}$$

If, on the other hand, $x^*_{1i} > 0$, then from (4.12) and (4.16), we get

$$\begin{aligned}
x^+_{1i} &\geq x^*_{1i} - U_1\mu^+ - U_4(\epsilon_{pen} + \mu)^2 \\
&= x^*_{1i} - U_1\mu^+ - U_4\frac{(\epsilon_{pen} + \mu)^2}{\mu^+}\mu^+ \\
&> 0 \quad \text{(from (4.10b)).}
\end{aligned}$$

Similar arguments can be applied to get $x_2^+ > 0, s^+ > 0, \lambda_{\mathcal{I}}^+ > 0$.

To prove that x^+ satisfies (2.8), we first observe that

$$\begin{aligned}\|z^+ - z^*\| &\le \|z^+ - z^*(\mu^+)\| + \|z^*(\mu^+) - z^*\| \\ &\le U_4(\epsilon_{pen} + \mu)^2 + U_1\mu^+ \quad \text{(from (4.15) and (4.12))} \\ &= U_4\frac{(\epsilon_{pen} + \mu)^2}{\mu^+}\mu^+ + U_1\mu^+ \\ &\le U_5\mu^+ \quad \text{(from (4.10b)).} \end{aligned} \tag{4.17}$$

Let $i \in \{1, \ldots, p\}$, and assume, without loss of generality, that $x_{1i}^* = 0$. Then,

$$\begin{aligned}|\min\{x_{1i}^+, x_{2i}^+\}| &= \min\{x_{1i}^+, x_{2i}^+\} \quad \text{(because } x_1^+ > 0, x_2^+ > 0) \\ &\le x_{1i}^+ = |x_{1i}^+ - x_{1i}^*| \\ &\le U_5\mu^+ \quad \text{(from (4.17))} \\ &= U_5\frac{\mu^+}{\epsilon_{comp}^+}\epsilon_{comp}^+ \le \epsilon_{comp}^+, \end{aligned}$$

where the last inequality follows from (4.10c). Since this argument applies to all $i \in \{1, \ldots, p\}$, we have that (2.8) is satisfied. This concludes the proof of part 1 of the theorem.

For part 2, we have that

$$\begin{aligned}\|z^+ - z^*\| &\le \|z^+ - z^*(\mu^+)\| + \|z^*(\mu^+) - z^*\| \\ &\le U_3\|z - z^*(\mu^+)\|^2 + U_1\mu^+ \quad \text{(from (4.14) and (4.12))} \\ &\le U_3\left(\|z - z^*\| + \|z^* - z^*(\mu^+)\|\right)^2 + U_1\mu^+ \\ &\le U_3\left(2\|z - z^*\|^2 + 2\|z^* - z^*(\mu^+)\|^2\right) + U_1\mu^+ \\ &\le 2U_3\|z - z^*\|^2 + 2U_3(U_1\mu^+)^2 + U_1\mu^+ \quad \text{(from (4.14))} \\ &\le U_6\left(\|z - z^*\|^2 + \mu^+\right). \end{aligned}$$

This implies that

$$\frac{\|z^+ - z^*\|}{\|z - z^*\|} \le U_6\left(\|z - z^*\| + \frac{\mu^+}{\|z - z^*\|}\right).$$

We apply the left-hand inequality in (4.13), evaluated at z and with barrier parameter 0, to get

$$\frac{\|z^+ - z^*\|}{\|z - z^*\|} \le U_6\left(\|z - z^*\| + \frac{1}{L_2}\frac{\mu^+}{\|F_0(z;\pi)\|}\right). \tag{4.18}$$

Note that, from (4.13), if $\|z - z^*\|$ is sufficiently small, so is $\|F_0(z;\pi)\|$, which in turn, by (4.11), implies that the second term in the right-hand side is also close to 0. Hence, if $\|z - z^*\|$ is sufficiently small, it follows that the new iterate z^+ is even closer to z^*. Moreover, by applying (4.18) recursively, we conclude that the iterates converge to z^*. From the same relation, it is clear that this convergence happens at a superlinear rate, which concludes the proof. □

Many practical updating rules for μ and ϵ_{pen} satisfy conditions (4.10a)–(4.11). For example, we can define $\epsilon_{pen} = \theta\mu$ with $\theta \in [0, \sqrt{|\mathcal{I}|})$. In this case, it is not

difficult to show [5] that (4.10a), (4.10b), and (4.11) are satisfied if we update μ by the rule

$$\mu^+ = \mu^{1+\delta}, \quad 0 < \delta < 1.$$

The same is true for the rule

$$\mu^+ = \|F_\mu(z;\pi)\|^{1+\delta}, \quad 0 < \delta < 1.$$

A simple choice for ϵ_{comp} that ensures (4.10c) is μ^γ, with $0 < \gamma < 1$.

5. Implementation and numerical results. We begin by describing two practical implementations of Algorithm I that use different strategies for updating the penalty parameter. The first algorithm, *Classic*, is described in Figure 3; it updates the penalty parameter only after the barrier problem is solved, and provided the complementarity value has decreased sufficiently as stipulated in step 3. We index by k the major iterates that satisfy (2.7) and (2.8); this notation is consistent with that of section 2. We use j to index the sequence of all minor iterates generated by the algorithm Classic. Since $\gamma \in (0,1)$, the tolerance ϵ^k_{comp} defined in step 1 converges to 0 more slowly than does $\{\mu^k\}$; this is condition (4.10c) in Theorem 4.5.

In the numerical experiments, we use $\gamma = 0.4$ for the following reason: The distance between iterates x^k and the solution x^* is proportional to $\sqrt{\mu^k}$, if primal-dual strict complementarity does not hold at x^*. By choosing the complementarity tolerance to be $\epsilon^k_{comp} = (\mu^k)^{0.4}$, we ensure that the test (2.8) can be satisfied in this case. All other details of the interior method are described below.

Algorithm Classic: A Practical Interior-Penalty Method for MPCCs

Initialization: Let $z^0 = (x^0, s^0, \lambda^0)$ be the initial primal and dual variables. Choose an initial penalty π^0 and a parameter $\gamma \in (0,1)$. Set $j = 0, k = 1$.

repeat (barrier loop)

1. Choose a barrier parameter μ^k, a stopping tolerance ϵ^k_{pen}, let $\epsilon^k_{comp} = (\mu^k)^\gamma$ and let $\pi^k = \pi^{k-1}$.
2. **repeat** (inner iteration)
   (a) Let $j \leftarrow j + 1$ and let the current point be $z^c = z^{j-1}$.
   (b) Using a globally convergent method, compute a primal-dual step d^j based on the KKT system (2.4), with $\mu = \mu^k, \pi = \pi^k$ and $z = z^c$.
   (c) Let $z^j = z^c + d^j$.
   until conditions (2.7) are satisfied for ϵ^k_{pen}.
3. **If** $\|\min\{x^j_1, x^j_2\}\| \le \epsilon^k_{comp}$, let $z^k = z^j$, set $k \leftarrow k + 1$;
   else set $\pi^k \leftarrow 10\pi^k$ and go to Step 2

until a stopping test for the MPCC is satisfied.

FIG. 3. *Description of the algorithm Classic.*

The second algorithm we implemented, *Dynamic*, is described in Figure 4. It is more flexible than Classic in that it allows changes in the penalty parameter at every iteration of the inner algorithm. The strategy of step 2(c) is based on the following considerations: If the complementarity pair is relatively small according to the preset tolerance ϵ^k_{comp}, then there is no need to increase π. Otherwise, we check whether the current complementarity value, $x_1^{jT} x_2^j$, is less than a fraction of the maximum

value attained in the m previous iterations (in our tests, we use $m = 3$ and $\eta = 0.9$). If not, we increase the penalty parameter. We believe that it is appropriate to look back at several previous steps, and not require decrease at every iteration, because the sequence $\{x_1^{jT} x_2^j\}$ is frequently nonmonotone, especially for problems in which primal-dual strict complementarity is violated (see, e.g., Figure 6(a)). Note that the algorithms Classic and Dynamic are both special cases of Algorithm I of section 2.

Algorithm Dynamic: A Practical Interior-Penalty Method for MPCCs

Initialization: Let $z^0 = (x^0, s^0, \lambda^0)$ be the initial primal and dual variables. Choose an initial penalty π^0, parameters $\gamma, \eta \in (0, 1)$, and an integer $m \geq 1$. Set $j = 0$, $k = 1$.

repeat (barrier loop)

1. Choose a barrier parameter μ^k, a stopping tolerance ϵ_{pen}^k and let $\epsilon_{comp}^k = (\mu^k)^\gamma$.
2. **repeat** (inner iteration)
   (a) Set $j \leftarrow j + 1$, let the current point be $z^c = z^{j-1}$, and let $\pi^j = \pi^{j-1}$.
   (b) Using a globally convergent method, compute a primal-dual step d^j based on the KKT system (2.4), with $\mu = \mu^k, \pi = \pi^j$ and $z = z^c$.
   (c) If $\| \min\{x_1^j, x_2^j\} \| > \epsilon_{comp}^k$ and

$$x_1^{jT} x_2^j > \eta \max \left\{ x_1^{jT} x_2^j, \ldots, x_1^{(j-m+1)T} x_2^{(j-m+1)} \right\}, \tag{5.1}$$

   then set $\pi^j \leftarrow 10\pi^j$, adjust λ^j and go to Step 2.
   until conditions (2.7) are satisfied for ϵ_{pen}^k.
3. **If** $\| \min\{x_1^j, x_2^j\} \| \leq \epsilon_{comp}^k$, let $z^k = z^j$ and $k = k + 1$
   else set $\pi^k \leftarrow 10\pi^k$ and go to Step 2

until a stopping test for the MPCC is satisfied.

FIG. 4. *Description of the algorithm Dynamic.*

We implemented these two algorithms as an extension of our MATLAB solver IPM-D. This solver is based on the interior algorithm for nonlinear programming described in [26], with one change: IPM-D handles negative curvature by adding a multiple of the identity to the Hessian of the Lagrangian, as in [25], instead of switching to conjugate-gradient iterations. We chose to work with IPM-D because it is a simple interior solver that does not employ the regularizations, scalings, and other heuristics used in production packages that alter the MPCC, making it harder to assess the impact of the approach proposed in this paper.

In our implementation, all details of the interior-point iteration, such as the update of the barrier parameter, the step selection, and the choice of merit function, are handled by IMP-D. The main point of this section is to demonstrate how to adapt an existing interior-point method to solve MPCCs efficiently and reliably.

We tested the algorithms on a collection of 74 problems, listed in Table 3, where we report the number of variables n (excluding slacks), the number of constraints m (excluding complementarity constraints), and the number of complementarity constraints p. These problems are taken from the MacMPEC collection [18]; we added a few problems to test the sensitivity of our implementations to bad scalings in the MPCC. All the methods tested were implemented in IPM-D, and since this MATLAB program is not suitable for very large problems, we restricted our test set to a sample

TABLE 3
Test problem characteristics.

Name	n	m	p	Name	n	m	p
bar-truss-3	29	22	6	bard1	5	1	3
bard3	6	3	2	bilevel1	10	9	6
bilevel3	12	7	4	bilin	8	1	6
dempe	4	2	2	design-cent-1	12	9	3
design-cent-4	22	9	12	desilva	6	2	2
df1	2	2	1	ex9.1.1	13	12	5
ex9.1.3	23	21	6	ex9.1.5	13	12	5
ex9.1.6	14	13	6	ex9.1.7	17	15	6
ex9.1.8	14	12	5	ex9.1.9	12	11	5
ex9.1.10	14	12	5	ex9.2.1	10	9	4
ex9.2.2	10	11	4	ex9.2.4	8	7	2
ex9.2.5	8	7	3	ex9.2.6	16	12	6
ex9.2.7	10	9	4	ex9.2.8	6	5	2
ex9.2.9	9	8	3	flp2	4	2	2
flp4-1	80	60	30	gauvin	3	0	2
gnash10	13	4	8	gnash11	13	4	8
gnash12	13	4	8	gnash13	13	4	8
gnash14	13	4	8	gnash15	13	4	8
gnash16	13	4	8	gnash17	13	4	8
gnash18	13	4	8	gnash19	13	4	8
hakonsen	9	8	4	hs044-i	20	14	10
incid-set1-16	485	491	225	incid-set2c-16	485	506	225
kth1	2	0	1	kth2	2	0	1
kth3	2	0	1	liswet1-050	152	103	50
outrata31	5	0	4	outrata32	5	0	4
outrata33	5	0	4	outrata34	5	0	4
pack-comp1-16	332	151	315	pack-comp2c-16	332	166	315
pack-rig1c-16	209	148	192	pack-rig2-16	209	99	192
pack-rig3-16	209	99	192	portfl-i-2	87	25	12
portfl-i-6	87	25	12	qpec-100-1	105	102	100
ralph1	2	0	1	ralph2	2	0	1
ralphmod	104	0	100	scale1	2	0	2
scale2	2	0	2	scale3	2	0	2
scale4	2	0	2	scale5	2	0	2
scholtes1	3	1	1	scholtes2	3	1	1
scholtes3	2	0	2	scholtes4	3	2	2
scholtes5	3	2	2	tap-09	86	68	32

of problems with fewer than 1,000 variables. We report results for four methods, which are labeled in the figures as follows:

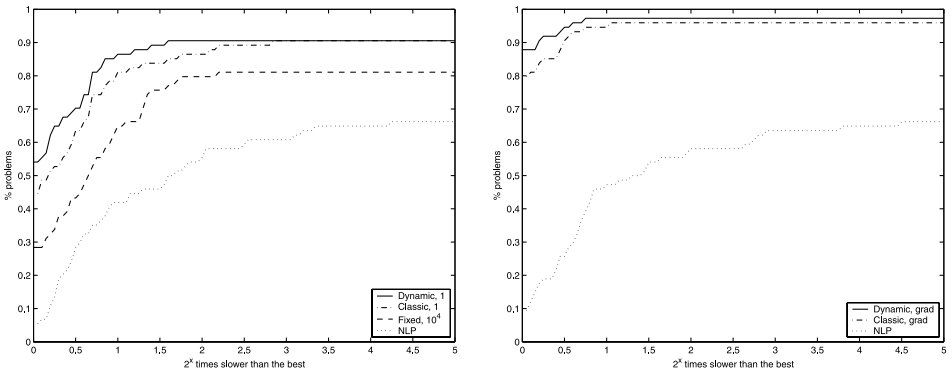
NLP is the direct application of the interior code IPM-D to the nonlinear programming formulation (1.3) of the MPCC.

Fixed is a penalty method in which IPM-D is applied to (2.1) with a fixed penalty of 10^4. The penalty parameter is not changed.

Classic is the algorithm given in Figure 3, implemented in the IPM-D solver.

Dynamic is the algorithm given in Figure 4, implemented in the IPM-D solver.

In Figure 5 we report results for these four methods in terms of total number of iterations (indexed by j). The figures use the logarithmic performance profiles described in [9]. An important choice in the algorithms Classic and Dynamic is the initial value of π. In Figure 5(a) we show results for $\pi^0 = 1$, and in Figure 5(b)



(a) $\pi = 1$, initially. (b) Initial π based on gradient norm.

FIG. 5. *Performance profiles.*

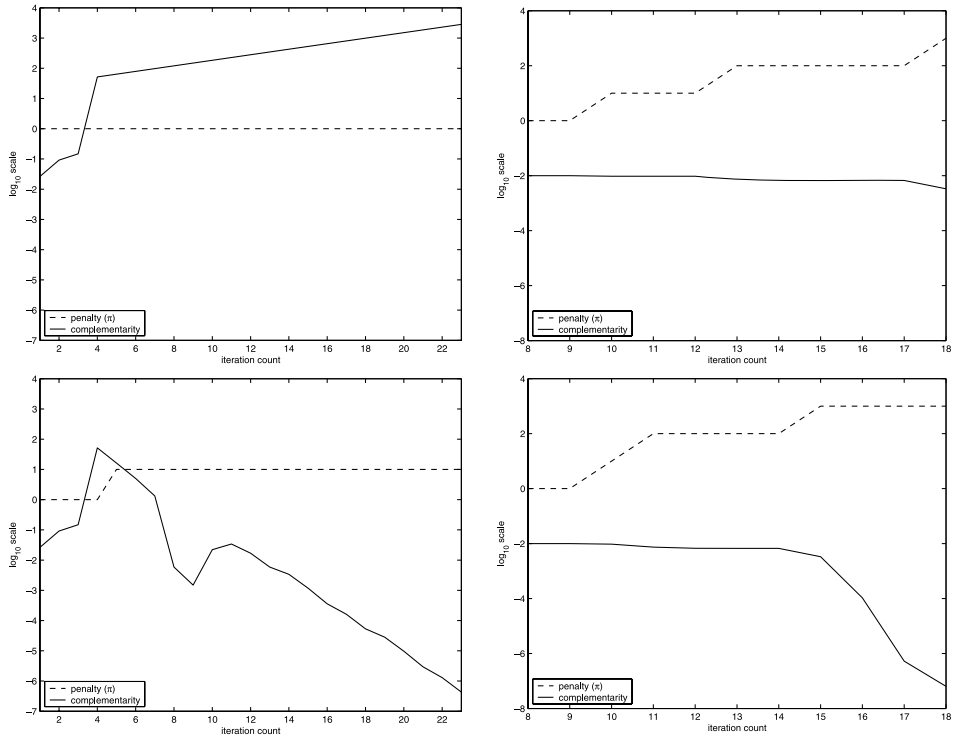
for $\pi^0 = \|\nabla f(x^0)\|$ (the latter rule is also used, for example, in the elastic phase of SNOPT [15]). Note that every time π is updated, a new barrier subproblem has to be solved, where the initial point is the current iterate and the barrier parameter is the current value of μ. The discrepancy in initial conditions when π is reset explains the difference in performance of the choices $\pi^0 = 1$ and $\pi^0 = \|\nabla f(x^0)\|$, for both Classic and Dynamic.

Comparing the results in Figure 5, we note that the direct application of the interior method, option NLP, gives the poorest results. Option Fixed (dashed curve in Figure 5(a)) is significantly more robust and efficient than option NLP, but it is clearly surpassed by the Classic and Dynamic methods. Option Fixed fails more often than Classic and Dynamic and it requires, in general, more iterations to solve each barrier problem. In extreme cases, such as `bar-truss-3`, Dynamic (with $\pi^0 = 1$) solves the first barrier problem in 15 iterations, whereas Fixed needs 43 iterations. Moreover, we frequently find that, near a solution, the algorithms Classic and Dynamic take one iteration per barrier problem, as expected, whereas Fixed keeps taking several steps to find a solution every time μ is updated.

Classic and Dynamic perform remarkably well with the seemingly naive initial value $\pi^0 = 1$ (Figure 5(a)). Both algorithms adjust π efficiently, especially Dynamic. The choice $\pi^0 = \|\nabla f(x^0)\|$, on the other hand, attempts to estimate the norm of the multipliers and can certainly be unreliable. Nonetheless, it performed very well on this test set. We note from Figure 5(b) that the performance of both algorithms improves for $\pi^0 = \|\nabla f(x^0)\|$.

The MacMPEC collection is composed almost exclusively of well-scaled problems, and `ralph2` is the only problem that becomes unbounded for the initial penalty (with either initialization of π). As a result, Dynamic does not differ significantly from Classic on this test set. We therefore take a closer look at the performance of these methods on problems `ralph2` and `scale1` discussed in section 2. We believe that the results for these examples support the choice of Dynamic over Classic for practical implementations.

Example 1 *(*`ralph2`*), revisited.* Figure 6(a) plots the complementarity measure $(x_1^{jT} x_2^j)$ (continuous line) and the value of the penalty parameter π^j (dashed line) for problem (2.9) (using a $\log_{10}$ scale). The top figure corresponds to Classic and the bottom figure to Dynamic; both used an initial penalty parameter of 1. Recall



(a) `ralph2`. (b) `scale1`.

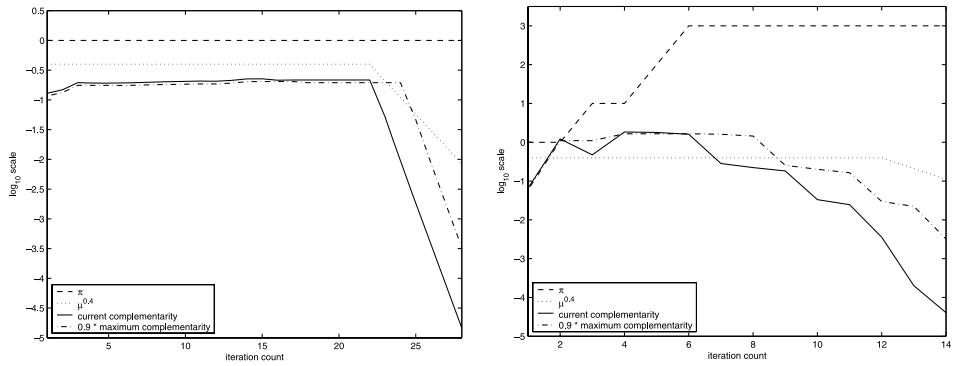
FIG. 6. *Evolution of penalty and complementarity values (*$\log_{10}$ *scale). Classic (top) versus Dynamic (bottom).*

that $\pi = 1$ gives rise to an unbounded penalty problem. The two algorithms perform identically up to iteration 4. Then, the Dynamic algorithm increases π, whereas the Classic algorithm never changes π, because it never solves the first barrier problem. Classic fails on this problem, and complementarity grows without bound. □

Example 2 *(*`scale1`*), revisited.* Problem (2.11) requires $\pi \geq 200$ so that the penalty problem recovers the solution of the MPCC. We again initialize Dynamic and Classic with $\pi = 1$. Figure 6(b) plots the complementarity measure and the penalty parameter values for both implementations. The two algorithms increase π three times (from 1 to 10, to 100, to 1000). While the Classic implementation (top figure) is performing the third update of π, the Dynamic implementation (bottom figure) has converged to the solution. The Dynamic algorithm detects earlier that complementarity has stagnated (and is not sufficiently small) and takes corrective action by increasing π. Not all plateaus mean that π needs to be changed, however, as we discuss next. □

To study in more detail the algorithm Dynamic, we consider two other problems, `bard3` and `bilin`, from the MacMPEC collection (we initialize the penalty parameter to 1, as before).

Example 3 (`bard3`). Figure 7(a) shows the results for problem `bard3`. The continuous line plots $x_1^{jT} x_2^j$, and the dashed-dotted line plots 0.9 times the maximum value of $x_1^{iT} x_2^i$ over the last three iterations. Note that the complementarity measure increases at the beginning and does not decrease during the first 20 iterations.



(a) Problem `bard3`. (b) Problem `bilin`.

FIG. 7. *Illustration of the Dynamic updating strategy.*

However, Dynamic does not increase the value of π (dashed line) because the value of complementarity is small enough, compared to the threshold $\mu^{0.4}$ (dotted line). This is the correct action; if the algorithm increased π simply because the maximum value of complementarity over the last three iterations is not decreasing, π would take on large values that would slow the iteration and could even cause failure. □

Example 4 (`bilin`). A different behavior is observed for problem `bilin`; see Figure 7(b). The value of complementarity (continuous line) not only lies above the line that plots 0.9 times the maximum complementarity over the last three iterations (dashed-dotted line), but is also above the line plotting $(\mu^j)^{0.4}$. Thus the penalty parameter is increased quickly (dashed line). The sufficient reduction condition is satisfied at iteration 3 but is then again violated, so π is increased again, until complementarity finally starts converging to zero. □

These results suggest that Dynamic constitutes an effective technique for handling the penalty parameter in interior-penalty methods for MPCCs.

We conclude this section by commenting on some of the failures of our algorithms. All implementations converge to a C-stationary point for problem `scale4` (which is a rescaling of problem `scholtes3`). We find it interesting that convergence to C-stationary points is possible in practice and is not simply allowed by the theory. We note that convergence to C-stationary points cannot be ruled out for SQP methods, and in this sense interior-point methods are no less robust than SQP methods applied to MPCCs. Another failure, discussed already, is problem `ralph2` for the algorithm Classic.

The rest of the failures can be attributed to various forms of problem deficiencies beyond the MPCC structure. All implementations have difficulties solving problems for which the minimizer is not a strongly stationary point, that is, problems for which there are no multipliers at the solution. This is the case in `ex9.2.2`, where our algorithms obtain good approximations of the solution but the penalty parameter diverges, and for `ralphmod`, where our algorithms fail to find a stationary point. These difficulties are not surprising because the algorithms strongly rely upon the existence of multipliers at the solution. SQP methods also fail to find strongly stationary solutions to these problems, and generate a sequence of multipliers that diverge to infinity.

Test problems in the groups `incid-set`$*$, `pack-rig`$*$, and `pack-comp`$*$ include degenerate constraints other than those defining complementarity. Our implementa-

tions are able to solve most of these problems, but the number of iterations is high, and the performance is very sensitive to changes in the implementation. In some of these problems our algorithms have difficulty making progress near the solution. Problem `tap-09` has a rank-deficient constraint Jacobian that causes difficulties for our algorithms. All of these point to the need for more general regularization schemes for interior methods that can cope with both MPCCs and with other forms of degeneracy. This topic is the subject of current investigation [6, 16].

6. Conclusions. Interior methods can be an efficient and robust tool for solving MPCCs, when appropriately combined with a regularization scheme. In this article, we have studied an interior-penalty approach and have carefully addressed issues related to efficiency and robustness. We have provided global and local convergence analysis to support the interior-penalty methods proposed here. We have also shown how to extend our global convergence results to interior methods based on the relaxation approach described by [19, 21].

We have presented two practical implementations. The first algorithm, Classic, is more flexible than the approach studied in [2, 17], which solves the penalty problem (2.1) with a fixed penalty parameter and then updates π if necessary. The approach in [2, 17] has the advantage that it can be used in combination with any off-the-shelf nonlinear programming solver; the disadvantage is that it can be very wasteful in terms of iterations if the initial penalty parameter is not appropriate. The second algorithm, Dynamic, improves on Classic by providing a more adaptive penalty update strategy. This can be particularly important in dealing with unbounded penalty problems and also yields an improvement in efficiency when the scaling of the problem complicates the detection of complementarity violation. The numerical results presented in this paper are highly encouraging. We plan to implement the penalty method for MPCCs in the KNITRO package, which will allow us to solve large-scale MPCCs.

The penalty methods considered here are designed specifically for MPCCs. However, lack of regularity other than that caused by complementarity constraints often occurs in practice, and a more general class of interior-penalty methods for degenerate nonlinear programs is the subject of current research [6, 16]. Some of the techniques proposed here may be useful in that more general context.

Acknowledgment. The authors are grateful to two anonymous referees for their helpful comments on this paper.

REFERENCES

[1] M. Anitescu, *On Solving Mathematical Programs with Complementarity Constraints as Nonlinear Programs*, Preprint ANL/MCS-P864-1200, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, 2000.

[2] M. Anitescu, *Global convergence of an elastic mode approach for a class of mathematical programs with complementarity constraints*, SIAM J. Optim., 16 (2005), pp. 120–145.

[3] H. Y. Benson, A. Sen, D. F. Shanno, and R. J. Vanderbei, *Interior-Point Algorithms, Penalty Methods and Equilibrium Problems*, Report ORFE-03-02, Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ, 2003.

[4] R. H. Byrd, M. E. Hribar, and J. Nocedal, *An interior point algorithm for large-scale nonlinear programming*, SIAM J. Optim., 9 (1999), pp. 877–900.

[5] 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, D. J. Higham, and G. A. Watson, eds., Addison–Wesley Longman, Harlow, UK, 1997, pp. 37–56.

[6] L. Chen and D. Goldfarb, *Interior-point ℓ_2-penalty methods for nonlinear programming with strong global convergence properties*, Math. Program., to appear.

[7] A. R. Conn, N. I. M. Gould, and Ph. L. Toint, *Trust Region Methods*, MPS/SIAM Ser. Optim. 1, SIAM, Philadelphia, 2000.

[8] V. DeMiguel, M. P. Friedlander, F. J. Nogales, and S. Scholtes, *A two-sided relaxation scheme for mathematical programs with equilibrium constraints*, SIAM J. Optim., 16 (2005), pp. 587–609.

[9] E. D. Dolan and J. J. Moré, *Benchmarking optimization software with performance profiles*, Math. Program., 91 (2002), pp. 201–213.

[10] M. C. Ferris and F. Tin-Loi, *On the solution of a minimum weight elastoplastic problem involving displacement and complementarity constraints*, Comput. Methods Appl. Mech. Engrg., 174 (1999), pp. 107–120.

[11] R. Fletcher and S. Leyffer, *Nonlinear programming without a penalty function*, Math. Program., 91 (2002), pp. 239–269.

[12] R. Fletcher and S. Leyffer, *Solving mathematical programs with complementarity constraints as nonlinear programs*, Optim. Methods Softw., 19 (2004), pp. 15–40.

[13] R. Fletcher, S. Leyffer, D. Ralph, and S. Scholtes, *Local convergence of SQP methods for mathematical programs with equilibrium constraints*, SIAM J. Optim., to appear.

[14] A. Forsgren, P. E. Gill, and M. H. Wright, *Interior methods for nonlinear optimization*, SIAM Rev., 44 (2002), pp. 525–597.

[15] P. E. Gill, W. Murray, and M. A. Saunders, *SNOPT: An SQP algorithm for large-scale constrained optimization*, SIAM J. Optim., 12 (2002), pp. 979–1006.

[16] N. I. M. Gould, D. Orban, and Ph. Toint, *An Interior-Point $l1$-Penalty Method for Nonlinear Optimization*, Tech. report RAL-TR-2003-022, Rutherford Appleton Laboratory Chilton, Oxfordshire, UK, 2003.

[17] X. M. Hu and D. Ralph, *Convergence of a penalty method for mathematical programming with complementarity constraints*, J. Optim. Theory Appl., 123 (2004), pp. 365–390.

[18] S. Leyffer, *MacMPEC: AMPL Collection of MPECs*, http://www.mcs.anl.gov/∼leyffer/MacMPEC (5 August 2005).

[19] X. Liu and J. Sun, *Generalized stationary points and an interior-point method for mathematical programs with equilibrium constraints*, Math. Program., 101 (2004), pp. 231–261.

[20] G. López-Calva, *Exact-Penalty Methods for Nonlinear Programming*, Ph.D. thesis, Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL, 2005.

[21] A. U. Raghunathan and L. T. Biegler, *An interior point method for mathematical programs with complementarity constraints (MPCCs)*, SIAM J. Optim., 15 (2005), pp. 720–750.

[22] D. Ralph and S. J. Wright, *Some properties of regularization and penalization schemes for MPECs*, Optim. Methods. Softw., 19 (2004), pp. 527–556.

[23] H. Scheel and S. Scholtes, *Mathematical programs with complementarity constraints: Stationarity, optimality and sensitivity*, Math. Oper. Res., 25 (2000), pp. 1–22.

[24] S. Scholtes, *Convergence properties of a regularization scheme for mathematical programs with complementarity constraints*, SIAM J. Optim., 11 (2001), pp. 918–936.

[25] R. J. Vanderbei and D. F. Shanno, *An interior point algorithm for nonconvex nonlinear programming*, Comput. Optim. Appl., 13 (1999), pp. 231–252.

[26] R. A. Waltz, J. L. Morales, J. Nocedal, and D. Orban, *An interior algorithm for nonlinear optimization that combines line search and trust region steps*, Math. Program., to appear.

SIAM J. OPTIM.
Vol. 17, No. 1, pp. 78–97
© 2006 Society for Industrial and Applied Mathematics

DISCRETE MONOTONIC OPTIMIZATION WITH APPLICATION TO A DISCRETE LOCATION PROBLEM*

HOANG TUY†, MICHEL MINOUX‡, AND N. T. HOAI-PHUONG†

Abstract. A general discrete optimization problem is investigated that includes integer polynomial programs as special cases. To exploit the discrete monotonic structure of these problems, a special class of cuts called monotonicity cuts are developed and then adjusted according to a suitable procedure to accommodate discrete requirements. As illustration, the method is applied to solve a discrete location problem which is also a variant of the well known engineering problem of design centering. Computational results are reported for instances of the latter problem with up to 100 variables and 500 constraints.

Key words. discrete optimization, differences of increasing functions, monotonic optimization, polyblock approximation, monotonicity cuts, branch-reduce-and-bound algorithm, discrete location, design centering

AMS subject classifications. 90C26, 65K05, 90C20, 90C30, 90C56, 78M50

DOI. 10.1137/04060932X

1. Introduction.

Throughout this paper, for any two vectors $x, y \in \mathbb{R}^n$ we write $x \leq y$ ($x < y$, resp.) to mean $x_i \leq y_i$ ($x_i < y_i$, resp.) for every $i = 1, \ldots, n$. If $a \leq b$ then the box $[a, b]$ ($(a, b]$, resp.) is the set of all $x \in \mathbb{R}^n$ satisfying $a \leq x \leq b$ ($a < x \leq b$, resp.).

A function $f : [a, b] \to \mathbb{R}$ is said to be *increasing* (*decreasing,* resp.) if

$$a \leq x \leq y \leq b \Rightarrow f(x) \leq f(y) \quad (f(x) \geq f(y), \text{ resp.}).$$

Monotonic functions, i.e., functions which are either increasing or decreasing, and more generally, *d.m. functions*, i.e., functions which are differences of monotonic functions, abound in economics and engineering: production function, cost function, profit function, performance function, etc. As can easily be proved (see [10]) the set of d.m. functions is dense in the space $\mathbf{C}([a, b])$ of continuous functions with the supnorm topology. Recently a general mathematical framework has been developed [10], [11], for the numerical study of monotonic optimization problems, i.e., mathematical programming problems described by means of monotonic or d.m. functions.

Since any generalized polynomial $P(x) = \sum_{\alpha \in I} c_\alpha x^\alpha$, where $c_\alpha \in \mathbb{R}$, $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n_+$, and $x^\alpha := x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}$, can be written as a difference of two generalized polynomials with positive coefficients,

$$P(x) = \sum_{\alpha \in I_+} c_\alpha x^\alpha - \sum_{\alpha \in I_-} (-c_\alpha) x^\alpha,$$

where $I_+ = \{\alpha \in I | c_\alpha > 0\}$, $I_- = \{\alpha \in I |\ c_\alpha < 0\}$, and each generalized polynomial with positive coefficients is obviously increasing on $\mathbb{R}^n_+$, the class of monotonic

*Received by the editors June 3, 2004; accepted for publication (in revised form) November 30, 2005; published electronically April 21, 2006.
http://www.siam.org/journals/siopt/17-1/60932.html

†Institute of Mathematics, 18 Hoang Quoc Viet Road, 10307 Hanoi, Vietnam (htuy@math.ac.vn, htphuong@math.ac.vn).

‡Université Paris VI, 4 Place Jussieu, 75005 Paris, France (Michel.Minoux@lip6.fr).

optimization problems includes as special cases generalized polynomial mathematical programs which have received increasing attention in recent years due to their applications in various fields (see, e.g., [7], [2], [15]).

It has been shown in [10] that any mathematical programming problem dealing with differences of increasing functions on a box $[a, b] \subset \mathbb{R}^n$ can be reduced to an equivalent constrained monotonic optimization problem of the following *canonical form*:

$$\text{(MO)} \qquad \max\{f(x)|\ g(x) \le 0 \le h(x),\ x \in [a, b]\},$$

where $f, g, h : [a, b] \to \mathbb{R}$ are given increasing functions on $[a, b]$. (Without loss of generality we may assume, as we will do in what follows, that $[a, b] \subset \mathbb{R}^n_+$.) An easily implementable cutting algorithm called the *polyblock algorithm* has been proposed for solving (MO). Applications of this approach to certain classes of difficult nonconvex optimization problems have demonstrated its efficiency when these problems, originally of large scale, can be converted into problems (MO) in low-dimensional space by a suitable change of variables [5], [12], [13], [15]. For larger problems a branch and bound approach, using polyblock approximation for bounding, has also been proposed and recently extended to a branch-reduce-and-bound method [11].

If a problem involves discrete constraints, e.g., boolean constraints like $x_i \in \{0, 1\},\ i = 1, \dots, s\ (s \le n)$, then these constraints can be written as $\sum_{i=1}^s x_i(1-x_i) \le 0,\ 0 \le x_i \le 1\ (i = 1, \dots, s)$, i.e., $\sum_{i=1}^s x_i - \sum_{i=1}^s x_i^2 \le 0,\ 0 \le x_i \le 1\ (i = 1, \dots, s)$, where the functions $\sum_{i=1}^s x_i,\ \sum_{i=1}^s x_i^2$ are increasing on $\mathbb{R}^n_+$. Therefore, a monotonic optimization problem with discrete constraints can in principle be reformulated and solved as one with only continuous monotonic constraints. However, so far this approach has never been implemented; moreover, its potential drawback is that, since the basic algorithms for continuous monotonic optimization are iterative procedures, by this approach only an approximate optimal solution can be computed in finitely many steps.

The aim of the present paper is to suggest an alternative, more practical, approach to monotonic optimization problems with discrete constraints. Specifically, given a box $[a, b] \subset \mathbb{R}^n_+$ with $a < b$, a finite set $S \subset \mathbb{R}^s_+,\ s \le n$, and increasing functions $f(x), g(x), h(x)$ on $[a, b]$, we will consider the general optimization problem

$$\max\{f(x)|\ g(x) \le 0 \le h(x),\ x \in [a, b],\ (x_1, \dots, x_s) \in S\}.$$

We will refer to this problem as the *canonical discrete monotonic optimization problem* (DMO). Setting

$$\text{(1)} \qquad G = \{x \in [a, b]|\ g(x) \le 0\}, H = \{x \in [a, b]|\ h(x) \ge 0\},$$

$$\text{(2)} \qquad S^* = \{x \in [a, b]|\ (x_1, \dots, x_s) \in S\},$$

we can rewrite it as

$$\text{(DMO)} \qquad \max\{f(x)|\ x \in G \cap H \cap S^*\}.$$

In what follows we propose to extend the basic algorithm for continuous monotonic optimization [11] to obtain an algorithm for (DMO) which is *finite* in the important special case when $s = n$ (so that S is a finite set in $\mathbb{R}^n_+$). It turns out that in this special case, by suitable modifications of the basic continuous algorithm, an exact optimal solution of (DMO) can be computed in finitely many iterations. Furthermore, for

certain continuous monotonic optimization problems, although the optimum may be known a priori to be achieved on a certain finite set S, quite often an exact optimal solution can be obtained only through infinitely many iterations of the basic iterative algorithm of the continuous approach. The discrete version to be proposed will help in many cases to turn this infinite procedure into a finite one, thus allowing the exact optimum in these continuous problems to be found in finitely many iterations, too.

The paper is organized as follows. First, in sections 2 and 3, we review some necessary concepts and results from monotonic optimization as presented in [10]. In section 4 two types of cuts exploiting the monotonic structure are described. The first type, introduced earlier in [10], is based on a special separation property of normal sets and plays a role very similar to that of convexity cuts in convex maximization. The second type, to be referred to as reduction cuts, is a further development of a procedure already used in [10] for reducing the size of the solution set of a system of monotonic inequalities. After this review of the essentials about continuous monotonic optimization, the next sections present the new theory of discrete monotonic optimization. A key operation, called S-adjustment, designed to accommodate the continuous monotonicity cuts to discrete constraints, is described in section 5. Next, a procedure for discrete optimization, to be referred to as the discrete polyblock algorithm, is presented in section 6. Some implementation issues are discussed, while the convergence of this algorithm in the general case $s \leq n$ and its finiteness when $s = n$ are established. To enhance efficiency for large scale problems, a branch-reduce-and-bound version of the method is developed in section 7. Finally in section 8, the method is specialized to solve a discrete maximin problem encountered in location, design centering, and some other applications. Computational experiments on problems with up to 100 variables and 500 constraints are reported which demonstrate the practicability of the method at least for this class of discrete optimization problems.

As a matter of notation, for any two vectors $x, y \in \mathbb{R}^n$, we write $u = x \vee y$, to mean $u_i = \max\{x_i, y_i\},\ i = 1, \ldots, n$, and $v = x \wedge y$, to mean $v_i = \min\{x_i, y_i\},\ i = 1, \ldots, n$; e^i denotes the ith unit vector of $\mathbb{R}^n$, i.e., a vector such that $e^i_i = 1, e^i_j = 0\ \forall j \neq i$, while $e \in \mathbb{R}^n$ is a vector of all ones, i.e., $e = \sum_{i=1}^n e^i$.

2. Some geometric concepts. In this and the next two sections we review some basic concepts and essential results of continuous monotonic optimization [10] which are needed for the development of discrete monotonic optimization. For the convenience of the reader, most of the proofs will be provided, although they are rather simple and can be found in [10].

A set $G \subset [a, b]$ is said to be *normal* if $x \in G \Rightarrow [a, x] \subset G$. A set $H \subset [a, b]$ is *conormal (reverse normal)* if $x \in H \Rightarrow [x, b] \subset H$. Thus the set $G = \{x \in [a, b] |\ g(x) \leq 0\}$ defined above (with an increasing function $g(x)$) is normal, whereas the set $H = \{x \in [a, b] |\ h(x) \geq 0\}$ (with an increasing function $h(x)$) is conormal.

Given a set $A \subset [a, b]$, the *normal hull* of A, written $A^\rceil$, is the smallest normal set containing A. The *conormal hull* of A, written $\lfloor A$, is the smallest conormal set containing A.

PROPOSITION 1. (i) *The normal hull of a set $A \subset [a, b] \subset R^n_+$ is the set $A^\rceil = \cup_{z \in A}[a, z]$. If A is compact, then so is $A^\rceil$.*

(ii) *The conormal hull of a set $A \subset [a, b] \subset R^n_+$ is the set $\lfloor A = \cup_{z \in A}[z, b]$. If A is compact, then so is $\lfloor A$.*

Proof. It suffices to prove (i), because the proof of (ii) is similar. Let $P = \cup_{z \in A}[a, z]$. Clearly P is normal and $P \supset A$; hence $P \supset A^\rceil$. Conversely, if $x \in P$, then $x \in [a, z]$ for some $z \in A \subset A^\rceil$; hence $x \in A^\rceil$ by normality of $A^\rceil$, so that $P \subset A^\rceil$

and therefore, $P = A^{\rceil}$. If A is compact, then A is contained in a ball B centered at 0, and if $x^k \in A^{\rceil}, k = 1, 2, \dots$, then since $x^k \in [a, z^k] \subset B$, there exists a subsequence $\{k_\nu\} \subset \{1, 2, \dots\}$ such that $z^{k_\nu} \to z^0 \in A,\ x^{k_\nu} \to x^0 \in [a, z^0]$, and hence $x^0 \in A^{\rceil}$, proving the compactness of $A^{\rceil}$. □

The normal hull of a finite set $T \subset [a, b]$ is called a *polyblock* P, with *vertex set* T. By Proposition 1, $P = \cup_{z \in T}[a, z]$. A vertex z of a polyblock is called *proper* if there is no vertex $z' \neq z$ "dominating" z, i.e., such that $z' \geq z$. An *improper* vertex or improper element of T is an element of T which is not a proper vertex. Obviously, a polyblock is fully determined by its proper vertex set; more precisely, *a polyblock is the normal hull of its proper vertices.*

Similarly, the conormal hull of a finite set $T \subset [a, b]$ is called a *copolyblock (reverse polyblock)* Q with *vertex set* T. By Proposition 1, $Q = \cup_{z \in T}[z, b]$. A vertex z of a copolyblock is called *proper* if there is no vertex $z' \neq z$ "dominated" by z, i.e., such that $z' \leq z$. An *improper* vertex or improper element of T is an element of T which is not a proper vertex. Obviously, a copolyblock is fully determined by its proper vertex set; more precisely, *a copolyblock is the conormal hull of its proper vertices.*

PROPOSITION 2. (i) *The intersection of finitely many polyblocks is a polyblock.*

(ii) *The intersection of finitely many copolyblocks is a copolyblock.*

Proof. If T_1, T_2 are the vertex sets of two polybocks P_1, P_2, respectively, then $P_1 \cap P_2 = (\cup_{z \in T_1}[a, z]) \cap (\cup_{y \in T_2}[a, y]) = \cup_{z \in T_1, y \in T_2}[a, z] \cap [a, y] = \cup_{z \in T_1, y \in T_2}[a, z \wedge y]$. Thus, $P_1 \cap P_2$ is a polyblock of vertex set $\{z \wedge y |\ z \in T_1, y \in T_2\}$. Similarly, if T_1, T_2 are the vertex sets of two copolyblocks Q_1, Q_2, respectively, then $Q_1 \cap Q_2 = \cup_{z \in T_1, y \in T_2}[z, b] \cap [y, b] = \cup_{z \in T_1, y \in T_2}[z \vee y, b]$, so $Q_1 \cap Q_2$ is a copolyblock with vertex set $\{z \vee y |\ z \in T_1, y \in T_2\}$. □

PROPOSITION 3. (i) *The maximum of an increasing function $f(x)$ over a polyblock is achieved at a proper vertex of this polyblock.*

(ii) *The minimum of an increasing function $f(x)$ over a copolyblock is achieved at a proper vertex of this copolyblock.*

Proof. We prove (i). Let $\bar{x}$ be a maximizer of $f(x)$ over a polyblock P. Since a polyblock is the normal hull of its proper vertices, there exists a proper vertex z of P such that $\bar{x} \in [a, z]$. Then $f(z) \geq f(\bar{x})$ because $z \geq \bar{x}$, so z also must be an optimal solution. The proof of (ii) is similar. □

LEMMA 4. (i) *If $a < x < b$, then the set $[a, b] \setminus (x, b]$ is a polyblock with vertices*

$$u^i = b + (x_i - b_i)e^i, \quad i = 1, \dots, n. \tag{3}$$

(ii) *If $a < x < b$, then the set $[a, b] \setminus [a, x)$ is a copolyblock with vertices*

$$v^i = a + (x_i - a_i)e^i, \quad i = 1, \dots, n.$$

Proof. We prove (i). Let $K_i = \{z \in [a, b] |\ x_i < z_i\}$. Since $(x, b] = \cap_{i=1,\dots,n} K_i$, we have $[a, b] \setminus (x, b] = \cup_{i=1,\dots,m}([a, b] \setminus K_i)$, proving the assertion because $[a, b] \setminus K_i = \{z |\ a_i \leq z_i \leq x_i,\ a_j \leq z_j \leq b_j\ \forall j \neq i\} = [a, u^i]$. The proof of (ii) is similar. □

Note that $u^1, \dots, u^n$ are the n vertices of the hyperrectangle $[x, b]$ that are adjacent to b, while $v^1, \dots, v^n$ are the n vertices of the hyperrectangle $[a, x]$ that are adjacent to a.

3. Monotonic functions. It has been shown in [10] that by simple transformations, any optimization problem dealing with d.m. functions can be converted to the canonical form (MO), where f, g, h are increasing functions. These transformations are based on the following properties of monotonic functions [10].

PROPOSITION 5. *If $f_i,\ i = 1, \dots, m$, are d.m. functions, then their upper and lower envelopes*

$$\max_{i=1,\dots,m} f_i(x), \qquad \min_{i=1,\dots,m} f_i(x)$$

are also d.m.

Proof. Let $f_i(x) = g_i(x) - h_i(x)$, where g_i, h_i are increasing. It is easily seen that

$$\max_{i=1,\dots,m}\{g_i - h_i\} = \sum_{k=1,\dots,m} g_k - \min_{i=1,\dots,m}\left\{\sum_{k\neq i} g_k + h_i\right\},$$

$$\min_{i=1,\dots,m}\{g_i - h_i\} = \sum_{k=1,\dots,m} g_k - \max_{i=1,\dots,m}\left\{\sum_{k\neq i} g_k + h_i\right\}.$$

The conclusion follows, because the sum and the upper and lower envelopes of finitely many increasing functions are obviously increasing. ∎

COROLLARY 6. *Any conjunctive or disjunctive system of d.m. inequalities is equivalent to a single d.m. inequality:*

$$f_i(x) \le 0\ \forall i = 1, \dots, m \qquad \Leftrightarrow \max_{i=1,\dots,m} f_i(x) \le 0,$$
$$f_i(x) \le 0 \text{ for at least one } i = 1, \dots, m \Leftrightarrow \min_{i=1,\dots,m} f_i(x) \le 0.$$

PROPOSITION 7. *Any constrained d.m. optimization problem*

$$\max\{f_0(x)|\ f_i(x) \le 0,\ i = 1, \dots, m,\ x \in [a, b] \subset \mathbb{R}^n_+\}, \tag{4}$$

with d.m. functions $f_i : [a, b] \to \mathbb{R},\ i = 0, 1, \dots, m$, can be transformed into a monotonic optimization problem in canonical form.

Proof. Since $f_0(x) = f_{01}(x) - f_{02}(x)$, with increasing functions $f_{01}(x), f_{02}(x)$, by writing the problem (4) as

$$\max\{f_{01}(x) + t - f_{02}(b)|\ t - f_{02}(b) + f_{02}(x) \le 0, f_i(x) \le 0, i = 1, \dots, m,$$
$$0 \le t \le f_{02}(b) - f_{02}(a), a \le x \le b\}$$

and changing the notation, one can assume that the objective function, i.e., the function $f_0(x)$ in (4), is increasing. Furthermore, by Corollary 6, the set of d.m. inequalities $f_i(x) \le 0,\ i = 1, \dots, m$, can be rewritten as a single inequality $g(x) - h(x) \le 0$, with increasing functions g, h. In turn, it is easily verified that the inequality $g(x) - h(x) \le 0$ has a solution $x \in [a, b]$ if and only if $h(b) - g(a) \ge 0$ and there exists $x_{n+1} \in \mathbb{R}$ such that

$$0 \le x_{n+1} \le h(b) - g(a), \quad g(x) + x_{n+1} \le h(b) \le h(x) + x_{n+1}.$$

Therefore, setting $z = (x, x_{n+1}),\ w(z) = f_0(x),\ u(z) = g(x) + x_{n+1} - h(b),\ v(z) = h(x) + x_{n+1} - h(b),\ c = (a, 0),$ *and* $d = (b, h(b) - g(a))$, we obtain the problem in z

$$\max\{w(z)|\ u(z) \le 0 \le v(z),\ z \in [c, d]\}$$

which is identical to (MO) except for the notation. ∎

4. Monotonicity cuts. A popular method for exploiting special structure in nonconvex optimization problems is to develop cuts that allow one to successively remove unfit portions of the region of the feasible set currently of interest.

For the problem (MO), two kinds of cuts can be introduced. Cuts of the first kind are based on a special separation property of normal sets similar to the separation property of convex sets and are, therefore, called *separation cuts*. Cuts of the second kind aim at tightening the box containing the feasible portion currently still of interest and are referred to as *reduction cuts*.

I. *Separation cut*. This cut was developed earlier in [5] and [10]. For a closed normal set G in $[a,b]$, a point $\bar{x} \in G$ is called an *upper boundary point* if the cone $K_{\bar{x}} := \{x|\ x > \bar{x}\}$ contains no point $x \in G$. The set of all upper boundary points of G is called its *upper boundary* and is denoted by $\partial^+ G$. Clearly, if $\bar{z} \in [a,b] \setminus G$, then the first point of G in the line segment joining $\bar{z}$ to a is an upper boundary point of G.

PROPOSITION 8 (see [10]). *Let G be a closed normal set in a box $[a,b]$, and $\bar{z} \in [a,b] \setminus G$. If $\bar{x}$ is any point on $\partial^+ G$ such that $\bar{x} < \bar{z}$ then the cone $K_{\bar{x}} := \{x|\ x > \bar{x}\}$ contains $\bar{z}$ but is disjoint from G.* (We say that this cone separates strictly $\bar{z}$ from G.)

Proof. If there were $x \in G$ such that $x > \bar{x}$, then by normality, $[\bar{x}, x] \subset G$; hence $G \cap K_{\bar{x}} \supset [\bar{x}, x] \cap K_{\bar{x}} \neq \emptyset$, conflicting with $\bar{x}$ being an upper boundary point. □

We shall refer to the cone $K_{\bar{x}}$ as a *separation cut with vertex* $\bar{x}$. Since, by Lemma 2, $[a,b] \setminus K_x = [a,b] \setminus (x,b]$ is a polyblock, Proposition 8 also says that for any $\bar{z} \in [a,b] \setminus G$, there exists a polyblock $P \supset G$ such that $z \notin P$.

The next corollary shows that with respect to compact normal sets, polyblocks play a role similar to that of polytopes with respect to compact convex sets.

COROLLARY 9 ([10]). *Any compact normal set $G \subset [a,b]$ is the intersection of a family of polyblocks.*

Proof. Clearly, $P_0 := [a,b]$ is a polyblock containing G. Let $\{P_i, i \in I\}$ be the family of all polyblocks containing G. We have $G = \cap_{i \in I} P_i$ because if there were $z \in \cap_{i \in I} P_i \setminus G$, there would exist, by Proposition 8, a polyblock $P \supset G$ (i.e., $P = P_i$ for some $i \in I$) such that $z \notin P$, a contradiction. □

Sometimes we need to perform a sequence of conjunctive separation cuts. The following proposition, which is an improved version of an earlier result ([10, Proposition 18]) indicates how to compute the proper vertex set of the resulting polyblock.

PROPOSITION 10. *Let P be a polyblock with proper vertex set $T \subset [a,b]$, let $x \in [a,b]$ be such that $T_* := \{z \in T|\ z > x\} \neq \emptyset$. For every $z \in T_*$ and every $i = 1, \ldots, n$, define $z^i = z + (x_i - z_i)e^i$. Then the vertex set of the polyblock $P \setminus (x,b]$ is*

$$T' = (T \setminus T_*) \bigcup \{z^i = z + (x_i - z_i)e^i|\ z \in T_*, i \in \{1, \ldots, n\}\}. \tag{5}$$

The improper elements of T' are those $z^i = z + (x_i - z_i)e^i$, with $z \in T_*$, for which there exists $y \in T$ such that $y \geq x$ and the set $J(z,y) := \{j|\ z_j > y_j\}$ has i as its unique element.

Proof. Since $[a,z] \cap (x,b] = \emptyset$ for every $z \in T \setminus T_*$, it follows that $P \setminus (x,b] = P_1 \cup P_2$, where P_1 is the polyblock generated by $T \setminus T_*$ and $P_2 = (\cup_{z \in T_*}[a,z]) \setminus (x,b] = \cup_{z \in T_*}([a,z] \setminus (x,b])$. Noting that $[a,b] \setminus (x,b]$ is a polyblock with vertices $u^i = b + (x_i - b_i)e^i,\ i = 1, \ldots, n$ (see (3)), we can then write $[a,z] \setminus (x,b] = [a,z] \cap ([a,b] \setminus (x,b]) = [a,z] \cap (\cup_{i=1,\ldots,n}[a,u^i]) = \cup_{i=1,\ldots,n}[a,z] \cap [a,u^i] = \cup_{i=1,\ldots,n}[a, z \wedge u^i]$; hence $P_2 = \cup\{[a, z \wedge u^i]|\ z \in T_*, i = 1, \ldots, n\}$, which shows that the vertex set of $P \setminus (x,b]$ is the set T' given by (5).

It remains to show that every $z \in T \setminus T_*$ is proper, while a $z^i = z + (x_i - z_i)e^i$ with $z \in T_*$ is improper if and only if $J(z,y) = \{i\}$ for some $y \in T$ such that $y \geq x$.

Since every $y \in T \setminus T_*$ is proper in T, while $z^i \leq z \in T$ for every $z^i = z + (x_i - z_i)e^i$, it is clear that every $y \in T \setminus T_*$ is proper in T'. Therefore, an improper element of T' must be some z^i such that $z^i \leq y$ for some $y \in T' \setminus \{z^i\}$. Two cases are possible: either $y \in T$ or $y \in T' \setminus T$. In the former case ($y \in T$), since obviously $x \leq z^i$ we must have $x \leq y$; furthermore, $z_j = z^i_j \leq y_j \ \forall j \neq i$; hence, since $z \not\leq y$, it follows that $z_i > y_i$, i.e., $J(z,y) = \{i\}$. In the latter case ($y \in T' \setminus T$), $z^i \leq y^l = y + (x_l - y_l)e^l$ for some $y \in T_*$ and some $l \in \{1, \ldots, n\}$. Then, since $y^l \neq z^i$, we must have $y \neq z$ and $z^i_j \leq y^l_j \ \forall j = 1, \ldots, n$. If $l = i$, then this implies that $z_j = z^i_j \leq y^i_j = y_j \ \forall j \neq i$; hence, since $z \not\leq y$ it follows that $z_i > y_i$ and $J(z,y) = \{i\}$. On the other hand, if $l \neq i$, then from $z^i_j \leq y^l_j \ \forall j = 1, \ldots, n$ we have $z_j \leq y_j \ \forall j \neq i$ and, again, since $z \not\leq y$, we must have $z_i > y_i$, so $J(z,y) = \{i\}$. Thus an improper $z^i = z + (x_i - z_i)e^i$ must satisfy $J(z,y) = \{i\}$ for some $y \in T$ such that $y \geq x$. Conversely, if $J(z,y) = \{i\}$ for some $y \in T$ such that $y \geq x$, then $z^i_j = z_j \leq y_j \ \forall j \neq i$; hence $z^i \leq y$ (because $z^i_i = x_i \leq y_i$) and so, noting that $y \in T'$ if $y \notin T_*$, while $y \leq y^i \in T'$ otherwise, we see that z^i is improper in T'. This completes the proof of the proposition. □

II. *Reduction cuts.* Consider any box $[p,q] \subset [a,b]$ and the problem (MO) restricted to $[p,q]$:

$$\max\{f(x) |\ g(x) \leq 0 \leq h(x),\ x \in [p,q]\}. \tag{6}$$

If a feasible solution of (MO) is known with objective function value γ, then we would like to recognize whether or not the box $[p,q]$ contains a feasible solution to (MO) with objective function value at least equal to γ. Furthermore, if a feasible solution x to (MO) satisfies $g(x) < 0$, then the intersection point x' of the half-line $\{p + \lambda(x-p) |\ \lambda > 0\}$ with the surface $g(x) = 0$ will give a better feasible solution to (MO), so we are interested only in feasible solutions satisfying $g(x) = 0$. Therefore, we can replace the problem (6) by

$$\max\{f(x) |\ g(x) \leq 0 \leq \min\{h(x), g(x), f(x) - \gamma\},\ x \in [p,q]\}. \tag{7}$$

Define

$$h_\gamma(x) := \min\{h(x), g(x), f(x) - \gamma\}, \tag{8}$$

$$H_\gamma := \{x \in [p,q] |\ h_\gamma(x) \geq 0\}. \tag{9}$$

PROPOSITION 11. (i) *Let* $h_\gamma(q) \geq 0$ *and* $p' = q - \sum_{i=1}^n \alpha_i(q_i - p_i)e^i$, *where*

$$\alpha_i = \sup\{\alpha |\ 0 \leq \alpha \leq 1,\ h_\gamma(q - \alpha(q_i - p_i)e^i) \geq 0\}, \quad i = 1, \ldots, n. \tag{10}$$

Then the box $[p', q]$ *still contains all feasible solutions of the problem* (7).

(ii) *Let* $g(p) \leq 0$ *and* $q' = p + \sum_{i=1}^n \beta_i(q_i - p_i)e^i$, *where*

$$\beta_i = \sup\{\beta |\ 0 \leq \beta \leq 1,\ g(p + \beta(q_i - p_i)e^i) \leq 0\}, \quad i = 1, \ldots, n. \tag{11}$$

Then the box $[p, q']$ *still contains all feasible solutions of the problem* (7).

Proof. It suffices to prove (i) because the proof of (ii) is similar. Since $p'_i = \alpha_i p_i + (1 - \alpha_i)q_i$ with $0 \leq \alpha_i \leq 1$, it follows that $p_i \leq p'_i \leq q_i \ \forall i = 1, \ldots, n$, i.e., $[p', q] \subset [p, q]$. For any $x \in H_\gamma \cap [p,q]$ we have, by conormality, $[x, q] \subset H_\gamma$, so if we define $x^i := q - (q_i - x_i)e^i$, then $x^i \in H_\gamma, i = 1, \ldots, n$. But $x_i \leq q_i$, so

$x^i = q - \alpha(q_i - p_i)e^i$ for some α (depending on i) such that $0 \le \alpha \le 1$. This implies that $\alpha \le \alpha_i$, i.e., $x^i \ge q - \alpha_i(q_i - p_i)e^i,\ i = 1, \dots, n$, and consequently $x \ge p'$, i.e., $x \in [p', q]$. Thus, $H_\gamma \cap [p, q] \subset H_\gamma \cap [p', q]$, which completes the proof because the converse inclusion is obvious from the fact that $[p', q] \subset [p, q]$. □

Clearly the box $[p', q]$ defined in (i) is obtained from $[p, q]$ by applying the cut $\cup_{i=1}^n \{x|\ x_i < p'_i\}$, while the box $[p, q']$ defined in (ii) is obtained from $[p, q]$ by applying the cut $\cup_{i=1}^n \{x|\ x_i > q'_i\}$. The former cut is referred to as a *lower cut with vertex* p' and the latter cut as an *upper cut with vertex* q'.

Given the problem (6) and a box $[p, q]$ satisfying $g(p) \le 0 \le h_\gamma(q)$, let p' be the vertex of the lower cut defined above.

If $g(p') \le 0$ and q' is the vertex of the upper cut for $[p', q]$, i.e., $q' = p' + \sum_{i=1}^n \beta_i(q_i - p'_i)e^i$, where

$$\beta_i = \sup\{\beta|\ 0 \le \beta \le 1,\ g(p' + \beta(q_i - p'_i)e^i) \le 0\}, \quad i = 1, \dots, n,$$

then the box $[p', q']$ is called a γ-*reduction* of $[p, q]$, written $[p', q'] = \text{red}_\gamma[p, q]$.

If $g(p') > 0$, then we set $\text{red}_\gamma[p, q] = \emptyset$.

By Proposition 11 the set $\text{red}_\gamma[p, q]$ still contains all feasible solutions x to (MO) satisfying $x \in [p, q], f(x) \ge \gamma$. In other words, by replacing the box $[p, q]$ with $\text{red}_\gamma[p, q]$ no feasible solution $x \in [p, q]$ satisfying $f(x) \ge \gamma$ is lost.

5. The *S*-adjustment operation. The above cuts are valid for the general monotonic optimization problem (MO). We now extend these cuts to the discrete optimization problem (DMO).

As we saw by Propositions 8 and 11, a separation cut is determined by the vertex $\bar{x} \in \partial^+ G$ of the cone to be removed, while a lower (upper) cut is determined by the vertex p' (or q') of the cone complementary to the part to be removed. Since the feasible set of (DMO) is smaller than that of (MO), these cuts can be adjusted to take account of the discrete constraint. The adjustment consists of moving the vertex of the cut (i.e., $\bar{x}, q'$, or p') to a suitable point inside, so that the new cut is deeper but still leaves unaffected the set of feasible solutions in $[p, q]$ with objective function value no less than γ.

The following proposition transforms the discrete problem (DMO) into a continuous problem with an implicitly defined monotonic constraint.

PROPOSITION 12. *Define* $\tilde{G} = (G \cap S^*)^\rceil$, *the normal hull of the set* $G \cap S^*$. *Then problem* (DMO) *is equivalent to*

$$\max\{f(x)|\ x \in \tilde{G} \cap H\}. \tag{12}$$

Proof. Since the feasible set of (DMO) is contained in the feasible set of (12), the optimal value of (DMO) cannot exceed that of problem (12). Conversely, if $\bar{x}$ solves (12), then $\bar{x} \in \tilde{G} \cap H$, but $\tilde{G} = \cup_{z \in G \cap S^*}[0, z]$ by Proposition 1, so $\bar{x} \in [0, z]$ for some $z \in G \cap S^*$; furthermore, since $\bar{x} \in H$ and $\bar{x} \le z$ we must also have $z \in H$ by conormality of H. Consequently, $\bar{x} \le z \in G \cap S^* \cap H$, and hence $f(\bar{x}) \le f(z)$ for some z feasible to (DMO). This implies that the optimal value of (12) cannot exceed that of (DMO). Therefore, the two problems (DMO) and (12) have the same optimal value. □

Solving problem (DMO) is thus reduced to solving (12) which is a monotonic optimization problem without explicit discrete constraint. The new difficulty, however, is how to handle the polyblock $\tilde{G}$ which is defined only implicitly as the normal hull of $G \cap S^*$.

Fortunately, as we will see shortly, the monotonicity cuts developed in the previous section for the continuous problem (MO) can be adjusted to yield valid cuts for the problem (12) equivalent to the discrete problem (DMO).

Consider a box $[p,q] \subset [a,b]$. Given any point $x \in [p,q]$, we define the *lower S-adjustment* of x to be the point

(13)
$$\lfloor x \rfloor_{S^*} = \tilde{x}, \text{ with } \tilde{x}_i = \begin{cases} \max\{y_i |\ y \in S^* \cup \{p\},\ y_i \le x_i\}, & i = 1,\dots,s, \\ x_i, & i = s+1,\dots,n \end{cases}$$

and the *upper S-adjustment* of x to be the point

(14)
$$\lceil x \rceil_{S^*} = \hat{x}, \text{ with } \hat{x}_i = \begin{cases} \min\{y_i |\ y \in S^* \cup \{q\},\ y_i \ge x_i\}, & i = 1,\dots,s, \\ x_i, & i = s+1,\dots,n. \end{cases}$$

A frequently encountered special case is when $S = S_1 \times \cdots \times S_s$ and every S_i is a finite set of real numbers. In this case

$$\tilde{x}_i = \begin{cases} \max\{\xi |\ \xi \in S_i \cup \{p_i\},\ \xi \le x_i\}, & i = 1,\dots,s, \\ x_i, & i = s+1,\dots,n, \end{cases} \tag{15}$$

$$\hat{x}_i = \begin{cases} \min\{\xi |\ \xi \in S_i \cup \{q_i\},\ x_i \le \xi\}, & i = 1,\dots,s, \\ x_i, & i = s+1,\dots,n. \end{cases} \tag{16}$$

(For example, if each S_i is the set of integers and $p_i, q_i \in S_i$, then $\tilde{x}_i$ is the largest integer no larger than x_i while $\hat{x}_i$ is the smallest integer no less than x_i.)

PROPOSITION 13. (i) *If $\bar{x}$ is the vertex of a separation cut for the problem* (MO), *then $\tilde{x} := \lfloor \bar{x} \rfloor_{S^*}$ is the vertex of a separation cut for the problem* (12).

(ii) *If q' is the vertex of an upper cut for* (MO), *then $\lfloor q' \rfloor_{S^*}$ is the vertex of an upper cut for the problem* (12).

(iii) *If p' is the vertex of a lower cut for* (MO), *then $\lceil p' \rceil_{S^*}$ is the vertex of a lower cut for the problem* (12).

Proof. We prove (i). If $\bar{x}$ is the vertex of a separation cut for the problem (MO), then $(\bar{x}, b] \cap G = \emptyset$. For any $y \in (\tilde{x}, b] \cap G \cap S^*$, since $y \in (\tilde{x}, b]$ we have $y_i > \tilde{x}_i$ for every $i = 1,\dots,n$. But $\tilde{x}_i = \bar{x}_i$ for $i = s+1,\dots,n$, so $y_i > \bar{x}_i,\ i = s+1,\dots,n$. On the other hand, since $y \in G \cap S^*$ while $(\bar{x}, b] \cap G = \emptyset$, it follows that $y \notin (\bar{x}, b]$; i.e., there is at least one $i_0 \in \{1,\dots,n\}$ such that $y_{i_0} \le \bar{x}_{i_0}$. Hence $i_0 \in \{1,\dots,s\}$. Since $y \in S^*$ and $y_{i_0} \le x_{i_0}$, it follows from the definition of $\tilde{x}$ that $\tilde{x}_{i_0} \ge y_{i_0}$, conflicting with $y_i > \tilde{x}_i\ \forall i$. Therefore $(\tilde{x}, b] \cap G \cap S^* = \emptyset$, as was to be proved. Analogously we can prove (ii) and (iii). □

Thus any monotonicity cut for the continuous problem (MO) can be adjusted to yield a monotonicity cut for the discrete problem (DMO).

6. The discrete polyblock algorithm. In an earlier paper [10] a polyblock approximation algorithm was developed for solving continuous monotonic optimization problems. Using the S-adjustment operation we now extend this algorithm to solve the discrete monotonic optimization problem (DMO).

Since the feasible set of (DMO) is obviously contained in the box $[a^*, b^*]$, where $a^* = \lceil a \rceil_{S^*}$ (upper S-adjustment of a), $b^* = \lfloor b \rfloor_{S^*}$ (lower S-adjustment of b), we can assume, without loss of generality, that $a = a^*, b = b^*$. If $g(a) > 0$, i.e., $a \notin G$,

then, since $g(x)$ is increasing, it follows that $g(x) > 0\ \forall x \in [a,b]$, and the problem is infeasible. Similarly, if $h(b) < 0$, i.e., $b \notin H$, then $h(x) < 0\ \forall x \in [a,b]$, and the problem is infeasible. Therefore, we also assume that

$$a \in G, \quad b \in H. \tag{17}$$

Now to solve (DMO) we will construct a sequence of polyblocks $P_0 \supset P_1 \supset \cdots$, together with a sequence of numbers $\gamma_0 \le \gamma_1 \le \cdots$, satisfying two conditions:

(C1) $\gamma_k = f(\hat{x}^k)$ for some $\hat{x}^k \in \tilde{G} \cap H$ if $\gamma_k > -\infty$,

(C2) $P_k \supset \tilde{G}_k \cap H$,

where $\tilde{G}_k = (G \cap S_k^*)^\rceil$, $S_k^* = \{x \in S^* |\ f(x) > \gamma_k\}$.

Start with an initial polyblock $P_0 \supset \tilde{G} \cap H$, e.g., $P_0 = [a,b]$, with vertex set $T_0 = \{b\}$ and $\gamma_0 = -\infty$. At iteration $k = 0, 1, \ldots$, let P_k be the current polyblock, T_k its proper vertex set, γ_k the current best value, and $\hat{x}^k$ the current best solution, satisfying (C1) and (C2). Perform the following transformation (*) of T_k:

(*) For every $v \in T_k$ define $\mathrm{red}^*_{\gamma_k}[a,v]$ to be the γ_k-reduction of the box $[a,v]$, S-adjusted as indicated above. If $\mathrm{red}^*_{\gamma_k}[a,v] = \emptyset$, then drop v (in particular delete any $v \notin H$); if $\mathrm{red}^*_{\gamma_k}[a,v] \neq \emptyset$, denote the highest vertex of $\mathrm{red}^*_{\gamma_k}[a,v]$ again by v, and if $f(v) \le \gamma_k$, then drop v.

Let $\tilde{T}_k$ be the set that results from T_k upon the transformation (*) (note that $\tilde{T}_k \subset H$). If $\tilde{T}_k = \emptyset$, the procedure terminates: the current best feasible solution is optimal (if $\gamma_k > -\infty$) or the problem is infeasible (if $\gamma_k = -\infty$). If $\tilde{T}_k \neq \emptyset$, let P_k be the polyblock with vertex set $\tilde{T}_k$ and select

$$v^k \in \mathrm{argmax}\{f(x) |\ x \in \tilde{T}_k\}.$$

Two cases may occur:

Case 1. $v^k \in G \cap S^*$. Since $v^k \in H$ we have $v^k \in G \cap H \cap S_k^*$, so v^k is a feasible solution of (DMO) with objective function value no less than γ_k. We set $\hat{x}^{k+1} = v^k, \gamma_{k+1} = f(v^k)$, and define P_{k+1} to be the polyblock with vertex set $T_{k+1} = \tilde{T}_k \setminus \{v^k\}$. Also in this case we define $\tilde{v}^k = v^k$.

Case 2. $v^k \notin G \cap S^*$. Then we find $x^k = \pi_G(v^k)$, the first point of G on the line segment joining v^k to a. If $x^k \in S_k^*$, then set $\tilde{v}^k = x^k$; otherwise, set $\tilde{v}^k = \lfloor x^k \rfloor_{S_k^*}$. Perform the cut with vertex at $\tilde{v}^k$, yielding a polyblock P_{k+1}. Compute the proper vertex set T_{k+1} of P_{k+1} according to Proposition 10.

If a new feasible solution has appeared with a better objective function value than γ_k, then let $\hat{z}^{k+1}$ be the best among such solutions, and set $\gamma_{k+1} = f(\hat{z}^{k+1})$; otherwise, set $\hat{z}^{k+1} = \hat{z}^k, \gamma_{k+1} = \gamma_k$.

PROPOSITION 14. *Let $\gamma_{k+1} = f(\hat{x}^{k+1})$, where $\hat{x}^{k+1}$ is the new current best feasible solution. The polyblock P_{k+1} still contains $G \cap H \cap S_{k+1}^*$ and $P_{k+1} \subset P_k \setminus (\tilde{v}^k, b]$ (so conditions* (C1), (C2) *still hold for $k \leftarrow k+1$).*

Proof. In Case 1, we have $P_{k+1} \supset P_k \setminus [a, v^k]$, but $P_k \supset G \cap H \cap S_k^*$, while $[0, v^k] \subset \{x \in P_k |\ f(x) \le f(v^k) = \gamma_{k+1}\}$, and hence $P_{k+1} \supset G \cap H \cap S_{k+1}^*$; furthermore, since $v^k = \tilde{v}^k, (v^k, b] \cap P_k = \emptyset$ (v^k is a proper vertex of P_k), we have $P_k \setminus (\tilde{v}^k, b] = P_k \supset P_{k+1}$. In Case 2, if $v^k \in G \setminus S^*$, then, since v^k is a proper vertex of P_k and $v^k \in G$, one must have $v^k \in \partial^+ G$, i.e., $(v^k, b] \cap G = \emptyset$; hence, by Proposition 13, $(\tilde{v}^k, b] \cap G \cap S_k^* = \emptyset$, and consequently, $P_{k+1} \supset G \cap H \cap S_k^* \supset G \cap H \cap S_{k+1}^*$. On the other hand, if $v^k \notin G$, then $(v^k, b] \cap G = \emptyset$; hence, again by Proposition 13, $(\tilde{v}^k, b] \cap G \cap S_k^* = \emptyset$, and $P_{k+1} \supset G \cap H \cap S_k^* \supset G \cap H \cap S_{k+1}^*$. That $P_{k+1} \subset P_k \setminus (\tilde{v}^k, b]$ follows from Proposition 10. □

Thus, P_{k+1} and γ_{k+1} will still satisfy (C1), (C2) (for $k \leftarrow k+1$). We can then go to iteration $k+1$.

In a formal way, assuming that $a, b \in S^*$, we can state the following algorithm.

ALGORITHM 1 (discrete polyblock algorithm).

Initialization. Take an initial polyblock $P_0 \supset G \cap H$, with proper vertex set T_0. Let $\hat{x}^0$ be the best feasible solution available (the current best feasible solution), $\gamma_0 := f(\hat{x}^0)$. If no feasible solution is available, let $\gamma_0 := -\infty$. Set $k := 0$.

Step 1. Perform the transformation (*) of T_k. Let $\tilde{T}_k$ be the resulting set ($\tilde{T}_k \subset H$). Reset $T_k := \tilde{T}_k$.

Step 2. If $T_k = \emptyset$, terminate: if $\gamma_k = -\infty$, the problem is infeasible; if $\gamma_k > -\infty$, $\hat{x}^k$ is an optimal solution.

Step 3. If $T_k \neq \emptyset$, select $v^k \in \operatorname{argmax}\{f(v) |\ v \in T_k\}$.

If $v^k \in G \cap S^*$, terminate: v^k is an optimal solution.

Step 4. If $v^k \in G \setminus S^*$, compute $\tilde{v}^k := \lfloor v^k \rfloor_{S_k^*}$ (using formula (13) for $S^* := S_k^*$).

If $v^k \notin G$, compute $x^k := \pi_G(v^k)$ and define $\tilde{v}^k := x^k$ if $x^k \in S_k^*$, $\tilde{v}^k := \lfloor x^k \rfloor_{S_k^*}$ if $x^k \notin S_k^*$.

Step 5. Let $T_{k,*} := \{z \in T_k |\ z > \tilde{v}^k\}$. Compute

$$T_k' := (T_k \setminus T_{k,*}) \bigcup \{z^{k,i} = z + (\tilde{v}_i^k - z_i)e^i |\ z \in T_{k,*}, i = 1, \dots, n\}. \tag{18}$$

Let T_{k+1} be the set obtained from T_k' by removing every z^i such that $\{j |\ z_j > y_j\} = \{i\}$ for some $z \in T_{k,*}$ and $y \in T_{k,*}^+$.

Step 6. Determine the new current best feasible solution $\hat{x}^{k+1}$ and $\gamma_{k+1} := f(\hat{x}^{k+1})$. Increment k and return to Step 1.

To prove the convergence of Algorithm 1 assume that either $s = n$ (so that S is a finite subset of $\mathbb{R}^n$) or there exists a constant $\alpha > 0$ such that

$$\min_{i=s+1,\dots,n} (x_i - a_i) \geq \alpha \quad \forall x \in H. \tag{19}$$

THEOREM 15. *If $s = n$, then Algorithm* 1 *is finite. If $s < n$ and condition* (19) *holds, then either Algorithm* 1 *is finite or it generates an infinite sequence of feasible solutions converging to an optimal solution.*

Proof. At each iteration k a pair $v^k, \tilde{v}^k$ is generated such that $v^k \notin (G \cap S_k^*)^\rceil$, $\tilde{v}^k \in (G \cap S_k^*)^\rceil$ and the rectangle $(\tilde{v}^k, b]$ contains no point of P_l with $l > k$ and hence no $\tilde{v}^l$ with $l > k$. Therefore, there can be no repetition in the sequence $\{\tilde{v}^0, \tilde{v}^1, \dots, \tilde{v}^k, \dots\} \subset S^*$. This implies finiteness of the algorithm in the case $s = n$ since then $S^* = S$ is finite. In the case $s < n$, if the sequence $\{v^k\}$ is infinite, it follows from the fact $(\tilde{v}_1^k, \dots, \tilde{v}_s^k) \in S^*$ that for some sufficiently large k_0,

$$\tilde{v}_i^k = \tilde{v}_i, \quad i = 1, \dots, s, \quad \forall k \geq k_0. \tag{20}$$

On the other hand, since $v^k \in H$ we have from (19) that

$$\min_{i=s+1,\dots,n} (v_i^k - a_i) \geq \alpha \quad \forall k. \tag{21}$$

We show that $v^k - x^k \to 0$ as $k \to \infty$. Suppose the contrary, that there exist $\eta > 0$ and an infinite sequence k_l such that $\|v^{k_l} - x^{k_l}\| \geq \eta > 0$ $\forall l$. For all $\mu > l$ we have $v^{k_\mu} \notin (\tilde{v}^{k_l}, v^{k_l}]$ because $P_{k_\mu} \subset P_{k_l} \setminus (\tilde{v}^{k_l}, b]$. Since $\tilde{v}_i^k = x_i^k$ $\forall i > s$, we then derive

$$\|v^{k_\mu} - v^{k_l}\| \geq \min_{i=s+1,\dots,n} |v_i^{k_l} - \tilde{v}_i^{k_l}| = \min_{i=s+1,\dots,n} |v_i^{k_l} - x_i^{k_l}|.$$

On the other hand, $\min\{v_i^{k_l} - a_i |\ i = s+1, \ldots, n\} \geq \alpha$ by (21) because $v^{k_l} \in H$, while x^{k_l} lies on the line segment joining a to v^{k_l}; thus

$$v_i^{k_l} - x_i^{k_l} = \frac{v_i^{k_l} - a_i}{\|v^{k_l} - a\|} \|v^{k_l} - x^{k_l}\| \geq \frac{\alpha\eta}{\|b - a\|} \quad \forall i > s.$$

Therefore,

$$\|v^{k_\mu} - v^{k_l}\| \geq \min_{i=s+1,\ldots,n} |v_i^{k_l} - x_i^{k_l}| \geq \frac{\alpha\eta}{\|b - a\|},$$

conflicting with the boundedness of the sequence $\{v^{k_l}\} \subset [a, b]$. Thus, $\|v^k - x^k\| \to 0$ and by boundedness we may assume, by passing to subsequences if necessary, that $x^k \to \bar{x},\ v^k \to \bar{x}$. Then, since $v^k \in H,\ x^k \in G\ \ \forall k$, it follows that $\bar{x} \in G \cap H$ and hence, $\bar{v} \in G \cap H \cap S^*$ for $\bar{v}$ defined by $\bar{v}_i = \tilde{v}_i\ (i = 1, \ldots, s),\ \bar{v}_i = \bar{x}_i\ (i = s+1, \ldots, n)$. Furthermore, $f(v^k) \geq f(v)\ \forall v \in P_k \supset G \cap H \cap S_k$; hence by letting $k \to +\infty$, we have $f(\bar{x}) \geq f(x)\ \forall x \in G \cap H \cap S_{\bar{\gamma}}$, for $\bar{\gamma} = \lim_{k\to\infty} \gamma_k$. The latter in turn implies that $\bar{v} \in \operatorname{argmax}\{f(x) |\ x \in G \cap H \cap S\}$, i.e., $\bar{v}$ is an optimal solution. □

Remark 1. A discrete monotonic *minimization* problem

$$\min\{f(x) |\ g(x) \leq 0 \leq h(x), x \in [a, b] \cap S^*\}$$

can be reduced, by the change of variables $x = b - y$, to the discrete maximization problem

$$\max\{\tilde{f}(y) |\ \tilde{h}(y) \leq 0 \leq \tilde{g}(y), y \in [0, b - a] \cap (b - S^*)\},$$

where $\tilde{f}(y) = -f(b-y),\ \tilde{g}(y) = -g(b-y),\ \tilde{h}(x) = -h(b-y)$ are increasing functions on $[0, b - a]$.

7. Branch-reduce-and-bound algorithm. Algorithm 1 can be interpreted as a branch and bound algorithm in which a node z of the branch and bound tree represents a box $[a, z]$ and branching is performed by splitting a node into n descendants, while the bound over a node z is taken to be $f(z)$. An attractive feature of this algorithm is that, though the bounds are rather rough, the convergence is guaranteed because monotonicity cuts are used in each step to progressively reduce the feasible portion currently of interest. However, since each node has n descendants, the vertex set T_k of the polyblock at iteration k may grow rapidly and reach a prohibitively large size. Furthermore, since this approach requires the problem to be put in the canonical form (MO), some preliminary transformations are necessary which, as we saw in section 3, would introduce a number of additional variables and might increase the dimension of the problem considerably. Therefore, for large scale problems not necessarily in the canonical form we propose an alternative branch and bound version of the algorithm in which each node has only a small number of descendants (typically two descendants).

Consider a discrete d.m. optimization problem in the general form

$$\text{(DDM)}, \qquad \max\{f(x) |\ g(x) - h(x) \leq 0, x \in [a, b] \cap S^*\},$$

where $f(x) = f^+(x) - f^-(x)$ and $f^+, f^-, g, h : \mathbb{R}^n_+ \to \mathbb{R}$ are increasing functions, while S^* is defined by (2); i.e., $S^* = \{x \in \mathbb{R}^n_+ |\ (x_1, \ldots, x_s) \in S\}$ with S being a given discrete subset of $\mathbb{R}^s_+$.

Instead of converting this problem to the canonical form (DMO), we now develop an alternative version of Algorithm 1 for solving it directly.

This alternative algorithm, to be referred to as a *branch-reduce-and-bound algorithm*, is a procedure involving in each iteration three basic operations: branching, reduction, and bounding.

We next describe these operations.

I. *Branching.* Starting from the initial box $M_1 = [a, b]$, we successively bisect it into smaller and smaller boxes using the following subdivision rule:

Let $M = [p, q]$ be a box candidate for subdivision. Compute the numbers $\delta(M) = \max_{i=1,\dots,n}(q_i - p_i) = q_{i_M} - p_{i_M}$, $r_{i_M} = (p_{i_M} + q_{i_M})/2$ and divide M into two boxes

$$M_+ = \{x \in M |\ x_{i_M} \geq (r_{i_M})^{X_i}\},$$
$$M_- = \{x \in M |\ x_{i_M} \leq (r_{i_M})_{X_i}\},$$

where $X_i = \{\xi \in \mathbb{R} |\ \xi = x_i, x \in S\}$ and r^{X_i}, r_{X_i} denote, respectively, the smallest element of $X_i \cup \{q_i\}$ no less than r and the largest element of $X_i \cup \{p_i\}$ no larger than r.

II. *Reduction.* At each iteration we will have a current best value γ of the objective function, together with a set of newly generated boxes that remain for exploration. If $M = [p, q]$ is such a box, to check whether M contains a feasible solution x with $f(x) \geq \gamma$, we use the following lemma.

LEMMA 16. *There exists a feasible solution $x \in [p, q]$ to* (DDM) *satisfying $f(x) \geq \gamma$ only if*

$$h(q) - g(p) \geq 0, \qquad f^+(q) - f^-(p) \geq \gamma, \tag{22}$$

and any such x must be contained in the box $[p', q'] \subset [p, q]$ defined by

$$p' = q - \sum_{i=1}^{n} \alpha_i (q_i - p_i) e^i, \qquad q' = p' + \sum_{i=1}^{n} \beta_i (q_i - p'_i) e^i, \tag{23}$$

where, for $i = 1, \dots, n$,

$$\alpha_i = \sup\{\alpha |\ 0 \leq \alpha \leq 1,\ h(q - \alpha(q_i - p_i)e^i) - g(p) \geq 0,$$
$$f^+(q - \alpha(q_i - p_i)e^i) - \gamma \geq f^-(p)\}, \tag{24}$$
$$\beta_i = \inf\{\beta |\ 0 \leq \beta \leq 1,\ g(p' + \beta(q_i - p'_i)e^i) - h(q) \leq 0,$$
$$f^-(p' + \beta(q_i - p'_i)e^i) \leq f^+(q) - \gamma\}. \tag{25}$$

Proof. Consider any $x \in [p, q]$ satisfying

$$g(x) - h(x) \leq 0, \qquad f(x) = f^+(x) - f^-(x) \geq \gamma. \tag{26}$$

Since $f^+(.), f^-(.), g(.), h(.)$ are increasing, $g(p) \leq g(x) \leq h(x) \leq h(q), f^+(q) \geq f^+(x) \geq f^-(x) + \gamma \geq f^-(p) + \gamma$; hence $h(q) \geq g(p), f^+(q) - \gamma \geq f^-(p)$, so (22) holds.

If $x \not\geq p'$, then there is i such that $x_i < p'_i = q_i - \alpha_i(q_i - p_i)$, i.e., $x_i = q_i - \alpha(q_i - p_i)$ with some $\alpha > \alpha_i$. By virtue of the definition of α_i, this implies that

$$\text{either} \quad h(q - (q_i - x_i)e^i) = h(q - \alpha(q_i - p_i)e^i) < g(p),$$
$$\text{or} \quad f^+(q - (q_i - x_i)e^i) = f^+(q - \alpha(q_i - p_i)e^i) < \gamma + f^-(p).$$

Since $x \le q-(q_i-x_i)e^i$, in the former case $h(x) \le h(q-(q_i-x_i)e^i) < g(p)$, conflicting with $h(x) \ge g(x) \ge g(p)$, while in the second case $f^+(x) \le f^+(q-(q_i-x_i)e^i) < \gamma + f^-(p)$, conflicting with $f^+(x) - \gamma \ge f^-(x) \ge f^-(p)$. Therefore, $x \ge p'$. In a similar way, from $x \in [p', q]$ and (26) we show that $x \le q'$, and thus $x \in [p', q']$. □

We shall denote by $\mathrm{red}^*_\gamma[p, q]$ the box obtained from $[p', q']$ by replacing p', q' with their lower and upper S-adjustments, respectively. It follows from the above that by replacing $[p, q]$ with $\mathrm{red}^*_\gamma[p, q]$ no feasible solution $x \in [p, q]$ with $f(x) \ge \gamma$ is lost.

Remark 2. When there are more than one d.m. constraint $g_j(x) - h_j(x) \le 0,\ j = 1, \dots, m$, the formulas (23)–(24) should be replaced by the following:

$$\alpha_i = \sup\{\alpha |\ 0 < \alpha \le 1,\ h_j(q - \alpha(q_i - p_i)e^i) \ge g_j(p),\ j = 1, \dots, m, \tag{27}$$
$$f^+(q - \alpha(q_i - p_i)e^i) - \gamma \ge f^-(p)\},$$

$$\beta_i = \sup\{\beta |\ 0 < \beta \le 1,\ g_j(p' + \beta(q_i - p'_i)e^i) \le h_j(q),\ j = 1, \dots, m, \tag{28}$$
$$f^-(p' + \beta(q_i - p'_i)e^i) \le f^+(q) - \gamma\}.$$

III. *Bounding.* For every given box $M = [p, q]$ compute a number $\mu(M)$ such that

$$\mu(M) \ge \gamma(M) := \max\{f(x) |\ g(x) - h(x) \le 0, x \in M \cap S^*\}.$$

To ensure convergence, this upper bound must be consistent in the sense that for any infinite nested sequence of boxes $\{M_{k_\nu}\}$ shrinking to a single point x^*,

$$\lim_{\nu \to +\infty} \mu(M_{k_\nu}) = f(x^*). \tag{29}$$

Since $f(x) = f^+(x) - f^-(x)$ with $f^+(x), f^-(x)$ increasing, an obvious bound is $f^+(q_{k_\nu}) - f^-(p_{k_\nu})$, and any bound such that

$$\mu(M_{k_\nu}) \le f^+(q_{k_\nu}) - f^-(p_{k_\nu}) \tag{30}$$

will satisfy (29). Therefore, one can always take $\mu(M) = f^+(q) - f^-(p)$ when a better bound is expensive to compute. In a formal way we can state the following algorithm.

ALGORITHM 2 (branch-reduce-and-bound algorithm for (DDM)).

Initialization. Let $\mathcal{P}_1 := \{M_1\}, M_1 := [a, b], \mathcal{R}_1 := \emptyset$. If some feasible solutions are available, let CBV denote the value of $f(x)$ at the best of them (current best value). Otherwise, set $CBV := -\infty$. Set $k := 1$.

Step 1. Apply S-adjusted reduction to reduce each box $[p, q] \in \mathcal{P}_k$. In particular delete every box $[p, q]$ such that $h(q) - g(p) < 0$. Let $\mathcal{P}'_k := \{\mathrm{red}^*_\gamma[p, q] |\ \ [p, q] \in \mathcal{P}_k\}$ for $\gamma = CBV$.

Step 2. For each box $M \in \mathcal{P}'_k$ compute a bound $\mu(M)$ satisfying (29).

Step 3. Let $\mathcal{S}_k := \mathcal{R}_k \cup \mathcal{P}'_k$. Update CBV, using the new feasible solutions encountered in Steps 1 and 2, if any. Delete every $M \in \mathcal{S}_k$ such that $\mu(M) < CBV$ and let $\mathcal{R}_{k+1}$ be the collection of remaining boxes.

Step 4. If $\mathcal{R}_{k+1} = \emptyset$, then terminate: if $CBV = -\infty$, the problem is infeasible; otherwise, CBV is the optimal value and the feasible solution $\bar{x}$ with $f(\bar{x}) = CBV$ is an optimal solution.

Step 5. If $\mathcal{R}_{k+1} \ne \emptyset$, let $M_k \in \mathrm{argmax}\{\mu(M) |\ M \in \mathcal{R}_{k+1}\}$. Divide M_k into two boxes according to the above described rule. Let $\mathcal{P}_{k+1}$ be the collection of these two subboxes of M_k.

Step 6. Increment k and return to Step 1.

THEOREM 17. *If $s = n$, Algorithm 2 terminates after finitely many iterations, yielding an optimal solution or establishing the infeasibility of the problem. If $s < n$, either Algorithm 2 terminates after finitely many iterations, yielding an optimal solution or establishing the infeasibility of the problem, or it generates an infinite sequence of feasible solutions converging to an optimal solution.*

Proof. If $s = n$, the set S^* is finite. But due to S-adjustment reduction, every box $M = [p, q] \in \mathcal{S}_k$ satisfies $p = \wedge\{x|\ x \in S^* \cap M\},\ q = \vee\{x|\ x \in S^* \cap M\}$. Therefore, the total number of nodes of the branch and bound tree is finite, which implies finiteness of the algorithm itself.

If $s < n$, the subdivision rule implies that the algorithm, whenever infinite, generates a sequence of boxes $\{M_{k_\nu} := [p^{k_\nu}, q^{k_\nu}]\}$ shrinking to a point $x^* = \lim q^{k_\nu} \in S^*$. Since the deletion rule in Step 3 implies, by condition (22) in Lemma 16, that $g(p^{k_\nu}) - h(q^{k_\nu}) \leq 0$, we must have $g(x^*) - h(x^*) \leq 0$; hence $x^* \in \Omega$, the feasible set of (DDM). On the other hand, by the selection rule in Step 5,

$$\mu(M_{k_\nu}) \geq \max\{\mu(M)|\ M \in \mathcal{R}_k\} \geq f(x) \quad \forall x \in \Omega,$$

while by (29), $\mu(M_{k_\nu}) \to f(x^*)$, and hence $f(x^*) \geq f(x)\ \forall x \in \Omega$; i.e., x^* is an optimal solution. If $\hat{x}^k$ is the current best solution at iteration k, then by passing to a subsequence if necessary we may assume $\hat{x}^k \to \hat{x}$ with $f(\hat{x}) = f(x^*)$; i.e., $\hat{x}$ is also an optimal solution. □

Remark 3. The above proof shows that Algorithm 2 converges, even if in Step 2 we always take $\mu(M) = f^+(q) - f^-(p)$ for every box $M = [p, q]$ (so that condition (29) is ensured). Furthermore, in the general case when there are more than one d.m. constraints, the algorithm can be applied without having to convert a system of several d.m. constraints into a single one (and accordingly increase the dimension of the problem). As was mentioned in Remark 2, it suffices in that case to replace, in the reduction operation, the formulas (23)–(24) by (27)–(28). We thus have an extremely simple method for solving any discrete optimization problem of the form (DDM).

Remark 4. The performance of Algorithm 2 critically depends on the reduction operation in Step 1 and also the bounding operation in Step 2. Therefore, although the bound $\mu(M) = f^+(q) - f^-(p)$ (for a box $M = [p, q]$) is sufficient for guaranteeing convergence, tighter bounds are often necessary to achieve reasonable efficiency.

For example, if a polytope D containing the feasible portion in $M = [p, q]$ can be found together with a linear overestimator $L(x)$ of $f(x)$ such that $L(x^M) = f(x^M)$ at a point $x^M \in M = [p, q]$, then an upper bound satisfying (29) is provided by the optimal value of the linear program

$$\max\{L(x)|\ x \in D \cap M\}.$$

In any case, if the problem has the form (DMO), then tighter bounds than $f(q)$ can be computed by either of the following methods based on polyblock approximation.

Method 1: Apply a truncated version of Algorithm 1 (i.e., perform a given number of k iterations of this algorithm) to problem $\max\{f(x)|\ x \in G \cap H \cap [p, q]\}$ (usually $k = 1$ or $k = 2$). If T_k is the vertex set of the last polyblock obtained, then let $\mu(M) = \max\{f(z)|\ z \in T_k\}$.

Method 2: Consider a grid $U = \{c^0, c^1, \ldots, c^n\} \subset \{x \in R^n_+|\ \sum_{i=1}^n x_i = 1\}$, for example, $U = \{c^0, c^1, \ldots, c^k\}$ with

$$c^0 = e/n, \quad c^k = \frac{(n+1)e - ne^k}{n^2}, \quad k = 1, \ldots, n.$$

For each $k = 0, 1, \ldots, n$, let $x^k = \pi(c^k) := p + \lambda_k c^k$, with $\lambda_k = \sup\{\alpha |\ p + \alpha c^k \in G\}$. Next construct a set T as follows:

Step 0. Let $T = \{u^1, \ldots, u^n\}$ with $u^i = q + (x_i^0 - q_i)e^i, i = 1, \ldots, n$. Set $k = 1$.

Step k. Compute $x^k = \pi(c^k)$, let $T_* = \{z \in T |\ z > x^k\}$, and compute

$$T' = (T \setminus T_*) \bigcup \{z^i = z + (x_i^k - z_i)e^i |\ z \in T_*, i \in \{1, \ldots, n\}\}$$

and from T' remove all z^i for which there exists $y \in T_*^+ := \{z \in T |\ z \geq x^k\}$ such that $\{j |\ z_j > y_j\} = \{i\}$. Reset T equal to the set of remaining elements of T'. If $k < n$, let $k \leftarrow k + 1$ and go back to Step k. If $k = n$, stop.

If T is the last set obtained by the above procedure then let $\mu(M) = \max\{f(z) |\ z \in T \cap H\}$.

8. Application: A discrete location problem. As an application, let us consider the following discrete location problem:

(DL) *Given m balls in $\mathbb{R}^n$ of centers $a^i \in R_{++}^n$ and radii $\alpha_i (i = 1, \ldots, m)$, and a bounded discrete set $S \subset \mathbb{R}_+^n$, find the largest ball that has center in S and is disjoint from any of these m balls.* In other words,

$$\begin{array}{l} \text{maximize } z \quad \text{subject to} \\ \|x - a^i\| - \alpha_i \geq z, \quad i = 1, \ldots, m, \\ x \in S \subset R_+^n,\ z \in \mathbb{R}_+. \end{array} \tag{31}$$

This problem is encountered in various applications. For example, in location theory (see, e.g., [4], [1]), it can be interpreted as a "maximin location problem": $a^i,\ i = 1, \ldots, m$, are the locations of m obnoxious facilities, and $\alpha_i > 0$ is the radius of the polluted region of facility i, while an optimal solution is a location $x \in S$ outside all polluted regions and as far as possible from the nearest of these obnoxious facilities. In engineering design (DL) appears as a variant of the "design centering problem" [16], [8], an important special case of which, when $\alpha_i = 0, i = 1, \ldots, m$, is the "largest empty ball problem" [6]: given m points $a^1, \ldots, a^m$ in $\mathbb{R}_{++}^n$ and a bounded set S, find the largest ball that has center in $S \subset \mathbb{R}_+^n$ and contains none of these points.

As a first step toward solving (DL) one can study the following feasibility problem:

(Q(r)) Given a number $r \geq 0$, find a point $x(r) \in S$ lying outside any one of the m balls of centers a^i and radii $\theta_i = \alpha_i + r$.

It has been shown in [14] that this problem can be reduced to

$$\max\{\|x\|^2 - h(x) |\ x \in S\},$$

where $h(x) = \max_{i=1,\ldots,m}(2\langle a^i, x\rangle + \theta_i^2 - \|a^i\|^2)$. Since both $\|x\|^2$ and $h(x)$ are obviously increasing functions, this is a discrete d.m. optimization problem that can be solved by the above presented method.

Clearly if $\bar{r}$ is the maximal value of r such that (Q(r)) is feasible, then $\bar{x} = x(\bar{r})$ will solve (DL). Noting that $\bar{r} \geq 0$ and for any $r > 0$ one has $\bar{r} \geq r$ or $\bar{r} < r$ according to whether (Q(r)) is feasible or not, the value $\bar{r}$ can be found by a Bolzano binary search scheme: starting from an interval $[0, s]$ containing $\bar{r}$, one reduces it by a half at each step by solving a (Q(r)) with a suitable r. Quite encouraging computational results with this scheme have been reported in [14], where each subproblem (Q(r)) was solved by a preliminary version of Algorithm 2 described in [3]. However, it turns out that more complete and much better results can be obtained by a direct application of Algorithm 2 to problem (31). Next we describe this method.

Observe that

$$\begin{aligned} &\|x-a^i\|-\alpha_i \geq z, \quad i=1,\dots,m,\\ \Leftrightarrow\ &\|x\|^2+\|a^i\|^2-2\langle a^i,x\rangle \geq (z+\alpha_i)^2, \quad i=1,\dots,m,\\ \Leftrightarrow\ &\max_{i=1,\dots,m}\{(z+\alpha_i)^2+2\langle a^i,x\rangle-\|a^i\|^2\} \leq \|x\|^2. \end{aligned}$$

Therefore, by setting

$$\varphi(x,z)=\max_{i=1,\dots,m}((z+\alpha_i)^2+2\langle a^i,x\rangle-\|a^i\|^2), \tag{32}$$

problem (31) can be restated as

$$\max\{z|\ \varphi(x,z)-\|x\|^2\leq 0,\ x\in[a,b]\cap S,\ z\geq 0\}, \tag{33}$$

where $\varphi(x,z)$ and $\|x\|^2$ are increasing functions and $[a,b]$ is a box containing S.

To apply Algorithm 2 for solving this problem, observe that the value of z is determined when x is fixed. Therefore, branching should be performed upon the variables x only, using the subdivision rule described in section 6.

Another key operation in Algorithm 2 is bounding: given a box $[p,q]\subset[a,b]$, compute an upper bound $\mu(M)$ for the optimal value of the subproblem

$$\max\{z|\ \varphi(x,z)-\|x\|^2\leq 0,\ x\in S\cap[p,q],\ 0\leq z\}. \tag{34}$$

If $CBV=r>0$ is the largest thus far known value of z at a feasible solution (x,z), then only feasible points (x,z) with $z>r$ are still of interest. Therefore, to obtain a tighter value of $\mu(M)$ one should consider, instead of (34), the problem

$$\max\{z|\ \varphi(x,z)-\|x\|^2\leq 0,\ x\in S\cap[p,q],\ z\geq r\}. \tag{DL(M,r)}$$

Because $\varphi(x,z)$ and $\|x\|^2$ are both increasing, the constraints of (DL(M,r)) can be relaxed to $\varphi(p,z)-\|q\|^2$, so an obvious upper bound is

$$c(p,q):=\max\{z|\ \varphi(p,z)-\|q\|^2\leq 0\}. \tag{35}$$

Although this bound is easy to compute (it is the zero of the increasing function $z\mapsto\varphi(p,z)-\|q\|^2$), it is often not sufficiently efficient. A better bound can be computed by solving a linear relaxation of (DL(M,r)) obtained by omitting the discrete constraint $x\in S$ and replacing $\varphi(x,z)$ with a linear underestimator. Since

$$(z+\alpha_i)^2\geq(r+\alpha_i)^2+2(r+\alpha_i)(z-r),$$

a linear underestimator of $\varphi(x,z)$ is

$$\psi(x,z):=\max_{i=1,\dots,n}\{2(r+\alpha_i)(z-r)+(r+\alpha_i)^2+2\langle a^i,x\rangle-\|a^i\|^2\}.$$

On the other hand,

$$\|x\|^2\leq\sum_{j=1}^{n}[(p_j+q_j)x_j-p_jq_j]\quad\forall x\in[p,q],$$

so by (32) the constraint $\varphi(x,z)-\|x\|^2\leq 0$ can be relaxed to

$$\psi(x,z)-\sum_{j=1}^{n}[(p_j+q_j)x_j-p_jq_j]\leq 0,$$

which is equivalent to the system of linear constraints

$$\max_{i=1,\dots,m}\{2(r+\alpha_i)(z-r)+(r+\alpha_i)^2+2\langle a^i,x\rangle-\|a^i\|^2\}\le\sum_{j=1}^{n}[(p_j+q_j)x_j-p_jq_j].$$

Therefore,

$$\text{(LP}(M,r)\text{)}\qquad \begin{aligned}\mu(M)=\max\Big\{z|\ & 2(r+\alpha_i)(z-r)+(r+\alpha_i)^2+2\langle a^i,x\rangle-\|a^i\|^2\\ & -\sum_{j=1}^{n}[(p_j+q_j)x_j-p_jq_j]\le 0,\ i=1,\dots,m,\\ & z\ge r,\ p\le x\le q\Big\}.\end{aligned}$$

The bounds can be further improved by using valid reduction operations. For this, observe that, since $\varphi(x,r)\le\varphi(x,z)$ for $r\le z$, the feasible set of (DL(M,r)) is contained in the set

$$\{x|\ \varphi(x,r)-\|x\|^2\le 0,\ x\in[p,q]\},$$

so, according to Lemma 16, if p',q' are defined by (23)–(25) with

$$g(x)=\varphi(x,r),\quad h(x)=\|x\|^2,$$

and $\tilde p=\lceil p'\rceil_{S^*},\tilde q=\lfloor q'\rfloor_{S^*}$, then the box $[\tilde p,\tilde q]\subset[p,q]$ is a valid reduction of $[p,q]$.

Thus, Algorithm 2 specialized to problem (DL) becomes the following algorithm.

Algorithm 3 (branch-reduce-and-bound algorithm for (DL)).

Initialization. Let $\mathcal{P}_1:=\{M_1\},M_1:=[a,b],\mathcal{R}_1:=\emptyset$. If some feasible solution $(\bar x,\bar z)$ is available, let $r:=\bar z$ be CBV, the current best value. Otherwise, let $r:=0$. Set $k:=1$.

Step 1. Apply S-adjusted reduction cuts to reduce each box $M:=[p,q]\in\mathcal{P}_k$. In particular delete any box $[p,q]$ such that $\varphi(p,r)-\|q\|^2\ge 0$. Let $\mathcal{P}'_k$ be the resulting collection of reduced boxes.

Step 2. For each $M:=[p,q]\in\mathcal{P}'_k$ compute $\mu(M)$ by solving LP(M,r). If LP(M,r) is infeasible or $\mu(M)=r$, then delete M. Let $\mathcal{P}^*_k:=\{M\in\mathcal{P}'_k|\ \mu(M)>r\}$. For every $M\in\mathcal{P}^*_k$ if a basic optimal solution of LP(M,r) can be S-adjusted to derive a feasible solution (x^M,z^M) with $z^M>r$, then use it to update CBV.

Step 3. Let $\mathcal{S}_k:=\mathcal{R}_k\cup\mathcal{P}^*_k$. Reset $r:=CBV$. Delete every $M\in\mathcal{S}_k$ such that $\mu(M)<r$ and let $\mathcal{R}_{k+1}$ be the collection of remaining boxes.

Step 4. If $\mathcal{R}_{k+1}=\emptyset$, then terminate: if $r=0$, the problem is infeasible; otherwise, r is the optimal value and the feasible solution $(\bar x,\bar z)$ with $\bar z=r$ is an optimal solution.

Step 5. If $\mathcal{R}_{k+1}\ne\emptyset$, let $M_k\in\operatorname{argmax}\{\mu(M)|\ M\in\mathcal{R}_{k+1}\}$. Divide M_k into two boxes according to the rule described in section 6. Let $\mathcal{P}_{k+1}$ be the collection of these two subboxes of M_k.

Step 6. Increment k and return to Step 1.

Theorem 18. *Algorithm* 3 *solves the problem* (DL) *in finitely many iterations.*

Proof. Although the variable z is not explicitly required to take on discrete values, this is actually a discrete variable, since the constraint (31) amounts to requiring that $z=\min_{i=1,\dots,m}(\|x-a^i\|-\alpha_i)$ for $x\in S$. The finiteness of Algorithm 3 then follows from Theorem 17. □

TABLE 1

Problems	n	m	It	Time	max$N$
1-10	10	200	81	0.683	35
10-20	10	400	118	2.197	52
21-30	20	200	99	1.655	52
31-40	20	400	183	7.383	98
41-50	30	100	178	2.097	86
51-60	30	200	208	5.449	97
61-70	40	100	199	3.155	108
71-80	40	200	298	11.442	163
81-90	40	300	619	40.442	340
91-100	40	400	526	55.911	265
101-110	40	500	504	78.847	297
111-120	50	100	344	6.759	157
121-130	50	200	618	34.077	331
131-140	50	300	711	69.859	432
141-150	50	400	851	123.864	489
151-160	50	500	1089	216.725	529
161-170	100	100	1511	63.520	852
171-180	100	200	5135	642.710	2949

TABLE 2

Problems	n	m	It		Time		max$N$	
			A	B	A	B	A	B
1-10	10	200	324	81	2.4	0.794	27	35
10-20	10	400	450	118	7.9	2.319	44	52
21-30	20	200	332	99	3.7	1.766	42	52
31-40	20	400	685	183	20.7	7.383	80	98
41-50	30	100	567	178	4.8	2.659	84	86

Algorithm 3 was coded in C++ and tested on a number of problem instances of dimension ranging from 10 to 100 (10 problems for each instance of n). Points a^j were randomly generated in the square $1000 \leq x_i \leq 90000,\ i = 1, \ldots, n$, while S was taken to be the lattice of points with integral coordinates. The program was run on a PC Pentium IV (2.53GHz with 256Mb of DDR RAM), with linear subproblems solved by the LP software CPLEX 8.0.

The computational results (with relative error ≤ 0.01 for the optimal value) are summarized in Table 1 with the following notation:

n: dimension,

m: number of given points a^j,

It: average number of iterations,

Time: average running time in seconds,

max$N$: average maximal number of active nodes of the branch and bound tree.

These results suggest that the method is quite practical for this class of discrete optimization problems. The computational cost increases much more rapidly with n (dimension of space or number of variables) than with m (number of balls). Also, when compared with the results preliminarily reported in [14], they show that the performance of the method can be drastically improved by using a more suitable monotonic reformulation. To give a more precise idea of the improvement, Table 2 reports the computational results obtained in the two versions of the method on the 50 first problems tested (the columns A, B refer to performances of the version used in [14] and the present version, respectively).

Finally, we note that, although variants of the problem (DL) have been known for years, the literature on numerical results for this problem appear to be very poor. To the best of our knowledge, prior to the paper [14], the problem was studied only for n and/or m very small (see, e.g., [16], [8], where, however, ellipsoids rather than balls were considered). A d.c. optimization method proposed in [6] for the largest empty ball problem has never been implemented. It seems that [14] was the first numerical study for the problem (DL) with fairly large values of n and m.

Acknowledgment. The authors are grateful to the referees for several helpful comments and suggestions.

REFERENCES

[1] H. Konno, P. T. Thach, and H. Tuy, *Optimization on Low Rank Nonconvex Structures*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1997.

[2] J. B. Lasserre, *Global optimization with polynomials and the problem of moments*, SIAM J. Optim., 11 (2001), pp. 796–817.

[3] M. Minoux and H. Tuy, *An Approach to Discrete Monotonic Optimization*, preprint, Institute of Mathematics, Hanoi, Vietnam, 2001.

[4] F. Plastria, *Continuous location problems*, in Facility Location, Z. Drezner and H. W. Hamacher, eds., Springer, Berlin, 1995, pp. 225–262.

[5] A. Rubinov, H. Tuy, and H. Mays, *Algorithm for a monotonic global optimization problem*, Optimization, 49 (2001), pp. 205–221.

[6] J. S. Shi and Y. Yamamoto, *A d.c. approach to the largest empty sphere problem in higher dimension*, in State of the Art in Global Optimization, C. Floudas and P. Pardalos, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996, pp. 395–411.

[7] N. Z. Shor, *Nondifferentiable Optimization and Polynomial Problems*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998.

[8] P. T. Thach, *The design centering problem as a D.C. programming problem*, Math. Programming, 41 (1988), pp. 229–248.

[9] H. Tuy, *Normal sets, polyblocks, and monotonic optimization*, Vietnam J. Math., 27 (1999), pp. 277–300.

[10] H. Tuy, *Monotonic optimization: Problems and solution approaches*, SIAM J. Optim., 11 (2000), pp. 464–494.

[11] H. Tuy, F. Al-Khayyal, and P. T. Thach, *Monotonic optimization: Branch and cut methods*, in Essays and Surveys on Global Optimization, C. Audet, P. Hansen, and G. Savard, eds., Springer, New York, 2005, pp. 39–78.

[12] H. Tuy and L. T. Luc, *A new approach to optimization under monotonic constraint*, J. Global Optim., 18 (2000), pp. 1–15.

[13] H. Tuy and N. T. Hoai Phuong, *A unified monotonic approach to generalized linear fractional programming*, J. Global Optim., 26 (2003), pp. 229–259.

[14] H. Tuy, N. D. Nghia, and L. S. Vinh, *A discrete location problem*, Acta Math. Vietnam., 28 (2003), pp. 185–199.

[15] H. Tuy, P. T. Thach, and H. Konno, *Optimization of polynomial fractional programming*, J. Global Optim., 29 (2004), pp. 19–44.

[16] L. M. Vidigal and S. W. Director, *A design centering problem algorithm for nonconvex regions of acceptability*, IEEE Trans. on Computer-aided Design of Integrated Circuits and Systems, 14 (1982), pp. 13–24.

SIAM J. OPTIM.
Vol. 17, No. 1, pp. 98–118
© 2006 Society for Industrial and Applied Mathematics

ON THE SOLUTION OF THE TIKHONOV REGULARIZATION OF THE TOTAL LEAST SQUARES PROBLEM*

AMIR BECK† AND AHARON BEN-TAL†

Abstract. *Total least squares* (TLS) is a method for treating an overdetermined system of linear equations $\mathbf{Ax} \approx \mathbf{b}$, where both the matrix $\mathbf{A}$ and the vector $\mathbf{b}$ are contaminated by noise. *Tikhonov regularization* of the TLS (TRTLS) leads to an optimization problem of minimizing the sum of fractional quadratic and quadratic functions. As such, the problem is nonconvex. We show how to reduce the problem to a single variable minimization of a function $\mathcal{G}$ over a closed interval. Computing a value and a derivative of $\mathcal{G}$ consists of solving a single trust region subproblem. For the special case of regularization with a squared Euclidean norm we show that $\mathcal{G}$ is unimodal and provide an alternative algorithm, which requires only one spectral decomposition. A numerical example is given to illustrate the effectiveness of our method.

Key words. total least squares, Tikhonov regularization, fractional programming, nonconvex optimization, trust region subproblem

AMS subject classifications. 65F20, 90C20, 90C32

DOI. 10.1137/050624418

1. Introduction.

Many problems in data fitting and estimation give rise to an overdetermined system of linear equations $\mathbf{Ax} \approx \mathbf{b}$, where both the matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ and the vector $\mathbf{b} \in \mathbb{R}^m$ are contaminated by noise. The total least squares (TLS) approach to this problem [11, 12, 19] is to seek a perturbation matrix $\mathbf{E} \in \mathbb{R}^{m\times n}$ and a perturbation vector $\mathbf{r} \in \mathbb{R}^m$ that minimize $\|\mathbf{E}\|^2 + \|\mathbf{r}\|^2$ subject to the consistency equation $(\mathbf{A}+\mathbf{E})\mathbf{x} = \mathbf{b}+\mathbf{r}$ (here and elsewhere in this paper a matrix norm is always the Frobenius norm and a vector norm is the Euclidean one). The TLS approach was extensively used in a variety of scientific disciplines such as signal processing, automatic control, statistics, physics, economic, biology, and medicine (see, e.g., [19] and the references therein). The TLS problem has essentially an explicit solution, expressed by the singular value decomposition of the augmented matrix $(\mathbf{A}, \mathbf{b})$ (see, e.g., [11, 19]).

In practical situations, the original (noise-free) linear system is often ill-conditioned. For example, this happens when the system is obtained via discretization of ill-posed problems such as integral equations of the first kind (see, e.g., [10] and the references therein). In these cases the least squares (LS) solution as well as the TLS solution can be physically meaningless, and thus regularization is essential for stabilizing the solution.

There are two well-established approaches (among many others) to stabilize the LS solution: (i) *Tikhonov regularization*, where a quadratic penalty is appended to the LS objective function [4, 33], and (ii) *regularized least squares* (abbreviated RLS and LSQI), where a quadratic constraint bounding the size of the solution is added [4, 8].

*Received by the editors February 15, 2005; accepted for publication (in revised form) December 7, 2005; published electronically April 21, 2006.
http://www.siam.org/journals/siopt/17-1/62441.html
†MINERVA Optimization Center, Department of Industrial Engineering and Management Technion, Israel Institute of Technology, Haifa 3200, Israel (becka@ie.technion.ac.il, abental@ie.technion.ac.il). The research of the first author was partially supported by ISF grant 729/04. The research of the second author was partially supported by BSF grant 2002038.

For the TLS problem the situation is different. Stabilization by introducing a quadratic constraint was extensively studied [1, 10, 14, 28, 24]. On the other hand, Tikhonov regularization of the TLS (TRTLS) problem has not yet been considered.

In this paper we adopt the Tikhonov regularization concept to stabilize the TLS solution; i.e., we consider the problem

$$\text{(TRTLS)} \quad \min_{\mathbf{E},\mathbf{r},\mathbf{x}} \left\{ \|\mathbf{E}\|^2 + \|\mathbf{r}\|^2 + \rho\|\mathbf{L}\mathbf{x}\|^2 : (\mathbf{A}+\mathbf{E})\mathbf{x} = \mathbf{b}+\mathbf{r} \right\}, \tag{1}$$

where $\mathbf{L} \in \mathbb{R}^{k\times n}, k \le n$, is a full row rank matrix and $\rho > 0$ is a penalty parameter. $\mathbf{L}$ is a matrix that defines a (semi)norm on the solution through which its "size" is measured. A common example where $\mathbf{L}$ is not square is when $\mathbf{L}$ is an approximation matrix of the first or second order derivative [10, 16, 18].

The main difficulty associated with problem (TRTLS) is its nonconvexity. Nevertheless, we show in this paper that the problem can be solved efficiently to global optimality. First, in section 2 we reduce problem (TRTLS) to one involving only the $\mathbf{x}$ variables:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \left\{ \frac{\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2}{\|\mathbf{x}\|^2+1} + \rho\|\mathbf{L}\mathbf{x}\|^2 \right\}. \tag{2}$$

In section 3 we derive an extremely mild condition for the attainability of an optimal solution to (2). An algorithm for solving problem (TRTLS) is then described in section 4. The algorithm consists of minimizing a single variable continuous (and differentiable under a mild condition) function $\mathcal{G}(\alpha)$ on a closed interval. Computing $\mathcal{G}(\alpha)$ and its derivative involves the solution of a single trust region subproblem. The interesting special case, where the matrix $\mathbf{L}$ in problem (TRTLS) is the identity matrix, is studied in section 5, where we prove that in this case $\mathcal{G}$ is unimodal and provide an alternative algorithm for solving the TRTLS problem requiring a single spectral decomposition. Finally, we provide in section 6 a detailed algorithm for the solution of the TRTLS problem (with a general regularization matrix) and demonstrate our method through an image deblurring example.

2. Simplified formulation of the TRTLS problem. In order to simplify problem (1), we use a derivation similar to the one used in [1].[1] Problem (TRTLS) can be written as a double minimization problem:

$$\min_{\mathbf{x}} \min_{\mathbf{E},\mathbf{r}} \left\{ \|\mathbf{E}\|^2 + \|\mathbf{r}\|^2 + \rho\|\mathbf{L}\mathbf{x}\|^2 : (\mathbf{A}+\mathbf{E})\mathbf{x} = \mathbf{b}+\mathbf{r} \right\}. \tag{3}$$

Consider the inner minimization problem

$$\min_{\mathbf{E},\mathbf{r}} \left\{ \|\mathbf{E}\|^2 + \|\mathbf{r}\|^2 + \rho\|\mathbf{L}\mathbf{x}\|^2 : (\mathbf{A}+\mathbf{E})\mathbf{x} = \mathbf{b}+\mathbf{r} \right\}. \tag{4}$$

The Lagrangian of problem (4) is given by

$$\mathcal{L}(\mathbf{E},\mathbf{r},\boldsymbol{\lambda}) = \|\mathbf{E}\|^2 + \|\mathbf{r}\|^2 + \rho\|\mathbf{L}\mathbf{x}\|^2 + 2\boldsymbol{\lambda}^T((\mathbf{A}+\mathbf{E})\mathbf{x} - \mathbf{b} - \mathbf{r}).$$

Note that problem (4) is a linearly constrained convex problem with respect to the variables $\mathbf{E}$ and $\mathbf{r}$. Thus, the KKT conditions are necessary and sufficient [3, Proposition 3.4.1], and we conclude that $(\mathbf{E},\mathbf{r})$ is an optimal solution of (4) if and only if

[1] We thank Marc Teboulle for his contribution to this derivation.

there exists $\boldsymbol{\lambda} \in \mathbb{R}^m$ such that

$$2\mathbf{E} + 2\boldsymbol{\lambda}\mathbf{x}^T = \mathbf{0} \quad (\nabla_{\mathbf{E}}\mathcal{L} = 0), \tag{5}$$

$$2\mathbf{r} - 2\boldsymbol{\lambda} = \mathbf{0} \qquad (\nabla_{\mathbf{r}}\mathcal{L} = 0), \tag{6}$$

$$(\mathbf{A} + \mathbf{E})\mathbf{x} = \mathbf{b} + \mathbf{r} \quad \text{(feasibility)}. \tag{7}$$

From (6) we have $\boldsymbol{\lambda} = \mathbf{r}$. Substituting this into (5) we have

$$\mathbf{E} = -\mathbf{r}\mathbf{x}^T. \tag{8}$$

Combining (8) with (7) we obtain $(\mathbf{A} - \mathbf{r}\mathbf{x}^T)\mathbf{x} = \mathbf{b} + \mathbf{r}$, so

$$\mathbf{r} = \frac{\mathbf{A}\mathbf{x} - \mathbf{b}}{\|\mathbf{x}\|^2 + 1} \tag{9}$$

and consequently

$$\mathbf{E} = -\frac{(\mathbf{A}\mathbf{x} - \mathbf{b})\mathbf{x}^T}{\|\mathbf{x}\|^2 + 1}. \tag{10}$$

Finally, by substituting (9) and (10) into the objective function of problem (4) we obtain that the value of problem (4) is equal to $\frac{\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2}{\|\mathbf{x}\|^2+1} + \rho\|\mathbf{L}\mathbf{x}\|^2$. Consequently, the TRTLS problem (1) reduces to

$$f^* = \min_{\mathbf{x}\in\mathbb{R}^n} \left\{ \mathcal{H}(\mathbf{x}) \equiv \frac{\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2}{\|\mathbf{x}\|^2 + 1} + \rho\|\mathbf{L}\mathbf{x}\|^2 \right\}. \tag{11}$$

For a given optimal solution $\mathbf{x}$ to the simplified TRTLS problem (11), the optimal pair $(\mathbf{E}, \mathbf{r})$ to the original TRTLS problem is given by (9) and (10).

3. Attainability of the minimum. In this section, we find a sufficient condition for the attainability of the minimum in (11). First, notice that if $k = n$, then $\mathbf{L}$ has full rank and as a result the objective function is a coercive function[2] and the minimum is attained (see [3]). On the other hand, if $k < n$, then the minimum in (11) might not be attained. This is illustrated by the following example.

Example. Consider problem (11) with data

$$m = 3, \quad n = 2, \quad \mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}, \ \mathbf{b} = \begin{pmatrix} 4 \\ 0 \\ 0 \end{pmatrix}, \ \mathbf{L} = \begin{pmatrix} 1 & 0 \end{pmatrix}, \ \rho = 1.$$

The TRTLS problem (11) in this case is

$$\min_{x_1,x_2} \left\{ \underbrace{\frac{(x_1-4)^2 + x_2^2}{1 + x_1^2 + x_2^2} + x_1^2}_{\mathcal{H}(x_1,x_2)} \right\}. \tag{12}$$

To show the nonattainment of the minimum, suppose on the contrary that the minimum is attained at a point (x_1^*, x_2^*). Notice that

$$(x_1^*)^2 \le \mathcal{H}(x_1^*, x_2^*) \le \mathcal{H}(0, x_2) \quad \forall x_2 \in \mathbb{R}.$$

[2] A real valued function $f : \mathbb{R}^n \to \mathbb{R}$ is coercive if $\lim_{\|\mathbf{x}\|\to\infty} f(\mathbf{x}) = \infty$.

Since $\mathcal{H}(0,x_2)=\frac{16+x_2^2}{1+x_2^2}\overset{x_2\to\infty}{\longrightarrow}1$ we conclude that $|x_1^*|\le 1$, which implies the inequality $(x_1^*-4)^2>1+(x_1^*)^2$. Therefore, the function $\varphi(y)=\mathcal{H}(x_1^*,y)=\frac{(x_1^*-4)^2+y^2}{1+(x_1^*)^2+y^2}+(x_1^*)^2$ is strictly decreasing and as a result we have, for example, $\mathcal{H}(x_1^*,x_2^*+1)<\mathcal{H}(x_1^*,x_2^*)$, which is a contradiction to the assumption that the minimum is attained at (x_1^*,x_2^*). We therefore conclude that the minimum (12) is not attained. ∎

Theorem 3.1 introduces a sufficient condition for the attainability of the minimum of the TRTLS problem (11).

THEOREM 3.1. *Consider problem* (11) *with* $\mathbf{A}\in\mathbb{R}^{m\times n},\mathbf{b}\in\mathbb{R}^m$, *and* $\mathbf{L}\in\mathbb{R}^{k\times n},n>k$. *Let* $\mathbf{F}\in\mathbb{R}^{n\times k}$ *be a matrix whose columns form an orthogonal basis for the null space of* $\mathbf{L}$. *If the following condition is satisfied,*

$$\lambda_{min}\begin{pmatrix}\mathbf{F}^T\mathbf{A}^T\mathbf{A}\mathbf{F} & \mathbf{F}^T\mathbf{A}^T\mathbf{b}\\ \mathbf{b}^T\mathbf{A}\mathbf{F} & \|\mathbf{b}\|^2\end{pmatrix}<\lambda_{min}(\mathbf{F}^T\mathbf{A}^T\mathbf{A}\mathbf{F}), \tag{13}$$

then

(i)

$$f^*\le\lambda_{min}\begin{pmatrix}\mathbf{F}^T\mathbf{A}^T\mathbf{A}\mathbf{F} & \mathbf{F}^T\mathbf{A}^T\mathbf{b}\\ \mathbf{b}^T\mathbf{A}\mathbf{F} & \|\mathbf{b}\|^2\end{pmatrix}; \tag{14}$$

(ii) *the minimum of* (11) *is attained.*

Proof. (i) Let $\mathbf{d}\in\mathbb{R}^{p+1}$ be an eigenvector corresponding to the minimum eigenvalue of the matrix

$$\mathbf{H}=\begin{pmatrix}\mathbf{F}^T\mathbf{A}^T\mathbf{A}\mathbf{F} & \mathbf{F}^T\mathbf{A}^T\mathbf{b}\\ \mathbf{b}^T\mathbf{A}\mathbf{F} & \|\mathbf{b}\|^2\end{pmatrix}.$$

Then

$$\frac{\mathbf{d}^T\mathbf{H}\mathbf{d}}{\|\mathbf{d}\|^2}=\lambda_{min}(\mathbf{H}). \tag{15}$$

d_{p+1} must be different from zero since otherwise we would have

$$\lambda_{min}(\mathbf{H})\overset{\mathbf{d}=(\tilde{\mathbf{d}}^T,0)^T}{=}\frac{\mathbf{d}^T\mathbf{H}\mathbf{d}}{\|\mathbf{d}\|^2}=\frac{\tilde{\mathbf{d}}^T\mathbf{F}^T\mathbf{A}^T\mathbf{A}\mathbf{F}\tilde{\mathbf{d}}}{\|\tilde{\mathbf{d}}\|^2}\ge\lambda_{min}(\mathbf{F}^T\mathbf{A}^T\mathbf{A}\mathbf{F}),$$

which is in contradiction to (13). Therefore, $d_{p+1}\neq 0$. Let $\mathbf{y}\in\mathbb{R}^p$ be such that $\frac{\mathbf{d}}{-d_{p+1}}=(\mathbf{y}^T,-1)^T$. Then

$$\begin{aligned}\lambda_{min}(\mathbf{H})&\overset{(15)}{=}\frac{\mathbf{d}^T\mathbf{H}\mathbf{d}}{\|\mathbf{d}\|^2}=\frac{\left(\frac{\mathbf{d}}{d_{n+1}}\right)^T\mathbf{H}\left(\frac{\mathbf{d}}{d_{n+1}}\right)}{\left\|\left(\frac{\mathbf{d}}{d_{n+1}}\right)\right\|^2}\\ &=\frac{\begin{pmatrix}\mathbf{y}^T & -1\end{pmatrix}\mathbf{H}\begin{pmatrix}\mathbf{y}\\ -1\end{pmatrix}}{\left\|\begin{pmatrix}\mathbf{y}^T & -1\end{pmatrix}\right\|^2}=\frac{\mathbf{y}^T\mathbf{F}^T\mathbf{A}^T\mathbf{A}\mathbf{F}\mathbf{y}-2\mathbf{y}^T\mathbf{F}^T\mathbf{A}^T\mathbf{b}+\|\mathbf{b}\|^2}{\|\mathbf{y}\|^2+1}\\ &\overset{\mathbf{F}^T\mathbf{F}=\mathbf{I},\mathbf{L}\mathbf{F}=0}{=}\frac{\mathbf{y}^T\mathbf{F}^T\mathbf{A}^T\mathbf{A}\mathbf{F}\mathbf{y}-2\mathbf{y}^T\mathbf{F}^T\mathbf{A}^T\mathbf{b}+\|\mathbf{b}\|^2}{\mathbf{y}^T\mathbf{F}^T\mathbf{F}\mathbf{y}+1}+\rho\|\mathbf{L}\mathbf{F}\mathbf{y}\|^2\\ &=\mathcal{H}(\mathbf{F}\mathbf{y})\ge f^*,\end{aligned}$$

thus proving (i). To prove (ii), suppose on the contrary that the minimum value of (11), f^*, is not attained, which implies that there exists a sequence $\mathbf{x}_k, k \geq 1$, such that

$$\|\mathbf{x}_k\| \to \infty, \quad \underbrace{q(\mathbf{x}_k) + h(\mathbf{x}_k)}_{\mathcal{H}(\mathbf{x}_k)} \to f^*, \tag{16}$$

where $q(\mathbf{x}_k) \equiv \frac{\|\mathbf{A}\mathbf{x}_k - \mathbf{b}\|^2}{\|\mathbf{x}_k\|^2+1}$ and $h(\mathbf{x}_k) \equiv \rho\|\mathbf{L}\mathbf{x}_k\|^2$. Since both the sequences $q(\mathbf{x}_k)$ and $\frac{\mathbf{x}_k}{\|\mathbf{x}_k\|}$ are bounded, there exists a subsequence $\mathbf{x}_{n_k}$ for which the subsequences $q(\mathbf{x}_{n_k})$ and $\frac{\mathbf{x}_{n_k}}{\|\mathbf{x}_{n_k}\|}$ converge to a finite value. That is, there exist η and $\mathbf{d}$ such that

$$q(\mathbf{x}_{n_k}) \to \eta, \quad \frac{\mathbf{x}_{n_k}}{\|\mathbf{x}_{n_k}\|} \to \mathbf{d}.$$

Now, from (16) it follows that

$$\frac{q(\mathbf{x}_{n_k}) + h(\mathbf{x}_{n_k})}{\|\mathbf{x}_{n_k}\|^2} \to 0$$

and since $q(\mathbf{x}_{n_k})$ is bounded we have that $\frac{h(\mathbf{x}_{n_k})}{\|\mathbf{x}_{n_k}\|^2} \to 0$. But, on the other hand, $\frac{h(\mathbf{x}_{n_k})}{\|\mathbf{x}_{n_k}\|^2} \to \rho\|\mathbf{L}\mathbf{d}\|^2$ and as a result we have that $\|\mathbf{L}\mathbf{d}\|^2 = 0$, which is equivalent to $\mathbf{d} \in \text{Null}(\mathbf{L})$. To summarize, we have found a subsequence $\mathbf{x}_{n_k}$ for which $q(\mathbf{x}_{n_k})$ converges and $\frac{\mathbf{x}_{n_k}}{\|\mathbf{x}_{n_k}\|} \to \mathbf{d}$, where $\mathbf{d} \in \text{Null}(\mathbf{L})$ and $\|\mathbf{d}\| = 1$. Now,

$$\begin{aligned}
f^* &= \lim_{k\to\infty} \{q(\mathbf{x}_{n_k}) + h(\mathbf{x}_{n_k})\} \\
&\overset{h(\cdot)\geq 0}{\geq} \lim_{k\to\infty} q(\mathbf{x}_{n_k}) = \lim_{k\to\infty} \frac{\|\mathbf{A}\mathbf{x}_{n_k} - \mathbf{b}\|^2}{\|\mathbf{x}_{n_k}\|^2 + 1} \\
&= \lim_{k\to\infty} \frac{\mathbf{x}_{n_k}^T\mathbf{A}^T\mathbf{A}\mathbf{x}_{n_k} - 2\mathbf{b}^T\mathbf{A}\mathbf{x}_{n_k} + \|\mathbf{b}\|^2}{\|\mathbf{x}_{n_k}\|^2 + 1} \\
&= \lim_{k\to\infty} \frac{\left(\frac{\mathbf{x}_{n_k}}{\|\mathbf{x}_{n_k}\|}\right)^T \mathbf{A}^T\mathbf{A}\left(\frac{\mathbf{x}_{n_k}}{\|\mathbf{x}_{n_k}\|}\right) - 2\frac{1}{\|\mathbf{x}_{n_k}\|}\mathbf{b}^T\mathbf{A}\left(\frac{\mathbf{x}_{n_k}}{\|\mathbf{x}_{n_k}\|}\right) + \frac{\|\mathbf{b}\|^2}{\|\mathbf{x}_{n_k}\|^2}}{1 + \frac{1}{\|\mathbf{x}_{n_k}\|^2}} \\
&= \mathbf{d}^T\mathbf{A}^T\mathbf{A}\mathbf{d}.
\end{aligned}$$

Since $\mathbf{d} \in \text{Null}(\mathbf{L})$ we can write $\mathbf{d} = \mathbf{F}\mathbf{v}$, and therefore we obtain the following lower bound on f^*:

$$f^* \geq \min_{\mathbf{v}^T\mathbf{F}^T\mathbf{F}\mathbf{v}=1} \mathbf{v}^T\mathbf{F}^T\mathbf{A}^T\mathbf{A}\mathbf{F}\mathbf{v} \overset{\mathbf{F}^T\mathbf{F}=\mathbf{I}}{=} \min_{\|\mathbf{v}\|^2=1} \mathbf{v}^T\mathbf{F}^T\mathbf{A}^T\mathbf{A}\mathbf{F}\mathbf{v} = \lambda_{min}(\mathbf{F}^T\mathbf{A}^T\mathbf{A}\mathbf{F}).$$

On the other hand, by condition (13), $\lambda_{min}(\mathbf{F}^T\mathbf{A}^T\mathbf{A}\mathbf{F}^T) > \lambda_{min}(\mathbf{H})$, and therefore we have that

$$f^* > \lambda_{\min}(\mathbf{H}),$$

which is a contradiction to part (i). □

Remarks.

1. Weak inequality is always satisfied in (13): the matrix in the right-hand side of (13) is a principal submatrix of the one in the left-hand side. Hence, by the interlacing theorem of eigenvalues [34, Theorem 7.8], weak inequality holds.
2. Condition (13) is invariant to the specific choice of the orthogonal basis of the null space of $\mathbf{L}$.
3. For $\mathbf{L} = \mathbf{0}$ problem (11) reduces to the classical TLS problem. In this case we can take $\mathbf{F} = \mathbf{I}$ in condition (13), which then reduces to the well-known condition [11, 19] for the attainability of the minimum in the TLS problem:

$$\lambda_{min}\begin{pmatrix} \mathbf{A}^T\mathbf{A} & \mathbf{A}^T\mathbf{b} \\ \mathbf{b}^T\mathbf{A} & \|\mathbf{b}\|^2 \end{pmatrix} < \lambda_{min}\left(\mathbf{A}^T\mathbf{A}\right). \tag{17}$$

   Incidentally, for the nonregularized version of problem (12), i.e.,

$$\min_{x_1,x_2}\left\{\frac{(x_1-4)^2+x_2^2}{1+x_1^2+x_2^2}\right\}, \tag{18}$$

   condition (17) does hold since

$$\lambda_{min}\begin{pmatrix} \mathbf{A}^T\mathbf{A} & \mathbf{A}^T\mathbf{b} \\ \mathbf{b}^T\mathbf{A} & \|\mathbf{b}\|^2 \end{pmatrix} = 0 < 1 = \lambda_{min}\left(\mathbf{A}^T\mathbf{A}\right)$$

   and indeed (18) attains an optimal solution $x_1^* = 4, x_2^* = 0$.
4. The TRTLS problem (12), for which nonattainability of the minimum was established, indeed does not satisfy condition (13). $\mathbf{F}$ can be chosen to be $\binom{0}{1}$, and we have

$$\lambda_{min}(\mathbf{F}^T\mathbf{A}^T\mathbf{A}\mathbf{F}) = 1$$

   and

$$\lambda_{min}\begin{pmatrix} \mathbf{F}^T\mathbf{A}^T\mathbf{A}\mathbf{F} & \mathbf{F}^T\mathbf{A}^T\mathbf{b} \\ \mathbf{b}^T\mathbf{A}\mathbf{F} & \|\mathbf{b}\|^2 \end{pmatrix} = \lambda_{min}\begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix} = 1.$$

4. Solving the TRTLS problem with general L. In this section we consider the TRTLS problem (11) with a full row rank $k \times n$ regularization matrix $\mathbf{L}$. We will assume that condition (13) is satisfied, and therefore the minimum is attained.

Problem (11) can be formulated as a double minimization problem in the following way:

$$\min_{\alpha\geq 1}\min_{\|\mathbf{x}\|^2=\alpha-1}\left\{\frac{\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2}{\alpha}+\rho\|\mathbf{L}\mathbf{x}\|^2\right\},$$

which can be written as

$$\min_{\alpha\geq 1}\{\mathcal{G}(\alpha)\}, \tag{19}$$

where

$$\mathcal{G}(\alpha) \equiv \min_{\|\mathbf{x}\|^2=\alpha-1}\left\{\frac{\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2}{\alpha}+\rho\|\mathbf{L}\mathbf{x}\|^2\right\}. \tag{20}$$

Calculating function values of $\mathcal{G}$ requires solving a minimization problem with a quadratic objective function and a norm equality constraint. In section 4.1 we briefly review known results on this problem including necessary and sufficient optimality conditions. In section 4.2 continuity and differentiability of $\mathcal{G}$ are established under standard second order sufficiency conditions. In section 4.3 an upper bound $\bar{\alpha}$ on the value of the optimal α is derived. Thus, the TRTLS problem (11) is reduced to a one dimensional minimization of $\mathcal{G}$ over a finite interval $[1, \bar{\alpha}]$.

4.1. Minimization of a quadratic function subject to a norm equality constraint. In this section we consider the minimization problem

$$\min_{\|\mathbf{x}\|^2=\beta} \left\{ \mathbf{x}^T \mathbf{Q} \mathbf{x} - 2\mathbf{f}^T \mathbf{x} + c \right\}. \tag{21}$$

We do not assume that $\mathbf{Q}$ is positive semidefinite, and therefore the objective function need not be convex. Problem (21) is the well-known *trust region subproblem* (TRS); it has been extensively studied from both theoretical and algorithmic aspects [2, 5, 7, 23, 27, 31].[3] Necessary and sufficient conditions for a (global) solution of (21) are well established [5, 7, 32].

THEOREM 4.1 (see [5, 7, 32]). *Consider problem* (21) *with a symmetric matrix* $\mathbf{Q} \in \mathbb{R}^{n \times n}, \mathbf{f} \in \mathbb{R}^n, c \in \mathbb{R}, \beta \in \mathbb{R}^+$. *Then* $\mathbf{x}^*$ *is an optimal solution of* (21) *if and only if there exists* $\lambda^* \in \mathbb{R}$ *such that*

$$(\mathbf{Q} - \lambda^* \mathbf{I})\mathbf{x}^* = \mathbf{f}, \tag{22}$$

$$\|\mathbf{x}^*\|^2 = \beta, \tag{23}$$

$$\mathbf{Q} - \lambda^* \mathbf{I} \succeq \mathbf{0}. \tag{24}$$

Moreover, if $\mathbf{f} \notin Null(\mathbf{Q} - \lambda_{\min}(\mathbf{Q})\mathbf{I})^{\perp}$, *then the solution of problem* (21) *is unique.*

Many algorithms have been suggested to solve the TRS. A solution based on the complete spectral decomposition can be found in [8]. For medium and large-scale problems the latter approach is not applicable. Thus, several methods have been devised for these scenarios [5, 7, 13, 23, 25, 30, 29].

4.2. Continuity and differentiability of $\mathcal{G}$.

4.2.1. Continuity. The continuity of $\mathcal{G}(\alpha)$ for $\alpha > 1$ follows from a theorem by Gauvin and Dubeau [9, Theorem 3.3] . The notation in [9] is quite different from the notation in this paper, and therefore we will present the three sufficient conditions for continuity of $\mathcal{G}$ at a point $\bar{\alpha}$ from [9] in our terminology (the quotation from [9] is in italic).

1. *The feasible set* $\{\mathbf{x} : \|\mathbf{x}\|^2 = \bar{\alpha} - 1\}$ *is nonempty.* This condition is naturally satisfied for $\bar{\alpha} > 1$.
2. *There exists* $\epsilon > 0$ *such that* $\bigcup_{|\alpha - \bar{\alpha}| < \epsilon} \{\mathbf{x} : \|\mathbf{x}\|^2 = \alpha - 1\}$ *is compact.* This is also true in our problem since the union is equal to $\{\mathbf{x} : \bar{\alpha} - 1 - \epsilon \leq \|\mathbf{x}\|^2 \leq \bar{\alpha} - 1 + \epsilon\}$, which is obviously compact.
3. *The Mangasarian–Fromovitz regularity conditions are satisfied (see* [22]*).* In our problem, this means that the gradient of the constraint is different from zero at the optimal solution, i.e., $\mathbf{x}^* \neq 0$. This is true for $\bar{\alpha} > 1$ since $\|\mathbf{x}^*\|^2 = \bar{\alpha} - 1$.

[3] The TRS is usually considered with an inequality constraint $\|\mathbf{x}\|^2 \leq \beta$ instead of an equality one; however, all known results can be trivially converted to the equality case.

What is left to prove is that $\mathcal{G}$ is continuous at $\alpha = 1$ (from the right). This is proved next.

LEMMA 4.1. *$\mathcal{G}$ is continuous at $\alpha = 1$ from the right.*

Proof. First, $\mathcal{G}(1) = \|\mathbf{b}\|^2$. Now, for every $\alpha > 1$ let $\mathbf{x}_\alpha$ be such that $\mathcal{H}(\mathbf{x}_\alpha) = \mathcal{G}(\alpha)$ and $\|\mathbf{x}_\alpha\|^2 = \alpha - 1$. Then

$$\begin{aligned}
|\mathcal{G}(\alpha) - \mathcal{G}(1)| &= |\mathcal{H}(\mathbf{x}_\alpha) - \|\mathbf{b}\|^2| \\
&= \left| \frac{\|\mathbf{A}\mathbf{x}_\alpha - \mathbf{b}\|^2}{\alpha} + \rho\|\mathbf{L}\mathbf{x}_\alpha\|^2 - \|\mathbf{b}\|^2 \right| \\
&= \left| \left(\frac{1}{\alpha} - 1\right)\|\mathbf{b}\|^2 + \frac{\mathbf{x}_\alpha^T\mathbf{A}^T\mathbf{A}\mathbf{x}_\alpha - 2\mathbf{b}^T\mathbf{A}\mathbf{x}_\alpha}{\alpha} + \rho\mathbf{x}_\alpha^T\mathbf{L}^T\mathbf{L}\mathbf{x}_\alpha \right| \\
&\leq \left(1 - \frac{1}{\alpha}\right)\|\mathbf{b}\|^2 + \left(\frac{\lambda_{max}(\mathbf{A}^T\mathbf{A})}{\alpha} + \rho\lambda_{max}(\mathbf{L}^T\mathbf{L})\right)\|\mathbf{x}_\alpha\|^2 \\
&\quad + 2\frac{\|\mathbf{A}^T\mathbf{b}\|}{\alpha}\|\mathbf{x}_\alpha\| \\
&\overset{\|\mathbf{x}_\alpha\|^2 = \alpha - 1}{=} \left(1 - \frac{1}{\alpha}\right)\|\mathbf{b}\|^2 + \left(\frac{\lambda_{max}(\mathbf{A}^T\mathbf{A})}{\alpha} + \rho\lambda_{max}(\mathbf{L}^T\mathbf{L})\right)(\alpha - 1) \\
&\quad + 2\frac{\|\mathbf{A}^T\mathbf{b}\|}{\alpha}\sqrt{\alpha - 1} \\
&\overset{\alpha \to 1^+}{\longrightarrow} 0.
\end{aligned}$$

Therefore, $\lim_{\alpha \to 1^+} \mathcal{G}(\alpha) = \mathcal{G}(1)$. ∎

COROLLARY 4.1. *$\mathcal{G}$ is continuous over $[1, \infty)$.*

4.2.2. Differentiability. The function $\mathcal{G}$ is of the general form

$$\mathcal{G}(\alpha) = \min_{g(\mathbf{x}) = \alpha - 1} f(\mathbf{x}, \alpha), \tag{25}$$

where

$$f(\mathbf{x}, \alpha) \equiv \mathbf{x}^T\mathbf{Q}_\alpha\mathbf{x} - 2\mathbf{f}_\alpha^T\mathbf{x} + c_\alpha$$

and

$$g(\mathbf{x}) = \|\mathbf{x}\|^2,\ \mathbf{Q}_\alpha = \frac{1}{\alpha}\mathbf{A}^T\mathbf{A} + \rho\mathbf{L}^T\mathbf{L},\ \mathbf{f}_\alpha = \frac{1}{\alpha}\mathbf{A}^T\mathbf{b},\ c_\alpha = \frac{1}{\alpha}\|\mathbf{b}\|^2. \tag{26}$$

The single variable function $\mathcal{G}$ is not necessarily differentiable. In this subsection we show that under a suitable condition, $\mathcal{G}$ is differentiable of any order.

Our argument is the same as the one used in the sensitivity analysis of minimization problems (see, e.g., [3, 26] and the references therein). Theorem 4.2 establishes the differentiability of $\mathcal{G}$ under a suitable regularity condition.

THEOREM 4.2. *For every $\alpha > 1$ that satisfies the condition*

$$\mathbf{f}_\alpha \notin \mathit{Null}(\mathbf{Q}_\alpha - \lambda_{min}(\mathbf{Q}_\alpha)\mathbf{I})^\perp, \tag{27}$$

$\mathcal{G}(\alpha)$ is differentiable of any order and its first derivative is given by

$$\mathcal{G}'(\alpha) = \lambda(\alpha) + f'_\alpha(\mathbf{x}(\alpha), \alpha) = \lambda(\alpha) - \frac{\|\mathbf{A}\mathbf{x}(\alpha) - \mathbf{b}\|^2}{\alpha^2}, \tag{28}$$

where $\mathbf{x}(\alpha)$ *and* $\lambda(\alpha)$ *are the unique solutions of the KKT conditions* (22) *and* (23).

Proof. Let $\alpha > 1$ be such that condition (27) is satisfied. By Theorem 4.1, condition (27) implies the uniqueness of the solution of the minimization problem (25). Consider the system of equations

$$(\mathbf{Q}_\alpha - \lambda \mathbf{I})\mathbf{x} = \mathbf{f}_\alpha, \tag{29}$$

$$\|\mathbf{x}\|^2 = \alpha - 1. \tag{30}$$

By Theorem 4.1, $x(\alpha)$ and $\lambda(\alpha)$ are the solutions of the system for the given α. The Jacobian matrix associated with the system of equations (29) and (30) with respect to $(\mathbf{x}, \lambda)$ at $(\mathbf{x}(\alpha), \lambda(\alpha))$ is given by

$$J = \begin{pmatrix} \mathbf{Q}_\alpha - \lambda(\alpha)\mathbf{I} & \mathbf{x}(\alpha) \\ \mathbf{x}(\alpha)^T & 0 \end{pmatrix}.$$

To show that J is nonsingular note first that condition (27) implies also that

$$\mathbf{Q}_\alpha - \lambda(\alpha)\mathbf{I} \succ \mathbf{0}. \tag{31}$$

This is true since (29) implies that $\mathbf{f}_\alpha \in \text{Range}(\mathbf{Q}_\alpha - \lambda(\alpha)\mathbf{I}) = \text{Null}(\mathbf{Q}_\alpha - \lambda(\alpha)\mathbf{I})^\perp$. This condition combined with (27) and (24) implies that $\lambda(\alpha) < \lambda_{min}(\mathbf{Q}_\alpha)$. To show the nonsingularity of J, we will prove that the only solution of the system

$$J \begin{pmatrix} \mathbf{w} \\ t \end{pmatrix} = \mathbf{0}, \quad \mathbf{w} \in \mathbb{R}^n,\ t \in \mathbb{R},$$

is the trivial solution. Indeed, the system can be written explicitly as

$$(\mathbf{Q}_\alpha - \lambda(\alpha)\mathbf{I})\mathbf{w} + 2t\mathbf{x}(\alpha) = \mathbf{0}, \tag{32}$$

$$2\mathbf{x}(\alpha)^T\mathbf{w} = \mathbf{0}. \tag{33}$$

Multiplying (32) by $\mathbf{w}^T$ from the left and using (33), we obtain $\mathbf{w}^T(\mathbf{Q}_\alpha - \lambda(\alpha)\mathbf{I})\mathbf{w} = 0$. Since $\mathbf{Q}_\alpha - \lambda(\alpha)\mathbf{I} \succ \mathbf{0}$ we conclude that $\mathbf{w} = 0$. Substituting this into (32) we have $t = 0$, proving the nonsingularity of J. Invoking the implicit function theorem, the differentiability of any order of $\mathbf{x}(\alpha)$ and $\lambda(\alpha)$ in a neighborhood of α follows. Now $\mathbf{x}(\alpha)$ and $\lambda(\alpha)$ satisfy the identities (in α)

$$f'_{\mathbf{x}}(\mathbf{x}(\alpha), \alpha) - \lambda(\alpha) g'_{\mathbf{x}}(\mathbf{x}(\alpha)) = 0, \tag{34}$$

$$g(\mathbf{x}(\alpha)) = \alpha - 1. \tag{35}$$

Differentiating both sides of (35) yields the equation

$$\dot{\mathbf{x}}(\alpha)^T g'_{\mathbf{x}}(\mathbf{x}(\alpha)) = 1. \tag{36}$$

Multiplying (34) from the left by $\dot{\mathbf{x}}(\alpha)^T$ we obtain

$$\dot{\mathbf{x}}(\alpha)^T f'_{\mathbf{x}}(\mathbf{x}(\alpha), \alpha) - \lambda(\alpha)\dot{\mathbf{x}}(\alpha)^T g'_{\mathbf{x}}(\mathbf{x}(\alpha)) = 0. \tag{37}$$

By substituting (36) into (37) we obtain

$$\dot{\mathbf{x}}(\alpha)^T f'_{\mathbf{x}}(\mathbf{x}(\alpha), \alpha) = \lambda(\alpha). \tag{38}$$

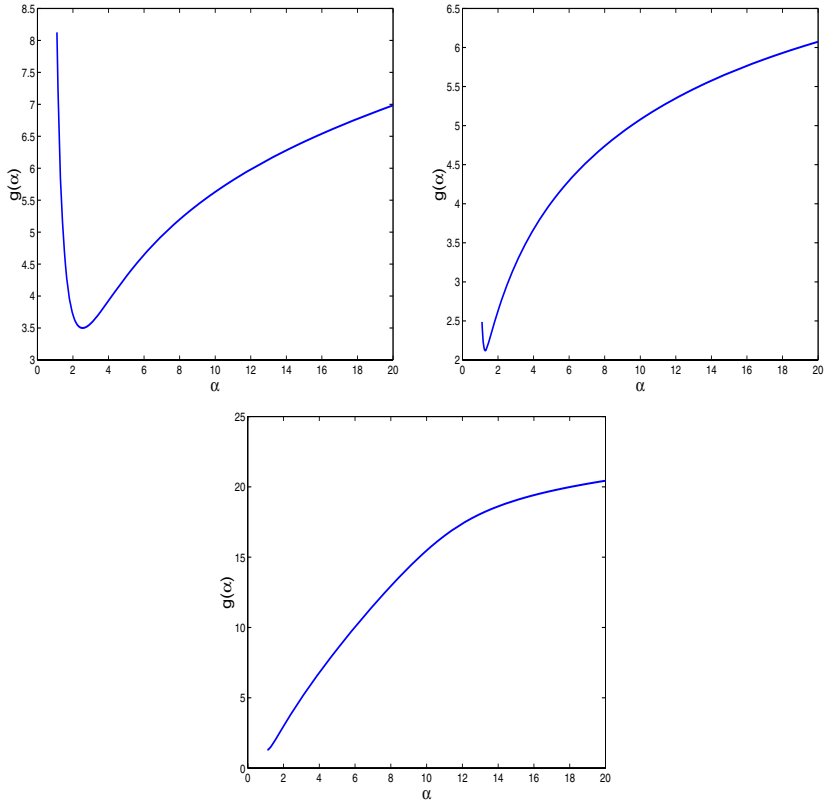


FIG. 1. *Examples of* $\mathcal{G}(\alpha)$.

$\mathcal{G}(\alpha)$ and its derivatives are given by

$$\begin{aligned} \mathcal{G}(\alpha) &= f(\mathbf{x}(\alpha), \alpha), \\ \mathcal{G}'(\alpha) &= \dot{\mathbf{x}}(\alpha)^T f'_{\mathbf{x}}(\mathbf{x}(\alpha), \alpha) + f'_{\alpha}(\mathbf{x}(\alpha), \alpha). \end{aligned} \tag{39}$$

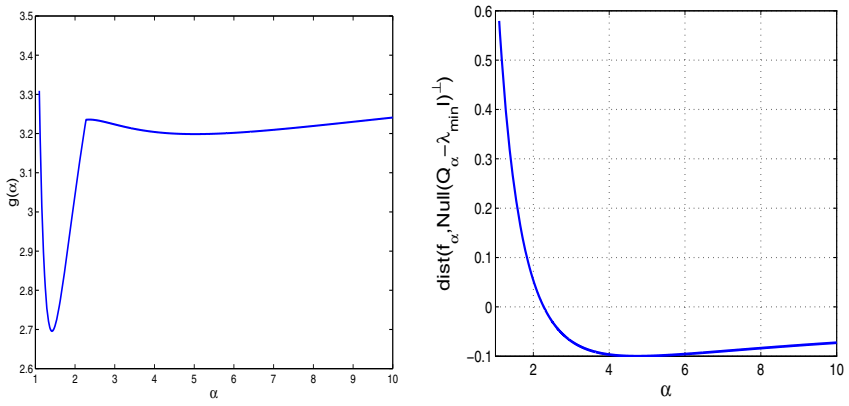
Substituting (38) into (39), the expression for the derivative (28) follows. ∎

Example. Some examples of $\mathcal{G}(\alpha)$ are given in Figure 1. These examples were randomly generated with dimensions $m = n = 4$ and $k = 3$.

In all of these examples $\mathcal{G}$ is continuous and differentiable. Note that in most examples the function $\mathcal{G}$ seems to be "well behaved" in the sense that it is strictly unimodal. A "bad" example is given in Figure 2(a), where we see an example of a nondifferentiable function. The point of nondifferentiability is $\bar{\alpha} = 2.275$. Figure 2(b) plots the quantity $dist(\mathbf{f}_\alpha, \text{Null}(\mathbf{Q}_\alpha - \lambda_{min}(\mathbf{Q}_\alpha)\mathbf{I})^\perp)$ versus α. It can be readily seen that the point in which the distance is zero is exactly the point $\bar{\alpha}$.

So far we have shown how to reduce the TRTLS problem (11) to a one dimensional problem $\min_{\alpha \geq 1} \mathcal{G}(\alpha)$. One of the problems frequently arising in one dimensional (line search) methods is determining an initial interval of search in which the optimum is known to reside. At this point, we have only shown that a lower bound on α is 1. Next we derive an upper bound.

4.3. Upper bound on the norm of optimal solutions. Let $\mathbf{x}^*$ be an optimal solution of problem (11). In this section we find an upper bound for $\|\mathbf{x}^*\|$. We recall



(a) (b)

FIG. 2. *An example of a nondifferentiable* $\mathcal{G}(\alpha)$.

the assumption that $\mathbf{L}$ is full row rank. In the case where $k = n$, it is very easy to bound the $\|\mathbf{x}^*\|$, as can be seen from the following lemma.

LEMMA 4.2. *Suppose that* $k = n$, *and let* $\mathbf{x}^*$ *be an optimal solution of* $\min_{\mathbf{x}\in\mathbb{R}^n} \mathcal{H}(\mathbf{x})$. *Then* $\|\mathbf{x}^*\|^2 \leq \frac{\|\mathbf{b}\|^2}{\rho\lambda_{min}(\mathbf{L}\mathbf{L}^T)}$.

Proof. First, notice that $\lambda_{min}(\mathbf{L}\mathbf{L}^T) > 0$ since $\mathbf{L}$ has full row rank. Now,

$$\rho\|\mathbf{L}\mathbf{x}^*\|^2 \leq \mathcal{H}(\mathbf{x}^*) \leq \mathcal{H}(0) = \|\mathbf{b}\|^2,$$

and the result follows from the simple observation that $\|\mathbf{L}\mathbf{x}^*\|^2 = (\mathbf{x}^*)^T\mathbf{L}^T\mathbf{L}\mathbf{x}^* \geq \lambda_{min}(\mathbf{L}\mathbf{L}^T)\|\mathbf{x}^*\|^2 > 0$. □

The case in which $k < n$ is much harder. In this case, we assume that condition (13) is satisfied.

THEOREM 4.3. *Suppose that condition* (13) *is satisfied, and let* $\mathbf{x}^*$ *be an optimal solution of* $\min_{\mathbf{x}\in\mathbb{R}^n} \mathcal{H}(\mathbf{x})$. *Then*

$$\|\mathbf{x}^*\|^2 \leq \max\left\{1, \frac{\|\mathbf{b}\|^2 + \lambda_{max}(\mathbf{A}^T\mathbf{A})(\delta + 2\sqrt{\delta}) + \|\mathbf{A}^T\mathbf{b}\|(\delta + 2\sqrt{\delta}) + l_1(1+\delta))}{l_1 - l_2}\right\}^2 + \delta, \tag{40}$$

where

$$l_2 = \lambda_{min}\begin{pmatrix} \mathbf{F}^T\mathbf{A}^T\mathbf{A}\mathbf{F} & \mathbf{F}^T\mathbf{A}^T\mathbf{b} \\ \mathbf{b}^T\mathbf{A}\mathbf{F} & \|\mathbf{b}\|^2 \end{pmatrix},$$
$$l_1 = \lambda_{min}(\mathbf{F}^T\mathbf{A}^T\mathbf{A}\mathbf{F}),$$

$\delta = \frac{l_2}{\lambda_{min}(\mathbf{L}\mathbf{L}^T)\rho}$, *and* $\mathbf{F}$ *is a matrix whose columns form an orthogonal base of Null*$(\mathbf{L})$.

Proof. Consider the decomposition

$$\mathbf{x}^* = \mathbf{F}\mathbf{v} + \mathbf{L}^T\mathbf{w}, \tag{41}$$

where $\mathbf{v} \in \mathbb{R}^{n-k}$ and $\mathbf{w} \in \mathbb{R}^n$ (such decomposition is possible since $\text{Null}(\mathbf{L}) = (\text{Range}(\mathbf{L}^T))^\perp$). Now,

$$\|\mathbf{x}^*\|^2 = \|\mathbf{v}\|^2 + \mathbf{w}^T\mathbf{L}\mathbf{L}^T\mathbf{w}. \tag{42}$$

By (14),

$$\mathcal{H}(\mathbf{x}^*) = f^* \leq l_2.$$

As a result,

$$\rho\|\mathbf{L}\mathbf{x}^*\|^2 \leq l_2. \tag{43}$$

Substituting (41) into (43) we obtain

$$\rho\mathbf{w}^T(\mathbf{L}\mathbf{L}^T)^2\mathbf{w} \leq l_2,$$

which implies the following inequality:

$$\mathbf{w}^T\mathbf{L}\mathbf{L}^T\mathbf{w} = \mathbf{w}^T(\mathbf{L}\mathbf{L}^T)^2\mathbf{w}\frac{\mathbf{w}^T\mathbf{L}\mathbf{L}^T\mathbf{w}}{\mathbf{w}^T(\mathbf{L}\mathbf{L}^T)^2\mathbf{w}} \leq \frac{l_2}{\rho}\lambda_{max}((\mathbf{L}\mathbf{L}^T)^{-1}(\mathbf{L}\mathbf{L}^T)(\mathbf{L}\mathbf{L}^T)^{-1}) = \delta. \tag{44}$$

We assume for now that $\|\mathbf{v}\| \geq 1$. Substituting the decomposition (41) into the objective function $\mathcal{H}$ we have

$$\begin{aligned}
&\mathcal{H}(\mathbf{x}^*)\\
&= \frac{\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|^2}{\|\mathbf{x}^*\|^2 + 1} + \rho\|\mathbf{L}\mathbf{x}^*\|^2\\
&\geq \frac{\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|^2}{\|\mathbf{x}^*\|^2 + 1} = \frac{\|\mathbf{A}(\mathbf{F}\mathbf{v} + \mathbf{L}^T\mathbf{w}) - \mathbf{b}\|^2}{\|\mathbf{F}\mathbf{v} + \mathbf{L}^T\mathbf{w}\|^2 + 1}\\
&= \frac{\mathbf{v}^T\mathbf{F}^T\mathbf{A}^T\mathbf{A}\mathbf{F}\mathbf{v} + 2\mathbf{v}^T\mathbf{F}^T\mathbf{A}^T\mathbf{A}\mathbf{L}^T\mathbf{w} - 2\mathbf{v}^T\mathbf{F}^T\mathbf{A}^T\mathbf{b} + \mathbf{w}^T\mathbf{L}\mathbf{A}^T\mathbf{A}\mathbf{L}^T\mathbf{w} - 2\mathbf{w}^T\mathbf{L}\mathbf{A}^T\mathbf{b} + \|\mathbf{b}\|^2}{1 + \|\mathbf{v}\|^2 + \mathbf{w}^T\mathbf{L}\mathbf{L}^T\mathbf{w}}\\
&= \frac{\frac{\mathbf{v}^T\mathbf{F}^T\mathbf{A}^T\mathbf{A}\mathbf{F}\mathbf{v}}{\|\mathbf{v}\|^2} + \beta}{1 + \gamma} \geq \frac{l_1 + \beta}{1 + \gamma},
\end{aligned}$$

where

$$\gamma = \frac{1 + \mathbf{w}^T\mathbf{L}\mathbf{L}^T\mathbf{w}}{\|\mathbf{v}\|^2},$$

$$\beta = \frac{2\mathbf{v}^T\mathbf{F}^T\mathbf{A}^T\mathbf{A}\mathbf{L}^T\mathbf{w} - 2\mathbf{v}^T\mathbf{F}^T\mathbf{A}^T\mathbf{b} + \mathbf{w}^T\mathbf{L}\mathbf{A}^T\mathbf{A}\mathbf{L}^T\mathbf{w} - 2\mathbf{w}^T\mathbf{L}\mathbf{A}^T\mathbf{b} + \|\mathbf{b}\|^2}{\|\mathbf{v}\|^2}.$$

We have thus proven that $\mathcal{H}(\mathbf{x}^*) \geq \theta$, where $\theta = \frac{l_1+\beta}{1+\gamma}$. Combining this with Theorem 3.1 and condition (13) we have $\theta \leq l_2 < l_1$. Now,

$$l_1 - l_2 \leq l_1 - \theta = |\theta - l_1| = \left|\frac{l_1 + \beta}{1 + \gamma} - l_1\right| = \left|\frac{\beta - l_1\gamma}{1 + \gamma}\right| \leq \beta + l_1\gamma. \tag{45}$$

Also,

(46)

$$\begin{aligned}
\gamma &\leq \frac{1 + \delta}{\|\mathbf{v}\|^2} \overset{\|\mathbf{v}\|\geq 1}{\leq} \frac{1 + \delta}{\|\mathbf{v}\|},\\
\beta &\leq \frac{2|\mathbf{v}^T\mathbf{F}^T\mathbf{A}^T\mathbf{A}\mathbf{L}^T\mathbf{w}| + 2|\mathbf{v}^T\mathbf{F}^T\mathbf{A}^T\mathbf{b}| + |\mathbf{w}^T\mathbf{L}\mathbf{A}^T\mathbf{A}\mathbf{L}^T\mathbf{w}| + 2|\mathbf{w}^T\mathbf{L}\mathbf{A}^T\mathbf{b}| + \|\mathbf{b}\|^2}{\|\mathbf{v}\|^2}\\
&\overset{(*)}{\leq} \frac{2}{\|\mathbf{v}\|}\left(\lambda_{max}(\mathbf{A}^T\mathbf{A})\sqrt{\delta} + \|\mathbf{A}^T\mathbf{b}\|\right) + \frac{1}{\|\mathbf{v}\|^2}\left(\|\mathbf{b}\|^2 + \lambda_{max}(\mathbf{A}^T\mathbf{A})\delta + 2\|\mathbf{A}^T\mathbf{b}\|\sqrt{\delta}\right)
\end{aligned}$$

$$\begin{aligned}
&\overset{\|\mathbf{v}\|\geq 1}{\leq} \frac{2}{\|\mathbf{v}\|}\left(\lambda_{max}(\mathbf{A}^T\mathbf{A})\sqrt{\delta}+\|\mathbf{A}^T\mathbf{b}\|\right)+\frac{1}{\|\mathbf{v}\|}\left(\|\mathbf{b}\|^2+\lambda_{max}(\mathbf{A}^T\mathbf{A})\delta+2\|\mathbf{A}^T\mathbf{b}\|\sqrt{\delta}\right)\\
&= \frac{1}{\|\mathbf{v}\|}\left(\|\mathbf{b}\|^2+\lambda_{max}(\mathbf{A}^T\mathbf{A})(\delta+2\sqrt{\delta})+\|\mathbf{A}^T\mathbf{b}\|(2+2\sqrt{\delta})\right),
\end{aligned} \tag{47}$$

where inequality (*) is true due to the Cauchy–Schwarz inequality and trivial linear algebra inequalities. For example, $|\mathbf{v}^T\mathbf{F}^T\mathbf{A}^T\mathbf{A}\mathbf{L}^T\mathbf{w}|$ is bounded as follows:

$$\begin{aligned}
|\mathbf{v}^T\mathbf{F}^T\mathbf{A}^T\mathbf{A}\mathbf{L}^T\mathbf{w}| &\overset{\text{C-S}}{\leq} \|\mathbf{F}\mathbf{v}\|\cdot\|\mathbf{A}^T\mathbf{A}\mathbf{L}^T\mathbf{w}\| \overset{\lambda_{max}(\mathbf{F})\leq 1}{\leq} \|\mathbf{v}\|\lambda_{max}(\mathbf{A}^T\mathbf{A})\|\mathbf{L}^T\mathbf{w}\|\\
&\overset{(44)}{\leq} \lambda_{max}(\mathbf{A}^T\mathbf{A})\|\mathbf{v}\|\sqrt{\delta}.
\end{aligned}$$

Using the upper bound on β (47) and the upper bound on γ (46), we conclude that if $\|\mathbf{v}\|\geq 1$, then

$$\begin{aligned}
l_1-l_2 &\overset{(45)}{\leq} \beta+l_1\gamma\\
&\leq \frac{1}{\|\mathbf{v}\|}\left(\|\mathbf{b}\|^2+\lambda_{max}(\mathbf{A}^T\mathbf{A})(\delta+2\sqrt{\delta})+\|\mathbf{A}^T\mathbf{b}\|(2+2\sqrt{\delta})+l_1(1+\delta)\right).
\end{aligned}$$

Therefore,

$$\|\mathbf{v}\|\leq \max\left\{1,\frac{\|\mathbf{b}\|^2+\lambda_{max}(\mathbf{A}^T\mathbf{A})(\delta+2\sqrt{\delta})+\|\mathbf{A}^T\mathbf{b}\|(\delta+2\sqrt{\delta})+l_1(1+\delta)}{l_1-l_2}\right\}. \tag{48}$$

Finally,

$$\begin{aligned}
&\|\mathbf{x}^*\|^2\\
&= \|\mathbf{v}\|^2+\|\mathbf{L}^T\mathbf{w}\|^2\\
&\overset{(44),(48)}{\leq} \max\left\{1,\frac{\|\mathbf{b}\|^2+\lambda_{max}(\mathbf{A}^T\mathbf{A})(\delta+2\sqrt{\delta})+\|\mathbf{A}^T\mathbf{b}\|(\delta+2\sqrt{\delta})+l_1(1+\delta)}{l_1-l_2}\right\}^2+\delta. \qquad \square
\end{aligned}$$

Remark. Recall that the sufficient condition for attainability is that $l_2<l_1$. Note that if l_2 is very close to l_1, then the upper bound on $\|\mathbf{x}^*\|^2$ might be very large.

5. The case $\mathbf{L}=\mathbf{I}$.

5.1. Strict unimodality of $\mathcal{G}$. In this section we show that in the case in which $\mathbf{L}=\mathbf{I}$, the function $\mathcal{G}$ defined in (20) has a very attractive property: *strictly unimodal*. A strictly unimodal function over an interval $[a,b]$ is a function that has a unique global minimum α^* and is strictly decreasing over $[a,\alpha^*]$ and strictly increasing over $[\alpha^*,b]$ (α^* can be equal to a or b and in that case the function is monotone). The fact that $\mathcal{G}$ is strictly unimodal implies that we can solve the one dimensional minimization problem efficiently (with, e.g., the golden section method; see [3]).

THEOREM 5.1. *Consider problem* (11) *with* $\mathbf{L}=\mathbf{I}$. *If* $\mathbf{A}^T\mathbf{b}\notin$ *Null*$(\mathbf{A}^T\mathbf{A}-\lambda_{\min}(\mathbf{A}^T\mathbf{A})\mathbf{I})^{\perp}$, *then* $\mathcal{G}$, *defined in* (20), *is differentiable for every* $\alpha>1$ *and strictly unimodal.*

Proof. First, by substituting $\mathbf{Q}_\alpha=\frac{1}{\alpha}\mathbf{A}^T\mathbf{A}+\rho\mathbf{I}$ and $\mathbf{f}_\alpha=\frac{1}{\alpha}\mathbf{A}^T\mathbf{b}$ into (27) we obtain the following sufficient condition for differentiability of $\mathcal{G}$ at α:

$$\mathbf{A}^T\mathbf{b}\notin \text{Null}(\mathbf{A}^T\mathbf{A}-\lambda_{\min}(\mathbf{A}^T\mathbf{A})\mathbf{I})^{\perp}.$$

Now, in order to prove the strict unimodality of $\mathcal{G}$, it is sufficient to prove the following property of $\mathcal{G}$: *if* $\mathcal{G}'(\alpha) = 0$, *then* $\mathcal{G}''(\alpha) > 0$. By differentiating both sides of (39), we obtain

$$\mathcal{G}''(\alpha) = \ddot{\mathbf{x}}(\alpha)^T f'_{\mathbf{x}}(\mathbf{x}(\alpha),\alpha) + \dot{\mathbf{x}}(\alpha)^T f''_{\mathbf{x}^2}(\mathbf{x}(\alpha),\alpha)\dot{\mathbf{x}}(\alpha) + 2\dot{\mathbf{x}}(\alpha)^T f''_{\mathbf{x}\alpha}(\mathbf{x}(\alpha),\alpha) + f''_{\alpha^2}(\mathbf{x}(\alpha),\alpha). \tag{49}$$

Differentiating (36), we have

$$\ddot{\mathbf{x}}(\alpha)^T g'_{\mathbf{x}}(\mathbf{x}(\alpha)) + \dot{\mathbf{x}}(\alpha)^T g''_{\mathbf{x}^2}(\mathbf{x}(\alpha))\dot{\mathbf{x}}(\alpha) = 0. \tag{50}$$

Therefore,

$$\begin{aligned}
\mathcal{G}''(\alpha) &= \mathcal{G}''(\alpha) - \lambda(\alpha)\cdot 0 \\
&\overset{(50)}{=} \mathcal{G}''(\alpha) - \lambda(\alpha)(\ddot{\mathbf{x}}(\alpha)^T g'_{\mathbf{x}}(\mathbf{x}(\alpha)) + \dot{\mathbf{x}}(\alpha)^T g''_{\mathbf{x}^2}(\mathbf{x}(\alpha))\dot{\mathbf{x}}(\alpha)) \\
&\overset{(49)}{=} \overbrace{\ddot{\mathbf{x}}(\alpha)^T \left(f'_{\mathbf{x}}(\mathbf{x}(\alpha),\alpha) - \lambda(\alpha) g'_{\mathbf{x}}(\mathbf{x}(\alpha))\right)}^{A} \\
&\quad + \overbrace{\dot{\mathbf{x}}(\alpha)^T \left(f''_{\mathbf{x}^2}(\mathbf{x}(\alpha),\alpha) - \lambda(\alpha) g''_{\mathbf{x}^2}(\mathbf{x}(\alpha))\right)\dot{\mathbf{x}}(\alpha)}^{B} \\
&\quad + \underbrace{2\dot{\mathbf{x}}(\alpha)^T f''_{\mathbf{x}\alpha}(\mathbf{x}(\alpha),\alpha) + f''_{\alpha^2}(\mathbf{x}(\alpha),\alpha)}_{C}.
\end{aligned}$$

By (34) we have $A = 0$ and

$$B = \dot{\mathbf{x}}(\alpha)^T \left(f''_{\mathbf{x}^2}(\mathbf{x}(\alpha),\alpha) - \lambda(\alpha) g''_{\mathbf{x}^2}(\mathbf{x}(\alpha))\right)\dot{\mathbf{x}}(\alpha) = \dot{\mathbf{x}}(\alpha)^T(\mathbf{Q}_\alpha - \lambda(\alpha)\mathbf{I})\dot{\mathbf{x}}(\alpha) \overset{(31)}{>} 0.$$

The latter inequality is true since by (36) $\dot{\mathbf{x}}(\alpha) \neq \mathbf{0}$. Suppose that $\mathcal{G}'(\alpha) = 0$; then

$$\dot{\mathbf{x}}(\alpha)^T f'_{\mathbf{x}}(\mathbf{x}(\alpha),\alpha) + f'_{\alpha}(\mathbf{x}(\alpha),\alpha) = 0,$$

which can also be written as

$$2\dot{\mathbf{x}}(\alpha)^T \left(\frac{\mathbf{A}^T(\mathbf{A}\mathbf{x}(\alpha) - \mathbf{b})}{\alpha}\right) - \frac{\|\mathbf{A}\mathbf{x}(\alpha) - \mathbf{b}\|^2}{\alpha^2} = -2\rho\dot{\mathbf{x}}(\alpha)^T \mathbf{L}^T\mathbf{L}\mathbf{x}(\alpha). \tag{51}$$

Now,

$$\begin{aligned}
C &= 2\dot{\mathbf{x}}(\alpha)^T f''_{\mathbf{x}\alpha}(\mathbf{x}(\alpha),\alpha) + f''_{\alpha^2}(\mathbf{x}(\alpha),\alpha) \\
&= -4\dot{\mathbf{x}}(\alpha)^T \frac{\mathbf{A}^T(\mathbf{A}\mathbf{x}(\alpha) - \mathbf{b})}{\alpha^2} + 2\frac{\|\mathbf{A}\mathbf{x}(\alpha) - \mathbf{b}\|^2}{\alpha^3} \\
&\overset{(51)}{=} 4\rho\frac{\dot{\mathbf{x}}(\alpha)^T\mathbf{L}^T\mathbf{L}\mathbf{x}(\alpha)}{\alpha}.
\end{aligned}$$

In our case $\mathbf{L} = \mathbf{I}$, and thus $C = 4\rho\frac{\dot{\mathbf{x}}(\alpha)^T\mathbf{x}(\alpha)}{\alpha} \overset{(36)}{=} \frac{2\rho}{\alpha} > 0$ and we conclude that, when $\mathcal{G}'(\alpha) = 0$, then $\mathcal{G}''(\alpha) = A + B + C > 0$, proving the unimodality property. ∎

5.2. Another approach to the case $\mathbf{L} = \mathbf{I}$.

5.2.1. The schematic algorithm. In the case $\mathbf{L} = \mathbf{I}$ the problem is given by

$$\min_{\mathbf{x}\in\mathbb{R}^n} \left\{ \mathcal{H}(\mathbf{x}) \equiv \frac{\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2}{\|\mathbf{x}\|^2+1} + \rho\|\mathbf{x}\|^2 \right\}. \tag{52}$$

We use the following simple observation, which goes back to Dinkelbach [6]: For every $t \in \mathbb{R}$, the following two statements are equivalent:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \mathcal{H}(\mathbf{x}) \le t,$$

$$\min_{\mathbf{x}\in\mathbb{R}^n} \{\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2 + \rho\|\mathbf{x}\|^4 + \rho\|\mathbf{x}\|^2 - t(\|\mathbf{x}\|^2+1)\} \le 0. \tag{53}$$

The minimization problem (53) also seems hard to solve; however, we will show in section 5.2.2 that it is in fact a very simple problem having essentially an explicit solution. Consider the function $\phi : \mathbb{R} \to \mathbb{R}$ defined by

$$\phi(t) = \min_{\mathbf{x}\in\mathbb{R}^n} \{\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2 + \rho\|\mathbf{x}\|^4 + \rho\|\mathbf{x}\|^2 - t(\|\mathbf{x}\|^2+1)\}.$$

We claim that ϕ is strictly decreasing. To prove this suppose that $t_1 < t_2$, and let $\mathbf{x}_{t_1} \equiv \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^n}\{\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2 + \rho\|\mathbf{x}\|^4 + \rho\|\mathbf{x}\|^2 - t_1(\|\mathbf{x}\|^2+1)\}$. Then

$$\begin{aligned}\phi(t_1) &= \|\mathbf{A}\mathbf{x}_{t_1}-\mathbf{b}\|^2 + \rho\|\mathbf{x}_{t_1}\|^4 + \rho\|\mathbf{x}_{t_1}\|^2 - t_1(\|\mathbf{x}_{t_1}\|^2+1)\\ &> \|\mathbf{A}\mathbf{x}_{t_1}-\mathbf{b}\|^2 + \rho\|\mathbf{x}_{t_1}\|^4 + \rho\|\mathbf{x}_{t_1}\|^2 - t_2(\|\mathbf{x}_{t_1}\|^2+1) \ge \phi(t_2).\end{aligned}$$

From the above observation we also have that $t^* \equiv \min_{\mathbf{x}\in\mathbb{R}^n} \mathcal{H}(\mathbf{x})$ is the unique root of $\phi(\cdot)$. Moreover, $t^* \in [0, \|\mathbf{b}\|^2]$ since

$$\phi(0) = \min_{\mathbf{x}\in\mathbb{R}^n} \{\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2 + \rho\|\mathbf{x}\|^4 + \rho\|\mathbf{x}\|^2\} \ge 0$$

and

$$\begin{aligned}\phi(\|\mathbf{b}\|^2) &= \min_{\mathbf{x}\in\mathbb{R}^n} \{\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2 + \rho\|\mathbf{x}\|^4 + (\rho-\|\mathbf{b}\|^2)\|\mathbf{x}\|^2 - \|\mathbf{b}\|^2\}\\ &\le \min_{\mathbf{x}\in\mathbb{R}^n} \{\|\mathbf{A}\mathbf{0}-\mathbf{b}\|^2 + \rho\|\mathbf{0}\|^4 + (\rho-\|\mathbf{b}\|^2)\|\mathbf{0}\|^2 - \|\mathbf{b}\|^2\} = 0.\end{aligned}$$

As a result, the optimal t^* can be found by, e.g., a simple bisection algorithm with an initial interval $[0, \|\mathbf{b}\|^2]$.

5.2.2. Solving the subproblem. The subproblem can also be written as

$$\min_{\mathbf{x}\in\mathbb{R}^n} \left\{\mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x} + (\rho-t)\|\mathbf{x}\|^2 + \rho\|\mathbf{x}\|^4 - 2\mathbf{b}^T\mathbf{A}\mathbf{x} + \|\mathbf{b}\|^2 - t\right\}.$$

Making the change of variables $\mathbf{x} = \mathbf{U}\mathbf{z}$, where $\mathbf{U}$ is orthogonal matrix diagonalizing $\mathbf{A}^T\mathbf{A}$, i.e., $\mathbf{U}^T\mathbf{A}^T\mathbf{A}\mathbf{U} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, the problem then reduces to

$$\min_{\mathbf{z}\in\mathbb{R}^n} \sum_{j=1}^{n} \left\{\lambda_j z_j^2 + (\rho-t)z_j^2 + \rho z_j^4 - 2f_j z_j\right\}, \tag{54}$$

where $\mathbf{f} = \mathbf{U}^T\mathbf{A}^T\mathbf{b}$. Note that since ρ canbe smaller than t, (54) might be a nonconvex problem. But, in fact, this does not really matter since this is a separable problem in its variables. Therefore, the solution of (54) requires solving n independent minimization problems:

$$\min_{\mathbf{z}_j\in\mathbb{R}} \left\{(\lambda_j + \rho - t)z_j^2 + \rho z_j^4 - 2f_j z_j\right\}. \tag{55}$$

The scalar objective function is a coercive function (since the dominating factor is z_j^4). Therefore, the minimum of (55) is attained at a point satisfying $g_j'(z_j) = 0$, where $g_j(z_j) = (\lambda_j + \rho - t)z_j^2 + \rho z_j^4 - 2f_j z_j$. Therefore, the minimum is attained at one of the real roots of

$$4\rho z_j^3 + 2(\lambda_j + \rho - t)z_j - 2f_j = 0. \tag{56}$$

This is a cubic equation and therefore can be solved explicitly by Cardano's formula. More precisely, the roots of the cubic equation $x^3 + 3Qx - 2R = 0$ are given by

$$x_1 = (R + \sqrt{Q^3 + R^2})^{1/3} + (R - \sqrt{Q^3 + R^2})^{1/3}$$

and

$$\begin{aligned} x_{2,3} = &-\frac{1}{2}\left[(R + \sqrt{Q^3 + R^2})^{1/3} + (R - \sqrt{Q^3 + R^2})^{1/3}\right] \\ &\pm \frac{\sqrt{3}}{2} i \left[(R + \sqrt{Q^3 + R^2})^{1/3} - (R - \sqrt{Q^3 + R^2})^{1/3}\right]. \end{aligned}$$

In any case, it has three real roots if $Q^3 + R^2 \leq 0$ and only one real root (and two complex roots) otherwise. The minimum of (55) is attained at one of the roots of the cubic equation (56). Therefore, the initial step of the algorithm is to diagonalize the matrix $\mathbf{A}^T\mathbf{A}$, and then a bisection algorithm is invoked to find the unique root of the strictly decreasing function ϕ. The calculation of a function value of ϕ requires solving n cubic equations.

The algorithm described in this section is summarized below.

ALGORITHM TRTLSI.

Input: $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{b} \in \mathbb{R}^m, \rho > 0$, and ϵ—a tolerance parameter.

Output: $\mathbf{x}^*$—a solution (up to some tolerance) of the TRTLS problem (11) with $\mathbf{L} = \mathbf{I}$.

1. **Set** $t_{\min} \leftarrow 0$ and $t_{\max} \leftarrow \|\mathbf{b}\|^2$.
2. Compute the spectral decomposition of $\mathbf{A}^T\mathbf{A}$: $\mathbf{U}^T\mathbf{A}^T\mathbf{A}\mathbf{U} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$.
3. **Set** $\mathbf{f} \leftarrow \mathbf{U}^T\mathbf{A}^T\mathbf{b}$.
4. **While** $|t_{\max} - t_{\min}| > \epsilon$ **repeat** steps (a), (b), and (c):
   (a) For every $j = 1, 2, \ldots, n$ compute the solutions $z_1^j, \ldots, z_{p_j}^j$ of the one dimensional cubic equation (56). Here p_j denotes the number of different real solutions of the jth cubic equation.
   (b) For every $j = 1, 2, \ldots, n$ **set**
   $$\beta_j \leftarrow \min_{k=1,\ldots,p_j} \{(\lambda_j + \rho - t)(z_k^j)^2 + \rho(z_k^j)^4 - 2f_j z_k^j\}.$$
   (c) **If** $\sum_{j=1}^n \beta_j - t < 0$, **then** $t_{\max} = t$; **else** $t_{\min} = t$.
5. **Set**
   $$m_j \leftarrow \operatorname*{argmin}_{k=1,\ldots,p_j} \{(\lambda_j + \rho - t)(z_k^j)^2 + \rho(z_k^j)^4 - 2f_j z_k^j\}.$$
6. Let $\mathbf{w}$ be such that $w_j = z_{m_j}^j$ for every $j = 1, \ldots, n$.
7. Set $\mathbf{x}^* = \mathbf{U}\mathbf{w}$.

The dominant computational effort when applying algorithm TRTLSI is the single calculation of the spectral decomposition of $\mathbf{A}^T\mathbf{A}$, which requires $O(n^3)$ operations. At each iteration the computational cost of solving n cubic equations is $O(n)$. For problems with up to several hundreds of variables, algorithm TRTLSI is therefore applicable. However, for problems with thousands or even tens of thousands of variables, algorithm TRTLSI cannot be implemented. Nevertheless, it is still possible to use the approach of solving the one dimensional minimization problem (19) since large-scale TRSs can be solved efficiently (see, e.g., [5, 7] and the references therein). A specific implementation of the algorithm for a general regularization matrix is given in the subsequent section.

6. Implementation and example. We have shown that solving the TRTLS problem (11) (for a general regularization matrix $\mathbf{L}$) reduces to a problem of solving a one dimensional minimization problem over a closed interval. The specific details of the algorithm (for a general regularization matrix) depend on the choice of the one dimensional solver and the selection of a method for solving the TRS. In section 6.1 we describe a specific implementation—algorithm TRTLSG. We then apply the proposed algorithm in section 6.2 to an image deblurring example.

6.1. A detailed algorithm for the TRTLS problem. We use the method of Moré and Sorensen for solving the TRS (21). The method is based on applying Newton's method to the problem

$$\frac{1}{\phi(\lambda)} - \frac{1}{\beta} = 0, \tag{57}$$

where $\phi(\lambda) \equiv \mathbf{f}^T(\mathbf{Q} - \lambda\mathbf{I})^{-1}\mathbf{f}$. The main computational effort at each iteration is the calculation of a Cholesky factorization of a matrix of the form $\mathbf{Q} - \lambda\mathbf{I}$. For large-scale problems the Cholesky factorization is not affordable, and other nondirect methods, such as Krylov subspace methods, can be employed (see, e.g., [29] and the references in [5, 7]). In our example $n = 1024$ so that Moré and Sorensen's method is appropriate.

Algorithm TRTLSG.

Input: $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{b} \in \mathbb{R}^m, \mathbf{L} \in \mathbb{R}^{k\times n}, \rho > 0$, and ϵ_1, ϵ_2—tolerance parameters.

Output: $\mathbf{x}^*$—a solution (up to some tolerance) of the TRTLS problem (11).

1. **Set** $\alpha_{\min} \leftarrow 1 + \epsilon_1$.
2. **If** $k = n$, **set** $\alpha_{\max}$ to be the upper bound given in Lemma 4.2; **else** $\alpha_{\max}$ is equal to the upper bound given in Theorem 4.3.
3. **While** $|\alpha_{\max} - \alpha_{\min}| > \epsilon_2$ **repeat** steps (a), (b), and (c):
   (a) **Set** $\alpha \leftarrow \frac{\alpha_{min}+\alpha_{max}}{2}$.
   (b) Solve the following TRS:

$$\min_{\|\mathbf{x}\|^2=\alpha-1} \left\{\mathbf{x}^T\mathbf{Q}_\alpha\mathbf{x} - 2\mathbf{f}_\alpha^T\mathbf{x}\right\},$$

   where $\mathbf{Q}_\alpha$ and $\mathbf{f}_\alpha$ are given in (26), and obtain a solution $\mathbf{x}(\alpha)$ and a multiplier $\lambda(\alpha)$ that satisfy conditions (22), (23), and (24) (with $\mathbf{Q} = \mathbf{Q}_\alpha, \mathbf{f}_\alpha, \mathbf{x}^* = \mathbf{x}(\alpha)$, and $\lambda^* = \lambda(\alpha)$).
   (c) **If** $\underbrace{\lambda(\alpha) - \frac{\|\mathbf{A}\mathbf{x}(\alpha) - \mathbf{b}\|^2}{\alpha^2}}_{\mathcal{G}'(\alpha)} > 0$, **then** $\alpha_{\max} = \alpha$; **else** $\alpha_{\min} = \alpha$.
4. Set $\mathbf{x}^* = \mathbf{x}(\alpha_{\max})$.

In our implementation the tolerance parameters take the values $\epsilon_1 = 10^{-1}$ and $\epsilon_2 = 10^{-6}$.

The one dimensional solver in algorithm TRTLSG is a simple bisection algorithm applied to the derivative of $\mathcal{G}(\alpha)$. To *guarantee* global convergence of the algorithm, the function $\mathcal{G}$ should be unimodal. For the case $\mathbf{L} = \mathbf{I}$ the unimodality property was proven in section 5.1. We observed through numerous random examples of the TRTLS problem of different dimensions ($4 \leq n, m, k \leq 1024$) that the unimodality property almost always holds even for $\mathbf{L} \neq \mathbf{I}$. The "bad" example in Figure 2 (with $m = n = 4, k = 3$) is an exceptional example. Moreover, for $n > 10$ we have not been able to find a single example which is not unimodal. Thus, for all practical purposes, algorithm TRTLSG finds the global optimum. If, for some reason, the function $\mathcal{G}$ is not unimodal, then algorithm TRTLSG does not necessarily converge to a global minimum and more sophisticated one dimensional global solvers should be employed.

6.2. Example. To illustrate the effectiveness of the TRTLS approach, we consider an image deblurring example. The TRTLS problems arising in this example were solved by algorithm TRTLSG implemented in MATLAB.

The choice of the regularization parameter ρ in our experiments was done by using the L-curve method [16, 21]. This method was originally devised as a method for choosing the regularization parameter for a regularized *least squares* problem. The L-curve is a plot of the L-norm $\|\mathbf{L}\mathbf{x}_\rho\|$ versus the residual $\|\mathbf{A}\mathbf{x}_\rho - \mathbf{b}\|$, where $\mathbf{x}_\rho$ is the solution of the regularization method with parameter ρ. The obtained plot usually has an L-shape appearance, and the chosen parameter is the one which is the closest to the left bottom corner. For the TLS problem, we follow the L-curve approach described in [24]: we plot the L-norm $\|\mathbf{L}\mathbf{x}_\rho\|^2$ versus the *fractional residual* $\|\mathbf{A}\mathbf{x}_\rho - \mathbf{b}\|^2/(1 + \|\mathbf{x}_\rho\|^2)$ for a various number of regularization parameters and pick the parameter closest to the L-shaped corner.

Let X be a 32×32 two dimensional image obtained from the sum of three harmonic oscillations:

$$X(z_1, z_2) = \sum_{l=1}^{3} a_i \cos(w_{l,1} z_1 + w_{l,2} z_2 + \phi_l), \quad \left(w_{l,i} = \frac{2\pi k_{l,i}}{n} \right), \ 1 \leq z_1, z_2 \leq 32,$$

where $k_{l,i} \in \mathbb{Z}^2$ (see Figure 3—true image). The specific values of the parameters are given in Table 1.

TABLE 1
Image parameters.

l	a_l	$w_{l,1}$	$w_{l,2}$	ϕ_l
1	1.3936	0.1473	0.0982	5.8777
2	0.5579	0.0982	0.0982	5.7611
3	0.8529	0.0491	0.0982	2.5778

We consider the square system

$$\mathbf{A}_{\text{true}}\mathbf{x}_{\text{true}} = \mathbf{b}_{\text{true}},$$

where $\mathbf{x}_{\text{true}} \in \mathbb{R}^{1024}$ is obtained by stacking the columns of the 32×32 image X. The vector $\mathbf{x}_{\text{true}}$ was normalized so that $\|\mathbf{x}_{\text{true}}\| = 1$. The 1024×1024 matrix $\mathbf{A}_{\text{true}}$ represents an atmospheric turbulence blur originating from [15] and implemented in the function blur(n,3) from the "Regularization Tools" [17]. The observed matrix

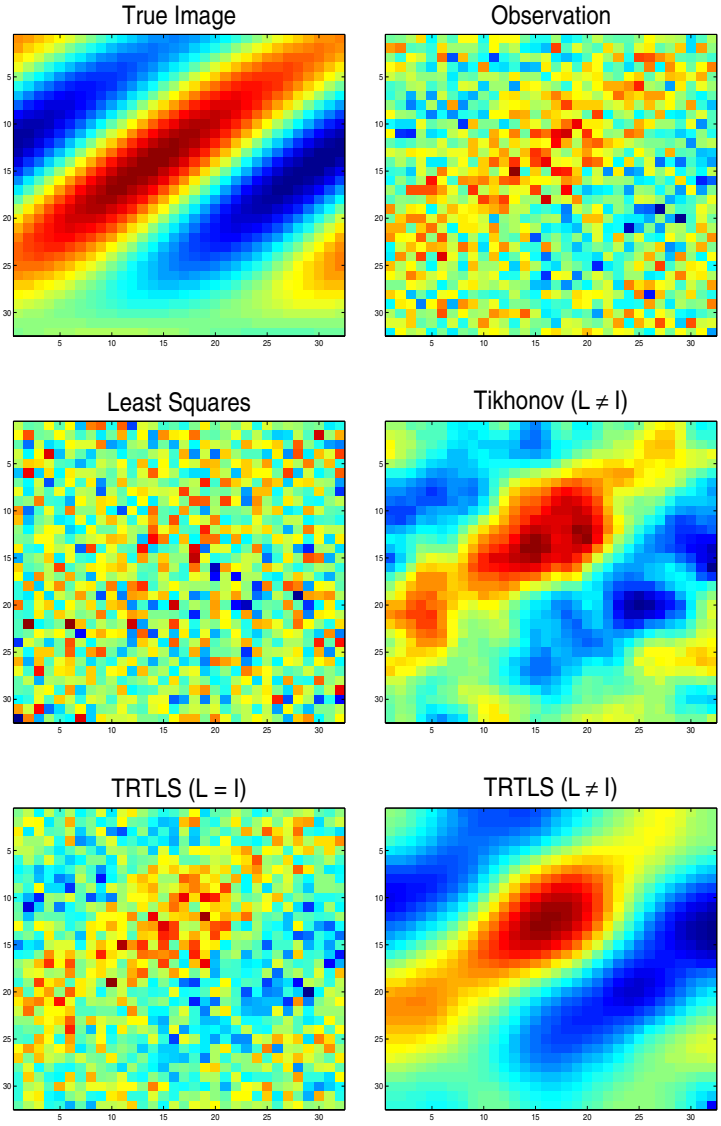


FIG. 3. *Results for different regularization solvers.*

and vector were generated by adding white noise to the data: $\mathbf{A} = \mathbf{A}_{true} + \sigma\mathbf{E}$ and $\mathbf{b} = \mathbf{b}_{\text{true}} + \sigma\mathbf{e}$, where each component of $\mathbf{E} \in \mathbb{R}^{1024\times 1024}$ and $\mathbf{e} \in \mathbb{R}^{1024}$ was generated from a standard normal distribution.

In our experiment the standard deviation σ was chosen to be 0.05, which results in a highly noisy image (see Figure 3—observation). The LS estimator was implemented in the function lsqr from [17]; it can be readily observed that it produces a poor image.

The choice of regularization matrix has a major influence on the quality of the obtained image. The solution of the TRTLS problem with *standard regularization* produces an unsatisfactory image (see Figure 3—TRTLS with $L = I$).

To produce a better result, we use a regularization matrix that accounts for the

smoothness property of this image. In particular, the matrix $\mathbf{L}$ was chosen to satisfy the relation

$$\mathbf{L}^T\mathbf{L} = \mathbf{R}^T\mathbf{R} + \mathbf{I}, \tag{58}$$

where $\mathbf{R}$ is a discrete approximation of the Laplace operator, which is a two dimensional convolution with the following mask:

$$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}.$$

This operator is standard in image processing [20]. With this choice of $\mathbf{L}$, the TRTLS algorithm gave the much better image (see Figure 3—TRTLS with $L \neq I$). We also compared our results to the one obtained by the classic Tikhonov regularization of the least squares, i.e., the solution of the minimization problem

$$\min_{\mathbf{x}}\{\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 + \rho\|\mathbf{L}\mathbf{x}\|^2\}$$

with the same regularization matrix given in (58). Tikhonov regularization of the *least squares* (see Figure 3—Tikhonov $\mathbf{L} \neq \mathbf{I}$) provides a better image than the least squares, but its quality is inferior to the one obtained by the corresponding TRTLSG algorithm.

Acknowledgment. We give special thanks to the referees for their constructive comments and suggestions.

REFERENCES

[1] A. Beck, A. Ben-Tal, and M. Teboulle, *Finding a global optimal solution for a quadratically constrained fractional quadratic problem with applications to the regularized total least squares*, SIAM J. Matrix Anal. Appl., to appear.

[2] A. Ben-Tal and M. Teboulle, *Hidden convexity in some nonconvex quadratically constrained quadratic programming*, Math. Programming, 72 (1996), pp. 51–63.

[3] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed., Athena Scientific, Belmont, MA, 1999.

[4] Å. Björck, *Numerical Methods for Least Squares Problems*, SIAM, Philadelphia, 1996.

[5] A. R. Conn, N. I. M. Gould, and P. L. Toint, *Trust-Region Methods*, MPS/SIAM Ser. Optim. 1, SIAM, Philadelphia, 2000.

[6] W. Dinkelbach, *On nonlinear fractional programming*, Management Sci., 13 (1967), pp. 492–498.

[7] C. Fortin and H. Wolkowicz, *The trust region subproblem and semidefinite programming*, Optim. Methods Softw., 19 (2004), pp. 41–67.

[8] W. Gander, G. H. Golub, and U. von Matt, *A constrained eigenvalue problem*, Linear Algebra Appl., 114/115 (1989), pp. 815–839.

[9] J. Gauvin and F. Dubeau, *Differential properties of the marginal function in mathematical programming*, Math. Programming Stud., 19 (1982), pp. 101–119.

[10] G. H. Golub, P. C. Hansen, and D. P. O'Leary, *Tikhonov regularization and total least squares*, SIAM J. Matrix Anal. Appl., 21 (1999), pp. 185–194.

[11] G. H. Golub and C. F. Van Loan, *An analysis of the total least-squares problem*, SIAM J. Numer. Anal., 17 (1980), pp. 883–893.

[12] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, MD, 1996.

[13] N. I. M. Gould, S. Lucidi, M. Roma, and P. L. Toint, *Solving the trust-region subproblem using the Lanczos method*, SIAM J. Optim., 9 (1999), pp. 504–525.

[14] H. Guo and R. Renaut, *A regularized total least squares algorithm*, in Total Least Squares and Errors-in-Variables Modeling, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2002, pp. 57–66.

[15] M. Hanke and P. C. Hansen, *Regularization methods for large-scale problems*, Surveys Math. Indust., 3 (1993), pp. 253–315.
[16] P. C. Hansen and D. P. O'Leary, *The use of the L-curve in the regularization of discrete ill-posed problems*, SIAM J. Sci. Comput., 14 (1993), pp. 1487–1503.
[17] P. C. Hansen, *Regularization tools: A Matlab package for analysis and solution of discrete ill-posed problems*, Numer. Algorithms, 6 (1994), pp. 1–35.
[18] P. C. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*, SIAM, Philadelphia, 1998.
[19] S. Van Huffel and J. Vandewalle, *The Total Least Squares Problem: Computational Aspects and Analysis*, Frontiers Appl. Math. 9, SIAM, Philadelphia, PA, 1991.
[20] A. K. Jain, *Fundamentals of Digital Image Processing*, Prentice–Hall, Englewood Cliffs, NJ, 1989.
[21] C. L. Lawson and R. J. Hanson, *Solving least squares problems*, Prentice–Hall, Englewood Cliffs, NJ, 1974.
[22] O. L. Mangasarian, *Nonlinear programming*, McGraw–Hill, New York, 1969.
[23] J. J. Moré and D. C. Sorensen, *Computing a trust region step*, SIAM J. Sci. Statist. Comput., 4 (1983), pp. 553–572.
[24] R. A. Renaut and H. Guo, *Efficient algorithms for solution of regularized total least squares*, SIAM J. Matrix Anal. Appl., 26 (2005), pp. 457–476.
[25] F. Rendel and H. Wolkowicz, *A semidefinite framework for trust region subproblems with applications to large scale minimization*, Math. Programming, 77 (1997), pp. 273–299.
[26] A. Shapiro, *Second order sensitivity analysis and asymptotic theory of parameterized nonlinear programs*, Math. Programming, 33 (1985), pp. 280–299.
[27] B. W. Silverman, *On the estimation of a probability density function by the maximum penalized likelihood method*, Ann. Statist., 10 (1982), pp. 795–810.
[28] D. Sima, S. Van Huffel, and G. H. Golub, *Regularized total least squares based on quadratic eigenvalue problem solvers*, BIT, 44 (2004), pp. 793–812.
[29] D. C. Sorensen, *Minimization of a large-scale quadratic function subject to a spherical constraint*, SIAM J. Optim., 7 (1997), pp. 141–161.
[30] T. Steihaug, *The conjugate gradient method and trust regions in large scale optimization*, SIAM J. Numer. Anal., 20 (1983), pp. 626–637.
[31] R. J. Stern and H. Wolkowicz, *Indefinite trust region subproblems and nonsymmetric eigenvalue perturbations*, SIAM J. Optim., 5 (1995), pp. 286–313.
[32] P. D. Tao and L. T. H. An, *A D.C. optimization algorithm for solving the trust-region subproblem*, SIAM J. Optim., 8 (1998), pp. 476–505.
[33] A. N. Tikhonov and V. Y. Arsenin, *Solution of Ill-Posed Problems*, V.H. Winston, Washington, DC, 1977.
[34] F. Zhang, *Matrix Theory*, Springer, New York, 1999.

SIAM J. OPTIM.
Vol. 17, No. 1, pp. 119–146
© 2006 Society for Industrial and Applied Mathematics

CONSTRAINT REDUCTION FOR LINEAR PROGRAMS WITH MANY INEQUALITY CONSTRAINTS*

ANDRÉ L. TITS†, P.-A. ABSIL‡, AND WILLIAM P. WOESSNER§

Abstract. Consider solving a linear program in standard form where the constraint matrix A is $m \times n$, with $n \gg m \gg 1$. Such problems arise, for example, as the result of finely discretizing a semi-infinite program. The cost per iteration of typical primal-dual interior-point methods on such problems is $O(m^2 n)$. We propose to reduce that cost by replacing the normal equation matrix, AD^2A^T, where D is a diagonal matrix, with a "reduced" version (of same dimension), $A_Q D_Q^2 A_Q^T$, where Q is an index set including the indices of M most nearly active (or most violated) dual constraints at the current iterate, with $M \geq m$ a prescribed integer. This can result in a speedup of close to $n/|Q|$ at each iteration. Promising numerical results are reported for constraint-reduced versions of a dual-feasible affine-scaling algorithm and of Mehrotra's predictor-corrector method [S. Mehrotra, *SIAM J. Optim.*, 2 (1992), pp. 575–601]. In particular, while it could be expected that neglecting a large portion of the constraints, especially at early iterations, may result in a significant deterioration of the search direction, it appears that the total number of iterations typically remains essentially constant as the size of the reduced constraint set is decreased down to some threshold. In some cases this threshold is a small fraction of the total set. In the case of the affine-scaling algorithm, global convergence and local quadratic convergence are proved.

Key words. linear programming, constraint reduction, column generation, primal-dual interior-point methods, affine scaling, Mehrotra's predictor-corrector

AMS subject classifications. 90C05, 65K05, 90C06, 90C34, 90C51

DOI. 10.1137/050633421

1. Introduction.

Consider a primal-dual linear programming pair in standard form, i.e.,

$$\min\ c^T x \text{ subject to } Ax = b,\ x \geq 0, \tag{1.1}$$

$$\max\ b^T y \text{ subject to } A^T y + s = c,\ s \geq 0, \tag{1.2}$$

where A has dimensions $m \times n$. The dual problem is equivalently written as

$$\max\ b^T y \text{ subject to } A^T y \leq c. \tag{1.3}$$

*Received by the editors June 10, 2005; accepted for publication (in revised form) December 9, 2005; published electronically April 21, 2006. The work of the first and third authors was supported in part by the National Science Foundation under grants DMI-9813057 and DMI-0422931, and by the US Department of Energy under grant DEFG0204ER25655. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation or those of the US Department of Energy.

http://www.siam.org/journals/siopt/17-1/63342.html

†Department of Electrical and Computer Engineering and Institute for Systems Research, University of Maryland, College Park, MD 20742 (andre@umd.edu).

‡Department of Mathematical Engineering, Université catholique de Louvain, Belgium (http://www.inma.ucl.ac.be/∼absil/). The work of this author was supported in part by Microsoft Research through a Research Fellowship at Peterhouse, Cambridge. Part of this work was done while this author was a Research Fellow with the Belgian National Fund for Scientific Research (Aspirant du F.N.R.S.) at the University of Liège, and while he was a Postdoctoral Associate with the School of Computational Science of Florida State University.

§Department of Computer Science and Institute for Systems Research, University of Maryland, College Park, MD 20742.

Most algorithms that have been proposed for the numerical solution of such problems belong to one of two classes: simplex methods and interior-point methods. For background on such methods, see, e.g., [NS96] and [Wri97]. In both classes of algorithms, the main computational task at each iteration is the solution of a linear system of equations. In the simplex case, the system has dimension m; in the interior-point case it has dimensions $2n + m$, but can readily be reduced ("normal equations") to one of size m at the cost of forming the matrix $H := AS^{-1}XA^T$. Here S and X are diagonal but vary from iteration to iteration, and the cost of forming H, when A is dense, is of the order of m^2n operations at each iteration.

The focus of the present paper is the solution of (1.1)–(1.2) when $n \gg m \gg 1$, i.e., when there are many more variables than equality constraints in the primal, many more inequality constraints than variables in the dual. This includes fine discretizations of "semi-infinite" problems of the form

$$\max\ b^T y \text{ subject to } a(\omega)^T y \le c(\omega)\ \forall \omega \in \Omega, \tag{1.4}$$

where, in the simplest cases, Ω is an interval of the real line. Network problems may also have a disproportionately large number of inequality constraints: For many network problems in dual form, there is one variable for each node of the network and one constraint for each arc or link, so that a linear program associated with a network with m nodes could have up to $O(m^2)$ constraints. Clearly, for such problems one iteration of a standard interior-point method would be computationally much more costly than one iteration of a simplex method. On the other hand, given the large number of vertices in the polyhedral feasible set of (1.3), the number of iterations needed to approach a solution with an interior-point method is likely to be significantly smaller than that needed when a simplex method is used.

Intuitively, when $n \gg m$, most of the constraints in (1.3) are of little or no relevance. Conceivably, if an interior-point search direction were computed based on a much smaller problem, with only a small subset of the constraints, significant progress could still be made toward a solution, provided this subset were astutely selected. Motivated by such consideration, in the present paper we aim at devising interior-point methods for the solution of (1.1)–(1.2) with $n \gg m \gg 1$, with drastically reduced computational cost per iteration. In a sense, such an algorithm would combine the best aspects of simplex methods and interior-point methods in the context of problems for which $n \gg m \gg 1$: each iteration would be effected at low computational cost, yet the iterates would follow an "interior" trajectory rather than being constrained to proceed along edges.

The issue of computing search directions for linear programs of the form (1.3) with $n \gg m$—or for semi-infinite linear programs (with a continuum of inequality constraints)—based on a small subset of the constraints has been an active area of research for many years. In most cases, the proposed schemes are based on logarithmic barrier ("primal") interior-point methods. In one approach, known as "column generation" (for the A matrix) or "build-up" (see, e.g., [Ye92, dHRT92, GLY94, Ye97]), constraints are added to (but never deleted from) the constraint set iteratively as they are deemed critical. In particular, the scheme studied in [Ye97] allows for more than one constraint (column) to be added at each step, and it is proved that the algorithm terminates in polynomial time with a bound whose dependence on the constraints is limited to those that are eventually included in the constraint set. In [Ye92, GLY94, Ye97], in the spirit of cutting-plane methods, the successive iterates are infeasible for (1.3) and the algorithm stops as soon as a feasible point is achieved;

while in the approach proposed in [dHRT92] all iterates are feasible for (1.3). Another approach is the "build-down" process (e.g., [Ye90]) by which columns of A are discarded when it is determined that the corresponding constraints are guaranteed not to be active at the solution. Both build-up [dHRT92] and build-down [Ye90] approaches were subsequently combined in [dHRT94], and a complexity analysis for the semi-infinite case was carried out in [LRT99].

In the present paper, a constraint reduction scheme is proposed in the context of *primal-dual* interior-point methods. Global and local quadratic convergence are proved in the case of a primal-dual affine-scaling (PDAS) method. (An early version of this analysis appeared in [Tit99].) Distinctive merits of the proposed scheme are its simplicity and the fact that it can be readily incorporated into other primal-dual interior-point methods. In the scheme's simplest embodiment, the constraint set is determined "from scratch" at the beginning of each iteration, rather than being updated in a build-up/build-down fashion. Promising numerical results are reported with constraint-reduced versions of the PDAS method and of Mehrotra's predictor-corrector (MPC) algorithm [Meh92]. Strikingly, while (consistent with conventional wisdom) the unreduced version of MPC significantly outperformed that of PDAS in our random experiments, the reduced version of PDAS performed essentially at the same level, in terms of CPU time, as that of MPC.

The remainder of the paper is organized as follows. In section 2, the basics of primal-dual interior-point methods are reviewed and the computational cost per iteration is analyzed, with special attention paid to possible gains to be achieved in certain steps by ignoring most constraints. Section 3 contains the heart of this paper's contribution. There, a dual-feasible PDAS algorithm is proposed that features a constraint-reduction scheme. Global and local quadratic convergence of this algorithm are proved, and numerical results are reported that suggest that, even with a simplistic implementation, the constraint-reduction scheme may lead to significant speedup. In section 4, promising numerical results are reported for a similarly reduced MPC algorithm, both with a dual-feasible initial point and with an infeasible initial point. Finally, section 5 is devoted to concluding remarks.

2. Preliminaries. Let

$$\mathbf{n} := \{1, 2, \ldots, n\};$$

for $i \in \mathbf{n}$, let $a_i \in \mathbb{R}^m$ denote the ith column of A; let F be the feasible set for (1.3), i.e.,

$$F := \{y : A^T y \leq c\},$$

and let $F^{\mathrm{o}} \subseteq \mathbb{R}^m$ denote the dual strictly feasible set

$$F^{\mathrm{o}} := \{y : A^T y < c\}.$$

Also, given $y \in F$, let $I(y)$ denote the index set of active constraints at y, i.e.,

$$I(y) := \{i \in \mathbf{n} : a_i^T y = c_i\}.$$

Given any index set $Q \subseteq \mathbf{n}$, let A_Q denote the $m \times |Q|$ matrix obtained from A by deleting all columns a_i with $i \notin Q$; similarly let x_Q and s_Q denote the vectors of size $|Q|$ obtained from x and s by deleting all entries x_i and s_i with $i \notin Q$. Further, following standard practice, let X denote $\mathrm{diag}(x_i, i \in \mathbf{n})$, and S denotes $\mathrm{diag}(s_i, i \in \mathbf{n})$.

When subscripts, superscripts, or diacritical signs are attached to x and s, they are inherited by x_Q, s_Q, X, and S. The rest of the notation is standard. In particular, $\|\cdot\|$ denotes the Euclidean norm.

Primal-dual interior-point algorithms use search directions based on the Newton step for the solution of the equalities in the Karush–Kuhn–Tucker (KKT) conditions for (1.2), or a perturbation thereof, while maintaining positivity of x and s. Given $\mu > 0$, the perturbed KKT conditions of interest are

$$A^T y + s - c = 0, \tag{2.1a}$$

$$Ax - b = 0, \tag{2.1b}$$

$$Xs = \mu e, \tag{2.1c}$$

$$x, s \geq 0, \tag{2.1d}$$

with $\mu = 0$ yielding the true KKT conditions. Given a current guess (x, y, s), the Newton step of interest is the solution to the linear system

$$\begin{bmatrix} 0 & A^T & I \\ A & 0 & 0 \\ S & 0 & X \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \\ \Delta s \end{bmatrix} = \begin{bmatrix} -r_c \\ -r_b \\ -Xs + \mu e \end{bmatrix}, \tag{2.2}$$

where

$$r_b := Ax - b, \quad r_c := A^T y + s - c$$

are the primal and dual *residuals*. Applying block Gaussian elimination to eliminate Δs yields the system (usually referred to as "augmented system")

$$\begin{bmatrix} 0 & A \\ XA^T & -S \end{bmatrix} \begin{bmatrix} \Delta y \\ \Delta x \end{bmatrix} = \begin{bmatrix} -r_b \\ -Xr_c + Xs - \mu e \end{bmatrix}, \tag{2.3a}$$

$$\Delta s = -A^T \Delta y - r_c. \tag{2.3b}$$

With $s > 0$, further elimination of Δx results in the "normal equations"

$$AS^{-1}XA^T \Delta y = -r_b + A(-S^{-1}Xr_c + x - \mu S^{-1}e), \tag{2.4a}$$

$$\Delta s = -A^T \Delta y - r_c, \tag{2.4b}$$

$$\Delta x = -x + \mu S^{-1}e - S^{-1}X\Delta s. \tag{2.4c}$$

Note that (2.4a) is equivalently written as

$$AS^{-1}XA^T \Delta y = b - AS^{-1}(Xr_c + \mu e).$$

For ease of reference, define the Jacobian and "augmented" Jacobian

$$J(A, x, s) := \begin{bmatrix} 0 & A^T & I \\ A & 0 & 0 \\ S & 0 & X \end{bmatrix}, \quad J_{\mathrm{a}}(A, x, s) := \begin{bmatrix} 0 & A \\ XA^T & -S \end{bmatrix}. \tag{2.5}$$

The following result is proven in the appendix.[1]

[1] Concerning the second claim of Lemma 1, only sufficiency is used in the convergence analysis, but the fact that the listed conditions are in fact necessary and sufficient may be of independent interest. We could not find this result (or even the sufficiency portion) in the literature, so are providing a proof for completeness. We would be grateful to anyone who would point us to a reference for the result.

LEMMA 1. *$J_a(A,x,s)$ is nonsingular if and only if $J(A,x,s)$ is. Further, suppose that $x \geq 0$ and $s \geq 0$.*[2] *Then $J(A,x,s)$ is nonsingular if and only if the following three conditions hold: (i) $|x_i| + |s_i| > 0$ for all i, (ii) $\{a_i : s_i = 0\}$ is linear independent, and (iii) $\{a_i : x_i \neq 0\}$ spans $\mathbb{R}^m$.*

In the next two sections, two types of primal-dual interior-point methods are considered: first, a dual-feasible (but primal-infeasible) PDAS algorithm, then a version of MPC. In the former, at each iteration the normal equations (2.4) are solved once, with $\mu = 0$, and $r_c = 0$. In the latter, the normal equations are solved twice per iteration with different right-hand sides.

We assume that A is dense. For large m and $n \gg m$, the bulk of the CPU cost is consumed by the solution of the normal equations (2.4). Indeed, the number of operations (per iteration) in other computations amounts to at most a small multiple of n. As for the operations involved in solving the normal equations, the operation count is roughly as follows:

- Forming $H := AS^{-1}XA^T$: m^2n;
- Forming $v := b - AS^{-1}(Xr_c + \mu e)$: $2mn$;
- Solving $H\Delta y = v$ (Cholesky factorization): $m^3/3$;
- Computing $\Delta s := -A^T\Delta y - r_c$: $2mn$;
- Computing $\Delta x := -x + S^{-1}(-X\Delta s + \mu e)$: $2n$.

(Because both algorithms we consider update y and s by taking a common step $\hat{t}$ along Δy and Δs, r_c is updated at no cost: the new value is $(1 - \hat{t})$ times the old value.)

The above suggests that maximum CPU savings should be obtained by replacing, in the definition of H, matrix A by its submatrix A_Q, corresponding to a suitably chosen index set Q. The cost of forming H would then be reduced to $m^2|Q|$ operations. In this paper, we investigate the effect of making that modification only, and leaving all else unchanged, so as to least "perturb" the original algorithms.

A central issue is then the choice of Q. Given $y \in \mathbb{R}^m$ and $M \geq m$, let $\mathcal{Q}_M(y)$ be the set of all subsets of $\mathbf{n}$ that contain the indexes of M leftmost components of $c - A^Ty$. More precisely (some components of $c - A^Ty$ may be equal, so "M leftmost" may not be uniquely defined), let

(2.6)
$$\mathcal{Q}_M(y) := \{Q \subseteq \mathbf{n} : \exists Q' \subseteq Q \text{ s.t. } |Q'| = M \text{ and } c_i - a_i^Ty \leq c_j - a_j^Ty \ \forall i \in Q', j \notin Q'\}.$$

Consequently, the statement "$Q \in \mathcal{Q}_M(y)$," to be used later, means that the index set Q contains the indices of M components of the vector $c - A^Ty$ that are smaller or equal to all other components of $c - A^Ty$. The convergence analysis of section 3.2 guarantees that our reduced PDAS algorithm will perform appropriately (under certain assumptions involving M) as long as Q is in $\mathcal{Q}_M(y)$. Given that $n \gg m$, this leaves a lot of leeway in choosing Q. We have two competing goals. On the one hand, we want $|Q|$ to be small enough that the iterations are significantly faster than when $Q = \mathbf{n}$. On the other hand, we want to include enough well chosen constraints that the iteration count remains low. In the numerical experiments we report toward the end of this paper, we restrict ourselves to a very simple scheme: we let Q be precisely the set of indexes of M leftmost components of $c - A^Ty$. Note that the "M leftmost" rule is inexpensive to apply: it takes at most $O(n \log n)$ operations—comparisons,

[2] The result still holds, and the same proof applies, under the milder but less intuitive assumption "$x_is_i \geq 0$ for all i." The result as stated is sufficient for our present purpose.

which are faster than additions or multiplications. (For small M, it takes even fewer comparisons.)

The following assumption will be needed in order for the proposed algorithms to be well defined.

ASSUMPTION 1. *All $m \times M$ submatrices of A have full row rank.*

LEMMA 2. *Suppose Assumption 1 holds. Let $x > 0$, $s > 0$, and $Q \subseteq \mathbf{n}$ with $|Q| \geq M$. Then $A_Q S_Q^{-1} X_Q A_Q^T$ is positive definite.*

Proof. Follows from positive definiteness of $S_Q^{-1} X_Q$ and full row rank of A_Q. ∎

3. A reduced, dual-feasible PDAS algorithm.

3.1. Algorithm statement. The proposed reduced primal-dual interior-point affine scaling (rPDAS) iteration is strongly inspired from the iteration described in [TZ94], a dual-feasible primal-dual iteration based on the Newton system discussed above, with $\mu = 0$ and $r_c = 0$. In particular, the normal equations for the algorithm of [TZ94] are given by

$$AS^{-1}XA^T\Delta y = b, \tag{3.1a}$$

$$\Delta s = -A^T\Delta y, \tag{3.1b}$$

$$\Delta x = -x - S^{-1}X\Delta s. \tag{3.1c}$$

The iteration focuses on the dual variables. Note that the iteration requires the availability of an initial $y^0 \in F^o$.

Iteration rPDAS.

Parameters. $\beta \in (0,1)$, $x_{\max} > 0$, $\underline{x} > 0$, integer M satisfying $m \leq M \leq n$.

Data. $y \in F^o$, $s := c - A^T y$, $x > 0$, with $x_i \leq x_{\max}$, $i = 1, \ldots, n$, $Q \in \mathcal{Q}_M(y)$.

Step 1. Compute search direction:

$$\text{Solve} \quad A_Q S_Q^{-1} X_Q A_Q^T \Delta y = b, \tag{3.2a}$$

$$\text{and compute} \qquad \Delta s := -A^T\Delta y, \tag{3.2b}$$

$$\Delta x := -x - S^{-1}X\Delta s. \tag{3.2c}$$

Set $\tilde{x} := x + \Delta x$ and, for $i \in \mathbf{n}$, set

$$(\tilde{x}_-)_i := \min\{\tilde{x}_i, 0\}.$$

Step 2. Updates:

(i) Compute the largest dual feasible step size

$$\bar{t} := \begin{cases} \infty & \text{if } \ \Delta s_i \geq 0 \ \ \forall i \in \mathbf{n}, \\ \min\{(-s_i/\Delta s_i) : \Delta s_i < 0, i \in \mathbf{n}\} & \text{otherwise.} \end{cases} \tag{3.3}$$

Set

$$\hat{t} := \min\{\max\{\beta\bar{t},\ \bar{t}\text{–}\|\Delta y\|\},\ 1\}. \tag{3.4}$$

Set $y^+ := y + \hat{t}\Delta y$, $s^+ := s + \hat{t}\Delta s$.

(ii) Set

$$x_i^+ := \min\{\max\{\min\{\|\Delta y\|^2 + \|\tilde{x}_-\|^2, \underline{x}\},\ \tilde{x}_i\},\ x_{\max}\}\ \forall i \in \mathbf{n}. \tag{3.5}$$

(iii) Pick $Q^+ \in \mathcal{Q}_M(y)$.

It should be noted that $(\Delta x_Q, \Delta y, \Delta s_Q)$ constructed by Iteration rPDAS also satisfies

$$\Delta s_Q = -A_Q^T \Delta y, \tag{3.6a}$$

$$\Delta x_Q = -x_Q - S_Q^{-1} X_Q \Delta s_Q, \tag{3.6b}$$

i.e., it satisfies the full set of normal equations associated with the constraint-reduced system. Equivalently, they satisfy the Newton system (with $\mu = 0$ and $r_c = 0$)

$$\begin{bmatrix} 0 & A_Q^T & I \\ A_Q & 0 & 0 \\ S_Q & 0 & X_Q \end{bmatrix} \begin{bmatrix} \Delta x_Q \\ \Delta y \\ \Delta s_Q \end{bmatrix} = \begin{bmatrix} 0 \\ b - A_Q x_Q \\ -X_Q s_Q \end{bmatrix}. \tag{3.7}$$

Remark 1. Primal update rule (3.5) is identical to the "dual" update rule used in [AT06] in the context of indefinite quadratic programming. (The "primal" problem in [AT06] can be viewed as a direct generalization of the *dual* (1.3).) As explained in [AT06], imposing the lower bound $\|\Delta y\|^2 + \|\tilde{x}_-\|^2$, which is a key to our global convergence analysis, precludes updating of x by means of a step in direction Δx; further, the specific form of this lower bound simplifies the global convergence analysis while, together with the bound $\bar{t} - \|\Delta y\|$ in (3.4), allowing for a quadratic convergence rate. Also key to our global convergence analysis (though in our experience not needed in practice) is the upper bound $x_{\max}$ imposed on all components of the primal variable x; it should be stressed that global convergence of the sequence of vectors $\tilde{x}$ to a solution is guaranteed regardless of the value of $x_{\max} > 0$. Finally, replacing in (3.5) $\min\{\|\Delta y\|^2 + \|\tilde{x}_-\|^2, \underline{x}\}$ simply with $\|\Delta y\|^2 + \|\tilde{x}_-\|^2$ would not affect the theoretical convergence properties of the algorithm. However, allowing small values of x_i^+ even when $\|\Delta y\|^2 + \|\tilde{x}_-\|^2$ is large proved beneficial in practice, especially in early iterations.

3.2. Convergence analysis. Before embarking on a convergence analysis, we introduce two more definitions. First, let $F^* \subseteq \mathbb{R}^m$ be the set of solutions of (1.3), i.e.,

$$F^* := \{y^* \in F : b^T y^* \geq b^T y \ \forall y \in F\}.$$

Of course, F^* is the set of y for which (2.1) holds with $\mu = 0$ for some $x, s \in \mathbb{R}^n$. Second, given $y \in F$, we will say that y is *stationary* for (1.3) whenever there exists $x \in \mathbb{R}^n$ such that

$$Ax = b \tag{3.8}$$

and

$$X(c - A^T y) = 0, \tag{3.9}$$

with no sign constraint imposed on x; equivalently, with $s := c - A^T y$ (≥ 0), (2.1a)–(2.1c) hold with $\mu = 0$ for some $x \in \mathbb{R}^n$. We will refer to such x as a *multiplier vector* associated with y. Clearly, every point in F^* is stationary, but not all stationary points are in F^*: in particular, all vertices of F are stationary.

3.2.1. Global convergence. We now show that, under certain nondegeneracy assumptions, the sequence of dual iterates generated by Iteration rPDAS converges to F^*. First, on the basis of Lemma 2, it is readily verified that, under Assumption 1, Iteration rPDAS is well defined. That it can be repeated *ad infinitum* then follows from the next proposition.

PROPOSITION 3. *Suppose Assumption* 1 *holds. Then Iteration rPDAS generates quantities with the following properties:* (i) $\Delta y \neq 0$ *if and only if* $b \neq 0$*;* (ii) $\hat{t} > 0$*,* $y^+ \in F^o$*,* $s^+ = c - A^T y^+ > 0$*, and* $x^+ > 0$*.*

Proof. The first claim is a direct consequence of Lemma 2 and (3.2a); the other claims are immediate. ∎

Now, let $y^0 \in F^o$, let $s^0 = c - A^T y^0$, let $x^0 > 0$, $Q^0 \subseteq \mathbf{n}$ with $|Q^0| \geq M$, and let $\{(x^k, y^k, s^k)\}$, $\{Q^k\}$, $\{\Delta y^k\}$, $\{\tilde{x}^k\}$, $\{\bar{t}^k\}$, and $\{\hat{t}^k\}$ be generated by successive applications of Iteration rPDAS starting at (x^0, y^0, s^0). Our analysis focuses on the dual sequence $\{y^k\}$.

In view of Proposition 3, $s^k = c - A^T y^k > 0$ for all k, so $y^k \in F^o$ for all k. We first note that, under no additional assumptions, the sequence of dual objective values is monotonic nondecreasing, strictly so if $b \neq 0$. This fact plays a central role in our global convergence analysis.

LEMMA 4. *Suppose Assumption* 1 *holds. If* $b \neq 0$*, then* $b^T \Delta y^k > 0$ *for all* k*. In particular,* $\{b^T y\}$ *is nondecreasing.*

Proof. The claim follows from (3.2a), Lemma 2, Proposition 3, and Step 2(i) of Iteration rPDAS. ∎

The remainder of the global convergence analysis is carried out under two additional assumptions. The first one implies that $\{y^k\}$ is bounded.

ASSUMPTION 2. *The dual solution set* F^* *is nonempty and bounded.*

Equivalently, the superlevel sets $\{y \in F : b^T y \geq \alpha\}$ are bounded for all α. Boundedness of $\{y^k\}$ then follows from its feasibility and monotonicity of $\{b^T y^k\}$ (Lemma 4 and Step 2(i) of Iteration rPDAS).

LEMMA 5. *Suppose Assumptions* 1 *and* 2 *hold; then* $\{y^k\}$ *is bounded.*

Our final nondegeneracy assumption ensures that small values of $\|\Delta y^k\|$ indicate that a stationary point of (1.3) is being approached (Lemma 6).

ASSUMPTION 3. *For all* $y \in F$*,* $\{a_i : i \in I(y)\}$ *is a linear independent set of vectors.*

LEMMA 6. *Suppose Assumptions* 1 *and* 3 *hold. Let* $y^* \in \mathbb{R}^m$ *and suppose that* K*, an infinite index set, is such that* $\{y^k\}$ *converges to* y^* *on* K*. If* $\{\Delta y^k\}$ *converges to zero on* K*, then* y^* *is stationary and* $\{\tilde{x}^k\}$ *converges to* x^* *on* K*, where* x^* *is the unique multiplier vector associated with* y^**.*

Proof. Suppose $\{\Delta y^k\} \to 0$ as $k \to \infty$, $k \in K$. Without loss of generality (by going down to a further subsequence if necessary), assume that, for some Q^*, $Q_k = Q^*$ for all $k \in K$. Equation (3.7) implies that

$$A_{Q^*} \tilde{x}^k_{Q^*} - b = 0 \quad \forall k \in K, \tag{3.10}$$

and (3.2b)–(3.2c) yield

$$x^k_i a_i^T \Delta y^k - s^k_i \tilde{x}^k_i = 0 \quad \forall i, \forall k. \tag{3.11}$$

Let $s^* = c - A^T y^*$, so $s^k \to s^*$ as $k \to \infty$, $k \in K$. Since $\{x^k\}$ is bounded ($x^k_i \in [0, x_{\max}]$ $\forall i$ by construction), it follows from (3.11) that for all $i \notin I(y^*)$ (i.e., all i for which $s^*_i > 0$), $\{\tilde{x}^k_i\} \to 0$ as $k \to \infty$, $k \in K$. Now, in view of Assumption 3 (linear

independence of the active constraints), $|I(y^*)| \leq m$ and, since $Q^k \in \mathcal{Q}_M(y^k)$ for all k, $M \geq m$, $I(y^*) \subseteq Q^*$. Hence, (3.10) yields

$$\sum_{i \in I(y^*)} \tilde{x}_i^k a_i - b \to 0 \text{ as } k \to \infty,\ k \in K,$$

and Assumption 3 implies that, for all $i \in I(y^*)$, $\{\tilde{x}_i^k\}$ converges on K, say, to x_i^*. Taking limits in (3.10)–(3.11) then yields

$$Ax^* - b = 0,$$

$$X^* s^* = 0,$$

implying that y^* is stationary, with multiplier vector x^*. Uniqueness of x^* again follows from Assumption 3. □

Proving that $\{y^k\}$ converges to F^* will be achieved in two main steps. The first objective is to show that $\{y^k\}$ converges to the set of *stationary points* of (1.3) (Lemma 9). This will be proved via a contradiction argument: if, for some infinite index set K, $\{y^k\}$ were to converge on K to a *nonsolution* point—for instance, to a nonstationary point—then $\{\Delta y^k\}$ would have to go to zero on K (Lemma 8), in contradiction with Lemma 6. The heart of the argument lies in the following lemma.

LEMMA 7. *Suppose Assumptions* 1, 2, *and* 3 *hold. Let K be an infinite index set such that*

$$\inf\{\|\Delta y^{k-1}\|^2 + \|\tilde{x}_-^{k-1}\|^2 :\ k \in K\} > 0.$$

Then $\{\Delta y^k\} \to 0$ as $k \to \infty,\ k \in K$.

Proof. In view of (3.5), for all $i \in \mathbf{n}$, x_i^k is bounded away from zero on K. Proceeding by contradiction, assume that, for some infinite index set $K' \subseteq K$, $\inf_{k \in K'} ||\Delta y^k|| > 0$. Since $\{y^k\}$ (see Lemma 5) and $\{x^k\}$ (see (3.5)) are bounded, we may assume, without loss of generality, that for some y^* and x^*, with $x_i^* > 0$ for all i, and some Q^* with $|Q^*| \geq M$,

$$\{y^k\} \to y^* \text{ as } k \to \infty, \quad k \in K',$$

$$\{x^k\} \to x^* \text{ as } k \to \infty, \quad k \in K',$$

and

$$Q^k = Q^* \quad \forall k \in K'.$$

Let $s^* := c - A^T y^*$; since $s^k = c - A^T y^k$ for all k, it follows that $\{s^k\} \to s^*$, $k \in K'$. Since in view of Lemma 1, of Assumptions 1 and 3, and of the fact that $x_i^* > 0$ for all i, the matrix $J(A_{Q^*}, x_{Q^*}^*, s_{Q^*}^*)$ is nonsingular, it follows from (3.7) that, for some v^* and $\tilde{x}_i^*$, $i \in Q^*$, with $v^* \neq 0$ (since $\inf_{k \in K'} \|\Delta y^k\| > 0$),

$$\{\Delta y^k\} \to v^* \text{ as } k \to \infty, \quad k \in K',$$

$$\{x_i^k\} \to \tilde{x}_i^* \text{ as } k \to \infty, \quad k \in K', \quad i \in Q^*.$$

In view of linear independence Assumption 3, since $\{s^k\} \to s^* = c - A^T y^*$, and since, by definition of $\mathcal{Q}_M$, $I(y^*) \subseteq Q^*$, it follows that s_i^k is bounded away from zero when $i \notin Q^*$, $k \in K'$. It then follows from (3.2b) and (3.2c) that, for some $\tilde{x}^*$,

$$\{\tilde{x}^k\} \to \tilde{x}^* \text{ as } k \to \infty, \quad k \in K'. \tag{3.12}$$

Now, Step 2(i) of Iteration rPDAS and (3.2c) yield

$$\bar{t}^k = -\frac{s_{i_k}^k}{\Delta s_{i_k}^k} = \frac{x_{i_k}^k}{\tilde{x}_{i_k}^k} \quad \text{for some } i_k,$$

for all $k \in K'$ such that $\bar{t}^k < \infty$. Since the components of $\{x^k\}$ are bounded away from zero on K' (since $x_i^* > 0$ for all i), it follows from (3.12) that $\bar{t}^k$ is bounded away from zero on K', and from Step 2(i) in Iteration rPDAS that the same holds for $\hat{t}^k$. Thus, for some $\underline{t} > 0$, $\hat{t}^k \geq \underline{t}$ for all $k \in K'$. Also, (3.2b)–(3.2c) yield, for all k,

$$\tilde{x}^k = (S^k)^{-1} X^k A^T \Delta y^k \tag{3.13}$$

which together with (3.2)(a) yields

$$b^T \Delta y^k = (\Delta y^k)^T A_{Q^k} X_{Q^k}^k (S_{Q^k}^k)^{-1} A_{Q^k}^T \Delta y^k = (\tilde{x}_{Q^k}^k)^T A_{Q^k}^T \Delta y^k \quad \forall k. \tag{3.14}$$

In view of Lemma 4, it follows that

$$b^T y^{k+1} = b^T (y^k + \hat{t}^k \Delta y^k) \geq b^T y^k + \underline{t} b^T \Delta y^k = b^T y^k + \underline{t} (A_{Q^k} \tilde{x}_{Q^k}^k)^T \Delta y^k \quad \forall k \in K'. \tag{3.15}$$

Now, since $v^* \neq 0$ and $|Q^*| \geq M$, it follows from Assumption 1 that $A_{Q^*}^T v^* \neq 0$. Further, taking limits in (3.13) as $k \to \infty$, $k \in K'$, we get

$$X_{Q^*}^* A_{Q^*}^T v^* - S_{Q^*}^* \tilde{x}_{Q^*}^* = 0.$$

Positivity of x_i^* and nonnegativity of s_i^* for all i then imply that $\left(\tilde{x}_{Q^*}^*\right)_i$ and $(A_{Q^*}^T v^*)_i$ have the same sign whenever the latter is nonzero, in which case the former is nonzero as well. It follows that $(\tilde{x}_{Q^*}^*)^T A_{Q^*}^T v^* > 0$. Thus there exists $\delta > 0$ such that $(A_{Q^*} \tilde{x}_{Q^*}^k)^T (\Delta y^k) > \delta$ for k large enough, $k \in K'$. Since, in view of Lemma 4, $\{b^T y^k\}$ is monotonic nondecreasing, it follows from (3.15) that $b^T y^k \to \infty$ as $k \to \infty$, a contradiction since y^k is bounded. ∎

LEMMA 8. *Suppose Assumptions* 1, 2, *and* 3 *hold. Suppose there exists an infinite index set K such that $\{y^k\}$ is bounded away from F^* on K. Then $\{\Delta y^k\}$ goes to zero on K.*

Proof. Let us again proceed by contradiction, i.e., suppose $\{\Delta y^k\}$ does not converge to zero as $k \to \infty$, $k \in K$. In view of Lemma 7, there exists an infinite index set $K' \subseteq K$ such that

$$\tilde{x}_-^{k-1} \to 0 \text{ as } k \to \infty, \ k \in K' \tag{3.16}$$

and

$$\Delta y^{k-1} \to 0 \text{ as } k \to \infty, \ k \in K'. \tag{3.17}$$

Further, since $\{y^k\}$ is bounded, and bounded away from F^*, there is no loss of generality in assuming that, for some $y^* \notin F^*$, $\{y^k\} \to y^*$ as $k \to \infty$, $k \in K'$.

Since $\|y^k - y^{k-1}\| = \|\hat{t}^{k-1}\Delta y^{k-1}\| \leq \|\Delta y^{k-1}\|$, it follows that $\{y^{k-1}\} \to y^*$ as $k \to \infty$, $k \in K'$ which implies, in view of (3.17) and of Lemma 6, that y^* is stationary and $\{\tilde{x}^{k-1}\} \to x^*$ as $k \to \infty$, $k \in K'$, where x^* is the corresponding multiplier vector. From (3.16) it follows that $x^* \geq 0$, thus that $y^* \in F^*$, a contradiction. ∎

LEMMA 9. *Suppose Assumptions* 1, 2, *and* 3 *hold. Then* $\{y^k\}$ *converges to the set of stationary points of* (1.3).

Proof. Suppose the claim does not hold. Because $\{y^k\}$ is bounded, there exist an infinite index set K and some nonstationary y^* such that $y^k \to y^*$ as $k \to \infty$, $k \in K$. By Lemma 6, $\{\Delta y^k\}$ does not converge to zero on K. This contradicts Lemma 8. ∎

We are now ready to embark on the final step in the global convergence analysis: prove convergence of $\{y^k\}$ to the *solution set* for (1.3). The key to this result is Lemma 11, which establishes that the multiplier vectors associated with all limit points of $\{y^k\}$ are the same. Thus, let

$$L := \left\{y \in \mathbb{R}^m : y \text{ is a limit point of } \{y^k\}\right\}.$$

(In view of Lemma 9, every $y \in L$ is a stationary point of (P).) The set L is bounded (since $\{y^k\}$ is bounded) and, as a limit set, it is closed, and thus compact. We first prove an auxiliary lemma.

LEMMA 10. *Suppose Assumptions* 1, 2, *and* 3 *hold. If* $\{y^k\}$ *is bounded away from* F^*, *then* L *is connected.*

Proof. Suppose the claim does not hold. Since L is compact, there must exist compact sets $E_1, E_2 \subset \mathbb{R}^n$, both nonempty, such that $L = E_1 \cup E_2$ and $E_1 \cap E_2 = \emptyset$. Thus $\delta := \min_{y \in E_1, y' \in E_2} \|y - y'\| > 0$. A simple contradiction argument based on the fact that $\{y^k\}$ is bounded shows that, for k large enough, $\min_{y \in L} \|y^k - y\| \leq \delta/3$, i.e., either $\min_{y \in E_1} \|y^k - y\| \leq \delta/3$ or $\min_{y \in E_2} \|y^k - y\| \leq \delta/3$. Moreover, since both E_1 and E_2 are nonempty (i.e., contain limit points of $\{y^k\}$), each of these situations occurs infinitely many times. Thus $K := \{k : \min_{y \in E_1} \|y^k - y\| \leq \delta/3, \min_{y \in E_2} \|y^{k+1} - y\| \leq \delta/3\}$ is an infinite index set and $\|\Delta y^k\| \geq \delta/3 > 0$ for all $k \in K$. In view of Lemma 8, this is a contradiction. ∎

LEMMA 11. *Suppose Assumptions* 1, 2, *and* 3 *hold. Suppose* $\{y^k\}$ *is bounded away from* F^*. *Let* $y, y' \in L$. *Let* x *and* x' *be the associated multiplier vectors. Then* $x = x'$.

Proof. Given any $y \in L$, let $x(y)$ be the multiplier vector associated with y, let $s(y) = c - A^T y$, and let $J(y)$ be the index set of "binding" constraints at y, i.e.,

$$J(y) = \{i \in \mathbf{n} : x_i(y) \neq 0\}.$$

We first show that, if $y, y' \in L$ are such that $J(y) = J(y')$, then $x(y) = x(y')$. Indeed, from (2.1b),

$$\sum_{j \in J(y)} x_j(y) a_j = b = \sum_{j \in J(y)} x_j(y') a_j,$$

and the claim follows from linear independence Assumption 3. To conclude the proof, we show that, for any $y, y' \in L$, $J(y) = J(y')$. Let $\tilde{y} \in L$ be arbitrary and let $E_1 := \{y \in L : J(y) = J(\tilde{y})\}$ and $E_2 := \{y \in L : J(y) \neq J(\tilde{y})\}$. We show that both E_1 and E_2 are closed. Let $\{\xi^\ell\} \subseteq L$ be a convergent sequence, say to $\hat{\xi}$, such that $J(\xi^\ell) = J$ for all ℓ, for some J. It follows from the first part of this proof that $x(\xi^\ell) = x$ for all ℓ for some x. Now, for all ℓ, $s_j(\xi^\ell) = 0$ for all j such that $x_j \neq 0$, so

that $s_j(\hat{\xi}) = 0$ for all j such that $x_j \neq 0$. Thus $J \subseteq I(\hat{\xi})$, and from linear independence Assumption 3 it follows that $x(\hat{\xi}) = x$ and thus $J(\hat{\xi}) = J$. Also, since L is closed, $\hat{\xi} \in L$. Thus, if $\{\xi^\ell\} \subseteq E_1$, then $\hat{\xi} \in E_1$ and, if $\{\xi^\ell\} \subseteq E_2$, then $\hat{\xi} \in E_2$, proving that both E_1 and E_2 are closed. Since E_1 is nonempty (it contains $\tilde{y}$), connectedness of L (Lemma 10) implies that E_2 is empty. Thus $J(y) = J(\tilde{y})$ for all $y \in L$, and the proof is complete. ∎

With all the tools in hand, we present the final theorem of this section. The essence of its proof is that if $\{y_k\}$ does not converge to F^*, complementary slackness will not be satisfied.

THEOREM 12. *Suppose Assumptions* 1, 2, *and* 3 *hold. Then* $\{y^k\}$ *converges to* F^*.

Proof. Proceeding again by contradiction, suppose that some limit point of $\{y^k\}$ is not in F^* and thus, since $y^k \in F$ for all k and since, in view of the monotonicity of $\{b^T y^k\}$ (Lemma 4), $b^T y^k$ takes on the same value at all limit points of $\{y^k\}$, that $\{y^k\}$ is bounded away from F^*. In view of Lemma 8, $\{\Delta y^k\} \to 0$. Let x^* be the common multiplier vector associated with all limit points of $\{y^k\}$ (see Lemma 11). A simple contradiction argument shows that Lemma 6 then implies that $\{\tilde{x}^k\} \to x^*$. Since $\{y^k\}$ is bounded away from F^*, $x^* \not\geq 0$. Let i_0 be such that $x^*_{i_0} < 0$. Then $\tilde{x}^k_{i_0} < 0$ for all k large enough. The definition of $\tilde{x}^k$ in Step 1 of Iteration rPDAS, together with (3.2c), then implies that $\Delta s^k_{i_0} > 0$ for k large enough, and it then follows from the update rule for s^k in Step 2(i) that, for k large enough,

$$0 < s^k_{i_0} < s^{k+1}_{i_0} < \cdots .$$

On the other hand, since $x^*_{i_0} < 0$, complementary slackness (3.9) implies that $(c - A^T\hat{y})_{i_0} = 0$ for all limit points $\hat{y}$ of $\{y^k\}$ and thus, since $\{y^k\}$ is bounded, $\{s^k_{i_0}\} \to 0$. This is a contradiction. ∎

3.2.2. Local rate of convergence. We prove q-quadratic convergence of the pair (x^k, y^k) (when $x_{\max}$ is large enough) under one additional assumption, which supersedes Assumption 2. (A sequence $\{z^k\}$ is said to converge q-quadratically to z^* if it converges to z^* and there exists a constant θ such that $\|z^{k+1} - z^*\| \leq \theta\|z^k - z^*\|^2$ for all k large enough.)

ASSUMPTION 4. *The dual solution set* F^* *is a singleton.*

Let y^* denote the unique solution to (1.3), i.e., $F^* = \{y^*\}$, let $s^* := c - A^T y^*$, and let x^* be the corresponding multiplier vector (unique in view of Assumption 3). Of course, under Assumptions 1, 3, and 4, it follows from Theorem 12 that $\{y^k\} \to y^*$ as $k \to \infty$. Further, under Assumption 3, Assumption 4 implies that strict complementarity holds, i.e.,

$$x^*_i > 0 \quad \forall i \in I(y^*). \tag{3.18}$$

Moreover, Assumption 4 implies that

$$\operatorname{span}(\{a_i : i \in I(y^*)\}) = \mathbb{R}^m. \tag{3.19}$$

LEMMA 13. *Suppose Assumptions* 1, 3, *and* 4 *hold and let* $Q \supseteq I(y^*)$ *and* $Q^{\mathrm{c}} = \mathbf{n} \setminus Q$. *Then* $J_{\mathrm{a}}(A_Q, x^*_Q, s^*_Q)$ *and* $J(A_Q, x^*_Q, s^*_Q)$ *are nonsingular and* $A^T_{Q^{\mathrm{c}}} y^* < c_{Q^{\mathrm{c}}}$.

Proof. The last claim is immediate. Since $s^* = c - A^T y^*$, it follows from linear independence Assumption 3 that $\{a_i : s_i = 0\}$ is linear independent, hence its subset

retaining only the columns whose indices are in Q is also linear independent. The first two claims now follow directly from Lemma 1. ∎

The following preliminary result is inspired from [PTH88, Proposition 4.2].

LEMMA 14. *Suppose Assumptions* 1, 3, *and* 4 *hold. Then* (i) $\{\Delta y^k\} \to 0$; (ii) $\{\tilde{x}^k\} \to x^*$; *and* (iii) *if* $x_i^* \leq x_{\max}$ *for all* $i \in \mathbf{n}$, *then* $\{x^k\} \to x^*$.

Proof. To prove Claim (i), proceed by contradiction. Specifically, suppose that, for some infinite index set K, $\inf_{k\in K}\|\Delta y^k\| > 0$. Without loss of generality, assume that, for some Q^*, $Q^k = Q^*$ for all $k \in K$. Since $Q^* \in \mathcal{Q}_M(y^k)$ for all $k \in K$, since $\{y^k\} \to y^*$ as $k \to \infty$, and since, in view of Assumption 3, $|I(y^*)| \leq m \leq M$, it must hold that $Q^* \supseteq I(y^*)$. On the other hand, Lemma 7 implies that there exists an infinite index set $K' \subseteq K$ such that $\{\Delta y^{k-1}\}_{k\in K'}$ and $\{\tilde{x}_-^{k-1}\}_{k\in K'}$ go to zero. In view of Lemma 6 it follows that $\{\tilde{x}^{k-1}\}_{k\in K'} \to x^*$. It then follows from (3.5) that, for all i, $\{x_i^k\}_{k\in K'} \to \xi_i^* := \min\{x_i^*, x_{\max}\}$. Since $Q^* \supseteq I(y^*)$, it follows from Lemma 1, Assumption 3, (3.18), and (3.19) that $J(A_{Q^*}, \xi_{Q^*}^*, s_{Q^*}^*)$ is nonsingular. Now note that, in view of (3.7), it holds that (see (2.5))

$$J(A_{Q^k}, x_{Q^k}^k, s_{Q^k}^k) \begin{bmatrix} \tilde{x}_{Q^k}^k \\ \Delta y^k \\ \Delta s_{Q^k}^k \end{bmatrix} = \begin{bmatrix} 0 \\ b \\ 0 \end{bmatrix} \quad \forall k \in K'. \tag{3.20}$$

On the other hand, by feasibility of x^* and complementarity slackness,

$$J(A_{Q^*}, \xi_{Q^*}^*, s_{Q^*}^*) \begin{bmatrix} x_{Q^*}^* \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ b \\ 0 \end{bmatrix}. \tag{3.21}$$

From (3.20) and (3.21), nonsingularity of $J(A_{Q^*}, \xi_{Q^*}^*, s_{Q^*}^*)$ thus implies that $\{\Delta y^k\} \to 0$ as $k \to \infty$, $k \in K'$, a contradiction since $K' \subseteq K$. Claim (i) is thus proved. Claim (ii) then directly follows from Lemma 6 and Claim (iii) follows from (3.5). ∎

To prove q-quadratic convergence of $\{(y^k, x^k)\}$, the following property of Newton's method will be used. It is borrowed from [TZ94, Proposition 3.10].

PROPOSITION 15. *Let* $\Phi : \mathbb{R}^n \to \mathbb{R}^n$ *be twice continuously differentiable and let* $\hat{z} \in \mathbb{R}^n$ *be such that* $\Phi(\hat{z}) = 0$ *and* $\frac{\partial\Phi}{\partial z}(\hat{z})$ *is nonsingular. Let* $\rho > 0$ *be such that* $\frac{\partial\Phi}{\partial z}(z)$ *is nonsingular whenever* $z \in B(\hat{z}, \rho) := \{z : \|z - \hat{z}\| \leq \rho\}$. *Let* $d^{\mathrm{N}} : B(\hat{z}, \rho) \to \mathbb{R}^n$ *be the Newton increment* $d^{\mathrm{N}}(z) := -\left(\frac{\partial\Phi}{\partial z}(z)\right)^{-1}\Phi(z)$. *Then given any* $c_1 > 0$ *there exists* $c_2 > 0$ *such that the following statement holds:*

For all $z \in B(\hat{z}, \rho)$ *and* $z^+ \in \mathcal{R}^n$ *such that, for each* $i \in \{1, \ldots, n\}$, *either*

(i) $|z_i^+ - \hat{z}_i| \leq c_1\|d^{\mathrm{N}}(z)\|^2$

or

(ii) $|z_i^+ - (z_i + d_i^{\mathrm{N}}(z))| \leq c_1\|d^{\mathrm{N}}(z)\|^2$,

it holds that

$$\|z^+ - \hat{z}\| \leq c_2\|z - \hat{z}\|^2. \tag{3.22}$$

We will apply this proposition to the equality portion of the KKT conditions (2.1) (with $\mu = 0$). Eliminating s from this system of equations yields $\Phi(x, y) = 0$, with Φ given by

$$\Phi(x, y) := \begin{bmatrix} Ax - b \\ X(c - A^T y) \end{bmatrix}. \tag{3.23}$$

It is readily verified that (2.3a), with $\mu = 0$, $r_c = 0$, and s replaced by $c - A^T y$, is the Newton iteration for the solution of $\Phi(x, y) = 0$. In particular, $J_{\mathrm{a}}(A, x, c - A^T y)$ is the Jacobian of $\Phi(x, y)$.

Of course, Iteration rPDAS does not make use of the Newton direction for Φ, since it is based on a reduced set of constraints. Lemma 16 below relates the direction computed in Step 1 of Iteration rPDAS to the Newton direction.

In what follows, we use z (possibly with subscripts, superscripts, or diacritical signs) to denote (x, y) (with the same subscripts, superscripts, or diacritical signs on both x and y). Also, let

$$G^o := \{z : x > 0, y \in F^{\mathrm{o}}\}$$

and, given $z \in G^o$ and $Q \in \mathcal{Q}_M(y)$, let $\Delta x(z, Q)$, $\Delta y(z, Q)$, $x^+(z, Q)$, $y^+(z, Q)$, $\tilde{x}(z, Q)$, $\bar{t}(z, Q)$, and $\hat{t}(z, Q)$ denote the quantities defined by Iteration rPDAS, and let $\Delta z(z, Q) := (\Delta x(z, Q), \Delta y(z, Q))$ and $z^+(z, Q) := (x^+(z, Q), y^+(z, Q))$. Further, let $Q^{\mathrm{c}} := \mathbf{n} \setminus Q$, let $d^{\mathrm{N}}(z) := \Delta z(z, \mathbf{n})$ denote the Newton increment for Φ, and, given $\rho > 0$, let $B(z^*, \rho) := \{z : \|z - z^*\| \le \rho\}$. The following lemma was inspired from an idea of O'Leary [O'L04]. (Existence of $\rho > 0$ follows from Lemma 13.)

LEMMA 16. *Suppose Assumptions* 1, 3, *and* 4 *hold. Let $\rho > 0$ be such that, for all $(x, y) \in B(z^*, \rho) \cap G^o$ and for all $Q \in \mathcal{Q}_M(y)$, $J_{\mathrm{a}}(A_Q, x_Q, c_Q - A_Q^T y)$ is nonsingular and $A_{Q^{\mathrm{c}}}^T y < c_{Q^{\mathrm{c}}}$. Then there exists $\gamma > 0$ such that, for all $z \in B(z^*, \rho) \cap G^o$, $Q \in \mathcal{Q}_M(y)$,*

$$\|\Delta z(z, Q) - d^{\mathrm{N}}(z)\| \le \gamma \|z - z^*\| \cdot \|d^{\mathrm{N}}(z)\|.$$

Proof. Let $z \in B(z^*, \rho) \cap G^o$, $Q \in \mathcal{Q}_M(y)$ and let $s := c - A^T y$. Then, $\Delta y(z, Q)$ and $\Delta x_Q(z, Q)$ satisfy (direct consequence of (3.7))

$$\begin{bmatrix} 0 & A_Q \\ X_Q A_Q^T & -S_Q \end{bmatrix} \begin{bmatrix} \Delta y(z, Q) \\ \Delta x_Q(z, Q) \end{bmatrix} = \begin{bmatrix} b - A_Q x_Q \\ X_Q s_Q \end{bmatrix}. \tag{3.24}$$

On the other hand, $\Delta y(z, \mathbf{n})$ and $\Delta x(z, \mathbf{n})$ satisfy (see (2.3a), with $\mu = 0$ and $r_c = 0$)

$$\begin{bmatrix} 0 & A \\ X A^T & -S \end{bmatrix} \begin{bmatrix} \Delta y(z, \mathbf{n}) \\ \Delta x(z, \mathbf{n}) \end{bmatrix} = \begin{bmatrix} b - Ax \\ Xs \end{bmatrix},$$

and eliminating $\Delta x_{Q^{\mathrm{c}}}(z, \mathbf{n})$ in the latter yields

$$\begin{bmatrix} A_{Q^{\mathrm{c}}} S_{Q^{\mathrm{c}}}^{-1} X_{Q^{\mathrm{c}}} A_{Q^{\mathrm{c}}}^T & A_Q \\ X_Q A_Q^T & -S_Q \end{bmatrix} \begin{bmatrix} \Delta y(z, \mathbf{n}) \\ \Delta x_Q(z, \mathbf{n}) \end{bmatrix} = \begin{bmatrix} b - A_Q x_Q \\ X_Q s_Q \end{bmatrix}. \tag{3.25}$$

Equating the left-hand sides of (3.24) and (3.25) yields

$$\begin{bmatrix} \Delta y(z, Q) \\ \Delta x_Q(z, Q) \end{bmatrix} = J_{\mathrm{a}}(A_Q, x_Q, c_Q - A_Q^T y)^{-1} \begin{bmatrix} A_{Q^{\mathrm{c}}} S_{Q^{\mathrm{c}}}^{-1} X_{Q^{\mathrm{c}}} A_{Q^{\mathrm{c}}}^T & A_Q \\ X_Q A_Q^T & -S_Q \end{bmatrix} \begin{bmatrix} \Delta y(z, \mathbf{n}) \\ \Delta x_Q(z, \mathbf{n}) \end{bmatrix}. \tag{3.26}$$

Expressing $\begin{bmatrix} \Delta y(z, \mathbf{n}) \\ \Delta x_Q(z, \mathbf{n}) \end{bmatrix}$ as $J_{\mathrm{a}}(A_Q, x_Q, c_Q - A_Q^T y)^{-1} J_{\mathrm{a}}(A_Q, x_Q, c_Q - A_Q^T y) \begin{bmatrix} \Delta y(z, \mathbf{n}) \\ \Delta x_Q(z, \mathbf{n}) \end{bmatrix}$ and subtracting from (3.26) then yields

$$\begin{bmatrix} \Delta y(z, Q) \\ \Delta x_Q(z, Q) \end{bmatrix} - \begin{bmatrix} \Delta y(z, \mathbf{n}) \\ \Delta x_Q(z, \mathbf{n}) \end{bmatrix}$$

$$= J_{\mathrm{a}}(A_Q, x_Q, c_Q - A_Q^T y)^{-1} \begin{bmatrix} A_{Q^{\mathrm{c}}} S_{Q^{\mathrm{c}}}^{-1} X_{Q^{\mathrm{c}}} A_{Q^{\mathrm{c}}}^T & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta y(z, \mathbf{n}) \\ \Delta x_Q(z, \mathbf{n}) \end{bmatrix}.$$

Further, it follows from (3.1b) and (3.1c) and from (3.2b) and (3.2c) that

$$\Delta x_{Q^{\mathrm{c}}}(z, Q) - \Delta x_{Q^{\mathrm{c}}}(z, \mathbf{n}) = S_{Q^{\mathrm{c}}}^{-1} X_{Q^{\mathrm{c}}} A_{Q^{\mathrm{c}}}^T (\Delta y(z, Q) - \Delta y(z, \mathbf{n})).$$

Since $S_{Q^{\mathrm{c}}}$ and $J_{\mathrm{a}}(A_Q, x_Q, c_Q - A_Q^T y)$ are continuous and nonsingular over the closed ball $B(z^*, \rho)$, and since $\|X_{Q^{\mathrm{c}}}\| = \|X_{Q^{\mathrm{c}}} - X_{Q^{\mathrm{c}}}^*\| \leq \|z - z^*\|$ (since $c_{Q^{\mathrm{c}}} - A_{Q^{\mathrm{c}}}^T z^* > 0$, i.e., $I(y^*) \cap Q^{\mathrm{c}}$ is empty), in view of the fact that $\{Q^{\mathrm{c}} : Q \in \mathcal{Q}_M(y)\}$ is finite (since $\mathcal{Q}_M(y)$ is) the claim follows. ∎

We are now ready to prove q-quadratic convergence.

THEOREM 17. *Suppose Assumptions* 1, 3, *and* 4 *hold. If* $x_i^* < x_{\max}$ *for all* $i \in \mathbf{n}$, *then* $\{(x^k, y^k)\}$ *converges to* (x^*, y^*) *q-quadratically.*

Proof. We aim at establishing that the conditions in Proposition 15 hold for Φ given by (3.23) and with $\hat{z} := z^*(= (x^*, y^*))$, $z^+ := z^+(z, Q)$, $Q \in \mathcal{Q}_M(y)$, and $\rho > 0$ small enough. First, note that

$$\frac{\partial \Phi}{\partial z}(z) = J_{\mathrm{a}}(A, x, c - A^T y),$$

so, in view of (3.18), (3.19), and linear independence Assumption 3, it follows from Lemma 1 that $\frac{\partial \Phi}{\partial z}(z^*)$ is nonsingular. Next, let $i \in I(y^*)$ and consider Step 2(ii) in Iteration rPDAS. Let $Q \supseteq I(y^*)$. From Lemma 13, we know that $J(A_Q, x_Q^*, s_Q^*)$ is nonsingular. Since, by feasibility of x^* and complementarity slackness,

$$J(A_Q, x_Q^*, s_Q^*) \begin{bmatrix} x_Q^* \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ b \\ 0 \end{bmatrix},$$

it follows from (3.7), (3.2), and the fact that $s_j := c_j - a_j^T y$ is bounded away from zero in a neighborhood of z^* for $j \notin Q$ (Lemma 13), that

$$\Delta y(z, Q) \to 0 \text{ as } z \to z^* \tag{3.27}$$

and

$$\tilde{x}(z, Q) \to x^* \text{ as } z \to z^*. \tag{3.28}$$

Note that, for $z \in G^o$ close enough to z^*, $Q \supseteq I(y^*)$ for all $Q \in \mathcal{Q}_M(y)$. Since $x_i^* > 0$ (from (3.18)) and $x^* \geq 0$, it follows that for $z \in G^o$ close enough to z^*,

$$\|\Delta y(z, Q)\|^2 + \|\tilde{x}_-(z, Q)\|^2 < \tilde{x}_i(z, Q) \quad \forall Q \in \mathcal{Q}_M(y),$$

which, in view of the update rule for x in Step 2(ii) of Iteration rPDAS, since $x_i^* < x_{\max}$, implies that, for $z \in G^o$ close enough to z^*,

$$x_i^+(z, Q) = \tilde{x}_i(z, Q) = x_i + \Delta x_i(z, Q) \quad \forall Q \in \mathcal{Q}_M(y),$$

yielding, for $z \in G^o$ close enough to z^*,

$$x_i^+(z, Q) - (x_i + \Delta x_i(z, \mathbf{n})) = \Delta x_i(z, Q) - \Delta x_i(z, \mathbf{n}) \quad \forall Q \in \mathcal{Q}_M(y). \tag{3.29}$$

In view of Lemma 16, it follows that, for z close enough to z^*, $z \in G^o$,

$$|x_i^+(z,Q) - (x_i + \Delta x_i(z,\mathbf{n}))| \leq \gamma \|z - z^*\| \cdot \|d^{\mathrm{N}}(z)\| \quad \forall Q \in \mathcal{Q}_M(y). \tag{3.30}$$

Next, let $i \notin I(y^*)$, so that $x_i^* = 0$, and again consider Step 2(ii) in Iteration rPDAS. Then for every $z \in G^o$ close enough to z^*, and every $Q \in \mathcal{Q}_M(y)$, either again

$$x_i^+(z,Q) = x_i + \Delta x_i(z,Q),$$

yielding again (3.29) and (3.30), or

$$x_i^+(z,Q) = \|\Delta y(z,Q)\|^2 + \|\tilde{x}_-(z,Q)\|^2, \leq \|\Delta z(z,Q)\|^2$$

yielding, since $x_i^* = 0$,

(3.31)
$$\|x_i^+(z,Q) - x_i^*\| \leq \|\Delta z(z,Q) - d^{\mathrm{N}}(z) + d^{\mathrm{N}}(z)\|^2 \leq (\gamma\|z - z^*\| \cdot \|d^{\mathrm{N}}(z)\| + \|d^{\mathrm{N}}(z)\|)^2$$

for every $z \in G^o$ close enough to z^*, $Q \in \mathcal{Q}_M(y)$. Finally, consider the "y" components of z. From (3.2b)–(3.2c) we know that, for $z \in G^o$ close enough to z^*, for all $Q \in \mathcal{Q}_M(y)$, and for all i such that $\Delta s_i(z,Q) := -a_i^T \Delta y(z,Q) \neq 0$,

$$\frac{s_i}{a_i^T \Delta y(z,Q)} = -\frac{s_i}{\Delta s_i(z,Q)} = \frac{x_i}{\tilde{x}_i(z,Q)}.$$

In view of Lemma 14(i), we conclude that, for $i \notin I(y^*)$,

$$\frac{|x_i|}{|\tilde{x}_i(z,Q)|} \to \infty \quad \text{as } z \to z^*, Q \in \mathcal{Q}_M(y).$$

Step 2(i) in Iteration rPDAS then yields

$$\bar{t}(z,Q) = \min\left\{ \frac{x_i}{\tilde{x}_i(z,Q)} : \ i \in I(y^*) \right\}$$

for $z \in G^o$ close enough to z^*, $Q \in \mathcal{Q}_M(y)$. Step 2(i) in Iteration rPDAS further yields, for $z \in G^o$ close enough to z^* (using (3.27)), $Q \in \mathcal{Q}_M(y)$,

$$\hat{t}(z,Q) = \min\left\{1, \frac{x_{i(z,Q)}}{\tilde{x}_{i(z,Q)}(z,Q)} - \|\Delta y(z,Q)\|\right\},$$

for some $i(z,Q) \in I(y^*)$. (Nonemptiness of $I(y^*)$ is insured by Assumption 4.) Thus, for $z \in G^o$ close enough to z^*, $Q \in \mathcal{Q}_M(y)$, and some $i(z,Q) \in I(y^*)$,

$$\begin{aligned}\|y^+(z,Q) - (y + \Delta y(z,Q))\| &= |\hat{t}(z,Q) - 1| \|\Delta y(z,Q)\| \\ &\leq \left| \ \|\Delta y(z,Q)\| + \frac{\tilde{x}_{i(z,Q)}(z,Q) - x_{i(z,Q)}}{\tilde{x}_{i(z,Q)}(z,Q)} \right| \ \|\Delta y(z,Q)\|.\end{aligned}$$

Since $x_i^* > 0$ for all $i \in I(y^*)$, it follows that for some $c_0 > 0$ and $z \in G^o$ close enough to z^*,

$$\begin{aligned}\|y^+(z,Q) - (y + \Delta y(z,Q))\| &\leq (\|\Delta y(z,Q)\| + c_0\|\Delta x(z,Q)\|)\|\Delta y(z,Q)\| \\ &\leq (1 + c_0)\|\Delta z(z,Q)\|^2 \\ &\leq (1 + c_0)(\|\Delta z(z,Q) - d^{\mathrm{N}}(z)\| + \|d^{\mathrm{N}}(z)\|)^2 \\ &\leq (1 + c_0)(\gamma\|z - z^*\| \cdot \|d^{\mathrm{N}}(z)\| + \|d^{\mathrm{N}}(z)\|)^2\end{aligned}$$

for all $Q \in \mathcal{Q}_M(y)$. It follows from Lemma 16 that

$$\|y^+(z,Q) - (y + \Delta y(z,\mathbf{n}))\| \le (1+c_0)(\gamma\|z-z^*\| \cdot \|d^{\mathrm{N}}(z)\| + \|d^{\mathrm{N}}(z)\|)^2$$

$$+\gamma\|z-z^*\| \cdot \|d^{\mathrm{N}}(z)\|$$

$$\le c_1 \max\{\|d^{\mathrm{N}}(z)\|^2, \|z-z^*\|^2\} \tag{3.32}$$

for some $c_1 > 0$ independent of z for all $z \in G^o$ close enough to z^*, $Q \in \mathcal{Q}_M(y)$. Equations (3.30), (3.31), and (3.32) are the key to the completion of the proof.

For $z \in G^o$ (close enough to z^*) such that $\|z-z^*\| \le \|d^{\mathrm{N}}(z)\|$, in view of Proposition 15, (3.30), (3.31), and (3.32) imply that, for some $c_2 > 0$ independent of z, and for all $Q \in \mathcal{Q}_M(y)$,

$$\|z^+(z,Q) - z^*\| \le c_2\|z-z^*\|^2.$$

On the other hand, for $z \in G^o$ close enough to z^* such that $\|d^{\mathrm{N}}(z)\| < \|z-z^*\|$, (3.30) yields, for some $c_3 > 0$ independent of z and for all $Q \in \mathcal{Q}_M(y)$,

$$|x_i^+(z,Q) - x_i^*| \le \gamma\|z-z^*\| \cdot \|d^{\mathrm{N}}(z)\| + \|x_i + \Delta x_i(z,\mathbf{n}) - x_i^*\| \le c_3\|z-z^*\|^2$$

for all $i \in I(y^*)$, where we have invoked quadratic convergence of the Newton iteration; (3.31) yields, for some $c_4 > 0$ independent of z,

$$|x_i^+(z,Q) - x_i^*| \le c_4\|z-z^*\|^2$$

for all $i \notin I(y^*)$; and (3.32) yields, for some $c_5 > 0$ independent of z,

$$\|y^+(z,Q) - y^*\| \le c_1\|z-z^*\|^2 + \|y_i + \Delta y(z,\mathbf{n}) - y^*\| \le c_5\|z-z^*\|^2.$$

In particular, they together imply again that, for all $z \in G^o$ close enough to z^*, $Q \in \mathcal{Q}_M(y)$,

$$\|z^+(z,Q) - z^*\| \le c_2\|z-z^*\|^2$$

for some $c_2 > 0$ independent of z. Since, in view of Lemma 14, z^k converges to z^*, it follows that z^k converges to z^* q-quadratically. □

3.3. Numerical results. Algorithm rPDAS was implemented in MATLAB and run on an Intel(R) Pentium(R) III CPU 733MHz machine with 256 KB cache, 512 MB RAM, Linux kernel 2.6.11, and MATLAB 7 (R14).[3]

Parameters were chosen as $\beta := 0.99$, $x_{\max} := 10^{15}$, $\underline{x} := 10^{-4}$. The code was supplied with strictly feasible initial dual points (i.e., $y^0 \in F^{\mathrm{o}}$). The initial primal vector, x^0, was chosen using the heuristic in [Meh92, p. 589] modified to accommodate dual feasibility. Specifically, $x^0 := \hat{x}^0 + \hat{\delta}_x$, where $\hat{x}^0$ is the minimum norm solution of $Ax = b$, $\hat{\delta}_x := \delta_x + (\hat{x}^0 + \delta_x e)^T s^0/(2\sum_{i=1}^n s_i^0)$, $\delta_x := \max\{-1.5 \cdot \min(x), 0\}$, where $s^0 := x - A^T y^0$ and e is the vector of all ones. The "M most active" heuristic (Q consists of the indexes of the M leftmost components of $c - A^T y$) was used to select the index set Q. The code uses MATLAB's "Cholesky-Infinity" factorization (`cholinc` function) to solve the normal equations (3.2a). A safeguard $s_i := \max\{10^{-14}, s_i\}$ was

[3]The code is available from the authors.

applied before Step 1; this prevents the matrix $A_Q S_Q^{-1} X_Q A_Q^T$ in (3.2a) from being excessively ill-conditioned and avoids inaccuracies in the ratios $s_i/\Delta s_i$ involved in (3.3) that could lead to unnecessarily small steps. We used a stopping criterion, adapted from [Meh92, p. 592], based on the error in the primal-dual equalities (2.1b)–(2.1a) and the duality gap. Specifically, convergence was declared when

$$\frac{\|b - Ax\|}{1 + \|x\|} + \frac{\|c - A^T y - s\|}{1 + \|s\|} + \frac{|c^T x - b^T y|}{1 + |b^T y|} < tol,$$

where *tol* was set to 10^{-8}. Notice that, in the case of iteration rPDAS, $\|c - A^T y - s\|$ vanishes throughout, up to numerical errors.

Execution times strongly depend on how the computation of $H^Q := A_Q D_Q^2 A_Q^T$ (where $D_Q^2 := S_Q^{-1} X_Q$ is diagonal), involved in (3.2a), is implemented in MATLAB. Storing the $|Q| \times |Q|$ matrix D_Q^2 as a full matrix is inefficient, or even impossible (even when $|Q|$ is much smaller than n, it may still be large). We are left with two options: Either (i) compute an auxiliary matrix $D_Q^2 A_Q^T$ using a "for" loop over the $|Q|$ rows of A_Q^T, then compute $A_Q * D_Q^2 A_Q^T$, or (ii) create a sparse matrix D_Q^2 using the `spdiags` function, then compute $A_Q * D_Q^2 * A_Q^T$. If the matrix A is in sparse form with sufficiently few nonzero elements, then the second option is (much) more efficient. If the matrix A is dense, then both options have comparable speed. Consequently, we used the second option in all experiments.

Numerical results obtained with algorithm rPDAS on several types of problems (to be discussed below) are presented in Figures 1 through 4. The points on the plots, as well as those on the plots of Figures 5 through 12 discussed in section 4, correspond to different runs on the same problem. The runs differ only by the number of constraints M that are retained in Q; this information is indicated on the horizontal axis in relative value. The rightmost point thus corresponds to the experiment without constraint reduction, while the points on the extreme left correspond to the most drastic constraint reduction. The plots are built as follows: the execution script picks progressively smaller values of M from a predefined list of values until it reaches the end of the list, or early, abnormal termination occurs. The latter was always caused by either (i) the number of iterations reaching a predefined limit of 100 (this is an ad hoc choice to stop the execution script when it reaches values of M where the number of iterations become high), or (ii) MATLAB generating `NaN` values, which happens when the normal matrix becomes numerically singular. Abnormal termination did occur in the numerical experiment presented in Figure 3 due to (ii) and in a few other instances due to (i).

In the lower plot of each figure, the vertical axis indicates CPU times to solution (total time, as well as time expended in the computation of H^Q and time used for the solution of the normal equation) as returned by the MATLAB function `cputime`. We emphasize that these results are valid only for the specific MATLAB implementation described above. Results could vary widely, depending on the programming language, the possible use of the BLAS, and the hardware. In contrast, the number of iterations shown on the upper plot has more meaning.

The first test problem is of the finely discretized semi-infinite type: the dual feasible set F is a polytope whose faces are tangent to the unit sphere. Contact points on the sphere were selected from the uniform distribution by first generating vectors of numbers distributed according to $\mathcal{N}(0, 1)$—normal distribution with mean zero and standard deviation one—and then normalizing these vectors. These points form the columns of A, and c was selected as the vector of all ones. Each entry of the objective vector, b, was chosen from $\mathcal{N}(0, 1)$, and y^0 was selected as the zero vector.

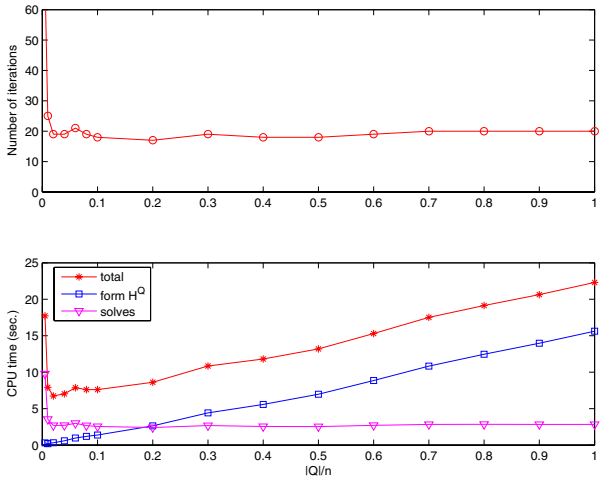


FIG. 1. *rPDAS on the problem with constraints tangent to the unit sphere.*

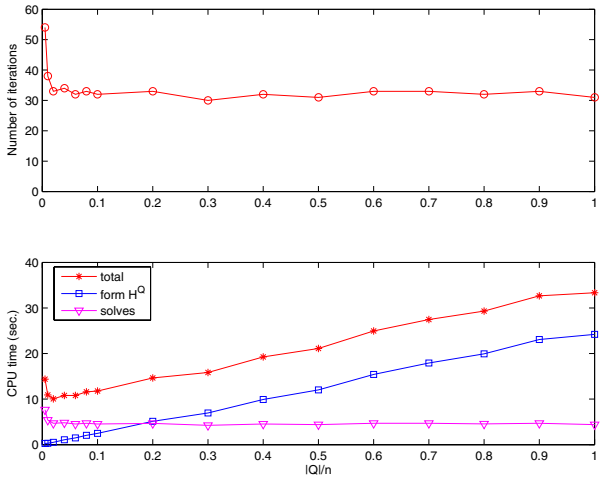


FIG. 2. *rPDAS on the "fully random" problem.*

This yields a problem that lends itself nicely to constraint reduction, since $n \gg m$ (we chose $m = 50$ and $n = 20000$), A is dense, and Assumption 1 on the full rank of submatrices of A holds for M as low as m. Numerical results are presented in Figure 1.

Arguably the most remarkable result in this paper is that observed on the upper plot of Figure 1 (and again in other figures discussed below): the number of iterations shows little variation over a significant range of values of $|Q|$. We tested the algorithm on several problems randomly generated, as explained above, and always observed that only very low values of $|Q|$ produce a significant increase in the number of iterations.

The second test problem is "fully random." The entries of A and b were generated from $\mathcal{N}(0, 1)$. To ensure a dual-feasible initial point, y^0 and s^0 were chosen from a uniform distribution on $(0, 1)$ and the vector c was generated by taking $c := A^T y^0 + s^0$. We again chose $m = 50$ and $n = 20000$. Results are displayed in Figure 2.

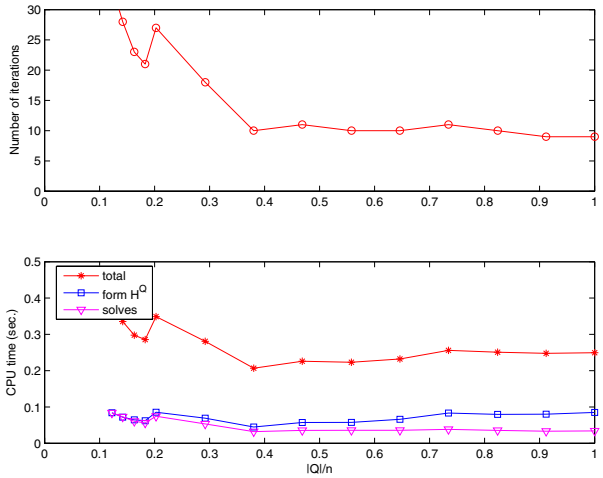


FIG. 3. *rPDAS on SCSD*1.

Note that these results are qualitatively similar to those of Figure 1. Here again, the number of iterations is stable over a wide range of values of $|Q|$. Experiments conducted on other test problems drawn from the same distribution produced similar results.

Next, we searched the Netlib LP library for problems where n is significantly greater than m and Assumption 1 is satisfied for reasonably small M. This left us with the SCSD problems. These problems, however, are very sparse. The computation of the normal matrix AD^2A^T involves only sparse matrix multiplications that can be performed efficiently and account only for a small portion of the total execution time. Therefore, the constraint reduction strategy, which focuses on reducing the cost of forming the normal matrix, has little effect on the overall execution time. (If the computation of H^Q is done with a `for` loop as explained above, then an important speedup is observed.) We tested algorithm rPDAS on SCSD1 ($m = 77$, $n = 760$) and SCSD6 ($m = 147$, $n = 1350$). For both problems, we set y^0 to $0 \in F^o$. Results are displayed in Figures 3 and 4. Here again, the number of iterations is quite stable over a wide range of values of $|Q|$.

4. A reduced MPC algorithm.

4.1. Algorithm statement. We consider a constraint-reduced version of Mehrotra's predictor-corrector (MPC) method [Meh92]—or rather of the simplified version of that algorithm found in [Wri97].

Iteration rMPC.

Parameters. $\beta \in (0,1)$, integer M satisfying $m \le M \le n$.

Data. $y \in \mathbb{R}^m$, $s > 0$, $x > 0$, $Q \in \mathcal{Q}_M(y)$, $\mu := x^T s/n$.

Step 1. Compute affine scaling direction:

Solve

$$A_Q S_Q^{-1} X_Q A_Q^T \Delta y = -r_b + A(-S^{-1} X r_c + x)$$

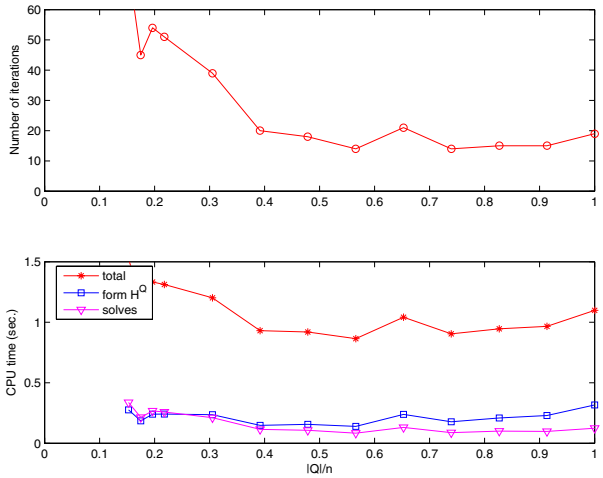


FIG. 4. *rPDAS on SCSD6.*

and compute

$$\Delta s := -A^T \Delta y - r_c,$$
$$\Delta x := -x - S^{-1} X \Delta s,$$

and let

$$t_{\text{aff}}^{\text{pri}} := \arg\max\{t \in [0,1] \mid x + t\Delta x \geq 0\},$$
$$t_{\text{aff}}^{\text{dual}} := \arg\max\{t \in [0,1] \mid s + t\Delta s \geq 0\}.$$

Step 2. Compute centering parameter:

$$\mu_{\text{aff}} := (x + t_{\text{aff}}^{\text{pri}} \Delta x)^T (s + t_{\text{aff}}^{\text{dual}} \Delta s)/n,$$
$$\sigma := (\mu_{\text{aff}}/\mu)^3.$$

Step 3. Compute centering/corrector direction:

Solve

$$A_Q S_Q^{-1} X_Q A_Q^T \Delta y^{\text{cc}} = -AS^{-1}(\sigma\mu e - \Delta X \Delta s)$$

and compute

$$\Delta s^{\text{cc}} := -A^T \Delta y^{\text{cc}},$$
$$\Delta x^{\text{cc}} := S^{-1}(\sigma\mu e - \Delta X \Delta s) - S^{-1} X \Delta s^{\text{cc}}.$$

Step 4. Compute MPC step:

$$\Delta x^{\text{mpc}} := \Delta x + \Delta x^{\text{cc}},$$
$$\Delta y^{\text{mpc}} := \Delta y + \Delta y^{\text{cc}},$$
$$\Delta s^{\text{mpc}} := \Delta s + \Delta s^{\text{cc}},$$

$$t^{\rm pri}_{\rm max} := \arg\max\{t \in [0,1] \mid x + t\Delta x^{\rm mpc} \geq 0\},$$
$$t^{\rm dual}_{\rm max} := \arg\max\{t \in [0,1] \mid s + t\Delta s^{\rm mpc} \geq 0\},$$

$$t^{\rm pri} := \min\{\beta t^{\rm pri}_{\rm max},\ 1\},$$
$$t^{\rm dual} := \min\{\beta t^{\rm dual}_{\rm max},\ 1\}.$$

Step 5. Updates:

$$x^+ := x + t^{\rm pri}\Delta x^{\rm mpc},$$
$$y^+ := y + t^{\rm dual}\Delta y^{\rm mpc},$$
$$s^+ := s + t^{\rm dual}\Delta s^{\rm mpc}.$$

Pick $Q^+ \in \mathcal{Q}_M(y^+)$.

As compared with the case of Iteration rPDAS, the speed-up per iteration achieved by rMPC over MPC is not as striking. This is due to the presence of two additional matrix-vector products in the iteration (see Step 3) for a total of three matrix-vector products per iteration. Further, these products involve the full A matrix and require $O(mn)$ flops, which can be substantial.

4.2. Numerical results: Dual-feasible initial point. We report on numerical results obtained with a MATLAB implementation of the reduced MPC (rMPC) algorithm.[4] The hardware, software, test problems, initial points, and presentation of the results are the same as in section 3.3. Figures 5, 6, 7, and 8 are the counterparts of Figures 1, 2, 3, and 4, respectively.

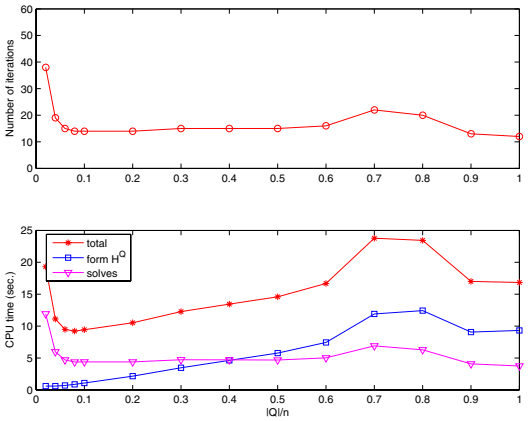


FIG. 5. *rMPC on the problem with constraints tangent to the unit sphere, with dual-feasible initial point.*

4.3. Numerical results: Infeasible initial point. We now report on numerical experiments that differ from the ones in section 4.2 only by the choice of the initial

[4]The code is available from the authors.

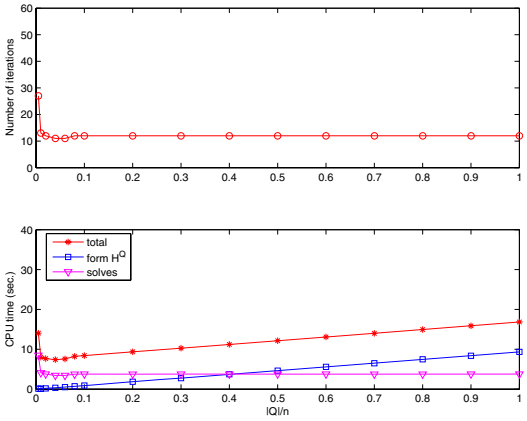


FIG. 6. *rMPC on the "fully random" problem, with dual-feasible initial point.*

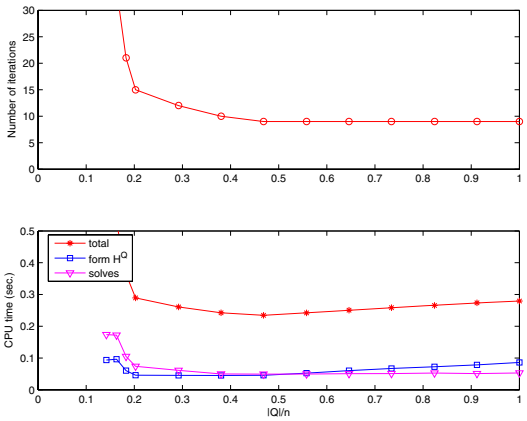


FIG. 7. *rMPC on SCSD*1*, with dual-feasible initial point.*

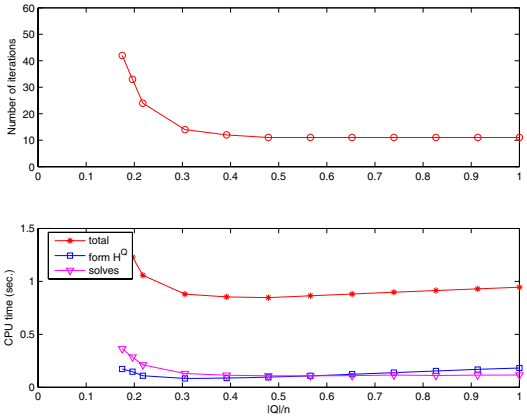


FIG. 8. *rMPC on SCSD*6*, with dual-feasible initial point.*

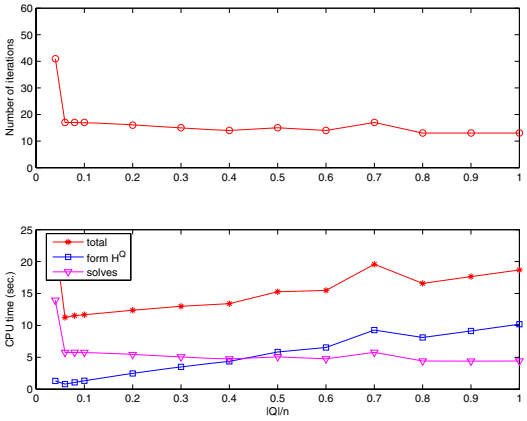


FIG. 9. *rMPC on the problem with constraints tangent to the unit sphere, with infeasible initial point.*

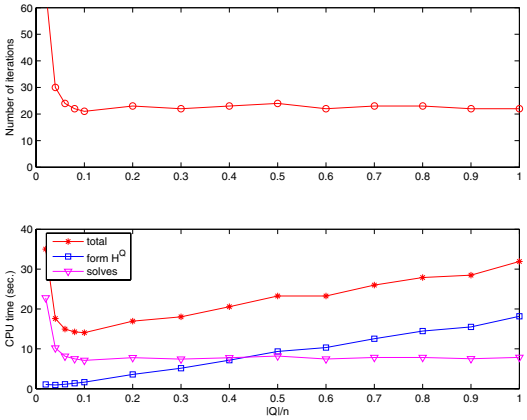


FIG. 10. *rMPC on the "fully random" problem, with infeasible initial point.*

variables. Here we select the initial variables as in [Meh92, p. 589], without modification. Consequently, there is no guarantee that the initial point will be dual-feasible; and indeed, in most experiments, the initial point was dual-infeasible (in addition to being primal-infeasible, as in all the previous experiments). Figures 9, 10, 11, and 12 are the counterparts of Figures 5, 6, 7, 8, respectively.

5. Discussion. In the context of primal-dual interior-point methods for linear programming, a scheme was proposed, aimed at significantly decreasing the computational effort at each iteration when solving problems which, when expressed in dual standard form, have many more constraints than (dual) variables. The core idea is to compute the dual search direction based only on a small subset of the constraints, carefully selected in an attempt to preserve the quality of the search direction. Global and local quadratic convergence was proved for a class of schemes in the case of a simple dual-feasible affine scaling algorithm. Promising numerical results were reported both on this "reduced" affine scaling algorithm and a similarly "reduced" version of the MPC algorithm, using a rather simplistic heuristic: for a prescribed $M > m$, keep only the M most nearly active (or most violated) constraints. In particular,

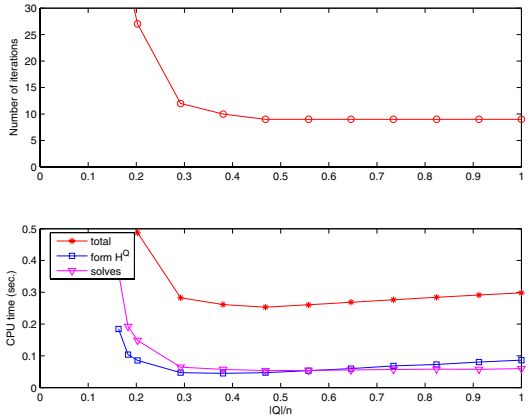


FIG. 11. *rMPC on SCSD1, with infeasible initial point.*

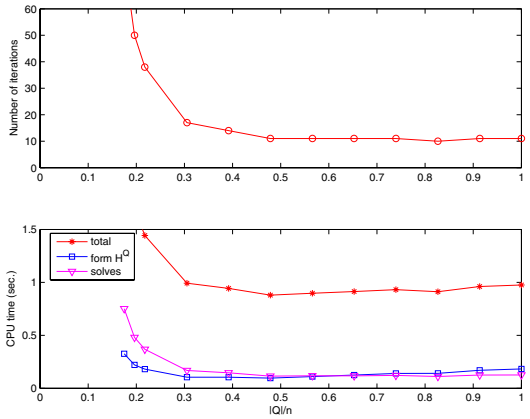


FIG. 12. *rMPC on SCSD6, with infeasible initial point.*

rather unexpectedly, it was observed that, on a number of problems, the number of iterations to solutions did not increase when the size of the reduced constraint set was decreased, down to a small fraction of the total number of constraints! Accordingly, all savings in computational effort per iteration directly translate to savings in total computational effort for the solution of the problem. Another interesting finding is that, in our MATLAB implementation, the reduced affine scaling algorihtm rPDAS works as well as the reduced MPC algorithm on random problems in terms of CPU time; however, CPU times may vary widely over implementations.

The (unreduced) MPC algorithm has remarkable invariance properties. Let $\{(x^k, y^k, s^k)\}$ be a sequence generated by MPC on the problem defined by (A, b, c). Let P be an invertible $m \times m$ matrix, let R be a diagonal positive-definite $n \times n$ matrix, let v belong to $\mathbb{R}^m$, and define $\mathbf{A} := PAR$, $\mathbf{b} := Pb$, $\mathbf{c} := Rc + RA^T P^T v$, $\mathbf{x}^0 := R^{-1}x^0$, $\mathbf{y}^0 := P^{-T}y^0 + v$, and $\mathbf{s}^0 := Rs^0$. Then the sequence $\{(\mathbf{x}^k, \mathbf{y}^k, \mathbf{z}^k)\}$ generated by MPC on the problem defined by $(\mathbf{A}, \mathbf{b}, \mathbf{c})$ satisfies $\mathbf{x}^k = R^{-1}x^k$, $\mathbf{y}^k = P^{-T}y^k + v$, and $\mathbf{s}^k = Rs^k$.[5] The reduced algorithm rMPC is still invariant under the action of

[5]Note, however, that the procedure recommended in [Meh92] for generating x^0 and s^0, while

P and v, but it is no longer invariant under the action of R, because the relation $\mathbf{s} = Rs$ affects the choice of the set Q. A simple way to recover invariance under R is to redefine $\mathcal{Q}_M$ based on $(c_i - a_i^T y)/s_i^0$ instead of $c_i - a_i^T y$.

The rPDAS and PDAS algorithms (rPDAS with $Q = \mathbf{n}$) have weaker invariance properties than MPC. While they are invariant under the action of v and of orthogonal P (that is, Euclidean transformations of the dual space), they are neither invariant under the action of nonorthogonal P, because of the presence of $\|\Delta y\|$ in (3.4) and (3.5), nor under the action of R, because of (3.5) containing the quantity $\|\tilde{x}_-\|$ and fixed bounds on x (and also, for rPDAS, because of the way $\mathcal{Q}_M$ is defined).[6] Algorithms rPDAS and PDAS can be modified to achieve other invariance properties. If $\|\Delta y\|$ is replaced[7] by $\|(\Delta Y^0)^{-1}\Delta y\|$, where $\Delta Y^0 = \operatorname{diag}(\Delta y_i^0,\ i = 1, \ldots, m)$, then the algorithms are invariant under v and nonsingular diagonal P. If instead $\|\Delta y\|$ is replaced by $\|\Delta y\|/\|\Delta y^0\|$, then the algorithms are invariant under Euclidean transformation and uniform scaling of the dual (i.e., P is a nonzero scalar multiple of an orthogonal matrix). If (3.5) is replaced by

$$x_i^+ := \min\{\max\{\min\{(\|\Delta y\|^2 + \|(X^0)^{-1}\tilde{x}_-\|^2)x_i^0, \underline{x}x_i^0\},\ \tilde{x}_i\},\ x_{\max}x_i^0\}\ \forall i \in \mathbf{n},$$

then PDAS is invariant under R; if, moreover, $\mathcal{Q}_M$ is redefined based on $(c_i - a_i^T y)/s_i^0$ instead of $c_i - a_i^T y$, then rPDAS becomes invariant under R, too.

We have focused on a constraint selection rule that requires that, at each iteration, the M "most nearly active" (or "most violated") constraints all be included in the reduced set. It should be clear, however, that nearness to activity can be measured differently for each constraint, and indeed differently at each iteration. In fact, only two conditions must be satisfied in order for our convergence analysis to go through: (i) A_Q must have full row rank at each iteration, which is required in order for the algorithm to be well defined, and (ii) constraints must be included in the reduced set whenever y is "close enough" to the corresponding constraint boundary.

Appendix. Proof of Lemma 1.

The first claim is a direct consequence of the equivalence between (2.2) and (2.3). Let us now prove the sufficiency portion of the second claim. Thus suppose conditions (i) through (iii) hold, and let $(\xi, \eta, \sigma)^T$ be in the nullspace of $J(A, x, s)$. We show that it must be identically zero. We have

$$A^T\eta + \sigma = 0, \tag{A.1}$$

$$A\xi = 0, \tag{A.2}$$

$$S\xi + X\sigma = 0. \tag{A.3}$$

Equation (A.1) yields

$$\xi^T A^T \eta + \xi^T \sigma = 0, \tag{A.4}$$

which, in view of (A.2), yields

$$\xi^T \sigma = 0. \tag{A.5}$$

invariant under the action of P and v, is not invariant under that of R.

[6] Note that simpler versions of PDAS that do not aim at superlinear convergence enjoy v, P, and R invariance as defined above (see, e.g., [MAR90]).

[7] Assuming that no component of Δy^0 vanishes.

Also, (A.3) yields

$$\sigma_i = -\frac{s_i}{x_i}\xi_i$$

whenever $x_i \neq 0$, and $\xi_i = 0$ otherwise (since $|x_i| + |s_i| > 0$), so that (A.5) yields

$$-\sum_{i:x_i\neq 0} \frac{s_i}{x_i}\xi_i^2 = 0.$$

Since $s_i/x_i \geq 0$ whenever $x_i \neq 0$, it follows that $s_i\xi_i = 0$ whenever $x_i \neq 0$. It then follows from (A.3) that $x_i\sigma_i = 0$ whenever $x_i \neq 0$, yielding

$$X\sigma = 0 \tag{A.6}$$

and, from (A.3),

$$S\xi = 0. \tag{A.7}$$

Since $\{a_i : s_i = 0\}$ is linear independent, it follows from (A.2) and (A.7) that $\xi = 0$. Equation (A.1) and (A.6) now yield $XA^T\eta = 0$, so that $a_i^T\eta = 0$ whenever $x_i \neq 0$. Since $\{a_i : x_i \neq 0\}$ spans $\mathbb{R}^m$, we conclude that $\eta = 0$. Finally, it now follows from (A.1) that $\sigma = 0$, concluding the proof of the sufficiency portion of the second claim.

As for the necessity portion of the second claim, first, inspection of the last n rows, then of the first n columns of $J(A, x, s)$, shows that the first two conditions are needed in order for $J(A, x, s)$ to be nonsingular. As for the third condition, suppose it does not hold, i.e., suppose that $\{a_i : x_i \neq 0\}$ does not span $\mathbb{R}^m$. Then there exists $\eta \neq 0$ such that $a_i^T\eta = 0$ for all i such that $x_i \neq 0$. Further, let $\xi := 0$ and let $\sigma := -A^T\eta$, so that $\sigma_i = 0$ for all i such that $x_i \neq 0$. It is readily checked that (ξ, η, σ) is in the nullspace of $J(A, x, s)$. Since $\eta \neq 0$, $J(A, x, s)$ must be singular. This completes the proof of the necessity portion of the second claim.

Acknowledgments. The authors wish to thank Dianne O'Leary for helpful discussions. Further, they wish to thank two anonymous referees for their careful reading of the paper and for their many helpful comments.

REFERENCES

[AT06] P.-A. Absil and A. L. Tits, *Newton-KKT interior-point methods for indefinite quadratic programming*, Comput. Optim. Appl., 2006, to appear.

[dHRT92] D. den Hertog, C. Roos, and T. Terlaky, *A build-up variant of the logarithmic barrier method for LP*, Oper. Res. Lett., 12 (1992), pp. 181–186.

[dHRT94] D. den Hertog, C. Roos, and T. Terlaky, *Adding and deleting constraints in the logarithmic barrier method for LP*, Advances in Optimization and Approximation, D. Z. Du and J. Sun, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1994, pp. 166–185.

[GLY94] J.-L. Goffin, Z.-Q. Luo, and Y. Ye, *On the complexity of a column generation algorithm for convex or quasiconvex feasibility problems*, in Large Scale Optimization, W. W. Hager, D. W. Hearn, and P. M. Pardalos, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1994, pp. 182–191.

[LRT99] Z.-Q. Luo, K. Roos, and T. Terlaky, *Complexity analysis of logarithmic barrier decomposition methods for semi-infinite linear programming*, Appl. Numer. Math., 29 (1999), pp. 379–394.

[MAR90] R. D. C. Monteiro, I. Adler, and M. G. C. Resende, *A polynomial-time primal-dual affine scaling algorithm for linear and convex quadratic programming and its power series extension*, Math. Oper. Res., 15 (1990), pp. 191–214.

[Meh92] S. Mehrotra, *On the implementation of a primal-dual interior point method*, SIAM J. Optim., 2 (1992), pp. 575–601.

[NS96] S. G. Nash and A. Sofer, *Linear and Nonlinear Programming*, McGraw-Hill, New York, 1996.

[O'L04] D. P. O'Leary, *private communication*, 2004.

[PTH88] 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.

[Tit99] A. L. Tits, *An Interior Point Method for Linear Programming, with an Active Set Flavor*, Technical report TR-99-47, Institute for Systems Research, University of Maryland, College Park, MD, 1999.

[TZ94] A. L. Tits and J. L. Zhou, *A simple, quadratically convergent algorithm for linear programming and convex quadratic programming*, Large Scale Optimization, W. W. Hager, D. W. Hearn, and P. M. Pardalos, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1994, pp. 411–427.

[Wri97] S. J. Wright, *Primal-dual Interior-point Methods*, SIAM, Philadelphia, 1997.

[Ye90] Y. Ye, *A "build-down" scheme for linear programming*, Math. Programming, 46 (1990), pp. 61–72.

[Ye92] Y. Ye, *A potential reduction algorithm allowing column generation*, SIAM J. Optim., 2 (1992), pp. 7–20.

[Ye97] Y. Ye, *Complexity analysis of the analytic center cutting plane method that uses multiple cuts*, Math. Programming, 78 (1997), pp. 85–104.

SIAM J. OPTIM.
Vol. 17, No. 1, pp. 147–158
© 2006 Society for Industrial and Applied Mathematics

ON THE STABILITY OF CONVEX-VALUED MAPPINGS AND THEIR RELATIVE BOUNDARY AND EXTREME POINTS SET MAPPINGS*

MIGUEL A. GOBERNA†, MAXIM I. TODOROV‡, AND VIRGINIA N. VERA DE SERIO§

Abstract. This paper deals with the transmission of the main stability properties (lower and upper semicontinuity in Berge sense, and closedness) from a given closed–convex-valued mapping to its corresponding relative boundary and extreme point set mappings, and vice versa. The domain of the mappings considered in this paper are locally metrizable spaces and the images range on Euclidean spaces. Important examples of the class of mappings considered in this paper are the feasible set mapping and the optimal set mapping of convex optimization problems, for which the space of parameters is the result of perturbing a given nominal problem.

Key words. stability theory, set-valued mappings, convex hull mappings, relative boundary mappings, extreme points set mappings

AMS subject classifications. 49K40, 28B20

DOI. 10.1137/050632476

1. Introduction.

The main objective of the paper is to analyze the relationships between important pairs of mappings, one of them being the convex hull of the other, which frequently arise in convex optimization (convex systems), where, as a consequence of measurement or roundoff errors, the nominal problem y_0 (system y_0) is usually replaced in practice by perturbed problems (systems, respectively) having the same structure. Let us denote by Y the set of all possible perturbed problems (systems) equipped with a certain pseudometric measuring the size of the perturbations and let $\mathcal{F} : Y \rightrightarrows \mathbb{R}^n$ be the set-valued mapping associating with each $y \in Y$ its feasible set or its optimal set (its solution set, respectively). Under mild conditions, $\mathcal{F}(y)$ is the convex hull of its boundary set $\operatorname{bd}\mathcal{F}(y)$, its relative boundary set $\operatorname{rbd}\mathcal{F}(y)$, and/or its extreme points set $\operatorname{extr}\mathcal{F}(y)$ for all $y \in Y$. We denote these mappings from Y to $\mathbb{R}^n$ as $\operatorname{bd}\mathcal{F}$, $\operatorname{rbd}\mathcal{F}$, and $\operatorname{extr}\mathcal{F}$, which are called *boundary mapping*, *relative boundary mapping*, and *extreme points set mapping* of $\mathcal{F}$, respectively. The connections between the stability properties of $\mathcal{F}$, $\operatorname{bd}\mathcal{F}$, and $\operatorname{extr}\mathcal{F}$ have been already analyzed in the particular context of linear semi-infinite systems ([3] and [4], respectively), where Y is equipped with the pseudometric of the uniform convergence.

Throughout this paper we consider given an arbitrary convex-valued mapping $\mathcal{F} : Y \rightrightarrows \mathbb{R}^n$, where the domain Y is a locally metrizable space (i.e., Y is equipped with the topology induced by an extended distance on Y, δ, taking values on $\mathbb{R}_+ \cup \{+\infty\}$), and its boundary mapping, relative boundary mapping, and extreme points set mapping,

*Received by the editors May 26, 2005; accepted for publication (in revised form) November 2, 2005; published electronically May 3, 2006.
http://www.siam.org/journals/siopt/17-1/63247.html

†University of Alicante, Statistics and Operations Research, Ctra. San Vicente s/n, Alicante 03071, Spain (mgoberna@ua.es). This author was supported by MCYT of Spain and FEDER of EU, grant BMF2002-04114-C02-01.

‡Actuary and Mathematics, UDLA, 72820 San Andrès Cholula, Puebla, Mexico. On leave from IMI-BAS, Sofia, Bulgaria (maxim.todorov@udlap.mx). This author was supported by CONACyT of Mexico, grant 44003.

§Universidad Nacional de Cuyo, Faculty of Economics, Campus UNCUYO, Mendoza 5500, Argentina (vvera@fcemail.uncu.edu.ar). This author was supported by SECYT-UNCuyo of Argentina, grant 987/02-R-04.

$\operatorname{bd}\mathcal{F}$, $\operatorname{rbd}\mathcal{F}$, and $\operatorname{extr}\mathcal{F}$. The relationships between $\mathcal{F}$ and $\operatorname{bd}\mathcal{F}$, assuming that $\mathcal{F} = \operatorname{conv}\operatorname{bd}\mathcal{F}$, have been studied in [5]. In the same vein, this paper considers the relationships between the stability properties of $\mathcal{F}$, $\operatorname{rbd}\mathcal{F}$, and $\operatorname{extr}\mathcal{F}$, assuming that $\mathcal{F} = \operatorname{conv}\operatorname{rbd}\mathcal{F}$ and $\mathcal{F} = \operatorname{conv}\operatorname{extr}\mathcal{F}$, respectively. The finite dimension of the image space plays a crucial role in those arguments based on the compactness of the unit sphere or on Carathéodory's theorem.

Some of these relationships are direct consequences of basic results about arbitrary mappings $\mathcal{A} : Y \rightrightarrows \mathbb{R}^n$ and their corresponding *convex hull mappings*, $\operatorname{conv}\mathcal{A} : Y \rightrightarrows \mathbb{R}^n$, which associates to each $y \in Y$ the convex hull of $\mathcal{A}(y)$, i.e., $(\operatorname{conv}\mathcal{A})(y) = \operatorname{conv}\mathcal{A}(y)$ for all $y \in Y$. Although some results on the transmission of stability properties between $\mathcal{A}$ and $\operatorname{conv}\mathcal{A}$ are already known (see, e.g., [6] and [1]), we provide proofs of other results which will be used in what follows. Thus, for each stability property, we start analyzing the relationships between $\mathcal{A}$ and $\operatorname{conv}\mathcal{A}$, and then we exploit the properties of the images of $\mathcal{F}$, $\operatorname{rbd}\mathcal{F}$, and $\operatorname{extr}\mathcal{F}$ in order to obtain the relationships between these mappings; section 3 deals with the lower semicontinuous (lsc) property and section 4 with the upper semicontinuous (usc) property and closedness.

Let us introduce some additional notation. Given $X \subset \mathbb{R}^n$, $\operatorname{aff} X$ denotes the affine hull of X. From the topological side, $\operatorname{bd} X$, $\operatorname{rbd} X$, $\operatorname{int} X$, $\operatorname{rint} X$, and $\operatorname{cl} X$ represent the boundary, the relative boundary, the interior, the relative interior, and the closure of X, respectively. If X is convex, its set of extreme points is denoted by $\operatorname{extr} X$. The Euclidean norm in $\mathbb{R}^n$ will be denoted by $\|.\|$ and the open ball centered at x and radius $\varepsilon > 0$ by $B(x;\varepsilon)$. If X is a convex set and $x \in X$, then

$$B(x;\varepsilon) \cap \operatorname{rbd} X = \emptyset \Longrightarrow B(x;\varepsilon) \cap \operatorname{aff} X \subset \operatorname{rint} X \tag{1.1}$$

for all $\varepsilon > 0$.

The standard simplex in $\mathbb{R}^{n+1}$ is

$$S := \left\{ (\lambda_1, \ldots, \lambda_{n+1}) \in \mathbb{R}^{n+1}_+ \mid \sum_{i=1}^{n+1} \lambda_i = 1 \right\}.$$

For the sake of completeness, we recall the stability concepts and some basic results for set-valued mappings that we shall consider in this paper. Let $\mathcal{M} : Y \rightrightarrows \mathbb{R}^n$ be a set-valued mapping with its domain $\operatorname{dom}\mathcal{M} := \{y \in Y \mid \mathcal{M}(y) \neq \emptyset\}$. The following semicontinuity concepts are due to Bouligand and Kuratowski (see [1, section 1.4]).

We say that $\mathcal{M}$ is *lower semicontinuous* at $y_0 \in Y$ *in the Berge sense* if, for each open set $W \subset \mathbb{R}^n$ such that $W \cap \mathcal{M}(y_0) \neq \emptyset$, there exists an open set $V \subset Y$, containing y_0, such that $W \cap \mathcal{M}(y) \neq \emptyset$ for each $y \in V$. Obviously, $\mathcal{M}$ is lsc at $y_0 \notin \operatorname{dom}\mathcal{M}$ and $y_0 \in \operatorname{int}\operatorname{dom}\mathcal{M}$ if $\mathcal{M}$ is lsc at $y_0 \in \operatorname{dom}\mathcal{M}$.

$\mathcal{M}$ is *upper semicontinuous* at $y_0 \in Y$ *in the Berge sense* if, for each open set $W \subset \mathbb{R}^n$ such that $\mathcal{M}(y_0) \subset W$, there exists an open set $V \subset Y$, containing y_0, such that $\mathcal{M}(y) \subset W$ for each $y \in V$. If $\mathcal{M}$ is usc at $y_0 \notin \operatorname{dom}\mathcal{M}$, then $y_0 \in \operatorname{int}(Y \setminus \operatorname{dom}\mathcal{M})$.

If $\mathcal{M}$ is simultaneously lsc and usc at y_0 we say that $\mathcal{M}$ is *continuous* at this point.

$\mathcal{M}$ is *closed* at $y_0 \in \operatorname{dom}\mathcal{M}$ if for all sequences $\{y_r\}_{r=1}^{\infty} \subset Y$ and $\{x_r\}_{r=1}^{\infty} \subset \mathbb{R}^n$ satisfying $x_r \in \mathcal{M}(y_r)$ for all $r \in \mathbb{N}$, $\lim_{r\to\infty} y_r = y_0$ and $\lim_{r\to\infty} x_r = x_0$ (in brief, $y_r \to y_0$ and $x_r \to x_0$) one has $x_0 \in \mathcal{M}(y_0)$. If $\mathcal{M}$ is usc at $y_0 \in \operatorname{dom}\mathcal{M}$ and $\mathcal{M}(y_0)$ is closed, then $\mathcal{M}$ is closed at y_0. Conversely, if $\mathcal{M}$ is closed and *locally bounded* at

$y_0 \in \operatorname{dom} \mathcal{M}$ (i.e., if there is a neighborhood of y_0, say V, and a bounded set $A \subset \mathbb{R}^n$ containing $\mathcal{M}(y)$ for every $y \in V$), then $\mathcal{M}$ is usc at y_0.

Finally, $\mathcal{M}$ is lsc (usc, closed, locally bounded) if it is lsc (usc, closed, locally bounded) at y for all $y \in Y$.

Without entering in details we would like to mention that there are other notions of lower and upper semicontinuity as lsc and usc in the sense of Hausdorff (see, e.g., [2]) or inner and outer semicontinuity (see, e.g., [8], where it is shown that the last two concepts are equivalent to lsc in Berge sense and closedness when $\mathcal{M}(y)$ is closed for all $y \in Y$).

2. Preliminaries. We say that $\mathcal{M} : Y \rightrightarrows \mathbb{R}^n$ is *locally convex* at $y_0 \in Y$ if there exists an open set $V \subset Y$, containing y_0, such that $\mathcal{M}(y)$ is convex for all $y \in V$. We shall use the following sufficient condition for $\mathcal{M}$ to be locally bounded.

PROPOSITION 2.1. *Let $\mathcal{M} : Y \rightrightarrows \mathbb{R}^n$ and let $y_0 \in \operatorname{dom} \mathcal{M}$ such that $\mathcal{M}(y_0)$ is bounded and $\mathcal{M}$ is lsc, closed, and locally convex at y_0. Then $\mathcal{M}$ is locally bounded and continuous at y_0.*

Proof. Let $r_0 \in \mathbb{N}$ such that

$$\mathcal{M}(y_0) \subset B(0_n; r_0). \tag{2.1}$$

Since $\mathcal{M}$ is lsc and locally convex at y_0 there exists an open set $V \subset Y$, containing y_0, such that $\mathcal{M}(y)$ is convex and

$$B(0_n; r_0) \cap \mathcal{M}(y) \neq \emptyset \text{ for each } y \in V. \tag{2.2}$$

If $\mathcal{M}$ is not locally bounded at y_0, given $r \in \mathbb{N}$ there exists $y_r \in Y$, with $\delta(y_r, y) \leq \frac{1}{r}$, such that $\mathcal{M}(y_r) \nsubseteq B(0_n; r)$. Thus there exists a sequence $\{x_r\}$ such that

$$x_r \in \mathcal{M}(y_r), \|x_r\| \geq r, r = 1, 2, \ldots .$$

Let $r_1 \geq r_0$ such that $y_r \in V$ for all $r \geq r_1$. In this case, due to (2.2), we can take $z_r \in B(0_n; r_0) \cap \mathcal{M}(y_r)$. Since $x_r \in \mathcal{M}(y_r) \setminus B(0_n; r_0)$ and $\mathcal{M}(y_r)$ is convex, there exists $u_r \in]x_r, z_r] := \{(1-\lambda) x_r + \lambda z_r \mid 0 < \lambda \leq 1\}$ such that

$$u_r \in \mathcal{M}(y_r), \|u_r\| = r_0, r \leq r_1. \tag{2.3}$$

By the compactness of the spheres in $\mathbb{R}^n$, there exists a subsequence $\{u_{r_k}\}$ such that $u_{r_k} \in \mathcal{M}(y_{r_k})$, $k = 1, 2, \ldots$, and $\lim_k u_{r_k} = u_0$, with $\|u_0\| = r_0$. Since $\mathcal{M}$ is closed at y_0 and $\lim_k y_{r_k} = y_0$, we must have $u_0 \in \mathcal{M}(y_0)$, which contradicts (2.1).

We have shown that $\mathcal{M}$ is locally bounded at y_0. Since we are assuming that $\mathcal{M}$ is closed at y_0, it is also usc at y_0. Hence it is continuous at y_0. ∎

The condition of $\mathcal{M}$ being locally convex above is not superfluous as the following example shows.

Example 2.2. If $Y = [0, 1]$ and $\mathcal{M} : Y \rightrightarrows \mathbb{R}$ is defined by $\mathcal{M}(y) = \{0, 1/y\}$ for $y \neq 0$ and $\mathcal{M}(0) = \{0\}$, then $\mathcal{M}$ is neither locally bounded nor continuous at $y_0 = 0$, in spite of $\mathcal{M}(y_0)$ being bounded and being $\mathcal{M}$ lsc and closed at y_0.

The *truncated mapping* of $\mathcal{M} : Y \rightrightarrows \mathbb{R}^n$ with radius $\rho > 0$ is $\mathcal{M}_\rho : Y \rightrightarrows \mathbb{R}^n$ defined such as

$$\mathcal{M}_\rho(y) := \mathcal{M}(y) \cap \operatorname{cl} B(0_n; \rho) \text{ for all } y \in Y.$$

The following result (Lemma 2 in [5]), which establishes the relationships between $\mathcal{M}$ and $\mathcal{M}_\rho$, will be useful in the next sections.

PROPOSITION 2.3. *Let $\mathcal{M} : Y \rightrightarrows \mathbb{R}^n$ and let $y_0 \in \operatorname{dom} \mathcal{M}$. Then the following*

statements hold:
(i) $\mathcal{M}$ *is closed at* y_0 *if and only if* $\mathcal{M}_\rho$ *is closed at* y_0 *for all* $\rho > 0$ *such that* $\mathcal{M}_\rho(y_0) \neq \emptyset$.
(ii) *If* $\mathcal{M}$ *is usc at* y_0 *and* $\mathcal{M}(y_0)$ *is closed, then* $\mathcal{M}_\rho$ *is usc at* y_0 *for all* $\rho > 0$ *such that* $\mathcal{M}_\rho(y_0) \neq \emptyset$.
(iii) *If* $\mathcal{M}$ *is usc at* y_0, *then there exist a positive scalar* $\overline{\rho}$ *and an open neighborhood of* y_0, V, *such that*

$$\mathcal{M}(y) \setminus \mathcal{M}_{\overline{\rho}}(y) \subset \mathcal{M}(y_0) \setminus \mathcal{M}_{\overline{\rho}}(y_0) \quad \text{for all } y \in V. \tag{2.4}$$

The converse statement holds when $\mathcal{M}$ *is closed at* y_0.
(iv) *If* $\mathcal{M}_\rho$ *is lsc at* y_0 *for every* ρ *such that* $\mathcal{M}(y_0) \cap B(0_n; \rho) \neq \emptyset$, *then* $\mathcal{M}$ *is lsc at* y_0. *The converse statement holds if* $\mathcal{M}(y_0)$ *is convex.*

As an immediate consequence of the following result we obtain characterizations of the identities $\mathcal{F} = \operatorname{conv} \operatorname{bd} \mathcal{F}$, $\mathcal{F} = \operatorname{conv} \operatorname{rbd} \mathcal{F}$, and $\mathcal{F} = \operatorname{conv} \operatorname{extr} \mathcal{F}$. Recall that an edge is a one-dimensional face whereas a half-flat is the intersection of a flat (also called affine manifold) with a closed halfspace which meets it, but does not contain it.

PROPOSITION 2.4. *Given a convex set* $F \subset \mathbb{R}^n$, *the following statements hold:*
(i) $F = \operatorname{conv} \operatorname{bd} F$ *if and only if* F *is a closed set which does not contain halfspaces.*
(ii) $F = \operatorname{conv} \operatorname{rbd} F$ *if and only if* F *is a closed set which does not contain half-flats of the same dimension.*
(iii) *If* $F = \operatorname{conv} \operatorname{extr} F$, *then* F *contains neither lines nor unbounded edges. The converse holds if* F *is closed.*

Proof. Obviously, if $F = \emptyset$, then

$$\operatorname{conv} \operatorname{bd} F = \operatorname{conv} \operatorname{rbd} F = \operatorname{conv} \operatorname{extr} F = \emptyset.$$

So we can assume that $F \neq \emptyset$ without loss of generality.

(i) It is a straightforward consequence of Lemma 2 in [3].

(ii) If $F = \operatorname{conv} \operatorname{rbd} F$, then $\operatorname{rbd} F \subset F$ and so F is closed for each $y \in Y$. If F contains a half-flat of the same dimension, then it is either a flat or a half-flat, with $\operatorname{conv} \operatorname{rbd} F \neq F$ in both cases.

Conversely, since F is a closed and convex set which is neither a flat nor a half-flat, then $F = \operatorname{conv} \operatorname{rbd} F$ by Theorem 2.6.12 in [9].

(iii) Suppose that $F = \operatorname{conv} \operatorname{extr} F$. $F \neq \emptyset$ entails $\operatorname{extr} F \neq \emptyset$ and so F does not contain lines. We shall obtain a contradiction assuming the existence of a halfline edge of F, say A.

Let $A = \{\overline{x} + \lambda v \mid \lambda \geq 0\}$ be an edge of F. Then $v \neq 0_n$ and $\overline{x} \in \operatorname{extr} F$. We shall prove that no element of $A \setminus \{\overline{x}\}$ belongs to $\operatorname{conv} \operatorname{extr} F$. We assume the contrary, i.e., that there exists $\lambda > 0$ such that $\overline{x} + \lambda v \in \operatorname{conv} \operatorname{extr} F$.

If $\overline{x} + \lambda v = x_1 \in \operatorname{extr} F$, then $x_1 = \frac{1}{2}\overline{x} + \frac{1}{2}(\overline{x} + 2\lambda v)$, with $\overline{x}, \overline{x} + 2\lambda v \in F$, making this impossible. Thus we can write $\overline{x} + \lambda v = \sum_{i=1}^{p} \lambda_i x_i$, where $p \geq 2$, $\sum_{i=1}^{p} \lambda_i = 1$ and $\lambda_i > 0$, and $x_i \in \operatorname{extr} F$, $i = 1, \ldots, p$, with $x_i \neq x_j$ if $i \neq j$. Then we can write

$$\overline{x} + \lambda v = \lambda_1 x_1 + (1 - \lambda_1) \sum_{i=2}^{p} \left(\frac{\lambda_i}{1 - \lambda_1} \right) x_i, \tag{2.5}$$

which yields x_1, $\sum_{i=2}^{p} \left(\frac{\lambda_i}{1-\lambda_1}\right) x_i \in A$ because A is a face of F. Since $A \cap \operatorname{extr} F = \{\overline{x}\}$, $x_1 = \overline{x}$, and so from (2.5) we get

$$\overline{x} + \frac{\lambda}{1-\lambda_1} v = \sum_{i=2}^{p} \left(\frac{\lambda_i}{1-\lambda_1} \right) x_i. \tag{2.6}$$

By taking into account again that A is a face of F, we get the following contradiction: $x_2, \ldots, x_p \in A \cap \operatorname{extr} F = \{\overline{x}\}$.

We have shown that $(A \setminus \{\overline{x}\}) \cap \operatorname{conv} \operatorname{extr} F = \emptyset$. Since $\emptyset \neq A \setminus \{\overline{x}\} \subset F$, we conclude that $\operatorname{conv} \operatorname{extr} F \subsetneq F$.

Conversely, if F is closed and does not contain lines, it is the convex hull of its extreme points and extreme directions (Corollary 2.6.15 in [9]). Since the assumption precludes the existence of extreme directions, we have $\operatorname{conv} \operatorname{extr} F = F$. □

Remark 2.5. According to Proposition 2.4, if $\mathcal{F} = \operatorname{conv} \operatorname{bd} \mathcal{F}$ ($\mathcal{F} = \operatorname{conv} \operatorname{rbd} \mathcal{F}$), then we have $\mathcal{F}_\rho = \operatorname{conv} \operatorname{bd} \mathcal{F}_\rho$ ($\mathcal{F}_\rho = \operatorname{conv} \operatorname{rbd} \mathcal{F}_\rho$, respectively) for all $\rho > 0$. Nevertheless, in the case of $\mathcal{F} = \operatorname{conv} \operatorname{extr} \mathcal{F}$, we need to show that $\mathcal{F}_\rho = \operatorname{conv} \operatorname{extr} \mathcal{F}_\rho$ because $\mathcal{F}$ could be not closed-valued. In order to do this, it is enough to prove that if $\mathcal{F}(y) := F = \operatorname{conv} \operatorname{extr} F$ and $x \in F_\rho$ with $\|x\| < \rho$, then $x \in \operatorname{conv} \operatorname{extr} F_\rho$. We can write

$$x = \sum_{j \in J} \lambda_j x_j, |J| < \infty, \sum_{j \in J} \lambda_j = 1, \lambda_j > 0 \text{ and } x_j \in \operatorname{extr} F \text{ for all } j \in J.$$

Let $I = \{j \in J \mid \|x_j\| > \rho\}$. If $I = \emptyset$, then $x_j \in [\operatorname{extr} F]_\rho \subset \operatorname{extr} F_\rho$ for all $j \in J$ and so $x \in \operatorname{conv} \operatorname{extr} F_\rho$. Otherwise take an arbitrary $k \in I$. Let $x'_k \in [x, x_k] \subset F$ such that $\|x'_k\| = \rho$, so that $x'_k \in \operatorname{extr} F_\rho$. If $x'_k = (1-\mu)\, x + \mu x_k$, with $0 < \mu < 1$, and we denote $y_j = x_j$ for all $j \in J$, $j \neq k$, and $y_k = x'_k$, we get an expression $x = \sum_{j \in J} \alpha_j y_j$, where $\sum_{j \in J} \alpha_j = 1, \alpha_j > 0$ and $y_j \in \operatorname{extr} F$ for all $j \in J$, but now the cardinality of the set $\{j \in J \mid \|y_j\| > \rho\}$ is $|I| - 1$. After $|I|$ iterations of this procedure we get x expressed as a convex combination of elements of $\operatorname{extr} F_\rho$. In fact, if Φ is any operator that transforms convex sets in $\mathbb{R}^n$ into sets in $\mathbb{R}^n$ satisfying $[\Phi(\mathcal{F})]_\rho \subset \Phi(\mathcal{F}_\rho) \subset \mathcal{F}_\rho$ and $\{x \in \mathcal{F}(y) \mid \|x\| = \rho\} \subset \Phi(\mathcal{F}_\rho(y))$ for all $y \in Y$, then

$$\mathcal{F} = \operatorname{conv} \Phi(\mathcal{F}) \Longrightarrow \mathcal{F}_\rho = \operatorname{conv} \Phi(\mathcal{F}_\rho).$$

Observe that $\Phi(\mathcal{F}) = \operatorname{bd} \mathcal{F}$, $\operatorname{rbd} \mathcal{F}$, and $\operatorname{extr} \mathcal{F}$ satisfy these conditions.

3. Lower semicontinuity. We shall use the following classical result ([6, Proposition 2.6]).

THEOREM 3.1. *If $\mathcal{A} : Y \rightrightarrows \mathbb{R}^n$ is lsc at $y_0 \in \operatorname{dom} \mathcal{A}$, then $\operatorname{conv} \mathcal{A}$ is also lsc at y_0.*

In particular, taking $\mathcal{A} = \operatorname{bd} \mathcal{F}$ we get the direct statement of Proposition 1 in [5], whose corresponding converse statement establishes that, if $\mathcal{F} = \operatorname{conv} \operatorname{bd} \mathcal{F}$ is lsc and closed at $y_0 \in \operatorname{dom} \mathcal{F}$, then $\operatorname{bd} \mathcal{F}$ is also lsc at y_0. The next two results are counterparts of this converse statement for $\operatorname{rbd} \mathcal{F}$ and $\operatorname{extr} \mathcal{F}$ (instead of $\operatorname{bd} \mathcal{F}$). Example 3 in [5], where $\operatorname{bd} \mathcal{F} = \operatorname{rbd} \mathcal{F} = \operatorname{extr} \mathcal{F}$, shows that the closedness of $\mathcal{F}$ is not superfluous in these results. The following example shows that, in general, if $\operatorname{conv} \mathcal{A}$ is lsc and closed at y_0, then $\mathcal{A}$ is not necessarily lsc at y_0. Accordingly, the proofs must appeal to the specific properties of the sets $\operatorname{rbd} \mathcal{F}(y)$ and $\operatorname{extr} \mathcal{F}(y)$.

Example 3.2. Let $\mathcal{A} : \mathbb{R} \rightrightarrows \mathbb{R}$ such that

$$\mathcal{A}(y) = \begin{cases} \{-1, 0, 1\}, & y = 0, \\ \{-1, 1\}, & y \neq 0. \end{cases}$$

It is easy to see that $\operatorname{conv}\mathcal{A}$ is constant (so that it is continuous and closed) whereas $\mathcal{A}$ is not lsc at $y_0 = 0$.

THEOREM 3.3. *Let $\mathcal{F} : Y \rightrightarrows \mathbb{R}^n$ be such that $\mathcal{F} = \operatorname{conv}\operatorname{rbd}\mathcal{F}$ and $\mathcal{F}$ is lsc and closed at $y_0 \in \operatorname{dom}\mathcal{F}$. Then $\operatorname{rbd}\mathcal{F}$ is lsc at y_0.*

Proof. Let us denote $\mathcal{R} = \operatorname{rbd}\mathcal{F}$. Since $\mathcal{F}(y_0)$ cannot be singleton (otherwise $\mathcal{R}(y_0) = \emptyset$, contradicting the assumptions), we have $|\mathcal{F}(y_0)| > 1$.

We assume that $\mathcal{R}$ is not lsc at y_0 and we shall obtain a contradiction. This assumption entails the existence of an open convex set W and a sequence $\{y_r\}$ such that $y_r \to y_0$,

$$W \cap \mathcal{R}(y_0) \neq \emptyset, \tag{3.1}$$

and

$$W \cap \mathcal{R}(y_r) = \emptyset, r = 1, 2, \ldots . \tag{3.2}$$

Since $y_0 \in \operatorname{int}\operatorname{dom}\mathcal{F}$, we can assume that $y_r \in \operatorname{dom}\mathcal{F}, r = 1, 2, \ldots$. By (3.1), we can choose a point $\widehat{x} \in W \cap \mathcal{R}(y_0)$. Fix $\overline{x} \in \operatorname{rint}\mathcal{F}(y_0)$. Then

$$\widehat{x} - \lambda(\overline{x} - \widehat{x}) \notin \mathcal{F}(y_0) \text{ for all } \lambda > 0. \tag{3.3}$$

Because $\mathcal{F}$ is lsc at y_0 and $\overline{x}, \widehat{x} \in \mathcal{F}(y_0)$, there exist two sequences, $\{\overline{x}_r\}$ and $\{\widehat{x}_r\}$, with $\overline{x}_r, \widehat{x}_r \in \mathcal{F}(y_r)$ for all r, $\overline{x}_r \to \overline{x}$, and $\widehat{x}_r \to \widehat{x}$. Let $\delta > 0$ such that $B(\widehat{x}; \delta) \subset W$ and take $r_0 \in \mathbb{N}$ such that $\widehat{x}_r \in B(\widehat{x}; \frac{\delta}{2})$ for all $r \geq r_0$. Given $r \geq r_0$, (3.2) yields $B(\widehat{x}; \frac{\delta}{2}) \cap \mathcal{R}(y_r) = \emptyset$ and so, by (1.1), $B(\widehat{x}; \frac{\delta}{2}) \cap \operatorname{aff}\mathcal{F}(y_r) \subset \mathcal{F}(y_r)$. Hence

$$\widehat{x}_r - \frac{\delta}{4\|\overline{x}_r - \widehat{x}_r\|}(\overline{x}_r - \widehat{x}_r) \in \mathcal{F}(y_r) \quad \text{for all } r \geq r_0.$$

Taking limits as $r \to \infty$ we get, by the closedness of $\mathcal{F}$ at y_0, that

$$\widehat{x} - \frac{\delta}{4\|\overline{x} - \widehat{x}\|}(\overline{x} - \widehat{x}) \in \mathcal{F}(y_0),$$

in contradiction with (3.3). □

THEOREM 3.4. *Let $\mathcal{F} : Y \rightrightarrows \mathbb{R}^n$ be such that $\mathcal{F} = \operatorname{conv}\operatorname{extr}\mathcal{F}$ and $\mathcal{F}$ is lsc and closed at $y_0 \in \operatorname{dom}\mathcal{F}$. Then $\operatorname{extr}\mathcal{F}$ is lsc at y_0.*

Proof. We denote $\mathcal{E} = \operatorname{extr}\mathcal{F}$ and consider two possible cases.

Case 1. $\mathcal{F}(y_0)$ is bounded.

$\mathcal{F}$ is locally bounded at y_0 according to Proposition 2.1. Let V be an open set in Y, $y_0 \in V$, and $\rho > 0$ such that $\mathcal{F}(y) \subset \operatorname{cl} B(0_n; \rho)$ for all $y \in V$.

We assume that $\mathcal{E}$ is not lsc at y_0 and we shall get a contradiction.

Let W be an open set and let $\{y_r\} \subset V$, with $y_r \to y_0$, be such that

$$W \cap \mathcal{E}(y_0) \neq \emptyset \tag{3.4}$$

and

$$W \cap \mathcal{E}(y_r) = \emptyset \text{ for all } r \in \mathbb{N}. \tag{3.5}$$

By (3.4) we can select a point $x_0 \in W \cap \mathcal{E}(y_0)$.

Given $k \in \mathbb{N}$, since $x_0 \in B(x_0; k^{-1}) \cap \mathcal{F}(y_0)$ and $\mathcal{F}$ is lsc at y_0, there exists $r_k \in \mathbb{N}$ such that $B(x_0; k^{-1}) \cap \mathcal{F}(y_{r_k}) \neq \emptyset$. We can assume that $\{y_{r_k}\}$ is a subsequence of $\{y_r\}$. Let

$$z_k \in B(x_0; k^{-1}) \cap \mathcal{F}(y_{r_k}), \quad k = 1, 2, \ldots . \tag{3.6}$$

For any $k \in \mathbb{N}$, we can write

$$z_k = \sum_{i=1}^{n+1} \lambda_i^k e_i^k, \quad (\lambda_1^k, \ldots, \lambda_{n+1}^k) \in S, \quad e_i^k \in \mathcal{E}(y_{r_k}), \quad i = 1, \ldots, n+1, \tag{3.7}$$

because $\mathcal{F}(y_{r_k}) = \operatorname{conv} \mathcal{E}(y_{r_k})$.

By the compactness of the simplex S, we can assume without loss of generality that $(\lambda_1^k, \ldots, \lambda_{n+1}^k) \to (\lambda_1, \ldots, \lambda_{n+1}) \in S$. Analogously, since for any $i \in \{1, \ldots, n+1\}$,

$$\left\{e_i^k\right\} \subset \mathcal{E}(y_{r_k}) \subset \mathcal{F}(y_{r_k}) \subset \operatorname{cl} B(0_n; \rho),$$

we can assume that $e_i^k \to e_i \in \operatorname{cl} B(0_n; \rho)$, $i = 1, \ldots, n+1$. Since $\mathcal{F}$ is closed at y_0 and $e_i^k \in \mathcal{F}(y_{r_k})$ for all $k \in \mathbb{N}$, we get $e_i \in \mathcal{F}(y_0)$. Now, taking $\lim_k$ in (3.7) and recalling (3.6), we obtain

$$x_0 = \sum_{i=1}^{n+1} \lambda_i e_i, \quad (\lambda_1^k, \ldots, \lambda_{n+1}^k) \in S, \quad e_i \in \mathcal{F}(y_0), \quad i = 1, \ldots, n+1. \tag{3.8}$$

Since $x_0 \in \mathcal{E}(y_0) = \operatorname{extr} \mathcal{F}(y_0)$, we must have in (3.8) all the coefficients $\lambda_i = 0$ except one, $\lambda_j = 1$, in which case $x_0 = e_j$. Since $e_j = \lim_k e_j^k$, $\{e_j^k\} \subset \mathcal{E}(y_{r_k}) \subset \mathbb{R}^n \backslash W$ by (3.5), and $\mathbb{R}^n \backslash W$ is closed, we have $x_0 = e_j \in \mathbb{R}^n \backslash W$, i.e., $x_0 \notin W$. This contradicts the selection of x_0 in $W \cap \mathcal{E}(y_0)$.

Case 2. $\mathcal{F}(y_0)$ is unbounded.

The plan of the proof is to consider the truncated mapping $\mathcal{F}_\rho$, for a certain $\rho > 0$. Since $\mathcal{F}_\rho = \operatorname{conv} \operatorname{extr} \mathcal{F}_\rho$ by Remark 2.5 and $\mathcal{F}_\rho(y_0)$ is bounded, we are in case 1 and so $\operatorname{extr} \mathcal{F}_\rho$ will be lsc at y_0. This will allow us to conclude that $\mathcal{E} = \operatorname{extr} \mathcal{F}$ is lsc at y_0.

First we show that if $\mathcal{E}_\rho$ is the truncated mapping of $\mathcal{E}$ of radius $\rho > 0$, then

$$\operatorname{extr} \mathcal{F}_\rho(y) = \mathcal{E}_\rho(y) \cup \{x \in \mathcal{F}(y) \mid \|x\| = \rho\} \text{ for all } y \in Y. \tag{3.9}$$

In fact, the inclusion $\operatorname{extr} \mathcal{F}_\rho(y) \supset \mathcal{E}_\rho(y) \cup \{x \in \mathcal{F}(y) \mid \|x\| = \rho\}$ is obvious. For the reverse inclusion take $x \in \operatorname{extr} \mathcal{F}_\rho(y)$ such that $\|x\| < \rho$. Assume that $x = \lambda u + (1-\lambda)v$ with $0 < \lambda < 1$ and $u, v \in \mathcal{F}(y), u \neq v$. We may assume without loss of generality that $\|u\|, \|v\| < \rho$ which contradicts the fact that x is an extreme point of $\mathcal{F}_\rho(y)$. Therefore, $x \in \mathcal{E}_\rho(y)$.

Now, in order to prove that $\mathcal{E}$ is lsc at y_0, assume that $\mathcal{E}$ is not. Then there exist $x_0 \in \mathcal{E}(y_0)$, $\delta > 0$, and a sequence $\{y_r\}$ such that $y_r \to y_0$ and $\mathcal{E}(y_r) \cap B(x_0; \delta) = \emptyset$ for every $r \in \mathbb{N}$. Take $\rho = \|x_0\| + \delta$ and observe that $x_0 \in \operatorname{extr} \mathcal{F}_\rho(y_0)$ according to (3.9). $\mathcal{F}_\rho$ is lsc and closed at y_0, and so, by case 1, $\operatorname{extr} \mathcal{F}_\rho$ is lsc at y_0, which implies that there exists a sequence $\{x_r\}$ such that $x_r \to x_0$, $x_r \in \operatorname{extr} \mathcal{F}_\rho(y_r)$, and $\|x_r\| < \rho$ for r large enough. This yields the contradiction $\mathcal{E}(y_r) \cap B(x_0; \delta) \neq \emptyset$. □

4. Upper semicontinuity and closedness. In contrast with the lower semicontinuity, the closedness of a set-valued mapping $\mathcal{A}$ is not inherited by $\operatorname{conv}\mathcal{A}$ (even though $\mathcal{A}=\operatorname{bd}\mathcal{F},\operatorname{rbd}\mathcal{F},\operatorname{extr}\mathcal{F}$, as Example 3 in [5] shows). On the other hand, Proposition 4 in [5] establishes that, if $\operatorname{bd}\mathcal{F}$ is usc at y_0, then $\mathcal{F}$ is usc at y_0. In this section we shall prove that a similar statement holds for $\operatorname{rbd}\mathcal{F}$, but not for $\operatorname{extr}\mathcal{F}$ even though $\operatorname{extr}\mathcal{F}$ is either locally bounded or closed (nevertheless, according to the next Theorem 4.3, these two properties together entail the upper semicontinuity and the closedness of $\mathcal{F}$).

Example 4.1. Let $\mathcal{E}:Y\rightrightarrows\mathbb{R}^2$, where $Y=[2,+\infty[$ and

$$\mathcal{E}(y)=\left\{x\in\mathbb{R}^2\mid\|x\|=1,x_1<y^{-1}\right\}\cup\{(y,0)\}\ \text{ for all } y\in Y.$$

It is easy to see that $\mathcal{E}$ is locally bounded and continuous but not closed at $y_0=2$, and that it is the extreme points set mapping of $\mathcal{F}=\operatorname{conv}\mathcal{E}$. We shall prove that $\mathcal{F}$ is not usc at y_0. Let

$$W:=\{x\in\mathbb{R}^2\mid\sqrt{3}\,|x_2|<2-x_1,x_1<2\}\cup B\left((2,0);\frac{1}{2}\right),$$

$\mathcal{F}(y_0)\subset W$. If $y>2$, then $\overline{x}=(1,\frac{1}{\sqrt{3}})\in\mathcal{F}(y)\setminus W$. Observe also that $\mathcal{F}$ cannot be closed at y_0 (because $\mathcal{F}(y_0)$ is not closed).

Example 4.2. Let $\mathcal{E}:\mathbb{R}\rightrightarrows\mathbb{R}^3$ be such that

$$\mathcal{E}(y)=\left\{(x_1,x_2,0)\in\mathbb{R}^3\mid x_2=x_1^2\right\}\cup\{(0,0,y)\}\ \text{ for all } y\in\mathbb{R}.$$

As in the previous example, $\mathcal{E}=\operatorname{extr}\mathcal{F}$ for $\mathcal{F}=\operatorname{conv}\mathcal{E}$ and $\mathcal{E}$ is continuous at $y_0=0$, but now $\mathcal{E}$ is also closed and $\mathcal{E}(y_0)$ is unbounded. In order to prove that $\mathcal{F}$ is not usc at y_0, let us consider the convex plane set $C:=\left\{x\in\mathbb{R}^2\mid x_2\geq x_1^2\right\}$ and the open set

$$W:=\mathbb{R}^3\setminus\left\{x\in\mathbb{R}^3\mid x_3\geq x_2^{-1},x_2>0\right\}.$$

Obviously, $\mathcal{F}(y_0)=C\times\{0\}\subset W$. Moreover, if $y>0$ and $y>4/r^2$ for $0\neq r\in\mathbb{R}$, we have

$$\left(0,\frac{r^2}{2},\frac{y}{2}\right)=\frac{1}{2}(0,0,y)+\frac{1}{4}[(-r,r^2,0)+(r,r^2,0)]\in\mathcal{F}(y)\setminus W,$$

so that $\mathcal{F}(y)\not\subseteq W$. Hence $\mathcal{F}$ is not usc at y_0.

Finally, we show that $\mathcal{F}$ is closed at y_0. Let $y_r\to y_0$ and $x^r\to x^0$ be such that $x^r\in\mathcal{F}(y_r)$, $r=1,2,\ldots$. Since $\mathcal{F}(y_r)=\operatorname{conv}[(C\times\{0\})\cup\{(0,0,y_r)\}]$, for any $r\in\mathbb{N}$, we can write

$$x^r=\lambda_r(c^r,0)+(1-\lambda_r)(0,0,y_r)=(\lambda_r c^r,(1-\lambda_r)y_r),c^r\in C,0\leq\lambda_r\leq 1.$$

Observe that $c^r\in C$ and $(0,0)\in C$ entail $\lambda_r c^r\in C$. On the other hand, $x_3^r=(1-\lambda_r)y_r\in\operatorname{conv}\{0,y_r\}$. Taking limits we get $x^0=\lim_r x^r\in C\times\{0\}=\mathcal{F}(y_0)$.

The next result is a reformulation of a well-known result ([1, Lemma 1.1.9]), taking into account the mentioned equivalence between closedness and outer semicontinuity.

THEOREM 4.3. *If $\mathcal{A}:Y\rightrightarrows\mathbb{R}^n$ is closed and locally bounded at $y_0\in\operatorname{dom}\mathcal{A}$, then $\operatorname{conv}\mathcal{A}$ is closed and usc at y_0.*

Observe that it is not possible to replace in Theorem 4.3 above the condition "$\mathcal{A}$ is closed and locally bounded at y_0" by just "$\mathcal{A}$ is closed and usc at y_0" (recall Example 4.2).

Given two set-valued mappings $\mathcal{M}, \mathcal{N} : Y \rightrightarrows \mathbb{R}^n$, we say that $\mathcal{M}$ is *contained* in $\mathcal{N}$ (in brief, $\mathcal{M} \subset \mathcal{N}$) *locally at* y_0 if there exists an open set $V \subset Y$, containing y_0, such that $\mathcal{M}(y) \subset \mathcal{N}(y)$ for all $y \in V$. We also define the *closure* of $\mathcal{M}$ as the mapping $\operatorname{cl}\mathcal{M} : Y \rightrightarrows \mathbb{R}^n$ such that $(\operatorname{cl}\mathcal{M})(y) = \operatorname{cl}\mathcal{M}(y)$ for all $y \in Y$.

COROLLARY 4.4. *Let $\mathcal{A} : Y \rightrightarrows \mathbb{R}^n$ and let $y_0 \in \operatorname{dom}\mathcal{A}$ be such that $\mathcal{A}(y_0)$ is bounded and $\mathcal{A}$ is usc at y_0. Then each of the following conditions guarantees that* $\operatorname{conv}\mathcal{A}$ *is closed and usc at* y_0*:*
(i) $\mathcal{A}(y_0)$ *is closed.*
(ii) $\operatorname{cl}\mathcal{A} \subset \operatorname{conv}\mathcal{A}$ *locally at* y_0.

Proof. (i) Since $\mathcal{A}$ is usc at y_0 and $\mathcal{A}(y_0)$ is bounded, then $\mathcal{A}$ is locally bounded at y_0. The conclusion follows from Theorem 4.3.

(ii) First we prove that $\operatorname{cl}\mathcal{A}$ is usc at y_0. In fact, given an open set W such that $\operatorname{cl}\mathcal{A}(y_0) \subset W$, we have

$$\mathcal{A}(y_0) \subset U := \operatorname{cl}\mathcal{A}(y_0) + B(0_n; \varepsilon),$$

where

$$\varepsilon := \frac{1}{2} d\left(\operatorname{cl}\mathcal{A}(y_0), \mathbb{R}^n \setminus W\right) > 0.$$

Since U is open, there exists an open set $V \subset Y$, $y_0 \in V$, such that $\mathcal{A}(y) \subset U$ for all $y \in V$. Then $\operatorname{cl}\mathcal{A}(y) \subset \operatorname{cl} U \subset W$.

Now we show that $\operatorname{conv}\mathcal{A}$ is usc at y_0.

Since $\operatorname{cl}\mathcal{A}$ is usc at y_0 and $\operatorname{cl}\mathcal{A}(y_0)$ is compact we can assert, applying statement (i) to $\operatorname{cl}\mathcal{A}$, that $\operatorname{conv}\operatorname{cl}\mathcal{A}$ is closed and usc at y_0. Since the assumption implies that $\operatorname{conv}\operatorname{cl}\mathcal{A} = \operatorname{conv}\mathcal{A}$ locally at y_0, we conclude that $\operatorname{conv}\mathcal{A}$ is closed and usc at y_0. ∎

The boundedness assumption in Corollary 4.4 is not superfluous even for the extreme points set mapping (recall again Example 4.2, where (i) holds).

Now, we give a condition that assures that if $\mathcal{A}$ is usc at y_0, then $\operatorname{conv}\mathcal{A}$ is usc at y_0 as well.

PROPOSITION 4.5. *Let $\mathcal{A} : Y \rightrightarrows \mathbb{R}^n$ and let $y_0 \in \operatorname{dom}\mathcal{A}$ be such that*

$$\operatorname{rbd}\operatorname{conv}\mathcal{A} \subset \mathcal{A} \subset \operatorname{conv}\operatorname{rbd}\operatorname{conv}\mathcal{A}$$

locally at y_0 and $\operatorname{conv}\mathcal{A}$ *is closed at y_0. If $\mathcal{A}$ is usc at y_0, then* $\operatorname{conv}\mathcal{A}$ *is usc at y_0.*

Proof. Let $\mathcal{F} := \operatorname{conv}\mathcal{A}$ and let $\mathcal{R} = \operatorname{rbd}\mathcal{F}$. We assume that $\mathcal{A}$ is usc at y_0.

Let V_1 be a neighborhood of y_0 such that $\mathcal{R}(y) \subset \mathcal{A}(y) \subset \operatorname{conv}\mathcal{R}(y)$ for all $y \in V_1$. Then we have $\mathcal{F}(y) = \operatorname{conv}\mathcal{R}(y)$ for all $y \in V_1$.

By Proposition 2.3, there exists $\overline{\rho} > 0$ and a neighborhood of y_0, $V_2 \subset V_1$, such that

$$\mathcal{A}(y) \setminus \mathcal{A}_{\overline{\rho}}(y) \subset \mathcal{A}(y_0) \setminus \mathcal{A}_{\overline{\rho}}(y_0) \quad \text{for all } y \in V_2. \tag{4.1}$$

We shall prove that we can replace $\mathcal{A}$ with $\mathcal{F}$ in (4.1), so that $\mathcal{F}$ will be usc at y_0 because $\mathcal{F}$ is closed at y_0 (again by Proposition 2.3). Let $\overline{y} \in V_2$ and $\overline{x}$ be such that

$$\overline{x} \in \mathcal{F}(\overline{y}) \text{ and } \|\overline{x}\| > \overline{\rho}.$$

If $\overline{x} \in \mathcal{A}(\overline{y})$, then $\overline{x} \in \mathcal{A}(\overline{y}) \setminus \mathcal{A}_{\overline{\rho}}(\overline{y})$ and so

$$\overline{x} \in \mathcal{A}(y_0) \setminus \mathcal{A}_{\overline{\rho}}(y_0) \subset \mathcal{A}(y_0) \subset \mathcal{F}(y_0).$$

Suppose that $\overline{x} \notin \mathcal{A}(\overline{y})$ and $\overline{x} \notin \mathcal{F}(y_0)$. Now, $\mathcal{R}(\overline{y}) \subset \mathcal{A}(\overline{y})$ implies that

$$\overline{x} \in \mathcal{F}(\overline{y}) \setminus \mathcal{A}(\overline{y}) \subset \mathcal{F}(\overline{y}) \setminus \mathcal{R}(\overline{y}) = \operatorname{rint} \mathcal{F}(\overline{y}). \tag{4.2}$$

Since $\mathcal{F}(y_0)$ is closed and convex, there exist $a \neq 0_n$ and a scalar α such that

$$a'\overline{x} = \alpha \text{ and } a'x < \alpha \text{ for all } x \in \mathcal{F}(y_0). \tag{4.3}$$

Consider the flat $H := \{x \in \operatorname{aff} \mathcal{F}(\overline{y}) \mid a'x = \alpha\}$. Obviously $a'c = 0$ for all $c \in H - \overline{x}$ (the linear subspace parallel to H).

We shall get a contradiction if we are able to prove that $H \subset \mathcal{F}(\overline{y})$. In fact, in this case if $a'x = \alpha$ for all $x \in \operatorname{aff} \mathcal{F}(\overline{y})$, then $H = \operatorname{aff} \mathcal{F}(\overline{y})$ and so $\mathcal{F}(\overline{y}) = \operatorname{aff} \mathcal{F}(\overline{y})$, i.e., $\mathcal{F}(\overline{y})$ is a flat. Otherwise $\mathcal{F}(\overline{y})$ is a half-flat. In both cases $\mathcal{F}(\overline{y}) \neq \operatorname{conv} \mathcal{R}(\overline{y})$ despite of $\overline{y} \in V_1$.

In order to prove that $H \subset \mathcal{F}(\overline{y})$ we associate with each $c \in (H - \overline{x}) \setminus \{0_n\}$ the halfline $S(c) := \{\overline{x} + \lambda c \mid \lambda \geq 0\} \subset H$. Now we prove that

$$S(c) \cap \operatorname{cl} B(0_n; \overline{\rho}) = \emptyset \Rightarrow S(c) \subset \operatorname{rint} \mathcal{F}(\overline{y}). \tag{4.4}$$

Assume that $S(c) \cap \operatorname{cl} B(0_n; \overline{\rho}) = \emptyset$ and $S(c) \not\subseteq \operatorname{rint} \mathcal{F}(\overline{y})$. By (4.2) we have

$$0 < \overline{\lambda} := \sup \{\lambda \in \mathbb{R}_+ \mid \overline{x} + \lambda c \in \operatorname{rint} \mathcal{F}(\overline{y})\} < +\infty.$$

Thus $\overline{x} + \overline{\lambda} c \in \mathcal{R}(\overline{y}) \subset \mathcal{A}(\overline{y})$ and, by (4.1), we have

$$\begin{aligned} \overline{x} + \overline{\lambda} c \in \mathcal{A}(\overline{y}) \setminus \operatorname{cl} B(0_n; \overline{\rho}) &= \mathcal{A}(\overline{y}) \setminus \mathcal{A}_{\overline{\rho}}(\overline{y}) \\ &\subset \mathcal{A}(y_0) \setminus \mathcal{A}_{\overline{\rho}}(y_0) \subset \mathcal{F}(y_0), \end{aligned}$$

so that by (4.3) $a'\overline{x} = \alpha$ and $a'\overline{x} = a'(\overline{x} + \overline{\lambda} c) < \alpha$. This is a contradiction.

Finally, we prove that $H \subset \mathcal{F}(\overline{y})$ by means of a discussion based on the set $C := H \cap \operatorname{cl} B(0_n; \overline{\rho})$.

If $C = \emptyset$, then H is the union of halflines emanating from $\overline{x}$ in all directions parallel to H, and these halflines are contained in $\operatorname{rint} \mathcal{F}(\overline{y})$, according to (4.4). Then $H \subset \operatorname{rint} \mathcal{F}(\overline{y}) \subset \mathcal{F}(\overline{y})$.

If $|C| = 1$, then all the halflines mentioned above are contained in $\operatorname{rint} \mathcal{F}(\overline{y})$, except one. Thus $H \subset \mathcal{F}(\overline{y})$.

If $|C| > 1$, then C is a closed ball in H and all the halflines in H emanating from $\overline{x}$ which do not meet C are contained in $\operatorname{rint} \mathcal{F}(\overline{y})$. Then $\mathcal{F}(\overline{y})$ contains the complement, relative to H, of a pointed cone with apex $\overline{x}$. Hence we have again $H \subset \mathcal{F}(\overline{y})$. □

Given $\mathcal{A} : Y \rightrightarrows \mathbb{R}^n$ and $\rho > 0$, we denote by $\mathcal{A}_\rho$ and by $(\operatorname{conv} \mathcal{A})_\rho$ the truncated mappings of $\mathcal{A}$ and $\operatorname{conv} \mathcal{A}$, respectively, with radius ρ. We also define the mapping $\mathcal{A}^\rho : Y \rightrightarrows \mathbb{R}^n$ such that

$$\mathcal{A}^\rho(y) = \mathcal{A}_\rho(y) \cup \{x \in \operatorname{conv} \mathcal{A}(y) \mid \|x\| = \rho\}.$$

If $\mathcal{F} = \operatorname{conv} \operatorname{rbd} \mathcal{F}$ ($\mathcal{F} = \operatorname{conv} \operatorname{bd} \mathcal{F}$), and $\mathcal{A} = \operatorname{rbd} \mathcal{F}$ ($\mathcal{A} = \operatorname{bd} \mathcal{F}$, respectively), then $(\operatorname{conv} \mathcal{A})_\rho = \operatorname{conv} \mathcal{A}^\rho$. The inclusion $(\operatorname{conv} \mathcal{A})_\rho \subset \operatorname{conv} \mathcal{A}^\rho$ follows from the fact that any convex combination $x = (1 - \lambda) u + \lambda v$, $0 \leq \lambda \leq 1$, $x, u, v \in \operatorname{conv} \mathcal{A}(y)$,

$\|x\| \leq \rho$ and $\|v\| > \rho$, can be expressed as $x = (1-\alpha)\,u + \alpha w$, where $0 \leq \alpha \leq 1$ and $w \in [x, v] \subset \operatorname{conv} \mathcal{A}(y)$, with $\|w\| = \rho$.

LEMMA 4.6. *Let $\mathcal{A} : Y \rightrightarrows \mathbb{R}^n$ and let $y_0 \in \operatorname{dom} \mathcal{A}$ be such that $\mathcal{A}(y_0)$ and $\operatorname{conv} \mathcal{A}(y_0)$ are closed and $\mathcal{A}$ is usc at y_0. Then $\{\rho > 0 \mid \mathcal{A}^\rho$ is closed at $y_0\}$ is unbounded.*

Proof. We will prove that, under the assumptions, $\mathcal{A}^\rho$ is closed at y_0 for all $\rho \in I := \{\rho > 0 \mid \mathcal{A}_\rho(y_0) \neq \emptyset\}$ (I is a halfline). We denote $\mathcal{F} = \operatorname{conv} \mathcal{A}$.

Let $\rho \in I$, $y_k \to y_0$ and $x_k \to x_0$ be such that $x_k \in \mathcal{A}^\rho(y_k)$, $k = 1, 2, \ldots$.

Since $\mathcal{A}^\rho(y_k) \subset \operatorname{cl} B(0_n; \rho)$ for all $k \in \mathbb{N}$, $\|x_0\| \leq \rho$.

If there exists an increasing sequence $\{k_r\} \subset \mathbb{N}$ such that $x_{k_r} \in \mathcal{A}(y_{k_r})$, $r = 1, 2, \ldots$, then $x_0 \in \mathcal{A}(y_0)$ (because $\mathcal{A}$ is closed at y_0) and so $x_0 \in \mathcal{A}_\rho(y_0) \subset \mathcal{A}^\rho(y_0)$.

Thus we can assume without loss of generality that $x_k \notin \mathcal{A}(y_k)$, $k = 1, 2, \ldots$.

Given $k \in \mathbb{N}$, we have $x_k \in \mathcal{A}^\rho(y_k) \setminus \mathcal{A}_\rho(y_k) \subset \{x \in \mathcal{F}(y_k) \mid \|x\| = \rho\}$. Since $\|x_k\| = \rho$ for all k, we have $\|x_0\| = \rho$.

If $x_0 \in \mathcal{F}(y_0)$, then $x_0 \in \mathcal{A}^\rho(y_0)$ and we have finished. So we assume that $x_0 \notin \mathcal{F}(y_0)$. Since this set is closed, $\varepsilon := \frac{1}{2} d(x_0, \mathcal{F}(y_0)) > 0$. Let us consider the open convex set $W := \mathcal{F}(y_0) + B(0_n; \varepsilon)$. Since $\mathcal{A}(y_0) \subset \mathcal{F}(y_0) \subset W$ and $\mathcal{A}$ is usc at y_0, there exists a neighborhood of y_0, say V, such that $\mathcal{A}(y) \subset W$ for all $y \in V$. Then, taking convex hulls, we get $\mathcal{F}(y) \subset W$ for all $y \in V$.

Let $k_0 \in \mathbb{N}$ be such that $y_k \in V$ for all $k \geq k_0$. For such a k we have $x_k \in \mathcal{A}^\rho(y_k) \subset \mathcal{F}(y_k) \subset W$ whereas $x_0 \notin \mathcal{F}(y_0)$, so that $d(x_k, x_0) \geq \varepsilon$. This contradicts $x_k \to x_0$. ∎

LEMMA 4.7. *Let $\mathcal{A} : Y \rightrightarrows \mathbb{R}^n$ be such that $(\operatorname{conv} \mathcal{A})_\rho = \operatorname{conv} \mathcal{A}^\rho$ for all $\rho > 0$ sufficiently large and let $y_0 \in \operatorname{dom} \mathcal{A}$ such that $\{\rho > 0 \mid \mathcal{A}^\rho$ is closed at $y_0\}$ is unbounded. Then $\operatorname{conv} \mathcal{A}$ is closed at y_0.*

Proof. Let $\mathcal{F} := \operatorname{conv} \mathcal{A}$. Let $y_r \to y_0$ and $x_r \to x_0$ be such that $x_r \in \mathcal{F}(y_r)$, $r = 1, 2, \ldots$.

Since the convergent sequence $\{x_r\}$ is bounded, and by the assumptions on $\{\mathcal{A}^\rho \mid \rho > 0\}$, there exists $\rho > 0$ such that $\|x_r\| \leq \rho$ for all $r \in \mathbb{N}$, $\mathcal{F}_\rho = \operatorname{conv} \mathcal{A}^\rho$ and $\mathcal{A}^\rho$ is closed at y_0. Since $\mathcal{A}^\rho$ is closed and locally bounded at y_0, by Theorem 4.3, $\mathcal{F}_\rho = \operatorname{conv} \mathcal{A}^\rho$ is closed and usc at y_0. Then, since $x_r \in \mathcal{F}_\rho(y_r)$ for all $r \in \mathbb{N}$, we have $x_0 \in \mathcal{F}_\rho(y_0) \subset \mathcal{F}(y_0)$. ∎

PROPOSITION 4.8. *Let $\mathcal{A} : Y \rightrightarrows \mathbb{R}^n$ be such that $(\operatorname{conv} \mathcal{A})_\rho = \operatorname{conv} \mathcal{A}^\rho$ for all $\rho > 0$ sufficiently large and let $y_0 \in \operatorname{dom} \mathcal{A}$ such that $\mathcal{A}(y_0)$ is closed,*

$$\operatorname{rbd} \operatorname{conv} \mathcal{A} \subset \mathcal{A} \subset \operatorname{conv} \operatorname{rbd} \operatorname{conv} \mathcal{A}$$

locally at y_0 and $\mathcal{A}$ is usc at y_0. Then $\operatorname{conv} \mathcal{A}$ is usc at y_0.

Proof. By assumption $\operatorname{rbd} \operatorname{conv} \mathcal{A}(y_0) \subset \mathcal{A}(y_0) \subset \operatorname{conv} \mathcal{A}(y_0)$, so that $\operatorname{conv} \mathcal{A}(y_0)$ is closed. Then, by Lemma 4.6, $\{\rho > 0 \mid \mathcal{A}^\rho$ is closed at $y_0\}$ is unbounded and, by Lemma 4.7, $\operatorname{conv} \mathcal{A}$ is closed at y_0. We conclude that $\operatorname{conv} \mathcal{A}$ is usc at y_0 by Proposition 4.5. ∎

THEOREM 4.9. *Let $\mathcal{F} : Y \rightrightarrows \mathbb{R}^n$ be such that $\mathcal{F} = \operatorname{conv} \operatorname{rbd} \mathcal{F}$ and $\operatorname{rbd} \mathcal{F}$ is usc at $y_0 \in \operatorname{dom} \mathcal{F}$. Then $\mathcal{F}$ is usc at y_0.*

Proof. It is a straightforward consequence of Proposition 4.8, taking $\mathcal{A} = \operatorname{rbd} \mathcal{F}$. ∎

The last four results are also valid replacing "rbd" everywhere with "bd" (see [5]). The final example illustrates the results in sections 3 and 4 and shows that there is no usc counterpart for Theorems 3.3 and 3.4.

Example 4.10. Let us identify the complex field $\mathbb{C}$ with $\mathbb{R}^2$ and let us take as Y the set of polynomials of degree $q \in \mathbb{N}$ (fixed) with complex coefficients equipped with the Euclidean distance on $\mathbb{R}^{2q+2}$. Given $y \in Y$, we denote by $\mathcal{A}(y)$ its set of complex zeros and by $\mathcal{F}(y)$ its convex hull, i.e., the polytope $\mathcal{F}(y) = \operatorname{conv} \mathcal{A}(y)$. By the fundamental theorem of algebra, $\mathcal{A}(y) \neq \emptyset$ for all $y \in Y$, so that $\operatorname{dom} \mathcal{A} = Y$. Let us denote by $\mathcal{B}$, $\mathcal{R}$, and $\mathcal{E}$ the boundary mapping, the relative boundary mapping, and the extreme points set mapping of $\mathcal{F}$, respectively. By Proposition 2.4, we have

$$\mathcal{F} = \operatorname{conv} \mathcal{B} = \operatorname{conv} \mathcal{R} = \operatorname{conv} \mathcal{E}.$$

$\mathcal{A}$ is lsc and usc as a consequence of a well-known consequence of Rolle's theorem for complex polynomials (see, e.g., [7]) and, since it has closed images, it is also closed. By Theorem 3.1 and Corollary 4.4, $\mathcal{F}$ is also lsc, usc, and closed. Consequently, $\mathcal{B}$, $\mathcal{R}$, and $\mathcal{E}$ are lsc by Propositions 1 in [4] and Theorems 3.3 and 3.4 in this paper (the direct proofs of these statements are rather involved). Now we show that $\mathcal{R}$ and $\mathcal{E}$ are neither usc nor closed if $q = 3$.

Let $y_0 = x^3 + x$, with $\mathcal{A}(y_0) = \{0, \pm i\}$, and let $y_r = x^3 - \frac{2}{r}x^2 + (1 + \frac{1}{r^2})x$, with $\mathcal{A}(y_r) = \left\{0, \frac{1}{r} \pm i\right\}$, $r = 1, 2, \ldots$. Obviously, $y_r \to y_0$. Taking the constant sequence $x_r = 0$, $r = 1, 2, \ldots$, we have $x_r \in \mathcal{E}(y_r) \subset \mathcal{F}(y_r)$ for all r, whereas $0 \notin \mathcal{E}(y_0) = \mathcal{R}(y_0) = \{\pm i\}$. Thus neither $\mathcal{R}$ nor $\mathcal{E}$ is closed (usc) at y_0.

Acknowledgment. The authors wish to thank the referees for their valuable comments and suggestions.

REFERENCES

[1] J.-P. Aubin and H. Frankowska, *Set-Valued Analysis*, Birkhäuser, Boston, 1990.
[2] B. Bank, J. Guddat, D. Klatte, B. Kummer, and K. Tammer, *Nonlinear Parametric Optimization*, Birkhäuser Verlag, Basel, Switzerland, 1983.
[3] M. A. Goberna, M. Larriqueta, and V. N. Vera de Serio, *On the stability of the boundary of the feasible set in linear optimization*, Set-Valued Anal., 11 (2003), pp. 203–223.
[4] M. A. Goberna, M. Larriqueta, and V. N. Vera de Serio, *On the stability of the extreme point set in linear optimization*, SIAM J. Optim., 15 (2005), pp. 1155–1169.
[5] M. A. Goberna, M. A. López, and M. I. Todorov, *On the stability of closed-convex-valued mappings and the associated boundaries*, J. Math. Anal. Appl., 306 (2005), pp. 502–515.
[6] E. Michael, *Continuous selections.* I., Ann. of Math., 63 (1956), pp. 361–382.
[7] J. E. Marsden and M. J. Hoffman, *Basic Complex Analysis*, W. H. Freeman, New York, 1987.
[8] R. Rockafellar and R. B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
[9] R. Webster, *Convexity*, Oxford University Press, Oxford, 1994.

SIAM J. OPTIM.
Vol. 17, No. 1, pp. 159–187
© 2006 Society for Industrial and Applied Mathematics

PATH-FOLLOWING METHODS FOR A CLASS OF CONSTRAINED MINIMIZATION PROBLEMS IN FUNCTION SPACE*

MICHAEL HINTERMÜLLER†‡ AND KARL KUNISCH‡

Abstract. Path-following methods for primal-dual active set strategies requiring a regularization parameter are introduced. Existence of a primal-dual path and its differentiability properties are analyzed. Monotonicity and convexity of the primal-dual path value function are investigated. Both feasible and infeasible approximations are considered. Numerical path-following strategies are developed and their efficiency is demonstrated by means of examples.

Key words. semismooth Newton methods, path-following methods, active set strategy, primal-dual methods

AMS subject classifications. 49M15, 49M37, 65K05, 90C33

DOI. 10.1137/040611598

1. Introduction.

Primal-dual active set strategies or, in some cases equivalently, semismooth Newton methods, were proved to be efficient methods for solving constrained variational problems in function space [1, 9, 10, 11, 12, 13]. In certain cases regularization is required, resulting in a family of approximating problems with more favorable properties than those of the original one, [12, 13]. In previous work [13] convergence, and in some cases rate of convergence, with respect to the regularization parameter was proved. In the numerical work the adaptation of these parameters was heuristic, however. The focus of the present investigation is on an efficient control of the regularization parameter in the primal-dual active set strategy for a class of constrained variational problems. To explain the involved issues we proceed mostly formally in this section and consider the problem

$$\left\{\begin{array}{l} \min \mathcal{J}(v) \quad \text{over } v \in X \\ \text{s.t. } G\,v \le \psi, \end{array}\right. \tag{1}$$

where $\mathcal{J}$ is a quadratic functional on a Hilbert space X, and $G\colon X \to Y$. It is assumed that $Y \subset L^2(\Omega)$ is a Hilbert lattice with ordering $\le$ induced by the natural ordering of $L^2(\Omega)$. We note that (1) subsumes problems of very different nature. For example, for the control constrained optimal control problem

$$\left\{\begin{array}{l} \min \frac{1}{2}|y-z|_{L^2}^2 + \frac{\alpha}{2}|u|_{L^2}^2 \\ \text{s.t. } -\Delta y = u \ \text{ in } \ \Omega, \quad y = 0 \ \text{ on } \ \partial\Omega, \\ \qquad u \le \psi \ \text{ a.e. in } \Omega, \end{array}\right.$$

with Ω a bounded domain in $\mathbb{R}^n$, $z \in L^2(\Omega)$, $\alpha > 0$, one can use $y = (-\Delta)^{-1}u$, where Δ denotes the Laplacian with homogenous Dirichlet boundary conditions, and arrive

*Received by the editors July 14, 2004; accepted for publication (in revised form) November 21, 2005; published electronically May 3, 2006. This research was partially supported by the Fonds zur Förderung der wissenschaftlichen Forschung under SFB 03 "Optimierung und Kontrolle."
http://www.siam.org/journals/siopt/17-1/61159.html
†Department of Computational and Applied Mathematics, Rice University, Houston, TX.
‡Institute of Mathematics and Scientific Computing, University of Graz, Graz, Austria (michael.hintermueller@uni-graz.at, karl.kunisch@uni-graz.at).

at

$$\begin{cases} \min \frac{1}{2}|(-\Delta)^{-1}u - z|^2 + \frac{\alpha}{2}|u|^2 \\ \text{s.t. } u \le \psi \text{ a.e. in } \Omega, \end{cases}$$

which is clearly of the form (1). For $\mathcal{J}(v) = \frac{1}{2}\int_\Omega |\nabla v|^2 dx - \int_\Omega f\, v$, $X = H_0^1(\Omega)$, and $G = I$ we obtain the classical obstacle problem. For state constrained control problems with $y \le \psi$ one has

$$\begin{cases} \min \frac{1}{2}|(-\Delta)^{-1}u - z|^2 + \frac{\alpha}{2}|u|^2 \\ \text{s.t. } (-\Delta)^{-1}u \le \psi \text{ a.e. in } \Omega, \end{cases}$$

which is also of the form (1). From the point of view of duality theory these three problems are very different. While it is straightforward to argue the existence of a Lagrange multiplier in $L^2(\Omega)$ for the control constrained optimal control problem, it is already more involved and requires additional assumptions to guarantee its existence in $L^2(\Omega)$ for obstacle problems, and for state constrained problems the Lagrange multiplier is only a measure. If we resort to a formal discussion, then in either of these cases we arrive at the optimality system of the form

$$\begin{cases} \mathcal{J}'(v) + G^*\lambda = 0, \\ \lambda = \max(0, \lambda + c(G(v) - \psi)\,) \end{cases} \tag{2}$$

for any fixed $c > 0$. Here, G^* denotes the adjoint of G. The second equation in (2) is equivalent to $\lambda \ge 0$, $G(v) \le \psi$, and $\lambda(G(v) - \psi) = 0$.

Continuing formally, the primal-dual active set strategy determines the active set at iteration level k by means of

$$\mathcal{A}_{k+1} = \{x \in \Omega \colon \lambda_k(x) + c(G(v_k)(x) - \psi(x)\,) > 0\},$$

assigns the inactive set $\mathcal{I}_{k+1} = \Omega \setminus \mathcal{A}_{k+1}$, and updates (v, λ) by means of

$$\begin{cases} \mathcal{J}'(v_{k+1}) + G^*\lambda_{k+1} = 0, \\ \lambda_{k+1} = 0 \ \text{ on } \ \mathcal{I}_{k+1}, \quad (G(v_{k+1}) - \psi)(x) = 0 \ \text{ for } \ x \in \mathcal{A}_{k+1}. \end{cases} \tag{3}$$

These auxiliary problems require special attention. For obstacle problems the constraint $v_{k+1} = \psi$ on $\mathcal{A}_{k+1}$ induces that the associated Lagrange multiplier λ_{k+1} is in general less regular than the Lagrange multiplier associated with $v \le \psi$ for the original problem; see, e.g., [13]. For problems with combined control and state constraints it may happen that due to the assignment on $\mathcal{I}_{k+1}$ and $\mathcal{A}_{k+1}$, (3) has no solution while the original problem does. For these reasons in, e.g., [9, 12, 13] the second equation in (2) was regularized, resulting in the family of equations

$$\begin{cases} \mathcal{J}'(v) + G^*\lambda = 0, \\ \lambda = \max(0, \bar{\lambda} + \gamma(G(v) - \psi)), \end{cases} \tag{4}$$

where $\bar{\lambda}$ is fixed, possibly $\bar{\lambda} = 0$, and $\gamma \in \mathbb{R}^+$. In the above-mentioned references it was shown that under appropriate conditions the solutions $(v_\gamma, \lambda_\gamma)$ to (4) exist, the quantity λ_γ enjoys extra regularity, and $(v_\gamma, \lambda_\gamma)$ converge to the solution of (2) as $\gamma \to \infty^+$.

In previous numerical implementations the increase of γ to infinity was heuristic. As the system (4) becomes increasingly ill-conditioned as γ tends to ∞, in this paper a framework for a properly controlled increase of γ-values will be developed in order to cope with the conditioning problem. In fact, in a typical algorithmic regime for solving (1) one uses the solution $(v_\gamma, \lambda_\gamma)$ to (4) for some γ as the initial guess for the solution to (4) for the updated γ-value $\gamma^+ > \gamma$. Typically, if $\gamma^+ \gg \gamma$, then $(v_\gamma, \lambda_\gamma)$ is only a poor approximation of $(v_{\gamma^+}, \lambda_{\gamma^+})$, which in addition to numerical linear algebra issues (like ill-conditioned system matrices) causes severe stability problems for iterative solvers for (4) such as semismooth Newton methods. Together with developing a new γ-update strategy, we aim at solving the auxiliary problems (4) only inexactly to keep the overall computational cost low. To this end we define neighborhoods of the path which allow inexact solutions and which contract in a controlled way towards the path as the iteration proceeds. Our work is inspired by concepts from path-following methods in finite dimensional spaces [4, 5, 16, 18, 19]. We first guarantee the existence of a sufficiently smooth path $\gamma \to (v_\gamma, \lambda_\gamma)$, with $\gamma \in (0, \infty)$ in appropriately chosen function spaces. Once the path is available it can be used as the basis for updating strategies of the path parameter. Given a current value γ_k, with associated primal and dual states $(v_{\gamma_k}, \lambda_{\gamma_k})$, the γ-update should be sufficiently large to make good progress towards satisfying the complementarity conditions. On the other hand, since we are not solving the problems along the path exactly, we have to use safeguards against steps which would lead us too far off the path. Of course, these goals are impeded by the fact that the path is not available numerically. To overcome this difficulty we use qualitative properties of the value function, like monotonicity and convexity, which can be verified analytically. These suggest the introduction of model functions which will be shown to approximate the value functional along the path very well. We use these model functions for our updating strategies of γ. In the case of exact path-following we can even prove convergence of the resulting strategy. In the present paper the program just described is carried out for a class of problems corresponding to contact problems. State constrained optimal control problems require a different approach that will be considered independently. As we shall see, the (infinite dimensional) parameter $\bar{\lambda}$ can be used to guarantee that the iterates of the primal variable are feasible. Further, it turns out that the numerical behavior of infeasible approximations is superior to the feasible ones from the point of view of iteration numbers.

Interior point methods also require an additional parameter, which, however, enters into (2) differently. For the problem under consideration here, the interior-point relaxation replaces the second equation in (2) by

$$\lambda(x)\,(\psi - G(v))(x) = \frac{1}{\gamma} \quad \text{for } x \in \Omega. \tag{5}$$

Path-following interior-point methods typically start strictly feasible, with iterates which are required to stay strictly feasible during the iterations while satisfying, or satisfying approximately, the first equation in (2) and (5). Path-following interior-point methods have not received much attention for infinite dimensional problems yet. In fact, we are aware of only [17], where such methods are analyzed for optimal control problems related to ordinary differential equations. For the problem classes that we outlined at the beginning of this section, the primal-dual active set strategy proved to be an excellent competitor to interior-point methods, as was demonstrated, for example, in [1] comparing these two methods.

This paper is organized as follows. Section 2 contains the precise problem formulation and the necessary background on the primal-dual active set strategy. The

existence and regularity of the primal-dual path is discussed in section 3. Properties of the primal-dual path value functional are analyzed in section 4. Section 5 contains the derivation of the proposed model functions for the primal-dual path value functional. Exact as well as inexact path-following algorithms are proposed in section 6, and their numerical behavior is discussed there as well.

2. Problem statement, regularization, and its motivation. We consider

$$\text{(P)} \qquad \begin{cases} \min \frac{1}{2}\, a(y,y) - (f,y) \quad \text{over} \quad y \in H_0^1(\Omega) \\ \text{s.t. } y \le \psi, \end{cases}$$

where $f \in L^2(\Omega)$, $\psi \in H^1(\Omega)$, with $\psi_{|\partial\Omega} \ge 0$, where Ω is a bounded domain in $\mathbb{R}^n$ with Lipschitz continuous boundary $\partial\Omega$. Throughout, $(\cdot,\cdot)$ denotes the standard $L_2(\Omega)$-inner product, and we assume that $a(\cdot,\cdot)$ is a bilinear form on $H_0^1(\Omega) \times H_0^1(\Omega)$ satisfying

$$a(v,v) \ge \nu |v|^2_{H_0^1} \ \text{ and } \ a(w,z) \le \mu |w|_{H^1} |z|_{H^1} \tag{6}$$

for some $\nu > 0$, $\mu > 0$ independent of $v \in H_0^1(\Omega)$ and $w, z \in H^1(\Omega)$. Here and throughout we use $|v|_{H_0^1} = |\nabla v|_{L^2}$ for $v \in H_0^1(\Omega)$, which defines a norm on $H_0^1(\Omega)$ due to Friedrichs' inequality, and $|w|_{H^1} = (|w|^2_{L^2} + |\nabla w|^2_{L^2})^{1/2}$ denotes the standard H^1-norm; see, e.g., [2]. Moreover, let $A\colon H_0^1(\Omega) \to H^{-1}(\Omega)$ be defined by

$$a(v,w) = \langle A\, v, w \rangle_{H^{-1}, H_0^1} \quad \text{for all } v, w \in H_0^1(\Omega).$$

It is well known that (P) admits a unique solution $y^* \in H_0^1(\Omega)$ with associated Lagrange multiplier $\lambda^* = -Ay^* + f$, satisfying the optimality system

$$\begin{cases} a(y^*, v) + \langle \lambda^*, v \rangle_{H^{-1}, H_0^1} = (f, v), \\ \langle \lambda^*, y^* - \psi \rangle_{H^{-1}, H_0^1} = 0, \quad y^* \le \psi, \langle \lambda^*, v \rangle \le 0 \ \text{ for all } \ v \le 0. \end{cases} \tag{7}$$

This also holds with $f \in H^{-1}(\Omega)$. Under well-known additional requirements on a, ψ, and Ω, as for example

$$\begin{cases} a(v,w) = \int_\Omega (\sum a_{ij} v_{x_i} w_{x_j} + d\, v\, w), \ \text{ with } \ a_{ij} \in C^1(\bar\Omega), d \in L^\infty(\Omega), \\ d \ge 0, \quad \psi \in H^2(\Omega), \partial\Omega \quad \text{is } \ C^{1,1}, \ \text{ or } \ \Omega \ \text{ is a convex polyhedron}, \end{cases} \tag{8}$$

we have $(y^*, \lambda^*) \in H^2(\Omega) \times L^2(\Omega)$, and the optimality system can be expressed as

$$\begin{cases} A\, y^* + \lambda^* = f \ \text{ in } \quad L^2(\Omega), \\ \lambda^* = (\lambda^* + c(y^* - \psi))^+ \quad \text{for some } \ c > 0, \end{cases} \tag{9}$$

where $(v)^+ = \max(0, v)$; for details see, e.g., [14].

Our aim is the development of Newton-type methods for solving (7) or (9), which is complicated by the system of inequalities in (7) and the nondifferentiable max-operator in (9). In the recent past significant progress was made in the investigation of semismooth Newton methods and primal-dual active set methods for coping with nondifferentiable functionals in infinite dimensional spaces; see, for instance, [10, 15]. A direct application of these techniques to (9) results in the following algorithm.

ALGORITHM A.

(i) Choose $c > 0$, (y_0, λ_0); set $k = 0$.
(ii) Set $\mathcal{A}_{k+1} = \{x \in \Omega\colon \lambda_k(x) + c(y_k(x) - \psi(x)) > 0\}$.
(iii) Compute $y_{k+1} = \arg\min \{\frac{1}{2}\, a(y, y) - (f, y)\colon y = \psi \text{ on } \mathcal{A}_{k+1}\}$.
(iv) Let λ_{k+1} be the Lagrange multiplier associated with the constraint in (iii), with $\lambda_{k+1} = 0$ on $\Omega \setminus \mathcal{A}_{k+1}$.
(v) Set $k := k + 1$ and go to (ii).

The optimality system for the variational problem in (iii) is given by

$$\text{(10)} \qquad \begin{cases} a(y_{k+1}, v) + \langle \lambda_{k+1}, v\rangle_{H^{-1}, H_0^1} = (f, v) \ \text{ for all } \ v \in H_0^1(\Omega), \\ y_{k+1} = \psi \ \text{ on } \ \mathcal{A}_{k+1}, \quad \lambda_{k+1} = 0 \ \text{ on } \ \mathcal{I}_{k+1} = \Omega \setminus \mathcal{A}_{k+1}. \end{cases}$$

This corresponds to (3) in our introductory discussion. The Lagrange multiplier associated with the constraint $y = \psi$ on $\mathcal{A}_{k+1}$ is in general only a distribution in $H^{-1}(\Omega)$ and is not in $L^2(\Omega)$. In fact, λ_{k+1} is related to the jumps in the normal derivatives of y across the interface between $\mathcal{A}_{k+1}$ and $\mathcal{I}_{k+1}$ [13]. This complicates the convergence analysis for Algorithm A since the calculus of Newton (or slant) differentiability [10] does not apply. We note that these difficulties are not present if (7) or (9) is discretized. However, they are crucial for the treatment of infinite dimensional problems, and as such they are generic. Analogous difficulties arise for state constrained optimization problems, for inverse problems with BV-regularization, and for elasticity problems with contract and friction, to mention a few. This suggests the introduction of regularized problems, which in our case are chosen as

$$(P_\gamma) \qquad \min \frac{1}{2}\, a(y, y) - (f, y) + \frac{1}{2\gamma} \int_\Omega |(\bar\lambda + \gamma(y - \psi))^+|^2 \quad \text{over } y \in H_0^1(\Omega),$$

where $\gamma > 0$ and $\bar\lambda \in L^2(\Omega)$, $\bar\lambda \geq 0$ are fixed. For later use we denote the objective functional of (P_γ) by $J(y; \gamma)$. The choice of $\bar\lambda$ will be used to influence the feasibility of the solution y_γ of (P_γ). Using Lebesgue's bounded convergence theorem to differentiate the max under the integral in $J(y; \gamma)$, the first order optimality condition associated with (P_γ) is given by

$$(\mathrm{OC}_\gamma) \qquad \begin{cases} a(y_\gamma, v) + (\lambda_\gamma, v) = (f, v) \ \text{ for all } \ v \in H_0^1(\Omega), \\ \lambda_\gamma = (\bar\lambda + \gamma(y_\gamma - \psi))^+, \end{cases}$$

where $(y_\gamma, \lambda_\gamma) \in H_0^1(\Omega) \times L^2(\Omega)$. With (8) holding, we have $y_\gamma \in H^2(\Omega)$. The primal-dual active set strategy, or equivalently the semismooth Newton method, for (P_γ) is given next. For its statement and for later use we introduce $\chi_{\mathcal{A}^{k+1}}$, the characteristic function of the set $\mathcal{A}^{k+1} \subseteq \Omega$.

ALGORITHM B.

(i) Choose $\bar\lambda \geq 0$, (y_0, λ_0); set $k = 0$.
(ii) Set $\mathcal{A}_{k+1} = \{x \in \Omega\colon \bar\lambda(x) + \gamma(y_k(x) - \psi(x)) > 0\}$, $\mathcal{I}_{k+1} = \Omega \setminus \mathcal{A}_{k+1}$.
(iii) Solve for $y_{k+1} \in H_0^1(\Omega)$: $a(y_{k+1}, v) + ((\bar\lambda + \gamma(y_{k+1} - \psi))\chi_{\mathcal{A}_{k+1}}, v) = (f, v)$ for all $v \in H_0^1(\Omega)$.
(iv) Set

$$\lambda_{k+1} = \begin{cases} 0 & \text{on } \mathcal{I}_{k+1}, \\ \bar\lambda + \gamma(y_{k+1} - \psi) & \text{on } \mathcal{A}_{k+1}. \end{cases}$$

Algorithm B was analyzed in [13], where global as well as locally superlinear convergence for every fixed $\gamma > 0$ were established. However, the choice and adaptation (increase) of γ was heuristic in [13] and earlier work. The focus of the present investigation is the automatic adaptive choice of γ. We shall utilize the following two results, which we recall from [13] where the proofs can also be found.

PROPOSITION 2.1. *The solutions* $(y_\gamma, \lambda_\gamma)$ *to* (OC_γ) *converge to* (y^*, λ^*) *in the sense that* $y_\gamma \to y^*$ *strongly in* $H_0^1(\Omega)$ *and* $\lambda_\gamma \rightharpoonup \lambda^*$ *weakly in* $H^{-1}(\Omega)$ *as* $\gamma \to \infty$.

We say that a satisfies the weak maximum principle if for any $v \in H_0^1(\Omega)$

$$a(v, v^+) \le 0 \text{ implies } v^+ = 0. \tag{11}$$

PROPOSITION 2.2. *Assume that* (11) *holds and let* $0 < \gamma_1 \le \gamma_2 < \infty$.

(a) *In the* infeasible case, *i.e., for* $\bar\lambda = 0$, *we have* $y^* \le y_{\gamma_2} \le y_{\gamma_1}$.

(b) *In the* feasible case, *i.e., if*

$$\bar\lambda \ge 0 \ \text{ and } \ \langle \bar\lambda - f + A\psi, v\rangle_{H^{-1}, H_0^1} \ge 0 \ \text{ for all } \ v \in H_0^1(\Omega), \tag{12}$$

with $v \ge 0$, *then* $y_{\gamma_1} \le y_{\gamma_2} \le y^* \le \psi$.

3. The primal-dual path. In this section we introduce the primal-dual path and discuss its smoothness properties.

DEFINITION 3.1. *The family of solutions* $\mathcal{C} = \{(y_\gamma, \lambda_\gamma) \colon \gamma \in (0, \infty)\}$ *to* (OC_γ), *considered as subset of* $H_0^1(\Omega) \times H^{-1}(\Omega)$, *is called the primal-dual path associated with* (P).

For $r \ge 0$ we further set $\mathcal{C}_r = \{(y_\gamma, \lambda_\gamma) \colon \gamma \in [r, \infty)\}$, and with some abuse of terminology we also refer to $\mathcal{C}_r$ as a path. In the following lemma we denote by $\hat y$ the solution to the unconstrained problem

$$\min J(y) = \frac{1}{2}\, a(y, y) - (f, y) \quad \text{ over } y \in H_0^1(\Omega). \tag{$\hat{\mathrm{P}}$}$$

Subsequently, in connection with convergence of a sequence in function space we use the subscript "weak" together with the space to indicate convergence in the weak sense.

LEMMA 3.2. *For each* $r > 0$ *the path* $\mathcal{C}_r$ *is bounded in* $H_0^1(\Omega) \times H^{-1}(\Omega)$, *with* $\lim_{\gamma\to\infty}(y_\gamma, \lambda_\gamma) = (y^*, \lambda^*)$ *in* $H_0^1(\Omega) \times H^{-1}(\Omega)_{weak}$. *For* $\bar\lambda = 0$ *the path* $\mathcal{C}_0$ *is bounded in* $H_0^1(\Omega) \times H^{-1}(\Omega)$, *with* $\lim_{\gamma\to 0^+}(y_\gamma, \lambda_\gamma) = (\hat y, 0)$ *in* $H_0^1(\Omega) \times L^2(\Omega)$.

Proof. From (OC_γ) we have for every $\gamma > 0$

$$a(y_\gamma, y_\gamma - y^*) + (\lambda_\gamma, y_\gamma - y^*) = (f, y_\gamma - y^*). \tag{13}$$

Since $\lambda_\gamma = \max(0, \bar\lambda + \gamma(y_\gamma - \psi)) \ge 0$ and $\psi - y^* \ge 0$ we have

$$\begin{aligned}(\lambda_\gamma, y_\gamma - y^*) &= \left(\lambda_\gamma, \frac{\bar\lambda}{\gamma} + y_\gamma - \psi + \psi - y^* - \frac{\bar\lambda}{\gamma}\right)\\ &\ge \frac{1}{\gamma}(\lambda_\gamma, \bar\lambda + \gamma(y_\gamma - \psi)) - \frac{1}{\gamma}(\lambda_\gamma, \bar\lambda)\\ &= \frac{1}{\gamma}\left[|\lambda_\gamma|_{L^2}^2 - (\lambda_\gamma, \bar\lambda)\right].\end{aligned}$$

Combined with (13) this implies that

$$a(y_\gamma, y_\gamma) + \frac{1}{\gamma}\,|\lambda_\gamma|_{L^2}^2 \le a(y_\gamma, y^*) + (f, y_\gamma - y^*) + \frac{1}{\gamma}\,(\bar\lambda, \lambda_\gamma). \tag{14}$$

This estimate, (6), (OC_γ), and the Poincaré–Friedrichs inequality imply that $\mathcal{C}_r$ is bounded in $H_0^1(\Omega) \times H^{-1}(\Omega)$ for every $r > 0$. In fact, for $\omega > 0$ satisfying $\omega|y|_{H^1}^2 \leq |y|_{H_0^1}^2$, we have

$$\begin{aligned}
\omega|y_\gamma|_{H^1}^2 + \frac{1}{\gamma}\,|\lambda_\gamma|_{L^2}^2 &\leq a(y_\gamma, y_\gamma) + \frac{1}{\gamma}\,|\lambda_\gamma|_{L^2}^2 \\
&\leq \mu|y_\gamma|_{H^1}\,|y^*|_{H^1} + |f|_{H^{-1}}\,(|y_\gamma|_{H^1} + |y^*|_{H^1}) + \frac{1}{\gamma}\,|\bar\lambda|_{L^2}\,|\lambda_\gamma|_{L^2} \\
&\leq \frac{\omega}{4}\,|y_\gamma|_{H^1}^2 + \frac{\mu^2}{\omega}\,|y^*|_{H^1}^2 + \frac{\omega}{2}|y_\gamma|_{H^1}^2 + \frac{1}{2\omega}|f|_{H^{-1}}^2 \\
&\quad + \frac{1}{2\gamma}|\lambda_\gamma|_{L^2}^2 + \frac{1}{2\gamma}|\bar\lambda|_{L^2}^2 + |f|_{H^{-1}}\,|y^*|_{H^1},
\end{aligned}$$

and hence

$$\frac{\omega}{4}|y_\gamma|_{H^1}^2 + \frac{1}{2\gamma}\,|\lambda_\gamma|_{L^2}^2 \leq \frac{\mu^2}{\omega}|y^*|_{H^1}^2 + \frac{1}{2\omega}\,|f|_{H^{-1}} + |f|_{H^{-1}}\,|y^*|_{H^1} + \frac{1}{2\gamma}\,|\bar\lambda|_{L^2}^2.$$

This estimate implies that $\{y_\gamma : \gamma \geq r\}$ is bounded in $H_0^1(\Omega)$ for every $r > 0$. The first equation of (OC_γ) implies that $\{\lambda_\gamma : \gamma \geq r\}$ is bounded in $H^{-1}(\Omega)$ as well. From Proposition 2.1 we have that $\lim_{\gamma\to\infty}(y_\gamma, \lambda_\gamma) = (y^*, \lambda^*)$ in $H_0^1(\Omega) \times H^{-1}(\Omega)_{weak}$. If $\bar\lambda = 0$, then from (14), (6), and (OC_γ) the path $\mathcal{C}_o$ is bounded in $H_0^1(\Omega) \times H^{-1}(\Omega)$ and $\lambda_\gamma \to 0$ in $L^2(\Omega)$ for $\gamma \to 0^+$. From (OC_γ) and the optimality condition for ($\hat{\mathrm{P}}$) we have

$$a(y_\gamma - \hat y, y_\gamma - \hat y) + (\lambda_\gamma, y_\gamma - \hat y) = 0,$$

and hence $\lim_{\gamma\to 0^+} y_\gamma = \hat y$ in $H_0^1(\Omega)$. ∎

PROPOSITION 3.3. *The path $\mathcal{C}_r$ is globally Lipschitz in $H_0^1(\Omega)\times H^{-1}(\Omega)$ for every $r > 0$. If $\bar\lambda = 0$, then $\mathcal{C}_0$ is globally Lipschitz continuous.*

Proof. Let $\gamma, \bar\gamma \in [r, \infty)$ be arbitrary. Then

$$A(y_\gamma - y_{\bar\gamma}) + (\bar\lambda + \gamma(y_\gamma - \psi))^+ - (\bar\lambda + \bar\gamma(y_{\bar\gamma} - \psi))^+ = 0.$$

Taking the inner-product with $y_\gamma - y_{\bar\gamma}$ and using the monotonicity and Lipschitz continuity (with constant $L = 1$) of $x \mapsto \max(0, x)$, we find

$$\begin{aligned}
a(y_\gamma - y_{\bar\gamma},\quad & y_\gamma - y_{\bar\gamma}) \\
&\leq \left|\left((\bar\lambda + \gamma(y_\gamma - \psi))^+ - (\bar\lambda + \bar\gamma(y_{\bar\gamma} - \psi))^+, y_\gamma - y_{\bar\gamma}\right)\right| \\
&\leq |\gamma - \bar\gamma|\,|y_\gamma - \psi|_{L^2}|y_\gamma - y_{\bar\gamma}|_{L^2}.
\end{aligned}$$

By Lemma 3.2 the set $\{y_\gamma\}_{\gamma\geq r}$ is bounded in $H_0^1(\Omega)$. Hence there exists $K_1 > 0$ such that

$$\nu|y_\gamma - y_{\bar\gamma}|_{H_0^1}^2 \leq K_1|\gamma - \bar\gamma| \cdot |y_\gamma - y_{\bar\gamma}|_{L^2},$$

and by Poincaré's inequality there exists $K_2 > 0$ such that

$$|y_\gamma - y_{\bar\gamma}|_{H_0^1} \leq K_2|\gamma - \bar\gamma| \quad \text{for all} \quad \gamma \geq r, \bar\gamma \geq r.$$

Let us recall here that $|y|_{H_0^1} = |\nabla y|_{L^2}$. Lipschitz continuity of $\gamma \mapsto \lambda_\gamma$ from $[r,\infty)$ to $H^{-1}(\Omega)$ follows from the first equation in (OC_γ). For $\bar\lambda = 0$ the set $\{y_\gamma\}_{\gamma\geq 0}$ is bounded in $H_0^1(\Omega)$. The remainder of the proof remains identical. ∎

LEMMA 3.4. *For every subset $I \subset [r,\infty)$, $r > 0$, the mapping $\gamma \mapsto \lambda_\gamma$ is globally Lipschitz from I to $L^2(\Omega)$.*

Proof. For $0 < \gamma_1 \leq \gamma_2$ we have by (OC_γ)

$$\begin{aligned}|\lambda_{\gamma_1} - \lambda_{\gamma_2}|_{L^2} &= |(\bar\lambda + \gamma_1(y_{\gamma_1} - \psi))^+ - (\bar\lambda + \gamma_2(y_{\gamma_2} - \psi))^+|_{L^2} \\ &\leq (K_3\gamma_1 + K_1 + |\psi|_{L^2})\,|\gamma_1 - \gamma_2|\end{aligned}$$

for some constant $K_3 > 0$. ∎

We shall use the following notation:

$$S_\gamma = \{x \in \Omega\colon \bar\lambda(x) + \gamma(y_\gamma - \psi)(x) > 0\}.$$

Further we set

$$g(\gamma) = \bar\lambda + \gamma(y_\gamma - \psi). \tag{15}$$

Since $\gamma \mapsto y_\gamma \in H_0^1(\Omega)$ is Lipschitz continuous by Proposition 3.3, there exists a weak accumulation point $\dot y (= \dot y_\gamma)$ of $\frac{1}{\bar\gamma - \gamma}(y_{\bar\gamma} - y_\gamma)$ as $\bar\gamma \to \gamma > 0$, which is also a strong accumulation point in $L^2(\Omega)$. Further $\frac{1}{\bar\gamma - \gamma}(g(\bar\gamma) - g(\gamma))$ has $\dot g(\gamma) := y_\gamma - \psi + \gamma\,\dot y_\gamma$ as a strong accumulation point in $L^2(\Omega)$ as $\bar\gamma \to \gamma$. In case $\bar\gamma$ approaches γ from above (or below), the associated accumulation points $\dot y_\gamma^r$ (or $\dot y_\gamma^l$) satisfy certain properties which are described next. In what follows we use $\dot g^r(\gamma)$ or $\dot g^l(\gamma)$ whenever $\dot y_\gamma$ in $\dot g(\gamma)$ is replaced by $\dot y_\gamma^r$ and $\dot y_\gamma^l$, respectively.

PROPOSITION 3.5. *Let $\gamma > 0$, and denote by $\dot y_\gamma^r$ any weak accumulation point of $\frac{1}{\bar\gamma - \gamma}(y_{\bar\gamma} - y_\gamma)$ in $H_0^1(\Omega)$ as $\bar\gamma \downarrow \gamma$. Set*

$$S_\gamma^+ = S_\gamma \cup \{x\colon \bar\lambda(x) + \gamma(y_\gamma(x) - \psi(x)) = 0 \wedge \dot g^r(\gamma)(x) \geq 0\}.$$

Then $\dot y_\gamma^r$ satisfies

$$a(\dot y_\gamma^r, v) + ((y_\gamma - \psi + \gamma\dot y_\gamma^r)\chi_{S_\gamma^+}, v) = 0 \quad \textit{for all} \quad v \in H_0^1(\Omega). \tag{16}$$

Proof. By (OC_γ) we have for every $v \in H_0^1(\Omega)$

$$a(y_{\bar\gamma} - y_\gamma, v) + ((\bar\lambda + \bar\gamma(y_{\bar\gamma} - \psi))^+ - (\bar\lambda + \gamma(y_\gamma - \psi))^+, v) = 0. \tag{17}$$

We multiply (17) by $(\bar\gamma - \gamma)^{-1}$ and discuss separately the two terms in (17). Clearly, we have

$$\lim_{\bar\gamma \downarrow \gamma} (\bar\gamma - \gamma)^{-1} a(y_{\bar\gamma} - y_\gamma, v) = a(\dot y_\gamma^r, v).$$

Here and below the limit is taken on the sequence of $\bar\gamma$-values, which provides the accumulation point. Lebesgue's bounded convergence theorem allows us to consider the pointwise limits of the integrands. Considering separately the cases $g(\gamma)(x) < 0$, $g(\gamma)(x) > 0$, and $g(\gamma)(x) = 0$, we have

$$\begin{aligned}&(\bar\gamma - \gamma)^{-1}((g(\bar\gamma))^+ - (g(\gamma))^+, v) \\ &\quad \to ((y_\gamma - \psi + \gamma\,\dot y_\gamma^r)\chi_{S_\gamma^+}, v) \ \text{ as } \ \bar\gamma \downarrow \gamma,\end{aligned} \tag{18}$$

which ends the proof. ∎

As a consequence of the proof we obtain the following result.

COROLLARY 3.6. *Let $\gamma > 0$, and denote by $\dot{y}_\gamma^l$ any weak accumulation point of $\frac{1}{\bar\gamma-\gamma}(y_{\bar\gamma} - y_\gamma)$ in $H_0^1(\Omega)$ as $\bar\gamma \uparrow \gamma$. Set $S_\gamma^- = S_\gamma \cup \{x\colon \bar\lambda(x) + \gamma(y_\gamma(x) - \psi(x)) = 0 \wedge \dot g^l(\gamma)(x) \geq 0\}$. Then $\dot{y}_\gamma^l$ satisfies*

$$a(\dot{y}_\gamma^l, v) + ((y_\gamma - \psi + \gamma\,\dot{y}_\gamma^l)\chi_{S_\gamma^-}, v) = 0 \quad \text{for all } v \in H_0^1(\Omega). \tag{19}$$

Another corollary of Proposition 3.5 treats the case $\bar\lambda = 0$.

COROLLARY 3.7. *Let $\bar\lambda = 0$, and assume that (11) holds. Then the right- and left- derivatives $\dot{y}_\gamma^r$ and $\dot{y}_\gamma^l$ of $\gamma \mapsto y_\gamma$, $\gamma \in (0,\infty)$, exist and are given by*

$$a(\dot{y}_\gamma^r, v) + ((y_\gamma - \psi + \gamma\,\dot{y}_\gamma^r)\chi_{\{y_\gamma > \psi\}}, v) = 0 \quad \text{for all } v \in H_0^1(\Omega), \tag{20}$$

$$a(\dot{y}_\gamma^l, v) + ((y_\gamma - \psi + \gamma\,\dot{y}_\gamma^l)\chi_{\{y_\gamma \geq \psi\}}, v) = 0 \quad \text{for all } v \in H_0^1(\Omega). \tag{21}$$

Proof. Let $\bar\gamma \downarrow \gamma$. By Proposition 2.2 any accumulation point $\dot{y}_\gamma^r$ of $(\bar\gamma-\gamma)^{-1}(y_{\bar\gamma} - y_\gamma)$ satisfies $\dot{y}_\gamma^r \leq 0$ and hence

$$S_\gamma^+ = \{x \in \Omega\colon y_\gamma(x) > \psi(x)\} \cup \{x \in \Omega\colon y_\gamma(x) = \psi(x) \wedge \dot{y}_\gamma^r(x) = 0\}.$$

Observe that

$$(y_\gamma - \psi + \gamma\dot{y}_\gamma^r)\chi_{S_\gamma^+} = (y_\gamma - \psi + \gamma\dot{y}_\gamma^r)\chi_{\{y_\gamma > \psi\}}.$$

This implies that every accumulation point $\dot{y}_\gamma^r$ satisfies (20). Since the solution to (20) is unique, the directional derivative from the right exists.

Similarly, if $\bar\gamma \uparrow \gamma$, by Proposition 2.2 we have $S_\gamma^- = \{x \in \Omega\colon y_\gamma(x) \geq \psi(x)\}$, and (21) follows. ∎

Henceforth we set

$$S_\gamma^\circ = \{x \in \Omega\colon \bar\lambda(x) + \gamma(y_\gamma - \psi)(x) = 0\}.$$

COROLLARY 3.8. *If $\operatorname{meas}(S_\gamma^\circ) = 0$, then $\gamma \mapsto y_\gamma \in H_0^1(\Omega)$ is differentiable at γ, and the derivative $\dot{y}_\gamma$ satisfies*

$$a(\dot{y}_\gamma, v) + ((y_\gamma - \psi + \gamma\,\dot{y}_\gamma)\chi_{S_\gamma}, v) = 0 \quad \textit{for all } v \in H_0^1(\Omega). \tag{22}$$

Proof. Let z denote the difference of two accumulation points of $(\bar\gamma - \gamma)^{-1}(y_{\bar\gamma} - y_\gamma)$ as $\bar\gamma \to \gamma$. As a consequence of (16) and (19)

$$a(z, v) + \gamma(z\chi_{S_\gamma}, v) = 0 \quad \text{for all } v \in H_0^1(\Omega).$$

This implies that $z = 0$ by (6). Consequently, accumulation points are unique, and by (16), (19) they satisfy (22). ∎

The assumption $\operatorname{meas}(S_\gamma^o) = 0$ in Corollary 3.8 reflects the lack of differentiability of the max-operation in (OC_γ).

4. The primal-dual path value functional. In this section we investigate the value function associated with (P_γ) and study its monotonicity and smoothness properties.

DEFINITION 4.1. *The functional*

$$\gamma \mapsto V(\gamma) = J(y_\gamma;\gamma) = \frac{1}{2}a(y_\gamma, y_\gamma) - (f, y_\gamma) + \frac{1}{2\gamma}|(\bar\lambda + \gamma(y_\gamma - \psi))^+|_{L^2}^2$$

defined on $(0,\infty)$ *is called the primal-dual path value functional.*

Let us start by studying first order differentiability properties of V.

PROPOSITION 4.2. *The value function* V *is differentiable with*

$$\dot V(\gamma) = -\frac{1}{2\gamma^2}\int_\Omega |(\bar\lambda + \gamma(y_\gamma - \psi))^+|^2 \;+\; \frac{1}{\gamma}\int_\Omega (\bar\lambda + \gamma(y_\gamma - \psi))^+(y_\gamma - \psi).$$

COROLLARY 4.3. *For* $\bar\lambda = 0$ *we have* $\dot V(\gamma) = \frac{1}{2}\int_\Omega |(y_\gamma - \psi)^+|^2 \geq 0$ *and* $\dot V(\gamma) > 0$ *unless* y_γ *is feasible. For* $\bar\lambda$ *satisfying* (12) *and with* (11) *holding, we have* $y_\gamma \leq \psi$ *and hence* $\dot V(\gamma) \leq 0$ *for* $\gamma \in (0,\infty)$.

In either of the two cases $\dot V(\gamma) = 0$ *implies that* y_γ *solves* ($\hat{\mathrm{P}}$).

Proof. We show only that $\dot V(\gamma) = 0$ implies that y_γ solves ($\hat{\mathrm{P}}$). The rest of the assertion follows immediately from Proposition 4.2.

If $\bar\lambda = 0$, then $\dot V(\gamma) = 0$ yields $y_\gamma \leq \psi$. Thus, $\lambda_\gamma = 0$, and hence y_γ solves ($\hat{\mathrm{P}}$).

If (11) and (12) are satisfied, then $y_\gamma \leq \psi$ and $\dot V(\gamma) = 0$ implies $\gamma(y_\gamma - \psi) \leq \bar\lambda + \gamma(y_\gamma - \psi) \leq 0$. As a consequence $\lambda_\gamma = 0$, and y_γ solves ($\hat{\mathrm{P}}$). ∎

Proof of Proposition 4.2. For $\bar\gamma, \gamma \in (0,\infty)$ we find

$$\begin{aligned}\frac{1}{2}\, & a(y_{\bar\gamma} + y_\gamma, y_{\bar\gamma} - y_\gamma) - (f, y_{\bar\gamma} - y_\gamma) \\ & + \tfrac{1}{2}((\bar\lambda + \bar\gamma(y_{\bar\gamma} - \psi))^+ + (\bar\lambda + \gamma(y_\gamma - \psi))^+, y_{\bar\gamma} - y_\gamma) = 0,\end{aligned} \tag{23}$$

and consequently

$$\begin{aligned} V(\bar\gamma) - V(\gamma) &= \frac{1}{2}a(y_{\bar\gamma}, y_{\bar\gamma}) - \frac{1}{2}a(y_\gamma, y_\gamma) - (f, y_{\bar\gamma} - y_\gamma) \\ &\quad + \frac{1}{2\bar\gamma}\int_\Omega |(\bar\lambda + \bar\gamma(y_{\bar\gamma} - \psi))^+|^2 - \frac{1}{2\gamma}\int_\Omega |(\bar\lambda + \gamma(y_\gamma - \psi))^+|^2 \\ &= \frac{1}{2\bar\gamma}\int_\Omega |(\bar\lambda + \bar\gamma(y_{\bar\gamma} - \psi))^+|^2 + \frac{1}{2\gamma}\int_\Omega -|(\bar\lambda + \gamma(y_\gamma - \psi))^+|^2 \\ &\quad + \frac{1}{2}\int_\Omega -((\bar\lambda + \bar\gamma(y_{\bar\gamma} - \psi))^+ + (\bar\lambda + \gamma(y_\gamma - \psi))^+)(y_{\bar\gamma} - y_\gamma) \\ &= \int_{P_{\bar\gamma}\cap P_\gamma} z + \int_{P_{\bar\gamma}\cap N_\gamma} z + \int_{P_\gamma \cap N_{\bar\gamma}} z = \mathcal{I}_1 + \mathcal{I}_2 + \mathcal{I}_3, \end{aligned}$$

where z stands for the sum of the kernels on the left of the above equalities,

$$P_\gamma = \{x\colon \bar\lambda + \gamma(y_\gamma - \psi) > 0\}, \qquad N_\gamma = \{x\colon \bar\lambda + \gamma(y_\gamma - \psi) < 0\},$$

and $P_{\bar\gamma}$, $N_{\bar\gamma}$ are defined analogously. For $\mathcal{I}_2$ we have

$$\begin{aligned} |\mathcal{I}_2| &\leq \frac{1}{2}\int_{P_{\bar\gamma}\cap N_\gamma} \frac{1}{\bar\gamma}(\bar\lambda + \bar\gamma(y_{\bar\gamma} - \psi))^2 + |\bar\lambda + \bar\gamma(y_{\bar\gamma} - \psi)|\,|y_{\bar\gamma} - y_\gamma| \\ &\leq \frac{1}{2}\int_\Omega \frac{1}{\bar\gamma}(\bar\gamma(y_{\bar\gamma} - \psi) - \gamma(y_\gamma - \psi))^2 + |y_{\bar\gamma} - y_\gamma|(|\bar\gamma y_{\bar\gamma} - \gamma\, y_\gamma| + |\bar\gamma - \gamma|\,|\psi|), \end{aligned}$$

and hence by Proposition 3.3

$$\lim_{\bar\gamma\to\gamma} \frac{1}{\bar\gamma - \gamma} \quad |\mathcal{I}_2| = 0. \tag{24}$$

Analogously one verifies that

$$\lim_{\bar\gamma\to\gamma} \frac{1}{\bar\gamma-\gamma} \quad |\mathcal{I}_3| = 0. \tag{25}$$

On $P_{\bar\gamma}\cap P_\gamma$ we have

$$\begin{aligned} z &= \frac{1}{2\bar\gamma}\,(\bar\lambda+\bar\gamma(y_{\bar\gamma}-\psi))^2 \\ &\qquad -\frac{1}{2\gamma}\,(\bar\lambda+\gamma(y_\gamma-\psi))^2 - \frac{1}{2}\,(2\bar\lambda+\bar\gamma(y_{\bar\gamma}-\psi)+\gamma(y_\gamma-\psi))(y_{\bar\gamma}-y_\gamma) \\ &= \frac{\gamma-\bar\gamma}{2\bar\gamma\gamma}\,(\bar\lambda+\bar\gamma(y_{\bar\gamma}-\psi))^2 \\ &\qquad +\frac{1}{2\gamma}\,\left[2\bar\lambda(\bar\gamma(y_{\bar\gamma}-\psi)-\gamma(y_\gamma-\psi))+\bar\gamma^2(y_{\bar\gamma}-\psi)^2-\gamma^2(y_\gamma-\psi)^2\right] \\ &\qquad -\frac{1}{2}\,(2\bar\lambda+\bar\gamma(y_{\bar\gamma}-\psi)+\gamma(y_\gamma-\psi))(y_{\bar\gamma}-y_\gamma) \\ &= \frac{\gamma-\bar\gamma}{2\bar\gamma\gamma}\,(\bar\lambda+\bar\gamma(y_{\bar\gamma}-\psi))^2+\frac{\bar\lambda}{\gamma}\,[\bar\gamma(y_{\bar\gamma}-\psi)-\gamma(y_{\bar\gamma}-\psi)] \\ &\qquad +\frac{1}{2}\,\left[\frac{\bar\gamma^2}{\gamma}(y_{\bar\gamma}-\psi)^2-\bar\gamma(y_{\bar\gamma}-\psi)^2+(\bar\gamma-\gamma)(y_{\bar\gamma}-\psi)(y_\gamma-\psi)\right], \end{aligned}$$

and thus on $P_{\bar\gamma}\cap P_{\bar\gamma}$

$$\begin{aligned} (\bar\gamma-\gamma)^{-1}z &= \frac{-1}{2\bar\gamma\gamma}(\bar\lambda+\bar\gamma(y_\gamma-\psi))^2+\frac{\bar\lambda}{\gamma}(y_{\bar\gamma}-\psi) \\ &\qquad +\frac{1}{2}\left[\frac{\bar\gamma}{\gamma}(y_{\bar\gamma}-\psi)^2+(y_{\bar\gamma}-\psi)(y_\gamma-\psi)\right]. \end{aligned}$$

By Lebesgue's bounded convergence theorem,

$$\begin{aligned} &\lim_{\bar\gamma\to\gamma}\frac{1}{\bar\gamma-\gamma}\;\mathcal{I}_1 = \lim_{\bar\gamma\to\gamma}\frac{1}{\bar\gamma-\gamma}\int_\Omega z\,\chi_{P_{\bar\gamma}\cap P_\gamma} \\ &= -\frac{1}{2\gamma^2}\int_\Omega((\bar\lambda+\gamma(y_\gamma-\psi))^+)^2+\frac{1}{\gamma}\int_\Omega(\bar\lambda+\gamma(y_\gamma-\psi))^+(y_\gamma-\psi). \end{aligned}$$

Together with (24) and (25), this implies the claim. ∎

Remark 4.1. Note that $\dot V$ is characterized without recourse to $\dot y_\gamma$.

The boundedness of $\{\gamma^2\dot V(\gamma)\}_{\gamma\ge0}$ is established next. In what follows we use $(v)^- = -\min(0,v)$.

PROPOSITION 4.4. *If $\bar\lambda = 0$ and $a(v^+,v^-) = 0$ for all $v\in H_0^1(\Omega)$, then $\{\gamma^2\dot V(\gamma)\}_{\gamma\ge0}$ is bounded. If* (11) *and* (12) *hold, then again $\{\gamma^2\dot V(\gamma)\}_{\gamma\ge0}$ is bounded.*

Proof. In the case $\bar\lambda = 0$ we have

$$a(y_\gamma-\psi,v)+\gamma((y_\gamma-\psi)^+,v) = (f,v)-a(\psi,v) \quad \text{for all } v\in H_0^1(\Omega).$$

Since $(y_\gamma-\psi)\in H_0^1(\Omega)$ and $a((y_\gamma-\psi)^+,(y_\gamma-\psi)^-) = 0$ we have, using (6) with $v = (y_\gamma-\psi)^+$,

$$\nu|(y_\gamma-\psi)^+|^2_{H_0^1(\Omega)}+\gamma|(y_\gamma-\psi)^+|^2_{L^2}\le|f|_{L^2}|(y_\gamma-\psi)^+|_{H_0^1}+\mu|\psi|_{H^1}|y_\gamma-\psi|_{H^1}.$$

This implies the existence of a constant K, depending on $|\psi|_{H^1}$ and $|f|_{L^2}$ but independent of $\gamma \geq 0$, such that $\gamma|(y_\gamma - \psi)^+|_{L^2} \leq K$. Since $\dot{V}(\gamma) = \frac{1}{2}\int_\Omega |(y_\gamma - \psi)^+|^2$ the claim follows.

Turning to the feasible case with (11) and (12) holding, we have that $y_\gamma \leq \psi$ for every $\gamma > 0$, and hence $(\bar{\lambda} + \gamma(y_\gamma - \psi))(x) > 0$ if and only if $\bar{\lambda}(x) > \gamma(\psi - y_\gamma)(x)$. Consequently,

$$\begin{aligned}|\dot{V}(\gamma)| &\leq \frac{1}{2\gamma^2}\int_\Omega |(\bar{\lambda} + \gamma(y_\gamma - \psi))^+|^2 + \frac{1}{\gamma}\int_\Omega (\bar{\lambda} + \gamma(y_\gamma - \psi))^+(\psi - y_\gamma) \\ &\leq \frac{3}{2\gamma^2}|\bar{\lambda}|^2_{L^2},\end{aligned}$$

which again implies the claim. ∎

Before we investigate $\ddot{V}$, we state a result which connects $\gamma\dot{V}(\gamma)$, $|y^* - y_\gamma|_{H^1_0}$, and $V^* - V(\gamma)$, where $V^* = \lim_{\gamma\to\infty} V(\gamma)$. It will be used in section 6.1 for designing a γ-update strategy.

PROPOSITION 4.5. *In the feasible and infeasible cases the following estimate holds true:*

$$|y^* - y_\gamma|^2_{H^1_0} \leq \frac{2}{\nu}\left(V^* - V(\gamma) - \gamma\dot{V}(\gamma)\right).$$

Proof. We have $V^* - V(\gamma) = J(y^*) - J(y_\gamma;\gamma)$ and

$$\begin{aligned}J(y^*) - J(y_\gamma;\gamma) \geq\ & \frac{\nu}{2}|y^* - y_\gamma|^2_{H^1_0} + a(y_\gamma, y^* - y_\gamma) - (f, y^* - y_\gamma) \\ & - \frac{1}{2\gamma}|(\bar{\lambda} + \gamma(y_\gamma - \psi))^+|^2_{L^2},\end{aligned}$$

where we have used (6). From (OC_γ) we have

$$a(y_\gamma, y^* - y_\gamma) - (f, y^* - y_\gamma) = -((\bar{\lambda} + \gamma(y_\gamma - \psi))^+, y^* - y_\gamma),$$

and hence

$$\begin{aligned}J(y^*) - J(y_\gamma;\gamma) \geq\ & \frac{\nu}{2}|y^* - y_\gamma|^2_{H^1_0} - ((\bar{\lambda} + \gamma(y_\gamma - \psi))^+, y^* - y_\gamma) \\ & - \frac{1}{2\gamma}|(\bar{\lambda} + \gamma(y_\gamma - \psi))^+|^2_{L^2} \\ \geq\ & \frac{\nu}{2}|y^* - y_\gamma|^2_{H^1_0} - \frac{1}{2\gamma}|(\bar{\lambda} + \gamma(y_\gamma - \psi))^+|^2_{L^2} \\ & + ((\bar{\lambda} + \gamma(y_\gamma - \psi))^+, y_\gamma - \psi) \\ =\ & \frac{\nu}{2}|y^* - y_\gamma|^2_{H^1_0} + \gamma\dot{V}(\gamma).\end{aligned}$$

This completes the proof. ∎

Below we shall assume that $y_\gamma - \psi \in C(\bar{\Omega})$. Recall that for dimension $n \leq 3$ and with (6) and (8) holding, we have $y_\gamma \in H^2(\Omega) \subset C(\bar{\Omega})$.

PROPOSITION 4.6. *Let $\dot{y}_\gamma$ denote any accumulation point of $(\bar{\gamma} - \gamma)^{-1}(y_{\bar{\gamma}} - y_\gamma)$ as $\bar{\gamma} \to \gamma$.*

(a) *If $\bar{\lambda} = 0$, $y_\gamma - \psi \in C(\bar{\Omega})$, and (8) is satisfied, then $\gamma \mapsto V(\gamma)$ is twice differentiable at γ with*

$$\ddot{V}(\gamma) = \int_\Omega (y_\gamma - \psi)^+\dot{y}_\gamma. \tag{26}$$

(b) *For arbitrary* $\bar\lambda$, *if* $\operatorname{meas}(S_\gamma^\circ)=0$, *then* $\gamma\mapsto V(\gamma)$ *is twice differentiable at* γ *with*

$$\begin{aligned}\ddot V(\gamma)&=\frac{1}{\gamma^3}\int_\Omega|(\bar\lambda+\gamma(y_\gamma-\psi))^+|^2\\&\quad-\frac{2}{\gamma^2}\int_\Omega(\bar\lambda+\gamma(y_\gamma-\psi))^+(y_\gamma-\psi)\\&\quad+\frac{1}{\gamma}\int_\Omega(y_\gamma-\psi)(y_\gamma-\psi+\gamma\,\dot y_\gamma)\chi_{S_\gamma}.\end{aligned}\tag{27}$$

Proof. (a) On the subsequence γ_n realizing the accumulation point, we have that $\lim_{n\to\infty}(\gamma_n-\gamma)^{-1}(\dot V(\gamma_n)-\dot V(\gamma))$ equals the right-hand side of (26). The claim will be established by verifying that the accumulation points $\dot y_\gamma$ restricted to $S_\gamma=\{x\colon y_\gamma(x)-\psi(x)>0\}$ are unique. Let z denote the difference of two accumulation points. By (16) and (19) we have

$$a(z,v)+\gamma(z,v)=0\ \text{ for all }\ v\in H_0^1(\Omega)\ \text{ with }\ v=0\ \text{ on }\ \Omega\setminus S_\gamma.$$

Using (8) and the fact that S_γ is an open set relative to Ω due to the continuity of $y_\gamma-\psi$, we find that $z=0$ in S_γ, as desired.

(b) Let $\dot y_\gamma$ denote any accumulation point of $(\bar\gamma-\gamma)^{-1}(y_{\bar\gamma}-y_\gamma)$ as $\bar\gamma\downarrow\gamma$, and recall the notation $g(\gamma)=\bar\lambda+\gamma(y_\gamma-\psi)$ and S_γ^+ from section 3. On the subsequence realizing the accumulation point we find

$$\begin{aligned}\lim_{\bar\gamma\to\gamma}\frac{1}{\bar\gamma-\gamma}(\dot V(\bar\gamma)-\dot V(\gamma))&=\frac{1}{\gamma^3}\int_\Omega|(\bar\lambda+\gamma(y_\gamma-\psi))^+|^2\\&\quad-\frac{2}{\gamma^2}\int_\Omega(\bar\lambda+\gamma(y_\gamma-\psi))^+(y_\gamma-\psi)\\&\quad+\frac{1}{\gamma}\int_\Omega(y_\gamma-\psi)(y_\gamma-\psi+\gamma\,\dot y_\gamma)\chi_{S_\gamma^+}.\end{aligned}\tag{28}$$

By assumption, $\operatorname{meas}(S_\gamma^\circ)=0$ and, hence the right-hand sides of (27) and (28) coincide. Since $\dot y_\gamma$ is unique by Corollary 3.8 the claim is established. ∎

5. Model functions. In this section we derive low-parameter families of functions which approximate the value functional V and share some of its qualitative properties. We will make use of these models in the numerics section when devising path-following algorithms.

5.1. Infeasible case. Throughout this subsection we assume (8) and

$$\bar\lambda=0,\ y_\gamma-\psi\in C(\bar\Omega)\quad\text{ for all }\ \gamma\in(0,\infty).\tag{29}$$

Observe that (8), together with the general assumption (6), implies (11). In fact, for any $v\in H_0^1(\Omega)$ we have $a(v,v^+)\ge\gamma|v^+|^2$, and hence $0\ge a(v,v^+)$ implies $v^+=0$.

PROPOSITION 5.1. *The value function* V *satisfies* $\dot V(\gamma)\ge0$ *and* $\ddot V(\gamma)\le0$ *for* $\gamma\in(0,\infty)$.

Proof. Proposition 4.2 implies that $\dot V(\gamma)\ge0$. Moreover, $y_{\gamma_2}\le y_{\gamma_1}$ for $\gamma_2\ge\gamma_1>0$ and hence $\dot y_\gamma\le0$ a.e. on S_γ. Consequently $\ddot V(\gamma)\le0$ by Proposition 4.6. ∎

A model function m for the value function V should reflect the sign properties of V and its derivatives. Moreover, $V(0)$ gives the value of $(\hat{\mathrm P})$, and hence we shall require that $m(0)=V(0)$. Finally from Lemma 3.2 we conclude that V is bounded on $[0,\infty)$. All these properties are satisfied by functions of the form

$$m(\gamma)=C_1-\frac{C_2}{E+\gamma}\tag{30}$$

with $C_1 \in \mathbb{R}$. Here $C_2 \geq 0$, $E > 0$ satisfy

$$m(0) = V(0) = C_1 - \frac{C_2}{E}. \tag{31}$$

Other choices for model functions are also conceivable, for example, $\gamma \to C_1 - \frac{C_1}{(E+\gamma)^r}$ with $r > 1$. Note, however, that the asymptotic behavior of the model in (30) is such that $\gamma^2 \dot{m}(\gamma)$ is bounded for $\gamma \to \infty$. This is consistent with the boundedness of $\gamma^2 \dot{V}(\gamma)$ for $\gamma \to \infty$ asserted in Proposition 4.4.

Another reason for choosing (30) is illustrated next. Choosing $v = (y_\gamma - \psi)^+$ in (OC_γ), we find

$$a(\dot{y}_\gamma, (y_\gamma - \psi)^+) + |(y_\gamma - \psi)^+|^2_{L^2} + \gamma \int_\Omega (y_\gamma - \psi)^+ \dot{y}_\gamma = 0. \tag{32}$$

For the following discussion we

$$\text{replace } a(\cdot,\cdot) \text{ by } E(\cdot,\cdot) \text{ with } E > 0 \text{ a constant, and } V \text{ by } m. \tag{33}$$

By Proposition 4.2 and (26) the following ordinary differential equation is obtained for m:

$$(E + \gamma)\, \ddot{m}(\gamma) + 2\, \dot{m}(\gamma) = 0. \tag{34}$$

The solutions to (34) are given by (30). To get an account for the quality of our model in (30) we refer to the left-hand plot of Figure 4 in section 6.

5.2. Feasible case. Throughout this subsection we assume

$$(11), \quad \bar{\lambda} \text{ satisfies (12)}, \quad \text{and meas}\,(S^\circ_\gamma) = 0 \ \text{ for all } \ \gamma \in (0, \infty). \tag{35}$$

PROPOSITION 5.2. *The value function V satisfies $\dot{V}(\gamma) \leq 0$ and $\ddot{V}(\gamma) \geq 0$ for $\gamma \in (0, \infty)$.*

Proof. By Proposition 2.2 we have $y_\gamma \leq \psi$ and hence $\dot{V}(\gamma) \leq 0$ by Proposition 4.2. A short computation based on (27) shows that

$$\ddot{V}(\gamma) = \frac{1}{\gamma^3} \int_\Omega \chi \bar{\lambda}^2 + \int_\Omega \chi (y_\gamma - \psi) \dot{y}_\gamma \geq \frac{1}{\gamma} \int_\Omega \chi (y_\gamma - \psi)^2 + \int_\Omega \chi\, (y_\gamma - \psi) \dot{y}_\gamma, \tag{36}$$

where χ is the characteristic function of the set $S_\gamma = \{\bar{\lambda} + \gamma(y_\gamma - \psi) > 0\}$. From (22) we have

$$\gamma |\dot{y}_\gamma|_{L^2(S_\gamma)} \leq |\psi - y_\gamma|_{L^2(S_\gamma)},$$

and hence $\ddot{V}(\gamma) \geq 0$. ∎

An immediate consequence is stated next.

LEMMA 5.3. *If the solution to the unconstrained problem is not feasible, then* $\lim_{\gamma \downarrow 0} V(\gamma) = \infty$.

Proof. Assume that $\lim_{\gamma \downarrow 0} V(\gamma)$ is finite. Then, using (P_γ), there exists a sequence $\gamma_n \to 0$ and $\tilde{y} \in H^1_0(\Omega)$ such that $y_{\gamma_n} \rightharpoonup \tilde{y}$ weakly in $H^1_0(\Omega)$, with y_{γ_n} the solution to (P_{γ_n}), and $\lambda_{\gamma_n} = \max(0, \bar{\lambda} + \gamma_n(y_n - \psi)) \to 0$ in $L^2(\Omega)$. Consequently $\tilde{y} \leq \psi$. Taking the limit with respect to n in (OC_{γ_n}), it follows that $\tilde{y} \leq \psi$ is the solution to ($\hat{\mathrm{P}}$), which contradicts our assumption. ∎

From Lemmas 3.2 and 5.3 and Proposition 5.2 it follows that $\gamma \mapsto V(\gamma)$, $\gamma \in (0,\infty)$, is a monotonically strictly decreasing convex function with $\lim_{\gamma\to 0^+} V(\gamma) = \infty$. All these properties are also satisfied by functions of the form

$$m(\gamma) = C_1 \; - \; \frac{C_2}{E+\gamma} \; + \; \frac{B}{\gamma}, \tag{37}$$

provided that $C_1 \in \mathbb{R}$, $C_2 \geq 0$, $E > 0$, $B > 0$, and $C_2 \leq B$.

We now give the motivation for choosing the model function m for V as in (37). From (22) with $v = (y_\gamma - \psi)\chi$ we get

$$a(\dot{y}_\gamma, (y-\psi)\chi) + \gamma(\dot{y}_\gamma\chi, y_\gamma - \psi) + ((y_\gamma - \psi)\chi, y_\gamma - \psi) = 0,$$

where $\chi = \chi_{S_\gamma}$. As in the infeasible case we replace $a(\cdot,\cdot)$ by $E(\cdot,\cdot)$, with E a constant, and using (22), we arrive at

$$(E+\gamma)(\dot{y}_\gamma\chi, v) + ((y_\gamma - \psi)\chi, v) = 0.$$

The choice $v = y_\gamma - \psi$ implies

$$(E+\gamma)(\dot{y}_\gamma\chi, y_\gamma - \psi) + ((y_\gamma - \psi)\chi, y_\gamma - \psi) = 0. \tag{38}$$

Note that $\dot{V}(\gamma)$ can be expressed as

$$\dot{V}(\gamma) = -\frac{1}{2\gamma^2}\int_\Omega \bar{\lambda}^2\chi + \frac{1}{2}\int_\Omega (y_\gamma - \psi)^2\chi. \tag{39}$$

Using (36) and (39) in (38), and replacing V by m, due to the substitution for $a(\cdot,\cdot)$, we find

$$(E+\gamma)\ddot{m} + 2\dot{m} - E\,\gamma^{-3}\int_\Omega \chi\,\bar{\lambda}^2 = 0.$$

We further replace $\int_\Omega \chi\bar{\lambda}^2$, which is a bounded quantity depending on γ, by $2B$, and obtain, as the ordinary differential equation that we propose for the model function m in the feasible case,

$$(E+\gamma)\ddot{m} + 2\dot{m} - 2\gamma^{-3}E\,B = 0. \tag{40}$$

The family of solutions is given by (37). In the right-hand plot of Figure 4 in section 6 we depict the approximation quality of $m(\gamma)$.

6. Path-following algorithms. In this section we study the basic Algorithm B together with a variety of adjustment schemes for the path parameter γ. For this purpose recall that, depending on the shift parameter $\bar{\lambda}$, the elements y_γ along the primal-dual path are feasible or infeasible. As we have seen in the previous section, this implies different models for approximating the value function V. We will see, however, that for $\gamma > 0$ in both cases similar strategies for updating γ may be used. When referring to the infeasible or feasible case, (29), respectively (35), is assumed to hold.

The subsequent discussion is based on the following two-dimensional test problems. We point out that the bound ψ in problem P1 below does not satisfy $\psi \in H^1(\Omega)$. However, as we shall see, the feasible and infeasible primal-dual path as well as the algorithms introduced subsequently still perform satisfactorily. We include this example since discontinuous obstacles are of practical relevance.

Test problem P1. We consider (8) with $a_{ij} = \delta_{ij}$, with δ_{ij} the Kronecker symbol, $d = 0$, and $\Omega = (0,1)^2$. We choose

$$f(x_1, x_2) = 500x_1 \sin(5x_1)\cos(x_2)$$

and $\psi \equiv 10$ on $\Omega \setminus K$, and $\psi \equiv 1$ on K with $K = \{x \in \Omega : \frac{1}{5} \le \|x - (\frac{1}{2}, \frac{1}{2})^\top\|_2 \le \frac{2}{5}\}$. The solution y^*, the obstacle ψ, and the active set $\mathcal{A}^*$ at the solution are shown in Figure 1.

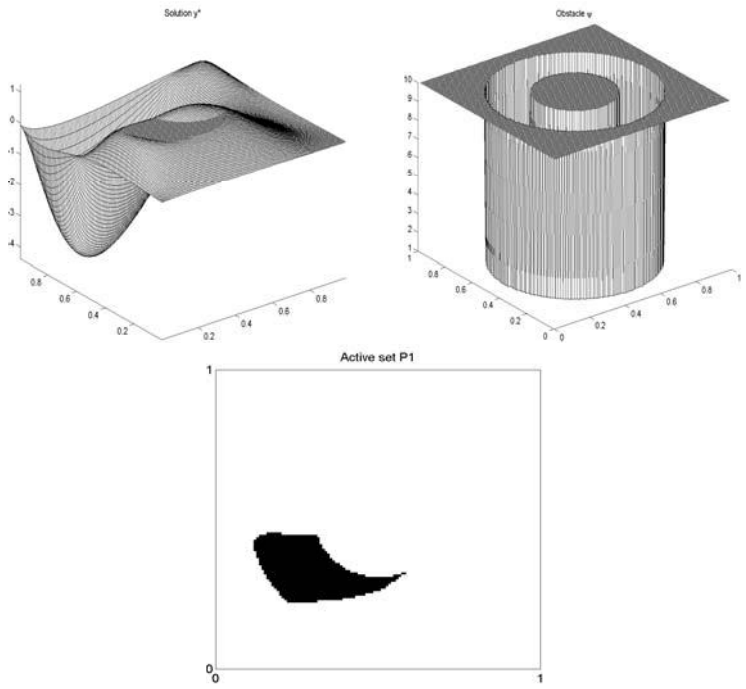


FIG. 1. *Optimal solution y^* (upper left plot), obstacle ψ (upper right plot), and the active set $\mathcal{A}^*$ (lower plot) for test problem* P1.

Test problem P2. Again we consider (8), with a_{ij}, d, and Ω as before, and define

$$y^\dagger := \begin{cases} x_1 & \text{on } T_1 := \{x \in \Omega : x_2 \le x_1 \wedge x_2 \le 1 - x_1\}, \\ 1 - x_2 & \text{on } T_2 := \{x \in \Omega : x_2 \le x_1 \wedge x_2 \ge 1 - x_1\}, \\ 1 - x_1 & \text{on } T_3 := \{x \in \Omega : x_2 \ge x_1 \wedge x_2 \ge 1 - x_1\}, \\ x_2 & \text{on } T_4 := \{x \in \Omega : x_2 \ge x_1 \wedge x_2 \le 1 - x_1\}. \end{cases} \tag{41}$$

The obstacle ψ is defined by $\psi \equiv y^\dagger$ on $S_1 := \{x \in \Omega : \|x - (\frac{1}{2}, \frac{1}{2})^\top\|_\infty \le \frac{1}{4}\}$, $\psi \equiv \frac{1}{4}$ on $S_2 \setminus S_1$, and

$$\psi := \begin{cases} 2x_1 & \text{on } T_1 \cap (\Omega \setminus S_2), \\ \frac{1}{4} - 2(x_2 - \frac{7}{8}) & \text{on } T_2 \cap (\Omega \setminus S_2), \\ \frac{1}{4} - 2(x_1 - \frac{7}{8}) & \text{on } T_3 \cap (\Omega \setminus S_2), \\ 2x_2 & \text{on } T_4 \cap (\Omega \setminus S_2), \end{cases}$$

with $S_2 := \{x \in \Omega : \|x - (\frac{1}{2}, \frac{1}{2})^\top\|_\infty \le \frac{3}{8}\}$. The forcing term is given by

$$(f, \phi)_{L^2} = \int_{\Omega^+} \phi(s)ds + (\chi_{S_1}, \phi)_{L^2} + \int_{S_1 \cap \Omega^+} \phi(s)ds \qquad \text{for all } \phi \in H_0^1(\Omega),$$

where $\Omega^+ := \{x \in \Omega : x_2 = x_1\} \cup \{x \in \Omega : x_2 = 1 - x_1\}$. We recall that for $\phi \in H_0^1(\Omega)$, $\Omega \subset \mathbb{R}^2$, the traces along smooth curves are well defined. The solution y^* is given by $y^* = y^\dagger$. The active or coincidence set at the solution is $\mathcal{A}^* = S_1$. The Lagrange multiplier $\lambda^* = f + \Delta y^*$ is in $H^{-1}(\Omega)$ and enjoys no extra regularity. In Figure 2 we display the optimal solution y^*, the obstacle ψ, and the active set $\mathcal{A}^*$.

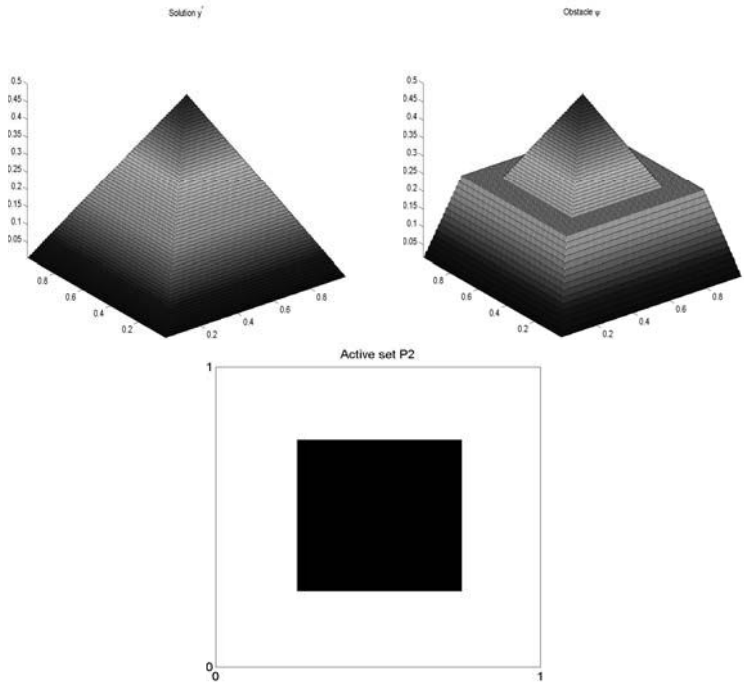


FIG. 2. *Optimal solution y^* (upper left plot), obstacle ψ (upper right plot), and the active set $\mathcal{A}^*$ (lower plot) for test problem* P2.

Test problem P3. For this test problem (8) is satisfied. We therefore obtain $y^* \in H^2(\Omega)$ and $\lambda^* \in L^2(\Omega)$. The coefficients a_{ij} and d as well as Ω are as before. The volume force f is given by $f = -\Delta v$ with $v(x_1, x_2) = \sin(3\pi x_1)\sin(3\pi x_2)$. Further, we have $\psi = \frac{1}{4} - \frac{1}{10}\sin(\pi x_1)\sin(\pi x_2)$. The optimal solution y^*, the Lagrange multiplier λ^*, and the active set at y^* are displayed in Figure 3.

Unless specified otherwise, the subsequent algorithms are initialized by $y_0 = (-\Delta)^{-1} f$, where $-\Delta$ denotes the Laplacian with homogeneous Dirichlet boundary conditions. The initial Lagrange multiplier is chosen as $\lambda_0 = \gamma_0 \chi_{\{y_0 > \psi\}}(y_0 - \psi)$.

The discretization of $-\Delta$ is based on the classical five-point finite difference stencil. We denote the mesh size by h, which we occasionally drop for convenience. The forcing term f in P2 is discretized by $f = -\Delta y^\dagger + \chi_{S_1} e + \chi_{S_1}(-\Delta y^\dagger)$, where e is the vector of all ones and χ_{S_1} represents a diagonal matrix with entry $(\chi_{S_1})_{ii} = 1$ for grid points $x_i \in S_1$ and $(\chi_{S_1})_{ii} = 0$ otherwise. Above $y^\dagger$ denotes the grid function corresponding to (41).

6.1. A strategy based on model functions—exact path-following. As outlined in section 5, there are good reasons to trust our model functions (30) and (37) in the infeasible and feasible cases, respectively. Let us start by focusing on the infeasible case. The model is given by $m(\gamma) = C_1 - C_2(E + \gamma)^{-1}$. For determining

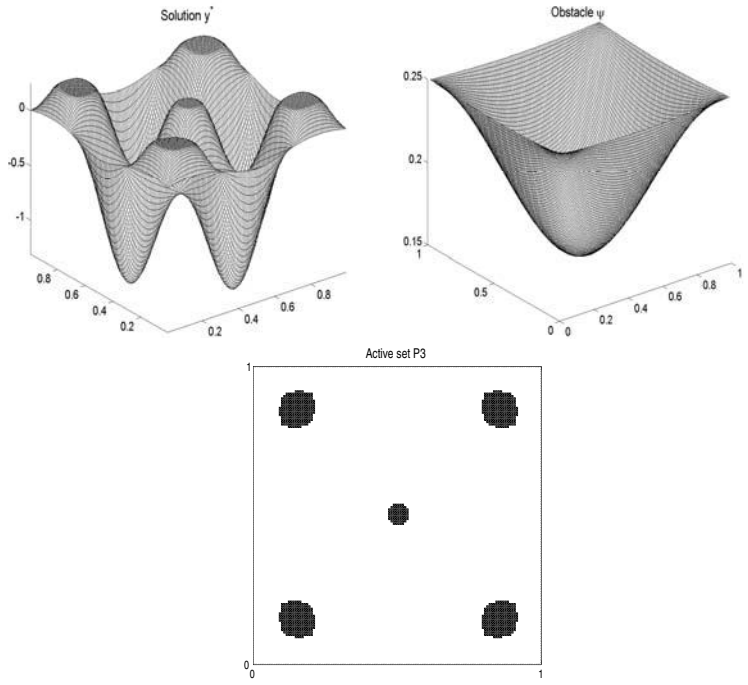


FIG. 3. *Optimal solution y^* (upper left plot), obstacle ψ (upper right plot), and the active set $\mathcal{A}^*$ (lower plot) for test problem* P3.

the three parameters C_1, C_2, and E, we use the information $V(0), V(\gamma), \dot{V}(\gamma)$, which, by Proposition 4.2, is available from one solve of the unconstrained problem ($\hat{\text{P}}$) and one solve for (Pγ). The conditions

$$m(0) = V(0), \quad m(\gamma) = V(\gamma), \quad \dot{m}(\gamma) = \dot{V}(\gamma) \tag{42}$$

yield

$$\begin{aligned} E &= \gamma^2 \dot{V}(\gamma) \left(V(\gamma) - V(0) - \gamma \dot{V}(\gamma) \right)^{-1}, \\ C_2 &= \gamma^{-1} E(E+\gamma) \left(V(\gamma) - V(0) \right), \\ C_1 &= V(0) + C_2 E^{-1}. \end{aligned} \tag{43}$$

We could have used an alternative reference value $\gamma_r \in (0, \gamma)$ and computed $m(\gamma_r) = V(\gamma_r)$ instead of $m(0) = V(0)$. In Figure 4 we compare $V(\gamma)$ to $m(\gamma)$ for different values of the coefficients (C_1, C_2, E). These coefficients depend on different values y_f for γ (in (42)) produced by Algorithm EP (see below) for problem P1. The solid line corresponds to $V(\gamma)$. The corresponding γ-values γ_f for (42) are depicted in the legend of the left plot in Figure 4. The dotted and dashed line belong to rather small γ-values, and the dashed-dotted and the circled lines to large γ_f in (42). As we can see, the dotted line is accurate in the range of relatively small γ_f, while the other lines are more accurate for large γ_f. From now on we consider only the choices $\gamma_r = 0$ and $\gamma = \gamma_k$ in (42) when updating γ_k.

Next we discuss properties of the model parameters E, C_1, C_2 according to (43). For this purpose assume that the solution $\hat{y}$ to ($\hat{\text{P}}$) is not feasible for (P). Then

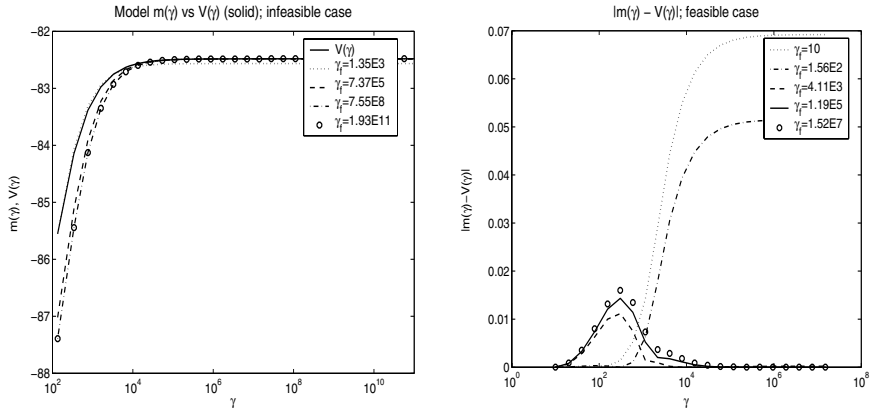


FIG. 4. *Left: Model $m(\gamma)$ vs. $V(\gamma)$ (solid) in the infeasible case for* P1. *Right: Model $m(\gamma)$ vs. $V(\gamma)$ in the feasible case.*

by Corollary 4.3 we have $\dot{V}(\gamma) > 0$ for all $\gamma > 0$. Consequently $V(\gamma) > V(0)$ and $V(\gamma) - V(0) - \gamma\dot{V}(\gamma) = -\int_0^\gamma \int_s^\gamma \ddot{V}(\sigma) d\sigma ds > 0$, and hence $E > 0$ and $C_2 > 0$ for all $\gamma \in (0, +\infty)$. This implies $m(\gamma) \leq C_1$ and $m(\gamma) \to C_1$ for $\gamma \to +\infty$.

We propose the following update strategy for γ: Let $\{\tau_k\}$ satisfy $\tau_k \in (0,1)$ for all $k \in \mathbb{N}$ and $\tau_k \downarrow 0$ as $k \to \infty$, and assume that $V(\gamma_k)$ is available. Then, given γ_k, the updated value γ_{k+1} should ideally satisfy

$$|V^* - V(\gamma_{k+1})| \leq \tau_k |V^* - V(\gamma_k)|. \tag{44}$$

Since V^* and $V(\gamma_{k+1})$ are unknown, we use $C_{1,k}$ and our model $m_k(\gamma) = C_{1,k} - C_{2,k}/(E_k + \gamma)$ at $\gamma = \gamma_{k+1}$ instead. Thus, (44) is replaced by

$$|C_{1,k} - m_k(\gamma_{k+1})| \leq \tau_k |C_{1,k} - V(\gamma_k)| =: \beta_k. \tag{45}$$

Solving the equation $C_{1,k} - m_k(\gamma_{k+1}) = \beta_k$, we obtain

$$\gamma_{k+1} = \frac{C_{2,k}}{\beta_k} - E_k. \tag{46}$$

In Theorem 6.1 we shall show that $\gamma_{k+1} \geq \kappa\gamma_k$, with $\kappa > 1$, independently of $k \in \mathbb{N}$.

Before we turn to the feasible case, we interpret (44) in view of Proposition 4.5 in the infeasible case. Recall that $V^* \geq V(\gamma)$, and observe that $|V^* - V(\gamma)| = \mathcal{O}(|y^* - y_\gamma|_{H_0^1})$. Proposition 4.5 yields

$$|y^* - y_\gamma|_{H_0^1}^2 \leq \frac{2}{\nu} (V^* - V(\gamma))$$

since $\dot{V}(\gamma) > 0$. Setting $\tau_k = \omega_k^2 |V^* - V(\gamma_k)|$, with $\omega_k \to 0$, in (44) yields

$$|y^* - y_{\gamma_{k+1}}|_{H_0^1}^2 \leq C_\tau \omega_k^2 |y^* - y_{\gamma_k}|_{H_0^1}^2.$$

Consequently, we obtain

$$\frac{|y^* - y_{\gamma_{k+1}}|_{H_0^1}}{|y^* - y_{\gamma_k}|_{H_0^1}} \leq C_\tau \omega_k,$$

which implies q-superlinear convergence of $\{y_{\gamma_k}\}$ in $H_0^1(\Omega)$.

In the feasible case, i.e., when $\bar{\lambda}$ satisfies (12), we use the model $m(\gamma) = C_1 - C_2(E+\gamma)^{-1} + B\gamma^{-1}$ with $C_2 \geq 0$ and $E, B > 0$; see (37). Let $\gamma_r > 0$, $\gamma_r \neq \gamma$, denote a reference γ-value; then we use the conditions

$$m(\gamma_r) = V(\gamma_r), \quad \dot{m}(\gamma_r) = \dot{V}(\gamma_r), \quad m(\gamma) = V(\gamma), \quad \dot{m}(\gamma) = \dot{V}(\gamma)$$

for fixing B, $C1$, $C2$, E. Solving the corresponding system of nonlinear equations, we get

$$E = \frac{\Big(\Big((\gamma_r - \gamma)(\dot{V}(\gamma_r)\gamma_r^2 + \dot{V}(\gamma)\gamma^2) + 2\gamma_r\gamma(V(\gamma) - V(\gamma_r)\Big)\Big)}{\Big((\dot{V}(\gamma)\gamma + \dot{V}(\gamma_r)\gamma_r)(\gamma - \gamma_r) + (\gamma_r + \gamma)(V(\gamma_r) - V(\gamma))\Big)}$$

and

$$B = \frac{\gamma_r^2\gamma^2\left((V(\gamma) - V(\gamma_r))^2 - \dot{V}(\gamma)\dot{V}(\gamma_r)(\gamma - \gamma_r)^2\right)}{\Big((\gamma - \gamma_r)^2(\dot{V}(\gamma_r)\gamma_r^2 + \dot{V}(\gamma)\gamma^2) + 2(\gamma - \gamma_r)\gamma_r\gamma(V(\gamma_r) - V(\gamma))\Big)}$$

Then the parameters C_1 and C_2 are given by

$$C_2 = (E+\gamma)^2\left(\frac{B}{\gamma^2} + \dot{V}(\gamma)\right),$$

$$C_1 = V(\gamma) + \frac{C_2}{E+\gamma} - \frac{B}{\gamma}.$$

In the right plot of Figure 4 we show $|m(\gamma) - V(\gamma)|$ with $m(\gamma)$ produced by the iterates of Algorithm EP for P1 similar to the infeasible case. Again we can see that our model yields a close approximation of the value function V.

If we require that (45) be satisfied in the feasible case, then we obtain the following update strategy for γ:

$$\gamma_{k+1} = -\frac{D_k}{2} + \sqrt{\frac{D_k^2}{4} + \frac{B_k E_k}{\beta_k}}, \tag{47}$$

where $D_k = E_k + (C_{2,k} - B_k)/\beta_k$. In Theorem 6.1 we shall establish $\gamma_{k+1} \geq \kappa\gamma_k$ for all $k \in \mathbb{N}_0$ with $\kappa > 1$ independent of k.

Next we describe an **e**xact **p**ath-following version of Algorithm B, which utilizes the update strategy (45) for updating γ.

ALGORITHM EP.

(i) Select γ_r. Compute $V(\gamma_r)$, and choose $\gamma_0 > \max(1, \gamma_r)$; set $k = 0$.
(ii) Apply Algorithm B to obtain y_{γ_k}.
(iii) Compute $V(\gamma_k)$, $\dot{V}(\gamma_k)$, and γ_{k+1} according to (46) in the infeasible case or (47) in the feasible case.
(iv) Set $k = k + 1$, and go to (ii).

Concerning the choice of γ_r note that in the infeasible case we have $\gamma_r \geq 0$, and in the feasible case $\gamma_r > 0$. Convergence of Algorithm EP is addressed next.

THEOREM 6.1. *Assume that the solution to* ($\hat{\text{P}}$) *is not feasible for* (P). *Then the iterates* γ_k *of Algorithm EP tend to* ∞ *as* $k \to \infty$, *and consequently* $\lim_{k\to\infty}(y_{\gamma_k}, \lambda_{\gamma_k}) = (y^*, \lambda^*)$ *in* $H_0^1(\Omega) \times H^{-1}(\Omega)_{weak}$.

Proof. Let us consider the infeasible case. Then (45) is equivalent to

$$0 < C_{1,k} - m_k(\gamma_{k+1}) < \tau_k(C_{1,k} - m_k(\gamma_k)). \tag{48}$$

Since $\gamma \mapsto m_k(\gamma)$ is strictly increasing and $\tau_k \in (0,1)$, it follows that $\gamma_{k+1} > \gamma_k$ for every $k = 0, 1, \ldots$. If $\lim_{k\to\infty} \gamma_k = \infty$, then $\lim_{k\to\infty}(y_{\gamma_k}, \lambda_{\gamma_k}) = (y^*, \lambda^*)$. Otherwise there exists $\bar\gamma$ such that $\lim_{k\to\infty} \gamma_k = \bar\gamma$. Since $\gamma \mapsto V(\gamma)$ and $\gamma \mapsto \dot V(\gamma)$ are continuous on $(0,\infty)$, it follows from (42) and (43) that $\lim_{k\to\infty} E_k = E(\bar\gamma)$, $\lim_{k\to\infty} C_{1,k} = C_1(\bar\gamma)$, and $\lim_{k\to\infty} C_{2,k} = C_2(\bar\gamma)$, where $E(\bar\gamma)$, $C_1(\bar\gamma)$, $C_2(\bar\gamma)$ are given by (43) with γ replaced by $\bar\gamma$. Taking the limit with respect to k in (48), we arrive at

$$\frac{C_2(\bar\gamma)}{E(\bar\gamma) + \bar\gamma} = 0,$$

which is impossible, since $C_2(\bar\gamma) > 0$ and $E(\bar\gamma) > 0$ if the solution to $(\hat{\mathrm{P}})$ is not feasible for (P). Thus $\lim_{k\to\infty} \gamma_k = \infty$. The feasible case is treated analogously. ∎

Numerically we stop the algorithm as soon as $\|(r_k^{1,h}, r_k^{2,h}, r_k^{3,h})^\top\|_2 \le \sqrt{\epsilon_M}$, where

$$\begin{aligned}
r_k^{1,h} &= \|y_{\gamma_k}^h + (-\Delta^h)^{-1}(\lambda_{\gamma_k}^h - f^h)\|_{H^{-1,h}} / \|f^h\|_{H^{-1,h}},\\
r_k^{2,h} &= \|\lambda_{\gamma_k}^h - \max(0, \lambda_{\gamma_k}^h + y_{\gamma_k}^h - \psi^h)\|_{H^{-1,h}},\\
r_k^{3,h} &= \|\max(0, y_{\gamma_k}^h - \psi^h)\|_{L_2^h},
\end{aligned}$$

and ϵ_M denotes the machine accuracy. Here $|\cdot|_{H^{-1,h}}$ denotes the discrete version of $|\cdot|_{H^{-1}}$. For some vector v it is realized as $|v|_{H^{-1}} = |\nabla^h(-\Delta^h)^{-1}v|_{L_2^h}$ with $|\cdot|_{L_2^h}$ the discrete L^2-norm and ∇^h a forward difference approximation of the gradient operator; see [8]. The inner iteration, i.e., Algorithm B for $\gamma = \gamma^k$, is terminated if successive active sets coincide or

$$\frac{\|-\Delta^h y_{\gamma_k}^{h,l} + \lambda_{\gamma_k}^{h,l} - f^h\|_{H^{-1,h}}}{\|f^h\|_{H^{-1,h}}} \le \sqrt{\epsilon_M}.$$

Here the superscript $l = l(k)$ denotes the iteration index of Algorithm B for fixed k. For a discussion and numerical results in the case where the approximation errors due to the discretization of the underlying function space problems are incorporated into the algorithmic framework, e.g., when stopping the algorithm, we refer to the next section 6.2.

The initialization of γ is as follows: In the infeasible case we propose a choice of γ_0 based on the deviation of the linearization of $V(\gamma)$ at $\gamma = \gamma_r$ from the objective value of the unconstrained problem $(\hat{\mathrm{P}})$ at the projection of y_{γ_r} onto the feasible set. In our realization of this heuristic we choose $\gamma_r = 0$ and compute $\hat y$, $V(0)$, and $\dot V(0)$. Then we set

$$\gamma_0 = \max\left\{1, \zeta\frac{J(y_b) - V(0)}{\dot V(0)}\right\}, \tag{49}$$

where $\zeta \in (0,1]$ is some fixed constant, $y_b(x) = \min(\hat y, \psi(x))$, and J denotes the objective function of $(\hat{\mathrm{P}})$. Note that $\hat y$ is the minimizer of the unconstrained problem $(\hat{\mathrm{P}})$. For the examples below we use $\zeta = 1$. In the feasible case we choose a reference value γ_r, e.g., $\gamma_r = 1$, and solve the path problem (P_γ). Then we choose

$$\gamma_0 = \gamma_r + \frac{J(\hat y) - V(\gamma_r)}{\dot V(\gamma_r)}, \tag{50}$$

where $\hat{y}$ denotes the minimizer of the discretized unconstrained problem $(\hat{\mathrm{P}})$. If $\hat{y}$ is not feasible for (P), then one has $J(\hat{y}) < V(\gamma_r)$ and hence $\gamma_0 > \gamma_r$.

When applied to P1, P2, and P3 for $h = 1/128$ and with $\tau_k = 0.01^{k+1}$, we obtain the results shown in Figure 5 and Table 6.1.

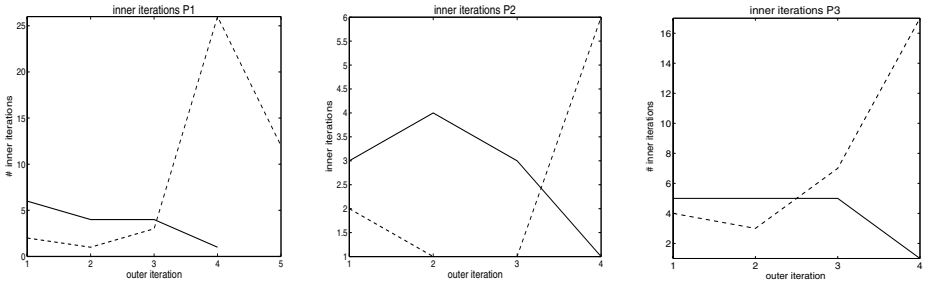


FIG. 5. *Number of inner iterations (vertical axis) per outer iteration for* P1 *(left plot),* P2 *(middle plot), and* P3 *(right plot); solid line – infeasible case, dashed line – feasible case.*

TABLE 6.1
Comparison of iteration counts.

	P1		P2		P3	
Version	# outer	# inner	# outer	# inner	# outer	# inner
Feasible	5	44	4	10	4	31
Infeasible	4	15	4	11	4	16

From our test runs, also for other test problems, we observe the following characteristics:

- For the feasible version the number of inner iterations exhibits an increasing tendency until a saturation value is reached, and then, unless the algorithm stops at an approximate solution, it starts to decrease. For the infeasible version we typically observe that the first couple of iterations require several inner iterations. As the outer iterations proceed the number of inner iterations drops eventually to one. We also tested less aggressive γ-updates compared to the ones used here, e.g., updates based on $\gamma_{k+1} = \xi\gamma_k$ with $\xi > 1$ fixed.
- The numerically observable convergence speed of y_{γ_k} towards y^* in $H_0^1(\Omega)$ is typically superlinear. This can be seen from Figure 6, where the plots for the discrete versions q_k^h of the quotients

$$q_k = \frac{|y_{\gamma_{k+1}} - y^*|_{H_0^1}}{|y_{\gamma_k} - y^*|_{H_0^1}}$$

are shown. Note that the vertical axis uses a logarithmic scale. In the first row, for P1 we depict the behavior of q_k^h for $h = 2^{-i}$, $i = 5, 6, 7, 8$, for the infeasible case (left plot) and the feasible case (right plot). We observe that the convergence rate is stable with respect to decreasing mesh size h. In the second row we see the behavior of q_k^h for P2 and P3, with $h = 2^{-7}$. Again, we observe a superlinear rate of convergence. With respect to decreasing h the same conclusion as for P1 holds true. These stability results provide a link between our function space theory and the numerical realization of the algorithms.

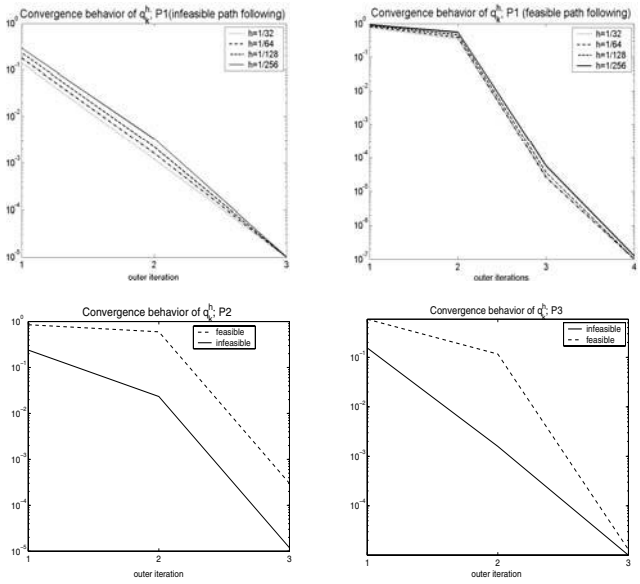


FIG. 6. *Discrete quotients* q_k^h *for* P1 *and various mesh sizes* h *(upper row) and for* P2 *(lower left) and* P3 *(lower right) for* $h = 1/128$.

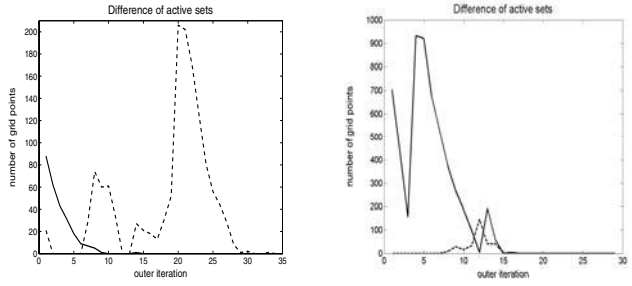


FIG. 7. *Difference in active sets for* P1 *and* P2*; solid line – infeasible case, dashed line – feasible case.*

- In connection with the convergence speed it is of interest how the detection process of the correct active set works. For the rather aggressive γ-updates used in Algorithm EP the difference between two successive active sets is zero typically only in the last iteration. However, if a less aggressive strategy for updating γ is used, then it is to be expected, that the difference of active sets might become zero earlier along the iteration. In Figure 7, for the strategy $\gamma_{k+1} = 2\gamma_k$, we show the difference of successive active sets; i.e., the vertical axis relates to the number of grid points that are in $\mathcal{A}_{k+1}$ but not in $\mathcal{A}_k$ and vice versa. We detect that for the infeasible case there exists an iteration index $\bar{k}$ after which the difference is constantly zero. This behavior is a strong indication that the correct active set was detected. It suggests that we fix this set $\mathcal{A}_{\bar{k}}$ and set $\bar{y}_{|\mathcal{A}_{\bar{k}}} = \psi_{|\mathcal{A}_{\bar{k}}}$, $\mathcal{I}_{\bar{k}} = \Omega \setminus \mathcal{A}_{\bar{k}}$, and $\bar{\lambda}_{\mathcal{I}_{\bar{k}}} = 0$. Then one computes $\bar{y}_{|\mathcal{I}_{\bar{k}}}$ and $\bar{\lambda}_{\mathcal{A}_{\bar{k}}}$ such that $a(\bar{y}, v) + \langle \bar{\lambda}, v \rangle_{H^{-1}, H_0^1} = (f, v)$ for all $v \in H_0^1(\Omega)$, and checks whether $(\bar{y}, \bar{\lambda})$ satisfies (7). If this is the case, then the solution is

found; otherwise $\gamma_{\bar{k}}$ is updated and the iteration continued. If we apply this technique for P1 in the infeasible case, then the algorithm stops at iteration 15 (35 inner iterations) with the exact discrete solution, as compared to 28 outer and 47 inner iterations without the additional stopping rule. There were four iterations where the additional system solve was necessary but without obtaining the numerical solution. Hence, with respect to system solves, the amount of work drops from 47 solves to 39 ($= 35 + 4$). A similar observation is true for P2 and P3. In the feasible case, however, this strategy yields no reduction of iterations. Here, typically the correct active set is determined in the last iteration (for large enough γ).

- The dependence of the iteration number on the mesh size of the discretization for P1 is depicted in Table 6.2 (those for P2 and P3 are similar). In parenthesis we show the number of inner iterations. The results clearly indicate that the outer iterations are mesh independent, while the number of inner iterations increases as the mesh size decreases. In the third row we display the results obtained by applying Algorithm A for the solution of the unregularized problem (P) with data according to P1. If we compare these results with those of the infeasible exact path-following algorithm, we find that for sufficiently small mesh sizes h the infeasible version of Algorithm EP requires significantly fewer iterations than does Algorithm A, which is also an infeasible algorithm. Also, the number of iterations required by Algorithm A exhibits a relatively strong dependence on h when compared to Algorithm EP in the infeasible case. Similar observations apply also to P2 and P3. This shows that taking into account the function space theoretic properties when regularizing problem (P) results in an algorithmic framework which performs stably with respect to decreasing mesh size of the discretization.

TABLE 6.2
Comparison of iteration counts for different mesh sizes.

	Mesh size h				
Version	1/16	1/32	1/64	1/128	1/256
EP feasible	5(19)	5(23)	5(30)	5(44)	5(72)
EP infeasible	4(8)	4(11)	4(13)	4(15)	4(19)
Algorithm A	4	8	14	26	48

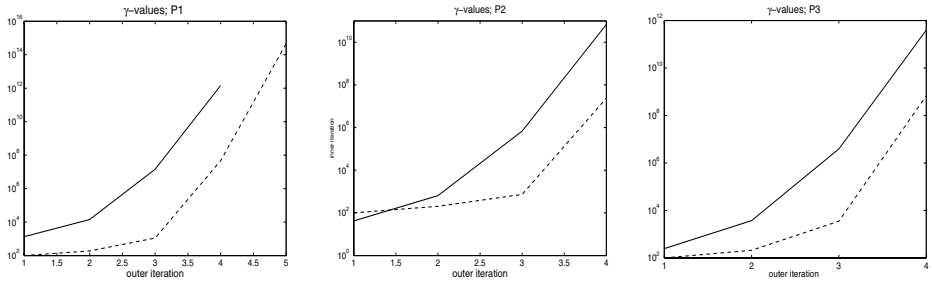


FIG. 8. *γ-updates; solid line – infeasible case, dashed line – feasible case.*

- From the plots in Figure 8, where the y-axis again has a logarithmic scale, it can be seen that our strategy (45) produces a rapidly increasing sequence

$\{\gamma_k\}$. The plots in Figure 8 depict the increase of γ_k as a function of the iteration number. The question arises of whether one could increase γ more rapidly. Numerical examples employing an ad hoc strategy show that if γ is increased too quickly, then the numerical error may prevent the residuals of the first order system from dropping below $\sqrt{\epsilon_M}$. This effect is due to the ill-conditioning of the linear systems for large γ. On the other hand, small increases in γ result in a slow convergence speed of Algorithm EP. Further, in our test runs and as can be seen from Figure 8, the feasible version of Algorithm EP is less aggressive in enlarging γ_k.

6.2. Inexact path-following. While exact path-following is primarily of theoretical interest, the development of inexact path-following techniques that keep the number of iterations as small as possible is of more practical importance. The strategy in the previous section relies on the fact that for every γ_k the corresponding point on the primal-dual path is computed. This, however, is not the case for inexact techniques and, as a consequence, a different update strategy for the path parameter γ is necessary. A common concept in inexact path-following methods is based on the definition of an appropriate neighborhood of the path; see, e.g., [3] and the references therein for a noninterior neighborhood-based path-following method, or [5, 16, 18, 19] for path-following techniques related to interior point methods. It is typically required that the primal-dual iterates stay within the neighborhood of the path, with the goal to reduce the computational burden while still maintaining convergence of the method.

We define

$$r_\gamma^1(y,\lambda) = \| -\Delta y + \lambda - f\|_{H^{-1}}, \tag{51a}$$

$$r_\gamma^2(y,\lambda) = \|\lambda - \max(0, \lambda + \gamma(y-\psi))\|_{H^{-1}}, \tag{51b}$$

and the neighborhood

$$\mathcal{N}(\gamma) := \left\{ (y,\lambda) \in H_0^1(\Omega) \times L^2(\Omega) : \|(r_\gamma^1(y,\lambda), r_\gamma^2(y,\lambda))^\top\|_2 \le \frac{\tau}{\sqrt{\gamma}} \right\} \tag{52}$$

in the infeasible case and

$$\mathcal{N}(\gamma) := \left\{ (y,\lambda) \in H_0^1(\Omega) \times L^2(\Omega) : \|(r_\gamma^1(y,\lambda), r_\gamma^2(y,\lambda))^\top\|_2 \le \frac{\tau}{\sqrt{\gamma}} \right.$$
$$\left. \wedge \frac{\partial}{\partial\gamma} J(y;\gamma) \le 0 \right\} \tag{53}$$

in the feasible case. Above, $\tau > 0$ denotes some fixed parameter. Note that adding the condition $\frac{\partial}{\partial\gamma} J(y;\gamma) \ge 0$ in (52) yields no further restriction, since this condition is automatically satisfied by the structure of $J(y;\gamma)$. We also point out that the conditions on the derivative of $J(y;\gamma)$ are included in (52) and (53), respectively, in order to qualitatively capture (up to first order) the analytical properties of the primal-dual path.

Next we specify our framework for an **i**nexact **p**ath-following algorithm.

ALGORITHM IP.

(i) Initialize γ_0 according to (49) in the infeasible case or (50) in the feasible case; set $k := 0$.

(ii) Apply Algorithm B to find $(y_{k+1}, \lambda_{k+1}) \in \mathcal{N}(\gamma_k)$.
(iii) Update γ_k to obtain γ_{k+1}.
(iv) Set $k = k + 1$, and go to (ii).

Note that if in step (ii) the path-problem (P_γ) is solved, then $r^1_\gamma(y_\gamma, \lambda_\gamma) = r^2_\gamma(y_\gamma, \lambda_\gamma) = 0$.

As is the case with primal-dual path-following interior point methods, the update strategy for γ in step (iii) of Algorithm IP is a delicate issue. If the increase of γ from one iteration to the next is rather small, then we follow the path closely, and the convergence speed is slow. If the γ-update is too aggressive, then step (ii) requires many iterations of Algorithm B to produce iterates in the neighborhood. We propose the following strategy, which performed very well in our numerical tests.

We introduce the *primal infeasibility measure* ρ^F and the *complementarity measure* ρ^C as follows:

$$\rho^F_{k+1} := \int_\Omega (y_{k+1} - \psi)^+ dx, \tag{54}$$

$$\rho^C_{k+1} := \int_{\mathcal{I}_{k+1}} (y_{k+1} - \psi)^+ dx + \int_{\mathcal{A}_{k+1}} (y_{k+1} - \psi)^- dx, \tag{55}$$

where $(\cdot)^- = -\min(0, \cdot)$ and $(\cdot)^+ = \max(0, \cdot)$. Note that at the optimal solution both measures vanish. Further, we point out that ρ^C is related to the duality measure well known from primal-dual path-following interior point methods. These measures are used in the following criterion for updating γ:

$$\gamma_{k+1} \geq \max\left(\gamma_k \max\left(\tau_1, \frac{\rho^F_{k+1}}{\rho^C_{k+1}}\right), \frac{1}{(\max(\rho^F_{k+1}, \rho^C_{k+1}))^q}\right) \tag{56}$$

with $\tau_1 > 1$ and $q \geq 1$. The first term in the outermost max-expression is used because of our observation that $\rho^F_{k+1} \geq \rho^C_{k+1}$ in the infeasible case. If ρ^C is small compared to ρ^F, we find that the iterates primarily lack feasibility as compared to complementarity. Therefore, a strong increase in γ, which aims at reducing constraint infeasibility, is favorable. If both measures are of almost the same size and rather small, then the second term in the outer max-expression should yield a significant increase in γ. Typically $q \in [\frac{3}{2}, 2]$ is chosen, which induces growth rates for γ.

If there is still a significant change in the active sets from one iteration to the next and the update γ_{k+1} based on (56) would be too large compared to γ_k, then many inner iterations would be necessary to keep track of the path, or very conservative γ-updates in the following iterations have to be chosen. We safeguard the γ-updates by utilizing our model function $m(\gamma)$, which was found to be a reliable tool. In fact, in updating γ, large deviations from $m(\gamma)$ are prohibited by comparing the value of the tangent to $J(y; \gamma)$ at $\gamma = \gamma_k$ with the actual model value. If necessary and as long as γ_{k+1} is much larger than γ_k, we reduce the actual γ-value until

$$|t_k(\gamma_{k+1}) - m_k(\gamma_{k+1})| \leq \tau_3 |J(y_{k+1}; \gamma_k) - J(y_k; \gamma_{k-1})| \tag{57}$$

with $0 < \tau_3 < 1$, $t_k(\gamma) = J(y_{k+1}; \gamma_k) + \frac{\partial J}{\partial \gamma}(y_{k+1}; \gamma_k)(\gamma - \gamma_k)$, and $m_k(\gamma)$ the model related to γ_k. Recall that $m_k(\gamma_k) = J(y_{k+1}; \gamma_k)$. The motivation of this strategy utilizes the good approximation qualities of our models. Indeed, for small γ the distance between t_k and m_k might be large, but so is $|J(y_{k+1}; \gamma_k) - J(y_k; \gamma_{k-1})|$ since the change in the function value is expected to be relatively large for small γ. For large γ, however, both difference measures tend to be small.

Concerning the numerical realization of Algorithm IP in the discrete setting we point out that by an a posteriori analysis of the discretization errors one finds that the norm of the residuals in (51a) and (51b) can be approximated typically to the order of h. This can be used as an upper bound for γ in the discrete versions of (52) and (53), respectively. However, since, on a fixed grid, our discrete versions of (P) and (P_γ) are consistent (as $\gamma \to \infty$) and admit unique solutions in $\mathbb{R}^{N_h}$, where $N_h \in \mathbb{N}$ depends on the mesh size of discretization h, it is of interest to consider $\gamma \to \infty$. On a fixed grid, this allows us also to study the behavior of our discretized algorithms as finite dimensional solvers for problems similar to the discrete versions of the ones under consideration. With respect to the latter aspect, below we report on test runs of Algorithm IP when applied to our test problems P1, P2, and P3. The parameters had values $q = 1.5$, $\tau_1 = 10$, $\tau_3 = 0.999$, $\tau = 1e6$. The stopping rule for the outer iteration is as before.

P1. The infeasible version of Algorithm IP requires 9 outer iterations and at most 2 inner iterations per outer iteration. In particular, in many iterations the criterion $(y_{k+1}, \lambda_{k+1}) \in \mathcal{N}(\gamma_k)$ was satisfied within 1 inner iteration. The feasible version of Algorithm IP stops after 11 iterations. With respect to inner iterations in the feasible case we note that more than 1 or 2 inner iterations were necessary only in the last 3 outer iterations with 3, 4, and 6 inner iterations, respectively. For both runs, the behavior of the measures ρ^F and ρ^C is shown in Figure 9. Note that the vertical scale is a logarithmic one. The left plot corresponds to the infeasible case. The feasibility measure ρ^F and the complementarity measure ρ^C are both convergent at a superlinear rate. In the feasible case, which is depicted in the right plot, we observe that ρ^C is only linearly convergent. In some iterations we have $\rho_k^F > 0$. However, the constraint violation is of the order of the machine precision and thus negligible.

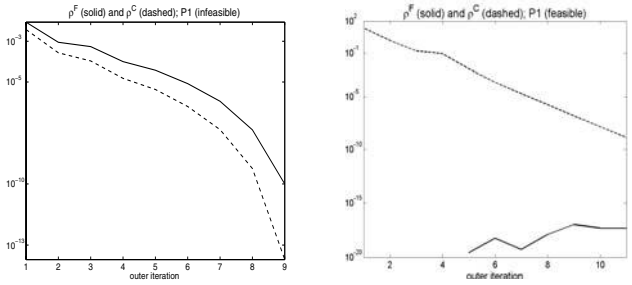


FIG. 9. *Behavior of the measures ρ^F (solid) and ρ^C (dashed) for* P1*, left plot – infeasible case, right plot – feasible case.*

P2. For this test problem the infeasible version of Algorithm IP required 11 iterations with one inner iteration per outer iteration. The feasible version needed 6 outer iterations and 9 inner iterations.

P3. The behavior of Algorithm IP for solving P3 is comparable to its behavior for P1 and P2. In fact, the infeasible version required 11 outer iterations and 11 inner iterations for solving the discrete problem. The feasible variant of Algorithm IP stopped successfully after 9 outer and 19 inner iterations. For the latter run, in the next-to-last iteration 5 inner iterations were necessary; otherwise at most 2 inner iterations were needed. With respect to the behavior of the decrease of the measures ρ^C and ρ^F, an observation similar to the one obtained from Figure 9 for P1 holds true. We remark only that in the feasible case ρ^C exhibits an almost superlinear

convergence behavior.

Compared to the exact path-following strategy of Algorithm EP, the inexact path-following concept of Algorithm IP is in many cases more efficient. In Table 6.3 we provide the number of outer and inner iterations for exact versus inexact path-following. In parenthesis we write the number of inner iterations.

TABLE 6.3
Comparison of iteration counts between exact and inexact path-following.

	Infeasible case			Feasible case		
	P1	P2	P3	P1	P2	P3
EP	4 (15)	4 (11)	4 (16)	5 (44)	4 (10)	4 (31)
IP	9 (12)	11 (11)	11 (11)	11 (25)	6 (9)	9 (19)

Finally we address the issue of how to incorporate the approximation error due to the discretization of function space quantities; see [6, 7]. First note that with (8) holding (which is the case for P3), the discretization of the residual in the definition of the neighborhoods (52), respectively (53), approximates the original one to the order of h. Hence, in our discrete version of Algorithm IP the neighborhood criterion

$$\|(r_\gamma^1(y,\lambda), r_\gamma^2(y,\lambda))^\top\|_2 \leq \frac{\tau}{\sqrt{\gamma}}$$

becomes

$$\|(r_\gamma^{1,h}(y,\lambda), r_\gamma^{2,h}(y,\lambda))^\top\|_2 \leq \max\left\{\sqrt{\epsilon_M}, \kappa_{\text{in}} h, \frac{\tau}{\sqrt{\gamma}}\right\},$$

with some constant $\kappa_{\text{in}} > 0$. We stop the outer iteration as soon as the discrete residual drops below $\max\{\kappa_{\text{out}} h, \sqrt{\epsilon_M}\}$, where $\kappa_{\text{out}} > 0$ is fixed. In our tests we use $\kappa_{\text{in}} = 1$ and $\kappa_{\text{out}} = 10$. Applying this strategy for the solution of P3, we obtain (outer) iteration numbers as displayed in Table 6.4. Here, in parenthesis we give the total number of inner iterations.

TABLE 6.4
Inexact path-following with h-dependent stopping of inner and outer iterations.

Version	Mesh size					
	1/16	1/32	1/64	1/128	1/256	1/512
IP	1 (1)	4 (4)	5 (5)	8 (8)	9 (10)	10 (10)

REFERENCES

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

[2] S. C. BRENNER AND L. R. SCOTT, *The Mathematical Theory of Finite Element Methods*, 2nd ed., Texts in Appl. Math. 15, Springer-Verlag, New York, 2002.

[3] X. CHEN AND P. TSENG, *Noninterior continuation methods for solving semidefinite complementarity problems*, Math. Program. Ser. A, 95 (2003), pp. 431–474.

[4] A. V. FIACCO AND G. P. MCCORMICK, *Nonlinear Programming, Sequential Unconstrained Minimization Techniques*, Classics in Appl. Math. 4, SIAM, Philadelphia, PA, 1990.

[5] A. FORSGREN, P. E. GILL, AND M. H. WRIGHT, *Interior methods for nonlinear optimization*, SIAM Rev., 44 (2002), pp. 525–597.

[6] C. Grossmann and A. A. Kaplan, *On the solution of discretized obstacle problems by an adapted penalty method*, Computing, 35 (1985), pp. 295–306.
[7] C. Grossmann and H.-G. Roos, *Numerik partieller Differentialgleichungen*, 2nd ed., Teubner Studienbücher Mathematik, B. G. Teubner, Stuttgart, 1994.
[8] W. Hackbusch, *Theorie und Numerik elliptischer Differentialgleichungen*, Teubner Verlag, Stuttgart, 1986.
[9] M. Hintermüller, *Inverse coefficient problems for variational inequalities: Optimality conditions and numerical realization*, M2AN Math. Model. Numer. Anal., 35 (2001), pp. 129–152.
[10] M. Hintermüller, K. Ito, and K. Kunisch, *The primal-dual active set strategy as a semi smooth Newton method*, SIAM J. Optim., 13 (2003), pp. 865–888.
[11] K. Ito and K. Kunisch, *Optimal control of elliptic variational inequalities*, Appl. Math. Optim., 41 (2000), pp. 343–364.
[12] K. Ito and K. Kunisch, *Semismooth Newton methods for state-constrained optimal control problems*, Systems Control Lett., 50 (2003), pp. 221–228.
[13] K. Ito and K. Kunisch, *Semismooth Newton methods for variational inequalities of the first kind*, M2AN Math. Model. Numer. Anal., 37 (2003), pp. 41–62.
[14] G. M. Troianiello, *Elliptic Differential Equations and Obstacle Problems*, Univ. Ser. Math., Plenum Press, New York, 1987.
[15] M. Ulbrich, *Semismooth Newton methods for operator equations in function spaces*, SIAM J. Optim., 13 (2003), pp. 805–842.
[16] R. J. Vanderbei, *Linear Programming, Foundations, and Extensions*, 2nd ed., Internat. Ser. Oper. Res. Management Sci. 37, Kluwer Academic Publishers, Boston, MA, 2001.
[17] M. Weiser, *Interior point methods in function space*, SIAM J. Control Optim., 44 (2005), pp. 1766–1786.
[18] S. J. Wright, *Primal-dual Interior-Point Methods*, SIAM, Philadelphia, PA, 1997.
[19] Y. Ye, *Interior Point Algorithms, Theory, and Analysis*, Wiley-Interscience, New York, 1997.

SIAM J. OPTIM.
Vol. 17, No. 1, pp. 188–217
© 2006 Society for Industrial and Applied Mathematics

MESH ADAPTIVE DIRECT SEARCH ALGORITHMS FOR CONSTRAINED OPTIMIZATION*

CHARLES AUDET† AND J. E. DENNIS, JR.‡

Abstract. This paper addresses the problem of minimization of a nonsmooth function under general nonsmooth constraints when no derivatives of the objective or constraint functions are available. We introduce the mesh adaptive direct search (MADS) class of algorithms which extends the generalized pattern search (GPS) class by allowing local exploration, called *polling*, in an asymptotically dense set of directions in the space of optimization variables. This means that under certain hypotheses, including a weak constraint qualification due to Rockafellar, MADS can treat constraints by the extreme *barrier* approach of setting the objective to infinity for infeasible points and treating the problem as unconstrained.

The main GPS convergence result is to identify limit points $\hat{x}$, where the Clarke generalized derivatives are nonnegative in a finite set of directions, called *refining directions*. Although in the unconstrained case, nonnegative combinations of these directions span the whole space, the fact that there can only be finitely many GPS refining directions limits rigorous justification of the barrier approach to finitely many linear constraints for GPS. The main result of this paper is that the general MADS framework is flexible enough to allow the generation of an asymptotically dense set of refining directions along which the Clarke derivatives are nonnegative.

We propose an instance of MADS for which the refining directions are dense in the hypertangent cone at $\hat{x}$ with probability 1 whenever the iterates associated with the refining directions converge to a single $\hat{x}$. The instance of MADS is compared to versions of GPS on some test problems. We also illustrate the limitation of our results with examples.

Key words. mesh adaptive direct search algorithms (MADS), convergence analysis, constrained optimization, nonsmooth analysis, Clarke derivatives, hypertangent, contingent cone

AMS subject classifications. 90C30, 90C56, 65K05, 49J52

DOI. 10.1137/040603371

1. Introduction.

We present and analyze a new *mesh adaptive direct search* (MADS) class of algorithms for minimizing a nonsmooth function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ under general constraints $x \in \Omega \neq \emptyset \subseteq \mathbb{R}^n$. For the form of the algorithm given here, the feasible region Ω may be defined through blackbox constraints given by an oracle, such as a computer code that returns a yes or no indicating whether or not a specified trial point is feasible.

In the unconstrained case, where $\Omega = \mathbb{R}^n$, this new class of algorithms occupies a position somewhere between the generalized pattern search (GPS) class [30], as organized in [8], and the Coope and Price frame-based methods [12]. A key advantage of MADS over GPS for both unconstrained and linearly constrained optimization is that local exploration of the space of variables is not restricted to a finite number of directions (called *poll directions*). This is the primary drawback of GPS algorithms in our opinion, and our main motivation in defining MADS was to overcome this

*Received by the editors January 20, 2004; accepted for publication (in revised form) November 28, 2005; published electronically May 3, 2006. The work of the first author was supported by FCAR grant NC72792 and NSERC grant 239436-01, and both authors were supported by AFOSR FA9550-04-1-0235, The Boeing Company, and ExxonMobil.

http://www.siam.org/journals/siopt/17-1/60337.html

†GERAD and Département de Mathématiques et de Génie Industriel, École Polytechnique de Montréal, C.P. 6079, Succ. Centre-ville, Montréal (Québec), H3C 3A7 Canada (Charles.Audet@gerad.ca, http://www.gerad.ca/Charles.Audet).

‡Computational and Applied Mathematics Department, Rice University, 8419 42nd Ave SW, Seattle, WA 98136 (dennis@rice.edu, http://www.caam.rice.edu/∼dennis).

restriction. MADS algorithms are frame-based methods. We propose a less general choice of frames than the choices allowed by Coope and Price. Our MADS frames are easy to implement, and they are specifically aimed at ensuring an asymptotically dense set of polling directions. We illustrate our ideas with an example algorithm that we call LTMADS because it is based on a random lower triangular matrix.

The convergence analysis here is based on Clarke's calculus [10] for nonsmooth functions. The analysis evolved from our previous work on GPS [3] where we gave a hierarchy of convergence results for GPS that show the limitations inherent in the restriction to finitely many directions. Specifically, we showed that for unconstrained optimization, GPS produces a limit point at which the gradient is zero if the function at that point is strictly differentiable [20]. Strict differentiability is just the requirement that the generalized gradient is a singleton, i.e., that $\partial f(\hat{x}) = \{\nabla f(\hat{x})\}$ in addition to the requirement that f is Lipschitz near $\hat{x}$. But if the function f is only Lipschitz near such a limit point $\hat{x}$, then Clarke's generalized directional derivatives [10] are provably nonnegative only for a finite set of directions $\hat{D} \subset \mathbb{R}^n$ (called the set of *refining directions*) whose nonnegative linear combinations span the whole space

$$f^\circ(\hat{x}; d) := \limsup_{y \to \hat{x},\ t \downarrow 0} \frac{f(y + td) - f(y)}{t} \geq 0 \quad \text{for all } d \in \hat{D}. \tag{1.1}$$

This result (1.1) for GPS is not as strong as stating that the generalized derivative is nonnegative for every direction in $\mathbb{R}^n$, i.e., that the limit point is a Clarke stationary point, or equivalently that $0 \in \partial f(\hat{x})$, the generalized gradient of f at $\hat{x}$ defined by

$$f^\circ(\hat{x}; v) \geq 0 \ \text{ for all } v \in \mathbb{R}^n \Leftrightarrow 0 \in \partial f(\hat{x}) := \{s \in \mathbb{R}^n : f^\circ(\hat{x}; v) \geq v^T s \text{ for all } v \in \mathbb{R}^n\}. \tag{1.2}$$

Example F in [2] shows that indeed the GPS algorithm does not necessarily produce a Clarke stationary point for Lipschitz functions because of the restriction to finitely many poll directions. This is so even if the gradient exists at the limit point $\hat{x}$. For the unconstrained case, this restriction can be overcome by assuming more smoothness for f, e.g., strict differentiability at $\hat{x}$ [3] as mentioned above.

However, even in the presence of simple bound constraints, the directional dependence of GPS cannot be overcome by any amount of smoothness, by using penalty functions, or by the use of the filter approach for handling constraints [4]. In contrast, MADS produces a limit point at which the Clarke derivatives are nonnegative for every direction in the tangent cone. The class of problems that MADS is designed for is similar but not the same as that of Lucidi, Sciandrone, and Tseng [23] and Price, Coope, and Dennis [27]. They also target nonlinear optimization but require that all functions be continuously differentiable and that the constraint derivatives be available.

Besides the advantages of an asymptotically dense set of refining directions, MADS can also treat a wide class of nonlinear constraints by the "barrier" approach. By this we mean that the algorithm is not applied directly to f but to the barrier function f_Ω, defined to be equal to f on Ω and $+\infty$ outside Ω. This way of rejecting infeasible points was shown to be effective for GPS with finitely many linear constraints by Lewis and Torczon [21]. However, their proof requires that the tangent cone generators of the feasible region at boundary points near an iterate be known at each iteration.

For LTMADS, a specific implementation of the general framework MADS, no special effort is needed for the barrier approach to be provably effective with probability 1 on nonlinear constraints satisfying a mild constraint qualification due to Rockafellar

[29]—that there exists a hypertangent vector at the limit point. A key advantage of the barrier approach is that one can avoid expensive function calls to f whenever a constraint is violated. Indeed, the question of feasibility of a trial point needs only a yes or no answer—the constraints do not need to be given by a known algebraic condition. Marsden [24] exploited this capability in an insightful way to avoid a significant number of full 3D LES turbulence simulations when the nonlinear constraints were violated in MADS applied to a trailing edge design problem.

The class of algorithms presented here differs significantly from previous GPS extensions [4, 22] to nonlinear constraints. Treating constraints as we do motivates us to use the generalization of the Clarke derivative presented in Jahn [18]. Jahn's approach is aimed at a case like ours where the evaluation of f is restricted to points in the feasible domain Ω. Thus instead of (1.1) we use the following definition of the Clarke generalized derivative at $\hat{x} \in \Omega$ in the direction $v \in \mathbb{R}^n$:

$$f^\circ(\hat{x}; v) := \limsup_{\substack{y \to \hat{x},\ y \in \Omega \\ t \downarrow 0,\ y + tv \in \Omega}} \frac{f(y + tv) - f(y)}{t}. \tag{1.3}$$

Both definitions (1.1) and (1.3) coincide when $\Omega = \mathbb{R}^n$ or when $\hat{x} \in int(\Omega)$.

The main theoretical objective of this paper is to show that under appropriate assumptions, a MADS algorithm produces a constrained Clarke stationary point, i.e., a limit point $\hat{x} \in \Omega$ satisfying the following necessary optimality condition:

$$f^\circ(\hat{x}; v) \ \geq \ 0 \text{ for all } v \in T_\Omega^{Cl}(\hat{x}), \tag{1.4}$$

where $T_\Omega^{Cl}(\hat{x})$ is the Clarke tangent cone to Ω at $\hat{x}$ (see [10] or Definition 3.5).

The paper is organized into two main parts. First, sections 2 and 3 present the abstract MADS algorithm class and its convergence analysis. The analysis revolves around three types of tangent cones. This allows us to tie some convergence results to local differentiability of the function f at limit points satisfying certain constraint qualifications. We present sufficient conditions under which (1.4) holds. We discuss the consequences of this when the algorithm is applied to an unconstrained problem, or when the set Ω is regular in the sense of Definition 3.7 or [10]. We also give a stronger constraint qualification ensuring that MADS produces a contingent KKT stationary point (Definition 3.11) if f is strictly differentiable. The reader will find a quite different algorithm analyzed using the same concepts in [15].

Then in sections 4 and 5, we give an instance of MADS along with numerical experiments to compare MADS with standard GPS. On an artificial example, where GPS is well known to stagnate, we show that MADS reaches the global optimum. We give a comparison on a parameter fitting problem in catalytic combustion kinetics on which we know that GPS performs well [17]. We also give an example illustrating the power of being able to handle even simple nonlinear constraints by the barrier approach. We also use this example to illustrate that MADS can cope surprisingly well as the dimension of the problem increases. The final example shows the value of randomly generated polling directions for a problem with a narrowing feasible region.

Notation. $\mathbb{R}, \mathbb{Z}$, and $\mathbb{N}$, respectively, denote the sets of real numbers, integers, and nonnegative integers. For $x \in \mathbb{R}^n$ and $\delta \in \mathbb{R}_+$, $B_\delta(x)$ denotes the open ball of radius δ centered at x. For a matrix D, the notation $d \in D$ indicates that d is a column of D. The iteration numbers are denoted by the index k.

2. Mesh adaptive direct search algorithms. MADS is an iterative feasible-point algorithm. Given an initial iterate $x_0 \in \Omega$, a MADS algorithm attempts to

locate a minimizer of the function f over Ω by evaluating f_Ω at some trial points. The algorithm does not require any derivative information for f. This is essential when ∇f is unavailable, either because it does not exist, or it cannot be accurately estimated due to noise in f or other reasons. At each iteration, a finite number of trial points are generated and the infeasible trial points are discarded. The objective function values at the feasible trial points are compared with the current incumbent value $f_\Omega(x_k)$, i.e., the best feasible objective function value found so far. Each of these trial points lies on the *current mesh*, constructed from a finite set of n_D directions $D \subset \mathbb{R}^n$ scaled by a *mesh size parameter* $\Delta_k^m \in \mathbb{R}_+$. Just as in GPS, this mesh is not actually constructed, it just underlies the algorithm.

There are two restrictions on the set D. First, D must be a positive spanning set [14], i.e., nonnegative linear combinations of its elements must span $\mathbb{R}^n$. Second, each direction $d_j \in D$ (for $j = 1, 2, \ldots, n_D$) must be the product Gz_j of some fixed nonsingular generating matrix $G \in \mathbb{R}^{n\times n}$ by an integer vector $z_j \in \mathbb{Z}^n$. For convenience, the set D is also viewed as a real $n \times n_D$ matrix.

DEFINITION 2.1. *At iteration k, the current mesh is defined to be the following union:*

$$M_k = \bigcup_{x\in S_k} \{x + \Delta_k^m Dz : z \in \mathbb{N}^{n_D}\},$$

where S_k is the set of points where the objective function f had been evaluated by the start of iteration k.

In the definition above, the mesh is defined to be the union of sets over S_k. Defining the mesh this way ensures that all previously visited points lie on the mesh, and that new trial points can be selected around any of them using the directions in D. This definition of the mesh is identical to the one in [4] and generalizes the one in [3].

The mesh is conceptual in the sense that it is never actually constructed. In practice, one can easily make sure that the strategy for generating trial points is such that they all belong to the mesh. One simply has to verify in Definition 2.1 that x belongs to S_k and that z is an integer vector. The objective of the iteration is to find a feasible trial mesh point with a lower objective function value than the current incumbent value $f_\Omega(x_k)$. Such a trial point is called an *improved mesh point*, and the iteration is called a *successful iteration*. There are no sufficient decrease requirements on the objective function value.

The evaluation of f_Ω at a trial point x is done as follows. First, the constraints defining Ω are tested to determine if x is feasible or not. Indeed, since some of the constraints defining Ω might be expensive or inconvenient to test, one would order the constraints to test the easiest ones first. If $x \notin \Omega$, then $f_\Omega(x)$ is set to $+\infty$ without evaluating $f(x)$, and perhaps without evaluating all the constraints defining Ω. In effect, this means we discard the infeasible trial points. On the other hand, if $x \in \Omega$, then $f(x)$ is evaluated. This remark may seem obvious, but it saves computation [24], and it is needed in the proof of Theorem 3.12.

Each iteration is divided into two steps. The first, called the SEARCH step, has the same flexibility as in GPS. It allows evaluation of f_Ω at any finite number of mesh points. Any strategy can be used in the SEARCH step to generate a finite number of trial mesh points. Restricting the SEARCH points to lie on the mesh is a way in which MADS is less general than the frame methods of Coope and Price [12]. The SEARCH is said to be empty when no trial points are considered. The drawback to the SEARCH flexibility is that it cannot be used in the convergence analysis—except to provide counterexamples as in [2]. More discussion of SEARCH steps is given in [1, 8].

When an improved mesh point is generated, then the iteration may stop, or it may continue if the user hopes to find a more improved mesh point. In either case, the next iteration will be initiated with a new incumbent solution $x_{k+1} \in \Omega$ with $f_\Omega(x_{k+1}) < f_\Omega(x_k)$ and with a mesh size parameter Δ^m_{k+1} equal to or larger than Δ^m_k (the exact rules for updating this parameter are presented in (2.1)). Coarsening the mesh when improvements in f_Ω are obtained can speed convergence.

Whenever the SEARCH step fails to generate an improved mesh point, then the second step, called the POLL, is invoked before terminating the iteration. The POLL step consists of a local exploration of the space of optimization variables near the current incumbent solution x_k. The difference between the MADS and the GPS algorithms lies exactly in this POLL step. For this reason, our numerical comparisons in what follows use empty, or very simple, SEARCH steps in order to illustrate the value of the MADS POLL step.

When the iteration fails in generating an improved mesh point, then the next iteration is initiated from any point $x_{k+1} \in S_{k+1}$ with $f_\Omega(x_{k+1}) = f_\Omega(x_k)$. There is usually a single such incumbent solution, and x_{k+1} is set to x_k. The mesh size parameter Δ^m_{k+1} is reduced to increase the mesh resolution in order to allow the evaluation of f at trial points closer to the incumbent solution. More precisely, given a fixed rational number $\tau > 1$, and two integers $w^- \le -1$ and $w^+ \ge 0$, the mesh size parameter is updated as follows:

$$\Delta^m_{k+1} = \tau^{w_k}\Delta^m_k \quad \text{for some } w_k \in \begin{cases} \{0, 1, \ldots, w^+\} & \text{if an improved mesh} \\ & \text{point is found} \\ \{w^-, w^- + 1, \ldots, -1\} & \text{otherwise.} \end{cases} \tag{2.1}$$

Everything up to this point in the section applies to both GPS and MADS. We now present the key difference between both classes of algorithms. For MADS, we introduce the *poll size parameter* $\Delta^p_k \in \mathbb{R}_+$ for iteration k. This new parameter dictates the magnitude of the distance from the trial points generated by the POLL step to the current incumbent solution x_k. In GPS, there is a single parameter to represent these quantities: $\Delta_k = \Delta^p_k = \Delta^m_k$. In MADS, the strategy for updating Δ^p_k must be such that $\Delta^m_k \le \Delta^p_k$ for all k, and moreover, it must satisfy

$$\lim_{k \in K} \Delta^m_k = 0 \text{ if and only if } \lim_{k \in K} \Delta^p_k = 0 \text{ for any infinite subset of indices } K. \tag{2.2}$$

An implementable updating strategy satisfying these requirements is presented in section 4.

We now move away from the GPS terminology, and toward that of Coope and Price because it is better suited to describe MADS. The set of trial points considered during the POLL step is called a *frame*. The frames of Coope and Price can be more general than MADS frames in a way not important to the present discussion. For this reason, we do not digress to discuss their general definition here [11].

The MADS frame is constructed using a current incumbent solution x_k (called the *frame center*) and the poll and mesh size parameters Δ^p_k and Δ^m_k to obtain a positive spanning set of directions D_k. Unlike GPS, generally the MADS set of directions D_k is not a subset of D.

DEFINITION 2.2. *At iteration k, the MADS frame is defined to be the set*

$$P_k = \{x_k + \Delta^m_k d : d \in D_k\} \subset M_k,$$

where D_k is a positive spanning set such that $0 \notin D_k$ and for each $d \in D_k$,

- *d can be written as a nonnegative integer combination of the directions in D: $d = Du$ for some vector $u \in \mathbb{N}^{n_{D_k}}$ that may depend on the iteration number k*
- *the distance from the frame center x_k to a frame point $x_k + \Delta_k^m d \in P_k$ is bounded above by a constant times the poll size parameter: $\Delta_k^m \|d\| \leq \Delta_k^p \max\{\|d'\| : d' \in D\}$*
- *limits (as defined in Coope and Price [11]) of the normalized sets $\mathcal{D}_k = \left\{\frac{d}{\|d\|} : d \in D_k\right\}$ are positive spanning sets.*

The set of all poll directions $\mathcal{D} = \bigcup_{k=1}^{\infty} \mathcal{D}_k$ is said to be asymptotically dense if the closure of the cone generated by $\mathcal{D}$ equals $\mathbb{R}^n$.

If the POLL step fails to generate an improved mesh point, then the frame is said to be a *minimal frame*, and the frame center x_k is said to be a *minimal frame center*. This leads to mesh refinement: $\Delta_{k+1} < \Delta_k$ (see (2.1)).

The algorithm is stated formally below. It is very similar to GPS, with differences in the POLL step, and in the new poll size parameter.

A GENERAL MADS ALGORITHM

- INITIALIZATION: Let $x_0 \in \Omega$, $\Delta_0^m \leq \Delta_0^p$, D, G, τ, w^- and w^+ satisfy the requirements given above.
  Set the iteration counter $k \leftarrow 0$.
- SEARCH AND POLL STEP: Perform the SEARCH and possibly the POLL steps (or only part of them) until an improved mesh point x_{k+1} is found on the mesh M_k (see Definition 2.1).
  - OPTIONAL SEARCH: Evaluate f_Ω on a finite subset of trial points on the mesh M_k.
  - LOCAL POLL: Evaluate f_Ω on the frame P_k (see Definition 2.2).
- PARAMETER UPDATE: Update Δ_{k+1}^m according to (2.1), and Δ_{k+1}^p so that (2.2) is satisfied.
  Set $k \leftarrow k + 1$ and go back to the SEARCH and POLL step.

The crucial distinction and advantage of MADS over GPS is that the MADS mesh size parameter Δ_k^m may go to zero more rapidly than Δ_k^p. Consequently, the directions in D_k used to define the frame may be selected in a way so that asymptotically they are not confined to a finite set. Note that in GPS both Δ_k^m and Δ_k^p are equal: a single parameter plays the role of the mesh and poll size parameters, and therefore, the number of positive spanning sets that can be formed by subsets of D is constant over all iterations.

For example, suppose that in $\mathbb{R}^2$ the set D is composed of the eight directions $\{(d_1, d_2)^T \neq (0, 0)^T : d_1, d_2 \in \{-1, 0, 1\}\}$. There are a total of eight distinct positive bases containing three directions that can be constructed from D. Figures 2.1 and 2.2 illustrate some possible frames in $\mathbb{R}^2$ for three values of Δ_k^m. The frames in Figure 2.1 are generated by a GPS instance, and are such that $\Delta_k^p = \Delta_k^m$. Regardless of k and of the mesh or poll size parameters, each direction in D_k is confined to be selected in D.

The frames in Figure 2.2 are generated by an instance of MADS with $\Delta_k^p = n\sqrt{\Delta_k^m}$. One can see that the new MADS algorithm may select the directions of D_k from a larger set. With the new algorithm, the frame may be chosen among the mesh points lying inside the square with the dark contour. We will present in section 4 an implementation of MADS ensuring that given any directions in $\mathbb{R}^n$, the

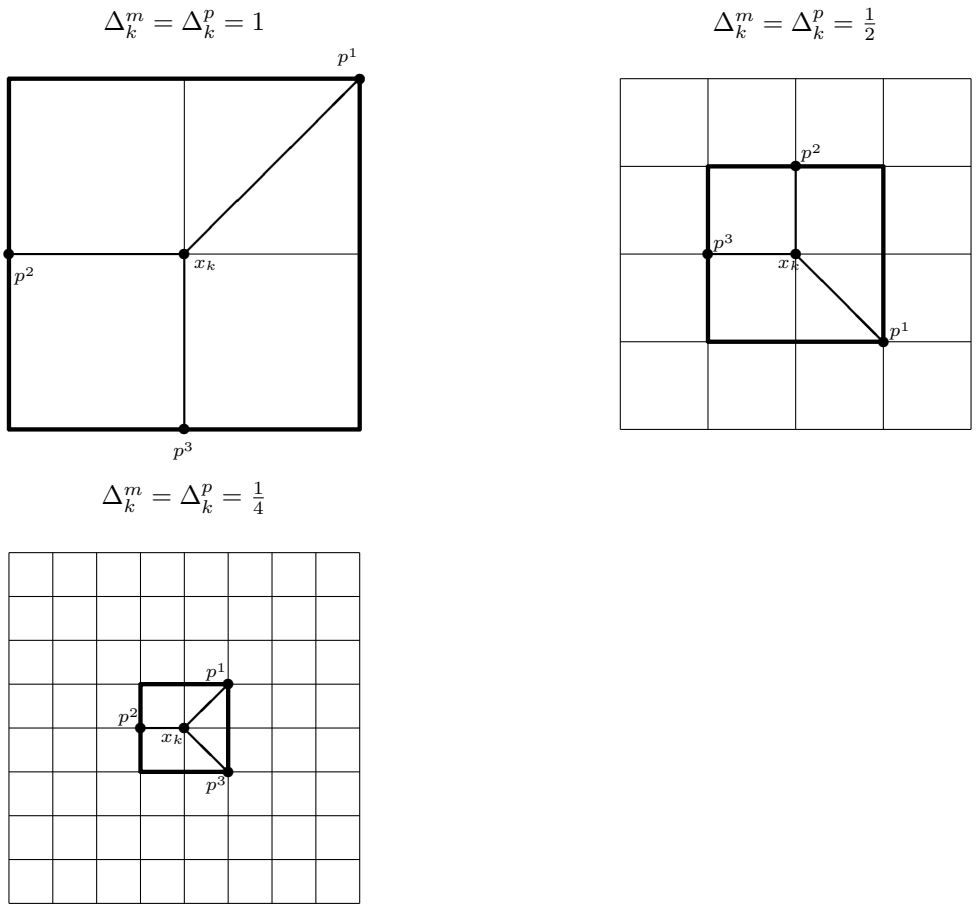


FIG. 2.1. *Example of GPS frames* $P_k = \{x_k + \Delta_k^m d : d \in D_k\} = \{p^1, p^2, p^3\}$ *for different values of* $\Delta_k^m = \Delta_k^p$. *In all three figures, the mesh* M_k *is the intersection of all lines.*

algorithm generates arbitrarily close poll directions, i.e., that the set of poll directions is asymptotically dense in $\mathbb{R}^n$.

We have presented above a general framework for MADS algorithms. The next section contains a detailed convergence analysis for that general framework. It presents sufficient conditions to ensure a hierarchy of convergence results based on the local differentiability of f (using the Clarke nonsmooth calculus) and on the local properties of Ω (using three types of tangent cones). The results rely on the assumption that a specific set of directions (called the refining directions—see Definition 3.2) be dense in a tangent cone. Then, in section 4 we propose a specific implementation called LTMADS, and give sufficient conditions to satisfy this assumption.

3. MADS convergence analysis. The convergence analysis below relies on the assumptions that $x_0 \in \Omega$, that $f(x_0)$ is finite, and that all iterates $\{x_k\}$ produced by the MADS algorithm lie in a compact set. Future work will relax the first assumption by incorporating the filter approach given in [4].

The section is divided into three subsections. The first recalls Torczon's [30] analysis of the behavior of the mesh size parameter and defines refining sequences as

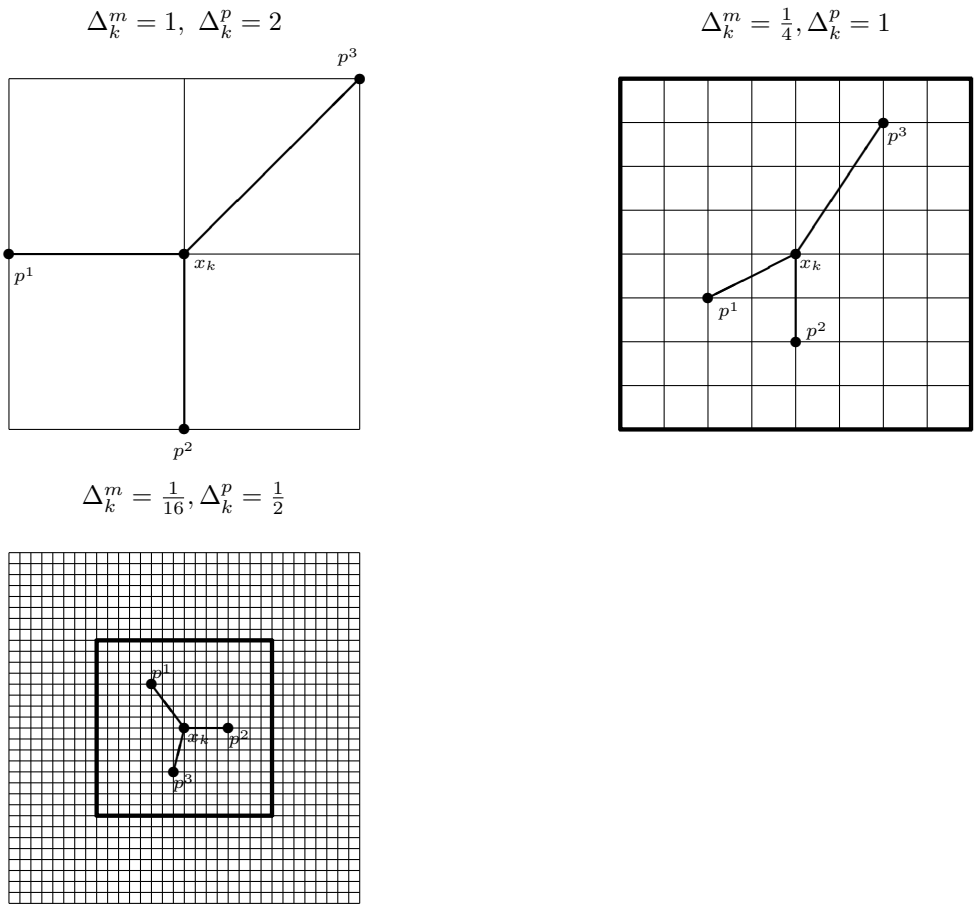


FIG. 2.2. *Example of MADS frames* $P_k = \{x_k + \Delta_k^m d : d \in D_k\} = \{p^1, p^2, p^3\}$ *for different values of* Δ_k^m *and* Δ_k^p. *In all three figures, the mesh* M_k *is the intersection of all lines.*

in [3]. It also defines the idea of a refining subsequence and a refining direction. The second subsection recalls the definitions of the hypertangent, Clarke, and contingent cones in addition to some results on generalized derivatives. The third contains a hierarchy of convergence results based on local properties of the feasible region Ω.

3.1. Preliminaries. Torczon [30] first showed the following result for unconstrained pattern search algorithms. Then Audet and Dennis [3] used the same technique for a description of GPS that is much closer to our description of MADS. The proof of this result for MADS is identical to that of GPS. The element necessary to the proof is that for any integer $N \geq 1$, the iterate x_N may be written as $x_N = x_0 + \sum_{k=0}^{N-1} \Delta_k^m D z_k$ for some vectors $z_k \in \mathbb{N}^{n_D}$. This is still true with our new way of defining the mesh and the frame (see Definitions 2.1 and 2.2).

PROPOSITION 3.1. *The poll and mesh size parameters produced by a MADS instance satisfy*

$$\liminf_{k \to +\infty} \Delta_k^p \ = \ \liminf_{k \to +\infty} \Delta_k^m \ = \ 0.$$

Price and Coope [26] propose a frame-based method for linearly constrained problems in which trial points are not confined to be on an underlying mesh. The price of this greater flexibility is that their convergence analysis is based upon the assumption that the frame size parameter (which plays a role similar to the MADS poll size parameter) goes to zero.

Since the mesh size parameter shrinks only at minimal frames, Proposition 3.1 guarantees that there are infinitely many minimal frame centers. The following definition specifies the subsequences of iterates and limit directions we use.

DEFINITION 3.2. *A subsequence of the MADS iterates consisting of minimal frame centers, $\{x_k\}_{k\in K}$ for some subset of indices K, is said to be a* refining subsequence *if $\{\Delta_k^p\}_{k\in K}$ converges to zero.*

Let $\hat{x}$ be the limit of a convergent refining subsequence. If the limit $\lim_{k\in L} \frac{d_k}{\|d_k\|}$ exists for some subset $L \subseteq K$ with poll direction $d_k \in D_k$, and if $x_k + \Delta_k^m d_k \in \Omega$ for infinitely many $k \in L$, then this limit is said to be a refining direction for $\hat{x}$.

It is shown in [3], that there exists at least one convergent refining subsequence. We now present some definitions that will be used later to guarantee the existence of refining directions.

3.2. Three types of tangent cones. Three different types of tangent cones play a central role in our analysis. Their definition, and equivalent ones, may be found in [29, 10, 18]. After presenting them, we supply an example where the three cones differ to illustrate some of our results. The first cone that we present is the hypertangent cone.

DEFINITION 3.3 (Hypertangent cone). *A vector $v \in \mathbb{R}^n$ is said to be a hypertangent vector to the set $\Omega \subseteq \mathbb{R}^n$ at the point $x \in \Omega$ if there exists a scalar $\epsilon > 0$ such that*

$$y + tw \in \Omega \qquad \text{for all } y \in \Omega \cap B_\epsilon(x), \quad w \in B_\epsilon(v) \quad \text{and} \quad 0 < t < \epsilon. \tag{3.1}$$

The set of hypertangent vectors to Ω at x is called the hypertangent cone to Ω at x *and is denoted by $T_\Omega^H(x)$.*

The hypertangent cone is a useful concept for understanding the behavior of the MADS algorithm. When analyzing MADS, we will be concerned with the following specific subsequences:

- minimal frame centers $x_k \to \hat{x}$;
- mesh size parameters $\Delta_k^m \searrow 0$ and step sizes $\Delta_k^m \|d_k\| \searrow 0$;
- normalized refining directions $\frac{d_k}{\|d_k\|} \to v \neq 0$.

These subsequences will be chosen in a way so that $x_k \in \Omega$ and $x_k + (\Delta_k^m \|d_k\|)\frac{d_k}{\|d_k\|} \in \Omega$. The connection with the hypertangent definition is obvious by noticing that the roles of y, t, and w are played by $x_k, \Delta_k^m\|d_k\|$, and $\frac{d_k}{\|d_k\|}$, respectively. The connection with the Clarke derivative (1.3) will be made explicit in Theorem 3.12.

Since the definition of a hypertangent is rather technical and crucial to our results, we will pause for a short discussion. The reader could easily show that if Ω is a full dimensional polytope defined by linear constraints, then every direction from a point $\hat{x} \in \Omega$ into the interior of Ω is a hypertangent. That follows immediately from the following result relating hypertangents to the constraint qualification suggested by Gould and Tolle [16]; see also [6] for a discussion of the Gould and Tolle constraint qualification and the closely related one of Mangasarian and Fromovitz.

THEOREM 3.4. *Let $C : \mathbb{R}^n \to \mathbb{R}^m$ be continuously differentiable at a point $\hat{x} \in \Omega = \{x \in \mathbb{R}^n \ : \ C(x) \leq 0\}$, and let $\mathcal{A}(\hat{x}) = \{i \in \{1, 2, \ldots, m\} : c_i(\hat{x}) = 0\}$ be the*

active set at $\hat{x}$. If $v \in \mathbb{R}^n$ is a hypertangent vector to Ω at $\hat{x}$, then $\nabla c_i(\hat{x})^T v < 0$ for each $i \in \mathcal{A}(\hat{x})$ such that $\nabla c_i(\hat{x}) \neq 0$. Furthermore, if $\nabla c_i(\hat{x})^T v < 0$ for each $i \in \mathcal{A}(\hat{x})$, then $v \in \mathbb{R}^n$ is a hypertangent vector to Ω at $\hat{x}$.

Proof. Let v be a hypertangent vector to Ω at $\hat{x}$. Then, there exists an $\epsilon > 0$ such that $\hat{x} + tv \in \Omega$ for any $0 < t < \epsilon$. Let $i \in \mathcal{A}(\hat{x})$. Continuous differentiability of c_i at $\hat{x}$ implies that

$$\nabla c_i(\hat{x})^T v = \lim_{t \to 0} \frac{c_i(\hat{x} + tv) - c_i(\hat{x})}{t} \leq 0.$$

It only remains to show that $\nabla c_i(\hat{x})^T v \neq 0$ when $\nabla c_i(\hat{x}) \neq 0$. Suppose by way of contradiction that $\nabla c_i(\hat{x})^T v = 0$ and $\nabla c_i(\hat{x}) \neq 0$. Since the hypertangent cone is an open set [29], for any nonnegative $\delta \in \mathbb{R}$ sufficiently small, $v + \delta \nabla c_i(\hat{x})$ is a hypertangent vector to Ω at $\hat{x}$. It follows that

$$0 \ \geq \ \nabla c_i(\hat{x})^T (v + \delta \nabla c_i(\hat{x})) \ = \ \delta \|\nabla c_i(\hat{x})\|_2^2 \ > \ 0,$$

which is a contradiction. Thus, $\nabla c_i(\hat{x})^T v < 0$ when $\nabla c_i(\hat{x}) \neq 0$.

To prove the converse, let $i \in \mathcal{A}(\hat{x})$ be such that $\nabla c_i(\hat{x}) \neq 0$ and $v \in \mathbb{R}^n$ be such that $\|v\| = 1$ and $\nabla c_i(\hat{x})^T v < 0$. The product $\nabla c_i(y)^T w$ is a continuous function at $(y;\ w) = (\hat{x};\ v)$, and so there is some $\epsilon_1 > 0$ such that

$$\nabla c_i(y)^T w < 0 \ \text{ for all } \ y \in B_{\epsilon_1}(\hat{x}) \text{ and } w \in B_{\epsilon_1}(v). \tag{3.2}$$

Take $\epsilon = \min\{1, \frac{\epsilon_1}{3}\}$ and let y, w be in $B_\epsilon(\hat{x})$ and $B_\epsilon(v)$, respectively, with $y \in \Omega$, and let $0 < t < \epsilon$. We will show that $y + tw \in \Omega$. Our construction ensures that $c_i(y) \leq 0$ and $\epsilon < \epsilon_1$, and so by the mean value theorem, we have

$$c_i(y + tw) \leq c_i(y + tw) - c_i(y) = \nabla c_i(y + \theta tw)^T (tw) \ \text{ for some } \theta \in [0, 1]. \tag{3.3}$$

But, $\|y + \theta tw - \hat{x}\| \leq \|y - \hat{x}\| + \theta t(\|w - v\| + \|v\|) < \epsilon + \epsilon(\epsilon + 1) \leq 3\epsilon \leq \epsilon_1$, thus $y + \theta tw \in B_{\epsilon_1}(\hat{x})$, and $w \in B_\epsilon(v) \subseteq B_{\epsilon_1}(v)$. It follows that (3.2) applies and therefore $\nabla c_i(y + \theta tw)^T w < 0$. This is combined with (3.3) and with the fact that $t > 0$ implies that $c_i(y + tw) \leq 0$. But c_i was any active component function, and so $C(y + tw) \leq 0$ and thus $y + tw \in \Omega$. ∎

We would like to culminate our hierarchy of convergence results by providing necessary conditions to ensure contingent stationarity. In order to do so, we present two other types of tangent cones.

DEFINITION 3.5 (Clarke tangent cone). *A vector $v \in \mathbb{R}^n$ is said to be a Clarke tangent vector to the set $\Omega \subseteq \mathbb{R}^n$ at the point x in the closure of Ω if for every sequence $\{y_k\}$ of elements of Ω that converges to x and for every sequence of positive real numbers $\{t_k\}$ converging to zero, there exists a sequence of vectors $\{w_k\}$ converging to v such that $y_k + t_k w_k \in \Omega$. The set $T_\Omega^{Cl}(x)$ of all Clarke tangent vectors to Ω at x is called the Clarke tangent cone to Ω at x.*

DEFINITION 3.6 (Contingent cone). *A vector $v \in \mathbb{R}^n$ is said to be a tangent vector to the set $\Omega \subseteq \mathbb{R}^n$ at the point x in the closure of Ω if there exists a sequence $\{y_k\}$ of elements of Ω that converges to x and a sequence of positive real numbers $\{\lambda_k\}$ for which $v = \lim_k \lambda_k(y_k - x)$. The set $T_\Omega^{Co}(x)$ of all tangent vectors to Ω at x is called the contingent cone (or sequential Bouligand tangent cone) to Ω at x.*

DEFINITION 3.7. *The set Ω is said to be regular at x provided $T_\Omega^{Cl}(x) = T_\Omega^{Co}(x)$.*

Any convex set is regular at each of its points [10]. Both $T_\Omega^{Co}(x)$ and $T_\Omega^{Cl}(x)$ are closed cones, and both $T_\Omega^{Cl}(x)$ and $T_\Omega^{H}(x)$ are convex cones. Moreover, $T_\Omega^{H}(x) \subseteq T_\Omega^{Cl}(x) \subseteq T_\Omega^{Co}(x)$. Rockafellar [29] showed that $T_\Omega^{H}(x) = int(T_\Omega^{Cl}(x))$ whenever $T_\Omega^{H}(x)$ is nonempty. Moreover, since the closure of the interior of a closed convex set is the set itself [28], it follows that $T_\Omega^{Cl}(x) = \mathrm{cl}(T_\Omega^{H}(x))$ whenever $T_\Omega^{H}(x)$ is nonempty.

3.3. Generalized derivatives. Recall that we are using Jahn's definition (1.3) of the Clarke derivative instead of (1.1), and therefore we cannot directly use the calculus theory developed in [10]. The next lemma and proposition extend previously known calculus results in the unconstrained case.

LEMMA 3.8. *Let f be Lipschitz near $\hat{x} \in \Omega$ with Lipschitz constant λ. If u and v belong to $T_\Omega^{H}(\hat{x})$, then*

$$f^\circ(\hat{x}; u) \geq f^\circ(\hat{x}; v) - \lambda\|u - v\|.$$

Proof. Let f be Lipschitz near $\hat{x} \in \Omega$ with Lipschitz constant λ and let u and v belong to $T_\Omega^{H}(\hat{x})$. Let $\epsilon > 0$ be such that $y + tw \in \Omega$ whenever $y \in \Omega \cap B_\epsilon(\hat{x})$, $w \in B_\epsilon(u) \cup B_\epsilon(v)$ and $0 < t < \epsilon$. This can be done by taking ϵ to be the smaller of the values for u and v guaranteed by the definition of a hypertangent. In particular, if $y \in \Omega \cap B_\epsilon(\hat{x})$ and if $0 < t < \epsilon$, then both $y + tu$ and $y + tv$ belong to Ω. This allows us to go from the first to the second equality of the following chain:

$$\begin{aligned}
f^\circ(\hat{x}; u) &= \limsup_{\substack{y \to \hat{x},\ y \in \Omega \\ t \downarrow 0,\ y + tu \in \Omega}} \frac{f(y+tu) - f(y)}{t} \\
&= \limsup_{\substack{y \to \hat{x},\ y \in \Omega \\ t \downarrow 0,\ y + tv \in \Omega}} \frac{f(y+tu) - f(y)}{t} \\
&= \limsup_{\substack{y \to \hat{x},\ y \in \Omega \\ t \downarrow 0,\ y + tv \in \Omega}} \frac{f(y+tv) - f(y)}{t} + \frac{f(y+tu) - f(y+tv)}{t} \\
&= f^\circ(\hat{x}; v) + \limsup_{\substack{y \to \hat{x},\ y \in \Omega \\ t \downarrow 0,\ y + tv \in \Omega}} \frac{f(y+tu) - f(y+tv)}{t} \geq f^\circ(\hat{x}; v) - \lambda\|u - v\|. \qquad \square
\end{aligned}$$

Based on the previous lemma, the next proposition shows that the Clarke generalized derivative is continuous with respect to v on the Clarke tangent cone. The result is necessary to the proofs of Theorems 3.12 and 3.13.

PROPOSITION 3.9. *Let f be Lipschitz near $\hat{x} \in \Omega$. If $T_\Omega^{H}(\hat{x}) \neq \emptyset$ and if $v \in T_\Omega^{Cl}(\hat{x})$ then*

$$f^\circ(\hat{x}; v) = \lim_{\substack{w \to v, \\ w \in T_\Omega^{H}(\hat{x})}} f^\circ(\hat{x}; w).$$

Proof. Let λ be a Lipschitz constant for f near $\hat{x} \in \Omega$ and let $\{w_k\} \subset T_\Omega^{H}(\hat{x})$ be a sequence of directions converging to a vector $v \in T_\Omega^{Cl}(\hat{x})$. By definition of the hypertangent cone, let $0 < \epsilon_k < \frac{1}{k}$ be such that

$$y + tw \in \Omega \text{ whenever } y \in \Omega \cap B_{\epsilon_k}(\hat{x}), w \in B_{\epsilon_k}(w_k) \text{ and } 0 < t < \epsilon_k. \tag{3.4}$$

We first show the inequality $f^\circ(\hat{x}; v) \leq \lim_k f^\circ(\hat{x}; w_k)$. Equation (3.4) implies that

$$
\begin{aligned}
f^\circ(\hat{x};v) = & \limsup_{\substack{y \to \hat{x},\ y \in \Omega \\ t \downarrow 0,\ y+tv \in \Omega}} \frac{f(y+tv)-f(y)}{t} \\
= & \limsup_{\substack{y \to \hat{x},\ y \in \Omega \\ t \downarrow 0,\ y+tv \in \Omega \\ y+tw_k \in \Omega}} \frac{f(y+tv)-f(y)}{t} \\
\leq & \limsup_{\substack{y \to \hat{x},\ y \in \Omega \\ t \downarrow 0,\ y+tw_k \in \Omega}} \frac{f(y+tw_k)-f(y)}{t} - \frac{f(y+tw_k)-f(y+tv)}{t} \\
= & \ f^\circ(\hat{x};w_k) \ + \limsup_{\substack{y \to \hat{x},\ y \in \Omega \\ t \downarrow 0,\ y+tw_k \in \Omega}} \frac{f(y+tw_k)-f(y+tv)}{t}.
\end{aligned}
$$

As k goes to infinity, $\left|\frac{f(y+tw_k)-f(y+tv)}{t}\right| \leq \lambda\|w_k - v\|$ goes to zero. Since $\{w_k\}$ was arbitrary in the hypertangent cone, it follows that

$$
f^\circ(\hat{x};v) \leq \lim_{\substack{w \to v, \\ w \in T_\Omega^H(\hat{x})}} f^\circ(\hat{x};w).
$$

Second, we show the reverse inequality: $f^\circ(\hat{x};v) \geq \lim_k f^\circ(\hat{x};w_k)$. Let us define $u_k = \frac{1}{k}w_k + (1-\frac{1}{k})v = w_k + (1-\frac{1}{k})(v-w_k)$. Since the hypertangent cone is a convex set, and since v lies in the closure of the hypertangent cone, it then follows that $u_k \in T_\Omega^H(\hat{x})$ for every $k = 1, 2, \ldots$

We now consider the generalized directional derivative

$$
f^\circ(\hat{x};u_k) = \limsup_{\substack{y \to \hat{x},\ y \in \Omega \\ t \downarrow 0,\ y+tu_k \in \Omega}} \frac{f(y+tu_k)-f(y)}{t}.
$$

The fact that $u_k \in T_\Omega^H(\hat{x})$ implies that there exists $y_k \in \Omega \cap B_{\epsilon_k}(\hat{x})$ and $0 < \frac{t_k}{k} < \epsilon_k$ such that $y_k + t_k u_k \in \Omega$ and

$$
\frac{f(y_k + t_k u_k) - f(y_k)}{t_k} \geq f^\circ(\hat{x};u_k) - \epsilon_k, \tag{3.5}
$$

where ϵ_k is the constant from (3.4). We now define the sequence $z_k = y_k + \frac{t_k}{k}w_k \in \Omega$ converging to $\hat{x}$, and the sequence of scalars $h_k = (1-\frac{1}{k})t_k > 0$ converging to zero. Notice that

$$
z_k + h_k v = y_k + t_k\left(\frac{1}{k}w_k + \left(1-\frac{1}{k}\right)v\right) = y_k + t_k u_k \ \in \ \Omega,
$$

and therefore

$$
\begin{aligned}
f^\circ(\hat{x};v) = & \limsup_{\substack{z \to \hat{x},\ z \in \Omega \\ h \downarrow 0,\ z+hv \in \Omega}} \frac{f(z+hv)-f(z)}{t} \\
& \geq \lim_k \frac{f(z_k+h_k v)-f(z_k)}{h_k} \\
& = \lim_k \frac{f(y_k+t_k u_k)-f(y_k)}{(1-\frac{1}{k})t_k} + \frac{f(y_k)-f(y_k+\frac{t_k}{k}w_k)}{(1-\frac{1}{k})t_k} \\
\text{by (3.5):} \quad & \geq \lim_k f^\circ(\hat{x};u_k) - \epsilon_k + \frac{f(y_k)-f(y_k+\frac{t_k}{k}w_k)}{(1-\frac{1}{k})t_k} \\
\text{by Lemma 3.8:} \quad & \geq \lim_k f^\circ(\hat{x};w_k) - \lambda\|u_k - w_k\| - \epsilon_k - \frac{\frac{\lambda}{k}\|w_k\|}{(1-\frac{1}{k})} \\
& = \lim_k f^\circ(\hat{x};w_k) - \lambda\|v - w_k\| - \tfrac{\lambda}{k}\|v\| \ = \ \lim_k f^\circ(\hat{x};w_k). \qquad \square
\end{aligned}
$$

Unfortunately, the above proposition is not necessarily true when the hypertangent cone is empty: $f^\circ(\hat{x};v)$ may differ from $\lim_{w\to v} f^\circ(\hat{x};w)$. The above proof breaks as we cannot show in (3.4) that $y+tw_k$ belongs to Ω when $y\in\Omega$ is close to $\hat{x}$ and when $t>0$ is small. The following example in $\mathbb{R}^2$ illustrates that in this case, the Clarke generalized derivative is not necessarily upper semicontinuous on the contingent cone.

Example 3.10. Consider a feasible region $\Omega\subset\mathbb{R}^2$ that is the union of

$$\Omega_1=\{(a,b)^T : a\geq 0,\ b\geq 0\} \ \text{ with } \ \Omega_2=\{(-a,b)^T : b=-a^2,\ a\geq 0\}.$$

One can verify that at the origin

$$T_\Omega^H(0)=\emptyset,\ \ T_\Omega^{Cl}(0)=\{(a,0)^T : a\geq 0\}\subset\Omega_1 \ \text{ and } \ T_\Omega^{Co}(0)=\Omega_1\cup\{(-a,0)^T : a\geq 0\},$$

and therefore Ω is not regular at the origin.

Consider the continuous concave function in $\mathbb{R}^2$: $f(a,b)=-\max\{0,a\}$. Notice that $f(a,b)=0$ for $(a,b)^T\in\Omega_2$, and $f(a,b)=-a\leq 0$ on Ω_1. We will show that $f^\circ(0;w)$ is nonnegative for w in the interior of the contingent cone but $f^\circ(0;e_1)=-1$ with $e_1=(1,0)^T$ in the Clarke tangent cone.

Let $w=(w_1,w_2)^T$ be any direction in $int(T_\Omega^{Co}(0))=int(\Omega_1)$. We will construct appropriate subsequences in order to compute a valid lower bound on $f^\circ(0;w)$. For every positive integer k, define

$$y_k=\left(\frac{-w_1}{k},\ \frac{-w_1^2}{k^2}\right)^T \ \text{ and } \ t_k=\frac{1}{k}.$$

One can easily check that $y_k\in\Omega_2\subset\Omega$, and hence $f(y_k)=0$ for every k. Also, for every $k>\frac{w_1^2}{w_2}$, we have $y_k+t_kw=\left(0,\ \frac{1}{k^2}(kw_2-w_1^2)\right)^T\in\Omega_1\subset\Omega$ is on the nonnegative b axis. It follows that $f(y_k+t_kw)=0$ for every such k, and so

$$f^\circ(0;w)\geq\lim_{k\to\infty}\tfrac{f(y_k+t_kw)-f(y_k)}{t_k} \ = \ \lim_{k\to\infty} k\cdot(0-0) \ = \ 0.$$

In particular, taking $w=(1,\epsilon)$, we have that $f^\circ(0;(1,\epsilon)^T)$ is nonnegative for any $\epsilon>0$.

However, let us compute the Clarke generalized directional derivative $f^\circ(0;e_1)$ at the origin in the direction $e_1=(1,0)^T\in T_\Omega^{Cl}(0)$. The origin cannot be approached by points $y_k=(a_k,b_k)^T\in\Omega$ with the properties that $b_k<0$, and $y_k+t_ke_1\in\Omega$ with $t_k>0$. This is easy to see from a picture because y_k would have to be in Ω_2, and then $y_k+t_ke_1$ cannot possibly be in Ω. A necessary condition for both sequences to be in Ω is that y_k belongs to Ω_1, where $f(a,b)=-a$. But then every difference quotient in the definition of $f^\circ(0;e_1)$ is -1, and therefore $f^\circ(0;e_1)=-1$.

This example shows that when the hypertangent cone at $\hat{x}$ is empty, the Clarke tangent cone is not. It is possible that $f^\circ(\hat{x};w)$ is nonnegative for every w in the interior of the contingent cone and drops discontinuously to a negative value on the boundary of the contingent cone: $f^\circ(\hat{x};e_1)<\limsup_{w\to e_1} f^\circ(\hat{x};w)$.

3.4. A hierarchy of convergence results for MADS. We now present different necessary optimality conditions based on the tangent cone definitions.

DEFINITION 3.11. *Let f be Lipschitz near $\hat{x}\in\Omega$. Then, $\hat{x}$ is said to be a Clarke or contingent stationary point of f over Ω, if $f^\circ(\hat{x};v)\geq 0$ for every direction v in the Clarke or contingent cone of Ω at $\hat{x}$, respectively.*

In addition, $\hat{x}$ is said to be a Clarke or contingent KKT stationary point of f over Ω, if $-\nabla f(\hat{x})$ exists and belongs to the polar of the Clarke or contingent cone of Ω at $\hat{x}$, respectively.

This leads to our basic result on refining directions from which all our hierarchy of results are derived. The proof of these results also illustrates the close connection between the MADS framework, the Clarke calculus, and the definition of a hypertangent vector.

THEOREM 3.12. *Let f be Lipschitz near a limit $\hat{x} \in \Omega$ of a refining subsequence, and let $v \in T_\Omega^H(\hat{x})$ be a refining direction for $\hat{x}$. Then the generalized directional derivative of f at $\hat{x}$ in the direction v is nonnegative,* i.e., *$f^\circ(\hat{x}; v) \geq 0$.*

Proof. Let $\{x_k\}_{k \in K}$ be a refining subsequence converging to $\hat{x}$ and $v = \lim_{k \in L} \frac{d_k}{\|d_k\|} \in T_\Omega^H(\hat{x})$ be a refining direction for $\hat{x}$, with $d_k \in D_k$ for every $k \in L$. Since f is Lipschitz near $\hat{x}$, Proposition 3.9 ensures that $f^\circ(\hat{x}; v) = \lim_{k \in L} f^\circ(\hat{x}; \frac{d_k}{\|d_k\|})$. But, for any $k \in L$, one can apply the definition of the Clarke generalized derivative with the roles of y and t played by x_k and $\Delta_k^m \|d_k\|$, respectively. Note that this last quantity indeed converges to zero since Definition 2.2 ensures that it is bounded above by $\Delta_k^p \max\{\|d'\| : d' \in D\}$, where D is a finite set of directions, and (2.2) states that Δ_k^p goes to zero. Therefore

$$\begin{aligned} f^\circ(\hat{x}; v) &\geq \limsup_{k \in L} \frac{f(x_k + \Delta_k^m \|d_k\| \frac{d_k}{\|d_k\|}) - f(x_k)}{\Delta_k^m \|d_k\|} \\ &= \limsup_{k \in L} \frac{f(x_k + \Delta_k^m d_k) - f(x_k)}{\Delta_k^m \|d_k\|} \;\geq\; 0. \end{aligned}$$

The last inequality follows from the fact that for each sufficiently large $k \in L$, $x_k + \Delta_k^m d_k \in \Omega$ and $f(x_k + \Delta_k^m d_k) = f_\Omega(x_k + \Delta_k^m d_k)$ was evaluated and compared by the algorithm to $f(x_k)$, but x_k is a minimal frame center, so the inequality holds. ∎

We now show that Clarke directional derivatives of f at the limit $\hat{x}$ of minimal frame centers, for meshes that get infinitely fine, are nonnegative for all directions in the hypertangent cone, i.e., we show that MADS generates a Clarke stationary point.

THEOREM 3.13. *Let f be Lipschitz near a limit $\hat{x} \in \Omega$ of a refining subsequence, and assume that $T_\Omega^H(\hat{x}) \neq \emptyset$. If the set of refining directions for $\hat{x}$ is dense in $T_\Omega^H(\hat{x})$, then $\hat{x}$ is a Clarke stationary point of f on Ω.*

Proof. The proof follows directly from Theorem 3.12 and Proposition 3.9. ∎

Note that even though the algorithm is applied to f_Ω instead of f, the convergence results are linked to the local smoothness of f and not f_Ω, which is obviously discontinuous on the boundary of Ω. This is because we use (1.3) as the definition of the Clarke generalized derivative instead of (1.1). The constraint qualification used in these results is that the hypertangent cone is nonempty at the feasible limit point $\hat{x}$. Further discussion of nonempty hypertangent cones is found in Rockafellar [29].

A corollary to this last theorem is that if f is strictly differentiable at $\hat{x}$, then it is a Clarke KKT point.

COROLLARY 3.14. *Let f be strictly differentiable at a limit $\hat{x} \in \Omega$ of a refining subsequence, and assume that $T_\Omega^H(\hat{x}) \neq \emptyset$. If the set of refining directions for $\hat{x}$ is dense in $T_\Omega^H(\hat{x})$, then $\hat{x}$ is a Clarke KKT stationary point of f over Ω.*

Proof. Strict differentiability ensures that the gradient $\nabla f(\hat{x})$ exists and that $\nabla f(\hat{x})^T v = f^\circ(\hat{x}; v)$ for all directions. It follows directly from the previous proposition that $-\nabla f(\hat{x})^T v \leq 0$ for every direction v in $T_\Omega^{Cl}(\hat{x})$, thus $\hat{x}$ is a Clarke KKT stationary point. ∎

Our next two results are based on the definition of set regularity (see Definition 3.7).

PROPOSITION 3.15. *Let f be Lipschitz near a limit $\hat{x} \in \Omega$ of a refining subsequence, and assume that $T_\Omega^H(\hat{x}) \neq \emptyset$. If the set of refining directions for $\hat{x}$ is dense in $T_\Omega^H(\hat{x})$, and if Ω is regular at $\hat{x}$, then $\hat{x}$ is a contingent stationary point of f over Ω.*

Proof. The definition of regularity of the set Ω ensures that $f^\circ(\hat{x}; w) \geq 0$ for all w in $T_\Omega^{Co}(\hat{x})$. ∎

The following result is the counterpart to Corollary 3.14 for contingent stationarity. The proof is omitted since it is essentially the same.

COROLLARY 3.16. *Let f be strictly differentiable at a limit $\hat{x} \in \Omega$ of a refining subsequence, and assume that $T_\Omega^H(\hat{x}) \neq \emptyset$. If the set of refining directions for $\hat{x}$ is dense in $T_\Omega^H(\hat{x})$, and if Ω is regular at $\hat{x}$, then $\hat{x}$ is a contingent KKT stationary point of f over Ω.*

Example F in [2] presents an instance of a GPS algorithm such that when applied to a given unconstrained optimization problem, it generates a single limit point $\hat{x}$ which is not a Clarke stationary point. In fact, it is shown that f is differentiable but not strictly differentiable at $\hat{x}$ and $\nabla f(\hat{x})$ is nonzero. This unfortunate circumstance is due to the fact that GPS uses a finite number of poll directions while MADS can use infinitely many.

The following result shows that the algorithm ensures strong optimality conditions for unconstrained optimization, or when $\hat{x}$ is in the interior of Ω.

THEOREM 3.17. *Let f be Lipschitz near a limit $\hat{x}$ of a refining subsequence. If $\Omega = \mathbb{R}^n$, or if $\hat{x} \in int(\Omega)$, and if the set of refining directions for $\hat{x}$ is dense in $\mathbb{R}^n$, then $0 \in \partial f(\hat{x})$.*

Proof. Let $\hat{x}$ be as in the statement of the result, then $T_\Omega^H(\hat{x}) = \mathbb{R}^n$. Combining Definition 3.11 and Theorem 3.13 with (1.2) yields the result. ∎

Newton's method uses second derivatives, and the standard analysis of Newton's method assumes Lipschitz continuity of the second derivatives. Correspondingly, MADS is an algorithm that uses only function values, and we assume only that the function f is Lipschitz near $\hat{x}$.

In the general statement of the algorithm we did not present a strategy that would guarantee a dense set of refining directions in the hypertangent cone. We want to keep the algorithm framework as general as possible. There are different strategies that could be used to generate a dense set of poll directions. The selection of the set D_k could be done in a deterministic way or may use some randomness. In the remainder of the paper, we present, analyze, and test one MADS strategy that uses some randomness. We do this because we have not found a deterministic strategy that achieves a satisfying distribution of poll directions when the process is terminated after a reasonable number of iterations.

4. Practical implementation—LTMADS.

We now present two variants of a stochastic implementation of the MADS algorithm. We call either variant LTMADS, because of the underlying lower triangular basis construction, and we show that with probability 1, the set of poll directions generated by the algorithm is dense in the whole space, and in particular in the hypertangent cone.

4.1. Implementable instances of a MADS algorithm.

Let $G = I$, the identity matrix, and let $D = [I \ -I], \tau = 4, w^- = -1$ and $w^+ = 1$ be the fixed algorithmic parameters. Choose $\Delta_0^m = 1, \Delta_0^p = 1$ to be the initial mesh and poll size

parameters, and define the update rules as follows:

$$\Delta_{k+1}^m = \begin{cases} \frac{\Delta_k^m}{4} & \text{if } x_k \text{ is a minimal frame center} \\ 4\Delta_k^m & \text{if an improved mesh point is found, and if } \Delta_k^m \leq \frac{1}{4} \\ \Delta_k^m & \text{otherwise.} \end{cases}$$

A consequence of these rules is that the mesh size parameter is always a power of 4 and never exceeds 1. Thus, $\frac{1}{\sqrt{\Delta_k^m}} \geq 1$ is always a nonnegative power of 2 and hence integral.

We now present a strategy to randomly generate the poll directions. In what follows, every random generation is done uniformly with equal probabilities. In order to ensure that the set of refining directions is dense in the hypertangent cone, one of these directions must be selected in a different way. This direction must depend only on the value of the mesh size parameter, and not on the iteration number. The direction is denoted by $b(\ell)$ where ℓ is an integer related to the mesh size parameter. An additional counter, called ℓ_c, is initially set to zero and is used to keep track of the values of ℓ for which $b(\ell)$ was created. The construction of $b(\ell)$ is as follows.

GENERATION OF THE DIRECTION $b(\ell)$ FOR A GIVEN NONNEGATIVE INTEGER ℓ.

- VERIFICATION IF $b(\ell)$ WAS ALREADY CREATED:
  If $\ell_c > \ell$, then exit this procedure with the existing vector $b(\ell) \in \mathbb{Z}^n$.
  Otherwise, set $\ell_c \leftarrow \ell_c + 1$, and continue to the next step.
- INDEX OF ENTRY WITH LARGEST COMPONENT:
  Let $\hat{\imath}$ be an integer randomly chosen in the set $N = \{1, 2, \ldots, n\}$.
- CONSTRUCTION OF $b(\ell)$:
  Randomly set $b_{\hat{\imath}}(\ell)$ to either plus or minus 2^ℓ, and $b_i(\ell)$ for $i \in N \setminus \{\hat{\imath}\}$ to be an integer in $\{-2^\ell + 1, -2^\ell + 2, \ldots, 2^\ell - 1\}$. Record $b(\ell)$ and exit this procedure.

The above procedure returns a vector $b(\ell) \in \mathbb{Z}^n$ such that all elements but one are integers between $-2^\ell + 1$ and $2^\ell - 1$. The other element is either -2^ℓ or 2^ℓ. Moreover, when two iterations have the same mesh size parameter, then the corresponding vectors $b(\ell)$ are identical.

To each mesh size parameter Δ_k^m, we assign an integer $\ell = -\log_4(\Delta_k^m) \in \mathbb{N}$ so that $\Delta_k^m = 4^{-\ell}$. Note that the mesh size parameter in LTMADS takes the values $1, \frac{1}{4}, \frac{1}{16}, \ldots$, and therefore ℓ is necessarily a nonnegative integer.

We now present a procedure that extends $b(\ell)$ to a positive spanning set of either $2n$ or $n+1$ poll directions. The procedure first generates an $(n-1) \times (n-1)$ lower triangular matrix, and then combines it with $b(\ell)$ to create a basis in $\mathbb{R}^n$. Finally, this basis is extended to a positive basis by either mirroring the directions (for a maximal $2n$ basis), or by taking the negative sum of the directions (for a $n+1$ basis).

The rows of a lower triangular matrix L are randomly permuted, and a row of zeroes is inserted in position $\hat{\imath}$, where $\hat{\imath}$ is the index defined in the construction of the vector $b(\ell)$. This results in a $n \times (n-1)$ matrix. The column $b(\ell)$ is appended to it, and this leads to a basis B in $\mathbb{R}^n$. The permutation of the rows ensures that the zeroes of the triangular matrix are not mostly located in the upper part of B. Afterwards, the columns of B are randomly permuted to ensure that the zeroes are not mostly located in the right part of B'. This construction ensures that $|\det(B)| = |\det(B')| = 2^{\ell n}$. The completion to a positive basis D_k appends to B' either the negative sum of the columns of B', or the negative of each column.

GENERATION OF THE POSITIVE BASIS D_k AND UPDATE OF Δ_k^p.

- CONSTRUCTION OF THE DIRECTION $b(\ell)$ AND INDEX $\hat{\imath}$:
  Let $\ell = -\log_4(\Delta_k^m)$, and construct $b(\ell)$ by the above procedure.
  Set $\hat{\imath}$ to be the integer in N such that $|b_{\hat{\imath}}(\ell)| = 2^\ell$.
- BASIS CONSTRUCTION IN $\mathbb{R}^{n-1}$:
  Let L be a lower triangular $(n-1) \times (n-1)$ matrix where each term on the diagonal is either plus or minus 2^ℓ, and the lower components are randomly chosen in $\{-2^\ell + 1, -2^\ell + 2, \ldots, 2^\ell - 1\}$.
  L is a basis in $\mathbb{R}^{n-1}$ with $|\det(L)| = 2^{\ell(n-1)}$.
- PERMUTATION OF THE LINES OF L, AND COMPLETION TO A BASIS IN $\mathbb{R}^n$:
  Let $\{p_1, p_2, \ldots, p_{n-1}\}$ be random permutations of the set $N \setminus \{\hat{\imath}\}$. Set

$$\begin{array}{ll} B_{p_i,j} = L_{i,j} & \text{for } i, j = 1, 2, \ldots, n-1 \\ B_{\hat{\imath},j} = 0 & \text{for } j = 1, 2, \ldots, n-1 \\ B_{i,n} = b_i(\ell) & \text{for } i = 1, 2, \ldots, n. \end{array}$$

  B is a basis in $\mathbb{R}^n$ with $|\det(B)| = 2^{\ell n}$.
- PERMUTATION OF THE COLUMNS OF B:
  Let $\{q_1, q_2, \ldots, q_n\}$ be random permutations of the set N.
  Set $B'_{i,q_j} = B_{i,j}$ for each i and j in N. B' is a basis in $\mathbb{R}^n$ with $|\det(B')| = 2^{\ell n}$.
- COMPLETION TO A POSITIVE BASIS:
  - Minimal positive basis: Set $D_k = [B' \; d]$ with $d_i = -\sum_{j \in N} B'_{ij}$.
    Set the poll size parameter to $\Delta_k^p = n\sqrt{\Delta_k^m} \geq \Delta_k^m$.
  - Maximal positive basis: Set $D_k = [B' \; -B']$.
    Set the poll size parameter to $\Delta_k^p = \sqrt{\Delta_k^m} \geq \Delta_k^m$.

The construction also ensures that $b(\ell)$ is necessarily a column of the positive basis D_k. Our convergence analysis will show that as k goes to infinity, the union of all directions $b(\ell)$ is dense in $\mathbb{R}^n$ with probability 1. We will also show that if the entire sequence of iterates converges, then the set of refining directions is also dense in $\mathbb{R}^n$ with probability 1.

The following example in $\mathbb{R}^5$ highlights the features of the positive basis construction.

Example 4.1. Consider an iteration k with $\Delta_k^m = \frac{1}{16}$. The step CONSTRUCTION OF THE DIRECTION $b(\ell)$ AND INDEX $\hat{\imath}$ fixed $\ell = -\log_4(\Delta_k^m) = 2$. Suppose that the randomly defined vector $b(\ell)$ is $(-3, 2, 4, -1, 0)^T$. It follows that $\hat{\imath} = 3$ since $b_3(\ell) = 4$. Observe that all other components of $b(\ell)$ are integers between $-2^2 + 1$ and $2^2 - 1$.

Suppose that the step BASIS CONSTRUCTION IN $\mathbb{R}^{n-1}$ generates the random lower triangular matrix

$$L = \begin{bmatrix} -4 & 0 & 0 & 0 \\ 3 & 4 & 0 & 0 \\ -1 & 2 & -4 & 0 \\ 1 & -2 & 0 & 4 \end{bmatrix} \in \mathbb{Z}^{4\times 4}.$$

Now, if the two permutation steps generate the row permutation vector $(p_1, p_2, p_3, p_4) = (4, 1, 2, 5)$, and the column permutation vector $(q_1, q_2, q_3, q_4, q_5) = (5, 1, 3, 2, 4)$, then the bases constructed from L and $b(\ell)$ are

$$B = \begin{bmatrix} \mathbf{3} & \mathbf{4} & \mathbf{0} & \mathbf{0} & -3 \\ \mathbf{-1} & \mathbf{2} & \mathbf{-4} & \mathbf{0} & 2 \\ 0 & 0 & 0 & 0 & 4 \\ \mathbf{-4} & \mathbf{0} & \mathbf{0} & \mathbf{0} & -1 \\ \mathbf{1} & \mathbf{-2} & \mathbf{0} & \mathbf{4} & 0 \end{bmatrix} \quad \text{and} \quad B' = \begin{bmatrix} \mathbf{4} & \mathbf{0} & \mathbf{0} & -3 & \mathbf{3} \\ \mathbf{2} & \mathbf{0} & \mathbf{-4} & 2 & \mathbf{-1} \\ 0 & 0 & 0 & 4 & 0 \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & -1 & \mathbf{-4} \\ \mathbf{-2} & \mathbf{4} & \mathbf{0} & 0 & \mathbf{1} \end{bmatrix}$$

(the entries copied from L appear in boldface characters). One may easily verify that $|\det(B)| = |\det(B')| = 4^5$ and that the four terms B'_{p_i,q_i} for $i = 1, 2, 3, 4$ as well as B'_{3,q_5} are equal to either 4 or -4.

Finally, depending on if the minimal or maximal positive basis is selected, the COMPLETION TO A POSITIVE BASIS step generates the set D_k composed of the columns of either

$$\begin{bmatrix} 4 & 0 & 0 & -3 & 3 & -4 \\ 2 & 0 & -4 & 2 & -1 & 1 \\ 0 & 0 & 0 & 4 & 0 & -4 \\ 0 & 0 & 0 & -1 & -4 & 5 \\ -2 & 4 & 0 & 0 & 1 & -3 \end{bmatrix}$$

$$\text{or} \quad \begin{bmatrix} 4 & 0 & 0 & -3 & 3 & -4 & 0 & 0 & 3 & -3 \\ 2 & 0 & -4 & 2 & -1 & -2 & 0 & 4 & -2 & 1 \\ 0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 & -4 & 0 \\ 0 & 0 & 0 & -1 & -4 & 0 & 0 & 0 & 1 & 4 \\ -2 & 4 & 0 & 0 & 1 & 2 & -4 & 0 & 0 & -1 \end{bmatrix}.$$

A key point of this construction is that any iteration with a mesh size parameter equal to $\frac{1}{16}$ will have $b(\ell)$ as the q_{5th} column of D_k. In this particular example, $b(\ell)$ is the fourth column of D_k. The other columns will usually differ from one iteration to another.

Since MADS is allowed to be *opportunistic* and end a POLL step as soon as a better point is found, we want to randomize the POLL directions. Thus, the purpose of the second step is to permute the rows of the matrix B so that the zeroes in the upper triangular part of the matrix are randomly positioned, and to permute the columns so that the dense column is not always the first in D_k. The name LTMADS is based on the lower triangular matrix at the heart of the construction of the frames.

The following result shows that the frames generated by the LTMADS algorithm satisfy the conditions of Definition 2.2.

PROPOSITION 4.2. *At each iteration k, the procedure above yields a D_k and a MADS frame P_k such that*

$$P_k = \{x_k + \Delta_k^m d : d \in D_k\} \subset M_k,$$

where M_k is given by Definition 2.1 *and D_k is a positive spanning set such that for each $d \in D_k$,*

- *d can be written as a nonnegative integer combination of the directions in D: $d = Du$ for some vector $u \in \mathbb{N}^{n_D}$ that may depend on the iteration number k;*
- *the distance from the frame center x_k to a frame point $x_k + \Delta_k^m d \in P_k$ is bounded above by a constant times the poll size parameter: $\Delta_k^m \|d\| \leq \Delta_k^p \max\{\|d'\| : d' \in D\}$;*

- *limits (as defined in Coope and Price* [11]*) of the normalized sets $\mathcal{D}_k$ are positive spanning sets.*

Proof. The first n columns of D_k form a basis of $\mathbb{R}^n$ because they are obtained by permuting rows and columns of the lower triangular matrix B, which is nonsingular because it has nonzero terms on the diagonal. Moreover, taking the last direction to be the negative of the sum of the others leads to a minimal positive basis, and combining the first n columns of D_k with their negatives gives a maximal positive basis [14].

Again by construction, D_k has all integral entries in the interval $[-2^\ell, 2^\ell]$ $\left(\text{with } 2^\ell = \frac{1}{\sqrt{\Delta_k^m}}\right)$, and so clearly each column d of D_k can be written as a nonnegative integer combination of the columns of $D = [I, -I]$. Hence, the frame defined by D_k is on the mesh M_k.

Now the ℓ_∞ distance from the frame center to any frame point is $\|\Delta_k^m d\|_\infty = \Delta_k^m \|d\|_\infty$. There are two cases. If the maximal positive basis construction is used, then $\Delta_k^m \|d\|_\infty = \sqrt{\Delta_k^m} = \Delta_k^p$. If the minimal positive basis construction is used, then $\Delta_k^m \|d\|_\infty \le n\sqrt{\Delta_k^m} = \Delta_k^p$. The proof of the second bullet follows by noticing that $\max\{\|d'\|_\infty : d' \in [I \ -I]\} = 1$.

The frame can be rewritten in the equivalent form $\{x_k + \sqrt{\Delta_k^m} v : v \in \mathcal{V}\}$, where $\mathcal{V}$ is a set whose columns are the same as those of B after permutation and multiplication by $\sqrt{\Delta_k^m}$.

Coope and Price [11] show that a sufficient condition for the third bullet to hold is that each element of $\mathcal{V}$ is bounded above and below by positive constants that are independent of k. This is trivial to show with our construction. Indeed, each entry of $\mathcal{V}$ lies between -1 and 1 and every term on the diagonal is ± 1. B is a triangular matrix, and therefore $|det(\mathcal{V})| = 1$. ∎

The frames given in Figure 2.2 were generated using minimal positive bases with direction sets D_k: $\{(-1,0)^T, (0,-1)^T, (1,1)^T\}$, $\{(-2,-1)^T, (0,-2)^T, (2,3)^T\}$, and $\{(-3,4)^T, (4,0)^T, (-1,-4)^T\}$. One can see that as Δ_k^m and Δ_k^p go to zero, the number of candidates for frame points increases rapidly. For the three examples illustrated in the figure, the number of distinct possible frames that LTMADS may choose from is 4, 20, and 44, respectively (the frames D_k are interpreted as sets and not matrices). For example, in the case depicted in the rightmost figure, the $\hat{\imath}th$ row of the matrix B is $[0 \ \pm 4]$, the other row is $[\pm 4 \ \beta]$, where β is an integer between -3 and 3. It follows that for a given $\hat{\imath}$, there are $2 \times 2 \times 7 = 28$ possibilities for B. The index $\hat{\imath}$ is either 1 or 2, and thus the number of possibilities for B' is 56. Permuting the columns does not change the points in the frames. However, some different values of B' lead to the same frame D_k when viewed as a set. For examples, when B' is

$$\begin{bmatrix} 4 & -1 \\ 0 & -4 \end{bmatrix} \text{ or } \begin{bmatrix} -3 & 4 \\ 4 & 0 \end{bmatrix},$$

then in both cases the set of directions in the frame D_k is $\{(-3,4)^T, (4,0)^T, (-1,-4)^T\}$. This leads to a total of 44 different frames.

In addition to an opportunistic strategy, i.e., terminating a POLL step as soon as an improved mesh point is detected, a standard trick we use in GPS to improve the convergence speed consists of promoting a successful poll direction to the top of the list of directions for the next POLL step. We call this *dynamic ordering* of the polling directions. This strategy cannot be directly implemented in MADS since

at a successful iteration $k-1$, the poll size parameter is increased, and therefore a step of Δ_k^m in the successful direction will often be outside the mesh. The way we mimic GPS dynamic ordering in MADS is that when the previous iteration succeeded in finding an improved mesh point, we execute a simple one point *dynamic search* in the next iteration as follows. Suppose that $f_\Omega(x_k) < f_\Omega(x_{k-1})$ and that d is the direction for which $x_k = x_{k-1} + \Delta_{k-1}^m d$. Then, the trial point produced by the SEARCH step is $s_k = x_{k-1} + 4\Delta_{k-1}^m d$. Note that with this construction, if $\Delta_{k-1}^m < 1$, then $s_k = x_{k-1} + \Delta_k^m d$ and otherwise, $s_k = x_{k-1} + 4\Delta_k^m d$. In both cases s_k lies on the current mesh M_k. If this SEARCH finds a better point, then we go on to the next iteration, but if not, then we proceed to the POLL step. The reader will see in the numerical results below that this seems to be a good strategy.

4.2. Convergence analysis. The convergence results in section 3.4 are based on the assumption that the set of refining directions for the limit of a refining sequence is asymptotically dense in the hypertangent cone at that limit. The following result shows that the above instances of LTMADS generates an asymptotically dense set of poll directions with probability 1. Therefore, the convergence results based on the local smoothness of the objective function f and on the local topology of the feasible region Ω can be applied to LTMADS.

THEOREM 4.3. *Let $\hat{x} \in \Omega$ be the limit of a refining subsequence produced by either instance of LTMADS. Then the set of poll directions for the subsequence converging to $\hat{x}$ is asymptotically dense in $T_\Omega^H(\hat{x})$ with probability* 1.

Proof. Let $\hat{x}$ be the limit of a refining subsequence $\{x_k\}_{k\in K}$ produced by one of the above instances of LTMADS (either with the minimal or maximal positive basis). Consider the sequence of positive bases $\{D_k\}_{k\in K}$. Each one of these bases is generated independently.

We use the notation $P[E]$ to denote the probability that E occurs. Let v be a direction in $\mathbb{R}^n$ with $\|v\|_\infty = 1$ such that $P[|v_j| = 1] \geq \frac{1}{n}$ and $P[v_j = 1 \mid |v_j| = 1] = P[v_j = -1 \mid |v_j| = 1] = \frac{1}{2}$. We will find a lower bound on the probability that a normalized direction in D_k is arbitrarily close to the vector v.

Let k be an index of K, and let $\ell = -\log_4(\Delta_k^m)$. Recall that in the generation of the positive basis D_k, the column $b(\ell)$ is such that $|b_{\hat{\imath}}(\ell)| = 2^\ell$, and the other components of $b(\ell)$ are random integers between $-2^\ell + 1$ and $2^\ell - 1$. Set $u = \frac{b(\ell)}{\|b(\ell)\|_\infty}$. It follows by construction that $u = 2^{-\ell} b(\ell)$ and $\|u\|_\infty = |u_{\hat{\imath}}| = 1$. We will now show for any $0 < \epsilon < 1$, that the probability that $\|u - v\|_\infty < \epsilon$ is bounded below by some nonnegative number independent of k, as $k \in K$ goes to infinity. Let us estimate the probability that $|u_j - v_j| < \epsilon$ for each j. For $j = \hat{\imath}$ we have

$$\begin{aligned} P[|u_{\hat{\imath}} - v_{\hat{\imath}}| < \epsilon] &\geq P[u_{\hat{\imath}} = v_{\hat{\imath}} = 1] + P[u_{\hat{\imath}} = v_{\hat{\imath}} = -1] \\ &= P[u_{\hat{\imath}} = 1] \times P[v_{\hat{\imath}} = 1] + P[u_{\hat{\imath}} = -1] \times P[v_{\hat{\imath}} = -1] \\ &\geq \frac{1}{2} \times \frac{1}{2n} + \frac{1}{2} \times \frac{1}{2n} \;=\; \frac{1}{2n}. \end{aligned}$$

For $j \in N \setminus \{\hat{\imath}\}$ we have

$$P[|u_j - v_j| < \epsilon] = P[v_j - \epsilon < u_j < v_j + \epsilon] \;=\; P\left[2^\ell(v_j - \epsilon) < b_j(\ell) < 2^\ell(v_j + \epsilon)\right].$$

We will use the fact that the number of integers in the interval $[2^\ell(v_j - \epsilon), 2^\ell(v_j + \epsilon)] \cap [-2^\ell + 1, 2^\ell - 1]$ is bounded below by the value $2^\ell\epsilon - 1$. Now, since the bases D_k are independently generated, and since $b_j(\ell)$ is an integer randomly chosen with equal

probability among the $2^{\ell+1}-1$ integers in the interval $[-2^\ell+1, 2^\ell-1]$, then it follows that

$$P[|u_j - v_j| < \epsilon] \geq \frac{2^\ell\epsilon - 1}{2^{\ell+1}-1} \;>\; \frac{2^\ell\epsilon - 1}{2^{\ell+1}} \;=\; \frac{\epsilon - 2^{-\ell}}{2}.$$

Recall that $\hat{x}$ is the limit of a refining subsequence, and so there exists an integer α such that $\sqrt{\Delta_k^m} = 2^{-\ell} \leq \frac{\epsilon}{2}$ whenever $\alpha \leq k \in K$, and so

$$P[|u_j - v_j| < \epsilon] \geq \frac{\epsilon - \sqrt{\Delta_k^m}}{2} \;\geq\; \frac{\epsilon}{4} \qquad \text{for any } k \in K \text{ with } k \geq \alpha.$$

It follows that

$$P[\|u - v\|_\infty < \epsilon] \;=\; \prod_{j=1}^{n} P[|u_j - v_j| < \epsilon] \geq \frac{\left(\frac{\epsilon}{4}\right)^{n-1}}{2n} \qquad \text{for any } k \in K \text{ with } k \geq \alpha.$$

We have shown when k is sufficiently large that $P[\|u - v\|_\infty < \epsilon]$ is larger than a strictly positive constant which is independent of Δ_k^m. Thus, there will be a poll direction in D_k for some $k \in K$ arbitrarily close to any direction $v \in \mathbb{R}^n$, and in particular to any direction $v \in T_\Omega^H(\hat{x})$. ∎

The proof of the previous result shows that the set of directions consisting of the $b(\ell)$ directions over all iterations is dense in $\mathbb{R}^n$. Nevertheless, we require the algorithm to use a positive spanning set at each iteration instead of a single poll direction. This ensures that any limit of a refining subsequence is the limit of minimal frame centers on meshes that get infinitely fine. At this limit point, the set of refining directions is generated from the set of poll directions which is dense in LTMADS and finite in GPS. Therefore with both MADS and GPS, the set of directions for which the Clarke generalized derivatives are nonnegative positively span the whole space. However, GPS does not allow the possibility that the set of refining directions is dense, since it is finite.

Finally, we give a condition that ensures dense MADS refining directions with probability 1.

THEOREM 4.4. *Suppose that the entire sequence of iterates produced by either instance of LTMADS converges to $\hat{x} \in \Omega$. Then the set of refining directions for the entire sequence of iterates is asymptotically dense in $T_\Omega^H(\hat{x})$ with probability* 1.

Proof. Let K be the set of indices of iterations that are minimal frame centers. If the entire sequence of iterates produced by an instance of LTMADS converges to $\hat{x} \in \Omega$, then the subsequence $\{x_k\}_{k\in K}$ also converges to $\hat{x}$. Therefore, $\{b(\ell)\}_{\ell=1}^\infty$ is a subsequence of refining directions. This subsequence was shown in Theorem 4.3 to be asymptotically dense in $T_\Omega^H(\hat{x})$ with probability 1. ∎

5. Numerical results. We consider four test problems in this section. Each problem is intended to make a point about MADS. We give results for GPS with a POLL step only and with a simple Latin hypercube SEARCH step. The GPS results all use a POLL ordering we have found to be advantageous in our experience using GPS.

The first problem is unconstrained, but GPS is well known to stagnate on this problem if it is given an unsuitable set of directions. MADS has no problem converging quickly to a global optimizer. The second problem is a bound constrained chemical engineering problem where GPS is known to behave well enough to justify publication of the results [17]. Still, on the whole, MADS does better. The third is a simple

nonlinearly constrained problem where GPS and another filter version of GPS are both known to converge short of an optimizer. As the theory given here predicts, MADS has no difficulty. We also use this problem to show that MADS does well as the number of variables increases.

The last example is such that the feasible region gets narrow very quickly. This is meant to be a test for any derivative-free feasible point algorithm—like GPS or MADS with the extreme barrier approach to constraints. MADS does better than GPS with the filter or the barrier, both of which stagnate due to the limitation of finitely many POLL directions. MADS stops making progress when the mesh size gets smaller than the precision of the arithmetic.

Of course, even when one tries to choose carefully, four examples are not conclusive evidence. However, we believe that these numerical results coupled with the more powerful theory for MADS make a good case for MADS versus GPS. In addition, there is the evidence in [24] that MADS was effective on a problem in which each function evaluation took weeks to perform. In [5] MADS was used in a context of identifying optimal algorithmic trust-region parameters. In [7] MADS was used to optimize spent potliner treatment process in aluminium production. Furthermore, MADS has recently been added as an option in the MATLAB GADS toolbox [25].

5.1. An unconstrained problem where GPS does poorly. Consider the unconstrained optimization problem in $\mathbb{R}^2$ presented in [19] where GPS algorithms are known to converge to nonstationary points:

$$f(x) \;=\; \big(1-\exp(-\|x\|^2)\big)\times\max\{\|x-c\|^2,\|x-d\|^2\},$$

where $c=-d=(30,40)^T$. Figure 5.1 shows level sets of this function. It can be shown that f is locally Lipschitz and strictly differentiable at its global minimizer $(0,0)^T$.

The GPS and MADS runs are initiated at $x_0=(-2.1,1.7)^T$, depicted by a diamond in the right part of Figure 5.1. The gradient of f exists and is nonzero at that point, and therefore both GPS and MADS will move away from it. Since there is some randomness involved in the MADS instance described in section 4.1, we ran it a total of 5 times, to see how it compares to our standard NOMAD [13] implementation of GPS. Figure 5.2 shows a log plot of the progress of the objective function value for each set of runs. All POLL steps were opportunistic, and the runs were stopped when a minimal frame with poll size parameter less than 10^{-10} was detected. For GPS, the maximal $2n$ positive basis refers to the set of positive and negative coordinate directions, and the two minimal $n+1$ positive bases are $\{(1,0)^T,(0,1)^T,(-1,-1)^T\}$ and $\{(1,0)^T,(-0.5,0.866025)^T,(-0.5,-0.866025)^T\}$, and the termination criteria are the same, i.e., when Δ_k drops below 10^{-10}.

Without a search strategy, every GPS run converged to a point on the line $x_2=-\frac{3}{4}x_1$, where f is not differentiable. These three limit points are denoted by stars in Figure 5.1. As proved in [3], the limit points for GPS satisfy the necessary optimality condition that the Clarke generalized directional derivatives are nonnegative for D at these limit points, but they are not local optimizers. One can see by looking at the level sets of f that no descent directions can be generated by the GPS algorithm using the above directions.

However, when adding a search strategy (by randomly selecting $2n$ mesh points at each SEARCH step) or when using LTMADS, all runs eventually generated good directions and converged to the origin, the global optimal solution. Figure 5.2 suggests that the MADS convergence is faster than GPS. Also, even if randomness appears in

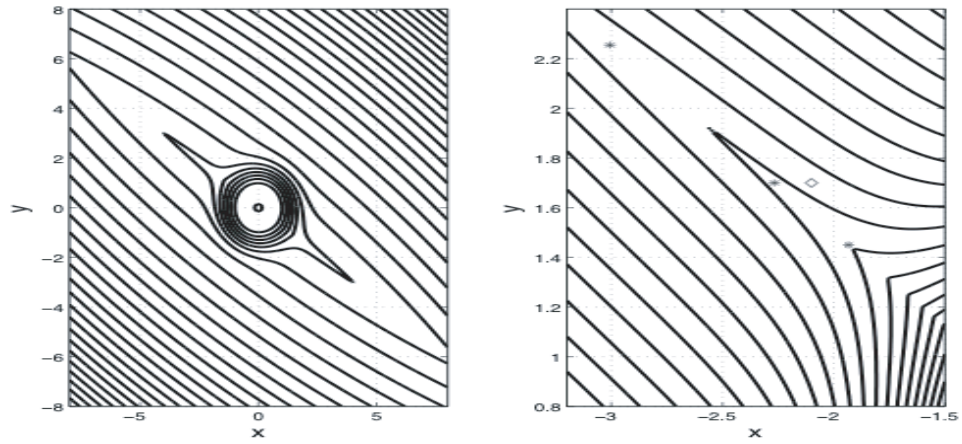


FIG. 5.1. *Level sets of* $f(x) = \left(1 - \exp(-\|x\|^2)\right) \times \max\{\|x - c\|^2, \|x - d\|^2\}$.

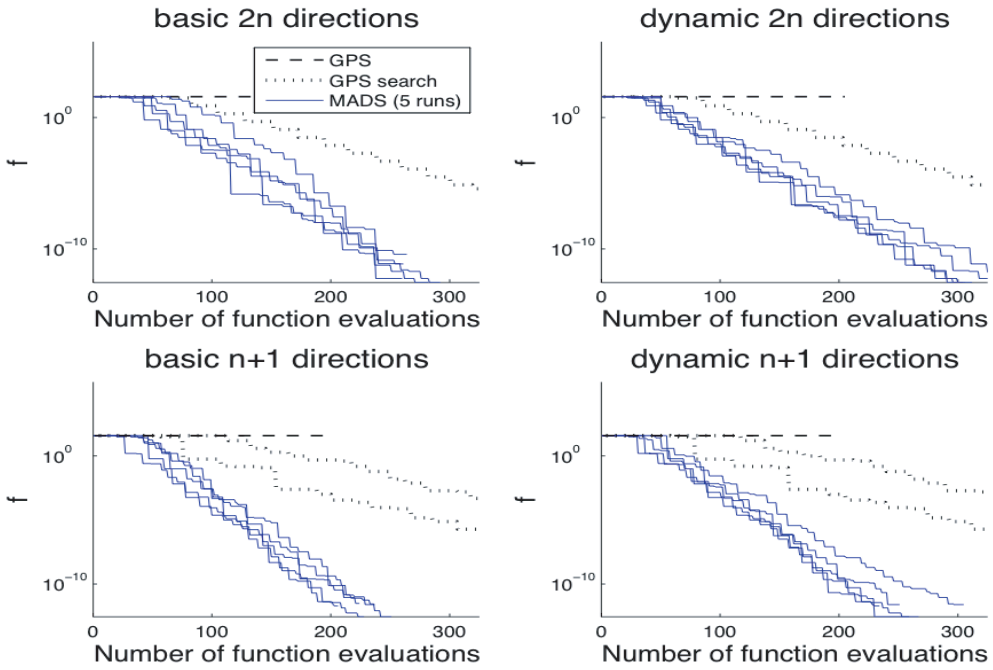


FIG. 5.2. *Progression of the objective function value vs the number of evaluations.*

these instances of LTMADS, the behavior of the algorithm is very stable in converging quickly to the origin.

5.2. A test problem where GPS does well. The academic example above was one of our motivations for developing MADS. We now apply MADS to an example from the chemical engineering literature for which GPS was shown to be preferable to a conjugate-direction approach. Hayes et al. [17] describe a method for evaluating the kinetic constants in a rate expression for catalytic combustion applications using experimental light-off curves. The method uses a transient one-dimensional single

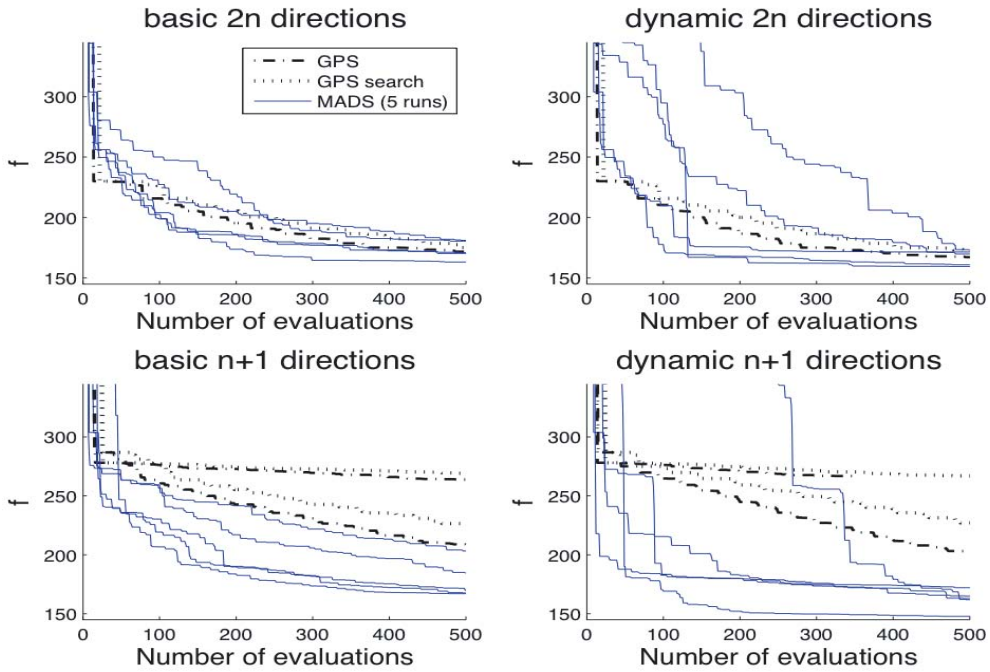


FIG. 5.3. *Data set* 1—*Progression of the objective function value vs the number of evaluations.*

channel monolith finite element reactor model to simulate reactor performance. The objective is to find the values of four parameters in a way such that the model estimates as closely as possible (in a weighted least square sense) an experimental conversion rate. This is a bound constrained nonsmooth optimization problem in $\mathbb{R}^4_+$, where the objective function measures the error between experimental data and values predicted by the model.

For the three sets of experimental data analyzed in [17], we compared the instances of GPS and MADS discussed above. The algorithms terminate whenever a minimal frame center with $\Delta_k^p \leq 2^{-6}$ (for MADS) or $\Delta_k \leq 2^{-6}$ (for GPS) is detected, or whenever 500 functions evaluations are performed, whichever comes first. Figures 5.3, 5.4, and 5.5 show the progression of the objective function value versus the number of evaluations for each data set.

The plots suggest that the objective function value decreases more steadily with GPS than with MADS. This is because GPS uses a fixed set of poll directions that we know to be an excellent choice for this problem. By allowing more directions, MADS eventually generates a steeper descent direction, and the dynamic runs capitalize on this by evaluating f further in that direction thus sharply reducing the objective function value in a few evaluations. In general, if the number of function evaluations is limited to a fixed number, then it appears that MADS with the dynamic strategy gives a better result than GPS.

For all three data sets, the dynamic runs are preferable to the basic runs. It also appears that for this problem, MADS runs with minimal $n+1$ directions perform better than the maximal $2n$ runs. GPS with a 2 point random search at each iteration systematically gave worst results than GPS without a search. In each of the three

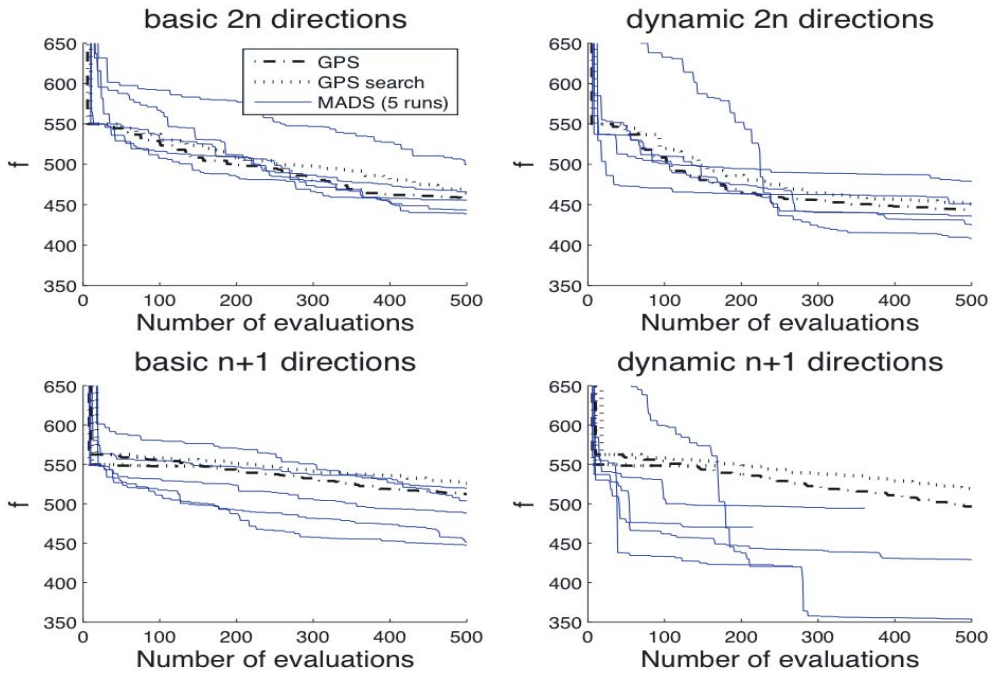


FIG. 5.4. *Data set* 2*—Progression of the objective function value vs the number of evaluations.*

data sets, the best overall solution was always produced by MADS with the dynamic $n+1$ directions.

The quality of the best solutions produced by GPS and MADS can be visualized in Figure 5.6 where the difference between the experimental and predicted conversions are plotted versus time. A perfect model with perfectly tuned parameters would have had a difference of zero everywhere. The superiority of the solution produced by MADS versus GPS is mostly visible for the second data set near the time 170 sec and the third data set near the time 190 sec where in both cases the fit is better by approximately 1%.

5.3. Linear optimization on an hypersphere. The third example shows again the difficulty caused by being restricted to a finite number of polling directions. It also illustrates the effect of dimension. This is a problem with a linear objective and strictly convex full-dimensional feasible region, surely the simplest nonlinearly constrained problem imaginable.

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & \sum_{i=1}^{n} x_i \\ \text{s.t.} \quad & \sum_{i=1}^{n} x_i^2 \le 3n. \end{aligned}$$

There is a single optimal solution to that problem: every component of the vector x is $-\sqrt{3}$ and the optimal value is $-\sqrt{3}n$.

The starting point is the origin, and the algorithm terminates when $\Delta_k^p \le 10^{-12}$ (for MADS) or $\Delta_k \le 10^{-12}$ (for GPS), or when the number of function evaluations exceeds $600n$, whichever comes first. The algorithm was run with four values of n.

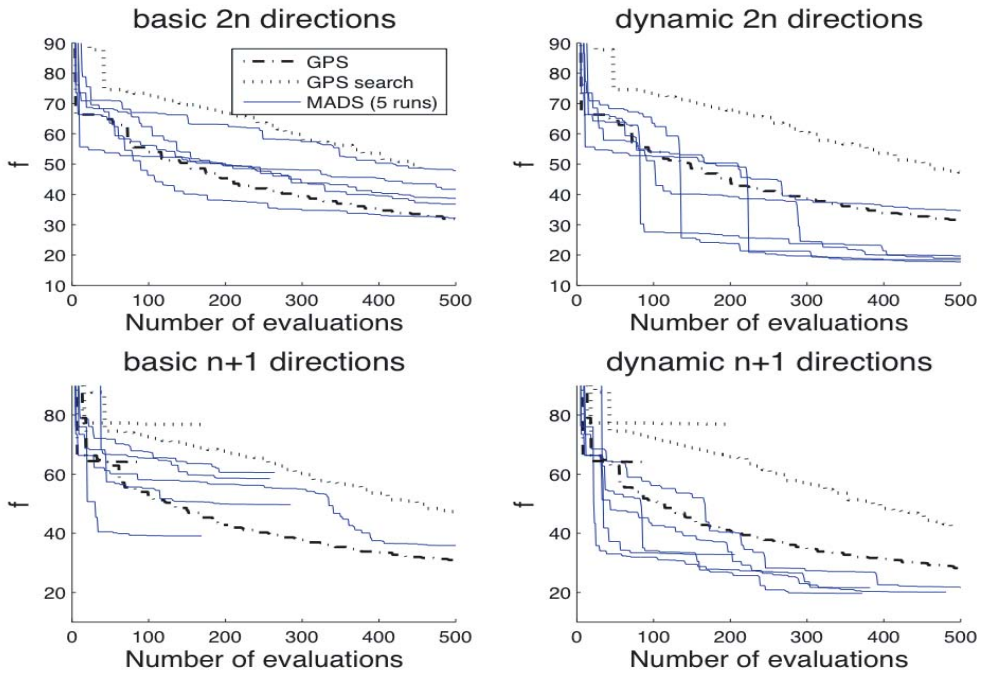


FIG. 5.5. *Data set* 3*—Progression of the objective function value vs the number of evaluations.*

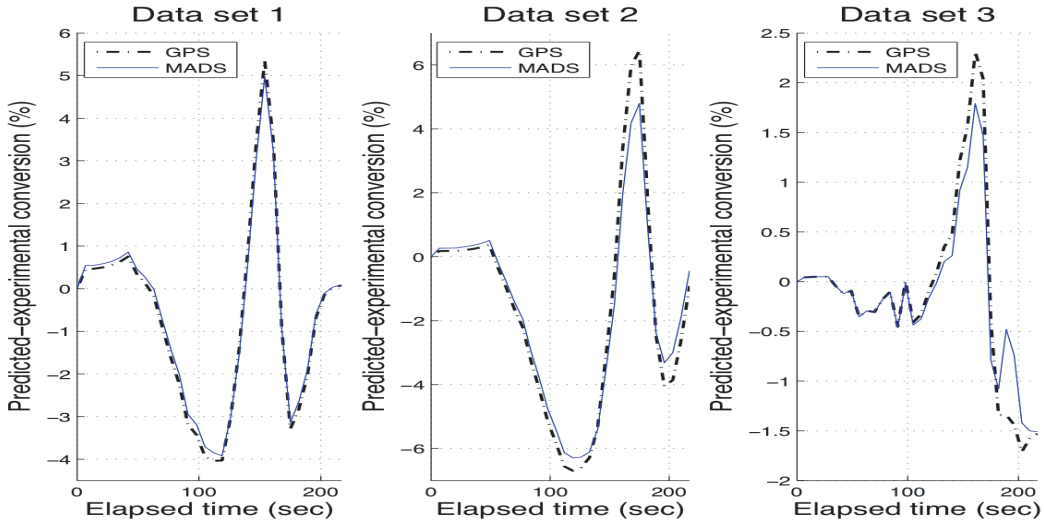


FIG. 5.6. *Conversion rate error versus time.*

For the GPS method we always used $D_k = D = [I, -I]$ with dynamic ordering. The GPS filter method is described in [4]. We used a search strategy, which we often use with the GPS filter method, consisting of a $5n$ point Latin hypercube sample at the first iteration, and a $n/5$ random search at other iterations.

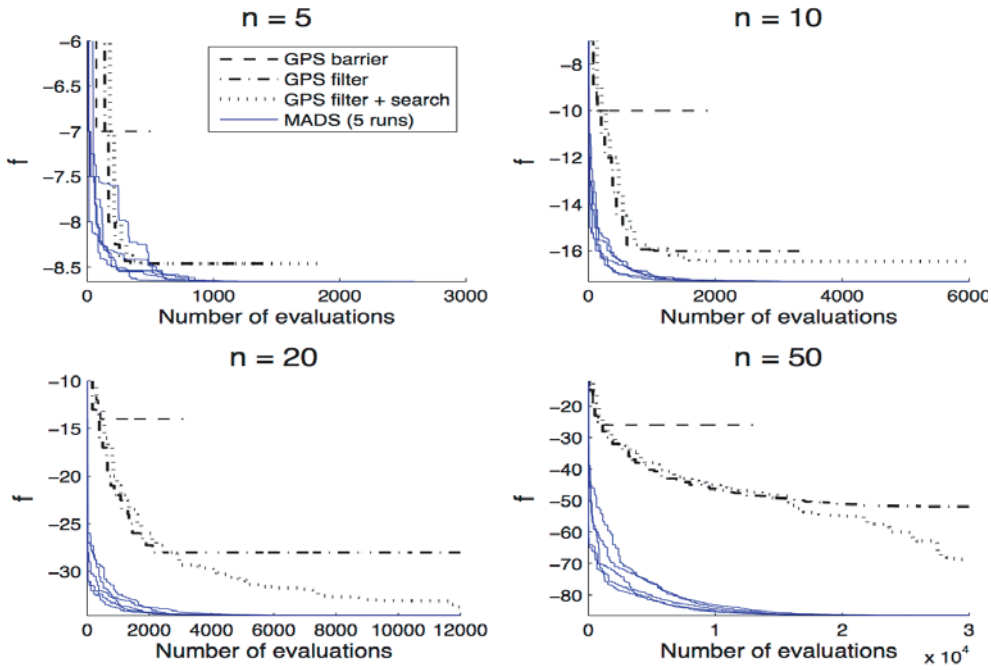


FIG. 5.7. *Progression of the objective function value vs the number of evaluations on an easy nonlinear problem.*

The behavior of LTMADS is comparable for every value of n. In every case, that MADS algorithm converged to the global optimal solution. The GPS barrier approach quickly moved to a point on the boundary of the domain and stalled there. The GPS filter approach was able to move away from that point, but it converged to a better suboptimal solution. The absence of a SEARCH strategy, and the restriction to a finite number of POLL directions traps the iterates at a nonoptimal solution. The addition of a random SEARCH strategy allows GPS with the filter, when n is $10, 20$, or 50, to move away from this solution, but it still was short of finding the optimal solution in the number of function calls allowed. The "GPS search" label below applies to the GPS with the filter and random SEARCH, because this combination performed better than GPS with a random SEARCH step. The progression of the runs is illustrated in Figure 5.7.

5.4. Numerical limitations. The optimal solution in this last example does not satisfy the hypotheses of any GPS or MADS theorems because it is located at $-\infty$. However, the example is intended to show how well the various algorithms track a feasible region that gets narrow quickly. Consider the following problem in $\mathbb{R}^2$:

$$\begin{aligned} \min_{x=(a,b)^T} \quad & a \\ \text{s.t.} \quad & e^a \le b \le 2e^a. \end{aligned}$$

The starting point is $(0,1)^T$, and the algorithm terminates when $\Delta_k^m < 10^{-323}$ (for MADS) or $\Delta_k < 10^{-323}$ (for GPS) i.e., when the mesh size parameter drops below the smallest positive representable number in double precision arithmetic. We admit that this is excessive, but we wanted to run the algorithms to their limits. The same strategies as in section 5.3 are used.

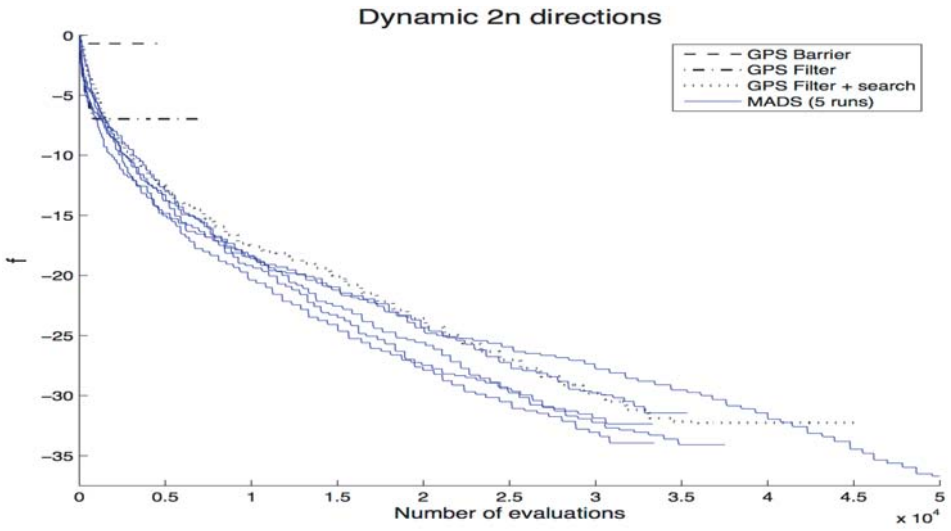


FIG. 5.8. *Progression of the objective function value versus the number of evaluations on a difficult nonlinear problem.*

The progression of the algorithms is illustrated in Figure 5.8. GPS with both the barrier and filter approaches to constraints converged quickly to points where the standard 2n basis does not contain a feasible descent direction. The filter GPS (without search) approach to constraints did better than the GPS barrier (without search) approach because it is allowed to become infeasible.

All 5 runs of the LTMADS method of the previous section ended with roughly the same solution, a point where $a \pm \Delta_k^p = a$ in finite arithmetic, which is all one can ask. The same behavior is observed for GPS with a random SEARCH (similar results were generated with or without the filter). The fact that LTMADS generates an asymptotically dense set of poll directions, and that a SEARCH step is conducted at each GPS iteration explain why both the GPS with a search and LTMADS do better than the GPS barrier or filter approach.

The feasible region is very narrow, and therefore it gets quite improbable that the MADS poll directions generate a feasible point. When such a feasible point is generated it is always very close to the frame center since the mesh and poll parameters are very small.

Even if the algorithm instances failed to solve this problem to optimality and converged to points that are not Clarke stationary points, the GPS and MADS convergence theory is not violated—yet. In all cases, there is a set of directions that positively span $\mathbb{R}^2$ such that for each direction either the Clarke generalized derivative is nonnegative or is an infeasible direction.

6. Discussion. GPS is a valuable algorithm, but the application of nonsmooth analysis techniques in [3] showed its limitations due to the finite choice of directions in [2]. MADS removes the GPS restriction to finitely many poll directions. We have long felt that this was the major impediment to stronger proofs of optimality for GPS limit points (and better behavior on nonsmooth problems), and in this paper we find more satisfying optimality conditions for MADS in addition to opening new options for handling nonlinear constraints.

It would be easy to define a GPS algorithm that contains a randomized selection of POLL directions. It would suffice to define the set of directions D to be large enough to contain more than one positive basis, as illustrated in Figure 2.1. But since that number would still remain finite and the directions would not fill the space, the theory would remain limited. Introducing randomness to GPS in this way or to MADS as in LTMADS does not make either into random search methods as viewed by that community. The interested reader can refer to [31] and Chapter 2 of [32] for the distinction.

MADS is a general framework which also contains both deterministic and randomized instances of polling direction choices. But most importantly, MADS allows infinitely many different polling directions. To illustrate the MADS generality, we proposed here a randomized way to choose polling directions. This method, which we called LTMADS, performed well, especially for a first implementation. Of course, this implies that the convergence analysis of LTMADS (and not that of the general framework MADS) requires probabilistic arguments.

We could have used a deterministic strategy to define our first instance of MADS, but the deterministic ways that we tried were such that when the algorithm was halted after finitely many iterations, the set of poll directions was often far from being uniformly distributed in $\mathbb{R}^n$. This convinced us to present LTMADS here rather than our early deterministic efforts despite being able to prove the same theorems without the need for probabilistic arguments.

We expect that more, and perhaps better, instances of MADS will be found, and we hope this paper will facilitate that. To have a MADS instance be backed by our convergence analysis, one needs to show that the new instance generates a dense set of refining directions.

When n is small, our examples suggested that GPS with a random SEARCH behaved similarly to MADS. Of course, it is well known [32] that random searches that sample the space using a uniform distribution get worse as the dimension of the problem increases, and the probability of improvement for a fixed dimension decreases as the function value decreases. Thus, for larger n, we would not recommend using a pure random search step with either GPS or MADS.

We think that the ideas here can be readily applied to choosing templates for implicit filtering [9], another very successful algorithm for nasty nonlinear problems.

7. Acknowledgments. Finally, we wish to thank Gilles Couture for coding NOMAD, the C++ implementation of MADS and GPS, and to acknowledge useful discussions with Andrew Booker, Mark Abramson, and Sébastien Le Digabel, and constructive comments by the associate editor Margaret Wright and by an anonymous referee.

REFERENCES

[1] M. A. Abramson, C. Audet, and J. E. Dennis, Jr., *Generalized pattern searches with derivative information*, Math. Program., 100 (2004), pp. 3–25.

[2] C. Audet, *Convergence results for pattern search algorithms are tight*, Optim. Eng., 5 (2004), pp. 101–122.

[3] C. Audet and J. E. Dennis, Jr., *Analysis of generalized pattern searches*, SIAM J. Optim., 13 (2003), pp. 889–903.

[4] C. Audet and J. E. Dennis, Jr., *A pattern search filter method for nonlinear programming without derivatives*, SIAM J. Optim., 14 (2004), pp. 980–1010.

[5] C. Audet and D. Orban, *Finding optimal algorithmic parameters using the mesh adaptive direct search algorithm*, SIAM J. Optim., to appear.

[6] M. Avriel, *Nonlinear Programming: Analysis and Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1976.
[7] V. Béchard, C. Audet, and J. Chaouki, *Robust optimization of chemical processes using a mads algorithm*, Technical report G-2005-16, Les Cahiers du Gerad, Montreal, Canada, 2005.
[8] 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.
[9] T. D. Choi and C. T. Kelley, *Superlinear convergence and implicit filtering*, SIAM J. Optim., 10 (2000), pp. 1149–1162.
[10] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983. Reissued in 1990 by SIAM Publications, Philadelphia, as Vol. 5 in the series Classics in Applied Mathematics.
[11] I. D. Coope and C. J. Price, *Frame based methods for unconstrained optimization*, J. Optim. Theory Appl., 107 (2000), pp. 261–274.
[12] 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.
[13] G. Couture, C. Audet, J. E. Dennis, Jr., and M. A. Abramson, *The NOMAD project*, http://www.gerad.ca/NOMAD/.
[14] C. Davis, *Theory of positive linear dependence*, Amer. J. Math., 76 (1954), pp. 733–746.
[15] D. E. Finkel and C. T. Kelley, *Convergence analysis of the direct algorithm*, Technical report, NCSU Mathematics Department, Raleigh, NC, 2004.
[16] F. J. Gould and J. W. Tolle, *Geometry of optimality conditions and constraint qualifications*, Math. Program., 2 (1972), pp. 1–18.
[17] R. E. Hayes, F. H. Bertrand, C. Audet, and S. T. Kolaczkowski, *Catalytic combustion kinetics: Using a direct search algorithm to evaluate kinetic parameters from light-off curves*, The Canadian Journal of Chemical Engineering, 81 (2003), pp. 1192–1199.
[18] J. Jahn, *Introduction to the Theory of Nonlinear Optimization*, Springer-Verlag, Berlin, 1996.
[19] T. G. Kolda, R. M. Lewis, and V. Torczon, *Optimization by direct search: new perspectives on some classical and modern methods*, SIAM Rev., 45 (2003), pp. 385–482.
[20] E. B. Leach, *A note on inverse function theorems*, Proc. Amer. Math. Soc., 12 (1961), pp. 694–697.
[21] R. M. Lewis and V. Torczon, *Rank ordering and positive bases in pattern search algorithms*, Technical report 96–71, Institute for Computer Applications in Science and Engineering, Mail Stop 132C, NASA 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] S. Lucidi, M. Sciandrone, and P. Tseng, *Objective-derivative-free methods for constrained optimization*, Math. Program., 92 (2002), pp. 37–59.
[24] A. L. Marsden, *Aerodynamic Noise Control by Optimal Shape Design*, Ph.D. thesis, Stanford University, Stanford, CA, 2004.
[25] Mathworks. Matlab gads toolbox. http://www.mathworks.com/products/gads/.
[26] C. J. Price and I. D. Coope, *Frames and grids in unconstrained and linearly constrained optimization: a nonsmooth approach*, SIAM J. Optim., 14 (2003), pp. 415–438.
[27] C. J. Price, I. D. Coope, and J. E. Dennis, Jr., *Direct search methods for nonlinearly constrained optimization using filters and frames*, Optim. Eng., 5 (2004), pp. 123–144.
[28] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[29] R. T. Rockafellar, *Generalized directional derivatives and subgradients of nonconvex functions*, Canad. J. Math., 32 (1980), pp. 257–280.
[30] V. Torczon, *On the convergence of pattern search algorithms*, SIAM J. Optim., 7 (1997), pp. 1–25.
[31] Z. B. Zabinsky and R. L. Smith, *Pure adaptive search in global optimization*, Math. Program., 53 (1992), pp. 323–338.
[32] Z. B. Zabinsky, *Stochastic Adaptive Search for Global Optimization*, Nonconvex Optimization and its Applications, 72, Kluwer Academic Publishers, Boston, 2003.

SIAM J. OPTIM.
Vol. 18, No. 4, pp. 1501–1503
© 2008 Society for Industrial and Applied Mathematics

ERRATUM: MESH ADAPTIVE DIRECT SEARCH ALGORITHMS FOR CONSTRAINED OPTIMIZATION*

CHARLES AUDET†, A. L. CUSTÓDIO‡, AND J. E. DENNIS, JR.§

Abstract. In [*SIAM J. Optim.*, 17 (2006), pp. 188–217] Audet and Dennis proposed the class of *mesh adaptive direct search* (MADS) *algorithms* for minimization of a nonsmooth function under general nonsmooth constraints. The notation used in the paper evolved since the preliminary versions, and, unfortunately, even though the statement of Proposition 4.2 is correct, it is not compatible with the final notation. The purpose of this note is to show that the proposition is valid.

Key words. mesh adaptive direct search algorithms, constrained optimization, nonsmooth optimization

AMS subject classifications. 90C30, 90C56, 65K05, 49J52

DOI. 10.1137/060671267

In [1] Audet and Dennis proposed the class of *mesh adaptive direct search* (MADS) *algorithms* for minimization of a nonsmooth function under general nonsmooth constraints. The paper contains a convergence analysis for this class of methods and proposes two variants of an implementable instance called LTMADS.

The proof that LTMADS is indeed an instance of MADS is not compatible with the notation used in the rest of the paper. We restate the proposition and propose a consistent proof.

PROPOSITION 0.1 (Proposition 4.2 of [1]). *At each iteration k, the procedure above yields a D_k and a MADS frame P_k such that*

$$P_k = \{x_k + \Delta_k^m d : d \in D_k\} \subset M_k,$$

where $\Delta_k^m > 0$ is the mesh size parameter, M_k is given by Definition 2.1 of [1], and D_k is a positive spanning set such that for each $d \in D_k$,

- *d can be written as a nonnegative integer combination of the directions in D: $d = Du$ for some vector $u \in \mathbb{N}^{n_D}$ that may depend on the iteration number k;*
- *the distance from the frame center x_k to a frame point $x_k + \Delta_k^m d \in P_k$ is bounded above by a constant times the poll size parameter: $\Delta_k^m \|d\|_\infty \leq \Delta_k^p \max\{\|d'\|_\infty : d' \in D\}$;*
- *limits (as defined in Coope and Price [2]) of convergent subsequences of the normalized sets $\overline{D_k} := \{\frac{d}{\|d\|_\infty} : d \in D_k\}$ are positive spanning sets.*

*Received by the editors October 2, 2006; accepted for publication (in revised form) July 1, 2007; published electronically January 16, 2008.

http://www.siam.org/journals/siopt/18-4/67126.html

†GERAD and Département de Mathématiques et de Génie Industriel, École Polytechnique de Montréal, C.P. 6079, Succ. Centre-ville, Montréal H3C 3A7, QC, Canada (Charles.Audet@gerad.ca, http://www.gerad.ca/Charles.Audet). This author's research was supported by NSERC grant 239436-01, AFOSR FA9550-04-1-0235, the Boeing Company, and ExxonMobil.

‡Departamento de Matemática, FCT-UNL, Quinta da Torre, 2829-516 Caparica, Portugal (alcustodio@fct.unl.pt, http://ferrari.dmat.fct.unl.pt/personal/alcustodio/). This author's research was supported by Centro de Matemática da Universidade de Coimbra and the FCT under grant POCI/MAT/59442/2004.

§Computational and Applied Mathematics Department, Rice University, 8419 42nd Ave. SW, Seattle, WA 98136 (dennis@rice.edu, http://www.caam.rice.edu/∼dennis). This author's research was supported by AFOSR FA9550-04-1-0235, the Boeing Company, and ExxonMobil.

Proof. In order to construct the set of directions D_k, the algorithm builds matrices at iteration k that should be called L_k, B_k, and B'_k. To ease the presentation, we omit the index k in the proof of the two first bullets. The index k reappears in the proof of the last bullet since this last result involves limits as k goes to infinity.

By the construction in [1], L is a lower triangular $(n-1)\times(n-1)$ matrix where each term on the diagonal is either plus or minus 2^ℓ, and the lower components are randomly chosen from the discrete set $\{-2^\ell+1,-2^\ell+2,\ldots,2^\ell-1\}$, with ℓ an integer that satisfies $2^\ell = 1/\sqrt{\Delta_k^m}$. The rules for updating the mesh size parameter Δ_k^m ensure that $\ell \in \mathbb{N}$. It follows that L is a basis in $\mathbb{R}^{n-1}$ with $|\det(L)| = 2^{\ell(n-1)}$. Let $\{p_1, p_2, \ldots, p_{n-1}\}$ be a random permutation of the set $\{1,2,\ldots,n\}\setminus\{\hat{\imath}\}$, where $\{\hat{\imath}\}$ is defined in [1]. The elements of the matrix B are defined as

$$\begin{array}{ll} B_{p_i,j} = L_{i,j} & \text{for } i,j = 1,2,\ldots,n-1,\\ B_{\hat{\imath},j} = 0 & \text{for } j = 1,2,\ldots,n-1,\\ B_{i,n} = b_i(\ell) & \text{for } i = 1,2,\ldots,n,\end{array}$$

where $b_i(\ell)$ is a vector that depends only on the value of the mesh size parameter and not on the iteration number (see section 4.1 of [1]). It follows that B is a permutation of the rows and the columns of a lower triangular matrix whose diagonal elements are either -2^ℓ or 2^ℓ. Therefore B is a basis in $\mathbb{R}^n$ and $|\det(B)| = 2^{\ell n}$.

The square matrix B' is obtained by permuting the columns of B, and therefore the columns of B' form a basis of $\mathbb{R}^n$. Furthermore, $|\det(B')| = |\det(B)| = 2^{\ell n}$.

One of the proposed versions of LTMADS uses a minimal positive basis at every iteration, and the other variant uses a maximal positive basis at every iteration. The columns of $[B' \;\; -b']$ with $b'_i = \sum_{j\in N} B'_{ij}$ define a minimal positive basis, and the columns of $[B' \;\; -B']$ define a maximal positive basis [3].

Therefore, if $D_k = [B' \;\; -b']$ or if $D_k = [B' \;\; -B']$, then all entries of D_k are integers in the interval $[-n2^\ell, n2^\ell]$ or in the interval $[-2^\ell, 2^\ell]$, respectively. It follows that each column d of D_k can be written as a nonnegative integer combination of the columns of $D = [I \;\; -I]$. Hence, the frame defined by D_k is on the mesh M_k.

Two cases must be considered to show the second bullet. Recall that with LTMADS, the poll size parameter Δ_k^p (see [1]) is defined differently depending on whether minimal or maximal positive bases are used. If the maximal positive basis construction is used, then $\|\Delta_k^m d\|_\infty = \Delta_k^m \|d\|_\infty = \sqrt{\Delta_k^m} = \Delta_k^p$. If the minimal positive basis construction is used, then $\|\Delta_k^m d\|_\infty = \Delta_k^m \|d\|_\infty \le n\sqrt{\Delta_k^m} = \Delta_k^p$. The proof of the second bullet follows by noticing that $\max\{\|d'\|_\infty : d' \in [I \;\; -I]\} = 1$.

To show the third bullet, we will verify that the limit of the normalized sets $\overline{D_k} := \{\frac{d}{\|d\|_\infty} : d \in D_k\}$ forms a positive basis. It suffices to show that the conditions (1a), (1b), and (C1) or (C2) of Coope and Price [2] hold.

- Conditions (1a) and (1b) ensure that the limit of any convergent subsequence of the sequence of bases $\overline{B'_k} := \{\frac{d}{\|d\|_\infty} : d \in B'_k\}$ is also a basis. Condition (1a) requires that $|\det(\overline{B'_k})|$ be bounded below by a positive constant that is independent of k. In our context, $|\det(\overline{B'_k})| = 1$ for all k, and therefore this condition is satisfied. Condition (1b) is also easily satisfied since normalized directions are used. It follows that the limit of $\overline{B'_k}$ is a basis.
- Conditions (C1) and (C2) involve the columns added to each basis B'_k to form a positive basis. In the case of the maximal bases, condition (C1) is easily satisfied. For the minimal bases, (C2) holds since all the structure constants ξ (again following the definition of Coope and Price [2]) satisfy $-1 \le \xi \le -\frac{1}{n}$.

This concludes the proof. ∎

REFERENCES

[1] C. Audet and J. E. Dennis, Jr., *Mesh adaptive direct search algorithms for constrained optimization*, SIAM J. Optim., 17 (2006), pp. 188–217.
[2] I. D. Coope and C. J. Price, *Frame based methods for unconstrained optimization*, J. Optim. Theory Appl., 107 (2000), pp. 261–274.
[3] C. Davis, *Theory of positive linear dependence*, Amer. J. Math., 76 (1954), pp. 733–746.

SIAM J. OPTIM.
Vol. 17, No. 1, pp. 218–242
© 2006 Society for Industrial and Applied Mathematics

SUMS OF SQUARES AND SEMIDEFINITE PROGRAM RELAXATIONS FOR POLYNOMIAL OPTIMIZATION PROBLEMS WITH STRUCTURED SPARSITY*

HAYATO WAKI†, SUNYOUNG KIM‡, MASAKAZU KOJIMA†, AND MASAKAZU MURAMATSU§

Abstract. Unconstrained and inequality constrained sparse polynomial optimization problems (POPs) are considered. A correlative sparsity pattern graph is defined to find a certain sparse structure in the objective and constraint polynomials of a POP. Based on this graph, sets of the supports for sums of squares (SOS) polynomials that lead to efficient SOS and semidefinite program (SDP) relaxations are obtained. Numerical results from various test problems are included to show the improved performance of the SOS and SDP relaxations.

Key words. polynomial optimization problem, sparsity, global optimization, Lagrangian relaxation, Lagrangian dual, sums of squares optimization, semidefinite program relaxation

AMS subject classifications. 15A15, 15A09, 15A23

DOI. 10.1137/050623802

1. Introduction.

Polynomial optimization problems (POPs) arise from various applications in science and engineering. Recent developments [9, 15, 18, 19, 22, 25, 27, 31, 32] in semidefinite program (SDP) and sums of squares (SOS) relaxations for POPs have attracted a lot of research from diverse directions. These relaxations have been extended to polynomial SDPs [11, 12, 17] and POPs over symmetric cones [20]. In particular, SDP and SOS relaxations have been popular for their theoretical convergence to the optimal value of a POP [22, 25]. From a practical point of view, improving the computational efficiency of SDP and SOS relaxations using the sparsity of polynomials in POPs has become an important issue [15, 19].

A polynomial f in real variables $x_1, x_2, \ldots, x_n$ of a positive degree d can have all monomials of the form $x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}$ with nonnegative integers α_i $(i = 1, 2, \ldots, n)$ such that $\alpha_i \geq 0$ and $\sum_{i=1}^n \alpha_i \leq d$; all monomials of different form add up to $\binom{n+d}{d}$. We call such a polynomial *fully dense*. When we examine polynomials in POPs from applications, we notice in many cases that they are *sparse* polynomials having a few or some of all possible monomials as defined in [19]. The sparsity provides a computational edge if it is handled properly when deriving SDP and SOS relaxations. More precisely, taking advantage of the sparsity of POPs is essential to obtaining an optimal value of a POP by applying SDP and SOS relaxations in practice.

For sparse POPs, generalized Lagrangian duals and their SOS relaxations were proposed in [15]. The relaxations are derived using SOS polynomials for the Lagrangian multipliers with sparsity similar to that of the associated constraint poly-

*Received by the editors February 3, 2005; accepted for publication (in revised form) December 8, 2005; published electronically May 12, 2006.
http://www.siam.org/journals/siopt/17-1/62380.html

†Department of Mathematical and Computing Sciences, Tokyo Institute of Technology, 2-12-1 Oh-Okayama, Meguro-ku, Tokyo 152-8552, Japan (Hayato.Waki@is.titech.ac.jp, kojima@is.titech.ac.jp).
‡Department of Mathematics, Ewha Women's University, 11-1 Dahyun-dong, Sudaemoon-gu, Seoul 120-750, Korea (skim@ewha.ac.kr). This author's research was supported by KRF 2003-041-C00038.
§Department of Computer Science, The University of Electro-Communications, Chofugaoka, Chofu-Shi, Tokyo 182-8585, Japan (muramatu@cs.uec.ac.jp).

nomials. Then the relaxations are converted into equivalent SDPs. As a result, the size of the resulting relaxations is reduced and computational efficiency is improved. This approach is shown to have an advantage in implementation over the SDP relaxation given in [22] whose size depends only on the degrees of objective and constraint polynomials of the POP.

The aim of this paper is to propose new practical SOS and SDP relaxations for a sparse POP and show their performance for various test problems. The framework of SOS and SDP relaxations presented here is based on the one proposed in [15]. The main idea here is that we define sparsity of a POP more precisely by finding a structure of the polynomials in the POP to obtain sparse SOS and SDP relaxations accordingly. Specifically, we introduce *correlative sparsity*, which is a special case of the sparsity [19] mentioned above; the correlative sparsity implies the sparsity, but the converse is not necessarily true. The correlative sparsity is described in terms of an $n \times n$ symmetric matrix $\boldsymbol{R}$, which we call the *correlative sparsity pattern matrix* (*csp matrix*) of the POP. Each element R_{ij} of the csp matrix $\boldsymbol{R}$ is either 0 or $\star$ representing a nonzero value. We assign $\star$ to every diagonal element R_{ii} $(i = 1, 2, \ldots, n)$, and also to each offdiagonal element $R_{ij} = R_{ji}$ $(1 \leq i < j \leq n)$ if and only if either (i) the variables x_i and x_j appear simultaneously in a term of the objective function or (ii) they appear in an inequality constraint. The csp matrix $\boldsymbol{R}$ constructed in this way represents the sparsity pattern of the Hessian matrix of the generalized Lagrangian function of [15] (or the Hessian matrix of the objective function in unconstrained cases) except for the diagonal elements; some diagonal elements of the Hessian matrix may vanish while $R_{ii} = \star$ $(i = 1, 2, \ldots, n)$ by definition. We say that the POP is *correlatively sparse* if the csp matrix $\boldsymbol{R}$ (or the Hessian matrix of the generalized Lagrangian function) is sparse.

From the csp matrix $\boldsymbol{R}$, it is natural to induce graph $G(N, E)$ with the node set $N = \{1, 2, \ldots, n\}$ and the edge set $E = \{\{i, j\} : R_{ij} = \star,\ i < j\}$ corresponding to the nonzero offdiagonal elements of $\boldsymbol{R}$. We call $G(N, E)$ the *correlative sparsity pattern graph* (*csp graph*) of the POP. We employ some results of graph theory regarding maximal cliques of chordal graphs [1]. A key idea in this paper is to use the maximal cliques of a chordal extension of the csp graph $G(N, E)$ to construct sets of supports for a sparse SOS relaxation. This idea is motivated by the recent work [5] that proposed positive semidefinite matrix completion techniques for exploiting sparsity in primal-dual interior-point methods for SDPs.

Theoretically, the proposed sparse SOS and SDP relaxations are not guaranteed to generate lower bounds of the same quality as the dense SDP relaxation [22] for general POPs. Practical experiences, however, show that the performance gap between the two relaxations is small as we will observe in section 6. In particular, the definition of a structured sparsity based on the csp matrix $\boldsymbol{R}$ and the csp graph $G(N, E)$ make it possible to achieve the same quality of lower bounds for quadratic optimization problems (QOPs) where all polynomials in the objective function and constraints are quadratic. More precisely, the proposed sparse relaxation of order 1 obtains lower bounds of the same quality as the dense SOS relaxation of order 1, as shown in section 4.5.

The remainder of the paper is organized as follows. After introducing basic notation and symbols of polynomials, we define SOS polynomials in section 2. In section 3, we first describe the dense SOS relaxation of unconstrained POPs and then the sparse SOS relaxation. We show how a csp matrix is defined from a given unconstrained POP and how a sparse SOS relaxation is constructed using the maximal cliques of a chordal extension of a csp graph induced from the csp matrix. Section 4 contains the

description of an SOS relaxation of an inequality constrained POP with a structured sparsity characterized by a csp matrix and a csp graph. We introduce a generalized Lagrangian dual for the inequality constrained POP and a sparse SOS relaxation. Section 5 discusses some additional techniques which enhance the practical performance of the sparse SOS relaxation such as computing optimal solutions, handling equality constraints, and scaling. Section 6 includes numerical results on various test problems. We show that the proposed sparse SOS and SDP relaxations exhibit much better performance in practice. Finally, we give concluding remarks in section 7.

2. Polynomials and SOS polynomials. Let $\mathbb{R}$ be the set of real numbers, and let $\mathbb{Z}_+$ be the set of nonnegative integers. $\mathbb{R}[\boldsymbol{x}]$ is the set of real-valued multivariate polynomials in x_i $(i = 1, 2, \ldots, n)$. Each polynomial $f \in \mathbb{R}[\boldsymbol{x}]$ is represented as $f(\boldsymbol{x}) = \sum_{\boldsymbol{\alpha}\in\mathcal{F}} c(\boldsymbol{\alpha})\boldsymbol{x}^{\boldsymbol{\alpha}}$, where $\mathcal{F} \subset \mathbb{Z}^n_+$ is a nonempty finite subset, $c(\boldsymbol{\alpha})$ $(\boldsymbol{\alpha} \in \mathcal{F})$ are real coefficients, and $\boldsymbol{x}^{\boldsymbol{\alpha}} = x_1^{\alpha_1}x_2^{\alpha_2}\cdots x_n^{\alpha_n}$. The *support* of f is defined by $\mathrm{supp}(f) = \{\boldsymbol{\alpha} \in \mathcal{F} : c(\boldsymbol{\alpha}) \neq 0\} \subset \mathbb{Z}^n_+$, and the *degree* of $f \in \mathbb{R}[\boldsymbol{x}]$ is defined by $\deg(f) = \max\{\sum_{i=1}^n \alpha_i : \boldsymbol{\alpha} \in \mathrm{supp}(f)\}$.

For every nonempty finite set $\mathcal{G} \subset \mathbb{Z}^n_+$, $\mathbb{R}[\boldsymbol{x}, \mathcal{G}]$ denotes the set of polynomials in x_i $(i = 1, 2, \ldots, n)$ whose support is in $\mathcal{G}$; i.e., $\mathbb{R}[\boldsymbol{x}, \mathcal{G}] = \{f \in \mathbb{R}[\boldsymbol{x}] : \mathrm{supp}(f) \subset \mathcal{G}\}$. We denote $\mathbb{R}[\boldsymbol{x}, \mathcal{G}]^2$ as the set of SOS polynomials in $\mathbb{R}[\boldsymbol{x}, \mathcal{G}]$. By construction, we see that $\mathrm{supp}(g) \subset \mathcal{G} + \mathcal{G}$ if $g \in \mathbb{R}[\boldsymbol{x}, \mathcal{G}]^2$, where $\mathcal{G} + \mathcal{G}$ denotes the Minkowski sum of two $\mathcal{G}$'s.

Let $\mathbb{R}^{\mathcal{G}}$ denote the $|\mathcal{G}|$-dimensional Euclidean space whose coordinates are indexed by $\boldsymbol{\alpha} \in \mathcal{G}$. Each vector of $\mathbb{R}^{\mathcal{G}}$ is denoted as $\boldsymbol{w} = (w_{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \mathcal{G})$. We use the symbol $\mathcal{S}(\mathcal{G})$ for the set of $|\mathcal{G}| \times |\mathcal{G}|$ symmetric matrices with coordinates $\boldsymbol{\alpha} \in \mathcal{G}$. Let $\mathcal{S}_+(\mathcal{G})$ be the set of positive semidefinite matrices in $\mathcal{S}(\mathcal{G})$; if $V \in \mathcal{S}_+(\mathcal{G})$,

$$\boldsymbol{w}^T\boldsymbol{V}\boldsymbol{w} = \sum_{\boldsymbol{\alpha}\in\mathcal{G}}\sum_{\boldsymbol{\beta}\in\mathcal{G}} V_{\boldsymbol{\alpha}\boldsymbol{\beta}}w_{\boldsymbol{\alpha}}w_{\boldsymbol{\beta}} \geq 0 \quad \text{for every } \boldsymbol{w} = (w_{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \mathcal{G}) \in \mathbb{R}^{\mathcal{G}}.$$

The symbol $\boldsymbol{u}(\boldsymbol{x}, \mathcal{G})$ is used for the $|\mathcal{G}|$-dimensional column vector consisting of elements $\boldsymbol{x}^{\boldsymbol{\alpha}}$ $(\boldsymbol{\alpha} \in \mathcal{G})$. Then, the set $\mathbb{R}[\boldsymbol{x}, \mathcal{G}]^2$ can be rewritten as

$$\mathbb{R}[\boldsymbol{x}, \mathcal{G}]^2 = \left\{\boldsymbol{u}(\boldsymbol{x}, \mathcal{G})^T\boldsymbol{V}\boldsymbol{u}(\boldsymbol{x}, \mathcal{G}) : \boldsymbol{V} \in \mathcal{S}_+(\mathcal{G})\right\}. \tag{2.1}$$

For more details, see [4, 25]. Let $N = \{1, 2, \ldots, n\}$, $\emptyset \neq C \subset N$, and

$$\mathcal{A}^C_\omega = \left\{\boldsymbol{\alpha} \in \mathbb{Z}^n_+ : \alpha_i = 0 \text{ if } i \notin C \text{ and } \sum_{i\in C}\alpha_i \leq \omega\right\}.$$

Then we observe that $\mathcal{A}^C_\omega + \mathcal{A}^C_\omega = \mathcal{A}^C_{2\omega}$ for every nonempty $C \subset N$ and $\omega \in \mathbb{Z}_+$.

3. SOS relaxations of unconstrained POPs. In this section, we consider an unconstrained POP,

$$\text{minimize } f_0(\boldsymbol{x}). \tag{3.1}$$

Let $\zeta^* = \inf\{f_0(\boldsymbol{x}) : \boldsymbol{x} \in \mathbb{R}^n\}$. Throughout this section, we assume that $\zeta^* > -\infty$. Then $\deg(f_0)$ must be an even integer, i.e., $\deg(f_0) = 2\omega_0$ for some $\omega_0 \in \mathbb{Z}_+$. By the lemma in section 3 of [29], we also know that $\mathcal{F}_0 = \mathrm{supp}(f_0) \subset \mathrm{conv}(\mathcal{F}^e_0)$, where $\mathcal{F}^e_0 = \{\boldsymbol{\alpha} \in \mathcal{F}_0 : \alpha_i \text{ is an even nonnegative integer } (i = 1, 2, \ldots, n)\}$.

3.1. An outline of sparse SOS relaxations. We first convert the POP (3.1) into an equivalent problem,

$$\text{maximize } \zeta \text{ subject to } f_0(\boldsymbol{x}) - \zeta \geq 0. \tag{3.2}$$

We fix a positive integer $\omega \geq \omega_0$, and replace the constraint of the problem (3.2) by an SOS constraint to obtain

$$\text{maximize } \zeta \text{ subject to } f_0(\boldsymbol{x}) - \zeta \in \mathbb{R}[\boldsymbol{x}, \mathcal{A}_\omega^N]^2. \tag{3.3}$$

The SOS optimization problem (3.3) serves as a relaxation of the POP (3.1). See [26] and the references therein for more details of this relaxation. We can rewrite the SOS constraint of (3.2) using the relation (2.1) as

$$f_0(\boldsymbol{x}) - \zeta = \boldsymbol{u}(\boldsymbol{x}, \mathcal{A}_\omega^N)^T \boldsymbol{V} \boldsymbol{u}(\boldsymbol{x}, \mathcal{A}_\omega^N) \quad \text{and} \quad \boldsymbol{V} \in \mathcal{S}_+(\mathcal{A}_\omega^N). \tag{3.4}$$

We call the parameter $\omega \in \mathbb{Z}_+$ in (3.3) the *(relaxation) order*. In fact, we can fix $\omega = \omega_0$ in the unconstrained case. Nevertheless, we regard ω as a parameter to be consistent with the notation of the constrained case.

We call a polynomial $f_0 \in \mathbb{R}[\boldsymbol{x}, \mathcal{A}_{2\omega}^N]$ *sparse* if the number of elements in its support $\mathcal{F}_0 = \text{supp}(f_0)$ is much smaller than the number of elements in $\mathcal{A}_{2\omega}^N$ that forms a support of fully dense polynomials in $\mathbb{R}[\boldsymbol{x}, \mathcal{A}_{2\omega}^N]$. When the objective function f_0 is a sparse polynomial in $\mathbb{R}[\boldsymbol{x}, \mathcal{A}_{2\omega}^N]$, the size of the SOS constraint (3.3) can be reduced by eliminating redundant elements from $\mathcal{A}_\omega^N$. In fact, by applying Theorem 1 of [29], $\mathcal{A}_\omega^N$ in problem (3.3) can be replaced by

$$\mathcal{G}_0^0 = \text{conv}\left\{\frac{\boldsymbol{\alpha}}{2} : \boldsymbol{\alpha} \in \mathcal{F}_0^e \bigcup \{\boldsymbol{0}\}\right\} \cap \mathbb{Z}_+^n \subset \mathcal{A}_\omega^N.$$

Note that $\{\boldsymbol{0}\}$ is added as the support for the real number variable ζ.

A method that can further reduce the size of the SOS optimization problem by eliminating redundant elements in $\mathcal{G}_0^0$ was proposed by Kojima, Kim, and Waki in [19]. We write the resulting SOS constraint from their method as

$$f_0(\boldsymbol{x}) - \zeta \in \mathbb{R}[\boldsymbol{x}, \mathcal{G}_0^*]^2, \tag{3.5}$$

where $\mathcal{G}_0^* \subset \mathcal{G}_0^0 \subset \mathcal{A}_\omega^N$ denotes the set obtained by applying the method.

We now outline a new sparse relaxation. Using the structure obtained from the correlative sparsity, we generate multiple support sets $\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_p \subset \mathbb{Z}_+^n$ such that

$$\mathcal{F}_0 \bigcup \{\boldsymbol{0}\} \subset \bigcup_{\ell=1}^p (\mathcal{G}_\ell + \mathcal{G}_\ell), \tag{3.6}$$

and replace the SOS constraint (3.5) by

$$f_0(\boldsymbol{x}) - \zeta \in \sum_{\ell=1}^p \mathbb{R}[\boldsymbol{x}, \mathcal{G}_\ell]^2, \tag{3.7}$$

where $\sum_{\ell=1}^p \mathbb{R}[\boldsymbol{x}, \mathcal{G}_\ell]^2 = \left\{\sum_{\ell=1}^p h_\ell : h_\ell \in \mathbb{R}[\boldsymbol{x}, \mathcal{G}_\ell]^2 \ (\ell = 1, 2, \ldots, p)\right\}$. The support of $f_0(\boldsymbol{x}) - \zeta$ is $\mathcal{F}_0 \bigcup \{\boldsymbol{0}\}$, while the support of each polynomial in $\sum_{\ell=1}^p \mathbb{R}[\boldsymbol{x}, \mathcal{G}_\ell]^2$ is contained in $\bigcup_{\ell=1}^p (\mathcal{G}_\ell + \mathcal{G}_\ell)$. Hence (3.6) is necessary for the SOS constraint (3.7) to be feasible although it is not sufficient. If the size of each $\mathcal{G}_\ell$ is much smaller than the size of $\mathcal{G}_0^*$ and if the number of the support sets p is not large, the size of the SOS constraint (3.7) is smaller than the size of the SOS constraint of (3.3).

3.2. Correlative sparsity pattern matrix. The sparsity considered here is measured by the number of different kinds of cross terms in the objective polynomial f_0. We will call this type of sparsity *correlative sparsity*. The correlative sparsity is represented with the $n \times n$ (symbolic, symmetric) *correlative sparsity pattern matrix* (abbreviated as *csp matrix*) $\boldsymbol{R}$ whose (i,j)th element R_{ij} is given by

$$R_{ij} = \begin{cases} \star & \text{if } i = j, \\ \star & \text{if } \alpha_i \geq 1 \ \text{ and } \alpha_j \geq 1 \ \text{ for some } \boldsymbol{\alpha} \in \mathcal{F}_0 = \mathrm{supp}(f_0), \\ 0 & \text{otherwise} \end{cases}$$

$(i = 1, 2, \ldots, n, \ j = 1, 2, \ldots, n)$. Here $\star$ stands for some nonzero element. If the csp matrix $\boldsymbol{R}$ of f_0 is sparse, then f_0 is sparse as defined in [19], but the converse is not true. We say that f_0 is *correlatively sparse* if the associated csp matrix is sparse. As mentioned in the introduction, the correlative sparsity of an objective function $f_0(\boldsymbol{x})$ is equivalent to the sparsity of its Hessian matrix with some additional nonzero diagonal elements.

3.3. Correlative sparsity pattern graphs. We describe a method to determine the sets of supports $\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_p$ for the target SOS relaxation (3.7) of the unconstrained POP (3.1). The basic idea is to use the structure of the csp matrix $\boldsymbol{R}$ and some results from graph theory.

Given a csp matrix $\boldsymbol{R}$, the undirected graph $G(N,E)$ with $N = \{1, 2, \ldots, n\}$ and $E = \{\{i,j\} : i,j \in N, \ i < j, \ R_{ij} = \star\}$ is called the *correlative sparsity pattern graph* (abbreviated as *csp graph*). Let $C_1, C_2, \ldots, C_p \subset N$ denote the maximal cliques of the csp graph $G(N,E)$. Then, choose the sets of supports $\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_p$ such that $\mathcal{G}_\ell = \mathcal{A}_\omega^{C_\ell}$ $(\ell = 1, 2, \ldots, p)$. We can easily verify that the relation (3.6) holds. However, the method described above for choosing $\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_p$ has a critical disadvantage since finding all maximal cliques of a graph is a difficult problem in general. In fact, finding a single maximum clique is an NP-hard problem. To resolve this difficulty, we generate a chordal extension $G(N,E')$ of the csp graph $G(N,E)$ and use the extended csp graph $G(N,E')$ instead of $G(N,E)$. See [1, 7] for chordal graphs and finding all maximal cliques.

Consequently, we obtain a sparse SOS relaxation of the POP (3.1):

$$\text{maximize} \quad \zeta \quad \text{subject to} \quad f_0(\boldsymbol{x}) - \zeta \in \sum_{\ell=1}^{p} \mathbb{R}[\boldsymbol{x}, \mathcal{A}_\omega^{C_\ell}]^2, \tag{3.8}$$

where C_ℓ $(\ell = 1, 2, \ldots, p)$ denote the maximal cliques of a chordal extension $G(N,E')$ of the csp graph $G(N,E)$.

There may be several different chordal extensions of a graph $G(N,E)$, and any of them is valid for deriving the sparse relaxation presented in this paper. The chordal extension with the least number of edges, called the minimum chordal extension, serves best for the resulting sparse relaxation. We remark that finding a chordal extension of a graph is equivalent to calculating symbolic sparse Cholesky factorization of its adjacency matrix; the resulting sparse matrix represents the chordal extension. The minimum chordal extension corresponds to the sparse Cholesky factorization with the minimum fill-ins. Finding the minimum chordal extension is difficult in general, but fortunately, several heuristics, such as the minimum degree ordering, are known to efficiently produce a good approximation. For more information on symbolic Cholesky factorization with minimum degree ordering and a chordal extension, see [6].

It should be noted that the number of the maximal cliques of $G(N, E')$ does not exceed n, which is equivalent to the number of nodes of the graph $G(N, E')$ as well as to the number of variables of the objective polynomial f_0.

Let us consider a few typical examples. Suppose that the objective polynomial function $f_0 \in \mathbb{R}[\boldsymbol{x}]_{2\omega}$ of the unconstrained POP (3.1) is a separable polynomial of the form $f_0(\boldsymbol{x}) = \sum_{i=1}^{n} h_i(x_i)$, where each $h_i(x_i)$ denotes a polynomial in a single variable $x_i \in \mathbb{R}$ with $\deg(h_i(x_i)) = 2\omega$. In this case, the csp matrix $\boldsymbol{R}$ becomes an $n \times n$ diagonal matrix so that $C_i = \{i\}$ $(i = 1, 2, \ldots, n)$. Hence we take $\mathcal{G}_\ell = \{\rho \boldsymbol{e}^\ell : \rho = 0, 1, 2, \ldots, \omega\}$ $(\ell = 1, 2, \ldots, n)$ in the sparse SOS relaxation (3.8). Here $\boldsymbol{e}^\ell \in \mathbb{R}^n$ denotes the ℓth unit vector with 1 at the ℓth coordinate and 0 elsewhere. The resulting SOS optimization problem inherits the separability from the separable polynomial objective function f_0, and is subdivided into n independent subproblems; each subproblem forms an SOS relaxation of the corresponding subproblem of the POP (3.1), minimizing $h_\ell(x_\ell)$ in a single variable. We remark here that if we directly apply the sparse SOS relaxation proposed in [19], we obtain the dense relaxation of the form (3.3). Therefore, this case shows a critical difference between the sparse SOS relaxation proposed in this paper and the one given in [19]. See Proposition 5.1 of [19] for more details.

Suppose that $f_0(\boldsymbol{x}) = \sum_{i=1}^{n-1}(a_i x_i^4 + b_i x_i^2 x_{i+1} + c_i x_i x_{i+1})$, where $a_i,\ b_i$, and c_i are nonzero real numbers $(i = 1, 2, \ldots, n-1)$. Then, the csp matrix turns out to be the $n \times n$ tridiagonal matrix which induces in fact a chordal graph; hence there is no need to extend. In this case, the maximal cliques of the chordal graph are $C_\ell = \{\ell,\ \ell+1\}$ $(\ell = 1, 2, \ldots, n-1)$.

For another example, let us consider $f_0(\boldsymbol{x}) = \sum_{i=1}^{n-1}(a_i x_i^4 + b_i x_i^2 x_n + c_i x_i x_n)$, where $a_i,\ b_i$, and c_i are nonzero real numbers $(i = 1, 2, \ldots, n-1)$. In this case, we have the csp matrix

$$\boldsymbol{R} = \begin{pmatrix} \star & 0 & \ldots & 0 & \star \\ 0 & \star & & 0 & \star \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \ldots & \star & \star \\ \star & \star & \ldots & \star & \star \end{pmatrix},$$

which gives a chordal graph with the maximal cliques $C_\ell = \{\ell,\ n\}$ $(\ell = 1, 2, \ldots, n-1)$.

4. SOS relaxations of inequality constrained POPs. Let $f_k \in \mathbb{R}[\boldsymbol{x}]$ $(k = 0, 1, 2, \ldots, m)$. Consider the following POP:

$$\text{(4.1)} \qquad \text{minimize} \quad f_0(\boldsymbol{x}) \quad \text{subject to} \quad f_k(\boldsymbol{x}) \geq 0 \ (k = 1, 2, \ldots, m).$$

Let $\zeta^* = \inf\{f_0(\boldsymbol{x}) : f_k(\boldsymbol{x}) \geq 0 \ (k = 1, 2, \ldots, m)\}$. With the correlative sparsity of the POP (4.1), we determine the generalized Lagrangian function with the same sparsity and proper sets of supports in an SOS relaxation. A sparse SOS relaxation is proposed in two steps. In the first step, we convert the POP (4.1) into an unconstrained minimization of the generalized Lagrangian function according to [15]. In the second step, we apply the sparse SOS relaxation given in the previous section for unconstrained POPs to the resulting minimization problem. A key point of utilizing the correlative sparsity of the POP (4.1) is that the POP (4.1) and its generalized Lagrangian function have the same correlative sparsity.

4.1. Correlative sparsity in inequality constrained POPs. Let $F_k = \{i : \alpha_i \geq 1 \text{ for some } \boldsymbol{\alpha} \in \mathrm{supp}(f_k)\}$ $(k = 1, 2, \ldots, m)$. Each F_k is regarded as the index set of variables x_i of the polynomial f_k. For example, if $n = 4$ and $f_k(\boldsymbol{x}) = x_1^3 + 3x_1x_4 - 2x_4^2$, then $F_k = \{1, 4\}$. Define the $n \times n$ (symbolic, symmetric) csp matrix $\boldsymbol{R}$ such that

$$R_{ij} = \begin{cases} \star & \text{if } i = j, \\ \star & \text{if } \alpha_i \geq 1 \text{ and } \alpha_j \geq 1 \text{ for some } \boldsymbol{\alpha} \in \mathrm{supp}(f_0), \\ \star & \text{if } i \in F_k \text{ and } j \in F_k \text{ for some } k \in \{1, 2, \ldots, m\}, \\ 0 & \text{otherwise.} \end{cases}$$

When the csp matrix $\boldsymbol{R}$ is sparse, we say that the POP (4.1) is correlatively sparse.

4.2. Generalized Lagrangian duals. The generalized Lagrangian function [15] is defined as

$$L(\boldsymbol{x}, \boldsymbol{\varphi}) = f_0(\boldsymbol{x}) - \sum_{k=1}^{m} \varphi_k(\boldsymbol{x}) f_k(\boldsymbol{x}),$$

where $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{\varphi} = (\varphi_1, \varphi_2, \ldots, \varphi_m) \in \Phi$, and

$$\Phi = \left\{ \boldsymbol{\varphi} = (\varphi_1, \varphi_2, \ldots, \varphi_m) : \begin{array}{l} \varphi_k \in \mathbb{R}[\boldsymbol{x}, \mathcal{A}_\omega^N]^2 \text{ for some } \omega \in \mathbb{Z}_+ \\ (k = 1, 2, \ldots, m) \end{array} \right\}.$$

Then, for each fixed $\boldsymbol{\varphi} \in \Phi$, the problem of minimizing $L(\boldsymbol{x}, \boldsymbol{\varphi})$ over $\boldsymbol{x} \in \mathbb{R}^n$ serves as a Lagrangian relaxation problem; its optimal value, $L^*(\boldsymbol{\varphi}) = \inf\{L(\boldsymbol{x}, \boldsymbol{\varphi}) : \boldsymbol{x} \in \mathbb{R}^n\}$, bounds the optimal value ζ^* of the POP (4.1) from below.

If our aim is to preserve the correlative sparsity of the POP (4.1) in the resulting SOS relaxation, we need to have the Lagrangian function L that inherits the correlative sparsity from the POP (4.1). Notice that $\boldsymbol{\varphi}$ can be chosen for this purpose. In [15], Kim, Kojima, and Waki proposed choosing a polynomial of the same variables as the variables x_i $(i \in F_k)$ in the polynomial f_k for each multiplier polynomial φ_k, i.e., $\mathrm{supp}(\varphi_k) \subset \{\boldsymbol{\alpha} \in \mathbb{Z}_+^n : \alpha_i = 0 \ (i \notin F_k)\}$. Let $\omega_k = \lceil \deg(f_k)/2 \rceil$ $(k = 0, 1, 2, \ldots, m)$ and $\omega_{\max} = \max\{\omega_k : k = 0, 1, \ldots, m\}$. For every nonnegative integer $\omega \geq \omega_{\max}$, define

$$\Phi_\omega = \left\{ \boldsymbol{\varphi} = (\varphi_1, \varphi_2, \ldots, \varphi_m) : \varphi_k \in \mathbb{R}[\boldsymbol{x}, \mathcal{A}_{\omega-\omega_k}^{F_k}]^2 \ (k = 1, 2, \ldots, m) \right\}.$$

Here the parameter $\omega \in \mathbb{Z}_+$ serves as the (relaxation) order of the SOS relaxation of the POP (4.1) that is derived in the next subsection. Then a generalized Lagrangian dual (with the Lagrangian multiplier $\boldsymbol{\varphi}$ restricted to Φ_ω) [15] is defined as

$$\text{maximize} \quad \zeta \quad \text{subject to} \quad L(\boldsymbol{x}, \boldsymbol{\varphi}) - \zeta \geq 0 \text{ and } \boldsymbol{\varphi} \in \Phi_\omega. \tag{4.2}$$

Let L_ω^* denote the optimal value of this problem: $L_\omega^* = \sup\{L^*(\boldsymbol{\varphi}) : \boldsymbol{\varphi} \in \Phi_\omega\}$. Then $L_\omega^* \leq \zeta^*$. If the POP (4.1) includes the box inequality constraint of the form $\rho - x_i^2 \geq 0$ $(i = 1, 2, \ldots, n)$ for some $\rho > 0$, we know by Theorem 3.1 of [15] that L_ω^* converges to ζ^* as $\omega \to \infty$.

4.3. Sparse SOS relaxations. We show how a sparse SOS relaxation is formulated using the sets of supports constructed from the csp matrix $\boldsymbol{R}$. Let $\omega \geq \omega_{\max}$ be fixed. Suppose that $\boldsymbol{\varphi} \in \Phi_\omega$. Then $L(\cdot, \boldsymbol{\varphi})$ forms a polynomial in x_i $(i = 1, 2, \ldots, n)$

with $\deg(L(\cdot,\boldsymbol{\varphi})) = 2\omega$. We also observe from the construction of the csp matrix $\boldsymbol{R}$ and Φ_ω that the polynomial $L(\cdot,\boldsymbol{\varphi})$ has the same csp matrix as the csp matrix $\boldsymbol{R}$ constructed for the POP (4.1). As in section 3.4, the csp matrix $\boldsymbol{R}$ induces the csp graph $G(N,E)$. By construction, we know that each F_k forms a clique of the csp graph $G(N,E)$. Let $C_1, C_2, \ldots, C_p$ be the maximal cliques of a chordal extension $G(N,E')$ of $G(N,E)$. Then, a sparse SOS relaxation of the POP (4.1) is written as

$$\text{(4.3)} \qquad \text{maximize} \quad \zeta \quad \text{subject to} \quad L(\boldsymbol{x},\boldsymbol{\varphi}) - \zeta \in \sum_{\ell=1}^{p} \mathbb{R}[\boldsymbol{x}, \mathcal{A}_\omega^{C_\ell}]^2 \ \text{ and } \boldsymbol{\varphi} \in \Phi_\omega.$$

Let ζ_ω denote the optimal objective value of this SOS optimization problem. Then $\zeta_\omega \le L_\omega^* \le \zeta^*$ for every $\omega \ge \omega_{\max}$, but the convergence of ζ_ω to ζ^* as $\omega \to \infty$ is not guaranteed in theory.

The above idea of the SOS relaxation of the constrained POP (4.1) using the generalized Lagrangian function stems from Putinar's lemma [28] and was first used in [22]. In fact, if we replace every index subset F_k of N $(k = 1, 2, \ldots, m)$ by the entire index set N and if we take $p = 1$ and $C_1 = N$, then the resulting SOS relaxation (4.3) of the POP (4.1) essentially coincides with the dense SOS relaxation (4.10) of Lasserre [22], and in this case, it was shown in [22] that $\zeta_\omega \to \zeta^*$ as $\omega \to \infty$ under moderate assumptions.

4.4. Primal approach. We have formulated a sparse SOS relaxation (4.3) of the inequality constrained POP (4.1) in the previous subsection. For numerical computation, we convert the SOS optimization problem (4.3) into an SDP, which serves as an SDP relaxation of the POP (4.1). We may regard this way of deriving an SDP relaxation from the POP (4.1) as the dual approach. We briefly mention below the so-called primal approach to the POP (4.1) whose sparsity is characterized by the csp matrix $\boldsymbol{R}$ and the csp graph $G(N,E)$. We use the same symbols and notation as in section 4.3. Let $\omega \ge \omega_{\max}$. To derive a primal SDP relaxation, we first transform the POP (4.1) into an equivalent polynomial SDP,

$$\text{(4.4)} \qquad \left.\begin{array}{ll} \text{minimize} & f_0(\boldsymbol{x}) \\ \text{subject to} & \boldsymbol{u}(\boldsymbol{x}, \mathcal{A}_{\omega-\omega_k}^{F_k})\boldsymbol{u}(\boldsymbol{x}, \mathcal{A}_{\omega-\omega_k}^{F_k})^T f_k(\boldsymbol{x}) \in \mathcal{S}_+(\mathcal{A}_{\omega-\omega_k}^{F_k}) \\ & (k = 1, 2, \ldots, m), \\ & \boldsymbol{u}(\boldsymbol{x}, \mathcal{A}_\omega^{C_\ell})\boldsymbol{u}(\boldsymbol{x}, \mathcal{A}_\omega^{C_\ell})^T \in \mathcal{S}_+(\mathcal{A}_\omega^{C_\ell}) \ (\ell = 1, 2, \ldots, p). \end{array}\right\}$$

The matrices $\boldsymbol{u}(\boldsymbol{x}, \mathcal{A}_{\omega-\omega_k}^{F_k})\boldsymbol{u}(\boldsymbol{x}, \mathcal{A}_{\omega-\omega_k}^{F_k})^T$ $(k = 1, 2, \ldots, m)$ and $\boldsymbol{u}(\boldsymbol{x}, \mathcal{A}_\omega^{C_\ell})\boldsymbol{u}(\boldsymbol{x}, \mathcal{A}_\omega^{C_\ell})^T$ $(\ell = 1, 2, \ldots, p)$ are positive semidefinite symmetric matrices of rank 1 for any $\boldsymbol{x} \in \mathbb{R}^n$, and have 1 as a diagonal element. These facts ensure the equivalence between the POP (4.1) and the polynomial SDP above. Let

$$\begin{aligned}
\widetilde{\mathcal{F}} &= \left(\bigcup_{\ell=1}^{p} \mathcal{A}_\omega^{C_\ell} \right) \setminus \{\boldsymbol{0}\}, \\
\widetilde{\mathcal{S}} &= \mathcal{S}(\mathcal{A}_{\omega-\omega_1}^{F_1}) \times \cdots \times \mathcal{S}(\mathcal{A}_{\omega-\omega_m}^{F_m}) \times \mathcal{S}(\mathcal{A}_\omega^{C_1}) \times \cdots \times \mathcal{S}(\mathcal{A}_\omega^{C_p}) \\
&\quad \text{(the set of block diagonal matrices of matrices in } \mathcal{S}(\mathcal{A}_{\omega-\omega_k}^{F_k}) \\
&\quad\ (k = 1, \ldots, m) \text{ and } \mathcal{S}(\mathcal{A}_\omega^{C_\ell}) \ (\ell = 1, \ldots, p) \text{ on their diagonal blocks)}, \\
\widetilde{\mathcal{S}}_+ &= \{\boldsymbol{M} \in \widetilde{\mathcal{S}} : \text{positive semidefinite}\}.
\end{aligned}$$

Then we can rewrite the polynomial SDP above as

$$\text{minimize} \quad \sum_{\boldsymbol{\alpha}\in\widetilde{\mathcal{F}}} \tilde{c}_0(\boldsymbol{\alpha})\boldsymbol{x}^{\boldsymbol{\alpha}} \quad \text{subject to} \quad \boldsymbol{M}(\boldsymbol{0}) + \sum_{\boldsymbol{\alpha}\in\widetilde{\mathcal{F}}} \boldsymbol{M}(\boldsymbol{\alpha})\boldsymbol{x}^{\boldsymbol{\alpha}} \in \widetilde{\mathcal{S}}_+$$

for some $\tilde{c}_0(\boldsymbol{\alpha}) \in \mathbb{R}$ ($\boldsymbol{\alpha} \in \widetilde{\mathcal{F}}$), $\boldsymbol{M}(\boldsymbol{0}) \in \widetilde{\mathcal{S}}$, and $\boldsymbol{M}(\boldsymbol{\alpha}) \in \widetilde{\mathcal{S}}$ ($\boldsymbol{\alpha} \in \widetilde{\mathcal{F}}$). Now, replacing each monomial $\boldsymbol{x}^{\boldsymbol{\alpha}}$ by a single real variable $y_{\boldsymbol{\alpha}}$, we have an SDP relaxation problem of (4.1):

$$\text{minimize} \quad \sum_{\boldsymbol{\alpha}\in\widetilde{\mathcal{F}}} \tilde{c}_0(\boldsymbol{\alpha})y_{\boldsymbol{\alpha}} \quad \text{subject to} \quad \boldsymbol{M}(\boldsymbol{0}) + \sum_{\boldsymbol{\alpha}\in\widetilde{\mathcal{F}}} \boldsymbol{M}(\boldsymbol{\alpha})y_{\boldsymbol{\alpha}} \in \widetilde{\mathcal{S}}_+. \tag{4.5}$$

We denote the optimal objective value by $\hat{\zeta}_\omega$.

The primal approach described in this section is based on the moment formulation proposed by [22], which is the dual to the SOS relaxation of the constrained POP (4.1). More precisely, if we replace every index subset F_k of N ($k = 1, 2, \ldots, m$) by the entire index set N and if we take $p = 1$ and $C_1 = N$, then we have the SDP relaxation of the POP (4.1) which corresponds to the SDP (4.6) of Lasserre [22]. In this case, the linearization of the matrix $\boldsymbol{u}(\boldsymbol{x}, \mathcal{A}^N_\omega)\boldsymbol{u}(\boldsymbol{x}, \mathcal{A}^N_\omega)^T$ forms the moment matrix $M_\omega(y)$ of the SDP (4.5) of Lasserre [22].

4.5. SOS and SDP relaxations of quadratic optimization problems with order 1. Consider a QOP

$$\left.\begin{array}{ll} \text{minimize} & \boldsymbol{x}^T\boldsymbol{Q}_0\boldsymbol{x} + 2\boldsymbol{q}_0^T\boldsymbol{x} \\ \text{subject to} & \boldsymbol{x}^T\boldsymbol{Q}_k\boldsymbol{x} + 2\boldsymbol{q}_k^T\boldsymbol{x} + \gamma_k \geq 0 \ (k = 1, 2, \ldots, m). \end{array}\right\} \tag{4.6}$$

Here $\boldsymbol{Q}_k$ denotes an $n \times n$ symmetric matrix, $\boldsymbol{q}_k \in \mathbb{R}^n$, and $\gamma_k \in \mathbb{R}$. In this case, we show that the proposed sparse SOS relaxation (4.3) of order $\omega = 1$ using any chordal extension of the csp graph $G(N, E)$ attains the same optimal value as the dense SOS relaxation [22] of order $\omega = 1$. This demonstrates an advantage of using the set of maximal cliques of a chordal extension of the csp graph $G(N, E)$ instead of the set of maximal cliques of $G(N, E)$ itself.

We formulate the dense [22] and sparse relaxations of order $\omega = 1$ using SOS polynomials from the dual side. Consider the Lagrangian dual of (4.6):

$$\text{maximize} \quad \zeta \quad \text{subject to} \quad L(\boldsymbol{x}, \boldsymbol{\varphi}) - \zeta \geq 0 \ (\forall \boldsymbol{x} \in \mathbb{R}^n) \ \text{ and } \boldsymbol{\varphi} \in \mathbb{R}^m_+, \tag{4.7}$$

where L denotes the Lagrangian function such that

$$L(\boldsymbol{x}, \boldsymbol{\varphi}) = \boldsymbol{x}^T \left(\boldsymbol{Q}_0 - \sum_{k=1}^m \varphi_k \boldsymbol{Q}_k\right)\boldsymbol{x} + 2\left(\boldsymbol{q}_0 - \sum_{k=1}^m \varphi_k \boldsymbol{q}_k\right)^T \boldsymbol{x} - \sum_{k=1}^m \varphi_k\gamma_k.$$

Then we replace the constraint $L(\boldsymbol{x}, \boldsymbol{\varphi}) - \zeta \geq 0$ ($\forall \boldsymbol{x} \in \mathbb{R}^n$) by an SOS condition $L(\boldsymbol{x}, \boldsymbol{\varphi}) - \zeta \in \mathbb{R}[\boldsymbol{x}, \mathcal{A}^N_1]^2$ to obtain the dense relaxation [22] of order $\omega = 1$,

$$\text{maximize} \quad \zeta \quad \text{subject to} \quad L(\boldsymbol{x}, \boldsymbol{\varphi}) - \zeta \in \mathbb{R}[\boldsymbol{x}, \mathcal{A}^N_1]^2 \ \text{ and } \boldsymbol{\varphi} \in \mathbb{R}^m_+. \tag{4.8}$$

Now consider the aggregated sparsity pattern matrix $\tilde{\boldsymbol{R}}$ over the coefficient matrices $\boldsymbol{Q}_0, \boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_m$ such that

$$\tilde{R}_{ij} = \begin{cases} \star & \text{if } i = j, \\ \star & \text{if } i \neq j \text{ and } [Q_k]_{ij} \neq 0 \ \text{ for some } k \in \{0, 1, 2, \ldots, m\}, \\ 0 & \text{otherwise,} \end{cases}$$

which coincides with the csp matrix of the Lagrangian function $L(\cdot,\boldsymbol{\varphi})$ with $\boldsymbol{\varphi} \in \mathbb{R}^m_+$. Let $G(N,E')$ be a chordal extension of the csp graph $G(N,E)$ from $\tilde{\boldsymbol{R}}$, and let C_ℓ $(\ell = 1,2,\ldots,p)$ be the maximal cliques of $G(N,E')$. Then we can apply the sparse relaxation (3.8) to the unconstrained minimization of the Lagrangian function $L(\cdot,\boldsymbol{\varphi})$ with $\boldsymbol{\varphi} \in \mathbb{R}^m_+$. Thus, replacing $\mathbb{R}[\boldsymbol{x},\mathcal{A}_1^N]^2$ in the dense SOS relaxation (4.8) by $\sum_{\ell=1}^p \mathbb{R}[\boldsymbol{x},\mathcal{A}_1^{C_\ell}]^2$, we obtain the sparse SOS relaxation

$$\left.\begin{array}{ll} \text{maximize} & \zeta \\ \text{subject to} & L(\boldsymbol{x},\boldsymbol{\varphi}) - \zeta \in \displaystyle\sum_{\ell=1}^{p} \mathbb{R}[\boldsymbol{x},\mathcal{A}_1^{C_\ell}]^2 \ \text{ and } \boldsymbol{\varphi} \in \mathbb{R}^m_+. \end{array}\right\} \tag{4.9}$$

Note that $L(\cdot,\boldsymbol{\varphi})$ is a quadratic function in $\boldsymbol{x} \in \mathbb{R}^n$ which results in the same csp graph $G(N,E)$ for each $\boldsymbol{\varphi} \in \mathbb{R}^m_+$, and that C_ℓ $(\ell = 1,2,\ldots,p)$ are the maximal cliques of a chordal extension $G(N,E')$ of the csp graph $G(N,E)$. Hence, if φ is chosen so that the Hessian matrix $\nabla_{xx}L(\boldsymbol{x},\varphi)$ of $L(\boldsymbol{x},\varphi)$ is positive semidefinite, $\nabla_{xx}L(\boldsymbol{x},\varphi)$ can be factorized using a Cholesky factorization such that $\nabla_{xx}L(\boldsymbol{x},\varphi) = \boldsymbol{M}\boldsymbol{M}^T$ for some $n \times n$ matrix $\boldsymbol{M}$ with the property $\{i \in N : M_{ij} \neq 0\} \subset C'_j$ for some maximal clique C'_j of $G(N,E')$ $(j = 1,2,\ldots,n)$. If in addition $L(\boldsymbol{x},\boldsymbol{\varphi}) - \zeta$ is an SOS polynomial or the constraint of the dense relaxation (4.8) is satisfied, then $L(\boldsymbol{x},\boldsymbol{\varphi}) - \zeta$ is represented as

$$L(\boldsymbol{x},\boldsymbol{\varphi}) - \zeta = (1,\boldsymbol{x}^T)\widetilde{\boldsymbol{M}}\widetilde{\boldsymbol{M}}^T \begin{pmatrix} 1 \\ \boldsymbol{x} \end{pmatrix} = \sum_{\ell=1}^{n} \left(\widetilde{\boldsymbol{M}}_{\cdot\ell+1}^T \begin{pmatrix} 1 \\ \boldsymbol{x} \end{pmatrix} \right)^2 + \alpha^2$$

for some $\alpha \geq 0$ and some $(1+n) \times (1+n)$ matrix $\widetilde{\boldsymbol{M}}$ of the form

$$\widetilde{\boldsymbol{M}} = \begin{pmatrix} \alpha & \boldsymbol{b} \\ \boldsymbol{0} & \boldsymbol{M} \end{pmatrix}.$$

Here $\widetilde{\boldsymbol{M}}_{\cdot\ell+1}$ denotes the $(\ell+1)$st column of $\widetilde{\boldsymbol{M}}$. It should be noted that each $\widetilde{\boldsymbol{M}}_{\cdot\ell+1}^T \binom{1}{\boldsymbol{x}}$ is an affine function whose support is contained in

$$\mathcal{A}_1^{C'_\ell} = \left\{ \boldsymbol{\alpha} \in \mathbb{Z}^n_+ : \alpha_i = 0 \ (i \notin C'_\ell) \right\}$$

as a polynomial. Therefore we have shown that the dense SOS relaxation (4.8) with order $\omega = 1$ is equivalent to the sparse SOS relaxation (4.9) with order $\omega = 1$.

5. Some technical issues.

5.1. Computing optimal solutions. Henrion and Lasserre [10] presented a linear algebra method that computes multiple optimal solutions of the POP (4.1). The moment matrix of full size induced from $\boldsymbol{u}(\boldsymbol{x},\mathcal{A}_\omega^N)\boldsymbol{u}(\boldsymbol{x},\mathcal{A}_\omega^N)^T$ plays an essential role in their method. In the proposed sparse relaxation, however, the moment matrix of full size is not available; instead multiple but partial moment matrices from $\boldsymbol{u}(\boldsymbol{x},\mathcal{A}_\omega^{C_\ell})\boldsymbol{u}(\boldsymbol{x},\mathcal{A}_\omega^{C_\ell})^T$ $(\ell = 1,2,\ldots,p)$, where the monomials in variables x_i $(i \in C_\ell)$ with degree up to ω are taken for the elements of the column vector $\boldsymbol{u}(\boldsymbol{x},\mathcal{A}_\omega^{C_\ell})$, are generated. As mentioned in the previous sections, we further apply the method [19] that eliminates redundant monomials from $\boldsymbol{u}(\boldsymbol{x},\mathcal{A}_\omega^{C_\ell})$ to reduce the size of the partial moment matrices. Because of these reasons, it is difficult to utilize the linear algebra method in the sparse relaxation.

We present a different technique. The basic idea is to perturb the POP (4.1) so that the projection of optimal solutions of the resulting primal SDP relaxation (4.5) onto the space of the variables x_i $(i = 1, 2, \ldots, n)$ consists of a unique point, which is the unique optimal solution of the perturbed POP. This technique is originally proposed in section 6.7 of [9]. We may assume without loss of generality that the objective polynomial function f_0 of the POP (4.1) is linear; if f_0 is not linear, we may replace $f_0(\boldsymbol{x})$ by a new variable x_0 and add the inequality constraint $f_0(\boldsymbol{x}) \leq x_0$.

We consider

$$\text{minimize } f_0(\boldsymbol{x}) + \boldsymbol{p}^T\boldsymbol{x} \text{ subject to } f_k(\boldsymbol{x}) \geq 0 \ (k = 1, 2, \ldots, m). \tag{5.1}$$

Here $\boldsymbol{p} \in \mathbb{R}^n$ denotes a perturbation vector. We then focus on the primal SDP relaxation of the perturbed POP (5.1), which can be described as the problem of minimizing $f_0(y_{e^1}, y_{e^2}, \ldots, y_{e^n}) + \sum_{i=1}^n p_i y_{e^i}$ subject to the constraint of the SDP (4.5). Define

$$\widetilde{D} = \{(y_{\boldsymbol{e}^1}, y_{\boldsymbol{e}^2}, \ldots, y_{\boldsymbol{e}^n}) \in \mathbb{R}^n : (y_{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \widetilde{\mathcal{F}}) \text{ is a feasible solution of (4.5)}\}.$$

Note that $\widetilde{D}$ is a convex subset of $\mathbb{R}^n$. Then the primal SDP relaxation of the perturbed POP (5.1) is equivalent to the convex program

$$\text{minimize } f_0(\boldsymbol{x}) + \boldsymbol{p}^T\boldsymbol{x} \text{ subject to } \boldsymbol{x} \in \widetilde{D}, \tag{5.2}$$

which may be regarded as the projection of the primal SDP relaxation of the perturbed POP (5.1) onto the space of the variables of (5.1). Now we assume a certain weak stability for the optimal solution set of the convex program (5.2): there exist $\epsilon > 0$ and $\rho > 0$ such that the optimal solution set of the convex program (5.2) is nonempty and is contained in the ball $B(\rho) = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\| \leq \rho\}$ for any perturbation with $\|\boldsymbol{p}\| \leq \epsilon$. In this case, we can replace the feasible region $\widetilde{D}$ of the convex program (5.2) by $\widetilde{D} \cap B(\rho)$. Now $\widetilde{D} \cap B(\rho)$ is convex and bounded. Hence, the convex program (5.2) has a unique solution for almost every $\boldsymbol{p}$ with $\|\boldsymbol{p}\| \leq \epsilon$ by Theorem 2.2.9 of [30].

Consequently, under the assumption on its optimal solution set, the convex program (5.2) has a unique optimal solution for almost every small $\boldsymbol{p}$. Suppose that

(a) $\boldsymbol{p}$ is sufficiently small;
(b) the convex program (5.2) has a unique optimal solution $\hat{\boldsymbol{x}}$; this means that $\hat{\boldsymbol{x}} = (\hat{y}_{e^1}, \hat{y}_{e^2}, \ldots, \hat{y}_{e^n})^T$ is obtained from any optimal solution $(\hat{y}_{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \widetilde{\mathcal{F}})$ of the primal SDP relaxation of the perturbed POP (5.1);
(c) the optimal value of the primal SDP relaxation of (5.1) coincides with the value $f_0(\hat{\boldsymbol{x}}) + \boldsymbol{p}^T\hat{\boldsymbol{x}}$; when f_0 is linear, this condition always holds;
(d) $\hat{\boldsymbol{x}}$ is a feasible solution of the perturbed POP (5.1).

Then $\hat{\boldsymbol{x}}$ is an optimal solution of the perturbed POP (5.1), which may be regarded as an approximate optimal solution of the original POP (4.1).

5.2. Equality constraints. Consider the POP

$$\left.\begin{array}{ll} \text{minimize} & f_0(\boldsymbol{x}) \\ \text{subject to} & f_k(\boldsymbol{x}) \geq 0 \ (k = 1, 2, \ldots, m), \ h_j(\boldsymbol{x}) = 0 \ (j = 1, 2, \ldots, q). \end{array}\right\} \tag{5.3}$$

Here $h_j \in \mathbb{R}[\boldsymbol{x}]$. Replacing each $h_j(\boldsymbol{x}) = 0$ by two inequality constraints $h_j(\boldsymbol{x}) \geq 0$ and $-h_j(\boldsymbol{x}) \geq 0$, we reduce the POP (5.3) to the inequality constrained POP:

$$\left.\begin{array}{ll} \text{minimize} & f_0(\boldsymbol{x}) \\ \text{subject to} & f_k(\boldsymbol{x}) \geq 0 \ (k = 1, 2, \ldots, m), \\ & h_j(\boldsymbol{x}) \geq 0, \ -h_j(\boldsymbol{x}) \geq 0 \ (j = 1, 2, \ldots, q). \end{array}\right\} \tag{5.4}$$

Let

$$\begin{aligned} \omega_k &= \lceil \deg(f_k)/2 \rceil \ (k = 0, 1, 2, \ldots, m), \\ \chi_j &= \lceil \deg(h_j)/2 \rceil \ (j = 1, 2, \ldots, q), \\ \omega_{\max} &= \max\{\omega_k \ (k = 0, 1, 2, \ldots, m), \ \chi_j \ (j = 1, 2, \ldots, q)\}, \\ F_k &= \{i : \alpha_i \geq 1 \text{ for some } \boldsymbol{\alpha} \in \mathrm{supp}(f_k)\} \ (k = 1, 2, \ldots, m), \\ H_j &= \{i : \alpha_i \geq 1 \text{ for some } \boldsymbol{\alpha} \in \mathrm{supp}(h_j)\} \ (j = 1, 2, \ldots, q). \end{aligned}$$

We construct the csp matrix $\boldsymbol{R}$ and the csp graph $G(N, E)$ of the POP (5.4). Let $C_1, C_2, \ldots, C_p$ be the maximal cliques of a chordal extension of $G(N, E)$, and let $\omega \geq \omega_{\max}$. Applying the SOS relaxation in section 4 to the POP (5.4), we have

$$\begin{array}{ll} \text{maximize} & \zeta \\ \text{subject to} & f_0(\boldsymbol{x}) - \displaystyle\sum_{k=1}^{m} \varphi_k(\boldsymbol{x}) f_k(\boldsymbol{x}) - \sum_{j=1}^{q} \left(\psi_j^+(\boldsymbol{x}) - \psi_j^-(\boldsymbol{x})\right) h_j(\boldsymbol{x}) - \zeta \\ & \in \sum_{\ell=1}^{p} \mathbb{R}[\boldsymbol{x}, \mathcal{A}_\omega^{C_\ell}]^2, \\ & \boldsymbol{\varphi} \in \Phi_\omega, \ \psi_j^+, \psi_j^- \in \mathbb{R}[\boldsymbol{x}, \mathcal{A}_{\omega-\chi_j}^{H_j}]^2 \ (j = 1, 2, \ldots, q). \end{array}$$

Since $\mathbb{R}[\boldsymbol{x}, \mathcal{A}_{\omega-\chi_j}^{H_j}]^2 - \mathbb{R}[\boldsymbol{x}, \mathcal{A}_{\omega-\chi_j}^{H_j}]^2 = \mathbb{R}[\boldsymbol{x}, \mathcal{A}_{2(\omega-\chi_j)}^{H_j}]$, this problem is equivalent to

$$\left.\begin{array}{ll} \text{maximize} & \zeta \\ \text{subject to} & f_0(\boldsymbol{x}) - \displaystyle\sum_{k=1}^{m} \varphi_k(\boldsymbol{x}) f_k(\boldsymbol{x}) - \sum_{j=1}^{q} \psi_j(\boldsymbol{x}) h_j(\boldsymbol{x}) - \zeta \\ & \in \sum_{\ell=1}^{p} \mathbb{R}[\boldsymbol{x}, \mathcal{A}_\omega^{C_\ell}]^2, \\ & \boldsymbol{\varphi} \in \Phi_\omega, \ \psi_j \in \mathbb{R}[\boldsymbol{x}, \mathcal{A}_{2(\omega-\chi_j)}^{H_j}] \ (j = 1, 2, \ldots, q). \end{array}\right\} \tag{5.5}$$

We can solve the SOS optimization problem (5.5) as an SDP with free variables.

5.3. Reducing the sizes of SOS relaxations. In [19], a method of reducing the size of the SOS relaxation is proposed by exploiting sparsity. The method consists of two phases. Suppose that, given an SOS polynomial f whose support is $\mathcal{F}$, we want to represent f using unknown polynomials $\phi_i \in \mathbb{R}[\boldsymbol{x}, \mathcal{G}]$ $(i = 1, 2, \ldots, k)$ with some support $\mathcal{G}$ such that $f = \sum_{i=1}^{k} \phi_i^2$. In phase 1 of the method in [19], we compute $\mathcal{G}^0 = \mathrm{conv}\left\{\frac{\boldsymbol{\alpha}}{2} : \boldsymbol{\alpha} \in \mathcal{F}^e\right\} \bigcap \mathbb{Z}_+^n$, where $\mathcal{F}^e = \{\boldsymbol{\alpha} \in \mathcal{F} : \alpha_i \text{ is even } (i = 1, 2, \ldots, n)\}$. It is known in [29] that $\mathrm{supp}(\phi_i) \subset \mathcal{G}^0$ for any SOS representation of $f = \sum_{i=1}^{k} \phi_i^2$. In phase 2, we eliminate redundant elements from $\mathcal{G}^0$ that are unnecessary in any SOS representation of f.

In the sparse SOS relaxations (3.8) and (4.3), we can apply phase 2 of the method with some modification to eliminate redundant elements from $\mathcal{A}_\omega^{C_\ell}$ $(\ell = 1, 2, \ldots, p)$. Let $\mathcal{F}$ denote the support of a polynomial f which we want to represent as

$$f = \sum_{\ell=1}^{p} \psi_\ell \quad \text{for some } \psi_\ell \in \mathbb{R}[\boldsymbol{x}, \mathcal{G}_\ell]^2 \ (\ell = 1, 2, \ldots, p). \tag{5.6}$$

The polynomial f can be either $f_0 - \zeta$ in the unconstrained POP (3.1), or $L(\cdot, \boldsymbol{\varphi}) - \zeta$ with $\boldsymbol{\varphi} \in \Phi_\omega$ in the constrained POP (4.1). In both cases, we assume that the family of supports $\mathcal{G}_\ell = \mathcal{A}_\omega^{C_\ell}$ $(\ell = 1, 2, \ldots, p)$ is sufficient to represent f as in (5.6); hence phase 1 is not implemented. Let $\mathcal{F}^e = \{\boldsymbol{\alpha} \in \mathcal{F} : \alpha_i \text{ is even } (i = 1, 2, \ldots, n)\}$. For each $\boldsymbol{\alpha} \in \bigcup_{\ell=1}^p \mathcal{G}_\ell$, we check whether the following relations are true:

$$2\boldsymbol{\alpha} \notin \mathcal{F}^e \ \text{ and } 2\boldsymbol{\alpha} \notin \bigcup_{\ell=1}^{p} \{\boldsymbol{\beta} + \boldsymbol{\gamma} : \boldsymbol{\beta} \in \mathcal{G}_\ell,\ \boldsymbol{\gamma} \in \mathcal{G}_\ell,\ \boldsymbol{\beta} \neq \boldsymbol{\alpha}\}.$$

If an $\boldsymbol{\alpha} \in \mathcal{G}_\ell$ satisfies these relations, we can eliminate $\boldsymbol{\alpha}$ from $\mathcal{G}_\ell$ and continue this process until no $\boldsymbol{\alpha} \in \bigcup_{\ell=1}^p \mathcal{G}_\ell$ satisfies these two relations. See [19] for more details.

5.4. Supports for Lagrange multiplier polynomials. In the generalized Lagrangian dual (4.2) and the sparse SOS relaxation (4.3), each multiplier polynomial φ_k has been chosen from the SOS polynomials with the support $\mathcal{A}_{\omega-\omega_k}^{F_k}$ to inherit the correlative sparsity from the original POP (4.1). For each k, let $J_k = \{\ell : F_k \subset C_\ell\}$ $(k = 1, 2, \ldots, m)$. By construction, $J_k \neq \emptyset$. Now we can replace the support $\mathcal{A}_{\omega-\omega_k}^{F_k}$ of SOS polynomials for φ_k by a union of $\mathcal{A}_{\omega-\omega_k}^{C_\ell}$ over some $\ell \in J_k$ in the sparse SOS relaxation (4.3). This modification strengthens the SOS relaxation (4.3) without losing much of the correlative sparsity of the other part.

5.5. Valid polynomial inequalities and their linearization. By adding appropriate valid polynomial inequalities to the constrained POP (4.1), we can strengthen its SDP relaxation (4.5). This idea has been used in many convex relaxation methods. See [18] and the references therein. We consider two types of valid polynomial inequalities that occur frequently in practice. These inequalities are used for some test problems in the numerical experiments in section 6. Suppose that (4.1) involves the nonnegative and upper bound constraints on all variables: $0 \leq x_i \leq \rho_i$ $(i = 1, 2, \ldots, n)$, where ρ_i denotes a nonnegative number $(i = 1, 2, \ldots, n)$. In this case, $0 \leq \boldsymbol{x}^{\boldsymbol{\alpha}} \leq \boldsymbol{\rho}^{\boldsymbol{\alpha}}$ $(\boldsymbol{\alpha} \in \widetilde{\mathcal{F}})$ form valid inequalities, where $\boldsymbol{\rho} = (\rho_1, \rho_2, \ldots, \rho_n) \in \mathbb{R}^n$. Therefore we can add their linearizations $0 \leq y_{\boldsymbol{\alpha}} \leq \boldsymbol{\rho}^{\boldsymbol{\alpha}}$ to the primal SDP relaxation (4.5). The complementarity condition $x_i x_j = 0$ is another example. If $\alpha_i \geq 1$ and $\alpha_j \geq 1$ for some $\boldsymbol{\alpha} \in \mathbb{Z}_+^n$, then $\boldsymbol{x}^{\boldsymbol{\alpha}} = 0$ forms a valid equality in this case; hence we can add $y_{\boldsymbol{\alpha}} = 0$ to the primal SDP relaxation or we can reduce the size of the primal SDP relaxation by eliminating the variable $y_{\boldsymbol{\alpha}} = 0$.

5.6. Scaling. High degree of polynomials in POPs can cause numerical problems. Even when the degrees of objective and constrained polynomials are small, the polynomial SDP (4.4) involves high degree monomials $\boldsymbol{x}^{\boldsymbol{\alpha}}$ as the order ω gets larger. Note that each variable $y_{\boldsymbol{\alpha}}$ corresponds to a monomial $\boldsymbol{x}^{\boldsymbol{\alpha}}$. More precisely, if $\boldsymbol{x}$ is a feasible solution of the POP (4.1), then $(y_{\boldsymbol{\alpha}} = \boldsymbol{x}^{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \widetilde{\mathcal{F}})$ is a feasible solution of the primal SDP relaxation (4.5) with the same objective value as (4.1). Therefore, if the magnitudes of some components of a feasible (or optimal) solution $\boldsymbol{x}$ of (4.1) are much larger (or smaller) than 1, the magnitude of some components of the corresponding solution $(y_{\boldsymbol{\alpha}} = \boldsymbol{x}^{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \widetilde{\mathcal{F}})$ can be huge (or tiny). This may be the source of numerical difficulties. To avoid such unbalanced magnitudes in the components of feasible (or optimal) solutions of the primal SDP relaxation (4.5), it would be ideal to scale the POP (4.1) so that the magnitudes of all nonzero components of optimal solutions of the scaled problem are near 1. Practically such an ideal scaling is impossible.

Here we restrict our discussion to a POP of the form (4.1) with additional finite lower and upper bounds on variables x_i $(i = 1, 2, \ldots, n)$: $\eta_i \leq x_i \leq \rho_i$ $(i = 1, 2, \ldots)$,

where η_i and ρ_i denote real numbers such that $\eta_i < \rho_i$. In this case, we can perform a linear transformation to the variables x_i such that $z_i = (x_i - \eta_i)/(\rho_i - \eta_i)$. Then we have objective and constrained polynomials $g_k \in \mathbb{R}[\boldsymbol{z}]$ $(k = 0, 1, \ldots, m)$ such that

$$g_k(z_1, z_2, \ldots, z_n) = f_k((\rho_1 - \eta_1)z_1 + \eta_1, (\rho_2 - \eta_2)z_2 + \eta_2, \ldots, (\rho_n - \eta_n)z_n + \eta_n).$$

We further normalize the coefficients of each $g_k \in \mathbb{R}[\boldsymbol{z}]$ such that $g'_k(\boldsymbol{z}) = g_k(\boldsymbol{z})/\nu_k$. Here ν_k denotes the maximum magnitude of the coefficients of the polynomial $g_k \in \mathbb{R}[\boldsymbol{z}]$ $(k = 0, 1, 2, \ldots, m)$. Consequently, we obtain a scaled POP which is equivalent to the POP (4.1) with the additional bounding constraint on variables x_i $(i = 1, 2, \ldots, n)$:

$$\left.\begin{array}{ll} \text{minimize} & g'_0(\boldsymbol{z}) \\ \text{subject to} & g'_k(\boldsymbol{z}) \geq 0 \ (k = 1, 2, \ldots, m), \ 0 \leq z_i \leq 1 \ (i = 1, 2, \ldots, n). \end{array}\right\} \tag{5.7}$$

We note that the scaled POP (5.7) provides the same csp matrix as the original POP (4.1). Furthermore, we can add the constraints $0 \leq y_{\boldsymbol{\alpha}} \leq 1$ $(\boldsymbol{\alpha} \in \widetilde{\mathcal{F}})$ to its primal SDP (4.5) to strengthen the relaxation. A similar technique can be found in [9].

6. Numerical results. In this section, we present numerical results of the proposed sparse relaxation for unconstrained and constrained problems. The focus is on verifying the efficiency of the sparse relaxation compared with the dense relaxation in [22]. The sparse and dense relaxations were implemented with MATLAB for constructing SDP problems and then a software package SeDuMi 1.05 was used to solve the SDP problems. All the experiments were done on a 2.4GHz AMD Opteron cpu with 8.0GB memory. Unconstrained problems that we deal with are benchmark test problems from [3, 21, 24] and randomly generated test problems with artificial correlative sparsity. Constrained test problems (section 6.2) are chosen from [8] and optimal control problems [2].

We employ the techniques described in section 5.1 for finding an optimal solution. In particular, we use the random perturbation techniques with the parameter $\epsilon = 10^{-5}$ in all the experiments presented here. After an optimal solution $\hat{\boldsymbol{y}}$ of an SDP relaxation of the POP is found by SeDuMi, the linear part $\hat{\boldsymbol{x}}$ is considered for a candidate of an optimal solution of the POP.

With regard to computing the accuracy of an obtained solution, we use the following for an unconstrained POP with an objective function f_0:

$$\epsilon_{\text{obj}} = \frac{|\text{the optimal value of SDP} - (f_0(\hat{\boldsymbol{x}}) + \boldsymbol{p}^T\hat{\boldsymbol{x}})|}{\max\{1, |f_0(\hat{\boldsymbol{x}}) + \boldsymbol{p}^T\hat{\boldsymbol{x}}|\}}.$$

Here $\boldsymbol{p} \in \mathbb{R}^n$ denotes a randomly generated perturbation vector such that $|p_j| < \epsilon = 10^{-5}$ $(j = 1, 2, \ldots, n)$. For an inequality and equality constrained POP of the form (5.3), we need another measure for feasibility in addition to ϵ_{obj} defined above. The following feasibility measure is used:

$$\epsilon_{\text{feas}} = \min\{f_k(\hat{\boldsymbol{x}}) \ (k = 1, \ldots, m), \ -|h_j(\hat{\boldsymbol{x}})| \ (j = 1, \ldots, q)\}.$$

We use the technique given in section 5.2 for every equality constrained problem and the technique in section 5.3 of reducing the size of an SOS relaxation for all test problems. In addition, we apply the techniques presented in sections 5.4, 5.5, and 5.6 to every constrained problem from the literature [8]. Specifically, we use $\cup_{\ell \in J_k} \mathcal{A}^{C_\ell}_{\omega - \omega_k}$ as the supports of ψ_k discussed in section 5.4.

TABLE 6.1
Notation.

n	the number of variables of a POP
d	the degree of a POP
sparse	cpu time in seconds consumed by the proposed sparse relaxation
dense	cpu time in seconds consumed by the dense relaxation [22]
cl.str	the structure of the maximal cliques
#clique	the average number of cliques found in randomly generated problems
#solved	the number of problems solved among randomly generated problems
max.cl	the number of the maximal cliques
max	the maximum of cpu time consumed by randomly generated problems
avr	the average of cpu time consumed by randomly generated problems
min	the minimum of cpu time consumed by randomly generated problems
cpu	cpu time in seconds
ω	the relaxation order

TABLE 6.2
Numerical results of the chained singular function and the Broyden banded function.

Chained singular function					Broyden banded function				
n	cl.str	$\epsilon_{\rm obj}$	sparse	dense	n	cl.str	$\epsilon_{\rm obj}$	sparse	dense
16	3*14	3.5e-7	0.6	3059.5	6	6*1	8.0e-9	11.3	11.6
40	3*38	8.4e-7	1.4	—	7	7*1	1.9e-8	69.5	69.5
100	3*98	5.5e-7	3.8	—	8	7*2	2.8e-8	164.1	373.7
200	3*198	3.0e-7	8.4	—	9	7*3	9.1e-8	240.3	1835.6
400	3*398	3.6e-7	19.3	—	10	7*4	6.2e-8	348.7	8399.4

Table 6.1 shows notation used in the description of numerical experiments in the following subsections. The notation "cl.str" indicates the structure of the maximal cliques obtained by applying MATLAB functions "symamd" and "chol" to the csp matrix. For example, $4*3+5*2$ means three cliques of size 4 and two cliques of size 5.

6.1. Unconstrained cases. The problems presented here are from the literature [3, 21, 24] and randomly generated problems. Table 6.2 displays the numerical results of the following two functions.

- The chained singular function [3]

$$f_{\rm cs}(\boldsymbol{x}) = \sum_{i\in J} \big((x_i + 10x_{i+1})^2 + 5(x_{i+2} - x_{i+3})^2 + (x_{i+1} - 2x_{i+2})^4 + 10(x_i - 10x_{i+3})^4\big),$$

  where $J = \{1, 3, 5, \ldots, n-3\}$ and n is a multiple of 4.
- The Broyden banded function [21]

$$f_{\rm Bb}(\boldsymbol{x}) = \sum_{i=1}^{n} \left(x_i(2 + 5x_i^2) + 1 - \sum_{j\in J_i} (1 + x_j)x_j \right)^2,$$

  where $J_i = \{j \mid j \neq i, \max(1, i-5) \leq j \leq \min(n, i+1)\}$.

The above two problems of relatively small size could be solved by the dense relaxation as shown in Table 6.2, and their results can be used for the comparison of the performance of the sparse and dense relaxations. In the case of the chained singular function $f_{\rm cs}$, its csp matrix $\boldsymbol{R}$ has nonzero elements near the diagonal; i.e.,

TABLE 6.3
Numerical results of Broyden tridiagonal function, the chained Wood function, and the generalized Rosenbrock function.

	Broyden tridiagonal			Chained Wood			Generalized Rosenbrock		
n	cl.str	$\epsilon_{\rm obj}$	cpu	cl.str	$\epsilon_{\rm obj}$	cpu	cl.str	$\epsilon_{\rm obj}$	cpu
600	3*598	9.1e-7	9.3	2*599	1.4e-5	0.9	2*599	3.9e-7	3.4
700	3*698	9.0e-7	10.9	2*699	1.6e-5	1.1	2*699	7.5e-9	4.0
800	3*798	2.2e-7	12.6	2*799	1.8e-5	1.3	2*799	2.1e-7	5.1
900	3*898	1.3e-7	14.4	2*899	3.4e-5	1.4	2*899	2.1e-7	5.6
1000	3*998	2.6e-7	16.0	2*999	3.8e-5	1.6	2*999	4.5e-7	5.9

$R_{ij} = 0$ if $|j - i| > 3$. This means that $f_{\rm cs}$ is correlatively sparse. The "cl.str" column of Table 6.2 shows that the sparsity can be detected correctly. As a result, the sparse relaxation is much more efficient than the dense relaxation. We could successfully solve the problem of 100 variables in a few seconds, while the dense relaxation could not handle the problem of 20 or 30 variables.

If we look at the result of the Broyden banded function $f_{\rm Bb}$ in Table 6.2, we observe that there is virtually no difference in performance between the proposed sparse and dense relaxations for $n = 6$ and $n = 7$. Because the csp matrix of this function has the bandwidth 7, it is fully dense when $n = 6$ and $n = 7$; the sparse relaxation is identical to the dense relaxation in these cases. As n increases, however, a sparse structure such as 7*2 for $n = 8$ can be found, and the sparse relaxation takes advantage of the structured sparsity, providing an optimal solution faster than the dense relaxation.

In Table 6.3, we present the numerical results of the following functions.

- The Broyden tridiagonal function [21]

$$f_{\rm Bt}(\boldsymbol{x}) = ((3-2x_1)x_1 - 2x_2 + 1)^2 + \sum_{i=2}^{n-1} ((3-2x_i)x_i - x_{i-1} - 2x_{i+1} + 1)^2 + ((3-2x_n)x_n - x_{n-1} + 1)^2.$$

- The chained Wood function [3]

$$f_{\rm cW}(\boldsymbol{x}) = 1 + \sum_{i\in J} \big(100(x_{i+1} - x_i^2)^2 + (1-x_i)^2 + 90(x_{i+3} - x_{i+2}^2)^2 + (1-x_{i+2})^2 + 10(x_{i+1} + x_{i+3} - 2)^2 + 0.1(x_{i+1} - x_{i+3})^2\big),$$

  where $J = \{1, 3, 5, \ldots, n-3\}$ and n is a multiple of 4.
- The generalized Rosenbrock function [24]

$$f_{\rm gR}(\boldsymbol{x}) = 1 + \sum_{i=2}^{n} \{100\left(x_i - x_{i-1}^2\right)^2 + (1-x_i)^2\}.$$

Each of the above three functions has a band structure in its csp matrix, and, therefore, the problems of large sizes can be handled efficiently. For example, the Broyden tridiagonal function $f_{\rm Bt}$ with 1000 variables could be solved in 16 seconds with the accuracy of 2.6e-07. Note that the solutions are accurate in all tested cases.

Next, we present the numerical results of randomly generated problems. The aim of the test using randomly generated problems is to observe the effects of increasing the

TABLE 6.4
Randomly generated polynomials with max.cl = 4 and 2d = 4.

n	#clique	max	avr	min	#solved
20	14.3	0.9	0.3	0.2	50/50
40	30.9	4.1	1.0	0.4	50/50
60	47.4	6.9	2.0	0.9	50/50
80	64.2	13.0	3.8	1.4	50/50
100	80.3	37.9	8.8	1.9	50/50

TABLE 6.5
Randomly generated polynomials with max.cl = 4 and n = 30.

$2d$	#clique	max	avr	min	#solved
4	22.7	1.1	0.6	0.3	50/50
6	22.9	18.9	5.1	1.4	50/50
8	22.7	624.2	74.7	7.9	50/50

number of variables, the degree of the polynomials, and the maximal size of cliques of the csp graph of a POP. The dense relaxation could not handle the randomly generated problems of the sizes reported here, and we include only the numerical results from the sparse relaxation.

Let us describe how an unconstrained problem with artificial correlative sparsity is generated randomly. We begin by constructing a chordal graph randomly such that the size of every maximal clique is not less than 2 and not greater than max.cl. From the chordal graph, we derive the set of maximal cliques $\{C_1, \ldots, C_\ell\}$ with $2 \leq |C_i| \leq$ max.cl $(i = 1, \ldots, \ell)$. We let $\boldsymbol{v}_{C_i}(\boldsymbol{x}) = (x_k^d \colon k \in C_i)$, where $2d$ is the degree of the polynomial, and generate a positive definite matrix $\boldsymbol{V}_i \in \mathcal{S}_{++}(C_i)$ and a vector $\boldsymbol{g}_i \in [-1, 1]^{|\mathcal{A}^{C_i}_{2d-1}|}$ $(i = 1, 2, \ldots, \ell)$ randomly such that the minimum eigenvalue σ of $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_\ell$ satisfies the relation

$$\sigma \geq \sum_{i=1}^{\ell} \left(\|\boldsymbol{g}_i\|_2 \sqrt{|\mathcal{A}^{C_i}_{2d-1}|} \right).$$

By using $\boldsymbol{V}_i$ and $\boldsymbol{g}_i$, we define the objective function:

$$f_{\rm rand}(\boldsymbol{x}) = \sum_{i=1}^{\ell} \left(\boldsymbol{v}_{C_i}(\boldsymbol{x})^T \boldsymbol{V}_i \boldsymbol{v}_{C_i}(\boldsymbol{x}) + \boldsymbol{g}_i^T \boldsymbol{u}\big(\boldsymbol{x}, \mathcal{A}^{C_i}_{2d-1}\big) \right).$$

This unconstrained POP is guaranteed to have an optimal solution in the compact set $\{\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n \mid \max_{i=1,\ldots,n} |x_i| \leq 1\}$. A scaling with the maximum of the absolute values of the coefficients of $f_{\rm rand}(\boldsymbol{x})$ is used in numerical experiments.

The numerical results are shown in Tables 6.4, 6.5, and 6.6. Table 6.4 exhibits how the sparse relaxation performs for a varying number of variables, Table 6.5 for raising the degree of the unconstrained problems, and Table 6.6 for increasing bounds of sizes of the cliques. For each choice of n, d, and max.cl, we generated 50 problems. Each column of #solved indicates the number of the problems whose optimal solutions were obtained with $\epsilon_{\rm obj} \leq 10^{-5}$ out of 50 problems. All problems tested were solved.

In Table 6.4, we notice that the number of cliques increases with n. For problems of large numbers of variables and cliques such as $n = 100$ and #clique= 80.3, the sparse relaxation provides optimal solutions in 8.8 average cpu seconds.

TABLE 6.6
Randomly generated polynomials with $2d = 4$ and $n = 30$.

max.cl	#clique	max	avr	min	#solved
4	22.7	1.1	0.6	0.3	50/50
6	20.0	31.5	6.4	1.3	50/50
8	17.3	497.9	79.8	4.0	50/50

The numerical results in Table 6.5 display the performance of the sparse relaxation for the problem of $n = 30$ with degrees up to 8. The maximum size of cliques is fixed to 4. As mentioned before, the size of the SDP relaxation of the POP of increasing degree becomes large rapidly even if the POP remains correlatively sparse. When $2d = 8$, the average cpu time is 74.7 and the maximum is 624.2.

A large size of cliques used during problem generation also increases the complexity of the problem as shown in Table 6.6. We tested with the maximum size of cliques 4, 6, and 8, and observe that cpu time to solve the corresponding problems grows very rapidly, e.g., 79.8 average cpu seconds and 497.9 maximum cpu seconds for max.cl = 8. From the increase of work measured by cpu time, we notice that the impact of the maximum size of cliques is comparable to that of degree, and bigger than that of the number of variables.

6.2. Constrained cases. In this subsection, we deal with the following constrained POPs:

- small-sized POPs from the literature [8],
- optimal control problems [2].

The numerical results on POPs from [8] are presented in Table 6.7. All problems are quadratic optimization problems except "alkyl," which involves polynomials of degree 3 in its equality constraints. We also added lower and upper bounds for the variables. In preliminary numerical experiments for some of the test problems, severe numerical difficulties occurred in badly scaled problems or problems with the complementarity condition. We incorporate all the techniques in sections 5.4, 5.5, and 5.6 into the dense and sparse relaxations for these problems.

In Table 6.7, ϵ'_{feas} denotes the feasibility for the scaled problems at the approximate optimal solutions obtained by the sparse and dense relaxations. We see that ϵ'_{feas} is small in most of the problems while the feasibility ϵ_{feas} for the original problems at the approximate optimal solutions becomes larger. The lower bounds obtained by the sparse relaxation are as good as the ones obtained by the dense relaxation except for the five problems ex5_2_2_cases1, 2, and 3, ex5_3_2, and ex9_1_4. In the first three cases, the dense relaxation with order $\omega = 2$ succeeds in computing accurate bounds while the sparse relaxation with order $\omega = 3$ computes accurate bounds with the same quality.

When we compare the performance of the sparse relaxation with the dense relaxation using these problems in Table 6.7, we observe that the sparse relaxation is much faster than the dense relaxation in large-dimensional problems. In some problems, however, the technique given in section 5.3 for reducing the sizes of relaxations worked so effectively that the difference between the dense and sparse relaxations decreased. For example, without incorporating this reduction technique, the sparse and dense relaxations of ex2_1_3 took 0.9 and 464.5 seconds, respectively, while 0.2 and 2.8 seconds, respectively, were consumed with the technique, as shown in Table 6.7. We will present more detailed comparison between the dense relaxation with the technique and the dense relaxation without the technique in section 6.3.

TABLE 6.7
Small-sized POPs. "h.ac" stands for "highly accurate" and means that the absolute value of the corresponding figure is less than 1.0e-9.

			Sparse				Dense		
Problem	n	ω	$\epsilon_{\rm obj}$	$\epsilon_{\rm feas}$	$\epsilon'_{\rm feas}$	cpu	$\epsilon_{\rm obj}$	$\epsilon'_{\rm feas}$	cpu
ex2_1_1	5	2	2e+0	h.ac	h.ac	0.1	2e+0	h.ac	0.1
ex2_1_1	5	3	3e-8	h.ac	h.ac	1.1	3e-8	h.ac	1.1
ex2_1_2	6	2	h.ac	h.ac	h.ac	0.1	h.ac	h.ac	0.2
ex2_1_3	13	2	h.ac	h.ac	h.ac	0.2	h.ac	h.ac	2.8
ex2_1_4	6	2	h.ac	h.ac	h.ac	0.1	h.ac	h.ac	0.1
ex2_1_5	10	2	h.ac	-1e-9	h.ac	1.2	h.ac	h.ac	1.2
ex2_1_8	24	2	h.ac	-1e-8	h.ac	82.8	h.ac	h.ac	419.3
ex3_1_1	8	2	9e-7	-2e+4	-6e-2	0.2	9e-7	-5e-2	0.9
ex3_1_1	8	3	9e-7	-1e-3	h.ac	3.3	9e-7	-3e-9	211.4
ex3_1_2	5	2	3e-8	h.ac	h.ac	0.2	9e-7	h.ac	0.2
ex5_2_2_case1	9	2	h.ac	-2e+1	-8e-2	1.0	h.ac	-2e-8	1.6
ex5_2_2_case1	9	3	h.ac	-7e-6	-5e-8	138.3	—	—	—
ex5_2_2_case2	9	2	h.ac	-7e+1	-3e-4	0.9	h.ac	-1e-7	1.4
ex5_2_2_case2	9	3	h.ac	-4e-4	-3e-5	131.9	—	—	—
ex5_2_2_case3	9	2	h.ac	-6e+1	-2e-1	0.8	h.ac	-2e-7	1.0
ex5_2_2_case3	9	3	h.ac	-3e-3	-1e-5	186.2	—	—	—
ex5_3_2	22	2	h.ac	-4e+0	-2e-1	24.4	h.ac	-6e-7	302.7
ex5_4_2	8	2	2e-6	-5e+5	-7e-1	0.3	2e-6	-7e-1	1.3
ex5_4_2	8	3	8e-7	-3e-2	-1e-7	4.0	8e-7	-3e-8	267.9
ex9_1_1	13	2	h.ac	-7e-9	-2e-9	0.7	h.ac	h.ac	2.7
ex9_1_2	10	2	h.ac	-5e-8	-3e-8	0.5	h.ac	h.ac	1.0
ex9_1_4	10	2	h.ac	-3e+0	-1e+0	1.4	h.ac	h.ac	1.2
ex9_1_4	10	3	h.ac	-6e+1	-1e+0	180.9	—	—	—
ex9_1_5	13	2	h.ac	-3e-6	-2e-6	0.7	h.ac	-1e-6	3.3
ex9_1_8	14	2	h.ac	-5e-9	-1e-9	0.3	h.ac	h.ac	1.9
ex9_2_1	10	2	2e-9	-9e-9	-3e-9	0.5	h.ac	h.ac	0.8
ex9_2_2	10	2	5e-6	-2e-5	-9e-6	0.5	5e-6	-8e-6	1.1
ex9_2_3	16	2	6e-4	-4e-2	-2e-2	0.8	6e-4	-1e-2	14.1
ex9_2_4	8	2	9e-6	-6e-8	-3e-8	0.2	9e-6	-1e-7	0.4
ex9_2_5	8	2	1e-9	-2e-9	h.ac	0.2	3e-9	-2e-9	0.3
ex9_2_6	16	2	2e-8	-2e-9	-1e-9	0.4	7e-4	-4e-5	2.4
ex9_2_7	10	2	h.ac	-3e-9	-1e-9	0.5	h.ac	h.ac	0.8
ex9_2_8	6	2	h.ac	h.ac	h.ac	0.1	h.ac	h.ac	0.1
alkyl	14	2	1e-2	-7e-1	-5e-2	1.8	7e-3	-4e-2	14.4
alkyl	14	3	6e-6	h.ac	h.ac	1923.1	—	—	—
st_bpaf1a	10	2	h.ac	-1e-8	-5e-9	0.6	h.ac	-2e-9	1.1
st_bpaf1b	10	2	h.ac	h.ac	h.ac	0.6	h.ac	h.ac	1.0
st_e05	5	2	1e-7	-4e-2	-2e-9	0.1	1e-7	h.ac	0.1
st_e07	10	2	h.ac	-9e-6	-1e-8	0.4	h.ac	-3e-9	1.5
st_jcbpaf2	10	2	h.ac	h.ac	h.ac	1.4	h.ac	h.ac	1.3

We present numerical results from discrete-time optimal control problems in [2]. The problem tested first (Problem 1 of [2]) is

$$\left.\begin{array}{ll} \min & \displaystyle\sum_{i=1}^{M-1}\left(\sum_{j=1}^{n_y}\left(y_{i,j}+\frac{1}{4}\right)^4+\sum_{j=1}^{n_x}\left(x_{i,j}+\frac{1}{4}\right)^4\right) \\ & \displaystyle+\sum_{j=1}^{n_y}\left(y_{M,j}+\frac{1}{4}\right)^4 \\ \text{subject to} & \boldsymbol{y}_{i+1}=\boldsymbol{A}\boldsymbol{y}_i+\boldsymbol{B}\boldsymbol{x}_i+(\boldsymbol{y}_i^T\boldsymbol{C}\boldsymbol{x}_i)\boldsymbol{e}\ (i=1,\ldots,M-1), \\ & \boldsymbol{y}_1=\boldsymbol{0},\ \boldsymbol{y}_i\in\mathbb{R}^{n_y}\ (i=1,\ldots,M), \\ & \boldsymbol{x}_i\in\mathbb{R}^{n_x}\ (i=1,\ldots,M-1), \end{array}\right\} \tag{6.1}$$

TABLE 6.8
Numerical results for the problem (6.1) *with* $(n_x, n_y, \mu) = (2, 4, 0)$. *The relaxation order* $\omega = 2$.

M	n	cl.str	ϵ_{obj}	ϵ_{feas}	cpu
6	30	4*1+5*4+6*2+7*3+8*6	2.8e-09	-1.6e-11	13.2
12	66	4*1+5*4+6*2+7*9+8*18	2.9e-09	-7.4e-12	40.9
18	102	4*1+5*4+6*2+7*15+8*30	5.7e-09	-9.4e-12	67.8
24	138	4*1+5*4+6*2+7*21+8*42	7.7e-09	-9.3e-12	95.6
30	174	4*1+5*4+6*2+7*27+8*54	9.5e-09	-9.1e-12	122.7

TABLE 6.9
Numerical results for the problem (6.1) *with* $(n_x, n_y, \mu) = (2, 4, 0.5)$. *The relaxation order* $\omega = 2$.

M	n	cl.str	ϵ_{obj}	ϵ_{feas}	cpu
6	30	5*2+7*4+10*3	3.4e-10	-1.0e-10	62.8
12	66	5*2+7*4+10*9	5.5e-09	-9.4e-10	168.8
18	102	5*2+7*4+10*15	8.1e-09	-9.7e-10	278.9
24	138	5*2+7*4+10*21	1.1e-08	-9.8e-10	390.2
30	174	5*2+7*4+10*27	1.3e-08	-9.9e-10	501.7

where $\boldsymbol{A} \in \mathbb{R}^{n_y \times n_y}$, $\boldsymbol{B} \in \mathbb{R}^{n_y \times n_x}$, and $\boldsymbol{C} \in \mathbb{R}^{n_y \times n_x}$ are given by

$$A_{i,j} = \begin{cases} 0.5 & \text{if } j = i, \\ 0.25 & \text{if } j = i+1, \\ -0.25 & \text{if } j = i-1, \\ 0 & \text{otherwise,} \end{cases} \qquad B_{i,j} = \frac{i-j}{n_y + n_x}, \quad C_{i,j} = \mu \frac{i+j}{n_y + n_x},$$

respectively. Here, $\boldsymbol{e}$ denotes a vector of ones in $\mathbb{R}^{n_y}$.

The numerical results of the problem (6.1) are shown in Tables 6.8, 6.9, and 6.10, which display the results of the problem (6.1) with $(n_x, n_y, \mu) = (2, 4, 0)$, the problem (6.1) with $(n_x, n_y, \mu) = (2, 4, 0.5)$ and the problem (6.1) with $(n_x, n_y, \mu) = (1, 2, 1)$, respectively. The values of n_x, n_y, and M determine the size of the problem, and μ is a parameter in $C_{i,j}$. The relaxation order 2 was used for all cases. As we increase M from 6 to 30, the number of variables becomes larger as indicated in the column of n. In all cases, the optimal solutions are obtained with good accuracy.

Depending on the choice of μ, it results in different clique structures and the size of the resulting SDP varies. This size can greatly affect the performance of the relaxations. When we take $\mu = 0$ in (6.1), the constraints are linear since $\boldsymbol{C} = \boldsymbol{O}$. Then, the cliques have smaller numbers of elements than the ones from the constraints with nonlinear terms, which enables the sparse relaxation to perform better in terms of cpu time. To see this, compare the column of cl.str of Table 6.8 with that of Table 6.9. For example, when $M = 30$ in Table 6.8, it took 122.7 cpu seconds to have an optimal solution, whereas it took 501.7 seconds in Table 6.9. Similarly, if we compare Tables 6.9 and 6.10, we notice that the size of the cliques in Table 6.10 is half the size of those in Table 6.9, while the cpu time of Table 6.10 is 1/100 of the cpu time of Table 6.9. This shows how much efficiency can be improved using appropriate clique structures.

In [2], the optimal values of the problem (6.1) with $\mu = 0, 0.5, 1$ and $n = 10, 50$ are shown as numerical results. We also solved the same problems with the sparse relaxation and obtained the same optimal values for each problem. Efficiency could not be compared because no cpu time was reported in [2].

TABLE 6.10
Numerical results for the problem (6.1) *with* $(n_x, n_y, \mu) = (1, 2, 1)$. *The relaxation order* $\omega = 2$.

M	n	cl.str	$\epsilon_{\rm obj}$	$\epsilon_{\rm feas}$	cpu
6	15	2*1+4*2+5*3	1.8e-10	-1.3e-10	0.4
12	33	2*1+4*2+5*9	1.1e-09	-3.8e-10	1.0
18	51	2*1+4*2+5*15	1.3e-09	-3.1e-10	1.7
24	69	2*1+4*2+5*21	7.6e-10	-1.6e-10	2.3
30	87	2*1+4*2+5*27	8.8e-10	-1.5e-10	3.0

TABLE 6.11
Numerical results from problem (6.2). *The relaxation order* $\omega = 1$.

M	n	cl.str	$\epsilon_{\rm obj}$	$\epsilon_{\rm feas}$	cpu
600	1198	2*1+3*598	3.5e-09	-2.1e-11	2.8
700	1398	2*1+3*698	3.4e-09	-1.8e-11	3.5
800	1598	2*1+3*798	2.5e-09	-1.2e-11	4.1
900	1798	2*1+3*898	2.1e-09	-8.8e-12	5.0
1000	1998	2*1+3*998	2.6e-09	-9.3e-12	5.5

The second problem (Problem 5 of [2]) is

$$\left.\begin{array}{ll} \text{min} & \dfrac{1}{M}\displaystyle\sum_{i=1}^{M-1}\left(y_i^2 + x_i^2\right) \\ \text{subject to} & y_{i+1} = y_i + \dfrac{1}{M}(y_i^2 - x_i) \quad (i = 1, \ldots, M-1), \quad y_1 = 1. \end{array}\right\} \tag{6.2}$$

Table 6.11 shows the results of (6.2) for various M.

From the column cl.str, we notice that the set of cliques has very few elements. The sparse relaxation can solve large-sized problems since they have plenty of correlative sparsity. In fact, the sparse relaxation provides optimal solutions for the problems with almost 2000 variables, where the size of the clique is 2 or 3.

From all numerical experiments in subsections 6.1 and 6.2, we have observed that the sparse relaxation is much faster than the dense relaxation with relatively accurate solutions. The sparse relaxation can handle large POPs with more than a hundred variables; this is impossible with the dense relaxation. The correlative sparsity has been the key to solve such large problems.

6.3. Performance comparison with some optimization software. In Table 6.12, we compare the performance of the sparse relaxation whose numerical results have been reported in the previous two subsections with optimization solvers LINGO (free version) [14], PENNON [16], LOQO [36], and GloptiPoly [9]. We have chosen the test problems that can show differences for the optimization solvers. We use the symbols ◯, △, and × to denote the accuracy of the optimal value obtained:

$$\begin{aligned} &\bigcirc\text{: } \frac{|\text{known best value} - \text{obtained value}|}{\max\{1, |\text{known best value}|\}} < \text{1.0e-5}, \\ &\triangle\text{: 1.0e-5} < \frac{|\text{known best value} - \text{obtained value}|}{\max\{1, |\text{known best value}|\}} < \text{1.0e-3}, \\ &\times\text{: otherwise.} \end{aligned}$$

LINGO is a global optimizer. The free version of LINGO was not able to handle large-sized problems; $n = 20$ was the largest size of the problems that it could handle. Although PENNON and LOQO are local optimizers, we include these solvers because they often attain global optimal values in the numerical experiments. We note that

TABLE 6.12

Comparison of LINGO, PENNON, LOQO, GloptiPoly, and the sparse relaxation. Bt and gR stand for Broyden tridiagonal and generalized Rosenbrock functions. The notation "na" is an abbreviation for "not applicable," and "-" means that "out of memory" resulted.

	n	LINGO	PENNON	LOQO	GloptiPoly	Sparse relax.
Bt	10	×	◯	◯	◯	◯
Bt	20	×	◯	◯	-	◯
Bt	40	na	◯	◯	-	◯
Bt	60	na	◯	×	-	◯
gR	10	◯	×	×	◯	△
gR	20	◯	×	×	-	△
gR	40	na	×	×	-	△
gR	60	na	×	×	-	△
ex2_1_1		×	×	×	◯	◯
ex5_2_2_case1		◯	◯	×	△	◯
ex5_2_2_case2		×	×	◯	◯	◯
ex5_2_2_case3		◯	◯	◯	◯	◯
ex9_2_5		◯	◯	×	◯	◯
st_e05		◯	×	×	×	◯
st_e07		×	◯	◯	◯	◯
st_bpaf1a		×	×	×	◯	◯
alkyl		◯	◯	◯	-	◯

TABLE 6.13

Comparison of our dense relaxation with reduction, without reduction, and GloptiPoly. "u.b" means that GloptiPoly returned the warning message that SeDuMi dual may be unbounded. "h.ac" stands for "highly accurate" and means that the absolute value of the corresponding figure is less than 1.0e-9.

	Dense w/reduction			Dense w/o reduction			GloptiPoly		
Problem	ϵ_{obj}	ϵ'_{feas}	cpu	ϵ_{obj}	ϵ'_{feas}	cpu	ϵ_{obj}	ϵ'_{feas}	cpu
ex2_1_1	4e-8	h.ac	1.1	2e-7	h.ac	5.7	9e-8	h.ac	2.7
ex5_2_2_case1	h.ac	-2e-8	1.5	h.ac	-1e-2	15.4	1e-5	-7e-4	7.0
ex5_2_2_case2	h.ac	-1e-7	1.4	h.ac	-1e-3	14.5	4e-7	-9e-5	9.7
ex5_2_2_case3	h.ac	-2e-7	1.0	h.ac	-4e-5	15.1	5e-7	-2e-5	7.9
ex9_2_5	3e-9	-2e-9	0.3	h.ac	h.ac	1.4	2e-9	h.ac	2.1
st_e05	h.ac	h.ac	0.2	h.ac	h.ac	0.3	u.b	—	2.7
st_e07	h.ac	-3e-9	1.6	h.ac	-1e-6	34.8	2e-9	h.ac	18.9
st_bpaf1a	h.ac	-2e-9	1.2	h.ac	-2e-9	38.5	h.ac	h.ac	21.1

× in a column of PENNON or LOQO does not necessarily mean that the corresponding local optimizer fails to "solve" the problem in its own context. GloptiPoly is a MATLAB implementation of the SDP relaxation method proposed by Lasserre [22] for solving POPs.

GloptiPoly needed larger memory as the number of variables increased for the Broyden tridiagonal and generalized Rosenbrock functions; hence, optimal values could not be obtained for $n \geq 20$. The problems ex5_2_2_case1 and st_e05 are badly scaled and could not be solved with GloptiPoly. We also note that the generalized Rosenbrock function has two distinct minimizers. As a result, the sparse relaxation with a perturbation resulted in less accurate approximate solutions. In contrast, GloptiPoly can obtain more accurate optimal solutions because of the special technique for detecting multiple optimal solutions developed by [10]. From Table 6.12, we observe that the proposed sparse relaxation provides more accurate optimal values than other software in most cases.

In Table 6.13, we compare our dense relaxation with that of GloptiPoly. Note that the dense relaxation without the reduction technique [19] stated in section 5.3 and

GloptiPoly are essentially the same relaxation; the difference lies in some additional techniques given in section 5 such as adding valid inequalities and scaling. Our dense relaxation was tested in two ways: with and without the reduction technique. In both cases, we did not perturb the objective function. The relaxation order $\omega = 2$ was used for all the problems, except ex2_2_1 where $\omega = 3$. It was shown that GloptiPoly is faster than the dense relaxation without the reduction in obtaining optimal values except for the problems ex9_2_5 and st_e05, and that the dense relaxation with the reduction is much faster than the others. The latter observation confirms that the reduction technique [19] is very effective in the dense relaxation.

7. Concluding discussions. The computational efficiency of the proposed sparse relaxations depends on the sparsity of a chordal extension of the csp graph. We note that the following two conditions are equivalent: (i) a chordal extension of the csp graph is sparse and (ii) Cholesky factorization of the Hessian matrix of the generalized Lagrangian function, or the Hessian matrix of the objective function in unconstrained problems, is sparse. When we compare the condition (ii) with the standard condition of traditional numerical methods, such as Newton's method for convex optimization, to be efficient for large-scale problems, we notice a difference between the generalized Lagrangian function and the usual Lagrangian function in their multipliers. SOS polynomials are the Lagrangian multipliers in the former whereas they are nonnegative real numbers in the latter. If a linear inequality constraint, the simplest polynomial constraint, is involved in a POP, it is multiplied by an SOS polynomial in the former. As a result, the Hessian matrix of the former can become denser than the Hessian matrix of the latter. In this sense, the condition (ii) in the proposed sparse relaxations is a stronger requirement on the sparsity in the POP than the standard condition for traditional numerical methods. This stronger requirement, however, can be justified if we understand the study of nonconvex and large-scale POPs in global optimization as a more complicated issue.

The proposed sparse relaxation for a correlatively sparse POP leads to an SDP that can maintain the sparsity for primal-dual interior-point methods. This is due to the fact that if a POP is correlatively sparse, the resulting SDP relaxation inherits the structured sparsity. In each iteration of a primal-dual interior-point method for solving an SDP, a system of linear equations, which is often called the Schur complement equation, is solved to compute a search direction. The coefficient matrix of this system is positive definite and fully dense in general. However, the sparse SDP relaxation of a correlatively sparse POP possesses sparsity in the coefficient matrix. This is an important advantage of our sparse relaxation. Among software packages implementing primal-dual interior-point methods, SeDuMi [33] handles SDPs with this sparsity in the coefficient matrix of the Schur complement equation while the current version of Semidefinite Programming Algorithm (SDPA) [37] developed by the authors' group is not equipped with the sparse Cholesky factorization for the Schur complement equation, showing slow performance for POPs with the correlative sparsity. This is the main reason that SeDuMi has been a choice for the numerical experiments instead of SDPA.

We encountered numerical difficulties during preliminary numerical experiments. The techniques presented in section 5 were very effective in overcoming the difficulties and in enhancing the performance of the sparse and dense relaxations. Some problems from [8], however, could not be solved because of numerical troubles resulting from SeDuMi. The failure has to be investigated more rigorously, but some SDPs generated as relaxations of POPs may be very difficult to solve. Additional techniques for

resolving this difficulty are to be developed, including the approach for formulating the SDP differently that was proposed in [23].

We mention that the proposed sparse SOS relaxation can be applied to the problems with rational objective functions in [13]. It is also interesting to extend the sparse SOS relaxation to optimization problems described with partially separable polynomial functions [34, 35]. These will be a future subject of study.

Acknowledgments. The authors would like to thank Phillipe L. Toint for discussion on a possible extension of the sparse SOS relaxation to partially separable optimization problems, and the anonymous referees for their valuable suggestions that considerably improved the presentation of this paper.

REFERENCES

[1] J. R. S. Blair and B. Peyton, *An introduction to chordal graphs and clique trees*, in Graph Theory and Sparse Matrix Computation, A. George, J. R. Gilbert, and J. W. H. Liu, eds., Springer-Verlag, New York, 1993, pp. 1–29.

[2] T. F. Coleman and A. Liao, *An efficient trust region method for unconstrained discrete-time optimal control problems*, Comput. Optim. Appl., 4 (1995), pp. 47–66.

[3] A. R. Conn, N. I. M. Gould, and Ph. L. Toint, *Testing a class of methods for solving minimization problems with simple bounds on the variables*, Math. Comp., 50 (1988), pp. 399–430.

[4] M. D. Choi, T. Y. Lam, and B. Reznick, *Sums of squares of real polynomials*, in K-Theory and Algebraic Geometry: Connections with Quadratic Forms and Division Algebras, Proc. Sympos. Pure Math., 58, AMS, Providence, RI, 1995, pp. 103–126.

[5] M. Fukuda, M. Kojima, K. Murota, and K. Nakata, *Exploiting sparsity in semidefinite programming via matrix completion* I: *General framework*, SIAM J. Optim., 11 (2000), pp. 647–674.

[6] D. R. Fulkerson and O. A. Gross, *Incidence matrices and interval graphs*, Pacific J. Math., 15 (1965), pp. 835–855.

[7] M. C. Golumbic, *Algorithmic Graph Theory and Perfect Graphs*, Academic Press, New York, 1980.

[8] *GLOBAL Library*, http://www.gamsworld.org/global/globallib.htm (2005).

[9] D. Henrion and J.-B. Lasserre, *GloptiPoly: Global optimization over polynomials with Matlab and SeDuMi*, ACM Trans. Math. Software, 29 (2003), pp. 165–194.

[10] D. Henrion and J.-B. Lasserre, *Detecting global optimality and extracting solutions in GloptiPoly*, in Positive Polynomials in Control, D. Henrion and A. Garulli, eds., Lecture Notes in Control and Inform. Sci. 312, Springer-Verlag, Berlin, 2005, pp. 293–310.

[11] D. Henrion and J.-B. Lasserre, *Convergent relaxations of polynomial matrix inequalities and static output feedback*, IEEE Trans. Automat. Control, 51 (2006), pp. 192–202.

[12] C. W. J. Hol and C. W. Scherer, *Sum of squares relaxations for polynomial semidefinite programming*, in Proceedings of the 16th International Symposium on Mathematical Theory of Networks and Systems (MTNS), Leuven, Belgium, 2004, pp. 1–10.

[13] D. Jibetean and E. de Klerk, *Global optimization of rational functions: A semidefinite programming approach*, Math. Program., 106 (2006), pp. 93–109.

[14] *LINGO*, http://www.lindo.com (2005).

[15] S. Kim, M. Kojima, and H. Waki, *Generalized Lagrangian duals and sums of squares relaxations of sparse polynomial optimization problems*, SIAM J. Optim., 15 (2005), pp. 697–719.

[16] M. Kočvara and M. Stingl, *PENNON: A code for convex nonlinear and semidefinite programming*, Optim. Methods Softw., 8 (2003), pp. 317–333.

[17] M. Kojima, *Sums of Squares Relaxations of Polynomial Semidefinite Programs*, Research Report B-397, Department of Mathematical and Computing Sciences, Tokyo Institute of Technology, Tokyo, 2003.

[18] M. Kojima, S. Kim, and H. Waki, *A general framework for convex relaxation of polynomial optimization problems over cones*, J. Oper. Res. Soc. Japan, 46 (2003), pp. 125–144.

[19] M. Kojima, S. Kim, and H. Waki, *Sparsity in sums of squares of polynomials*, Math. Program., 103 (2005), pp. 45–62.

[20] M. Kojima and M. Muramatsu, *An Extension of Sums of Squares Relaxations to Polynomial Optimization Problems over Symmetric Cones*, Research Report B-406, Department of Mathematical and Computing Sciences, Tokyo Institute of Technology, Tokyo, 2004.

[21] J. J. More, B. S. Garbow, and K. E. Hillstrom, *Testing unconstrained optimization software*, ACM Trans. Math. Software, 7 (1981), pp. 17–41.
[22] J. B. Lasserre, *Global optimization with polynomials and the problems of moments*, SIAM J. Optim., 11 (2001), pp. 796–817.
[23] J. Löfberg and P. A. Parrilo, *From coefficients to samples: A new approach to SOS optimization*, in Proceedings of the 43rd IEEE Conference on Decision and Control, 2004, pp. 3154–3159.
[24] S. G. Nash, *Newton-type minimization via the Lanczos method*, SIAM J. Numer. Anal., 21 (1984), pp. 770–788.
[25] P. A. Parrilo, *Semidefinite programming relaxations for semialgebraic problems*, Math. Program., 96 (2003), pp. 293–320.
[26] P. A. Parrilo and B. Sturmfels, *Minimizing polynomial functions*, in Algorithmic and Quantitative Real Algebraic Geometry, DIMACS Ser. Discrete Math. Theoret. Comput. Sci. 60, AMS, Providence, RI, 2003, pp. 83–99.
[27] S. Prajna, A. Papachristodoulou, and P. A. Parrilo, *SOSTOOLS: Sum of Squares Optimization Toolbox for MATLAB—User's Guide*, Control and Dynamical Systems, California Institute of Technology, Pasadena, CA, 2002; available online from http://www.mit.edu/∼parrilo/sostools/.
[28] M. Putinar, *Positive polynomials on compact semi-algebraic sets*, Indiana Univ. Math. J., 42 (1993), pp. 969–984.
[29] B. Reznick, *Extremal psd forms with few terms*, Duke Math. J., 45 (1978), pp. 363–374.
[30] R. Schneider, *Convex Bodies: The Brunn-Minkowski Theory*, Cambridge University Press, Cambridge, UK, 1993.
[31] M. Schweighofer, *Optimization of polynomials on compact semialgebraic sets*, SIAM J. Optim., 15 (2005), pp. 805–825.
[32] N. Z. Shor, *Dual quadratic estimates in polynomial and Boolean programming*, Ann. Oper. Res., 25 (1990), pp. 163–168.
[33] J. F. Sturm, *Using SeDuMi* 1.02*, a MATLAB toolbox for optimization over symmetric cones*, Optim. Methods Softw., 11/12 (1999), pp. 625–653.
[34] A. Griewank and Ph. L. Toint, *On the unconstrained optimization of partially separable functions*, in Nonlinear Optimization, 1981, M. J. D. Powell, ed., Academic Press, London, 1982, pp. 301–312.
[35] A. Griewank and Ph. L. Toint, *Partitioned variable metric updates for large structured optimization problems*, Numer. Math., 39 (1982), pp. 119–137.
[36] R. J. Vanderbei, *LOQO: An interior point code for quadratic programming*, Optim. Methods Softw., 11/12 (1999), pp. 451–484.
[37] M. Yamashita, K. Fujisawa, and M. Kojima, *Implementation and evaluation of SDPA* 6.0 *(Semidefinite Programming Algorithm* 6.0*)*, Optim. Methods Softw., 18 (2003), pp. 491–505.

SIAM J. OPTIM.
Vol. 17, No. 1, pp. 243–258
© 2006 Society for Industrial and Applied Mathematics

GENERALIZED LEVITIN–POLYAK WELL-POSEDNESS IN CONSTRAINED OPTIMIZATION*

X. X. HUANG† AND X. Q. YANG‡

Abstract. In this paper, we consider Levitin–Polyak-type well-posedness for a general constrained optimization problem. We introduce generalized Levitin–Polyak well-posedness and strongly generalized Levitin–Polyak well-posedness. Necessary and sufficient conditions for these types of well-posedness are given. Relations among these types of well-posedness are investigated. Finally, we consider convergence of a class of penalty methods and a class of augmented Lagrangian methods under the assumption of strongly generalized Levitin–Polyak well-posedness.

Key words. constrained optimization, generalized minimizing sequence, generalized Levitin–Polyak well-posedness, penalty-type methods

DOI. 10.1137/040614943

1. Introduction.

The study of well-posedness originates from Tykhonov [26] in dealing with unconstrained optimization problems. Its extension to the constrained case was developed by Levitin and Polyak [18]. Since then, various notions of well-posedness have been defined and extensively studied (see, e.g., [22, 6, 24, 28, 29, 9, 24, 30]). It is worth noting that recent research on well-posedness has been extended to vector optimization problems (see, e.g., [3, 20, 21, 12, 13, 7]).

Let (X, d_1) and (Y, d_2) be two metric spaces, and let $X_1 \subset X$ and $K \subset Y$ be two nonempty and closed sets. Consider the following constrained optimization problem:

$$\begin{aligned} (\mathrm{P}) \qquad & \min f(x) \\ & \text{s.t. } x \in X_1, \quad g(x) \in K, \end{aligned}$$

where $f : X \to R^1$ is a lower semicontinuous function and $g : X \to Y$ is a continuous function. Denote by X_0 the set of feasible solutions of (P), i.e.,

$$X_0 = \{x \in X_1 : g(x) \in K\}.$$

Denote by $\bar{X}$ and $\bar{v}$ the optimal solution set and the optimal value of (P), respectively. Throughout the paper, we always assume that $X_0 \neq \emptyset$ and $\bar{v} > -\infty$.

Let (Z, d) be a metric space and $Z_1 \subset Z$. We denote by $d_{Z_1}(z) = \inf\{d(z, z') : z' \in Z_1\}$ the distance from the point z to the set Z_1.

Levitin–Polyak (LP) well-posedness of (P) in the usual sense (when the optimal set of (P) is not necessarily a singleton) says that, for any sequence $\{x_n\} \subset X_1$ satisfying (i) $d_{X_0}(x_n) \to 0$ and (ii) $f(x_n) \to \bar{v}$, there exist a subsequence $\{x_{n_k}\}$ of $\{x_n\}$ and some $\bar{x} \in \bar{X}$ such that $x_{n_k} \to \bar{x}$.

*Received by the editors September 14, 2004; accepted for publication (in revised form) December 22, 2005; published electronically May 12, 2006. This work was supported by the Research Grants Council of Hong Kong (BQ-654), the National Science Foundation of China, and a small grant from Fudan University, China.

http://www.siam.org/journals/siopt/17-1/61494.html

†School of Management, Fudan University, Shanghai 200433, China, and Department of Mathematics and Computer Science, Chongqing Normal University, Chongqing 400047, China (xxhuang@fudan.edu.cn).

‡Department of Applied Mathematics, Hong Kong Polytechnic University, Kowloon, Hong Kong (mayangxq@polyu.edu.uk).

It should be noted that many optimization algorithms, such as penalty-type methods, e.g., penalty function methods and augmented Lagrangian methods, terminate when the constraint is approximately satisfied; i.e., $d_K(g(\bar{x})) \leq \epsilon$ for some $\epsilon > 0$ sufficiently small, and $\bar{x}$ is taken as an approximate solution of (P). These methods may generate sequences $\{x_n\} \subset X_1$ that satisfy $d_K(g(x_n)) \to 0$, not necessarily $d_{X_0}(x_n) \to 0$, as shown in the following simple example.

Example 1.1. Let $\alpha > 0$. Let $X = R^1$, $X_1 = R^1_+$, $K = R^1_-$, and

$$f(x) = \begin{cases} -x^\alpha & \text{if } x \in [0,1]; \\ -1/x^\alpha & \text{if } x \geq 1, \end{cases}$$

$$g(x) = \begin{cases} x & \text{if } x \in [0,1]; \\ 1/x^2 & \text{if } x \geq 1. \end{cases}$$

Consider the following penalty problem:

$$(PP_\alpha(n)) \quad \min_{x \in X_1} f(x) + n\left[\max\{0, g(x)\}\right]^\alpha, \quad n \in N.$$

It is easily verified that $x_n = 2^{1/\alpha} n^{1/\alpha}$ is the unique global solution to $(PP_\alpha(n))$ for each $n \in N$. Note that $X_0 = \{0\}$. It follows that we have $d_K(g(x_n)) = 1/(2^{2/\alpha} n^{2/\alpha}) \to 0$, while $d_{X_0}(x_n) = 2^{1/\alpha} n^{1/\alpha} \to +\infty$.

Thus, it is useful to consider sequences that satisfy $d_K(g(x_n)) \to 0$ instead of $d_{X_0}(x_n) \to 0$ as $n \to \infty$ in order to study convergence of penalty-type methods.

The sequence $\{x_n\}$ satisfying (i) and (ii) above is called an LP minimizing sequence. In what follows, we introduce two more types of generalized LP well-posedness.

DEFINITION 1.1. (P) *is called LP well-posedness in the generalized sense if, for any sequence* $\{x_n\} \subset X_1$ *satisfying* (i) $d_K(g(x_n)) \to 0$ *and* (ii) $f(x_n) \to \bar{v}$, *there exist a subsequence* $\{x_{n_k}\}$ *of* $\{x_n\}$ *and some* $\bar{x} \in \bar{X}$ *such that* $x_{n_k} \to \bar{x}$. *The sequence* $\{x_n\}$ *is called a generalized LP minimizing sequence.*

DEFINITION 1.2. (P) *is called LP well-posedness in the strongly generalized sense if, for any sequence* $\{x_n\} \subset X_1$ *satisfying* (i) $d_K(g(x_n)) \to 0$ *and* (ii) $\limsup_{n\to+\infty} f(x_n) \leq \bar{v}$, *there exist a subsequence* $\{x_{n_k}\}$ *of* $\{x_n\}$ *and some* $\bar{x} \in \bar{X}$ *such that* $x_{n_k} \to \bar{x}$. *The sequence* $\{x_n\}$ *is called a weakly generalized LP minimizing sequence.*

Remark 1.1. (i) The study of well-posedness for optimization problems with explicit constraints dates back to [17] when the abstract set X_1 does not appear. In [17], it was assumed that X is a Banach space and Y is a Banach space ordered by a closed and convex cone with some special properties; see [17] for details. What is worth emphasizing is that [17] studied only the case when (P) is a convex program. However, it is well known that penalty-type methods such as penalization methods and augmented Lagrangian methods are mostly developed for constrained nonconvex optimization problems. This is the main motivation of this paper.

(ii) The LP well-posedness in the strongly generalized sense defined above was called well-posedness in the strongly generalized sense in [17], while a weakly generalized LP minimizing sequence in the above definition is called a generalized minimizing sequence in [17].

(iii) It is obvious that LP well-posedness in the strongly generalized sense implies LP well-posedness in the generalized sense because a generalized LP minimizing sequence is a weakly generalized LP minimizing sequence.

(iv) If there exists some $\delta_0 > 0$ such that g is uniformly continuous on the set

$$\{x \in X_1 : d_{X_0}(x) \leq \delta_0\},$$

then it is not difficult to see that LP well-posedness in the generalized sense implies LP well-posedness.

(v) Any one type of (generalized) LP well-posedness defined above implies that the optimal set $\bar{X}$ of (P) is nonempty and compact.

The paper is organized as follows. In section 2, we investigate characterizations and criteria for the three types of (generalized) LP well-posednesses. In section 3, we establish relations among the three types of (generalized) LP well-posednesses. In section 4, we obtain convergence of a class of penalty methods and a class of augmented Lagrangian methods under the assumption of strongly generalized LP well-posedness.

2. Necessary and sufficient conditions for three types of (generalized) LP well-posedness. In this section, we present some criteria and characterizations for the three types of (generalized) LP well-posedness defined in section 1.

Consider the following statement:

(1)
$[\bar{X} \neq \emptyset$ and, for any LP minimizing sequence (resp., generalized LP minimizing sequence, weakly generalized LP minimizing sequence) $\{x_n\}$, we have $d_{\bar{X}}(x_n) \to 0]$.

The proof of the following proposition is elementary and thus omitted.

PROPOSITION 2.1. *If* (P) *is LP well-posed (resp., LP well-posed in the generalized sense and LP well-posed in the strongly generalized sense), then* (1) *holds. Conversely, if* (1) *holds and* $\bar{X}$ *is compact, then* (P) *is LP well-posed (resp., LP well-posed in the generalized sense and LP well-posed in the strongly generalized sense).*

Consider a real-valued function $c = c(t, s)$ defined for $t, s \geq 0$ sufficiently small, such that

$$c(t,s) \geq 0 \quad \forall t, s, \quad c(0,0) = 0, \tag{2}$$

$$s_k \to 0, t_k \geq 0, c(t_k, s_k) \to 0 \text{ imply } t_k \to 0. \tag{3}$$

THEOREM 2.1. *If* (P) *is LP well-posed, then there exists a function* c *satisfying* (2) *and* (3) *such that*

$$|f(x) - \bar{v}| \geq c(d_{\bar{X}}(x), d_{X_0}(x)) \quad \forall x \in X_1. \tag{4}$$

Conversely, suppose that $\bar{X}$ *is nonempty and compact, and* (4) *holds for some* c *satisfying* (2) *and* (3). *Then* (P) *is LP well-posed.*

Proof. Define

$$c(t,s) = \inf\{|f(x) - \bar{v}| : x \in X_1, d_{\bar{X}}(x) = t, d_{X_0}(x) = s\}.$$

It is obvious that $c(0,0) = 0$. Moreover, if $s_n \to 0$, $t_n \geq 0$ and $c(t_n, s_n) \to 0$, then there exists a sequence $\{x_n\} \subset X_1$ with

$$d_{\bar{X}}(x_n) = t_n, \tag{5}$$

$$d_{X_0}(x_n) = s_n \tag{6}$$

such that

$$|f(x_n) - \bar{v}| \to 0. \tag{7}$$

Note that $s_n \to 0$. Equations (6) and (7) jointly imply that $\{x_n\}$ is an LP minimizing sequence. By Proposition 2.1, we have $t_n \to 0$. This completes the proof of the first half of the theorem. Conversely, let $\{x_n\}$ be an LP minimizing sequence. Then, by (4), we have

$$|f(x_n) - \bar{v}| \geq c(d_{\bar{X}}(x_n), d_{X_0}(x_n)) \quad \forall x \in X_1. \tag{8}$$

Let

$$t_n = d_{\bar{X}}(x_n), \quad s_n = d_{X_0}(x_n).$$

Then $s_n \to 0$. In addition, $|f(x_n) - \bar{v}| \to 0$. These facts together with (8) as well as the properties of the function c imply that $t_n \to 0$. By Proposition 2.1, we see that (P) is LP well-posed. □

THEOREM 2.2. *If* (P) *is LP well-posed in the generalized sense, then there exists a function c satisfying* (2) *and* (3) *such that*

$$|f(x) - \bar{v}| \geq c(d_{\bar{X}}(x), d_K(g(x))) \quad \forall x \in X_1. \tag{9}$$

Conversely, suppose that $\bar{X}$ is nonempty and compact, and (9) *holds for some c satisfying* (2) *and* (3). *Then* (P) *is LP well-posed in the generalized sense.*

Proof. The proof is almost the same as that of Theorem 2.1. The only difference lies in the proof of the first part of Theorem 2.1. Here we define

$$c(t, s) = \inf\{|f(x) - \bar{v}| : x \in X_1, d_{\bar{X}}(x) = t, d_K(g(x)) = s\}. \qquad \square$$

Next we give a necessary and sufficient condition in the form of Furi and Vignoli [10] to characterize the LP well-posedness in the strongly generalized sense.

Let

$$\Omega(\epsilon) = \{x \in X_1 : f(x) \leq \bar{v} + \epsilon, d_K(g(x)) \leq \epsilon\}.$$

Let (X, d_1) be a complete metric space. Recall that the Kuratowski measure of noncompactness for a subset A of X is defined as

$$\alpha(A) = \inf\left\{\epsilon > 0 : A \subset \bigcup_{1 \leq i \leq n} C_i, \text{ for some } C_i, diam(C_i) \leq \epsilon\right\},$$

where $diam(C_i)$ is the diameter of C_i defined by

$$diam(C_i) = \sup\{d_1(x_1, x_2) : x_1, x_2 \in C_i\}.$$

The next theorem can be proved analogously to [17, Theorem 5.5].

THEOREM 2.3. *Let (X, d_1) be a complete metric space and f be bounded below on X_0. Then* (P) *is LP well-posed in the strongly generalized sense if and only if*

$$\alpha(\Omega(\epsilon))) \to 0 \text{ as } \epsilon \to 0.$$

DEFINITION 2.1. *Let Z be a topological space and $Z_1 \subset Z$ be nonempty. Suppose that $h : Z \to R^1 \cup \{+\infty\}$ is an extended real-valued function. h is said to be level-compact on Z_1 if, for any $s \in R^1$, the subset $\{z \in Z_1 : h(z) \leq s\}$ is compact.*

For any $\delta \geq 0$, define

$$X_1(\delta) = \{x \in X_1 : d_K(g(x)) \leq \delta\}. \tag{10}$$

The following proposition gives sufficient conditions that guarantee LP well-posedness in the strongly generalized sense.

PROPOSITION 2.2. *Let one of the following conditions hold.*

(i) *There exists $\delta_0 > 0$ such that $X_1(\delta_0)$ is compact.*

(ii) *f is level-compact on X_1.*

(iii) *X is a finite dimensional normed space and*

$$\lim_{x \in X_1, \|x\| \to +\infty} \max\{f(x), d_K(g(x))\} = +\infty. \tag{11}$$

(iv) *There exists $\delta_0 > 0$ such that f is level-compact on $X_1(\delta_0)$.*

Then (P) *is LP well-posed in the strongly generalized sense.*

Proof. Let $\{x_n\} \subset X_1$ be a weakly generalized LP minimizing sequence. Then

$$\limsup_{n \to +\infty} f(x_n) \leq \bar{v}, \tag{12}$$

$$d_K(g(x_n)) \to 0. \tag{13}$$

The proof of (i) is elementary. It is obvious that condition (ii) implies (iv). Now we show that (iii) implies (iv). Indeed, we need only to show that for any $s \in R^1$ and any $\delta > 0$, the set

$$A = \{x \in X_1(\delta) : f(x) \leq s\}$$

is bounded since X is a finite dimensional space. Suppose to the contrary that there exist $\delta > 0$, $s > 0$, and $\{x'_n\} \subset X_1(\delta)$ such that

$$\|x'_n\| \to +\infty \text{ and } f(x'_n) \leq s.$$

By $\{x'_n\} \subset X_1(\delta)$, we have $\{x'_n\} \subset X_1$ and

$$d_K(g(x'_n)) \leq \delta.$$

As a result,

$$\max\{f(x'_n), d_K(g(x'_n))\} \leq \max\{s, \delta\},$$

contradicting (11).

Thus, we need only to prove that if (iv) holds, then (P) is LP well-posed in the strongly generalized sense. By (13), it is apparent that we can assume without loss of generality that $\{x_n\} \subset X_1(\delta_0)$. By (12), we can assume without loss of generality that

$$\{x_n\} \subset \{x \in X_1 : f(x) \leq \bar{v} + 1\}.$$

By the level-compactness of f on $X_1(\delta_0)$, we deduce that there exist a subsequence $\{x_{n_k}\}$ of $\{x_n\}$ and $\bar{x} \in X_1$ such that $x_{n_k} \to \bar{x}$. It is obvious from (13) that $\bar{x} \in X_0$. Furthermore, from (12), we deduce that $f(\bar{x}) \leq \bar{v}$. So we have $f(\bar{x}) = \bar{v}$. That is, $\bar{x} \in \bar{X}$. Hence, (P) is LP well-posed in the strongly generalized sense. ∎

Now we consider the case when Y is a normed space and K is a closed and convex cone with nonempty interior $\text{int}K$. Arbitrarily fix an $e \in \text{int}K$. Let $t \geq 0$ and consider the following perturbed problem of (P):

$$(P_t) \qquad \min f(x)$$
$$\text{s.t. } x \in X_1, \quad g(x) \in K - te. \tag{14}$$

Let

$$X_2(t) = \{x \in X_1 : g(x) \in K - te\}. \tag{15}$$

PROPOSITION 2.3. *Let one of the following conditions hold.*
(i) *There exists $t_0 > 0$ such that $X_2(t_0)$ is compact.*
(ii) *f is level-compact on X_1.*
(iii) *X is a finite dimensional normed space and*

$$\lim_{x \in X_1, \|x\| \to +\infty} \max\{f(x), d_K(g(x))\} = +\infty.$$

(iv) *There exists $t_0 > 0$ such that f is level-compact on $X_2(t_0)$.*
Then (P) *is LP well-posed in the strongly generalized sense.*
Proof. The proof is similar to that of Proposition 2.2. □
Now we make the following assumption.

ASSUMPTION 2.1. *X is a finite dimensional normed space, Y is a normed space, $X_1 \subset X$ is a nonempty, closed, and convex set, $K \subset Y$ is a closed, and convex cone with nonempty interior* $\text{int}K$ *and $e \in \text{int}K$, f and g are continuous on X_1, f is a convex function on X_1, and g is K-concave on X_1 (namely, for any $x_1, x_2 \in X_1$ and any $\theta \in (0,1)$, there holds that $g(\theta x_1 + (1-\theta)x_2) - \theta g(x_1) - (1-\theta)g(x_2) \in K$).*

It is obvious that under Assumption 2.1, (P) is a convex program.
The next lemma can be proved similarly to that of [16, Proposition 2.4].

LEMMA 2.1. *Let Assumption* 2.1 *hold. Then the following two statements are equivalent.*
(i) *The optimal set $\bar{X}$ of* (P) *is nonempty and compact.*
(ii) *For any $t \geq 0$, f is level-compact on the set $X_2(t)$.*

THEOREM 2.4. *Let Assumption* 2.1 *hold. Then* (P) *is LP well-posed in the strongly generalized sense if and only if the optimal set $\bar{X}$ of* (P) *is nonempty and compact.*

Proof. The sufficiency part follows directly from Lemma 2.1 and Proposition 2.3, while the necessity part is obvious by Remark 1.1. □
The next two lemmas will be used to derive Theorem 2.5.

LEMMA 2.2 (see [1]). *Let (Z, d) be a complete metric space and $h : Z \to R^1 \cup \{+\infty\}$ be lower semicontinuous and bounded below. Let $\epsilon > 0$. Suppose that $z_0 \in Z$ satisfies $h(z_0) \leq \inf\{h(z) : z \in Z\} + \epsilon$. Then there exists $z_\epsilon \in Z$ such that*
(i) *$h(z_\epsilon) \leq h(z_0)$;*
(ii) *$d(z_\epsilon, z_0) \leq \sqrt{\epsilon}$;*
(iii) *$h(z_\epsilon) < h(z) + \sqrt{\epsilon} d(z, z_\epsilon)$ $\forall z \in Z \backslash \{z_\epsilon\}$.*

LEMMA 2.3. *Let Y be a normed space and $K \subset Y$ be a closed and convex cone with* $\text{int}K \neq \emptyset$ *and $e \in \text{int}K$. Suppose that $\{y_n\} \subset Y$. Then $d_K(y_n) \to 0$ if and only if there exists a sequence $\{t_n\} \subset R^1_+$ with $t_n \to 0$ such that $y_n \in K - t_n e$.*

Proof. For the necessity part, from $d_K(y_n) \to 0$, we have $\{u_n\} \subset K$ such that $\|y_n - u_n\| \to 0$. Let $y'_n = y_n - u_n$. Then $\|y'_n\| \to 0$. Let $t_n = \sqrt{\|y'_n\|}$. Then $\{t_n\} \subset R^1_+$,

$t_n \to 0$ and $y_n'/t_n \to 0$. Since $e \in \text{int}K$, it follows that $e + y_n'/t_n \in K$ when n is sufficiently large. Consequently, $y_n' \in K - t_n e$. Hence, $y_n = u_n + y_n' \in K - t_n e$.

For the sufficiency part, as $y_n \in K - t_n e$, we have $y_n + t_n e \in K$. Thus,

$$d_K(y_n) \le \|y_n - (y_n + t_n e)\| = t_n \|e\|.$$

Hence, $d_K(y_n) \to 0$. □

Suppose that K is a cone. We denote by K^* the positive polar cone of K, i.e.,

$$K^* = \{\mu \in Y^* : \mu(u) \ge 0 \ \forall u \in K\}.$$

THEOREM 2.5. *Assume that X is a Banach space, Y is a normed space, and $X_1 \subset X$ is nonempty, closed, and convex. $K \subset Y$ is a closed and convex cone with $\text{int}K \ne \emptyset$ and $e \in \text{int}K$. Suppose that $f : X \to R^1$ is convex and continuously differentiable on X_1 and $g : X \to Y$ is K-concave and continuously differentiable on X_1. Let Slater constraint qualification for* (P) *hold: there exists $x_0 \in X_1$ such that $g(x_0) \in \text{int}K$. Assume that the optimal set $\bar{X}$ of* (P) *is nonempty. Further assume that there exists a convergent subsequence of $\{x_n\}$ for any sequences $\{x_n\} \subset X_1$ and $\{\mu_n\} \subset K^*$ satisfying the following.*

(i) $\lim_{n \to +\infty} d_K(g(x_n)) = 0$.

(ii) *There exists a subsequence $\{\mu_{n_k}\}$ such that $\mu_{n_k} = 0 \ \forall k$ or $\lim_{n \to +\infty} \mu_n(g(x_n))/\|\mu_n\| = 0$.*

(iii) $\lim_{n \to +\infty} d_{(-N_{X_1}(x_n))}(\bigtriangledown f(x_n) - \mu_n(\bigtriangledown g(x_n))) = 0$, *where $N_{X_1}(x_n)$ is the normal cone of X_1 at x_n.*

Then, (P) is LP well-posed in the strongly generalized sense.

Proof. Suppose that $\bar{x} \in \bar{X}$. Since Slater constraint qualification holds, we have $\bar{\mu} \in K^*$ such that

$$f(\bar{x}) \le f(x) - \bar{\mu}(g(x)) \quad \forall x \in X_1 \tag{16}$$

and

$$\bar{\mu}(g(\bar{x})) = 0. \tag{17}$$

Let $\{x_n\} \subset X_1$ be a weakly generalized LP minimizing sequence for (P). Then, by Lemma 2.3,

$$\limsup_{n \to +\infty} f(x_n) \le \bar{v} \tag{18}$$

and

$$g(x_n) \in K - t_n e \tag{19}$$

for some $\{t_n\} \subset R^1_+$ with $t_n \to 0$. From (16), we have

$$f(\bar{x}) \le f(x) - \bar{\mu}(g(x)) \quad \forall x \in X_2(t_n).$$

Note that

$$-\bar{\mu}(g(x)) \le t_n \bar{\mu}(e) \quad \forall x \in X_2(t_n).$$

Thus,

$$f(\bar{x}) \le f(x) + t_n \bar{\mu}(e) \quad \forall x \in X_2(t_n). \tag{20}$$

Hence,

$$\inf_{x\in X_2(t_n)} f(x) > -\infty. \tag{21}$$

The combination of (19) and (20) gives

$$f(\bar{x}) \le f(x_n) + t_n\bar{\mu}(e).$$

Consequently,

$$f(\bar{x}) \le \liminf_{n\to+\infty} f(x_n).$$

This together with (18) yields

$$\lim_{n\to+\infty} f(x_n) = f(\bar{x}). \tag{22}$$

This combined with (20) implies that there exists $\epsilon_n \to 0^+$ such that

$$f(x_n) \le f(x) + \epsilon_n \quad \forall x \in X_2(t_n).$$

Note that $X_2(t_n) \subset X$ is nonempty and closed. $(X_2(t_n), \|\cdot\|)$ can be seen as a complete (metric) subspace of X. Applying Lemma 2.2, we obtain

$$x'_n \in X_2(t_n) \tag{23}$$

such that

$$\|x_n - x'_n\| \le \sqrt{\epsilon_n} \tag{24}$$

and

$$f(x'_n) \le f(x) + \sqrt{\epsilon_n}\|x - x'_n\| \quad \forall x \in X_2(t_n). \tag{25}$$

Note that Slater constraint qualification also holds for the following constrained optimization problem:

$$\begin{aligned} (P_n) \qquad & \min\ f(x) + \sqrt{\epsilon_n}\|x - x'_n\| \\ & \text{s.t. } x \in X_1, \quad g(x) \in K - t_n e, \end{aligned}$$

and by (25), x'_n is an optimal solution of (P_n). Hence, there exists $\mu_n \in K^*$ such that

$$0 \in \bigtriangledown f(x'_n) - \mu_n(\bigtriangledown g(x'_n)) + \sqrt{\epsilon_n}B^* + N_{X_1}(x'_n) \tag{26}$$

and

$$\mu_n(g(x'_n) + t_n e) = \mu_n(g(x'_n)) + t_n\mu_n(e) = 0, \tag{27}$$

where B^* is the closed unit ball of X^*. Equation (26) implies that

$$\lim_{n\to+\infty} d_{(-N_{X_1}(x'_n))}(\bigtriangledown f(x'_n) - \mu_n(\bigtriangledown g(x'_n))) = 0. \tag{28}$$

From (27), we see that if there does not exist a subsequence $\{\mu_{n_k}\}$ such that $\mu_{n_k} = 0\ \forall k$, then

$$\lim_{n\to+\infty} \mu_n(g(x_n))/\|\mu_n\| = 0. \tag{29}$$

The combination of (24), (28), and (29) implies that $\{x'_n\}$ and $\{\mu_n\}$ satisfy conditions (i)–(iii) of the theorem. Thus, $\{x'_n\}$ has a subsequence $\{x'_{n_k}\}$ which converges to some $\bar{x}' \in X_0$. From (24), we deduce that $x_{n_k} \to \bar{x}' \in X_0$. This combined with (22) implies $\bar{x}' \in \bar{X}$. Hence, (P) is LP well-posed in the strongly generalized sense. □

Remark 2.1. Conditions (i)–(iii) of Theorem 2.5 can be seen as the well-known Palais–Smale condition (C) [1] in the case of constrained optimization.

3. Relations among three types of (generalized) LP well-posedness. Simple relationships among the three types of LP well-posedness were mentioned in Remark 1.1. Now we investigate further relationships among them.

The proof of next theorem is elementary and is omitted.

THEOREM 3.1. *Suppose that there exist $\delta > 0$, $\alpha > 0$, and $c > 0$ such that*

$$d_{X_0}(x) \le c d_K^{\alpha}(g(x)) \quad \forall x \in X_1(\delta), \tag{30}$$

where $X_1(\delta)$ is defined by (10). *If* (P) *is LP well-posed, then* (P) *is LP well-posed in the generalized sense.*

Remark 3.1. Equation (30) is an error bound condition for the set X_0 in terms of the residual function

$$r(x) = d_K(g(x)) \quad \forall x \in X_1.$$

When $X = R^l, Y = R^m$, $X_1 = X$, and $X_0 \neq \emptyset$, by Theorem 5 of [23], (30) holds if and only if, for any $y \in R^m$ with $\|y\| \le \delta$,

$$\Psi(y) \subset \Psi(0) + c\|y\|^{\alpha} B,$$

where

$$\Psi(y) = \{x \in R^l : g(x) \in K + y\}, \quad y \in R^m,$$

and B is the closed unit ball of Y. Sufficient conditions guaranteeing (30) were given in numerous papers on error bounds for systems of inequalities and metric regularity of set-valued maps (when (30) holds locally with $\alpha = 1$) in finite and infinite dimensional spaces (see, e.g., [5, 8, 18] and the references therein).

DEFINITION 3.1 (see [4]). *Let W be a topological space and $F : W \to 2^X$ be a set-valued map. F is said to be upper Hausdorff semicontinuous (u.H.c.) at $w \in W$ if, for any $\epsilon > 0$, there exists a neighborhood U of w such that $F(U) \subset B(F(w), \epsilon)$, where, for $Z \subset X$ and $r > 0$,*

$$B(Z, r) = \{x \in X : d_Z(x) \le r\}.$$

DEFINITION 3.2 (see [1]). *Let W be a topological space and $F : W \to 2^X$ be a set-valued map. F is said to be upper semicontinuous (u.s.c.) in the Berge's sense at $w \in W$ if, for any neighborhood Ω of $F(w)$, there exists a neighborhood U of w such that $F(U) \subset \Omega$.*

It is obvious that the notion of u.s.c. (in Berge's sense) is stronger than u.H.c.

Clearly, $X_1(\delta)$ given by (10) can be seen as a set-valued map from R^1_+ to X. The next two theorems use conditions similar to those for the general stability results presented in section 3 of [4], where the uniform continuity of the objective function around the feasible set and the u.H.c. of the perturbation set-valued map were considered.

THEOREM 3.2. *Assume that the set-valued map $X_1(\delta)$ defined by* (10) *is u.H.c. at $0 \in R^1_+$. If* (P) *is LP well-posed, then* (P) *is LP well-posed in the generalized sense.*

Proof. Let $\{x_n\} \subset X_1$ be a generalized LP minimizing sequence. That is,

$$f(x_n) \to \bar{v}, \tag{31}$$

$$d_K(g(x_n)) \to 0. \tag{32}$$

Equation (32), together with the u.H.c. of $X_1(\delta)$ at 0, implies that $d_{X_0}(x_n) \to 0$. This fact combined with (31) implies that $\{x_n\}$ is an LP minimizing sequence. Thus,

there exist a subsequence $\{x_{n_k}\}$ of $\{x_n\}$ and some $\bar{x} \in \bar{X}$ such that $x_{n_k} \to \bar{x}$. Hence, (P) is LP well-posed in the generalized sense. □

THEOREM 3.3. *Assume that there exists $\epsilon_0 > 0$ such that f is uniformly continuous on $B(X_0, \epsilon_0)$ and the set-valued map $X_1(\delta)$ is u.H.c. at 0. If (P) is LP well-posed, then it is LP well-posed in the strongly generalized sense.*

Proof. Let $\{x_n\}$ be a weakly generalized LP minimizing sequence. That is,

$$\limsup_{n\to+\infty} f(x_n) \le \bar{v}, \tag{33}$$

$$d_K(g(x_n)) \to 0. \tag{34}$$

Note that $X_1(\delta)$ is u.H.c. at 0. This fact together with (34) implies that $d_{X_0}(x_n) \to 0$. Note that f is uniformly continuous on $B(X_0, \epsilon_0)$. It follows that

$$\liminf_{n\to+\infty} f(x_n) \ge \bar{v}. \tag{35}$$

The combination of (33) and (35) yields that

$$f(x_n) \to \bar{v}.$$

Hence, $\{x_n\}$ is an LP minimizing sequence. Thus, there exist a subsequence $\{x_{n_k}\}$ of $\{x_n\}$ and some $\bar{x} \in \bar{X}$ such that $x_{n_k} \to \bar{x}$. So, (P) is LP well-posed in the strongly generalized sense. □

Let $\delta \ge 0$. Consider the perturbed problem of (P):

$$(P_\delta) \qquad \begin{aligned} &\min\ f(x) \\ &\text{s.t. } x \in X_1, \quad d_K(g(x)) \le \delta. \end{aligned}$$

Denote by $v_1(\delta)$ the optimal value of (P_δ). Clearly, $v_1(0) = \bar{v}$.

THEOREM 3.4. *Consider problems* (P) *and* (P_δ). *Suppose that* (P) *is LP well-posed in the generalized sense and*

$$\liminf_{\delta\to 0^+} v_1(\delta) = \bar{v}. \tag{36}$$

Then (P) *is LP well-posed in the strongly generalized sense.*

Proof. Let $\{x_n\} \subset X_1$ be a weakly generalized LP minimizing sequence. Then

$$\limsup_{n\to+\infty} f(x_n) \le \bar{v} \tag{37}$$

and

$$\lim_{n\to+\infty} d_K(g(x_n)) = 0.$$

Let $\delta_n = d_K(g(x_n))$. Then x_n is feasible for (P_{δ_n}). Thus,

$$v_1(\delta_n) \le f(x_n).$$

Passing to the lower limit, we get

$$\liminf_{n\to+\infty} v_1(\delta_n) \le \liminf_{n\to+\infty} f(x_n).$$

This together with (37) and (36) yields

$$\lim_{n\to+\infty} f(x_n) = \bar{v}.$$

It follows that $\{x_n\}$ is a generalized LP minimizing sequence. Thus, there exist a subsequence $\{x_{n_k}\}$ of $\{x_n\}$ and some $\bar{x} \in \bar{X}$ such that $x_{n_k} \to \bar{x}$. So, (P) is LP well-posed in the strongly generalized sense. □

Remark 3.2. If the set-valued map $X_1(\delta)$ defined by (10) is u.s.c. at $0 \in R^1_+$, by Theorem 4.2.3 (1) of [2], (36) holds. In this case, the generalized LP well-posedness of (P) implies the strongly generalized LP well-posedness of (P).

Now let Y be a normed space and $y \in Y$. Consider the following perturbed problem of (P):

$$(P_y) \qquad \begin{aligned} &\min f(x) \\ &\text{s.t. } x \in X_1, \quad g(x) \in K + y. \end{aligned}$$

Denote by

$$X_3(y) = \{x \in X_1 : g(x) \in K + y\} \tag{38}$$

the feasible set of (P_y) and $v_3(y)$ the optimal value of (P_y). Here we note that if $X_3(y) = \emptyset$, we set $v_3(y) = +\infty$. It is obvious that $X_3(y)$ can be seen as a set-valued map from Y to X. Corresponding to Theorems 3.2–3.4, respectively, we have the following theorems.

THEOREM 3.5. *Assume that Y is a normed space and that the set-valued map $X_3(y)$ is u.H.c. at $0 \in Y$. If* (P) *is LP well-posed, then* (P) *is LP well-posed in the generalized sense.*

THEOREM 3.6. *Assume that Y is a normed space and that there exists $\epsilon_0 > 0$ such that f is uniformly continuous on $B(X_0, \epsilon_0)$ and the set-valued map $X_3(y)$ is u.H.c. at $0 \in Y$. If* (P) *is LP well-posed, then it is LP well-posed in the strongly generalized sense.*

THEOREM 3.7. *Assume that Y is a normed space. Consider problems* (P) *and* (P_y). *Suppose that* (P) *is LP well-posed in the generalized sense and*

$$\liminf_{y\to 0} v_3(y) = \bar{v}. \tag{39}$$

Then (P) *is LP well-posed in the strongly generalized sense.*

Similar to Remark 3.2, when the set-valued map X_3 is u.s.c. at $0 \in Y$, then (39) holds. Thus, the generalized LP well-posedness of (P) implies its strongly generalized LP well-posedness.

In the special case when K is a closed and convex cone with nonempty interior $\mathrm{int}K$, arbitrarily fix an $e \in \mathrm{int}K$. It is obvious that $X_2(t)$ defined by (15) can be seen as a set-valued map from R^1_+ to X. Denote by $v_2(t)$ the optimal value of (P_t).

THEOREM 3.8. *Assume that K is a closed and convex cone with nonempty interior* $\mathrm{int}K$ *and that the set-valued map $X_2(t)$ is u.H.c. at $0 \in R^1_+$. If* (P) *is LP well-posed, then* (P) *is LP well-posed in the generalized sense.*

THEOREM 3.9. *Assume that K is a closed and convex cone with nonempty interior* $\mathrm{int}K$ *and that there exists $\epsilon_0 > 0$ such that f is uniformly continuous on $B(X_0, \epsilon_0)$ and the set-valued map $X_2(t)$ is u.H.c. at $0 \in R^1_+$. If* (P) *is LP well-posed, then it is LP well-posed in the strongly generalized sense.*

THEOREM 3.10. *Assume that K is a closed and convex cone with nonempty interior* $\mathrm{int}K$. *Consider problems* (P) *and* (P_t). *Suppose that* (P) *is LP well-posed in the generalized sense and*

$$\liminf_{t\to 0^+} v_2(t) = \bar{v}. \tag{40}$$

Then (P) *is LP well-posed in the strongly generalized sense.*

Again, as noted in Remark 3.2, when the set-valued map X_2 is u.s.c. at $0 \in R^1_+$, then (39) holds. Thus, the generalized LP well-posedness of (P) implies its strongly generalized LP well-posedness.

4. Applications to penalty-type methods. In this section, we consider the convergence of a class of penalty methods and a class of augmented Lagrangian methods under the assumption of strongly generalized LP well-posedness of (P).

4.1. Penalty methods. Let $\alpha > 0$. Consider the following penalty problem:

$$(PP_\alpha(r)) \quad \min_{x\in X_1} f(x) + r d_K^\alpha(g(x)), \quad r > 0.$$

Denote by $v_4(r)$ the optimal value of $(PP_\alpha(r))$. It is clear that

$$v_4(r) \le \bar{v} \quad \forall r > 0. \tag{41}$$

Remark 4.1. When $\alpha \in (0,1)$, $X = R^l$, $Y = R^m$, $K = R^{m_1}_- \times \{0_{m-m_1}\}$, where $m \ge m_1$ and 0_{m-m_1} is the origin of the space R^{m-m_1}, this class of penalty functions was applied to the study of mathematical programs with equilibrium constraints [19]. Necessary and sufficient conditions for the exact penalization of this class of penalty functions were derived in [14]. This class of penalty methods was also applied to mathematical programs with complementarity constraints [27] and nonlinear semidefinite programs [15]. An important advantage of this class of penalty methods is that it requires weaker conditions to guarantee its exact penalization property than the usual l_1 penalty function method (see [19]).

THEOREM 4.1. *Let* $0 < r_n \to +\infty$. *Consider problems* (P) *and* $(PP_\alpha(r_n))$. *Assume that there exist* $\bar{r} > 0$ *and* $m_0 \in R^1$ *such that*

$$f(x) + \bar{r} d_K^\alpha(g(x)) \ge m_0 \quad \forall x \in X_1. \tag{42}$$

Let $0 < \epsilon_n \to 0$. Suppose that each $x_n \in X_1$ satisfies

$$f(x_n) + r_n d_K^\alpha(g(x_n)) \le v_4(r_n) + \epsilon_n. \tag{43}$$

Further assume that (P) is LP well-posed in the strongly generalized sense. Then there exist a subsequence $\{x_{n_k}\}$ of $\{x_n\}$ and some $\bar{x} \in \bar{X}$ such that $x_{n_k} \to \bar{x}$.

Proof. From (41) and (43), we have

$$f(x_n) \le \bar{v} + \epsilon_n.$$

Thus,

$$\limsup_{n\to+\infty} f(x_n) \le \bar{v}. \tag{44}$$

Moreover, from (41)–(43), we deduce that

$$f(x_n) + \bar{r} d_K^\alpha(g(x_n)) + (r_n - \bar{r}) d_K^\alpha(g(x_n)) \le \bar{v} + \epsilon_n.$$

Thus,

$$m_0 + (r_n - \bar{r}) d_K^{\alpha}(g(x_n)) \leq \bar{v} + \epsilon_n,$$

implying

$$d_K(g(x_n)) \leq \left[\frac{\bar{v} + \epsilon_n - m_0}{r_n - \bar{r}} \right]^{1/\alpha}.$$

Passing to the limit, we get

$$\lim_{n \to +\infty} d_K(g(x_n)) = 0. \tag{45}$$

It follows from (44) and (45) that $\{x_n\}$ is a weakly generalized LP minimizing sequence. Hence, there exist a subsequence $\{x_{n_k}\}$ of $\{x_n\}$ and some $\bar{x} \in \bar{X}$ such that $x_{n_k} \to \bar{x}$. □

4.2. Augmented Lagrangian methods. Let (X, d_1) be a metric space, let $Y = R^m$, and let $K \subset Y$ be a nonempty, closed, and convex set. Let $\sigma : R^m \to R^1 \cup \{+\infty\}$ be an augmenting function; namely, it is a lower semicontinuous, convex function satisfying

$$\min_{y \in R^m} \sigma(y) = 0 \text{ and } \sigma \text{ attains its unique minimum at } y = 0.$$

Following Example 11.46 in [25], we define the dualizing parametrization function by setting $X = X_1$ and $\theta = \delta_K$:

$$\bar{f}(x, u) = f(x) + \delta_{X_1}(x) + \delta_K(g(x) + u),$$

where δ_A is the indicator function of a subset A of a space Z, i.e.,

$$\delta_A(a) = \begin{cases} 0 & \text{if } a \in A, \\ +\infty & \text{if } a \in Z \backslash A. \end{cases}$$

Constructing the augmented Lagrangian as in Definition 11.55 of [25], we obtain the augmented Lagrangian:

$$\bar{l}(x, y, r) = \inf_{u \in R^m} \left\{ \bar{f}(x, u) + r\sigma(u) - \langle y, u \rangle \right\}, x \in X, y \in R^m, r > 0.$$

The augmented Lagrangian problem is

$$(ALP(y, r)) \quad \min_{x \in X} \bar{l}(x, y, r), \quad y \in R^m, r > 0.$$

Denote by $v_5(y, r)$ the optimal value of $(ALP(y, r))$.

We have the following result.

THEOREM 4.2. *Let $\{y_n\} \subset R^m$ be bounded and $0 < r_n \to +\infty$. Consider* (P) *and $(ALP(y_n, r_n))$. Assume that there exist $(\bar{y}, \bar{r}) \in R^m \times (0, +\infty)$ and $m_0 \in R^1$ such that*

$$\bar{l}(x, \bar{y}, \bar{r}) \geq m_0 \quad \forall x \in X. \tag{46}$$

Let $0 < \epsilon_n \to 0$. Suppose that each x_n satisfies

$$\bar{l}(x_n, y_n, r_n) \leq v_5(y_n, r_n) + \epsilon_n, \tag{47}$$

$v_5(y_n, r_n) > -\infty \ \forall n$, and (P) *is LP well-posed in the strongly generalized sense. Then there exist a subsequence $\{x_{n_k}\}$ of $\{x_n\}$ and some $\bar{x} \in \bar{X}$ such that $x_{n_k} \to \bar{x}$.*

Proof. By the definition of $\bar{l}(x, y, r)$, it is easy to see that

$$\bar{l}(x, y, r) = f(x) \quad \forall x \in X_0.$$

It follows that

$$v_5(y, r) \leq \bar{v} \quad \forall y \in R^m, r > 0.$$

Thus,

$$v_5(y_n, r_n) \leq \bar{v} \quad \forall n. \tag{48}$$

By the definition of $\bar{l}(x_n, y_n, r_n)$ and (47), $\{x_n\} \subset X_1$ and there exists $\{u_n\} \subset R^m$ satisfying

$$g(x_n) + u_n \in K \quad \forall n \tag{49}$$

such that

$$f(x_n) + r_n\sigma(u_n) - \langle y_n, u_n\rangle \leq v_5(y_n, r_n) + 2\epsilon_n. \tag{50}$$

This combined with (46) and (48) implies that

$$(r_n - \bar{r})\sigma(u_n) - \langle y_n - \bar{y}, u_n\rangle \leq \bar{v} + 2\epsilon_n - m_0. \tag{51}$$

We assert that $\{u_n\}$ is bounded. Otherwise, we assume without loss of generality that $\|u_n\| \to +\infty$. Since the lower semicontinuous and convex function σ has a unique minimum, by Proposition 3.2.5 in IV of [11] and Corollary 3.27 of [25], $\liminf_{n\to+\infty} \sigma(u_n)/\|u_n\| > 0$. As $\{y_n\}$ is bounded, (51) cannot hold. So, $\{u_n\}$ should be bounded. Assume without loss of generality that $u_n \to u_0$. We deduce from (51) that

$$\sigma(u_0) \leq \liminf_{n\to+\infty} \sigma(u_n) = 0.$$

It follows that $u_0 = 0$. We deduce from (48) and (50) that

$$f(x_n) - \langle y_n, u_n\rangle \leq \bar{v} + 2\epsilon_n.$$

Passing to the limit, we get

$$\limsup_{n\to+\infty} f(x_n) \leq \bar{v}.$$

From (49) and the fact that $u_n \to 0$, we obtain

$$\lim_{n\to+\infty} d_K(g(x_n)) = 0.$$

Thus, $\{x_n\}$ is a weakly generalized LP minimizing sequence. Hence, there exist a subsequence $\{x_{n_k}\}$ of $\{x_n\}$ and some $\bar{x} \in \bar{X}$ such that $x_{n_k} \to \bar{x}$. □

Acknowledgment. The authors would like to thank the two anonymous referees for their detailed and constructive comments which have helped to improve some results and the presentation of the paper.

REFERENCES

[1] J. P. Aubin and I. Ekeland, *Applied Nonlinear Analysis*, John Wiley and Sons, New York, 1984.
[2] B. Bank, J. Guddat, D. Klatte, B. Kummer, and K. Tammer, *Nonlinear Parametric Optimization*, Akademie-Verlag, Berlin, 1982.
[3] E. Bednarczuk, *Well-Posedness of Vector Optimization Problems*, Lecture Notes in Econom. and Math. Systems 294, Springer, Berlin, 1987, pp. 51–61.
[4] E. Bednarczuk and J. P. Penot, *Metrically well-set minimization problems*, Appl. Math. Optim., 26 (1992), pp. 273–285.
[5] P. Bosch, A. Jourani, and R. Henrion, *Sufficient conditions for error bounds and applications*, Appl. Math. Optim., 50 (2004), pp. 161–181.
[6] G. Beer and R. Lucchetti, *The epi-distance topology: Continuity and stability results with application to convex optimization problems*, Math. Oper. Res., 17 (1992), pp. 715–726.
[7] S. Deng, *Coercivity properties and well-posedness in vector optimization*, RAIRO Oper. Res., 37 (2003), pp. 195–208.
[8] A. L. Dontchev and R. T. Rockafellar, *Regularity properties and conditioning in variational analysis and optimization*, Set-Valued Anal., 12 (2004), pp. 79–109.
[9] A. L. Dontchev and T. Zolezzi, *Well-Posed Optimization Problems*, Lecture Notes in Math. 1543, Springer, Berlin, 1993.
[10] M. Furi and A. Vignoli, *About well-posed minimization problems for functionals in metric spaces*, J. Optim. Theory Appl., 5 (1970), pp. 225–229.
[11] J. B. Hiriart-Urruty and C. Lemarechal, *Convex Analysis and Minimization Algorithms*, Springer, Berlin, 1993.
[12] X. X. Huang, *Extended well-posed properties of vector optimization problems*, J. Optim. Theory Appl., 106 (2000), pp. 165–182.
[13] X. X. Huang, *Extended and strongly extended well-posedness of set-valued optimization problems*, Math. Methods Oper. Res., 53 (2001), pp. 101–116.
[14] X. X. Huang and X. Q. Yang, *A unified augmented Lagrangian approach to duality and exact penalization*, Math. Oper. Res., 282 (2003), pp. 533–55.
[15] X. X. Huang, X. Q. Yang, and K. L. Teo, *A lower order penalization approach to nonlinear semidefinite programs*, J. Optim. Theory Appl., to appear.
[16] X. X. Huang, X. Q. Yang, and K. L. Teo, *Characterizing nonemptiness and compactness of the solution set of a convex vector optimization problem with cone constraints and applications*, J. Optim. Theory Appl., 123 (2004), pp. 391–407.
[17] A. S. Konsulova and J. P. Revalski, *Constrained convex optimization problems-well-posedness and stability*, Numerical Funct. Anal. Optim., 15 (1994), pp. 889–907.
[18] E. S. Levitin and B. T. Polyak, *Convergence of minimizing sequences in conditional extremum problems*, Soviet Math. Dokl., 7 (1966), pp. 764–767.
[19] Z. Q. Luo, J. S. Pang, and D. Ralph, *Mathematical Programs with Equilibrium Constraints*, Cambridge University Press, New York, 1996.
[20] P. Loridan, *Well-posed vector optimization*, in Recent Developments in Well-Posed Variational Problems, Math. Appl. 331, R. Lucchetti and J. Revalski, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1995, pp. 171–192.
[21] R. Lucchetti, *Well-Posedness Towards Vector Optimization*, Lecture Notes in Econom. and Math. Systems 294, Springer, Berlin, 1987.
[22] R. Lucchetti and J. Revalski, eds., *Recent Developments in Well-Posed Variational Problems*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1995.
[23] J. S. Pang, *Error bounds in mathematical programming*, Math. Programming, 79 (1997), pp. 299–332.
[24] J. P. Revalski, *Hadamard and strong well-posedness for convex programs*, SIAM J. Optim., 7 (1997), pp. 519–526.
[25] R. T. Rockafellar and R. J-B. Wets, *Variational Analysis*, Springer, New York, 1998.
[26] A. N. Tykhonov, *On the stability of the functional optimization problem*, U.S.S.R. Comput. Math. and Math. Phys., 6 (1966), pp. 28–33.
[27] X. Q. Yang and X. X. Huang, *Lower order penalty methods for mathematical programs with complementarity constraints*, Optim. Methods Softw., 19 (2004), pp. 693–720.

[28] T. Zolezzi, *Well-posedness criteria in optimization with application to the calculus of variations*, Nonlinear Anal., 25 (1995), pp. 437–453.
[29] T. Zolezzi, *Extended well-posedness of optimization problems*, J. Optim. Theory Appl., 91 (1996), pp. 257–266.
[30] T. Zolezzi, *Well-posedness and optimization under perturbations*, Ann. Oper. Res., 101 (2001), pp. 351–361.

SIAM J. OPTIM.
Vol. 17, No. 1, pp. 259–286
© 2006 Society for Industrial and Applied Mathematics

LOCAL CONVERGENCE OF SQP METHODS FOR MATHEMATICAL PROGRAMS WITH EQUILIBRIUM CONSTRAINTS*

ROGER FLETCHER†, SVEN LEYFFER‡, DANNY RALPH§, AND STEFAN SCHOLTES§

Abstract. Recently, nonlinear programming solvers have been used to solve a range of mathematical programs with equilibrium constraints (MPECs). In particular, sequential quadratic programming (SQP) methods have been very successful. This paper examines the local convergence properties of SQP methods applied to MPECs. SQP is shown to converge superlinearly under reasonable assumptions near a strongly stationary point. A number of examples are presented that show that some of the assumptions are difficult to relax.

Key words. nonlinear programming, sequential quadratic programming (SQP), mathematical programs with equilibrium constraints (MPEC), mathematical programs with complementarity constraints (MPCC), equilibrium constraints

AMS subject classifications. 90C30, 90C33, 90C55, 49M37, 65K10

DOI. 10.1137/S1052623402407382

1. Introduction.

We consider mathematical programs with equilibrium constraints (MPECs) of the form

$$\begin{array}{ll} \text{minimize} & f(z) \\ \text{subject to} & c_{\mathcal{E}}(z) = 0, \\ & c_{\mathcal{I}}(z) \geq 0, \\ & 0 \leq z_1 \ \perp \ z_2 \geq 0, \end{array} \tag{1.1}$$

where $z = (z_0, z_1, z_2)$ is a decomposition of the problem variables into controls $z_0 \in \mathbb{R}^n$ and states $(z_1, z_2) \in \mathbb{R}^{2p}$. The equality constraints $c_i(z) = 0,\ i \in \mathcal{E}$, are abbreviated as $c_{\mathcal{E}}(z) = 0$, and similarly, $c_{\mathcal{I}}(z) \geq 0$ represents the inequality constraints. Problems of this type arise frequently in applications; see [7, 16, 17] for references. (Problem (1.1) is also referred to as a mathematical program with complementarity constraints (MPCC).)

Clearly, an MPEC with a more general complementarity condition such as

$$0 \leq G(z) \perp H(z) \geq 0 \tag{1.2}$$

can be written in the form (1.1) by introducing slack variables. One can easily show that the reformulated MPEC has the same properties (such as constraint qualifications or second-order conditions) as the original MPEC. In this sense, nothing is lost by introducing slack variables.

*Received by the editors May 10, 2002; accepted for publication (in revised form) December 27, 2005; published electronically May 12, 2006. Research for this work was carried out while the second author was at the University of Dundee and was supported by EPSRC grant GR/M59549.
http://www.siam.org/journals/siopt/17-1/40738.html
†Department of Mathematics, University of Dundee, Dundee, DD1 4HN, UK (fletcher@maths.dundee.ac.uk).
‡Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439 (leyffer@mcs.anl.gov).
§The Judge Institute, University of Cambridge, Cambridge, CB2 1AG, UK (d.ralph@jims.cam.ac.uk, s.scholtes@jims.cam.ac.uk).

One attractive way of solving (1.1) is to consider its equivalent nonlinear programming (NLP) formulation,

$$\begin{array}{ll} \text{minimize} & f(z) \\ \text{subject to} & c_{\mathcal{E}}(z) = 0, \\ & c_{\mathcal{I}}(z) \geq 0, \\ & z_1 \geq 0, \\ & z_2 \geq 0, \\ & z_1^T z_2 \leq 0, \end{array} \tag{1.3}$$

and solve (1.3) with existing NLP solvers. This paper examines the local convergence properties of sequential quadratic programming (SQP) methods applied to (1.3).

The NLP (1.3) obviously has no feasible point that satisfies the inequalities strictly. This fact implies that the Mangasarian–Fromovitz constraint qualification (MFCQ) is violated at every feasible point; see [4, 19]. There are other, MPEC-specific constraint qualifications, such as the MPEC-LICQ explained below, which guarantee the existence of multipliers at local optima of (1.3) and are not overly stringent; see [21]. MFCQ, however, is a sufficient condition for stability of an NLP, and the lack thereof has been advanced as a theoretical argument against the use of standard NLP solvers for MPECs.

Numerical experience with (1.3) has also been disappointing. Bard [2] reports failure on 50–70% of some bilevel problems for a gradient projection method. Conn, Gould, and Toint [5] and Ferris and Pang [7] attribute certain failures of lancelot to the fact that the problem contains a complementarity constraint. In contrast, Fletcher and Leyffer [10] recently reported encouraging numerical results on a large collection of MPECs [15]. They solved over 150 MPECs with an SQP solver and observed *quadratic* convergence for all but two problems. The two problems that did not give quadratic convergence violate certain MPEC regularity conditions and are rather pathological. The present work complements these numerical observations by giving a theoretical explanation for the good performance of the SQP method on apparently ill-posed problems of the type (1.3). We show that SQP is guaranteed to converge quadratically near a stationary point under relatively mild assumptions.

Recently, researchers have expressed renewed interest in the global convergence of algorithms for MPECs. Scholtes [20] analyzes a regularization scheme in which a sequence of parametric NLPs is solved. Fukushima and Tseng [11] analyze an algorithm that computes approximate KKT points for a sequence of active sets.

The paper also complements the recently renewed interest in the convergence properties of SQP under weaker assumptions. See, for example, [8, 13, 22]. These studies suggest modifications to enable SQP solvers to handle NLP problems for which the constraint gradients are linearly dependent at the solution and/or for which strict complementarity fails to hold.

Anitescu [1] extends Wright's analysis [22] to NLPs with unbounded multiplier sets. The fact that (1.3) violates MFCQ implies that the multiplier set at stationary solutions will be unbounded. Anitescu's work therefore applies to MPECs in the given form. However, his assumptions differ from ours, and neither set of assumptions is implied by the others. Most notably, Anitescu assumes that the QP solver employs an elastic mode, relaxing constraint linearizations if they are inconsistent. We do not require such a modification and provide a local analysis of the SQP method in its pure form.

In this paper, we argue that the introduction of slack variables is not just a convenience but plays an important role in ensuring convergence. In section 7.2 we present an example with a nonlinear complementarity constraint for which SQP converges to a nonstationary point. All QP approximations remain consistent during the solve. With the introduction of slack variables, on the other hand, SQP converges to a stationary point. Of course, this does not mean that the use of slacks makes an elastic mode or a feasibility restoration unnecessary. The example in section 2.2 clearly shows that NLP solvers must be able to handle inconsistent QPs.

This paper is organized as follows. The next section gives a few simple motivating examples that highlight the key ideas of our approach and illustrate the numerical difficulties associated with MPECs. In section 3 we review optimality conditions and constraint qualifications for MPECs. Section 4 shows that the optimality conditions of the MPEC and its equivalent NLP are related by a simple formula. In section 5 we show that SQP converges quadratically in two distinct situations. The first arises when SQP is started close to a complementary stationary point. If the starting point is not complementary, then we show convergence under the assumption that all QP subproblems remain consistent. Sufficient conditions for this assumption are introduced in section 6. In section 7 we present small examples that illustrate the necessity of some of these assumptions. We conclude by briefly emphasizing the importance of degeneracy handling at the QP level and pointing to future research directions.

Notation. Throughout the paper, $g(z) = \nabla f(z)$ is the objective gradient and the constraint gradients are denoted by $a_i(z) = \nabla c_i(z)$. Superscripts refer to the point at which functions or gradients are evaluated, for example, $a_i^{(k)} = a_i(z^{(k)}) = \nabla c_i(z^{(k)})$. The Jacobian matrices are denoted by $A_{\mathcal{E}} := [a_i]_{i\in\mathcal{E}}$ and $A_{\mathcal{I}} := [a_i]_{i\in\mathcal{I}}$, respectively.

2. Examples. The fact that the NLP formulation (1.3) of an MPEC violates MFCQ at any feasible point implies that (1.3) has certain features that pose numerical challenges to NLP solvers.

1. The active constraint normals are *linearly dependent* at any feasible point.
2. The set of multipliers is *unbounded*.
3. Arbitrarily close to a stationary point, the linearizations of (1.3) can be *inconsistent*.

These features are illustrated by the following examples. The examples also motivate the analysis in subsequent sections. The main conclusion of this section is that while MPECs possess these unpleasant properties, they arise in a well-structured way that allows SQP solvers to tackle MPECs successfully.

In the remainder of this paper, `*.mod` refers to the AMPL model of the problem in MacMPEC, an AMPL collection of MPECs [15].

2.1. Dependent constraint normals and unbounded multipliers. In this section we use a small example from Jiang and Ralph [14] (see also `jr*.mod`) to illustrate the key idea of our approach. Consider the two MPECs

$$\begin{cases} \underset{z}{\text{minimize}} & f_i(z) \\ \text{subject to} & 0 \le z_2 \ \perp \ z_2 - z_1 \ge 0 \end{cases} \tag{2.1}$$

with $f_1(z) = (z_1 - 1)^2 + z_2^2$ and $f_2(z) = z_1^2 + (z_2 - 1)^2$. The problems differ only in their objectives. The solution to both problems is $z^* = (1/2, 1/2)^T$; see Figure 1.

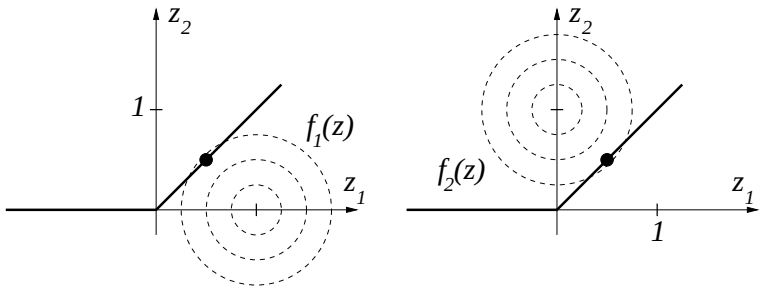


FIG. 1. *MPEC examples* 1 *and* 2.

The equivalent NLP problem to these MPECs is given by

$$\left\{\begin{array}{lll} \underset{z}{\text{minimize}} & f_i(z) & \text{multiplier} \\ \text{subject to} & z_2 \geq 0, & \nu \geq 0, \\ & z_2 - z_1 \geq 0, & \lambda \geq 0, \\ & z_2\,(z_2 - z_1) \leq 0, & \xi \geq 0. \end{array}\right. \tag{2.2}$$

The first-order conditions for these NLPs differ only in the objective gradient and are

$$\begin{pmatrix} -1 \\ 1 \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} 1 \\ -1 \end{pmatrix} \;=\; \lambda^* \begin{pmatrix} -1 \\ 1 \end{pmatrix} \;-\; \xi^* \begin{pmatrix} -\frac{1}{2} \\ \frac{1}{2} \end{pmatrix}.$$

Clearly, the two active constraint normals are linearly dependent. Since $z_2^* = \frac{1}{2} > 0$ it follows that $\nu^* = 0$. The multiplier sets, given by

$$\mathcal{M}_1 \;=\; \left\{(\lambda,\xi) \mid \xi \geq 0,\; \lambda - \frac{1}{2}\xi = 1\right\},$$

$$\mathcal{M}_2 \;=\; \left\{(\lambda,\xi) \mid \lambda \geq 0,\; -\lambda + \frac{1}{2}\xi = 1\right\},$$

are unbounded, as expected. The sets are shown in Figure 2.

This situation is typical for MPECs that satisfy a strong stationarity condition (see Definition 3.3). The multiplier set is a ray, and there is exactly one degree of freedom in the choice of multipliers.

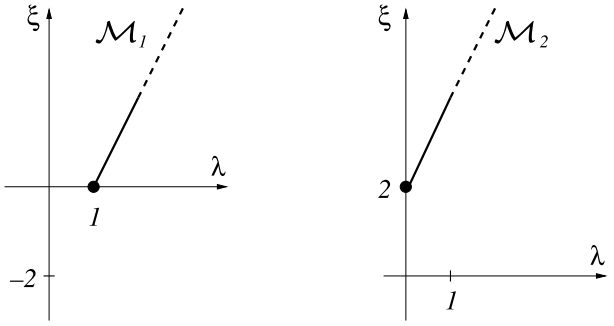


FIG. 2. *Multiplier sets of MPEC examples* 1 *and* 2.

Note, however, that if we restrict our attention to multipliers that correspond to a *linearly independent set* of constraint normals, then the following reduced sets are obtained:

$$\tilde{\mathcal{M}}_1 = \left\{\begin{pmatrix} 1 \\ 0 \end{pmatrix}\right\},$$

$$\tilde{\mathcal{M}}_2 = \left\{\begin{pmatrix} 0 \\ 2 \end{pmatrix}\right\}.$$

These multipliers are bounded and well behaved. We should expect SQP to converge if started near such a stationary point. The KKT multipliers that correspond to a solution with linearly independent strictly active constraints are illustrated by the black circles in Figure 2. The half-line shows the unbounded multiplier set.

Observe that in the first example, $\lambda \geq 0$ at the solution, which implies that this is also the solution for the NLP with the complementarity condition removed. In the second example, no $\lambda \geq 0$ can on its own satisfy the stationarity conditions, and $\xi > 0$ is required. If we had interpreted $z_2 - z_1 \geq 0$ as an equality constraint, then we could have chosen $\lambda = -1$ in the stationarity conditions. However, an NLP solver would never return $\lambda < 0$ for an inequality constraint, and hence $\xi = 2$ ensures that the stationarity conditions are satisfied.

The effect of the multiplier of the complementarity constraint is to relax the condition that $\lambda, \nu \geq 0$ for what is essentially an equality constraint. This is exploited in section 4, where we show that certain MPEC multipliers correspond to multipliers of (1.3). This situation is typical for MPECs under certain assumptions. The key idea is to show that SQP converges to a solution provided the QP solver chooses a linearly independent basis.

2.2. Inconsistent linearizations. The following example illustrates a possible pitfall for NLP solvers attempting to solve MPECs as NLPs. Consider `sl4.mod`:

$$\begin{cases} \underset{z}{\text{minimize}} & z_1 + z_2 \\ \text{subject to} & z_2^2 \geq 1, \\ & 0 \leq z_1 \ \perp \ z_2 \geq 0. \end{cases} \tag{2.3}$$

Its solution is $z^* = (0,1)^T$ with NLP multipliers $\lambda^* = 0.5$ of $z_2^2 \geq 1$, $\nu_1^* = 1$ of $z_1 \geq 0$, and $\xi^* = 0$ of $z_1 z_2 \leq 0$. In particular, this solution is a strongly stationary point (see Definition 3.3). However, linearizing the constraints about a point that satisfies the simple bounds and is *arbitrarily close to the solution*, such as $z^{(0)} = (\epsilon, 1-\delta)^T$ (with $\epsilon, \delta > 0$), gives a QP that is *inconsistent*. The linearizations are

$$(1-\delta)^2 + 2(1-\delta)(z_2 - (1-\delta)) \geq 1, \tag{2.4}$$

$$z_1 \geq 0,$$

$$z_2 \geq 0,$$

$$(1-\delta)\epsilon + (1-\delta)(z_1 - \epsilon) + \epsilon(z_2 - (1-\delta)) \leq 0. \tag{2.5}$$

One can show that

$$(2.4) \Rightarrow z_2 \geq \frac{1 + (1-\delta)^2}{2(1-\delta)} > 1,$$

$$(2.5) \Rightarrow z_2 \leq 1 - \delta < 1,$$

which indicates that the QP approximation is inconsistent. This is also observed during our filter solves (we enter restoration at this point).

Clearly, any NLP solver hoping to tackle MPECs will have to deal with this situation. The solver snopt [12] uses an *elastic mode* that relaxes the linearizations of the QP; filter [9] has a restoration phase. In section 5 convergence of SQP methods without modifications is analyzed. This analysis is closer in spirit to the results obtained using filter.

3. Optimality conditions for MPECs. This section reviews stationarity concepts for MPECs in the form (1.1) and introduces a second-order condition. It follows loosely the development of Scheel and Scholtes [19], although the presentation is slightly different.

Given two index sets $\mathcal{Z}_1,\ \mathcal{Z}_2 \subset \{1,\dots,p\}$ with

$$\mathcal{Z}_1 \cup\ \mathcal{Z}_2\ =\ \{1,\dots,p\}\,, \tag{3.1}$$

we denote their respective complements in $\{1,\dots,p\}$ by $\mathcal{Z}_1^c$ and $\mathcal{Z}_2^c$. For any such pair of index sets, we define the *relaxed NLP corresponding to the MPEC* (1.1) as

$$\begin{array}{lll} \underset{z}{\text{minimize}} & f(z) & \\ \text{subject to} & c_{\mathcal{E}}(z) = 0, & \\ & c_{\mathcal{I}}(z) \geq 0, & \\ & z_{1j} = 0 & \forall j \in \mathcal{Z}_2^c, \\ & z_{2j} = 0 & \forall j \in \mathcal{Z}_1^c, \\ & z_{1j} \geq 0 & \forall j \in \mathcal{Z}_2, \\ & z_{2j} \geq 0 & \forall j \in \mathcal{Z}_1. \end{array} \tag{3.2}$$

Concepts such as constraint qualifications, stationarity, and a second-order condition for MPECs will be defined in terms of the relaxed NLPs. The term "relaxed NLP" stems from the observation that if z^* is a local solution of a relaxed NLP (3.2) *and* satisfies complementarity ${z_1^*}^T z_2^* = 0$, then z^* is also a local solution of the original MPEC (1.1). One can naturally associate with every feasible point $\hat{z} = (\hat{z}_0, \hat{z}_1, \hat{z}_2)$ of the MPEC a relaxed NLP (3.2) by choosing $\mathcal{Z}_1$ and $\mathcal{Z}_2$ to contain the indices of the vanishing components of $\hat{z}_1$ and $\hat{z}_2$, respectively. In contrast to [19], our definition of the relaxed NLP is independent of a specific point; however, it will occasionally be convenient to identify the above sets of vanishing components associated with a specific point $\hat{z}$, in which case we denote them by $\mathcal{Z}_1(\hat{z})$, $\mathcal{Z}_2(\hat{z})$, or use suitable superscripts. Note that for these sets the condition (3.1) is equivalent to $\hat{z}_1^T \hat{z}_2 = 0$.

The indices that are both in $\mathcal{Z}_1$ and $\mathcal{Z}_2$ are referred to as the *biactive components* (or second-level degenerate indices) and are denoted by

$$\mathcal{D}(z) := \mathcal{Z}_1(z) \cap \mathcal{Z}_2(z) \ \text{ or } \ \mathcal{D} := \mathcal{Z}_1 \cap \mathcal{Z}_2.$$

Obviously, in view of (3.1), $(\mathcal{Z}_1^c, \mathcal{Z}_2^c, \mathcal{D})$ is a partition of $\{1,\dots,p\}$. A solution z^* to the problem (1.1) is said to be *second-level nondegenerate* if $\mathcal{D}(z^*) = \emptyset$.

First, the linear independence constraint qualification (LICQ) is extended to MPECs.

DEFINITION 3.1. *Let $z_1, z_2 \geq 0$, and define*

$$\mathcal{Z}_j := \{i\ :\ z_{ji} = 0\} \quad \textit{for}\ \ j = 1, 2.$$

The MPEC (1.1) *is said to satisfy an* MPEC-LICQ *at z if the corresponding relaxed NLP* (3.2) *satisfies an LICQ.*

In [19], four stationarity concepts are introduced for MPEC (1.1). The stationarity definition that allows the strongest conclusions is Bouligand or B-stationarity.

DEFINITION 3.2. *A point z^* is called* Bouligand*, or* B-stationary*, if $d = 0$ solves the linear program with equilibrium constraints (LPEC) obtained by linearizing f and c about z^*,*

$$\begin{array}{ll} \underset{d}{\text{minimize}} & {g^*}^T d \\ \text{subject to} & c^*_{\mathcal{E}} + {A^*_{\mathcal{E}}}^T d = 0, \\ & c^*_{\mathcal{I}} + {A^*_{\mathcal{E}}}^T d \geq 0, \\ & 0 \leq z^*_1 + d_1 \ \perp \ z^*_2 + d_2 \geq 0. \end{array}$$

We note that B-stationarity implies feasibility because if $d = 0$ solves the above LPEC, then $c^*_{\mathcal{E}} = 0$, $c^*_{\mathcal{I}} \geq 0$, and $0 \leq z^*_1 \ \perp \ z^*_2 \geq 0$. B-stationarity is difficult to check because it involves the solution of an LPEC that is a combinatorial problem and may require the solution of an exponential number of LPs, unless *all* these LPs share a common multiplier vector. Such a common multiplier vector exists if an MPEC-LICQ holds.

The results of this paper relate to the following notion of strong stationarity.

DEFINITION 3.3. *A point z^* is called* strongly stationary *if there exist multipliers λ, $\hat{\nu}_1$, and $\hat{\nu}_2$ such that*

$$\begin{aligned} g^* - \left[{A^*_{\mathcal{E}}}^T \ : \ {A^*_{\mathcal{I}}}^T\right]\lambda - \begin{pmatrix} 0 \\ \hat{\nu}_1 \\ \hat{\nu}_2 \end{pmatrix} &= 0, \\ c^*_{\mathcal{E}} &= 0, \\ c^*_{\mathcal{I}} &\geq 0, \\ z^*_1 &\geq 0, \\ z^*_2 &\geq 0, \\ z^*_{1j} = 0 \ \ \textit{or} \ \ z^*_{2j} &= 0, \\ \lambda_{\mathcal{I}} &\geq 0, \\ c^*_i \lambda_i &= 0, \\ z^*_{1j} \hat{\nu}_{1j} &= 0, \\ z^*_{2j} \hat{\nu}_{2j} &= 0, \\ \textit{if } z^*_{1j} = z^*_{2j} = 0, \ \ \textit{then} \ \ \hat{\nu}_{1j} \geq 0 \ \ \textit{and} \ \ \hat{\nu}_{2j} &\geq 0, \end{aligned} \tag{3.3}$$

where $g^* = \nabla f(z^*)$, $A^*_{\mathcal{E}} = \nabla c^T_{\mathcal{E}}(x^*)$, and $A^*_{\mathcal{I}} = \nabla c^T_{\mathcal{I}}(x^*)$.

Note that (3.3) are the stationarity conditions of the relaxed NLP (3.2) at z^*. B-stationarity is equivalent to strong stationarity if the MPEC-LICQ holds (e.g., [19]).

Next, a second-order sufficient condition (SOSC) for MPECs is given. Since strong stationarity is related to the relaxed NLP (3.2), it seems plausible to use the same NLP to define a second-order condition. For this purpose, let $\mathcal{A}^*$ denote the set of active constraints of (3.2) and $\mathcal{A}^*_+ \subset \mathcal{A}^*$ the set of active constraints with nonzero multipliers (some could be negative). Let A denote the matrix of active constraint normals, that is,

$$A = \begin{bmatrix} & & 0 & & 0 \\ A^*_{\mathcal{E}} \ : \ A^*_{\mathcal{I}\cap\mathcal{A}^*} \ : & & I^*_1 & : & 0 \\ & & 0 & & I^*_2 \end{bmatrix} =: \ [a^*_i]_{i\in\mathcal{A}^*},$$

where $A^*_{\mathcal{I}\cap\mathcal{A}^*}$ are the active inequality constraint normals and

$$I_1^* := [e_i]_{i\in\mathcal{Z}_1^*} \quad \text{and} \quad I_2^* := [e_i]_{i\in\mathcal{Z}_2^*}$$

are parts of the $p \times p$ identity matrices corresponding to active bounds. Define the set of feasible directions of zero slope of the relaxed NLP (3.2) as

$$S^* = \{s \mid s \neq 0\ ,\ {g^*}^T s = 0\ ,\ {a_i^*}^T s = 0\ ,\ i \in \mathcal{A}_+^*\ ,\ {a_i^*}^T s \geq 0\ ,\ i \in \mathcal{A}^*\backslash\mathcal{A}_+^*\}.$$

We can now give an MPEC-SOSC. This condition is also sometimes referred to as the strong-SOSC.

DEFINITION 3.4. *A strongly stationary point z^* with multipliers $(\lambda^*, \hat{\nu}_1^*, \hat{\nu}_2^*)$ satisfies the* MPEC-SOSC *if for every direction $s \in S^*$ it follows that*

$$s^T \nabla^2 \mathcal{L}^* s > 0,$$

where $\nabla^2\mathcal{L}^*$ is the Hessian of the Lagrangian of (3.2) *evaluated at $(z^*, \lambda^*, \hat{\nu}_1^*, \hat{\nu}_2^*)$.*

The definitions of this section are readily extended to the case where a more general complementarity condition such as (1.2) is used. Moreover, any reformulation using slacks preserves all of these definitions. In that sense, there is no loss of generality in assuming that slacks are being used.

4. Strong stationarity and NLP stationarity. This section shows that there exists a relationship between strong stationarity of the MPEC (1.1) and NLP stationarity conditions for (1.3). In particular, their respective multipliers are shown to be related by a simple formula.

The NLP stationarity conditions of (1.3) are that there exist multipliers $\mu := (\lambda, \nu_1, \nu_2, \xi)$ such that

$$\begin{aligned}
g(z) - \left[A_{\mathcal{E}}^T(z) : A_{\mathcal{I}}^T(z)\right]\lambda - \begin{pmatrix} 0 \\ \nu_1 \\ \nu_2 \end{pmatrix} + \xi \begin{pmatrix} 0 \\ z_2 \\ z_1 \end{pmatrix} &= 0, \\
c_{\mathcal{E}}(z) &\geq 0, \\
c_{\mathcal{I}}(z) &\geq 0, \\
z_1 &\geq 0, \\
z_2 &\geq 0, \\
z_1^T z_2 &\leq 0, \\
\lambda_{\mathcal{I}} &\geq 0, \\
\nu_1 &\geq 0, \\
\nu_2 &\geq 0, \\
\xi &\geq 0, \\
c_i(z)\lambda_i &= 0, \\
z_{1j}\nu_{1j} &= 0, \\
z_{2j}\nu_{2j} &= 0.
\end{aligned} \tag{4.1}$$

The complementarity condition $\xi z_1^T z_2 = 0$ is implied by the feasibility of z_1, z_2. This condition has been omitted.

We examine the difference between (4.1) and the strong-stationarity condition (3.3). In (3.3), the multipliers $\hat{\nu}_1$ and $\hat{\nu}_2$ may be negative for components that satisfy

second-level nondegeneracy, while in (4.1) $\nu_1 \geq 0, \nu_2 \geq 0$ is required. We will relate the multipliers of (3.3) and (4.1) to show that stationarity in both senses is equivalent.

The main observation in proving the following result is that the first-order condition of (4.1) can be written as

$$g(z) - \left[A_{\mathcal{E}}^T(z) : A_{\mathcal{I}}^T(z)\right] \lambda - \begin{pmatrix} 0 \\ \nu_1 - \xi z_2 \\ 0 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ \nu_2 - \xi z_1 \end{pmatrix} = 0,$$

which is equivalent to the corresponding first-order condition in (3.3) if

$$\hat{\nu}_1 = \nu_1 - \xi z_2, \tag{4.2}$$

$$\hat{\nu}_2 = \nu_2 - \xi z_1. \tag{4.3}$$

PROPOSITION 4.1. *A point z is strongly stationary in the MPEC* (1.1) *if and only if it is a stationary point of the NLP* (1.3).

Proof. First we show that (4.1) $\Rightarrow$ (3.3) by distinguishing three cases:

(a) If $z_{1j} > 0$, then $z_{2j} = 0 = \nu_{1j}$ from complementarity and slackness. From (4.2) it follows that $\hat{\nu}_{1j} = 0$ and $\hat{\nu}_{2j} = \nu_{2j} - \xi z_{1j}$ satisfies (3.3).

(b) If $z_{2j} > 0$, then transpose the above argument.

(c) If $z_{1j} = z_{2j} = 0$, then (4.2) and (4.3) imply that $\hat{\nu}_{1j} = \nu_{1j} \geq 0$ and $\hat{\nu}_{2j} = \nu_{2j} \geq 0$. Combining (a)–(c), one sees that (4.1) implies (3.3).

Next we show that (3.3) $\Rightarrow$ (4.1) by distinguishing three cases:

(d) If $z_{1j} > 0$, then $\hat{\nu}_{1j} = 0$ and $z_{2j} = 0$. This implies that $\nu_{1j} = \xi z_{2j} + \hat{\nu}_{1j} = 0 \geq 0$ for any ξ. To ensure that $\nu_{2j} = \xi z_{1j} + \hat{\nu}_{2j}$ is nonnegative, we need to choose ξ such that $\xi z_{1j} + \hat{\nu}_{2j} \geq 0 \; \forall j$, or equivalently that $\xi \geq -\hat{\nu}_{2j}/z_{1j} \; \forall j$.

(e) If $z_{2j} > 0$, then transpose the above argument.

(f) If $z_{1j} = z_{2j} = 0$, then $\nu_{1j} = \hat{\nu}_{1j} \geq 0$ and $\nu_{2j} = \hat{\nu}_{2j} \geq 0$, for any ξ.

From parts (d) and (e) it follows that choosing ξ to be at least

$$\xi = \max\left\{0, \max_{i \in \mathcal{Z}_2^c} \frac{-\hat{\nu}_{1i}}{z_{2i}^*}, \max_{i \in \mathcal{Z}_1^c} \frac{-\hat{\nu}_{2i}}{z_{1i}^*}\right\} \tag{4.4}$$

will ensure that $\nu_1, \nu_2 \geq 0$. Examining the expressions on the right-hand side of (4.4), one can see that ξ is bounded. Combining cases (d) to (f) it follows that (3.3) implies (4.1). ∎

The interesting point about the proof is that it relates the multiplier ξ to the fact that the NLP conditions (4.1) are more restrictive in the sense that they enforce $\nu_1, \nu_2 \geq 0$, while $\hat{\nu}_1, \hat{\nu}_2$ may be negative. In a way, ξ compensates for this: if, for instance, $\hat{\nu}_{1j} < 0$, then $z_{2j} > 0$, and we can get the corresponding NLP multiplier $\nu_{1j} = \hat{\nu}_{1j} + \xi z_{2j}$ nonnegative by choosing ξ sufficiently large.

Clearly, any value $\hat{\xi} > \xi$ in (4.4) would also satisfy the stationarity conditions (4.1), and this is how the unboundedness of the multiplier set arises. However, any such $\hat{\xi} > \xi$ would not correspond to a *basic solution*, in the sense that the constraint normals corresponding to nonzero multipliers are linearly dependent. The main argument in our convergence analysis is to show that an SQP solver that works with a nonsingular basis will pick the multiplier defined in (4.4).

DEFINITION 4.2. *The multiplier defined by* (4.4) *is referred to as the* basic multiplier.

The terminology of this definition is justified by the following lemma, which shows that if MPEC-LICQ holds, then the MPEC multipliers and the multiplier in (4.4) are unique and correspond to a linearly independent set of constraint normals.

LEMMA 4.3. *If MPEC-LICQ holds at a local minimizer of* (1.1), *then it is strongly stationary, and the multipliers in* (3.3) *and the basic multiplier defined by* (4.4) *are unique. Moreover, the set of constraint normals corresponding to nonzero multipliers is linearly independent.*

Proof. MPEC-LICQ implies the uniqueness of the MPEC multipliers (3.3); see [19]. The uniqueness of the MPEC multiplier implies that all expressions on the right-hand side of (4.4) are unique, hence implying the uniqueness of ξ. Finally, the uniqueness of the corresponding NLP multipliers follows from (4.2) and (4.3) (if the NLP multipliers were not unique, then we could find other MPEC multipliers).

To show that the constraint normals corresponding to nonzero multipliers are linearly independent, we distinguish two cases: $\xi = 0$ and $\xi > 0$.

If $\xi = 0$, then the linear independence of constraint normals corresponding to nonzero multipliers follows from MPEC-LICQ.

If $\xi > 0$, then there exists at least one component $i \in \mathcal{Z}_1^c$ or $i \in \mathcal{Z}_2^c$ such that $\nu_{2i} = 0$ or $\nu_{1i} = 0$.

It remains to show that the set of constraint normals corresponding to nonzero multipliers is linearly independent. By MPEC-LICQ, this is true for all but the complementarity constraint. Then we can exchange the normal of the complementarity constraint for any normal whose multiplier is driven to zero by (4.4) and (4.2) or (4.3) in the basis as explained in Lemma 5.8 below. ∎

The conclusions of this section can be readily extended to cover the case where the complementarity condition is of the more general form (1.2).

5. Local convergence of SQP methods. This section shows that SQP methods converge quadratically near a strongly stationary point under mild conditions. Section 7 discusses the assumptions and provides counterexamples for situations where (some of) these assumptions are not satisfied. In particular, we are interested in the situation where $z^{(k)}$ is close to a strongly stationary point, z^*, but ${z_1^{(k)}}^T z_2^{(k)}$ is *not* necessarily zero. SQP then solves a sequence of quadratic programming approximations, given by

$$(QP^k) \qquad \begin{cases} \underset{d}{\text{minimize}} & {g^{(k)}}^T d + \frac{1}{2} d^T W^{(k)} d \\ \text{subject to} & c_{\mathcal{E}}^{(k)} + {A_{\mathcal{E}}^{(k)}}^T d = 0, \\ & c_{\mathcal{I}}^{(k)} + {A_{\mathcal{I}}^{(k)}}^T d \geq 0, \\ & z_1^{(k)} + d_1 \geq 0, \\ & z_2^{(k)} + d_2 \geq 0, \\ & {z_1^{(k)}}^T z_2^{(k)} + {z_2^{(k)}}^T d_1 + {z_1^{(k)}}^T d_2 \leq 0, \end{cases}$$

where $W^{(k)} = \nabla^2 \mathcal{L}(z^{(k)}, \mu^{(k)})$ is the Hessian of the Lagrangian of (1.3) and $\mu^{(k)} = (\lambda^{(k)}, \nu_1^{(k)}, \nu_2^{(k)}, \xi^{(k)})$. The last constraint of (QP^k) is the linearization of the complementarity condition $z_1^T z_2 \leq 0$.

Assumption 5.1. The following assumptions are made:

[A1] f and c are twice Lipschitz continuously differentiable.

[A2] The MPEC (1.1) satisfies an MPEC-LICQ (Definition 3.1).

[A3] z^* is a strongly stationary point of (1.1) with multipliers $\lambda^*, \nu_1^*, \nu_2^*$ (Definition 3.3), and z^* satisfies the MPEC-SOSC (Definition 3.4).

[A4] $\lambda_i^* \neq 0 \; \forall i \in \mathcal{E}^*$, $\lambda_i^* > 0 \; \forall i \in \mathcal{A}^* \cap \mathcal{I}$, and both $\nu_{1j}^* > 0$ and $\nu_{2j}^* > 0 \; \forall j \in \mathcal{D}^*$.

[A5] The QP solver always chooses a linearly independent basis.

The most restrictive assumption is strong stationarity in [A3], which follows if z^* is a local minimizer from [A2]. That is, [A3] (or [A2]) removes the combinatorial nature of the problem. It is not clear that [A2] can readily be relaxed in the present context, since it allows us to check B-stationarity by solving exactly one LP or QP. Without assumption [A2] it would not be possible to verify B-stationarity without solving several LPs (one for every possible combination of second-level degenerate indices $i \in \mathcal{D}^*$). It seems unlikely, therefore, that a method that solves only a single LP or QP per outer iteration can be shown to be convergent to B-stationary points for problems that violate MPEC-LICQ. Note that we do *not* assume that the MPEC (1.1) is second-level nondegenerate; in other words, we do *not* assume that $z_1^* + z_2^* > 0$.

The strict complementarity Assumption [A4] can in fact be weakened for all the results of section 5.1 to require positivity only of the biactive multipliers ν_{1j}^* and ν_{2j}^*, because Proposition 5.2, which underlies our convergence analysis there, does not require $\lambda_i^* \neq 0 \; \forall i \in \mathcal{E}^*$ and $\lambda_i^* > 0 \; \forall i \in \mathcal{A}^* \cap \mathcal{I}$; see [3]. Section 5.2, however, requires all the conditions of [A4]. Assumption [A5] is a reasonable assumption in practice, as most modern SQP solvers are based on active set QP solvers that guarantee this.

This section is divided into two parts. First, we consider the case where complementarity is satisfied at a point sufficiently close to a stationary point. This case corresponds to the situation where all iterates (ultimately) remain on the same face of $0 \leq z_1 \perp z_2 \geq 0$. The key idea is to show that SQP applied to (1.3) behaves identical to SQP applied to (3.2).

The second case considered arises when ${z_1^{(k)}}^T z_2^{(k)} > 0$ for all iterates k. In this case, the previous ideas cannot be applied, and a separate proof is required. We make the additional assumption that all QP subproblems remain consistent. This assumption appears to be rather strong, especially in light of example (2.3), which shows that the QP approximation may be inconsistent *arbitrarily close to a solution*. However, we will give several sufficient conditions for it later that show that it is not unduly restrictive.

5.1. Local convergence for exact complementarity. In this section we make the following additional assumption:

[A6] For some k we have that ${z_1^{(k)}}^T z_2^{(k)} = 0$ and $(z^{(k)}, \mu^{(k)})$ is sufficiently close to a strongly stationary point.

Assumption [A6] implies that the correct face has been identified except for degenerate or biactive constraints. Thus, for given index sets $\mathcal{Z}_j = \{i : z_{ji}^{(k)} = 0\}$, $j = 1, 2$, the following holds:

$$\begin{aligned} z_{1j}^{(k)} &= 0 && \forall j \in \mathcal{Z}_2^c, \\ z_{2j}^{(k)} &= 0 && \forall j \in \mathcal{Z}_1^c, \\ z_{1j}^{(k)} &= 0 \quad \text{and} \quad z_{2j}^{(k)} = 0 && \forall j \in \mathcal{D}. \end{aligned}$$

In particular, it is *not* assumed that the biactive complementarity constraints $\mathcal{D}^*$ are active at $z^{(k)}$. Thus it may be possible that $\mathcal{Z}_1 \neq \mathcal{Z}_1^*$ (and similarly for $\mathcal{Z}_2$). However, it will be shown that the biactive constraints become active after one step of the SQP method as a consequence of [A4] (the positivity of biactive multipliers); see Proposition 5.2.

An important consequence of [A6] is that $\mathcal{Z}_1$ and $\mathcal{Z}_2$ satisfy

$$\begin{aligned} \mathcal{Z}_1^{*^c} \subset \mathcal{Z}_1^c \subset \mathcal{Z}_1^{*^c} \cup \mathcal{D}^*, \\ \mathcal{Z}_2^{*^c} \subset \mathcal{Z}_2^c \subset \mathcal{Z}_2^{*^c} \cup \mathcal{D}^*, \\ \mathcal{D} \subset \mathcal{D}^*; \end{aligned} \tag{5.1}$$

in other words, the indices $\mathcal{Z}_1^{*^c}$ and $\mathcal{Z}_2^{*^c}$ of the nondegenerate complementarity constraints have been identified correctly.

The key idea of the proof is to show that SQP applied to (1.3) is equivalent to SQP applied to the relaxed NLP (3.2) on a face. For a given partition $(\mathcal{Z}_1^c, \mathcal{Z}_2^c, \mathcal{D})$, an SQP step for (3.2) is obtained by solving the following QP:

$$(QP_R(z^{(k)}) \qquad \begin{cases} \underset{d}{\text{minimize}} & g^{(k)^T} d + \frac{1}{2} d^T \widehat{W^{(k)}} d & \\ \text{subject to} & c_{\mathcal{E}}^{(k)} + A_{\mathcal{E}}^{(k)^T} d = 0, & \\ & c_{\mathcal{I}}^{(k)} + A_{\mathcal{I}}^{(k)^T} d \geq 0, & \\ & d_{1j} = 0 & \forall j \in \mathcal{Z}_2^c, \\ & d_{2j} = 0 & \forall j \in \mathcal{Z}_1^c, \\ & z_{1j}^{(k)} + d_{1j} \geq 0 & \forall j \in \mathcal{Z}_2, \\ & z_{2j}^{(k)} + d_{2j} \geq 0 & \forall j \in \mathcal{Z}_1, \end{cases}$$

where

$$\widehat{W^{(k)}} = \nabla^2 f(z^{(k)}) - \sum \lambda_i^{(k)} \nabla^2 c_i(z^{(k)}) = W^{(k)} - \xi^{(k)} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & I \\ 0 & I & 0 \end{bmatrix}$$

is the Hessian of the Lagrangian of the relaxed NLP (3.2). Note that the relaxed NLP (3.2) is never set up nor is $(QP_R(z^{(k)}))$ ever solved. These two problems are merely used in the convergence proof. The key idea is to show that SQP applied to the ill-conditioned NLP (1.3) is equivalent to SQP applied to the well-behaved relaxed NLP (3.2), given by the sequence defined by $(QP_R(z^{(k)}))$.

The following proposition states the fact that SQP applied to the relaxed NLP converges quadratically and identifies the correct index sets $\mathcal{Z}_1^*$ and $\mathcal{Z}_2^*$ in one step.

PROPOSITION 5.2. *Let Assumptions* [A1]–[A6] *hold, and consider the relaxed NLP for any index sets* $\mathcal{Z}_1$, $\mathcal{Z}_2$ *(satisfying* (5.1) *by virtue of* [A6]*). Then it follows that*

1. *there exists a neighborhood* U *of* $(z^*, \lambda^*, \nu_1^*, \nu_2^*)$ *and a sequence of iterates generated by SQP applied to the relaxed NLP* (3.2), $\{(z^{(l)}, \lambda^{(l)}, \nu_1^{(l)}, \nu_2^{(l)})\}_{l>k}$, *that lies in* U *and converges Q-quadratically to* $(z^*, \lambda^*, \nu_1^*, \nu_2^*)$*;*
2. *the sequence* $\{z^{(l)}\}_{l>k}$ *converges Q-superlinearly to* z^**; and*
3. $\mathcal{Z}_1^{(l)} = \mathcal{Z}_1^*$ *and* $\mathcal{Z}_2^{(l)} = \mathcal{Z}_2^*$ *for* $l > k$.

Proof. The relaxed NLP satisfies LICQ and an SOSC. Therefore, there exists a neighborhood U of $(z^*, \lambda^*, \nu_1^*, \nu_2^*)$ such that for any $(z^{(l)}, \lambda^{(l)}, \nu_1^{(l)}, \nu_2^{(l)}) \in U$, there exists an SQP iterate $(z^{(l+1)}, \lambda^{(l+1)}, \nu_1^{(l+1)}, \nu_2^{(l+1)})$ that also lies in U; and any sequence of SQP iterates $\{z^{(l)}\}_{l>k} \subset U$ converges at second-order rate when applied to the relaxed NLP. In fact part 1 is a standard result whose proof can be found, for instance, in [6, Theorem 15.2.2] or in [3]. Part 2 is due to [3]. Part 3 follows from the fact

that SQP identifies the correct active set in one step by the strict complementarity assumption [A4]. ∎

Next, we show that the sequence of steps generated by SQP applied to the relaxed NLP (3.2) is identical to the sequence of steps generated by applying SQP to the equivalent NLP (1.3), provided that $z_1^{(k)^T} z_2^{(k)} = 0$, that is, [A6] holds for some k. If $z_1^{(k)^T} z_2^{(k)} = 0$, then an SQP step for (1.3) is obtained by solving the following (QP^k) with $z_1^{(k)^T} z_2^{(k)} = 0$ in the last constraint.

The two QPs (QP^k) and $(QP_R(z^{(k)}))$ have different constraints and Hessians. The Hessian of (QP^k) is

$$W^{(k)} = \widehat{W^{(k)}} + \xi^{(k)} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & I \\ 0 & I & 0 \end{bmatrix}.$$

Despite these differences, however, one can show that the two QPs have the same solution (from which second-order convergence follows). The following lemma shows that the constraint sets are the same.

LEMMA 5.3. *Let Assumptions* [A1]–[A6] *hold. Then, a step d is feasible in (QP^k) if and only if it is feasible in $(QP_R(z^{(k)}))$.*

Proof. The constraint sets differ only in the way in which indices $j \in \mathcal{Z}_2^c$ and $j \in \mathcal{Z}_1^c$ are handled. Thus it suffices to consider those constraints.

(a) *Let d be feasible in $(QP_R(z^{(k)}))$.* Then it follows in particular that d satisfies

$$\begin{aligned} d_{1j} &= 0 \quad \forall j \in \mathcal{Z}_2^c, \\ d_{2j} &= 0 \quad \forall j \in \mathcal{Z}_1^c. \end{aligned}$$

If these constraints are split into two inequalities, we have that d satisfies

$$d_{1j} \geq 0 \quad \forall j \in \mathcal{Z}_2^c, \tag{5.2}$$

$$d_{1j} \leq 0 \quad \forall j \in \mathcal{Z}_2^c, \tag{5.3}$$

$$d_{2j} \geq 0 \quad \forall j \in \mathcal{Z}_1^c, \tag{5.4}$$

$$d_{2j} \leq 0 \quad \forall j \in \mathcal{Z}_1^c. \tag{5.5}$$

Summing (5.3) over all $j \in \mathcal{Z}_2^c$ weighted with $z_{2j}^{(k)} > 0$ and (5.5) over all $j \in \mathcal{Z}_1^c$ weighted with $z_{1j}^{(k)} > 0$, it follows that d satisfies the last constraint of (QP^k) (the simple bounds follow from (5.2) and (5.4)).

(b) *Let d be feasible in (QP^k).* Since $z_{2j}^{(k)} > 0 \ \forall j \in \mathcal{Z}_2^c$ and $z_{1j}^{(k)} > 0 \ \forall j \in \mathcal{Z}_1^c$, it follows from [A6] that $z_{1j}^{(k)} = 0 \ \forall j \in \mathcal{Z}_2^c$ and that $z_{2j}^{(k)} = 0 \ \forall j \in \mathcal{Z}_1^c$. Thus, (QP^k) contains the constraints

$$d_{1j} \geq 0 \ \forall j \in \mathcal{Z}_2^c \quad \text{and} \quad d_{2j} \geq 0 \ \forall j \in \mathcal{Z}_1^c.$$

By [A6], the linearization of the complementarity constraint in (QP^k) simplifies to

$$\sum_{j \in \mathcal{Z}_2^c} z_{2j}^{(k)} d_{1j} + \sum_{j \in \mathcal{Z}_1^c} z_{1j}^{(k)} d_{2j} \leq 0.$$

Since $z_{2j}^{(k)} > 0$, and $z_{1j}^{(k)} > 0$ in this sum, it follows that

$$d_{1j} \leq 0 \ \forall j \in \mathcal{Z}_2^c \quad \text{and} \quad d_{2j} \leq 0 \ \forall j \in \mathcal{Z}_1^c.$$

Thus, d is feasible in $(QP_R(z^{(k)}))$. ∎

Next, we show that near z^*, the stationary points of (QP^k) and $(QP_R(z^{(k)}))$ are identical.

LEMMA 5.4. *Under Assumptions* [A1]–[A6], *any stationary point of* $(QP_R(z^{(k)}))$ *near zero is also a stationary point of* (QP^k), *and vice versa.*

Proof. From above, $(QP_R(z^{(k)}))$ and (QP_k) share the same feasible set. Consider a feasible point $\bar{d}$ that satisfies $\bar{d}_{ij} = 0$ for $j \in \mathcal{D}$, $i = 1, 2$. Since all feasible points d have $d_{1j} = 0$ for $j \in \mathcal{Z}_2^c$, then $\bar{d}_{1j} = 0$ for $j \in \mathcal{Z}_2^c \cup \mathcal{D} = \mathcal{Z}_1$. Likewise, $\bar{d}_{2j} = 0$ for $j \in \mathcal{Z}_2$. That is, we have orthogonality between $\bar{d}_1$ and d_2, and between $\bar{d}_2$ and d_1, for any feasible d. Thus the gradients of the objective functions of $(QP_R(z^{(k)}))$ and (QP_k) at $\bar{d}$, which differ only by $\xi^{(k)}(0, \bar{d}_2, \bar{d}_1)$, cannot be distinguished on the feasible set. It follows that $\bar{d}$ is stationary for $(QP_R(z^{(k)}))$ if and only if it is stationary for (QP_k). To complete the proof we show that any stationary point, near zero, of either QP does indeed satisfy the above conditions on $\bar{d}$.

From part 3 of Proposition 5.2, any stationary point $\bar{d}$ of $(QP_R(z^{(k)}))$, near zero, satisfies $z_{ij}^{(k)} + \bar{d}_{ij} = 0$ for $j \in \mathcal{D}^*$ and $i = 1, 2$. As $\mathcal{D}^* \supset \mathcal{D} \subset \mathcal{Z}_1$, we get for $j \in \mathcal{D}$ that $\bar{d}_{1j} = 0$. Likewise $\bar{d}_{2j} = 0$ for $j \in \mathcal{D}$.

Conversely let $\bar{d}$ be a stationary point, near zero, of (QP_k). The associated KKT conditions include

$$g^{(k)} + W^{(k)}\bar{d} - \left[{A_{\mathcal{E}}^{(k)}}^T : {A_{\mathcal{I}}^{(k)}}^T \right] \bar{\lambda} - \begin{pmatrix} 0 \\ \bar{\nu}_1 \\ \bar{\nu}_2 \end{pmatrix} + \bar{\xi} \begin{pmatrix} 0 \\ z_2^{(k)} \\ z_1^{(k)} \end{pmatrix} = 0$$

for some multipliers $\bar{\lambda}, \bar{\nu}_1, \bar{\nu}_2, \bar{\xi}$. As $z^{(k)}$ and $\bar{d}$ approach z^* and zero, respectively, where $\mu^{(k)}$ is within a given radius of $(\lambda^*, \nu_1^*, \nu_2^*, 0)$, we deduce from MPEC-LICQ and the first equation of (3.3) that $\bar{\nu}_1 - \bar{\xi} z_2^{(k)}$ and $\bar{\nu}_2 - \bar{\xi} z_1^{(k)}$ approach ν_1^* and ν_2^*, respectively. Therefore Assumption [A4], with nonnegativity of $z_1^{(k)}$, $z_2^{(k)}$, and $\bar{\xi}$, ensures that $\bar{\nu}_{ij} > 0$, hence $z_{ij}^{(k)} + \bar{d}_{ij} = 0$, for $j \in \mathcal{D}^*$ and $i = 1, 2$. The argument that $\bar{d}_{ij} = 0$ for $j \in \mathcal{D}$ and $i = 1, 2$ is given in the previous paragraph. ∎

LEMMA 5.5. *Let Assumptions* [A1]–[A6] *hold. Let* $(\lambda, \widehat{\nu}_1, \widehat{\nu}_2)$ *be the multipliers of* $(QP_R(z^{(k)}))$ *(corresponding to a step d near zero). Then it follows that the multipliers of* (QP^k), *corresponding to the same step d, are* $\mu = (\lambda, \nu_1, \nu_2, \xi)$, *where*

$$\xi = \max\left(0, \max_{j \in \mathcal{Z}_1 \setminus \mathcal{D}} \frac{-\widehat{\nu_{1j}} - \xi^{(k)} d_{2j}}{z_{2j}^{(k)}}, \max_{j \in \mathcal{Z}_2 \setminus \mathcal{D}} \frac{-\widehat{\nu_{2j}} - \xi^{(k)} d_{1j}}{z_{1j}^{(k)}}\right), \tag{5.6}$$

$$\nu_{1j} = \widehat{\nu_{1j}} > 0 \quad \forall j \in \mathcal{D}, \tag{5.7}$$

$$\nu_{2j} = \widehat{\nu_{2j}} > 0 \quad \forall j \in \mathcal{D}, \tag{5.8}$$

$$\nu_{1j} = \widehat{\nu_{1j}} + \xi^{(k)} d_{2j} + \xi z_{2j}^{(k)} \quad \forall j \in \mathcal{Z}_1 \setminus \mathcal{D}, \tag{5.9}$$

$$\nu_{2j} = \widehat{\nu_{2j}} + \xi^{(k)} d_{1j} + \xi z_{1j}^{(k)} \quad \forall j \in \mathcal{Z}_2 \setminus \mathcal{D}. \tag{5.10}$$

Conversely, given a solution d and multipliers μ *of* (QP^k), (5.7)–(5.10) *show how to construct multipliers so that* $(d, \lambda, \widehat{\nu}_1, \widehat{\nu}_2)$ *solves* $(QP_R(z^{(k)}))$.

Proof. If $z^{(k)}$ is sufficiently close to z^*, then the sign of the multipliers in (5.7) and (5.8) follows from [A4], and the value for the multipliers of (QP^k) follows similarly to Proposition 4.1. Similarly, the multipliers of (QP^k) in (5.9) and (5.10) are nonnegative by construction and satisfy first-order conditions by Lemma 5.4. ∎

Next, we show that both QPs have the same (unique) solution in a neighborhood of $d = 0$.

LEMMA 5.6. *The solution d of $(QP_R(z^{(k)}))$ is the only strict local minimizer in a neighborhood of $d = 0$ that is independent of k, and its corresponding multipliers (λ, ν_1, ν_2) are unique. Moreover, d is also the only strict local minimizer in a neighborhood of $d = 0$ of (QP^k).*

Proof. The result for $(QP_R(z^{(k)}))$ is due to Robinson [18] (see also Conn, Gould, and Toint [6]), since the relaxed NLP satisfies [A1]–[A4]. The statement for (QP^k) follows in two parts. First-order conditions are established in Lemma 5.5. Second-order conditions for (QP^k) follow from second-order conditions of $(QP_R(z^{(k)}))$, as we explain now. The critical cone at a stationary point is the set of directions in the tangent cone to the feasible set that are orthogonal to the gradient of the objective function. From Lemma 5.3 and the proof of Lemma 5.4, it can be seen that the critical cones of $(QP_R(z^{(k)}))$ and (QP^k) coincide; denote this cone C. Next, we use a standard fact that relates an inequality constraint with a positive multiplier to any direction d in the critical cone of (QP^k), namely, $d_{ij} = 0$ if $\bar{\nu}_{ij} > 0$. Hence, using the proof of Lemma 5.4, we have that $d_1^T d_2 = 0$ for $d \in C$. It follows that the Hessian matrices of $(QP_R(z^{(k)}))$ and (QP^k) are indistinguishable on C, hence that the SOSC of the former transfers to the latter, and the stationary point must be a local minimizer of the latter. ∎

The following theorem summarizes the results of this section. As remarked earlier, these results holds under a weak version of Assumption [A4] in which only positivity of the biactive multipliers ν_{1j}^* and ν_{2j}^* is required.

THEOREM 5.7. *If Assumptions* [A1]–[A6] *hold, then SQP applied to* (1.3) *generates a sequence $\{(z^{(k)}, \lambda^{(k)}, \nu_1^{(k)}, \nu_2^{(k)}, \xi^{(k)})\}_{l>k}$ that converges Q-quadratically to a solution $\{(z^*, \lambda^*, \nu_1^*, \nu_2^*, \xi^*)\}$ of* (4.1), *satisfying strong stationarity. Moreover, the sequence $\{z^{(k)}\}_{l>k}$ converges Q-superlinearly to z^* and $z_1^{(l)^T} z_2^{(l)} = 0 \ \forall l \geq k$.*

Proof. Under Assumptions [A1]–[A4], SQP converges quadratically when applied to the relaxed NLP (3.2); see Proposition 5.2. Lemmas 5.3–5.6 show that the sequence of iterates generated by this SQP method is equivalent to the sequence of steps generated by SQP applied to (1.3). This implies Q-superlinear convergence of $\{z^{(k)}\}_{l>k}$. Convergence of the multipliers follows by considering (5.6)–(5.10). Clearly, the multipliers in (5.7) and (5.8) converge, as they are just the multipliers of the relaxed NLP, which converge by virtue of Proposition 5.2. Now observe that (5.6) becomes

$$\widehat{\xi}^{(k+1)} = \max\left(0, \max_{j \in \mathcal{Z}_2^{*c}} \frac{-\widehat{\nu_{1j}^{(k+1)}} - \xi^{(k)} d_{2j}^{(k)}}{z_{2j}^{(k)}}, \max_{j \in \mathcal{Z}_1^{*c}} \frac{-\widehat{\nu_{2j}^{(k+1)}} - \xi^{(k)} d_{1j}^{(k)}}{z_{1j}^{(k)}}\right).$$

The right-hand side of this expression converges, since $\widehat{\nu_{1j}^{(k+1)}}, \widehat{\nu_{2j}^{(k+1)}}$ and $z_{1j}^{(k)}, z_{2j}^{(k)}$ converge and $d_{1j}^{(k)}, d_{2j}^{(k)} \to 0$. Note that the limit of (5.6) is the basic multiplier (4.4). Finally, (5.9) and (5.10) converge to (4.2) and (4.3) by a similar argument.

Now $z_1^{(l)^T} z_2^{(l)} = 0 \ \ \forall l \geq k$ follows from the convergence of SQP for the relaxed NLP (3.2) and the fact that SQP retains feasibility with respect to linear constraints. Assumption [A4] ensures that $d_{1j}^{(k)} = d_{2j}^{(k)} = 0 \ \forall j \in \mathcal{D}^*$, since $\nu_{1j}^{(k)}, \nu_{2j}^{(k)} > 0$ for biactive complementarity constraints. Thus SQP will not move out of the corner but will stay on the same face. ∎

5.2. Local convergence for nonzero complementarity. This section shows that SQP converges superlinearly even if complementarity does not hold at the starting point, that is, if $z_1^{(k)^T} z_2^{(k)} > 0$. Example (2.3) shows that the QP approximations can be inconsistent arbitrarily close to a stationary point. To avoid this problem, we make the following assumption, which often holds in practice.

[A7] All QP approximations (QP^k) are consistent.

This is clearly an undesirable assumption because it makes an assumption about the progress of the method. However, we show in the next section that this assumption is satisfied for some important practical applications.

Without loss of generality, we assume that $\mathcal{Z}_1^{*^c} = \emptyset$, that is, we will assume that the solution has the form $z_1^* = 0$ and $z_2^* = (0, z_{22}^*)$ and that $z_{22}^* > 0$. This assumption greatly simplifies the notation.

Our convergence analysis is concerned with showing that for any "basic" active set, SQP converges. To this end, we introduce the set of basic constraints

$$\mathcal{B}(z) := \mathcal{E} \cup \mathcal{I} \cap \mathcal{A}^* \cup \mathcal{Z}_1(z) \cup \mathcal{Z}_2(z) \cup \{z_1^T z_2 = 0\}$$

and the set of strictly active constraints (defined in terms of the basic multiplier, μ)

$$\mathcal{B}_+(z) \ := \ \{i \in \mathcal{B}(z) \mid \mu_i \neq 0\}.$$

Moreover, we let $B_+^{(k)}$ denote the matrix of strictly active constraint normals at $z = z^{(k)}$, namely,

$$B_+^{(k)} \ := \ \left[a_i^{(k)}\right]_{i \in \mathcal{B}_+(z^{(k)})}.$$

Note that Lemma 4.3 shows that the optimal multiplier is unique. However, it may be possible that for some iterates $\mathcal{B}_+^{(k)} \neq \mathcal{B}_+(z^*)$, and our analysis will have to allow for this.

The failure of any constraint qualification at a solution z^* of the equivalent NLP (1.3) implies that the active constraint normals at z^* are linearly dependent. However, the linear dependence occurs in a special form that can be exploited to prove fast convergence.

LEMMA 5.8. *Let Assumptions* [A1]–[A4] *hold, and let z^* be a solution of the MPEC* (1.1). *Let $\mathcal{I}^*$ denote the set of* active *inequalities $c_{\mathcal{I}}(x)$, and consider the matrix of active constraint normals at z^*,*

$$B^* \ = \ \begin{bmatrix} & & 0 & 0 & 0 \\ A_{\mathcal{E}}^* & A_{\mathcal{I}^*}^* & I & 0 & \begin{pmatrix} 0 \\ z_{22}^* \end{pmatrix} \\ & & 0 & \begin{bmatrix} I \\ 0 \end{bmatrix} & \begin{pmatrix} 0 \\ 0 \end{pmatrix} \end{bmatrix}, \tag{5.11}$$

where we have assumed without loss of generality that $\mathcal{Z}_1^{c^*} = \emptyset$. Note that the last column is the gradient of the complementarity constraint.

Then it follows that B is linearly dependent and any submatrix of columns of B has full rank, provided that it contains $[A_{\mathcal{E}}^* \ A_{\mathcal{I}}^*]$ and that either the last column of B is missing or any column corresponding to $z_{12} = 0$ is missing.

Proof. The fact that the columns of B are linearly dependent is clear by looking at the last three columns of B. Assumption [A2], MPEC-LICQ, implies that B without

the last column has full rank. The final statement follows by exchanging any column corresponding to $z_{12}^* = 0$ with the final column of B and observing that $z_{22}^* > 0$. ∎

The proof shows that in order to obtain a linearly independent basis, any column of $z_{12} = 0$ can be exchanged with the normal of the complementarity constraint. This idea is precisely what lies behind (4.2) and (4.3). The corresponding basic multipliers are shown as dots in Figure 2.

Next, we show that if we are close to z^* and the QP solver chooses the full basis B, then exact complementarity holds for all subsequent iterations. Thus, in this case the development of the previous section shows second-order convergence.

LEMMA 5.9. *Let $z^{(k)}$ be sufficiently close to z^*, and let Assumptions* [A1]–[A5] *and* [A7] *hold. If the QP solver chooses the full basis B^k, given by*

$$B^{(k)} = \begin{bmatrix} & & 0 & 0 & 0 \\ A_{\mathcal{E}}^{(k)} & A_{\mathcal{I}^*}^{(k)} & I & 0 & \begin{pmatrix} z_{21}^{(k)} \\ z_{22}^{(k)} \end{pmatrix} \\ & & 0 & \begin{bmatrix} I \\ 0 \end{bmatrix} & \begin{pmatrix} z_{11}^{(k)} \\ z_{12}^{(k)} \end{pmatrix} \end{bmatrix},$$

then it follows that $z_1^{(k)^T} z_2^{(k)} > 0$ and that after the QP step, $z_1^{(k+1)^T} z_2^{(k+1)} = 0$.

Proof. Assume that $z_1^{(k)^T} z_2^{(k)} = 0$, and seek a contradiction. Since $z^{(k)}$ is sufficiently close to z^*, it follows that there exists $\tau > 0$ such that $z_{22}^{(k)} \geq \tau > 0$. Hence, $z_{12}^{(k)} = 0$. Now consider the final three columns of $B^{(k)}$, and observe that if $z_{12}^{(k)} = 0$, then the last column lies in the range of the other two. Hence the basis would be singular, thus contradicting Assumption [A5], and so $z_1^{(k)^T} z_2^{(k)} > 0$.

Now, $z_1^{(k+1)^T} z_2^{(k+1)} = 0$ follows simply by observing that the full basis B implies that $0 = z_1^{(k)} + d_1 = z_1^{(k+1)}$. ∎

Thus, once a full basis is chosen, the corresponding step will give $z_1^{(k+1)^T} z_2^{(k+1)} = 0$ for a point close to z^*. Second-order convergence then follows from Theorem 5.7.

COROLLARY 5.10. *Let $z^{(k)}$ be sufficiently close to z^*, and let Assumptions* [A1]–[A5] *and* [A7] *hold. If the QP solver chooses the full basis B, then it follows that SQP converges quadratically from iteration $k+1$.*

In the remainder we can therefore concentrate on the case in which the full basis B is never chosen and $z_1^{(k)^T} z_2^{(k)} > 0$ for all iterates k (otherwise, we have convergence from the results of the previous section).

Next, we show that for $z^{(k)}$ sufficiently close to z^*, the basis at $z^{(k)}$ contains both $\mathcal{E}$ and $\mathcal{I}^*$.

LEMMA 5.11. *Let $z^{(k)}$ be sufficiently close to z^*, and let Assumptions* [A1]–[A5] *and* [A7] *hold. Then it follows that the optimal basis B of (QP^k) contains the normals $A_{\mathcal{E}}^{(k)}$ and $A_{\mathcal{I}^*}^{(k)}$ of active constraints at the solution.*

Proof. The proof follows by considering the gradient of the Lagrangian of (QP^k),

$$0 = g^{(k)} + \hat{W}^{(k)} d^{(k)} - \left[A_{\mathcal{E}}^{(k)^T} : A_{\mathcal{I}}^{(k)^T} \right] \lambda^{(k+1)} - \begin{pmatrix} 0 \\ \nu_1^{(k+1)} - \xi^{(k+1)} z_2^{(k)} \\ \nu_2^{(k+1)} - \xi^{(k+1)} z_1^{(k)} \end{pmatrix} + \xi^{(k)} \begin{pmatrix} 0 \\ d_2^{(k)} \\ d_1^{(k)} \end{pmatrix},$$

where $\hat{W}^{(k)}$ is the Hessian of the Lagrangian without the term corresponding to the complementarity constraint (the last term above). For $z^{(k)}$ sufficiently close to z^*, it follows from [A4] that $\lambda_i^{(k+1)} \neq 0 \ \forall i \in \mathcal{E} \cup \mathcal{I}^*$. ∎

Thus, as long as the QP approximations remain consistent, the optimal basis of (QP^k) will be a subset of B satisfying the conditions in Lemma 5.9. The key idea is now to show that for any such basis, there exists an equality constrained problem for which SQP converges quadratically. Since there exist only a finite number of bases, this implies convergence for SQP.

We now introduce the *reduced NLP*, which is an equality constraint NLP. Its constraints correspond to a linearly independent subset of the basis B^* in (5.11) of Lemma 5.8:

$$\begin{array}{lll} \text{minimize} & f(z) & \\ \text{subject to} & c_{\mathcal{E}}(z) = 0, & \\ & c_{\mathcal{I}^*}(z) = 0, & \\ & z_{11} = 0, & \\ & z_{21} = 0, & \\ & z_{12} = 0 & \Big\} \text{ subset of } B^* \text{ satisfying Lemma 5.8.} \\ & z_1^T z_2 = 0 & \end{array} \tag{5.12}$$

The next lemma shows that any reduced NLP satisfies an LICQ and an SOSC.

LEMMA 5.12. *Let Assumptions* [A1]–[A4] *and* [A7] *hold. Then it follows that any reduced NLP satisfies an LICQ and an SOSC.*

Proof. Lemma 5.8 shows that the normals of the equality constraints of each reduced NLP are linearly independent. The SOSC follows from the MPEC-SOSC and the observation that the MPEC and the reduced NLP have the same nullspace. ∎

Thus, applying SQP to the reduced NLP results in second-order convergence.

PROPOSITION 5.13. *Let Assumptions* [A1]–[A4] *and* [A7] *hold. Then it follows that SQP applied to any reduced NLP converges locally and quadratically to* (z^*, μ^*).

Proof. Lemma 5.12 shows that the reduced NLP satisfies LICQ and SOSC. Therefore, convergence of SQP follows. In particular, it follows that for a given reduced NLP corresponding to a basis $\mathcal{B}$, there exists a constant $c_{\mathcal{B}} > 0$ such that

$$\left\| \left(z^{(k+1)}, \mu^{(k+1)}\right) - \left(z^*, \mu^*\right) \right\| \le c_{\mathcal{B}} \left\| \left(z^{(k)}, \mu^{(k)}\right) - \left(z^*, \mu^*\right) \right\|^2. \tag{5.13}$$ ∎

Summarizing the results of this section, we obtain the following theorem.

THEOREM 5.14. *Let Assumptions* [A1]–[A5] *and* [A7] *hold. Then it follows that SQP applied to the NLP formulation* (1.3) *of the MPEC* (1.1) *converges quadratically near a solution* (z^*, μ^*).

Proof. Proposition 5.13 shows that SQP converges quadratically for any possible choice of basis $\mathcal{B}$, and Assumption [A7] shows that (QP^k) is consistent and remains consistent. Therefore, there exists a basis for which quadratic convergence follows. Thus, for each basis, a step is computed that satisfies a contraction condition like (5.13) for a constant $c_{\mathcal{B}} > 0$ that depends on the basis. Since there exists a finite number of bases, this condition holds also for $c = \max c_{\mathcal{B}}$ independent of the basis, and SQP converges quadratically independent of the basis. ∎

5.3. Discussion of the proofs. An interesting observation about the convergence proofs of this section is that if ${z_1^{(k)}}^T z_2^{(k)} = 0$, then the actual value of $\xi^{(k)}$ has no effect on the step computed by SQP. This shows that the curvature information contained in the complementarity constraint $z_1^T z_2 \le 0$ is not important. Consequently, one could omit this contribution to the Hessian of the Lagrangian. This can be easily implemented, and convergence results follow along lines similar to the observation above.

The conclusions and proofs presented in this section also carry through for linear complementarity constraints but *not* for general nonlinear complementarity constraints. The reason is that the implication

$$z_1^{(k)^T} z_2^{(k)} = 0 \Rightarrow z_1^{(k+1)^T} z_2^{(k+1)} = 0$$

holds for linear complementarity problems but *not* for nonlinear complementarity problems, because in general an SQP method would move off a nonlinear constraint. This is one reason for introducing slacks to deal with complementarity of the form (1.2).

Similar conclusions can easily be derived for other NLP formulations of the MPEC (1.1). For instance, the complementarity constraint in (1.3) can be replaced by

$$z_{1j} z_{2j} \leq 0 \quad \forall j = 1, \ldots, p.$$

In this case, a similar construction to (5.6) is possible, where $\widehat{\xi}$ is replaced by a vector of complementarity multipliers, one for each constraint. Equations (4.2) and (4.3) then become componentwise conditions and, similarly, (5.9) and (5.10). In addition, one can now see that a basis that satisfies the conditions of Lemma 5.9 satisfies a complementarity condition between the multipliers ξ_i and ν_{1i} (and ν_{2i}).

The strongest assumption in the present convergence analysis is Assumption [A7], namely, that all (QP^k) remain consistent. We show in the next section that this assumption holds for several interesting cases. We also show that a simple restoration procedure always ensures consistency after one step.

6. Sufficient conditions for consistency of (QP^k). Example (2.3) shows that the QP approximation to an MPEC can be inconsistent arbitrarily close to a stationary point. This section gives two situations in which consistency of (QP^k) can be guaranteed under Assumptions [A1]–[A5]. The first such situation arises when there are no general constraints on control and state variables. Next, we show that one step of a simple restoration procedure is guaranteed to find an iterate with $z_1^{(k)^T} z_2^{(k)} = 0$, thus ensuring consistency.

6.1. Vertical complementarity constraints. This section shows that the QP approximations (QP^k) are consistent arbitrarily close to a strongly stationary point, provided that the MPEC has the following form:

$$\begin{array}{ll} \text{minimize} & f(z) \\ \text{subject to} & c(z_0) = 0, \\ & 0 \leq G(z) \ \perp \ H(z) \geq 0, \end{array} \tag{6.1}$$

where $G, H : \mathbb{R}^{n+2p} \to \mathbb{R}^p$ are twice continuously differentiable. We note that the general constraints are on the control variables only and that the only complementarity constraint takes the form of a vertical complementarity constraints. This case was brought to our attention by Mihai Anitescu.

In this section, we make the following additional assumption, which is related to the mixed P_0 property (e.g., [16]).

[A8] The matrix $[\nabla c(z_0^*) : \nabla G(z^*) : \nabla H(z^*)]$ has full rank.

The motivation for considering this form of problem (6.1) is that the simple complementarity constraint $0 \leq z_1 \ \perp \ z_2 \geq 0$ always produces feasible linearization if there are no other constraints on z_1, z_2.

To see the relationship between Assumption [A8] and the mixed P_0 property, consider the equivalent MPEC with slacks defined by

$$\begin{array}{ll} \text{minimize} & f(z) \\ \text{subject to} & F(z,s) = 0, \\ & 0 \le s_1 \ \perp \ s_2 \ge 0, \end{array} \tag{6.2}$$

where

$$F(z,s) = \begin{pmatrix} c(z_0) \\ G(z) - s_1 \\ H(z) - s_2. \end{pmatrix}.$$

One can see that a sufficient condition for Assumption [A8] is that the Jacobian matrix

$$\begin{bmatrix} \nabla_{s_1} F \\ \nabla_{s_2} F \\ \nabla_z F \end{bmatrix} = \begin{bmatrix} 0 & -I & 0 \\ 0 & 0 & -I \\ & \nabla_z F & \end{bmatrix}$$

satisfy the mixed P_0 property. This assumption has been used, for instance, in the convergence analysis of MPEC solvers and holds for a range of test problems, such as those arising from obstacle or packaging problems [17, Chapter 9].

LEMMA 6.1. *Let Assumptions* [A1]–[A5] *and* [A8] *hold. Then it follows that* (QP^k) *is consistent* $\forall z^{(k)}$ *in a neighborhood of* z^* *where* $G^{(k)^T} H^{(k)} \ge 0$. *If, in addition, the functions* $G(z)$ *and* $H(z)$ *are convex, then* $G^{(k+1)^T} H^{(k+1)} \ge 0$.

Proof. Let $z^{(k)}$ be sufficiently close to z^* so that the Jacobian matrix

$$\left[\nabla c(z_0^{(k)}) : \nabla G(z^{(k)}) : \nabla H(z^{(k)})\right]$$

has full rank.

The linearizations of the QP approximation to (6.1) has the following constraints:

$$c^{(k)} + \nabla c^{(k)^T} d_0 = 0, \tag{6.3}$$

$$G^{(k)} + \nabla G^{(k)^T} d \ge 0, \tag{6.4}$$

$$H^{(k)} + \nabla H^{(k)^T} d \ge 0, \tag{6.5}$$

$$G^{(k)^T} H^{(k)} + G^{(k)^T} \nabla H^{(k)^T} d + H^{(k)^T} \nabla G^{(k)^T} d \le 0. \tag{6.6}$$

We need to show that these constraints are consistent. By [A8] it follows that there exists $\hat{d}$ such that constraints (6.3)–(6.5) hold with equality (this corresponds to the origin in the $G - H$ coordinate system).

It can be shown that $\hat{d}$ is also feasible in (6.6). The constraints (6.3) and (6.4) hold with equality, thus implying that $\nabla G^{(k)^T} \hat{d} = -G^{(k)}$ and $\nabla H^{(k)^T} \hat{d} = -H^{(k)}$. Substituting these last two equations into (6.6) simplifies that constraint to

$$G^{(k)^T} H^{(k)} + G^{(k)^T} \nabla H^{(k)^T} \hat{d} + H^{(k)^T} \nabla G^{(k)^T} \hat{d} \ = \ -G^{(k)^T} H^{(k)} \ \le \ 0,$$

where the last inequality follows from the assumption that $G^{(k)^T} H^{(k)} \ge 0$.

To show that the QP step d^* maintains nonnegative complementarity, we observe that for $z^{(k)}$ sufficiently close to z^*, SQP converges and identifies the correct active set. Thus, there exists a partition

$$\mathcal{G} := \{i \ : \ G_i(z^*) = 0\} \quad \text{and} \quad \mathcal{H} := \{i \ : \ H_i(z^*) = 0\},$$

and

$$G_i^{(k)} + \nabla G_i^{(k)^T} d^* = 0, i \in \mathcal{G}, \tag{6.7}$$

$$H_i^{(k)} + \nabla H_i^{(k)^T} d^* = 0, i \in \mathcal{H}. \tag{6.8}$$

Note that d^* is feasible for an LPEC approximation, because $\mathcal{G} \cup \mathcal{H} \supset \{1, \ldots, p\}$ implies that

$$\left(G^{(k)} + \nabla G^{(k)^T} d^*\right)^T \left(H^{(k)} + \nabla H^{(k)^T} d^*\right) = 0. \tag{6.9}$$

Hence, if $G(z)$ and $H(z)$ are convex, it follows that

$$G^{(k+1)} = G(z^{(k)} + d) \geq G^{(k)} + \nabla G^{(k)^T} d,$$

and similarly for $H^{(k+1)}$. Combining this with (6.9) implies that

$$G^{(k+1)^T} H^{(k+1)} \geq \left(G^{(k)} + \nabla G^{(k)^T} d\right)^T \left(H^{(k)} + \nabla H^{(k)^T} d\right) \geq 0. \qquad \square$$

The main conclusion of this section is that Assumption [A8] turns out to be satisfied for a range of practical problems as long as the vertical complementarity problem has certain properties. This assumption is satisfied, for instance, for obstacle and packaging problems.

6.2. Feasibility restoration for complementarity. This section examines the properties of (QP^k) where $z_1^{(k)^T} z_2^{(k)} > 0$. In this case, (QP^k) may be inconsistent. This section describes a simple restoration procedure that can be invoked if (QP^k) is inconsistent. The procedure finds a new iterate $z^{(k+1)}$ with $z_1^{(k+1)^T} z_2^{(k+1)} = 0$. Thus, after one step, all subsequent iterates retain feasibility of the QP approximations by virtue of Theorem 5.7.

If (QP^k) is inconsistent, then we consider solving the following LP:

$$(LP_F^k) \qquad \begin{cases} \underset{d,\theta}{\text{minimize}} & \theta \\ \text{subject to} & c_{\mathcal{E}}^{(k)} + A_{\mathcal{E}}^{(k)^T} d = 0, \\ & c_{\mathcal{I}}^{(k)} + A_{\mathcal{I}}^{(k)^T} d \geq 0, \\ & z_1^{(k)} + d_1 \geq 0, \\ & z_2^{(k)} + d_2 \geq 0, \\ & z_1^{(k)^T} z_2^{(k)} + z_2^{(k)^T} d_1 + z_1^{(k)^T} d_2 \leq \theta. \end{cases}$$

It follows from Assumption [A2] that any QP approximation to the relaxed NLP (3.2) is consistent for $z^{(k)}$ sufficiently close to z^* and thus that (LP_F^k) is consistent (since it is a relaxation of the relaxed QP). If $z^{(k)}$ is far away from z^*, then clearly (LP_F^k) need not be consistent. In that case we enter a restoration phase.

The following lemma shows that the solution d of (LP_F^k) satisfies $(z_1^{(k)}+d_1)^T(z_2^{(k)}+d_2) = 0$. The key idea of the proof is to show that the optimal active set includes $\mathcal{Z}_1$ and $\mathcal{Z}_2$.

LEMMA 6.2. *Let Assumptions* [A1]–[A5] *hold, and assume that $z^{(k)}$ is sufficiently close to z^* so that the linearizations of $c_{\mathcal{E}}(z)$, $c_{\mathcal{I}}(z)$ are consistent and $z_1^{(k)}, z_2^{(k)} \geq 0$.*

Then it follows that (LP_F^k) *has a solution* d *such that* $z^{(k+1)} = z^{(k)} + d$ *satisfies* ${z_1^{(k+1)}}^T z_2^{(k+1)} = 0$.

Proof. Assume without loss of generality that $\mathcal{Z}_1^c = \emptyset$, namely, that $z_1^* = 0$, and consider the dual feasibility conditions of (LP_F^k) (primal feasibility follows from Assumption [A2]),

$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} - \left[\begin{array}{c|cc|c} & 0 & 0 & 0 \\ A_{\mathcal{A}}^{(k)} & I_1 & 0 & z_2^{(k)} \\ & 0 & I_2 & z_1^{(k)} \\ \hline 0 & 0 & 0 & -1 \end{array}\right] \begin{pmatrix} \lambda_{\mathcal{A}} \\ \nu_1 \\ \nu_2 \\ \xi \end{pmatrix} = 0, \tag{6.10}$$

where $A_{\mathcal{A}}^{(k)}$ is the matrix of active constraint normals of $c_{\mathcal{E}}(z)$, $c_{\mathcal{I}}(z)$ at z^*, $I_2 = [e_i]_{i \in \mathcal{Z}_2}$ and $I_1 = [e_i]_{i \in \mathcal{Z}_1}$.

It follows immediately that $\xi = -1$ and that this active set gives rise to a primal feasible solution. Moreover, the columns of the basis matrix in (6.10) are linearly independent by Assumption [A2]. Thus there exists a unique solution to (6.10). Assumption [A2] implies in particular that the following block of (6.10) has full column rank:

$$\left[\begin{array}{c|cc} & 0 & 0 \\ A_{\mathcal{A}}^{(k)} & I_1 & 0 \\ & 0 & I_2 \end{array}\right].$$

This implies that the block of $A_{\mathcal{A}}^{(k)}$ corresponding to the first equation in (6.10) has full column rank, and thus $\lambda_{\mathcal{A}} = 0$ follows. This implies that

$$\nu_1 = z_2^{(k)} \geq 0 \quad \text{and} \quad \nu_2 = z_1^{(k)} \geq 0.$$

Complementary slackness of (LP_F^k) implies that ${z_1^{(k+1)}}^T z_2^{(k+1)} = 0$. To see how this follows, consider three cases:

Case 1. $i \in \mathcal{Z}_2^c$ implies that $z_{2i}^{(k)} > 0$. This implies that $\nu_{1i} > 0$, and thus $z_{1i}^{(k)} + d_{1i} = 0$.

Case 2. $i \in \mathcal{Z}_2$ and $z_{1i}^{(k)}, z_{2i}^{(k)} > 0$. This implies that $\nu_{1i}, \nu_{2i} > 0$, and thus $z_{1i}^{(k)} + d_{1i} = 0$ and $z_{2i}^{(k)} + d_{2i} = 0$.

Case 3. $i \in \mathcal{Z}_2$ and $z_{1i}^{(k)} > 0$ but $z_{2i}^{(k)} = 0$. This implies that $\nu_{2i} > 0$, and thus $z_{1i}^{(k)} + d_{1i} = 0$. The case where $z_{1i}^{(k)} = 0$ but $z_{2i}^{(k)} > 0$ is analogous.

Putting all three cases together and recalling that $\mathcal{Z}_1 = \emptyset$, one then has that ${z_1^{(k+1)}}^T z_2^{(k+1)} = 0$.

It remains to prove that there exist multipliers λ with $\lambda_{\mathcal{I}} \geq 0$ such that (6.10) holds. If $\lambda_{\mathcal{I} \cap \mathcal{A}} \geq 0$, there is nothing to show. Hence assume that there exists a multiplier $\lambda_i < 0$ for $i \in \mathcal{I} \cap \mathcal{A}$. Then one can perform an iteration of an active set method on (LP_F^k) that will not remove any columns of I_1 or I_2 from the basis. Since (LP_F^k) is bounded ($\theta > 0$, since (QP^k) is inconsistent), after a finite number of such pivots a basis is found with ν_1, ν_2 as above, and the conclusion follows. ∎

Solving (LP_F^k), if (QP^k) is inconsistent, is related to the elastic mode of snopt. In the elastic mode, some of the constraints are relaxed and an l_∞-QP is solved. The application of snopt to MPECs is explored in [1]. Unlike snopt, however, the present restoration will occur only at one iteration.

An alternative to solving (LP_F^k) would be to move $z^{(k)}$ onto the "nearest" axis. This is the effect of (LP_F^k), as can be seen from Lemma 6.2. However, solving (LP_F^k) avoids the need to choose tolerances to break ties between "close" values.

We note that this restoration does not address the wider issue of *global* convergence. It may be possible that the solution to (LP_F^k) is not acceptable to the global convergence criterion of the SQP method. Clearly, this possibility has to be taken into account in designing a globally convergent SQP method. This is beyond the scope of the present paper, which deals exclusively with local convergence issues.

7. Discussion of assumptions. This section discusses some of the assumptions made in the proof above. In particular, examples are presented showing that SQP will fail to converge at second-order rate if some or all of the assumptions are removed. The following table shows which assumptions seem difficult to remove. Below, each example is presented in turn.

Example	[A2]	MFCQ	[A3]	Slacks	SOSC	Comments
`scholtes4`	no	yes	no	yes	yes	$\xi \to \infty$, linear convergence
`sl2`	yes	yes	yes	no	yes	$\xi \to \infty$, nonstationary limit
`ralph2`	yes	yes	yes	yes	no	$\xi < \infty$, linear convergence

7.1. Unbounded multipliers and slow convergence. The following MPEC shows that if we remove Assumption [A2] and, in particular, Assumption [A3], then the NLP multipliers are not bounded (and may not even exist). Despite this, SQP converges linearly to the solution in the example presented here, although quadratic convergence is lost.

Consider the following MPEC (`scholtes4.mod`) from MacMPEC (see also [19]):

$$(P) \qquad \begin{cases} \underset{z}{\text{minimize}} & z_1 + z_2 - z_0 \\ \text{subject to} & -4z_1 + z_0 \le 0, \\ & -4z_2 + z_0 \le 0, \\ & 0 \le z_1 \ \perp \ z_2 \ge 0\,, \end{cases}$$

whose optimal solution is $z^* = (0,0,0)^T$. Writing (P) as an NLP gives

$$(P') \qquad \begin{cases} \underset{z}{\text{minimize}} & z_1 + z_2 - z_0 & & \text{multiplier} \\ \text{subject to} & -4z_1 + z_0 \le 0, & & \lambda_1 \ge 0, \\ & -4z_2 + z_0 \le 0, & & \lambda_2 \ge 0, \\ & z_1 z_2 \le 0, & & \xi \ge 0, \\ & z_1 \ge 0, & & \nu_1 \ge 0, \\ & z_2 \ge 0, & & \nu_2 \ge 0. \end{cases}$$

Next, we show that SQP converges linearly for this problem.

PROPOSITION 7.1. *SQP applied to (P') generates the following sequence of iterates:*

$$z^{(k)} = \begin{pmatrix} 2^{2-k} \\ 2^{-k} \\ 2^{-k} \end{pmatrix},\ \lambda^{(k)} = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix},\ \xi^{(k)} = 2^{k-1} + \xi^{(k-1)}/2 = \sum_{j=0}^{k-1} 2^{(k-1)-2j}$$

for suitable starting values (e.g., $z = (4,1,1)^T$). Moreover, SQP converges linearly.

Proof. the proof is by induction. The assertion holds trivially for $k = 0$ (i.e., the starting point). Now assume the assertion holds for k, and show that it also holds for $k + 1$. At iteration k, SQP solves the following QP for a step d:

$$(QP^{(k)}) \qquad \begin{cases} \underset{z}{\text{minimize}} & d_1\xi^{(k)}d_2 + d_1 + d_2 - d_0 \\ \text{subject to} & -4d_1 + d_0 \le 0, \\ & -4d_2 + d_0 \le 0, \\ & z_1^{(k)}z_2^{(k)} + z_2^{(k)}d_1 + z_1^{(k)}d_2 \le 0, \\ & z_1^{(k)} + d_1 \ge 0, \\ & z_2^{(k)} + d_2 \ge 0. \end{cases}$$

We note that all QP approximations are consistent and that the first three constraints are active. Subtracting the second from the first constraint, we have that $d_1 = d_2$. Substituting into the third constraint, we get $d_1 = d_2 = -2^{-(k+1)}$, from which it follows that $d_0 = 4(-2^{-(k+1)})$. We verify the KKT conditions of $(QP^{(k)})$:

$$0 = \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix} + \begin{pmatrix} 0 \\ -2^{-(k+1)}\xi^{(k)} \\ -2^{-(k+1)}\xi^{(k)} \end{pmatrix} + \lambda_1 \begin{pmatrix} 1 \\ -4 \\ 0 \end{pmatrix} + \lambda_2 \begin{pmatrix} 1 \\ 0 \\ -4 \end{pmatrix} + \xi \begin{pmatrix} 0 \\ 2^{-k} \\ 2^{-k} \end{pmatrix}.$$

Subtracting the second from the first equation shows that $\lambda_1 = \lambda_2$. Substituting into the third equation then verifies that $\lambda_1^{(k+1)} = \lambda_2^{(k+1)} = \frac{1}{2}$.

Finally, the second equation shows $\xi^{(k)} = 2^{k-1} + \xi^{(k-1)}/2$, the recurrence relation for ξ. The explicit formula for ξ follows easily. The iterates clearly converge linearly to the solution. ∎

Note that (P) satisfies an MPEC-MFCQ [20] but violates an MPEC-LICQ (as can be seen easily by observing that four constraints are active at the solution). In addition, (P) fails to satisfy strong complementarity. For strong complementarity, it would be necessary that $\lambda_i \ge 0$ and $\nu_i \ge 0$, since $z_1 = z_2 = 0$. Checking the first-order condition,

$$0 = \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix} + \lambda_1 \begin{pmatrix} 1 \\ -4 \\ 0 \end{pmatrix} + \lambda_2 \begin{pmatrix} 1 \\ 0 \\ -4 \end{pmatrix} - \hat{\nu}_1 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} - \hat{\nu}_2 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix},$$

one can see that the system is underdetermined. Setting $\lambda_1 = t$, we obtain $\lambda_2 = 1 - t$, $\nu_1 = 1 - 4t$, and $\nu_2 = -3 + 4t$. The condition $\nu_i \ge 0$ now implies that $t \le \frac{1}{4}$ *and* $t \ge \frac{3}{4}$, which cannot hold simultaneously. Thus the solution of (P) is not strongly stationary.

The linear inequalities always ensure that $z_1^{(0)} = z_2^{(0)} \ge 0$, and the above analysis goes through for alternative starting points. It is not clear what would happen if we allowed $z_1 < 0$, but sensible NLP solvers will always project the starting point into the set of linear constraints (or at least the set of box constraints). The solvers filter, snopt, and lancelot behave in this way.

7.2. Formulations without slacks. The next example shows that SQP methods can converge to nonstationary points if slacks are not added to replace nonlinear complementarity conditions. Consider the following MPEC (`s12.mod`) from

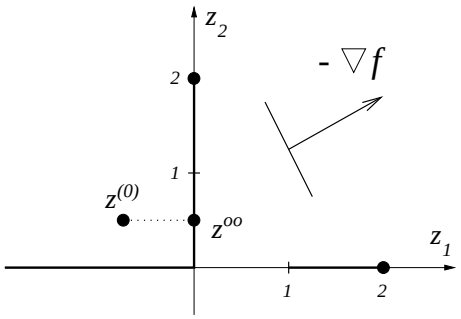


FIG. 3. *Example* `sl2`.

MacMPEC, which involves a nonlinear expression in the complementarity condition:

$$(P) \qquad \begin{cases} \underset{z}{\text{minimize}} & -z_1 - \frac{1}{2}z_2 \\ \text{subject to} & z_1 + z_2 \le 2, \\ & 0 \le z_1^2 - z_1 \ \perp \ z_2 \ge 0. \end{cases}$$

The problem has a global solution at $z^* = (2,0)^T$ with $f^* = -2$ and a local solution at $z^* = (0,2)^T$ with $f^* = -1$. Both solutions satisfy Assumptions [A1]–[A4]. The feasible set is illustrated by the bold lines in Figure 3.

Starting at $z^{(0)} = (-\epsilon, t)^T$ gives convergence to the nonstationary point $z^\infty = (0,t)^T$, where $t \ge 0$ is arbitrary. Moreover, one can show that $\xi \to \infty$ and that *both* the complementarity constraint *and* $0 \le z_1^2 + z_1$ remain in the active set. Thus, the active set is singular in the limit. Nevertheless, second-order convergence is observed!

It is straightforward to prove quadratic convergence to a nonstationary limit. Let $z^{(k)} = (-\epsilon, t)^T$ with $t \le 1$. Then the following problem is solved for a step of the SQP method:

$$(P) \qquad \begin{cases} \underset{d}{\text{minimize}} & -d_1 - \frac{1}{2}d_2 \\ \text{subject to} & d_1 + d_2 \le 2 + \epsilon - t, \\ & (\epsilon^2 + \epsilon) - (2\epsilon + 1)d_1 \ge 0, \\ & t + d_2 \ge 0, \\ & t(\epsilon^2 + \epsilon) - t(2\epsilon + 1)d_1 + (\epsilon^2 + \epsilon) \le 0, \end{cases}$$

whose solution is

$$d = \begin{pmatrix} \frac{\epsilon^2+\epsilon}{2\epsilon+1} \\ 0 \end{pmatrix}, \quad \xi = \frac{1}{2(\epsilon^2+\epsilon)}, \quad \nu_1 = \frac{1}{2\epsilon+1} + \xi t.$$

One can see that $z^{(k+1)} = (-\mathcal{O}(\epsilon^2), t)^T$ and quadratic convergence occurs to $z^\infty = (0,t)^T$. On the other hand, the multiplier ξ clearly diverges to infinity. Note that including the Hessian of the Lagrangian leads to a similar conclusion. This example shows that it is not sufficient to trigger the elastic mode only when QP become inconsistent. Clearly, the elastic mode is also required if the multipliers become too large. The introduction of slacks avoids the need for the elastic mode in this example.

When a slack variable is introduced, SQP converges quadratically. The SQP solver `filter` exhibits this behavior, while `lancelot` and `loqo` converge even for the problem *without* slacks. The reason for this apparently better behavior is that both introduce slacks internally before solving the problem!

Another reason for using slacks (rather than linear or even nonlinear complementarity) is that SQP solvers maintain linear feasibility throughout the iteration. Thus they *guarantee* that $z_1^{(k)} \geq 0,\ z_2^{(k)} \geq 0$ for all iterations k in *exact* arithmetic. In *inexact* arithmetic, one can truncate QP steps such that $z_1^{(k)} \geq 0,\ z_2^{(k)} \geq 0$ for all iterations k. This approach is *not* possible for general *linear* complementarity conditions even if iterative refinement were used.

Thus the use of slacks ensures that ${z_1^{(k)}}^T z_2^{(k)} \geq 0$ for all iterations k, and the trivial pitfall of [4], where it was observed that perturbing the right-hand side of the complementarity constraint to $-\epsilon$ renders an inconsistent QP, cannot occur.

7.3. Lack of second-order condition. The following MPEC (`ralph2.mod`) shows that if the second-order sufficient condition [A3] is violated, then SQP may converge only linearly:

$$(P) \qquad \begin{cases} \underset{z}{\text{minimize}} & z_1^2 + z_2^2 - 4z_1 z_2 \\ \text{subject to} & 0 \leq z_1 \ \perp \ z_2 \geq 0. \end{cases}$$

The problem has a global solution at $(0,0)$. Starting at $z = (1,1)$ causes SQP to converge linearly to the solution. Note that (P) also violates any upper-level strict complementarity condition.

The MPEC-SOSC is stronger than needed for MPECs in the sense that the set of directions over which positive curvature is required for SQP is larger than the set of MPEC-feasible directions. We illustrate this by the following example. The set of MPEC-feasible directions at $(0,0)$ is

$$S_M^* = \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\},$$

while the set of directions over which curvature is required to be positive for SQP to converge is the whole positive orthant (i.e., $\text{conv}(S_M^*)$). The linear rate of convergence is due to the fact that the curvature in the direction $(1,1)$ is negative.

8. Conclusions and future work. We have presented a convergence analysis that shows that SQP methods converge quadratically when applied to the NLP equivalent of an MPEC. This analysis goes some way toward explaining the extraordinary success of SQP solvers applied to MPECs, as we have observed. The result is remarkable because MPECs violate the MFCQ.

Conditions are identified under which local second-order convergence occurs. These conditions include the assumption that all QP approximations remain consistent. It can be shown that this assumption always holds if ${z_1^{(k)}}^T z_2^{(k)} = 0$ (i.e., for iterates which satisfy complementarity), and this is often observed in practice. We have also shown that MPECs whose lower-level problem is a certain vertical complementarity problem generate consistent QP approximations. Further we have given a restoration phase that ensures that this can always be guaranteed sufficiently close to a solution.

We have also experimented with an alternative to the restoration problem. In this approach, the linearization of the complementarity condition is relaxed as

$$(8.1) \qquad {z_1^{(k)}}^T z_2^{(k)} + {z_2^{(k)}}^T d_1 + {z_1^{(k)}}^T d_2 \leq \delta \left({z_1^{(k)}}^T z_2^{(k)} \right)^{1+\kappa},$$

where $0 < \delta, \kappa < 1$ are constants. Note that the perturbation to the right-hand side of the complementarity condition is $o(\|d_{NR}\|)$, where d_{NR} is the Newton step. This form of perturbation allows the superlinear convergence proof to be extended by virtue of the Dennis–Moré characterization theorem.

However, the perturbation alone is not sufficient to guarantee consistency of (QP^k). The following example illustrates the need for further assumptions. Consider the following feasible set:

$$z_1 + z_2 - 1 \geq 0\,, \quad 0 \leq z_1 \perp z_2 \geq 0.$$

It is easy to see that for any $z = (\epsilon^4, 1-\epsilon)$, the (QP^k) relaxed by using (8.1) with $\delta = \kappa = 0.5$ is inconsistent. Note that if we restrict our attention to points z that satisfy the linear constraints (e.g., $z = (\epsilon, 1-\epsilon)$), then (QP^k) using (8.1) is consistent in a neighborhood of $z = (0, 1)$. Thus (8.1) seems to ensure consistency of (QP^k) as long as $z^{(k)}$ satisfies the linearizations of $c_{\mathcal{E}}(z)$, $c_{\mathcal{I}}(z)$ about $z^{(k-1)}$. Unfortunately, we have been unable to prove any general results along those lines. Such a proof would clearly allow us to bootstrap a convergence of SQP for MPECs with the relaxed equation (8.1).

We finish this paper with some observations on the role of degeneracy and point to some future work. It has been observed that any QP approximation about a feasible point of (1.3) is degenerate. Moreover, approximations about points that satisfy $z_1^T z_2 = \epsilon > 0$ are near-degenerate, and we would expect this property to play a role in the SQP method. In our numerical experiment we use two SQP solvers, snopt and filter.

The solver `snopt` uses `EXPAND` to handle degeneracy. This procedure perturbs the bounds of (QP^k) to *remove* degeneracy. Some numerical experiments suggest that this is not the best way to handle degeneracy in the case of MPECs. The QP solver in filter, `bqpd`, applies a different methodology to handle degeneracy. It *creates* degeneracy whenever *near-degeneracy* is detected and then handles the degenerate situation. This approach has two implications:

1. If *exact degeneracy* exists (i.e., if ${z_1^{(k)}}^T z_2^{(k)} = 0$), then `bqpd` will deal with it.
2. If *near-degeneracy* exists (i.e., if ${z_1^{(k)}}^T z_2^{(k)} = \epsilon > 0$), then `bqpd` *creates degeneracy* by perturbing the bound ϵ to zero. This has the effect of *pushing* the solution onto the axis. As we have shown, this is a favorable situation for SQP methods.

Future work will focus on relaxing some assumptions and providing a global convergence analysis. Some numerical results suggest that SQP converges under even weaker assumptions than those made above, and it may be possible to pursue the ideas of [22] in this context. Another important question concerns the global convergence of SQP methods. Anitescu [1] provides a framework for convergence (possibly under additional assumptions) of Sl_∞QP methods. However, the numerical results suggest that a similar proof may be possible for filter methods.

REFERENCES

[1] M. Anitescu, *Global convergence of an elastic mode approach for a class of mathematical programs with complementarity constraints*, SIAM J. Optim., 16 (2005), pp. 120–145.
[2] J. F. Bard, *Convex two-level optimization*, Math. Program., 40 (1988), pp. 15–27.
[3] J. F. Bonnans, *Local convergence analysis of Newton-type methods for variational inequalities and nonlinear programming*, Appl. Math. Optim., 29 (1994), pp. 161–186.

[4] Y. Chen and M. Florian, *The nonlinear bilevel programming problem: Formulations, regularity and optimality conditions*, Optimization, 32 (1995), pp. 193–209.
[5] A. R. Conn, N. I. M. Gould, and Ph. L. Toint, *Numerical experiments with the* lancelot *package (Release A) for large–scale nonlinear optimization*, Math. Program., 73 (1996), pp. 73–110.
[6] A. R. Conn, N. I. M. Gould, and Ph. L. Toint, *Trust-Region Methods*, MPS/SIAM Ser. Optim. 1, SIAM, Philadelphia, 2000.
[7] M. C. Ferris and J. S. Pang, *Engineering and economic applications of complementarity problems*, SIAM Rev., 39 (1997), pp. 669–713.
[8] A. Fischer, *Modified Wilson method for nonlinear programs with nonunique multipliers*, Math. Oper. Res., 24 (1999), pp. 699–727.
[9] R. Fletcher and S. Leyffer, *Nonlinear programming without a penalty function*, Math. Program., 91 (2002), pp. 239–270.
[10] R. Fletcher and S. Leyffer, *Solving mathematical programs with complementarity constraints as nonlinear programs*, Optim. Methods Softw., 19 (2004), pp. 15–40.
[11] M. Fukushima and P. Tseng, *An implementable active-set algorithm for computing a B-stationary point of the mathematical program with linear complementarity constraints*, SIAM J. Optim., 12 (2002), pp. 724–739.
[12] P. E. Gill, W. Murray, and M. A. Saunders, *SNOPT: An SQP algorithm for large-scale constrained optimization*, SIAM J. Optim., 12 (2002), pp. 979–1006.
[13] W. W. Hager, *Stabilized sequential quadratic programming*, Comput. Optim. Appl., 12 (1999), pp. 253–273.
[14] H. Jiang and D. Ralph, *QPECgen, a MATLAB generator for mathematical programs with quadratic objectives and affine variational inequality constraints*, Comput. Optim. Appl., 13 (1999), pp. 25–59.
[15] S. Leyffer, *MacMPEC: AMPL collection of MPECs*; available online from www.mcs.anl.gov/∼leyffer/MacMPEC/, 2000.
[16] Z.-Q. Luo, J.-S. Pang, and D. Ralph, *Mathematical Programs with Equilibrium Constraints*, Cambridge University Press, Cambridge, UK, 1996.
[17] J. Outrata, M. Kocvara, and J. Zowe, *Nonsmooth Approach to Optimization Problems with Equilibrium Constraints*, Kluwer Academic, Dordrecht, The Netherlands, 1998.
[18] S. M. Robinson, *Perturbed Kuhn-Tucker points and rates of convergence for a class of nonlinear programming algorithms*, Math. Program., 7 (1974), pp. 1–16.
[19] H. Scheel and S. Scholtes, *Mathematical program with complementarity constraints: Stationarity, optimality and sensitivity*, Math. Oper. Res., 25 (2000), pp. 1–22.
[20] S. Scholtes, *Convergence properties of a regularization scheme for mathematical programs with complementarity constraints*, SIAM J. Optim., 11 (2001), pp. 918–936.
[21] S. Scholtes and M. Stöhr, *How stringent is the linear independence assumption for mathematical programs with complementarity constraints?*, Math. Oper. Res., 26 (2001), pp. 851–863.
[22] S. J. Wright, *Modifying SQP for degenerate problems*, SIAM J. Optim., 13 (2002), pp. 470–497.

SIAM J. OPTIM.
Vol. 17, No. 1, pp. 287–310
© 2006 Society for Industrial and Applied Mathematics

AN ITERATIVE SOLVER-BASED INFEASIBLE PRIMAL-DUAL PATH-FOLLOWING ALGORITHM FOR CONVEX QUADRATIC PROGRAMMING*

ZHAOSONG LU†, RENATO D. C. MONTEIRO‡, AND JEROME W. O'NEAL§

Abstract. In this paper we develop a long-step primal-dual infeasible path-following algorithm for convex quadratic programming (CQP) whose search directions are computed by means of a preconditioned iterative linear solver. We propose a new linear system, which we refer to as the *augmented normal equation* (ANE), to determine the primal-dual search directions. Since the condition number of the ANE coefficient matrix may become large for degenerate CQP problems, we use a maximum weight basis preconditioner introduced in [A. R. L. Oliveira and D. C. Sorensen, *Linear Algebra Appl.*, 394 (2005), pp. 1–24; M. G. C. Resende and G. Veiga, *SIAM J. Optim.*, 3 (1993), pp. 516–537; P. Vaida, *Solving Linear Equations with Symmetric Diagonally Dominant Matrices by Constructing Good Preconditioners*, Tech. report, Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, 1990] to precondition this matrix. Using a result obtained in [R. D. C. Monteiro, J. W. O'Neal, and T. Tsuchiya, *SIAM J. Optim.*, 15 (2004), pp. 96–100], we establish a uniform bound, depending only on the CQP data, for the number of iterations needed by the iterative linear solver to obtain a sufficiently accurate solution to the ANE. Since the iterative linear solver can generate only an approximate solution to the ANE, this solution does not yield a primal-dual search direction satisfying all equations of the primal-dual Newton system. We propose a way to compute an inexact primal-dual search direction so that the equation corresponding to the primal residual is satisfied exactly, while the one corresponding to the dual residual contains a manageable error which allows us to establish a polynomial bound on the number of iterations of our method.

Key words. convex quadratic programming, iterative linear solver, maximum weight basis preconditioner, primal-dual path-following methods, interior-point methods, augmented normal equation, inexact search directions, polynomial convergence

AMS subject classifications. 65F10, 65F35, 90C20, 90C25, 90C51

DOI. 10.1137/04060771X

1. Introduction.

In this paper we develop an interior-point long-step primal-dual infeasible path-following (PDIPF) algorithm for convex quadratic programming (CQP) whose search directions are computed by means of an iterative linear solver. We will refer to this algorithm as an *inexact* algorithm, in the sense that the Newton system which determines the search direction will be solved only approximately at each iteration. The problem we consider is

$$\min_x \left\{ \frac{1}{2} x^T Q x + c^T x : \quad Ax = b,\ x \geq 0 \right\}, \tag{1}$$

*Received by the editors May 3, 2004; accepted for publication (in revised form) December 5, 2005; published electronically May 19, 2006.
http://www.siam.org/journals/siopt/17-1/60771.html

†Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213-3890 (zhaolu@andrew.cmu.edu). This author was supported in part by NSF grant CCR-0203113 and ONR grant N00014-03-1-0401.

‡School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0205 (monteiro@isye.gatech.edu). This author was supported in part by NSF grants CCR-0203113, CCF-0430644, and INT-9910084 and ONR grants N00014-03-1-0401 and N00014-05-1-0183.

§Research, Modelling, & Design Group, Delta Technology, 1001 International Boulevard, Department 709, Atlanta, GA 30354 (jerome.w.oneal@delta.com). This author was supported in part by the NDSEG Fellowship Program sponsored by the Department of Defense.

where the data are $Q \in \Re^{n\times n}$, $A \in \Re^{m\times n}$, $b \in \Re^m$, and $c \in \Re^n$, and the decision vector is $x \in \Re^n$. We also assume that Q is positive semidefinite and that a factorization $Q = VE^2V^T$ is explicitly given, where $V \in \Re^{n\times l}$ and E is an $l \times l$ positive diagonal matrix.

A similar algorithm for solving the special case of linear programming (LP), i.e., problem (1) with $Q = 0$, was developed by Monteiro and O'Neal in [16]. The algorithm studied in [16] is essentially the long-step PDIPF algorithm studied in [9, 28], the only difference being that the search directions are computed by means of an iterative linear solver. We refer to the iterations of the iterative linear solver as the *inner iterations* and to the ones performed by the interior-point method itself as the *outer iterations*. The main step of the algorithm studied in [9, 16, 28] is the computation of the primal-dual search direction $(\Delta x, \Delta s, \Delta y)$, whose Δy component can be found by solving a system of the form $AD^2A^T\Delta y = g$, referred to as the *normal equation*, where $g \in \Re^m$ and the positive diagonal matrix D depends on the current primal-dual iterate. In contrast to [9, 28], the algorithm studied in [16] uses an iterative linear solver to obtain an approximate solution to the normal equation. Since the condition number of the normal matrix AD^2A^T may become excessively large on degenerate LP problems (see e.g., [13]), the maximum weight basis (MWB) preconditioner T introduced in [19, 22, 25] is used to better condition this matrix, and an approximate solution of the resulting equivalent system with coefficient matrix $TAD^2A^TT^T$ is then computed. By using a result obtained in [17], which establishes that the condition number of $TAD^2A^TT^T$ is uniformly bounded by a quantity depending only on A, Monteiro and O'Neal [16] showed that the number of inner iterations of the algorithm in [16] can be uniformly bounded by a constant depending on n and A.

In the case of CQP, the standard normal equation takes the form

$$A(Q + X^{-1}S)^{-1}A^T\Delta y = g \tag{2}$$

for some vector g. When Q is not diagonal, the matrix $(Q+X^{-1}S)^{-1}$ is not diagonal, and hence the coefficient matrix of (2) does not have the form required for the result of [17] to hold. To remedy this difficulty, we develop in this paper a new linear system, referred to as the *augmented normal equation* (ANE), to determine a portion of the primal-dual search direction. This equation has the form $\tilde{A}\tilde{D}^2\tilde{A}^Tu = w$, where $w \in \Re^{m+l}$, $\tilde{D}$ is an $(n+l)\times(n+l)$ positive diagonal matrix, and $\tilde{A}$ is a 2×2 block matrix of dimension $(m+l)\times(n+l)$ whose blocks consist of A, V^T, the zero matrix, and the identity matrix (see (21)). As was done in [16], a MWB preconditioner $\tilde{T}$ for the ANE is computed and an approximate solution of the resulting preconditioned equation with coefficient matrix $\tilde{T}\tilde{A}\tilde{D}^2\tilde{A}^T\tilde{T}^T$ is generated using an iterative linear solver. Using the result of [17], which claims that the condition number of $\tilde{T}\tilde{A}\tilde{D}^2\tilde{A}^T\tilde{T}^T$ is uniformly bounded regardless of $\tilde{D}$, we obtain a uniform bound (depending only on $\tilde{A}$) on the number of inner iterations performed by the iterative linear solver to find a desirable approximate solution to the ANE (see Theorem 3.5).

Since the iterative linear solver can generate only an approximate solution to the ANE, it is clear that not all equations of the Newton system, which determines the primal-dual search direction, can be satisfied simultaneously. In the context of LP, Monteiro and O'Neal [16] proposed a recipe to compute an inexact primal-dual search direction so that the equations of the Newton system corresponding to the primal and dual residuals were both satisfied. In the context of CQP, such an approach is no longer possible. Instead, we propose a way to compute an inexact primal-dual search direction so that the equation corresponding to the primal residual is satisfied

exactly, while the one corresponding to the dual residual contains a manageable error which allows us to establish a polynomial bound on the number of outer iterations of our method. Interestingly, the presence of this error on the dual residual equation implies that the primal and dual residuals go to zero at different rates. This is a unique feature of the convergence analysis of our algorithm in that it contrasts with the analysis of other interior-point PDIPF algorithms, where the primal and dual residuals are required to go to zero at the same rate.

The use of inexact search directions in interior-point methods has been extensively studied in the context of cone programming problems (see e.g., [1, 2, 7, 11, 12, 15, 18, 29]). Moreover, the use of iterative linear solvers to compute the primal-dual Newton search directions of interior-point path-following algorithms has also been extensively investigated in [1, 3, 4, 7, 12, 18, 19, 20, 22, 24]. For feasibility problems of the form $\{x \in \mathcal{H}_1 : \mathcal{A}x = b, x \in \mathcal{C}\}$, where $\mathcal{H}_1$ and $\mathcal{H}_2$ are Hilbert spaces, $\mathcal{C} \subseteq \mathcal{H}_1$ is a closed convex cone satisfying some mild assumptions, and $\mathcal{A} : \mathcal{H}_1 \to \mathcal{H}_2$ is a continuous linear operator, Renegar [21] has proposed an interior-point method where the Newton system that determines the search directions is approximately solved by performing a uniformly bounded number of iterations of the conjugate gradient (CG) method. To our knowledge, no one has used the ANE system in the context of CQP to obtain either an exact or inexact primal-dual search direction.

Our paper is organized as follows. In subsection 1.1, we give the terminology and notation which will be used throughout our paper. Section 2 describes the outer iteration framework for our algorithm and the complexity results we have obtained for it, along with presenting the ANE as a means to determine the search direction. In section 3, we discuss the use of iterative linear solvers to obtain a suitable approximate solution to the ANE and the construction of an inexact search direction based on this solution. Section 4 gives the proofs of the results presented in sections 2 and 3. Finally, we present some concluding remarks in section 5.

1.1. Terminology and notation. Throughout this paper, uppercase roman letters denote matrices, lowercase roman letters denote vectors, and lowercase Greek letters denote scalars. We let $\Re^n$, $\Re^n_+$, and $\Re^n_{++}$ denote the set of n-dimensional vectors having real, nonnegative real, and positive real components, respectively. Also, we let $\Re^{m\times n}$ denote the set of $m \times n$ matrices with real entries. For a vector $v \in \Re^n$, we let $|v|$ denote the vector whose ith component is $|v_i|$ for every $i = 1, \ldots, n$, and we let $\mathrm{Diag}(v)$ denote the diagonal matrix whose ith diagonal element is v_i for every $i = 1, \ldots, n$. In addition, given vectors $u \in \Re^m$ and $v \in \Re^n$, we denote by (u, v) the vector $(u^T, v^T)^T \in \Re^{m+n}$.

Certain matrices bear special notation, namely the matrices X, ΔX, S, D, and $\tilde{D}$. These matrices are the diagonal matrices corresponding to the vectors x, Δx, s, d, and $\tilde{d}$, respectively, as described in the previous paragraph. The symbol 0 will be used to denote a scalar, vector, or matrix of all zeros; its dimensions should be clear from the context. Also, we denote by e the vector of all 1's, and by I the identity matrix; their dimensions should be clear from the context.

For a symmetric positive definite matrix W, we denote its condition number by $\kappa(W)$, i.e., its maximum eigenvalue divided by its minimum eigenvalue. We will denote sets by uppercase calligraphic letters (e.g., $\mathcal{B}, \mathcal{N}$). For a finite set $\mathcal{B}$, we denote its cardinality by $|\mathcal{B}|$. Given a matrix $A \in \Re^{m\times n}$ and an ordered set $\mathcal{B} \subseteq \{1, \ldots, n\}$, we let $A_{\mathcal{B}}$ denote the submatrix whose columns are $\{A_i : i \in \mathcal{B}\}$ arranged in the same order as $\mathcal{B}$. Similarly, given a vector $v \in \Re^n$ and an ordered set $\mathcal{B} \subseteq \{1, \ldots, n\}$, we let $v_{\mathcal{B}}$ denote the subvector consisting of the elements $\{v_i : i \in \mathcal{B}\}$ arranged in the same

order as $\mathcal{B}$.

We will use several different norms throughout the paper. For a vector $z \in \Re^n$, $\|z\| = \sqrt{z^T z}$ is the Euclidian norm, $\|z\|_1 = \sum_{i=1}^n |z_i|$ is the "1-norm," and $\|z\|_\infty = \max_{i=1,\ldots,n} |z_i|$ is the "infinity norm." For a matrix $V \in \Re^{m\times n}$, $\|V\|$ denotes the operator norm associated with the Euclidian norm: $\|V\| = \max_{z:\|z\|=1} \|Vz\|$. Finally, $\|V\|_F$ denotes the Frobenius norm: $\|V\|_F = (\sum_{i=1}^m \sum_{j=1}^n V_{ij}^2)^{1/2}$.

2. Outer iteration framework. In this section, we introduce our PDIPF algorithm based on a class of inexact search directions and discuss its iteration complexity. This section is divided into two subsections. In subsection 2.1, we discuss an exact PDIPF algorithm, which will serve as the basis for the inexact PDIPF algorithm given in subsection 2.2, and we give its iteration complexity result. We also present an approach based on the ANE to determine the Newton search direction for the exact algorithm. To motivate the class of inexact search directions used by our inexact PDIPF algorithm, we describe in subsection 2.2 a framework for computing an inexact search direction based on an approximate solution to the ANE. We then introduce the class of inexact search directions, state a PDIPF algorithm based on it, and give its iteration complexity result.

2.1. An exact PDIPF algorithm and the ANE. Consider the following primal-dual pair of CQP problems:

$$\min_x \left\{ \frac{1}{2} x^T V E^2 V^T x + c^T x : \quad Ax = b,\ x \geq 0 \right\}, \tag{3}$$

$$\max_{(\hat{x},s,y)} \left\{ -\frac{1}{2} \hat{x}^T V E^2 V^T \hat{x} + b^T y : \quad A^T y + s - V E^2 V^T \hat{x} = c,\ s \geq 0 \right\}, \tag{4}$$

where the data are $V \in \Re^{n\times l}$, $E \in \mathrm{Diag}(\Re^l_{++})$, $A \in \Re^{m\times n}$, $b \in \Re^m$, and $c \in \Re^n$, and the decision variables are $x \in \Re^n$ and $(\hat{x}, s, y) \in \Re^n \times \Re^n \times \Re^m$. We observe that the Hessian matrix Q is already given in factored form $Q = VE^2V^T$.

It is well known that if x^* is an optimal solution for (3) and $(\hat{x}^*, s^*, y^*)$ is an optimal solution for (4), then (x^*, s^*, y^*) is also an optimal solution for (4). Now, let $\mathcal{S}$ denote the set of all vectors $w := (x, s, y, z) \in \Re^{2n+m+l}$ satisfying

$$Ax = b, \quad x \geq 0, \tag{5}$$

$$A^T y + s + Vz = c, \quad s \geq 0, \tag{6}$$

$$Xs = 0, \tag{7}$$

$$EV^T x + E^{-1} z = 0. \tag{8}$$

It is clear that $w \in \mathcal{S}$ if and only if x is optimal for (3), (x, s, y) is optimal for (4), and $z = -E^2V^Tx$. (Throughout this paper, the symbol w will always denote the quadruple (x, s, y, z), where the vectors lie in the appropriate dimensions; similarly, $\Delta w = (\Delta x, \Delta s, \Delta y, \Delta z)$, $w^k = (x^k, s^k, y^k, z^k)$, $\bar{w} = (\bar{x}, \bar{s}, \bar{y}, \bar{z})$, etc.)

We observe that the presentation of the PDIPF algorithm based on exact Newton search directions in this subsection differs from the classical way of presenting it in that we introduce an additional variable z as above. Clearly, it is easy to see that the variable z is completely redundant and can be eliminated, thereby reducing the method described below to the usual way of presenting it. The main reason for introducing the variable z is due to the development of the ANE presented at the end of this subsection.

We will make the following two assumptions throughout the paper.

Assumption 1. A has full row rank.

Assumption 2. The set $\mathcal{S}$ is nonempty.

For a point $w \in \Re^{2n}_{++} \times \Re^{m+l}$, let us define

$$\mu := \mu(w) \;=\; x^T s/n, \tag{9}$$

$$r_p := r_p(w) \;=\; Ax - b, \tag{10}$$

$$r_d := r_d(w) \;=\; A^T y + s + Vz - c, \tag{11}$$

$$r_V := r_V(w) \;=\; EV^T x + E^{-1} z, \tag{12}$$

$$r := r(w) \;=\; (r_p(w), r_d(w), r_V(w)). \tag{13}$$

Moreover, given $\gamma \in (0,1)$ and an initial point $w^0 \in \Re^{2n}_{++} \times \Re^{m+l}$, we define the following neighborhood of the central path:

$$\mathcal{N}_{w^0}(\gamma) := \left\{ w \in \Re^{2n}_{++} \times \Re^{m+l} : Xs \geq (1-\gamma)\mu e,\ r = \eta r^0 \right. \\ \left. \text{for some } 0 \leq \eta \leq \min\left[1, \frac{\mu}{\mu_0}\right] \right\}, \tag{14}$$

where $r := r(w)$, $r^0 := r(w^0)$, $\mu := \mu(w)$, and $\mu_0 := \mu(w^0)$.

We are now ready to state the PDIPF algorithm based on exact Newton search directions.

Exact PDIPF algorithm.

1. **Start:** Let $\epsilon > 0$ and $0 < \underline{\sigma} \leq \overline{\sigma} < 1$ be given. Let $\gamma \in (0,1)$ and $w^0 \in \Re^{2n}_{++} \times \Re^{m+l}$ be such that $w^0 \in \mathcal{N}_{w^0}(\gamma)$. Set $k = 0$.
2. **While** $\mu_k := \mu(w^k) > \epsilon$ **do**
   (a) Let $w := w^k$ and $\mu := \mu_k$; choose $\sigma := \sigma_k \in [\underline{\sigma}, \overline{\sigma}]$.
   (b) Let $\Delta w = (\Delta x, \Delta s, \Delta y, \Delta z)$ denote the solution of the linear system

$$A\Delta x = -r_p, \tag{15}$$

$$A^T \Delta y + \Delta s + V \Delta z = -r_d, \tag{16}$$

$$X\Delta s + S\Delta x = -Xs + \sigma\mu e, \tag{17}$$

$$EV^T \Delta x + E^{-1}\Delta z = -r_V. \tag{18}$$

   (c) Let $\tilde{\alpha} = \operatorname{argmax}\{\alpha \in [0,1] : w + \alpha'\Delta w \in \mathcal{N}_{w^0}(\gamma),\ \forall \alpha' \in [0,\alpha]\}$.
   (d) Let $\bar{\alpha} = \operatorname{argmin}\left\{(x + \alpha\Delta x)^T (s + \alpha\Delta s) : \alpha \in [0, \tilde{\alpha}]\right\}$.
   (e) Let $w^{k+1} = w + \bar{\alpha}\Delta w$, and set $k \leftarrow k + 1$.

   End (while)

A proof of the following result, under slightly different assumptions, can be found in [28].

Theorem 2.1. *Assume that the constants γ, $\underline{\sigma}$, and $\overline{\sigma}$ are such that*

$$\max\left\{\gamma^{-1}, (1-\gamma)^{-1}, \underline{\sigma}^{-1}, (1-\overline{\sigma})^{-1}\right\} = \mathcal{O}(1),$$

and that the initial point $w^0 \in \Re^{2n}_{++} \times \Re^{m+l}$ satisfies $(x^0, s^0) \geq (x^*, s^*)$ for some $w^* \in \mathcal{S}$. Then, the exact PDIPF algorithm finds an iterate $w^k \in \Re^{2n}_{++} \times \Re^{m+l}$ satisfying $\mu_k \leq \epsilon\mu_0$ and $\|r^k\| \leq \epsilon\|r^0\|$ within $\mathcal{O}(n^2 \log(1/\epsilon))$ iterations.

A few approaches have been suggested in the literature for computing the Newton search direction (15)–(18). Instead of using one of them, we will discuss below a new

approach, referred to in this paper as the ANE approach, that we believe to be suitable not only for direct solvers but especially for iterative linear solvers, as we will see in section 3.

Let us begin by defining the following matrices:

$$D := X^{1/2}S^{-1/2}, \tag{19}$$

$$\tilde{D} := \begin{pmatrix} D & 0 \\ 0 & E^{-1} \end{pmatrix} \in \Re^{(n+l)\times(n+l)}, \tag{20}$$

$$\tilde{A} := \begin{pmatrix} A & 0 \\ V^T & I \end{pmatrix} \in \Re^{(m+l)\times(n+l)}. \tag{21}$$

Suppose that we first solve the following system of equations for $(\Delta y, \Delta z)$:

$$\tilde{A}\tilde{D}^2\tilde{A}^T \begin{pmatrix} \Delta y \\ \Delta z \end{pmatrix} = \tilde{A} \begin{pmatrix} x - \sigma\mu S^{-1}e - D^2 r_d \\ 0 \end{pmatrix} + \begin{pmatrix} -r_p \\ -E^{-1}r_V \end{pmatrix} =: h. \tag{22}$$

This system is what we refer to as the ANE. Next, we obtain Δs and Δx according to

$$\Delta s = -r_d - A^T\Delta y - V\Delta z, \tag{23}$$

$$\Delta x = -D^2\Delta s - x + \sigma\mu S^{-1}e. \tag{24}$$

Clearly, the search direction $\Delta w = (\Delta x, \Delta s, \Delta y, \Delta z)$ computed as above satisfies (16) and (17) in view of (23) and (24). Moreover, it also satisfies (15) and (18) due to the fact that by (20)–(24), we have that

$$\begin{aligned}
\tilde{A} \begin{pmatrix} \Delta x \\ E^{-2}\Delta z \end{pmatrix} &= \tilde{A} \begin{pmatrix} -D^2\Delta s - x + \sigma\mu S^{-1}e \\ E^{-2}\Delta z \end{pmatrix} \\
&= \tilde{A} \begin{pmatrix} D^2 r_d + D^2 A^T\Delta y + D^2 V\Delta z - x + \sigma\mu S^{-1}e \\ E^{-2}\Delta z \end{pmatrix} \\
&= \tilde{A} \begin{pmatrix} D^2 A^T\Delta y + D^2 V\Delta z \\ E^{-2}\Delta z \end{pmatrix} + \tilde{A} \begin{pmatrix} D^2 r_d - x + \sigma\mu S^{-1}e \\ 0 \end{pmatrix} \\
&= \tilde{A}\tilde{D}^2\tilde{A}^T \begin{pmatrix} \Delta y \\ \Delta z \end{pmatrix} + \tilde{A} \begin{pmatrix} D^2 r_d - x + \sigma\mu S^{-1}e \\ 0 \end{pmatrix} \\
&= \begin{pmatrix} -r_p \\ -E^{-1}r_V \end{pmatrix}. \qquad (25)
\end{aligned}$$

Theorem 2.1 assumes that Δw is the exact solution of (22), which is usually obtained by computing the Cholesky factorization of the coefficient matrix of the ANE. In this paper, we will consider a variant of the exact PDIPF algorithm whose search directions are approximate solutions of (22) and ways of determining these inexact search directions by means of a suitable preconditioned iterative linear solver.

2.2. An inexact PDIPF algorithm for CQP. In this subsection, we describe a PDIPF algorithm based on a family of search directions that are approximate solutions to (15)–(18) and discuss its iteration complexity properties.

Clearly, an approximate solution to the ANE can yield only an approximate solution to (15)–(18). In order to motivate the class of inexact search directions used by the PDIPF algorithm presented in this subsection, we present a framework for

obtaining approximate solutions to (15)–(18) based on an approximate solution to the ANE.

Suppose that the ANE is solved only inexactly, i.e., that the vector $(\Delta y, \Delta z)$ satisfies

$$\tilde{A}\tilde{D}^2\tilde{A}^T \begin{pmatrix} \Delta y \\ \Delta z \end{pmatrix} = h + f \tag{26}$$

for some error vector f. If Δs and Δx were computed by (23) and (24), respectively, then it is clear that the search direction Δw would satisfy (16) and (17). However, (15) and (18) would not be satisfied, since by an argument similar to (25), we would have that

$$\begin{aligned} \tilde{A} \begin{pmatrix} \Delta x \\ E^{-2}\Delta z \end{pmatrix} = \cdots &= \tilde{A}\tilde{D}^2\tilde{A}^T \begin{pmatrix} \Delta y \\ \Delta z \end{pmatrix} + \tilde{A} \begin{pmatrix} D^2 r_d - x + \sigma\mu S^{-1} e \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} -r_p \\ -E^{-1} r_V \end{pmatrix} + f. \end{aligned}$$

Instead, suppose we use (23) to determine Δs as before, but now we determine Δx as

$$\Delta x = -D^2\Delta s - x + \sigma\mu S^{-1} e - S^{-1} p, \tag{27}$$

where the correction vector $p \in \Re^n$ will be required to satisfy some conditions which we will now describe.

To motivate the conditions on p, we note that (23), (26), and (27) imply that

(28)

$$\begin{aligned} &\tilde{A} \begin{pmatrix} \Delta x \\ E^{-2}\Delta z \end{pmatrix} + \begin{pmatrix} r_p \\ E^{-1} r_V \end{pmatrix} \\ &= \tilde{A} \begin{pmatrix} -D^2\Delta s - x + \sigma\mu S^{-1} e - S^{-1} p \\ E^{-2}\Delta z \end{pmatrix} + \begin{pmatrix} r_p \\ E^{-1} r_V \end{pmatrix} \\ &= \tilde{A} \begin{pmatrix} D^2 r_d + D^2 A^T \Delta y + D^2 V \Delta z - x + \sigma\mu S^{-1} e - S^{-1} p \\ E^{-2}\Delta z \end{pmatrix} + \begin{pmatrix} r_p \\ E^{-1} r_V \end{pmatrix} \\ &= \tilde{A}\tilde{D}^2 \begin{pmatrix} A^T \Delta y + V \Delta z \\ \Delta z \end{pmatrix} + \tilde{A} \begin{pmatrix} D^2 r_d - x + \sigma\mu S^{-1} e \\ 0 \end{pmatrix} - \tilde{A} \begin{pmatrix} S^{-1} p \\ 0 \end{pmatrix} \\ &\qquad + \begin{pmatrix} r_p \\ E^{-1} r_V \end{pmatrix} \\ &= \tilde{A}\tilde{D}^2\tilde{A}^T \begin{pmatrix} \Delta y \\ \Delta z \end{pmatrix} + \tilde{A} \begin{pmatrix} D^2 r_d - x + \sigma\mu S^{-1} e \\ 0 \end{pmatrix} - \tilde{A} \begin{pmatrix} S^{-1} p \\ 0 \end{pmatrix} + \begin{pmatrix} r_p \\ E^{-1} r_V \end{pmatrix} \\ &= f - \tilde{A} \begin{pmatrix} S^{-1} p \\ 0 \end{pmatrix}. \end{aligned}$$

Based on the above equation, one is naturally tempted to choose p so that the right-hand side of (28) is zero, and consequently (15) and (18) are satisfied exactly. However, the existence of such p cannot be guaranteed and, even if it exists, its magnitude might not be sufficiently small to yield a search direction which is suitable for the development of a polynomially convergent algorithm. Instead, we consider an alternative approach where p is chosen so that the first component of (28) is zero and the second

component is small. More specifically, by partitioning $f = (f_1, f_2) \in \Re^m \times \Re^l$, we choose $p \in \Re^n$ such that

$$AS^{-1}p = f_1. \tag{29}$$

It is clear that p is not uniquely defined. Note that (21) implies that (29) is equivalent to

$$f \;=\; \tilde{A}\begin{pmatrix} S^{-1}p \\ E^{-1}q \end{pmatrix}, \tag{30}$$

where $q := E(f_2 - V^T S^{-1}p)$. Then, using (21), (28), and (30), we conclude that

$$\begin{aligned} \tilde{A}\begin{pmatrix} \Delta x \\ E^{-2}\Delta z \end{pmatrix} + \begin{pmatrix} r_p \\ E^{-1}r_V \end{pmatrix} &= f - \tilde{A}\begin{pmatrix} S^{-1}p \\ E^{-1}q \end{pmatrix} + \tilde{A}\begin{pmatrix} 0 \\ E^{-1}q \end{pmatrix} \\ &= \tilde{A}\begin{pmatrix} 0 \\ E^{-1}q \end{pmatrix} = \begin{pmatrix} 0 \\ E^{-1}q \end{pmatrix}, \end{aligned} \tag{31}$$

from which we see that the first component of (28) is set to 0 and the second component is exactly $E^{-1}q$.

In view of (23), (27), and (31), the above construction yields a search direction Δw satisfying the following modified Newton system of equations:

$$A\Delta x = -r_p, \tag{32}$$

$$A^T\Delta y + \Delta s + V\Delta z = -r_d, \tag{33}$$

$$X\Delta s + S\Delta x = -Xs + \sigma\mu e - p, \tag{34}$$

$$EV^T\Delta x + E^{-1}\Delta z = -r_V + q. \tag{35}$$

As far as the outer iteration complexity analysis of our algorithm is concerned, all we require of our inexact search directions is that they satisfy (32)–(35) and that p and q be relatively small in the following sense.

DEFINITION 1. *Given a point $w \in \Re^{2n}_{++} \times \Re^{m+l}$ and positive scalars τ_p and τ_q, an inexact direction Δw is referred to as a* (τ_p, τ_q)-search direction *if it satisfies* (32)–(35) *for some p and q satisfying $\|p\|_\infty \le \tau_p\mu$ and $\|q\| \le \tau_q\sqrt{\mu}$, where μ is given by* (9).

We next define a generalized central path neighborhood which is used by our inexact PDIPF algorithm. Given a starting point $w^0 \in \Re^{2n}_{++} \times \Re^{m+l}$ and parameters $\eta \ge 0$, $\gamma \in [0,1]$, and $\theta > 0$, define the following set:

$$\mathcal{N}_{w^0}(\eta,\gamma,\theta) = \left\{ w \in \Re^{2n}_{++} \times \Re^{m+l} : \begin{array}{ccc} Xs \ge (1-\gamma)\mu e, & & (r_p, r_d) = \eta(r_p^0, r_d^0), \\ \|r_V - \eta r_V^0\| \le \theta\sqrt{\mu}, & & \eta \le \mu/\mu_0 \end{array} \right\}, \tag{36}$$

where $\mu = \mu(w)$, $\mu_0 = \mu(w^0)$, $r = r(w)$, and $r^0 = r(w^0)$. The generalized central path neighborhood is then given by

$$\mathcal{N}_{w^0}(\gamma,\theta) \;=\; \bigcup_{\eta\in[0,1]} \mathcal{N}_{w^0}(\eta,\gamma,\theta). \tag{37}$$

We observe that the neighborhood given by (37) agrees with the neighborhood given by (15) when $\theta = 0$.

We are now ready to state our inexact PDIPF algorithm.

INEXACT PDIPF ALGORITHM.

1. **Start:** Let $\epsilon > 0$ and $0 < \underline{\sigma} \leq \overline{\sigma} < 4/5$ be given. Choose $\gamma \in (0,1)$, $\theta > 0$, and $w^0 \in \Re^{2n}_{++} \times \Re^{m+l}$ such that $w^0 \in \mathcal{N}_{w^0}(\gamma, \theta)$. Set $k = 0$.
2. **While** $\mu_k := \mu(w^k) > \epsilon$ **do**
   (a) Let $w := w^k$ and $\mu := \mu_k$; choose $\sigma \in [\underline{\sigma}, \overline{\sigma}]$.
   (b) Set

$$\tau_p = \gamma\sigma/4 \quad \text{and} \tag{38}$$

$$\tau_q = \left[\sqrt{1 + (1 - 0.5\gamma)\,\sigma} - 1\right]\theta. \tag{39}$$

   (c) Set $r_p = Ax - b$, $r_d = A^T y + s + Vz - c$, $r_V = EV^T x + E^{-1}z$, and $\eta = \|r_p\|/\|r_p^0\|$.
   (d) Compute a (τ_p, τ_q)-search direction Δw.
   (e) Compute $\tilde{\alpha} := \operatorname{argmax}\{\alpha \in [0,1] : w + \alpha'\Delta w \in \mathcal{N}_{w^0}(\gamma,\theta),\ \forall \alpha' \in [0,\alpha]\}$.
   (f) Compute $\bar{\alpha} := \operatorname{argmin}\{(x + \alpha\Delta x)^T(s + \alpha\Delta s) : \alpha \in [0,\tilde{\alpha}]\}$.
   (g) Let $w^{k+1} = w + \bar{\alpha}\Delta w$, and set $k \leftarrow k + 1$.

   End (while)

The following result gives a bound on the number of iterations needed by the inexact PDIPF algorithm to obtain an ϵ-solution to the KKT conditions (5)–(8). Its proof will be given in subsection 4.2.

THEOREM 2.2. *Assume that the constants γ, $\underline{\sigma}$, $\overline{\sigma}$, and θ are such that*

$$\max\left\{\gamma^{-1},\, (1-\gamma)^{-1},\, \underline{\sigma}^{-1},\, \left(1 - \frac{5}{4}\overline{\sigma}\right)^{-1}\right\} = \mathcal{O}(1),\ \ \theta = \mathcal{O}(\sqrt{n}), \tag{40}$$

and that the initial point $w^0 \in \Re^{2n}_{++} \times \Re^{m+l}$ satisfies $(x^0, s^0) \geq (x^*, s^*)$ for some $w^* \in \mathcal{S}$. Then, the inexact PDIPF algorithm generates an iterate $w^k \in \Re^{2n}_{++} \times \Re^{m+l}$ satisfying $\mu_k \leq \epsilon\mu_0$, $\|(r_p^k, r_d^k)\| \leq \epsilon\|(r_p^0, r_d^0)\|$, and $\|r_V^k\| \leq \epsilon\|r_V^0\| + \epsilon^{1/2}\theta\mu_0^{1/2}$ within $\mathcal{O}\left(n^2 \log(1/\epsilon)\right)$ iterations.

3. Determining an inexact search direction via an iterative solver. The results in subsection 2.2 assume we can obtain a (τ_p, τ_q)-search direction Δw, where τ_p and τ_q are given by (38) and (39), respectively. In this section, we will describe a way to obtain a (τ_p, τ_q)-search direction Δw using a uniformly bounded number of iterations of a suitable preconditioned iterative linear solver applied to the ANE. It turns out that the construction of this Δw is based on the recipe given at the beginning of subsection 2.2, together with a specific choice of the perturbation vector p.

This section is divided into two subsections. In subsection 3.1, we introduce the MWB preconditioner which will be used to precondition the ANE. In addition, we also introduce a family of iterative linear solvers used to solve the preconditioned ANE. Subsection 3.2 gives a specific approach for constructing a pair (p, q) satisfying (30), and an approximate solution to the ANE so that the recipe described at the beginning of subsection 2.2 yields a (τ_p, τ_q)-search direction Δw. It also provides a uniform bound on the number of iterations that any member of the family of iterative linear solvers needs to perform to obtain such a direction Δw when applied to the preconditioned ANE.

3.1. MWB preconditioner and a family of solvers. In this subsection we introduce the MWB preconditioner, and we discuss its use as a preconditioner in

solving the ANE via a family of iterative linear solvers. Since the condition number of the ANE matrix $\tilde{A}\tilde{D}^2\tilde{A}^T$ may "blow up" for points w near an optimal solution, the direct application of a generic iterative linear solver for solving the ANE without first preconditioning it is generally not effective. We discuss a natural remedy to this problem which consists of using a preconditioner $\tilde{T}$, namely the MWB preconditioner, such that $\kappa(\tilde{T}\tilde{A}\tilde{D}^2\tilde{A}^T\tilde{T}^T)$ remains uniformly bounded regardless of the iterate w. Finally, we analyze the complexity of the resulting approach to obtain a suitable approximate solution to the ANE.

We start by describing the MWB preconditioner. Its construction essentially consists of building a basis B of $\tilde{A}$ which gives higher priority to the columns of $\tilde{A}$ corresponding to larger diagonal elements of $\tilde{D}$. More specifically, the MWB preconditioner is determined by the following algorithm.

MAXIMUM WEIGHT BASIS ALGORITHM.

Start: Given $\tilde{d} \in \Re^{(n+l)}_{++}$, and $\tilde{A} \in \Re^{(m+l)\times(n+l)}$ such that $\text{rank}(\tilde{A}) = m + l$,

1. Order the elements of $\tilde{d}$ so that $\tilde{d}_1 \geq \cdots \geq \tilde{d}_{n+l}$; order the columns of $\tilde{A}$ accordingly.
2. Let $\mathcal{B} = \emptyset$, $j = 1$.
3. **While** $|\mathcal{B}| < m + l$ **do**
   (a) If $\tilde{A}_j$ is linearly independent of $\{\tilde{A}_i : i \in \mathcal{B}\}$, set $\mathcal{B} \leftarrow \mathcal{B} \cup \{j\}$.
   (b) $j \leftarrow j + 1$.
4. Return to the original ordering of $\tilde{A}$ and $\tilde{d}$; determine the set $\mathcal{B}$ according to this ordering and set $\mathcal{N} := \{1, \ldots, n + l\}\backslash\mathcal{B}$.
5. Set $B := \tilde{A}_{\mathcal{B}}$ and $\tilde{D}_{\mathcal{B}} := \text{Diag}(\tilde{d}_{\mathcal{B}})$.
6. Let $\tilde{T} = \tilde{T}(\tilde{A}, \tilde{d}) := \tilde{D}_{\mathcal{B}}^{-1}B^{-1}$.

end

Note that the above algorithm can be applied to the matrix $\tilde{A}$ defined in (21) since this matrix has full row rank due to Assumption 1. The MWB preconditioner was originally proposed by Vaidya [25] and Resende and Veiga [22] in the context of the minimum cost network flow problem. In this case, $\tilde{A} = A$ is the node-arc incidence matrix of a connected digraph (with one row deleted to ensure that $\tilde{A}$ has full row rank), the entries of $\tilde{d}$ are weights on the edges of the graph, and the set $\mathcal{B}$ generated by the above algorithm defines a maximum spanning tree on the digraph. Oliveira and Sorensen [19] later proposed the use of this preconditioner for general matrices $\tilde{A}$. Boman et al. [5] have proposed variants of the MWB preconditioner for diagonally dominant matrices, using the fact that they can be represented as $D_1 + AD_2A^T$, where D_1 and D_2 are nonnegative diagonal and positive diagonal matrices, respectively, and A is a node-arc incidence matrix.

For the purpose of stating the next result, we now introduce some notation. Let us define

$$\varphi_{\tilde{A}} := \max\{\|B^{-1}\tilde{A}\|_F : B \text{ is a basis of } \tilde{A}\}. \tag{41}$$

The constant $\varphi_{\tilde{A}}$ is related to the well-known condition number $\bar{\chi}_{\tilde{A}}$ (see [26]), defined as

$$\bar{\chi}_{\tilde{A}} := \sup\{\|\tilde{A}^T(\tilde{A}\tilde{E}\tilde{A}^T)^{-1}\tilde{A}\tilde{E}\| : \tilde{E} \in \text{Diag}(\Re^{(n+l)}_{++})\}.$$

Specifically, $\varphi_{\tilde{A}} \leq (n + l)^{1/2}\bar{\chi}_{\tilde{A}}$, in view of the facts that $\|C\|_F \leq (n + l)^{1/2}\|C\|$ for any matrix $C \in \Re^{(m+l)\times(n+l)}$ and, as shown in [23] and [26],

$$\bar{\chi}_{\tilde{A}} = \max\{\|B^{-1}\tilde{A}\| : B \text{ is a basis of } \tilde{A}\}.$$

The following result, which establishes the theoretical properties of the MWB preconditioner, follows as a consequence of Lemmas 2.1 and 2.2 of [17].

PROPOSITION 3.1. *Let $\tilde{T} = \tilde{T}(\tilde{A}, \tilde{d})$ be the preconditioner determined according to the maximum weight basis algorithm, and define $W := \tilde{T}\tilde{A}\tilde{D}^2\tilde{A}^T\tilde{T}^T$. Then, $\|\tilde{T}\tilde{A}\tilde{D}\| \leq \varphi_{\tilde{A}}$ and $\kappa(W) \leq \varphi_{\tilde{A}}^2$.*

Note that the bound $\varphi_{\tilde{A}}^2$ on $\kappa(W)$ is independent of the diagonal matrix $\tilde{D}$ and depends only on $\tilde{A}$. This will allow us to obtain a uniform bound on the number of iterations needed by any member of the family of iterative linear solvers described below to obtain a suitable approximate solution of (22). This topic is the subject of the remainder of this subsection.

Instead of dealing directly with (22), we consider the application of an iterative linear solver to the preconditioned ANE:

$$Wu = v, \tag{42}$$

where

$$W := \tilde{T}\tilde{A}\tilde{D}^2\tilde{A}^T\tilde{T}^T, \qquad v := \tilde{T}h. \tag{43}$$

For the purpose of our analysis below, the only thing we will assume regarding the iterative linear solver when applied to (42) is that it generates a sequence of iterates $\{u^j\}$ such that

$$\|v - Wu^j\| \;\leq\; c(\kappa)\left[1 - \frac{1}{\psi(\kappa)}\right]^j \|v - Wu^0\| \;\; \forall\; j = 0, 1, 2, \ldots, \tag{44}$$

where c and ψ are positive, nondecreasing functions of $\kappa \equiv \kappa(W)$.

Examples of solvers which satisfy (44) include the steepest descent (SD) and CG methods, with the values for $c(\kappa)$ and $\psi(\kappa)$ given in Table 3.1.

TABLE 3.1

Solver	$c(\kappa)$	$\psi(\kappa)$
SD	$\sqrt{\kappa}$	$(\kappa+1)/2$
CG	$2\sqrt{\kappa}$	$(\sqrt{\kappa}+1)/2$

The justification for Table 3.1 follows from section 7.6 and Exercise 10 of section 8.8 of [14].

The following result gives an upper bound on the number of iterations that any iterative linear solver satisfying (44) needs to perform to obtain a *ξ-approximate solution of* (42), i.e., an iterate u^j such that $\|v - Wu^j\| \leq \xi\sqrt{\mu}$ for some constant $\xi > 0$.

PROPOSITION 3.2. *Let u^0 be an arbitrary starting point. Then, a generic iterative linear solver with a convergence rate given by* (44) *generates an iterate u^j satisfying $\|v - Wu^j\| \leq \xi\sqrt{\mu}$ in*

$$\mathcal{O}\left(\psi(\kappa)\log\left(\frac{c(\kappa)\|v - Wu^0\|}{\xi\sqrt{\mu}}\right)\right) \tag{45}$$

iterations, where $\kappa \equiv \kappa(W)$.

Proof. Let j be any iteration such that $\|v - Wu^j\| > \xi\sqrt{\mu}$. We use relation (44) and the fact that $1 + \omega \le e^\omega$ for all $\omega \in \Re$ to observe that

$$\xi\sqrt{\mu} < \|v - Wu^j\| \;\le\; c(\kappa)\left[1 - \frac{1}{\psi(\kappa)}\right]^j \|v - Wu^0\| \;\le c(\kappa)\exp\left\{\frac{-j}{\psi(\kappa)}\right\}\|v - Wu^0\|.$$

Rearranging the first and last terms of the inequality, it follows that

$$j \;<\; \psi(\kappa)\log\left(\frac{c(\kappa)\|v - Wu^0\|}{\xi\sqrt{\mu}}\right),$$

and the result is proven. ∎

From Proposition 3.2, it is clear that different choices of u^0 and ξ lead to different bounds on the number of iterations performed by the iterative linear solver. In subsection 3.2, we will describe a suitable way of selecting u^0 and ξ so that (i) the bound (45) is independent of the iterate w and (ii) the approximate solution $\tilde{T}^T u^j$ of the ANE, together with a suitable pair (p, q), yields a (τ_p, τ_q)-search direction Δw through the recipe described in subsection 2.2.

3.2. Computation of the inexact search direction Δw. In this subsection, we use the results of subsections 2.2 and 3.1 to build a (τ_p, τ_q)-search direction Δw, where τ_p and τ_q are given by (38) and (39), respectively. In addition, we describe a way of choosing u^0 and ξ which ensures that the number of iterations of an iterative linear solver satisfying (44) applied to the preconditioned ANE is uniformly bounded by a constant depending on n and $\varphi_{\tilde{A}}$.

Suppose that we solve (42) inexactly according to subsection 3.1. Then our final solution u^j satisfies $Wu^j - v = \tilde{f}$ for some vector $\tilde{f}$. Letting

$$\begin{pmatrix} \Delta y \\ \Delta z \end{pmatrix} = \tilde{T}^T u^j, \tag{46}$$

we easily see from (43) that (26) is satisfied with $f := \tilde{T}^{-1}\tilde{f}$. We can then apply the recipe of subsection 2.2 to this approximate solution, using the pair (p, q) which we will now describe.

First, note that (30) with f as defined above is equivalent to the system

$$\tilde{f} \;=\; \tilde{T}\tilde{A}\begin{pmatrix} S^{-1}p \\ E^{-1}q \end{pmatrix} \;=\; \tilde{T}\tilde{A}\tilde{D}\begin{pmatrix} (XS)^{-1/2} & 0 \\ 0 & I \end{pmatrix}\begin{pmatrix} p \\ q \end{pmatrix}. \tag{47}$$

Now, let $\mathcal{B} = (\mathcal{B}_1, \ldots, \mathcal{B}_{m+l})$ be the ordered set of basic indices computed by the MWB algorithm applied to the pair $(\tilde{A}, \tilde{d})$ and note that, by step 6 of this algorithm, the $\mathcal{B}_i$th column of $\tilde{T}\tilde{A}\tilde{D}$ is the ith unit vector for every $i = 1, \ldots, m + l$. Then, the vector $t \in \Re^{n+l}$ defined as $t_{\mathcal{B}_i} = \tilde{f}_i$ for $i = 1, \ldots, m + l$ and $t_j = 0$ for every $j \notin \{\mathcal{B}_1, \ldots, \mathcal{B}_{m+l}\}$ clearly satisfies

$$\tilde{f} \;=\; \tilde{T}\tilde{A}\tilde{D}\; t. \tag{48}$$

We then obtain a pair $(p, q) \in \Re^n \times \Re^l$ satisfying (30) by defining

$$\begin{pmatrix} p \\ q \end{pmatrix} \;:=\; \begin{pmatrix} (XS)^{1/2} & 0 \\ 0 & I \end{pmatrix} t. \tag{49}$$

It is clear from (49) and the fact that $\|t\| = \|\tilde{f}\|$ that

$$\|p\| \;\le\; \|XS\|^{1/2}\|\tilde{f}\|, \qquad \|q\| \;\le\; \|\tilde{f}\|. \tag{50}$$

As an immediate consequence of this relation, we obtain the following result.

LEMMA 3.3. *Suppose that $w \in \Re^{2n}_{++} \times \Re^{m+l}$ and positive scalars τ_p and τ_q are given. Assume that w^j is a ξ-approximate solution of* (42) *or, equivalently, $\tilde{f} \le \xi\sqrt{\mu}$, where $\xi := \min\{n^{-1/2}\tau_p, \tau_q\}$. Let Δw be determined according to the recipe given in subsection* 2.2 *using the approximate solution* (46) *and the pair (p, q) given by* (49). *Then Δw is a (τ_p, τ_q)-search direction.*

Proof. It is clear from the previous discussion that Δw and the pair (p, q) satisfy (32)–(35). Next, relation (50) and the facts that $\xi \le n^{-1/2}\tau_p$ and $\|XS\|^{1/2} \le \sqrt{n\mu}$ imply that

$$\|p\|_\infty \;\le\; \|p\| \;\le\; \|XS\|^{1/2}\|\tilde{f}\| \;\le\; \sqrt{n\mu}\,\xi\sqrt{\mu} \;\le\; \tau_p\mu.$$

Similarly, (50) and the fact that $\xi \le \tau_q$ imply that $\|q\| \le \tau_q\sqrt{\mu}$. Thus, Δw is a (τ_p, τ_q)-search direction as desired. ∎

Lemma (3.3) implies that to construct a (τ_p, τ_q)-search direction Δw as in step 2(d) of the inexact PDIPF algorithm, it suffices to find a ξ-approximate solution to (42), where

$$\xi \;:=\; \min\left\{ \frac{\gamma\sigma}{4\sqrt{n}}\,,\; \left[\sqrt{1 + \left(1 - \frac{\gamma}{2}\right)\sigma} - 1\right]\theta \right\}. \tag{51}$$

We next describe a suitable way of selecting u^0 so that the number of iterations required by an iterative linear solver satisfying (44) to find a ξ-approximate solution of (42) can be uniformly bounded by a universal constant depending only on the quantities n and $\varphi_{\tilde{A}}$. First, compute a point $\bar{w} = (\bar{x}, \bar{s}, \bar{y}, \bar{z})$ such that

$$\tilde{A}\begin{pmatrix} \bar{x} \\ E^{-2}\bar{z} \end{pmatrix} = \begin{pmatrix} b \\ 0 \end{pmatrix}, \qquad A^T\bar{y} + \bar{s} + V\bar{z} = c. \tag{52}$$

Note that vectors $\bar{x}$ and $\bar{z}$ satisfying the first equation in (52) can be easily computed once a basis of $\tilde{A}$ is available (e.g., the one computed by the maximum weight basis algorithm in the first outer iteration of the inexact PDIPF algorithm). Once $\bar{y}$ is arbitrarily chosen, a vector $\bar{s}$ satisfying the second equation of (52) is immediately available. We then define

$$u^0 \;=\; -\eta\,\tilde{T}^{-T}\begin{pmatrix} y^0 - \bar{y} \\ z^0 - \bar{z} \end{pmatrix}. \tag{53}$$

The following lemma gives a bound on the size of the initial residual $\|Wu^0 - v\|$. Its proof will be given in subsection 4.1.

LEMMA 3.4. *Assume that $\tilde{T} = \tilde{T}(\tilde{A}, \tilde{d})$ is given and that $w^0 \in \Re^{2n}_{++} \times \Re^{m+l}$ and $\bar{w}$ are such that $(x^0, s^0) \ge |(\bar{x}, \bar{s})|$ and $(x^0, s^0) \ge (x^*, s^*)$ for some $w^* \in \mathcal{S}$. Further, assume that $w \in \mathcal{N}_{w^0}(\gamma, \theta)$ for some $\gamma \in [0, 1]$ and $\theta > 0$, and that W, v, and u^0 are given by* (43) *and* (53), *respectively. Then, the initial residual in* (44) *satisfies $\|v - Wu^0\| \le \Psi\sqrt{\mu}$, where*

$$\Psi \;:=\; \left[\frac{7n + \theta^2/2}{\sqrt{1-\gamma}} + \theta\right]\varphi_{\tilde{A}}. \tag{54}$$

As an immediate consequence of Proposition 3.2 and Lemmas 3.3 and 3.4, we can bound the number of inner iterations required by an iterative linear solver satisfying (44) to yield a (τ_p, τ_q)-search direction Δw.

THEOREM 3.5. *Assume that ξ is defined in* (51)*, where σ, γ, θ are such that*

$$\max\{\sigma^{-1}, \gamma^{-1}, (1-\gamma)^{-1}, \theta, \theta^{-1}\}$$

is bounded by a polynomial of n. Assume also that $w^0 \in \Re^{2n}_{++} \times \Re^{m+l}$ and $\bar{w}$ are such that $(x^0, s^0) \geq |(\bar{x}, \bar{s})|$ and $(x^0, s^0) \geq (x^*, s^*)$ for some $w^* \in \mathcal{S}$. Then, a generic iterative linear solver with a convergence rate given by (44) *generates a ξ-approximate solution, which leads to a (τ_p, τ_q)-search direction Δw in*

$$\mathcal{O}\left(\psi(\varphi^2_{\tilde{A}}) \log\left(c(\varphi^2_{\tilde{A}}) n \varphi_{\tilde{A}}\right)\right) \tag{55}$$

iterations. As a consequence, the SD and CG methods generate this approximate solution u^j in $\mathcal{O}(\varphi^2_{\tilde{A}} \log(n\varphi_{\tilde{A}}))$ and $\mathcal{O}(\varphi_{\tilde{A}} \log(n\varphi_{\tilde{A}}))$ iterations, respectively.

Proof. The proof of the first part of Theorem 3.5 immediately follows from Propositions 3.1 and 3.2 and Lemmas 3.3 and 3.4. The proof of the second part of Theorem 3.5 follows immediately from Table 3.1 and Proposition 3.1. ∎

Using the results of sections 2 and 3, we see that the number of "inner" iterations of an iterative linear solver satisfying (44) is uniformly bounded by a constant depending on n and $\varphi_{\tilde{A}}$, while the number of "outer" iterations in the inexact PDIPF algorithm is polynomially bounded by a constant depending on n and $\log \epsilon^{-1}$.

4. Technical results. This section is devoted to the proofs of Lemma 3.4 and Theorem 2.2. Subsection 4.1 presents the proof of Lemma 3.4, and subsection 4.2 presents the proof of Theorem 2.2.

4.1. Proof of Lemma 3.4. In this subsection, we will provide the proof of Lemma 3.4. We begin by establishing three technical lemmas.

LEMMA 4.1. *Suppose that $w^0 \in \Re^{2n}_{++} \times \Re^{m+l}$, $w \in \mathcal{N}_{w^0}(\eta, \gamma, \theta)$ for some $\eta \in [0,1]$, $\gamma \in [0,1]$, and $\theta > 0$, and $w^* \in \mathcal{S}$. Then*

$$(x - \eta x^0 - (1-\eta)x^*)^T (s - \eta s^0 - (1-\eta)s^*) \geq -\frac{\theta^2}{4}\mu. \tag{56}$$

Proof. Let us define $\tilde{w} := w - \eta w^0 - (1-\eta)w^*$. Using the definitions of $\mathcal{N}_{w^0}(\eta, \gamma, \theta)$, r, and $\mathcal{S}$, we have that

$$\begin{aligned} A\tilde{x} &= 0, \\ A^T\tilde{y} + \tilde{s} + V\tilde{z} &= 0, \\ V^T\tilde{x} + E^{-2}\tilde{z} &= E^{-1}(r_V - \eta r_V^0). \end{aligned}$$

Multiplying the second relation by $\tilde{x}^T$ on the left and using the first and third relations along with the fact that $w \in \mathcal{N}_{w^0}(\eta, \gamma, \theta)$, we see that

$$\begin{aligned} \tilde{x}^T\tilde{s} = -\tilde{x}^T V\tilde{z} &= [E^{-2}\tilde{z} - E^{-1}(r_V - \eta r_V^0)]^T\tilde{z} = \|E^{-1}\tilde{z}\|^2 - (E^{-1}\tilde{z})^T(r_V - \eta r_V^0) \\ &\geq \|E^{-1}\tilde{z}\|^2 - \|E^{-1}\tilde{z}\|\|r_V - \eta r_V^0\| = \left(\|E^{-1}\tilde{z}\| - \frac{\|r_V - \eta r_V^0\|}{2}\right)^2 - \frac{\|r_V - \eta r_V^0\|^2}{4} \\ &\geq -\frac{\|r_V - \eta r_V^0\|^2}{4} \geq -\frac{\theta^2}{4}\mu. \end{aligned}$$

∎

LEMMA 4.2. *Suppose that* $w^0 \in \Re^{2n}_{++} \times \Re^{m+l}$ *such that* $(x^0, s^0) \geq (x^*, s^*)$ *for some* $w^* \in \mathcal{S}$. *Then, for any* $w \in \mathcal{N}_{w^0}(\eta, \gamma, \theta)$ *with* $\eta \in [0,1]$, $\gamma \in [0,1]$, *and* $\theta > 0$, *we have*

$$\eta(x^T s^0 + s^T x^0) \leq \left(3n + \frac{\theta^2}{4}\right)\mu. \tag{57}$$

Proof. Using the fact $w \in \mathcal{N}_{w^0}(\eta, \gamma, \theta)$ and (56), we obtain

$$\begin{aligned} x^T s - \eta(x^T s^0 + s^T x^0) + \eta^2 {x^0}^T s^0 - (1-\eta)(x^T s^* + s^T x^*) \\ + \eta(1-\eta)({x^*}^T s^0 + {s^*}^T x^0) + (1-\eta)^2 {x^*}^T s^* \ \geq \ -\frac{\theta^2}{4}\mu. \end{aligned}$$

Rearranging the terms in this equation and using the facts that $\eta \leq x^T s / {x^0}^T s^0$, ${x^*}^T s^* = 0$, $(x,s) \geq 0$, $(x^*, s^*) \geq 0$, $(x^0, s^0) > 0$, $\eta \in [0,1]$, $x^* \leq x^0$, and $s^* \leq s^0$, we conclude that

$$\begin{aligned} \eta(x^T s^0 + s^T x^0) &\leq \eta^2 {x^0}^T s^0 + x^T s + \eta(1-\eta)({x^*}^T s^0 + {s^*}^T x^0) + \frac{\theta^2}{4}\mu \\ &\leq \eta^2 {x^0}^T s^0 + x^T s + 2\eta(1-\eta){x^0}^T s^0 + \frac{\theta^2}{4}\mu \\ &\leq 2\eta {x^0}^T s^0 + x^T s + \frac{\theta^2}{4}\mu \\ &\leq 3x^T s + \frac{\theta^2}{4}\mu \ = \ \left(3n + \frac{\theta^2}{4}\right)\mu. \qquad \square \end{aligned}$$

LEMMA 4.3. *Suppose* $w^0 \in \Re^{2n}_{++} \times \Re^{m+l}$, $w \in \mathcal{N}_{w^0}(\eta, \gamma, \theta)$ *for some* $\eta \in [0,1]$, $\gamma \in [0,1]$, *and* $\theta > 0$, *and* $\bar{w}$ *satisfies* (52). *Let* W, v, *and* u^0 *be given by* (43) *and* (53), *respectively. Then,*

$$Wu^0 - v \ = \ \tilde{T}\tilde{A}\begin{pmatrix} -x + \sigma\mu S^{-1}e + \eta(x^0 - \bar{x}) + \eta D^2(s^0 - \bar{s}) \\ E^{-1}(r_V - \eta r_V^0) \end{pmatrix}. \tag{58}$$

Proof. Using the fact that $w \in \mathcal{N}_{w^0}(\eta, \gamma, \theta)$ along with (21), (36), and (52), we easily obtain that

$$\begin{aligned} \begin{pmatrix} r_p \\ E^{-1}r_V \end{pmatrix} &= \begin{pmatrix} \eta r_p^0 \\ \eta E^{-1} r_V^0 + E^{-1}(r_V - \eta r_V^0) \end{pmatrix} \\ &= \eta\tilde{A}\begin{pmatrix} x^0 - \bar{x} \\ E^{-2}(z^0 - \bar{z}) \end{pmatrix} + \tilde{A}\begin{pmatrix} 0 \\ E^{-1}(r_V - \eta r_V^0) \end{pmatrix}, \end{aligned} \tag{59}$$

$$s^0 - \bar{s} = -A^T(y^0 - \bar{y}) - V(z^0 - \bar{z}) + r_d^0. \tag{60}$$

Using relations (20), (21), (43), (36), (53), (59), and (60), we obtain

$$\begin{aligned}
Wu^0 - v &= \tilde{T}\tilde{A}\tilde{D}^2\tilde{A}^T\tilde{T}^T u^0 - \tilde{T}\tilde{A}\begin{pmatrix} x - \sigma\mu S^{-1}e - D^2 r_d \\ 0 \end{pmatrix} + \tilde{T}\begin{pmatrix} r_p \\ E^{-1}r_V \end{pmatrix} \\
&= -\eta\tilde{T}\tilde{A}\tilde{D}^2\tilde{A}^T\begin{pmatrix} y^0 - \bar{y} \\ z^0 - \bar{z} \end{pmatrix} - \tilde{T}\tilde{A}\begin{pmatrix} x - \sigma\mu S^{-1}e - \eta D^2 r_d^0 \\ 0 \end{pmatrix} + \tilde{T}\begin{pmatrix} r_p \\ E^{-1}r_V \end{pmatrix} \\
&= -\eta\tilde{T}\tilde{A}\begin{pmatrix} D^2A^T(y^0 - \bar{y}) + D^2V(z^0 - \bar{z}) - D^2 r_d^0 \\ E^{-2}(z^0 - \bar{z}) \end{pmatrix} \\
&\quad - \tilde{T}\tilde{A}\begin{pmatrix} x - \sigma\mu S^{-1}e \\ 0 \end{pmatrix} + \tilde{T}\begin{pmatrix} r_p \\ E^{-1}r_V \end{pmatrix}, \\
&= -\eta\tilde{T}\tilde{A}\begin{pmatrix} -D^2(s^0 - \bar{s}) \\ E^{-2}(z^0 - \bar{z}) \end{pmatrix} - \tilde{T}\tilde{A}\begin{pmatrix} x - \sigma\mu S^{-1}e \\ 0 \end{pmatrix} \\
&\quad + \eta\tilde{T}\tilde{A}\begin{pmatrix} x^0 - \bar{x} \\ E^{-2}(z^0 - \bar{z}) \end{pmatrix} + \tilde{T}\tilde{A}\begin{pmatrix} 0 \\ E^{-1}(r_V - \eta r_V^0) \end{pmatrix},
\end{aligned}$$

which yields (58), as desired. ∎

We now turn to the proof of Lemma 3.4.

Proof. Since $w \in \mathcal{N}_{w^0}(\gamma, \theta)$, we have that $x_i s_i \geq (1-\gamma)\mu$ for all i, which implies

$$\|(XS)^{-1/2}\| \;\leq\; \frac{1}{\sqrt{(1-\gamma)\mu}}. \tag{61}$$

Note that $\|Xs - \sigma\mu e\|$, when viewed as a function of $\sigma \in [0,1]$, is convex. Hence, it is maximized at one of its endpoints, which, together with the facts $\|Xs - \mu e\| < \|Xs\|$ and $\sigma \in [\underline{\sigma}, \overline{\sigma}] \subset [0,1]$, immediately implies that

$$\|Xs - \sigma\mu e\| \;\leq\; \|Xs\| \;\leq\; \|Xs\|_1 \;=\; x^T s \;=\; n\mu. \tag{62}$$

Using the fact that $(x^0, s^0) \geq |(\bar{x}, \bar{s})|$ together with Lemma 4.2, we obtain that

$$\begin{aligned}
\eta\|S(x^0 - \bar{x}) + X(s^0 - \bar{s})\| &\leq \eta\big\{\|S(x^0 - \bar{x})\| + \|X(s^0 - \bar{s})\|\big\} \;\leq\; 2\eta\big\{\|Sx^0\| + \|Xs^0\|\big\} \\
&\leq 2\eta(x^T s^0 + x^T s^0) \;\leq\; \left(6n + \frac{\theta^2}{2}\right)\mu. \qquad (63)
\end{aligned}$$

Since $w \in \mathcal{N}_{w^0}(\gamma, \theta)$, there exists $\eta \in [0,1]$ such that $w \in \mathcal{N}_{w^0}(\eta, \gamma, \theta)$. It is clear that the requirements of Lemma 4.3 are met, so (58) holds. By (19), (20), and (58), we see that

$$\begin{aligned}
\|v - Wu^0\| &= \left\|\tilde{T}\tilde{A}\tilde{D}\begin{pmatrix} (XS)^{-1/2}\{Xs - \sigma\mu e - \eta[S(x^0 - \bar{x}) + X(s^0 - \bar{s})]\} \\ r_V - \eta r_V^0 \end{pmatrix}\right\| \\
&\leq \|\tilde{T}\tilde{A}\tilde{D}\|\Big\{\|(XS)^{-1/2}\|\Big[\|Xs - \sigma\mu e\| + \eta\|X(s^0 - \bar{s}) + S(x^0 - \bar{x})\|\Big] \\
&\quad + \|r_V - \eta r_V^0\|\Big\}, \\
&\leq \varphi_{\tilde{A}}\left\{\frac{1}{\sqrt{(1-\gamma)\mu}}\left[n\mu + \left(6n + \frac{\theta^2}{2}\right)\mu\right] + \theta\sqrt{\mu}\right\} \;=\; \Psi\sqrt{\mu},
\end{aligned}$$

where the last inequality follows from Proposition 3.1, relations (61), (62), (63), and the assumption that $w \in \mathcal{N}_{w^0}(\gamma, \theta)$. ∎

4.2. "Outer" iteration results—Proof of Theorem 2.2. In this subsection, we will present the proof of Theorem 2.2. Specifically, we will show that the inexact PDIPF algorithm obtains an ϵ-approximate solution to (5)–(8) in $\mathcal{O}(n^2 \log(1/\epsilon))$ outer iterations.

Throughout this section, we use the following notation:

$$w(\alpha) := w + \alpha \Delta w, \quad \mu(\alpha) := \mu(w(\alpha)), \quad r(\alpha) := r(w(\alpha)).$$

LEMMA 4.4. *Assume that Δw satisfies* (32)–(35) *for some $\sigma \in \Re$, $w \in \Re^{2n+m+l}$, and $(p, q) \in \Re^n \times \Re^l$. Then, for every $\alpha \in \Re$, we have*

(a) $X(\alpha)s(\alpha) = (1-\alpha)Xs + \alpha\sigma\mu e - \alpha p + \alpha^2 \Delta X \Delta s$;
(b) $\mu(\alpha) = [1 - \alpha(1-\sigma)]\mu - \alpha p^T e/n + \alpha^2 \Delta x^T \Delta s/n$;
(c) $(r_p(\alpha), r_d(\alpha)) = (1-\alpha)(r_p, r_d)$;
(d) $r_V(\alpha) = (1-\alpha) r_V + \alpha q$.

Proof. Using (34), we obtain

$$\begin{aligned} X(\alpha)s(\alpha) &= (X + \alpha\Delta X)(s + \alpha \Delta s) \\ &= Xs + \alpha(X\Delta s + S\Delta x) + \alpha^2 \Delta X \Delta s \\ &= Xs + \alpha(-Xs + \sigma\mu e - p) + \alpha^2 \Delta X \Delta s \\ &= (1-\alpha)Xs + \alpha\sigma\mu e - \alpha p + \alpha^2 \Delta X \Delta s, \end{aligned}$$

thereby showing that (a) holds. Left multiplying the above equality by e^T and dividing the resulting expression by n, we easily conclude that (b) holds. Statement (c) can be easily verified by means of (32) and (33), while statement (d) follows from (35). ∎

LEMMA 4.5. *Assume that Δw satisfies* (32)–(35) *for some $\sigma \in \Re$, $w \in \Re^{2n}_{++} \times \Re^{m+l}$, and $(p, q) \in \Re^n \times \Re^l$ such that $\|p\|_\infty \le \gamma\sigma\mu/4$. Then, for every scalar α satisfying*

$$0 \le \alpha \le \min\left\{1, \frac{\sigma\mu}{4\,\|\Delta X \Delta s\|_\infty}\right\}, \tag{64}$$

we have

$$\frac{\mu(\alpha)}{\mu} \;\ge\; 1 - \alpha. \tag{65}$$

Proof. Since $\|p\|_\infty \le \gamma\sigma\mu/4$, we easily see that

$$|p^T e/n| \;\le\; \|p\|_\infty \;\le\; \sigma\mu/4. \tag{66}$$

Using this result and Lemma 4.4(b), we conclude for every α satisfying (64) that

$$\begin{aligned} \mu(\alpha) &= [1 - \alpha(1-\sigma)]\mu - \alpha p^T e/n + \alpha^2 \Delta x^T \Delta s/n \\ &\ge [1 - \alpha(1-\sigma)]\mu - \frac{1}{4}\alpha\sigma\mu + \alpha^2 \Delta x^T \Delta s/n \\ &\ge (1-\alpha)\mu + \frac{1}{4}\alpha\sigma\mu - \alpha^2 \|\Delta X \Delta s\|_\infty \\ &\ge (1-\alpha)\mu. \qquad \blacksquare \end{aligned}$$

LEMMA 4.6. *Assume that Δw is a (τ_p, τ_q)-search direction, where τ_p and τ_q are given by* (38) *and* (39)*, respectively. Assume also that $\sigma > 0$ and that $w \in \mathcal{N}_{w^0}(\gamma, \theta)$*

with $w^0 \in \Re^{2n}_{++} \times \Re^{m+l}$, $\gamma \in [0,1]$, *and* $\theta \geq 0$. *Then,* $w(\alpha) \in \mathcal{N}_{w^0}(\gamma,\theta)$ *for every scalar* α *satisfying*

$$0 \leq \alpha \leq \min\left\{1\,,\ \frac{\gamma\sigma\mu}{4\,\|\Delta X\Delta s\|_\infty}\right\}. \tag{67}$$

Proof. Since $w \in \mathcal{N}_{w^0}(\gamma,\theta)$, there exists $\eta \in [0,1]$ such that $w \in \mathcal{N}_{w^0}(\eta,\gamma,\theta)$. We will show that $w(\alpha) \in \mathcal{N}_{w^0}((1-\alpha)\eta,\gamma,\theta) \subseteq \mathcal{N}_{w^0}(\gamma,\theta)$ for every α satisfying (67).

First, we note that $(r_p(\alpha), r_d(\alpha)) = (1-\alpha)\eta(r_p^0, r_d^0)$ by Lemma 4.4(c) and the definition of $\mathcal{N}_{w^0}(\eta,\gamma,\theta)$. Next, it follows from Lemma 4.5 that (65) holds for every α satisfying (64), and hence (67) due to $\gamma \in [0,1]$. Thus, for every α satisfying (67), we have

$$(1-\alpha)\eta \leq \ \frac{\mu(\alpha)}{\mu}\eta \ \leq \ \frac{\mu(\alpha)}{\mu}\,\frac{\mu}{\mu_0} \ = \ \frac{\mu(\alpha)}{\mu_0}. \tag{68}$$

Now, it is easy to see that for every $u \in \Re^n$ and $\tau \in [0,n]$, there holds $\|u - \tau(u^Te/n)e\|_\infty \leq (1+\tau)\|u\|_\infty$. Using this inequality twice, the fact that $w \in \mathcal{N}_{w^0}(\eta,\gamma,\theta)$, relation (38), and statements (a) and (b) of Lemma 4.4, we conclude for every α satisfying (67) that

$$\begin{aligned}
&X(\alpha)s(\alpha) - (1-\gamma)\,\mu(\alpha)e \\
&= (1-\alpha)\,[Xs - (1-\gamma)\mu e] + \alpha\gamma\sigma\mu e - \alpha\left[p - (1-\gamma)\left(\frac{p^Te}{n}\right)e\right] \\
&\quad + \alpha^2\left[\Delta X\Delta s - (1-\gamma)\left(\frac{\Delta x^T\Delta s}{n}\right)e\right] \\
&\geq \alpha\left[\gamma\sigma\mu - \left\|p - (1-\gamma)\frac{p^Te}{n}\,e\right\|_\infty - \alpha\left\|\Delta X\Delta s - (1-\gamma)\frac{\Delta x^T\Delta s}{n}\,e\right\|_\infty\right]e \\
&\geq \alpha\,(\gamma\sigma\mu - 2\|p\|_\infty - 2\alpha\|\Delta X\Delta s\|_\infty)\,e \ \geq \ \alpha\left(\gamma\sigma\mu - \frac{1}{2}\gamma\sigma\mu - \frac{1}{2}\gamma\sigma\mu\right)e \ = \ 0.
\end{aligned}$$

Next, by Lemma 4.4(d), we have that

$$r_V(\alpha) = (1-\alpha)r_V + \alpha q = \ (1-\alpha)\eta r_V^0 + \hat{a},$$

where $\hat{a} = (1-\alpha)(r_V - \eta r_V^0) + \alpha q$. To complete the proof, it suffices to show that $\|\hat{a}\| \leq \theta\sqrt{\mu(\alpha)}$ for every α satisfying (67). By using equation (39) and Lemma 4.4(b) along with the facts that $\|r_V - \eta r_V^0\| \leq \theta\sqrt{\mu}$ and $\alpha \in [0,1]$, we have

$$\begin{aligned}
\|\hat{a}\|^2 - \theta^2\mu(\alpha) &= (1-\alpha)^2\|r_V - \eta r_V^0\|^2 + 2\alpha(1-\alpha)[r_V - \eta r_V^0]^Tq + \alpha^2\|q\|^2 - \theta^2\mu(\alpha) \\
&\leq (1-\alpha)^2\theta^2\mu + 2\alpha(1-\alpha)\theta\sqrt{\mu}\|q\| + \alpha^2\|q\|^2 \\
&\quad - \theta^2\left\{[1-\alpha(1-\sigma)]\mu - \alpha\frac{p^Te}{n} + \alpha^2\frac{\Delta x^T\Delta s}{n}\right\} \\
&\leq \alpha^2\|q\|^2 + 2\alpha\theta\sqrt{\mu}\|q\| - \alpha\theta^2\sigma\mu + \alpha\theta^2\frac{p^Te}{n} - \alpha^2\theta^2\frac{\Delta x^T\Delta s}{n} \\
&\leq \alpha\left[\|q\|^2 + 2\theta\sqrt{\mu}\|q\| - \left(1 - \frac{\gamma}{4}\right)\theta^2\sigma\mu + \theta^2\alpha\|\Delta X\Delta s\|_\infty\right] \\
&\leq \alpha\left[\|q\|^2 + 2\theta\sqrt{\mu}\|q\| - \left(1 - \frac{\gamma}{2}\right)\theta^2\sigma\mu\right] \ \leq \ 0,
\end{aligned}$$

where the last inequality follows from the quadratic formula and the fact that $\|q\| \leq \tau_q$, where τ_q is given by (39). ∎

Next, we derive a lower bound on the step size of the inexact PDIPF algorithm.

LEMMA 4.7. *In every iteration of the inexact PDIPF algorithm, the step length* $\bar{\alpha}$ *satisfies*

$$\bar{\alpha} \;\geq\; \min\left\{1, \frac{\min\{\gamma\sigma, 1-\frac{5}{4}\sigma\}\mu}{4\,\|\Delta X \Delta s\|_\infty}\right\} \tag{69}$$

and

$$\mu(\bar{\alpha}) \;\leq\; \left[1-\left(1-\frac{5}{4}\,\sigma\right)\frac{\bar{\alpha}}{2}\right]\mu. \tag{70}$$

Proof. We know that Δw is a (τ_p, τ_q)-search direction in every iteration of the inexact PDIPF algorithm, where τ_p and τ_q are given by (38) and (39). Hence, by Lemma 4.6, the quantity $\tilde{\alpha}$ computed in step (g) of the inexact PDIPF algorithm satisfies

$$\tilde{\alpha} \;\geq\; \min\left\{1, \frac{\gamma\sigma\mu}{4\,\|\Delta X \Delta s\|_\infty}\right\}. \tag{71}$$

Moreover, by (66), it follows that the coefficient of α in the expression for $\mu(\alpha)$ in Lemma 4.4(b) satisfies

$$\begin{aligned} -(1-\sigma)\mu - \frac{p^T e}{n} \;\leq\; -(1-\sigma)\mu + \|p\|_\infty \;&\leq\; -(1-\sigma)\mu + \frac{1}{4}\,\gamma\sigma\mu \\ &=\; -\left(1-\frac{5}{4}\,\sigma\right)\mu \;<\; 0, \end{aligned} \tag{72}$$

since $\sigma \in (0, 4/5)$. Hence, if $\Delta x^T \Delta s \leq 0$, it is easy to see that $\bar{\alpha} = \tilde{\alpha}$ and hence that (69) holds in view of (71). Moreover, by Lemma 4.4(b) and (72), we have

$$\mu(\bar{\alpha}) \;\leq\; [1-\bar{\alpha}(1-\sigma)]\mu - \bar{\alpha}\frac{p^T e}{n} \;\leq\; \left[1-\left(1-\frac{5}{4}\,\sigma\right)\bar{\alpha}\right]\mu \;\leq\; \left[1-\left(1-\frac{5}{4}\,\sigma\right)\frac{\bar{\alpha}}{2}\right]\mu,$$

showing that (70) also holds. We now consider the case where $\Delta x^T \Delta s > 0$. In this case, we have $\bar{\alpha} = \min\{\alpha_{\min}\,,\, \tilde{\alpha}\}$, where $\alpha_{\min}$ is the unconstrained minimum of $\mu(\alpha)$. It is easy to see that

$$\alpha_{\min} = \frac{n\mu(1-\sigma) + p^T e}{2\Delta x^T \Delta s} \;\geq\; \frac{n[\mu(1-\sigma) - \frac{1}{4}\sigma\mu]}{2\Delta x^T \Delta s} \;\geq\; \frac{\mu(1-\frac{5}{4}\sigma)}{2\,\|\Delta X \Delta s\|_\infty}.$$

The last two observations together with (71) imply that (69) holds in this case too. Moreover, since the function $\mu(\alpha)$ is convex, it must lie below the function $\phi(\alpha)$ over the interval $[0, \alpha_{\min}]$, where $\phi(\alpha)$ is the affine function interpolating $\mu(\alpha)$ at $\alpha = 0$ and $\alpha = \alpha_{\min}$. Hence,

$$\mu(\bar{\alpha}) \;\leq\; \phi(\bar{\alpha}) \;=\; \left[1-(1-\sigma)\frac{\bar{\alpha}}{2}\right]\mu - \bar{\alpha}\frac{p^T e}{2n} \;\leq\; \left[1-\left(1-\frac{5}{4}\,\sigma\right)\frac{\bar{\alpha}}{2}\right]\mu, \tag{73}$$

where the second inequality follows from (72). We have thus shown that $\bar{\alpha}$ satisfies (70). ∎

Our next task will be to show that the step size $\bar{\alpha}$ remains bounded away from zero. In view of (69), it suffices to show that the quantity $\|\Delta X \Delta s\|_\infty$ can be suitably bounded. The next lemma addresses this issue.

LEMMA 4.8. *Let $w^0 \in \Re^{2n}_{++} \times \Re^{m+l}$ be such that $(x^0, s^0) \geq (x^*, s^*)$ for some $w^* \in \mathcal{S}$, and let $w \in \mathcal{N}_{w^0}(\gamma, \theta)$ for some $\gamma \geq 0$ and $\theta \geq 0$. Then, the inexact search direction Δw used in the inexact PDIPF algorithm satisfies*

$$\begin{aligned} \max(\|D^{-1}\Delta x\|, \|D\Delta s\|) \leq & \left(1 - 2\sigma + \frac{\sigma^2}{1-\gamma}\right)^{1/2} \sqrt{n\mu} \\ & + \frac{1}{\sqrt{1-\gamma}} \left(\frac{\gamma\sigma}{4}\sqrt{n} + 6n + \frac{\theta^2}{2}\right) \sqrt{\mu} + \theta\sqrt{\mu}. \end{aligned} \tag{74}$$

Proof. Since $w \in \mathcal{N}_{w^0}(\gamma, \theta)$, there exists $\eta \in [0, 1]$ such that $w \in \mathcal{N}_{w^0}(\eta, \gamma, \theta)$. Let $\widetilde{\Delta w} := \Delta w + \eta(w^0 - w^*)$. Using relations (32), (33), (35), and the fact that $w \in \mathcal{N}_{w^0}(\eta, \gamma, \theta)$, we easily see that

$$A\widetilde{\Delta x} = 0, \tag{75}$$

$$A^T\widetilde{\Delta y} + \widetilde{\Delta s} + V\widetilde{\Delta z} = 0, \tag{76}$$

$$V^T\widetilde{\Delta x} + E^{-2}\widetilde{\Delta z} = E^{-1}(q - r_V + \eta r_V^0). \tag{77}$$

Premultiplying (76) by $\widetilde{\Delta x}^T$ and using (75) and (77), we obtain

$$\begin{aligned} \widetilde{\Delta x}^T\widetilde{\Delta s} = -\widetilde{\Delta x}^T V\widetilde{\Delta z} \;&=\; [E^{-2}\widetilde{\Delta z} - E^{-1}(q - r_V + \eta r_V^0)]^T\widetilde{\Delta z} \\ &= \|E^{-1}\widetilde{\Delta z}\|^2 - (q - r_V + \eta r_V^0)^T(E^{-1}\widetilde{\Delta z}) \\ &\geq \|E^{-1}\widetilde{\Delta z}\|^2 - \|q - r_V + \eta r_V^0\|\,\|E^{-1}\widetilde{\Delta z}\| \;\geq\; -\frac{\|q - r_V + \eta r_V^0\|^2}{4}. \end{aligned} \tag{78}$$

Next, we multiply (34) by $(XS)^{-1/2}$ to obtain $D^{-1}\Delta x + D\Delta s = H(\sigma) - (XS)^{-1/2}p$, where $H(\sigma) := -(XS)^{1/2}e + \sigma\mu(XS)^{-1/2}e$. Equivalently, we have that

$$D^{-1}\widetilde{\Delta x} + D\widetilde{\Delta s} = H(\sigma) - (XS)^{-1/2}p + \eta\left[D(s^0 - s^*) + D^{-1}(x^0 - x^*)\right] \;=:\; g.$$

Taking the squared norm of both sides of the above equation and using (78), we obtain

$$\begin{aligned} \|D^{-1}\widetilde{\Delta x}\|^2 + \|D\widetilde{\Delta s}\|^2 = \|g\|^2 - 2\widetilde{\Delta x}^T\widetilde{\Delta s} \;&\leq\; \|g\|^2 + \frac{\|q - r_V + \eta r_V^0\|^2}{2} \\ &\leq \left(\|g\| + \frac{\|q\| + \|r_V - \eta r_V^0\|}{\sqrt{2}}\right)^2 \;\leq\; \left(\|g\| + \theta\sqrt{\mu}\right)^2, \end{aligned}$$

since $\|q\| + \|r_V - \eta r_V^0\| \leq \left[\sqrt{2} - 1\right]\theta\sqrt{\mu} + \theta\sqrt{\mu} = \sqrt{2}\theta\sqrt{\mu}$ by (36), (39), and the fact that $1 + (1 - \gamma/2)\sigma \leq 2$. Thus, we have

$$\begin{aligned} &\max(\|D^{-1}\widetilde{\Delta x}\|\,,\, \|D\widetilde{\Delta s}\|) \;\leq\; \|g\| + \theta\sqrt{\mu} \\ &\leq\; \|H(\sigma)\| + \|(XS)^{-1/2}\|\,\|p\| + \eta\left[\|D(s^0 - s^*)\| + \|D^{-1}(x^0 - x^*)\|\right] + \theta\sqrt{\mu}. \end{aligned}$$

This, together with the triangle inequality, the definitions of D and $\widetilde{\Delta w}$, and the fact that $w \in \mathcal{N}_{w^0}(\eta, \gamma, \theta)$, implies that

(79)
$$\begin{aligned}
&\max(\|D^{-1}\Delta x\|, \|D\Delta s\|) \\
&\le \|H(\sigma)\| + \|(XS)^{-1/2}\|\,\|p\| + 2\eta\left[\|D(s^0 - s^*)\| + \|D^{-1}(x^0 - x^*)\|\right] + \theta\sqrt{\mu} \\
&\le \|H(\sigma)\| + \|(XS)^{-1/2}\|\,\|p\| + 2\eta\|(XS)^{-1/2}\|\left[\|X(s^0 - s^*)\| + \|S(x^0 - x^*)\|\right] + \theta\sqrt{\mu} \\
&\le \|H(\sigma)\| + \frac{1}{\sqrt{(1-\gamma)\mu}}\left[\|p\| + 2\eta\left(\|X(s^0 - s^*)\| + \|S(x^0 - x^*)\|\right)\right] + \theta\sqrt{\mu}.
\end{aligned}$$

It is well known (see, e.g., [10]) that

$$\|H(\sigma)\| \le \left(1 - 2\sigma + \frac{\sigma^2}{1-\gamma}\right)^{1/2}\sqrt{n\mu}. \tag{80}$$

Moreover, using the fact that $s^* \le s^0$ and $x^* \le x^0$ along with Lemma 4.2, we obtain

$$\eta\left(\|X(s^0 - s^*)\| + \|S(x^0 - x^*)\|\right) \;\le\; \eta(s^T x^0 + x^T s^0) \;\le\; \left(3n + \frac{\theta^2}{4}\right)\mu. \tag{81}$$

The result now follows by noting that $\|p\| \le \sqrt{n}\|p\|_\infty$ and by incorporating inequalities (80), (81), and (38) into (79). ∎

We are now ready to prove Theorem 2.2.

Proof. Let Δw^k denote the search direction, and let $r^k = r(w^k)$ and $\mu_k = \mu(w^k)$ at the kth iteration of the inexact PDIPF algorithm. Clearly, $w^k \in \mathcal{N}_{w^0}(\gamma, \theta)$. Hence, using Lemma 4.8, assumption (40), and the inequality

$$\|\Delta X^k \Delta s^k\|_\infty \;\le\; \|\Delta X^k \Delta s^k\| \;\le\; \|(D^k)^{-1}\Delta x^k\|\,\|D^k \Delta s^k\|,$$

we easily see that $\|\Delta X^k \Delta s^k\|_\infty = \mathcal{O}(n^2)\mu_k$. Using this conclusion together with assumption (40) and Lemma 4.7, we see that, for some universal constant $\beta > 0$, we have

$$\mu_{k+1} \;\le\; \left(1 - \frac{\beta}{n^2}\right)\mu_k \quad \forall k \ge 0.$$

Using this observation and some standard arguments (see, for example, Theorem 3.2 of [27]), we easily see that the inexact PDIPF algorithm generates an iterate $w^k \in \mathcal{N}_{w^0}(\gamma, \theta)$ satisfying $\mu_k/\mu_0 \le \epsilon$ within $\mathcal{O}\left(n^2 \log(1/\epsilon)\right)$ iterations. The theorem now follows from this conclusion and the definition of $\mathcal{N}_{w^0}(\gamma, \theta)$. ∎

5. Concluding remarks. We have shown that the long-step PDIPF algorithm for LP based on an iterative linear solver presented in [16] can be extended to the context of CQP. This was not immediately obvious at first since the standard normal equation for CQP does not fit into the mold required for the results of [17] to hold. By considering the ANE, we were able to use the results about the MWB preconditioner developed in [17] in the context of CQP. Another difficulty we encountered was the proper choice of the starting iterate u^0 for the iterative linear solver. By choosing $u^0 = 0$ as in the LP case, we obtain $\|v - Wu^0\| = \|v\|$, which can only be shown to be $\mathcal{O}(\max\{\mu, \sqrt{\mu}\})$. In this case, for every $\mu > 1$, Proposition 3.2 would guarantee that the number of inner iterations of the iterative linear solver is

$$\mathcal{O}\left(\psi(\varphi_{\tilde{A}}^2)\max\left\{\log\left(c(\varphi_{\tilde{A}}^2)n\varphi_{\tilde{A}}\right), \log\mu\right\}\right),$$

a bound which depends on the logarithm of the current duality gap. On the other hand, Theorem 3.5 shows that choosing u^0 as in (53) results in a bound that does not depend on the current duality gap.

We observe that under exact arithmetic, the CG algorithm applied to $Wu = v$ generates an exact solution in at most $m + l$ iterations (since $W \in \Re^{(m+l)\times(m+l)}$). It is clear, then, that the bound (55) is generally worse than the well-known finite termination bound for CG. However, our results in section 3 were given for a family of iterative linear solvers, only one member of which is CG. Also, under finite precision arithmetic, the CG algorithm loses its finite termination property, and its convergence rate behavior in this case is still an active topic of research (see, e.g., [8]). Certainly, the impact of finite precision arithmetic on our results is an interesting open issue.

Clearly, the MWB preconditioner is not suitable for dense CQP problems since, in this case, the cost to construct the MWB is comparable to the cost to form and factorize $\tilde{A}\tilde{D}^2\tilde{A}^T$, and each inner iteration would require $\Theta((m+l)^2)$ arithmetic operations, the same cost as a forward and backward substitution. There are, however, some classes of CQP problems for which the method proposed in this paper might be useful. One class of problems for which PDIPF methods based on MWB preconditioners might be useful are those for which bases of $\tilde{A}$ are sparse, but the ANE coefficient matrices $\tilde{A}\tilde{D}^2\tilde{A}^T$ are dense; this situation generally occurs in sparse CQP problems for which n is much larger than $m + l$. Other classes of problems for which our method might be useful are network flow problems. The paper [22] developed interior-point methods for solving the minimum cost network flow problem based on iterative linear solvers with maximum spanning tree preconditioners. Related to this work, we believe that the following two issues could be investigated: (i) whether the incorporation of the correction term p defined in (29) in the algorithm implemented in [22] will improve the convergence of the method; (ii) whether our algorithm might be efficient for network flow problems where the costs associated with the arcs are quadratic functions of the arc flows. Identification of other classes of CQP problems which could be efficiently solved by the method proposed in this paper is another topic for future research.

Regarding the second question above, it is easy to see (after a suitable permutation of the variables) that $V^T = \begin{pmatrix} I & 0 \end{pmatrix}$ and E^2 is a positive diagonal matrix whose diagonal elements are the positive quadratic coefficients. In this case, it can be shown that $\tilde{A}$ is totally unimodular; hence $\varphi_{\tilde{A}}^2 \leq (m+l)(n-m+1)$ by Cramer's rule (see [17]).

The usual way of defining the dual residual is as the quantity

$$R_d := A^T y + s - VE^2V^T x - c,$$

which, in view of (11) and (12), can be written in terms of the residuals r_d and r_V as

$$R_d = r_d - VEr_V. \tag{82}$$

Note that, along the iterates generated by the inexact PDIPF algorithm, we have $r_d = \mathcal{O}(\mu)$ and $r_V = \mathcal{O}(\sqrt{\mu})$, implying that $R_d = \mathcal{O}(\sqrt{\mu})$. Hence, while the usual primal residual converges to 0 according to $\mathcal{O}(\mu)$, the usual dual residual does so according to $\mathcal{O}(\sqrt{\mu})$. This is a unique feature of the convergence analysis of our algorithm in that it contrasts with the analysis of other interior-point PDIPF algorithms, where the primal and dual residuals are required to go to zero at the same rate. The convergence analysis under these circumstances is possible due to the specific form of the $\mathcal{O}(\sqrt{\mu})$-term present in (82), i.e., one that lies in the range space of VE.

CQP problems where V is explicitly available arise frequently in the literature. One important example arises in portfolio optimization (see [6]), where the rank of V is often small. In such problems, l represents the number of observation periods used to estimate the data for the problem. We believe that the inexact PDIPF algorithm could be of particular use for this type of application.

REFERENCES

[1] K. M. Anstreicher, *Linear programming in $\mathcal{O}(n^3/(\ln n)\mathrm{L})$ operations*, SIAM J. Optim., 9 (1999), pp. 803–812.

[2] V. Baryamureeba and T. Steihaug, *On the convergence of an inexact primal-dual interior point method for linear programming*, in Large-Scale Scientific Computing, Lecture Notes in Comput. Sci. 3743, Springer-Verlag, Berlin, 2006, pp. 629–637.

[3] V. Baryamureeba, T. Steihaug, and Y. Zhang, *Properties of a Class of Preconditioners for Weighted Least Squares Problems*, Tech. report 99-16, Department of Computational and Applied Mathematics, Rice University, Houston, 1999.

[4] L. Bergamaschi, J. Gondzio, and G. Zilli, *Preconditioning indefinite systems in interior point methods for optimization*, Comput. Optim. Appl., 28 (2004), pp. 149–171.

[5] E. G. Boman, D. Chen, B. Hendrickson, and S. Toledo, *Maximum-weight-basis preconditioners*, Numer. Linear Algebra Appl., 11 (2004), pp. 695–721.

[6] T. J. Carpenter and R. J. Vanderbei, *Symmetric indefinite systems for interior-point methods*, Math. Programming, 58 (1993), pp. 1–32.

[7] R. W. Freund, F. Jarre, and S. Mizuno, *Convergence of a class of inexact interior-point algorithms for linear programs*, Math. Oper. Res., 24 (1999), pp. 50–71.

[8] A. Greenbaum, *Iterative Methods for Solving Linear Systems*, Frontiers Appl. Math. 17, SIAM, Philadelphia, 1997.

[9] M. Kojima, N. Megiddo, and S. Mizuno, *A primal-dual infeasible-interior-point algorithm for linear programming*, Math. Programming, 61 (1993), pp. 263–280.

[10] 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, Springer-Verlag, New York, 1989, pp. 29–47.

[11] M. Kojima, M. Shida, and M. Shindoh, *Search directions in the SDP and monotone SDLCP: Generalization and inexact computation*, Math. Program., 85 (1999), pp. 51–80.

[12] J. Korzak, *Convergence analysis of inexact infeasible-interior-point algorithms for solving linear programming problems*, SIAM J. Optim., 11 (2000), pp. 133–148.

[13] V. V. Kovacevic-Vujcic and M. D. Asic, *Stabilization of interior-point methods for linear programming*, Comput. Optim. Appl., 14 (1999), pp. 331–346.

[14] D. G. Luenberger, *Linear and Nonlinear Programming*, Addison–Wesley, Reading, MA, 1984.

[15] S. Mizuno and F. Jarre, *Global and polynomial-time convergence of an infeasible-interior-point algorithm using inexact computation*, Math. Program., 84 (1999), pp. 357–373.

[16] R. D. C. Monteiro and J. W. O'Neal, *Convergence Analysis of a Long-Step Primal-Dual Infeasible Interior-Point LP Algorithm Based on Iterative Linear Solvers*, Tech. report, Georgia Institute of Technology, Atlanta, 2003.

[17] R. D. C. Monteiro, J. W. O'Neal, and T. Tsuchiya, *Uniform boundedness of a preconditioned normal matrix used in interior point methods*, SIAM J. Optim., 15 (2004), pp. 96–100.

[18] Y. Nesterov and A. Nemirovskii, *Interior Point Polynomial Algorithms in Convex Programming*, Stud. Appl. Math. 13, SIAM, Philadelphia, 1995.

[19] A. R. L. Oliveira and D. C. Sorensen, *A new class of preconditioners for large-scale linear systems from interior point methods for linear programming*, Linear Algebra Appl., 394 (2005), pp. 1–24.

[20] L. F. Portugal, M. G. C. Resende, G. Veiga, and J. J. Judice, *A truncated primal-infeasible dual-feasible network interior point method*, Networks, 35 (2000), pp. 91–108.

[21] J. Renegar, *Condition numbers, the barrier method, and the conjugate-gradient method*, SIAM J. Optim., 6 (1996), pp. 879–912.

[22] M. G. C. Resende and G. Veiga, *An implementation of the dual affine scaling algorithm for minimum-cost flow on bipartite uncapacitated networks*, SIAM J. Optim., 3 (1993), pp. 516–537.

[23] M. J. Todd, L. Tunçel, and Y. Ye, *Characterizations, bounds, and probabilistic analysis of two complexity measures for linear programming problems*, Math. Program., 90 (2001), pp. 59–69.

[24] K.-C. Toh and M. Kojima, *Solving some large scale semidefinite programs via the conjugate residual method*, SIAM J. Optim., 12 (2002), pp. 669–691.
[25] P. Vaidya, *Solving Linear Equations with Symmetric Diagonally Dominant Matrices by Constructing Good Preconditioners*, Tech. report, Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, 1990.
[26] S. A. Vavasis and Y. Ye, *A primal-dual interior point method whose running time depends only on the constraint matrix*, Math. Program., 74 (1996), pp. 79–120.
[27] S. J. Wright, *Primal-Dual Interior-Point Methods*, SIAM, Philadelphia, 1997.
[28] Y. Zhang, *On the convergence of a class of infeasible interior-point methods for the horizontal linear complementarity problem*, SIAM J. Optim., 4 (1994), pp. 208–227.
[29] G. Zhou and K.-C. Toh, *Polynomiality of an inexact infeasible interior point algorithm for semidefinite programming*, Math. Program., 99 (2004), pp. 261–282.

SIAM J. OPTIM.
Vol. 17, No. 2, pp. 311–330
© 2006 Society for Industrial and Applied Mathematics

OPTIMIZING OVER CONSECUTIVE 1'S AND CIRCULAR 1'S CONSTRAINTS*

DORIT S. HOCHBAUM† AND ASAF LEVIN‡

Abstract. We consider packing and covering optimization problems over constraints in consecutive and circular 1's. Such problems arise in the context of shift scheduling, and in problems related to interval graphs. Previous approaches to this problem depended on solving several minimum cost network flow problems. We devise here substantially more efficient and strongly polynomial algorithms based on parametric shortest paths approaches. The objective function in the covering and packing problems is to either minimize or maximize the number of sets that satisfy the constraints. The various problems studied are classified according to whether the constraints are all consecutive 1's or if there are also circular 1's constraints, and according to whether the constraints are all of covering type; all of packing type, or mixed. The running time of our algorithm for a pure covering all consecutive 1's constraints problem on n variables and m constraints is $O(m+n)$. For the pure packing problem with consecutive 1's constraints we present an $O(m+n\log n)$ time algorithm. For the "mixed" case with both covering and packing consecutive 1's constraints we present an $O(mn)$ time algorithm. An $O(mn+n^2\log n)$-time algorithm is presented for the case where the constraints are *circular* (consecutive 1's constraint is also circular) of pure type—either all covering constraints or all packing constraints. Finally, we show an $O(n\min\{mn, n^2\log n + m\log^2 n\})$ time algorithm for the most general problem of mixed covering and packing case where the constraints are *circular*. All our algorithms are strongly polynomial and improve on the nonstrongly polynomial parametric minimum cost network flow or the (strongly polynomial) linear programming known approaches.

Key words. circular scheduling problems, consecutive 1's constraints, parametric optimization, parametric shortest path

AMS subject classifications. 90C27, 68Q25

DOI. 10.1137/040603048

1. Introduction.

Here we study optimization problems over constraints with $0,1$ coefficients that have the consecutive and circular 1's property. The basic consecutive 1's problem is formulated for a given set of m pairs, or intervals, $A = A_{cover} \cup A_{pack}$. Each interval (p,q) for $p<q$ corresponds to the $0,1$ vector, $[0,\ldots,0,1,\ldots,1,0,\ldots,0]$ with the positions $p+1,\ldots,q$ with value 1 and all others with value 0. A constraint (p,q) is said to be in consecutive 1's if it is of the covering form $\sum_{j=p+1}^{q} x_j \geq b_{pq}$ or packing form, $\sum_{j=p+1}^{q} x_j \leq b_{pq}$. The formulation of the consecutive 1's problem on a mixed set of covering and packing constraints is

$$\text{(Consec)} \qquad \begin{array}{lll} \min & \sum_{j=1}^{n} x_j & \\ \text{subject to:} & \sum_{j=p+1}^{q} x_j \geq b_{pq} & \text{for } 0\leq p<q\leq n \text{ and } (p,q)\in A_{cover} \\ & \sum_{j=p+1}^{q} x_j \leq b_{pq} & \text{for } 0\leq p<q\leq n \text{ and } (p,q)\in A_{pack} \\ & x_j \geq 0 \ \text{ integer} & j=1,\ldots,n. \end{array}$$

Although not explicitly stated, the packing and covering constraints allow modeling of upper and lower bounds on the variables, respectively. So the nonnegativity need

*Received by the editors January 11, 2004; accepted for publication (in revised form) December 6, 2005; published electronically May 19, 2006. Research supported in part by NSF awards DMI-0085690 and DMI-0084857.

http://www.siam.org/journals/siopt/17-2/60304.html

†Department of Industrial Engineering and Operations Research and Walter A. Haas School of Business, University of California, Berkeley, CA 94720 (hochbaum@ieor.berkeley.edu).

‡Department of Statistics, The Hebrew University, Jerusalem, Israel (levinas@mscc.huji.ac.il).

not be listed explicitly, as we have in the formulation. When $A_{pack} = \emptyset$ the problem is the well known *set cover* problem.

When substituting the variables $y_q = \sum_{i=1}^{q} x_i$ in (Consec) the constraints of the problem become $y_q - y_p \geq b_{pq}$ and $y_q - y_p \leq b_{pq}$. (Detailed discussion of this transformation is provided in section 2.) Such constraints are recognized as the dual of the minimum cost network flow problem, and previous techniques for solving the problem are indeed based on solving minimum cost network flow problems.

The circular 1's problem (Circular) includes, in addition to consecutive 1's constraints, also at least one constraint of the type

$$\sum_{j=1}^{q'} x_j + \sum_{j=p'+1}^{n} x_j \geq b_{q'p'} \quad \text{for } 0 \leq q' < p' \leq n,$$

or

$$\sum_{j=1}^{q'} x_j + \sum_{j=p'+1}^{n} x_j \leq b_{q'p'} \quad \text{for } 0 \leq q' < p' \leq n.$$

Such constraints, called *circular*, are characterized by having an entry 1 in the first and last columns of the constraint coefficient. That is, a circular constraint corresponding to (p', q') for $p' > q'$ is represented by a $0,1$ vector $[1, \ldots, 1, 0, \ldots, 0, 1, \ldots, 1]$ with positions $q'+1$ through p' having 0 value, and the rest are 1. We refer to the problem with constraints that include consecutive 1's and circular 1's as (Circular).

We study here problems (Consec) and (Circular) with either $\max \sum_{j=1}^{n} x_j$ or $\min \sum_{j=1}^{n} x_j$ as the objective function. These problems have applications ranging from problems on interval and circular-arc graphs, to staff scheduling. Problems on interval and circular-arc graphs that can be modeled using (Consec) and (Circular) include the minimum dominating set where all the constraints are covering constraints with the right-hand side equaling 1, and the maximum independent set where all the constraints are packing constraints with the right-hand side equaling 1. The reader is referred to [BOR80] for details on the application of staff scheduling.

The recognition problem of whether the constraint matrix is of consecutive 1's or circular 1's type is polynomially solvable. Booth and Lueker, [BL76] showed that given a $0,1$ matrix of size $m \times n$ with f 1's, one can verify in linear time, $O(m+n+f)$, whether the matrix has the consecutive 1's property . It is also possible to test quickly whether the matrix has the circular 1's property in O(m+n+2f), [Boo75].

We present here combinatorial and strongly polynomial time algorithms which are not based on flow and yield improved run times for (Consec) and (Circular) as reported in Table 1.1.

Veinott and Wagner [VW62] studied the problem (Consec) and established its relationship to the minimum cost network flow problem. We use this relationship to conclude the known result for mixed (Consec) problem reported in Table 1.1. The pure covering (Consec) problem and the pure packing (Consec) problem were solved by Tamir [Ta03]. He considered these problems when each row of the constraint matrix corresponds to a neighborhood in a tree. For this general problem he designed an $O((m+n)\log^2 n)$ time algorithm, and noted that in the special case of neighborhoods on a line (i.e., each row is consecutive 1's) the time complexity reduced to $O((m+n)\log n)$. His results were obtained for the problem where the coefficient matrix is not given as a set of intervals of the consecutive 1's, and therefore the $\log n$ factor is essential to transform the problem into the set of intervals we have as our input.

TABLE 1.1
Complexity of algorithms for optimizing over consecutive 1*'s and circular* 1*'s matrices.*

Problem	**Known best result**	**Running time here**
Consecutive covering constraints	$O((m+n)\log n)$ [Ta03]	$O(m+n)$
Consecutive packing constraints	$O((m+n)\log n)$ [Ta03]	$O(m+n\log n)$
Consecutive packing and covering constraints	MCNF [VW62]	$O(mn)$
Circular covering constraints	$O(mn\log n)$ [KO81]	$O(nm+n^2\log n)$
Circular packing constraints	LP, or $O(\log b \cdot MCNF)$ [BOR80]	$O(nm+n^2\log n)$
Circular packing and covering constraints	LP, or $O(\log b \cdot MCNF)$ [BOR80]	$O(n^2\min\{m, n\log n\})$

Legend: **MCNF**, the complexity of solving minimum cost network flow. **LP**, the complexity of linear programming with $0,1$ constraint coefficients. $b=\sum_{(p,q)\in A} b_{pq}$.

Shah and Farach-Colton [SF02] presented an $O((m+n)\log n)$ time algorithm that finds the optimal solution value (not the solution itself) of the pure covering and pure packing problems where each row of the constraint matrix corresponds to a neighborhood in a tree.

Bartholdi, Orlin, and Ratliff [BOR80] were the first to propose a polynomial time algorithm for a generalization of the (Circular) problem in which the cost coefficients are not identical, and demonstrated that the linear optimization problem over constraints with circular 1's is solvable in polynomial time as well although the constraint matrix is no longer totally unimodular. Although they studied the problem with only covering constraints, their results hold also for problems with packing constraints as well, as we show in section 4. The running time of their algorithm is $\log b$ times the complexity of solving a minimum cost network flow problem (MCNF), where b is the sum of the right-hand sides. A second algorithm they devised solves the problem by calling twice to the linear programming relaxation (in the second linear programming relaxation they fix the value of $\sum_j x_j$). They showed that for the special objective function $\min \sum_j x_j$ and pure covering, it is enough to solve a single linear programming relaxation and then round-up the fractional solution vector to obtain an optimal integral solution. A similar result holds also for pure packing constraints but the fractional solution vector is rounded-down. For the mixed case we need to check the two possible solutions obtained by rounding-up and by rounding-down the optimal fractional solution. One of these solutions is guaranteed to be feasible if the problem itself is feasible. Then, we need to check whether these solutions are feasible, and if both are feasible the optimal integral solution is the better one.

A combinatorial linear program with all entries in the constraint matrix that are "small" is solvable in strongly polynomial time, [Tar86]. Thus the second, LP-based algorithm of [BOR80] is strongly polynomial. The drawback of employing this linear programming algorithm is that in order to achieve strongly polynomial time one has to use the Ellipsoid method which is neither efficient nor practical. We will show how the first algorithm of [BOR80] that is based on MCNF can be transformed into a strongly-polynomial time algorithm, albeit with run time that is still inferior to the run time of the algorithm reported here.

Karp and Orlin [KO81] solved problem (Circular) when all the constraints are covering constraints using parametric shortest path method in $O(mn\log n)$ time.

A related problem to the ones studied here is the optimization over constraints with circular 1's in *columns*. Hochbaum and Levin [HL03] showed that this problem is

at least as difficult as the *exact matching* problem, and thus harder than the problem in circular 1's in *rows* investigated here. They gave a 2-approximation algorithm for the problem, and presented an $O(n^3 \log B + n^4)$-time algorithm for the special case where right-hand sides are uniform and equal to B.

Paper overview. In section 2 we provide a description of Veinott and Wagner's [VW62] transformation. In section 3, we describe the parametric method of Megiddo [Meg83] and Cole [Col87] which is used to improve the running time of several of the algorithms presented. In section 4 and in the appendix we show how the results of Bartholdi, Orlin, and Ratliff [BOR80] are extended to the mixed case where there are packing constraints and covering constraints (rather than the pure covering constraints they investigated). We also devise a strongly-polynomial variant of the [BOR80]'s algorithm based on the parametric method. In section 5, we address the problem (Consec) with pure covering constraints and present an $O(m + n)$-time algorithm that solves it, and in section 6 we derive for (Consec) with pure packing constraints an $O(m + n \log n)$-time algorithm. In section 7 we show that pure packing problems are at least as difficult as pure covering problems. In section 8 we show an $O(mn)$-time algorithm for the *mixed* problem (Consec). In section 9 we discuss the pure packing circular 1's problem and present an algorithm with complexity $O(mn + n^2 \log n)$. Finally, in section 10 we present an algorithm for the general circular 1's case with both covering and packing constraints. The total time complexity of that algorithm is $O(n \min\{mn, n^2 \log n + m \log^2 n\})$.

Notation. For $i < j$, $[i, j]$ is the interval of integers $\{i, i+1, \ldots, j\}$. We use the notation convention $\mathbf{x}$ to refer to the *vector* $\{x_j\}_{j=1}^n$. The vector $\mathbf{e}_j$ is the vector of $n-1$ zeros and one 1 in the jth position.

2. The transformation and definitions. Veinott and Wagner [VW62] suggested the following transformation for problems on consecutive ones. Let the set of variables y_j be defined as follows: $y_0 = x_0 \equiv 0$, $y_j = \sum_{i=0}^{j} x_i$ for $j = 1, 2, \ldots, n$. The set of constraints of (Consec) in terms of the new variables are

$$\begin{aligned} y_j - y_i &\geq b_{ij} \quad \forall\ (i,j) \in A_{cover}, \\ y_j - y_i &\leq b_{ij} \quad \forall\ (i,j) \in A_{pack}, \\ y_j - y_{j-1} &\geq 0 \quad \forall j. \end{aligned}$$

This set of constraints has one 1 and one -1 in each row. The coefficient matrix of such a set of constraints is totally unimodular, and furthermore it forms the constraints of the dual of MCNF problem. This implies the following polynomial time algorithm for (Consec): Take the dual of the transformed problem, solve it with a minimum cost network flow procedure, and construct the dual solution (node potentials, in minimum cost network flow terminology) by a shortest paths procedure. We note that the dual node potentials are computed explicitly as part of Orlin's [Orl93] minimum cost network flow procedure. Therefore, if we use Orlin's algorithm, then the last step of computing the shortest paths is redundant.

Using the transformation above for the (Circular) problem, a circular 1's covering constraint $\sum_{j=1}^{q'} x_j + \sum_{j=p'+1}^{n} x_j \geq b_{q'p'}$ is mapped into $y_{q'} - y_{p'} + y_n \geq b_{q'p'}$. So circular constraints have, in addition to one 1 and one -1, also an additional coefficient 1 for the variable y_n. This renders the constraint matrix no longer totally unimodular. Writing the transformed constraint as $y_{q'} - y_{p'} \geq b_{q'p'} - y_n$, or equivalently as $y_{p'} - y_{q'} \leq y_n - b_{q'p'}$, it is possible to treat y_n as a parameter and solve the problem

as a parametric dual of minimum cost network flow. Thus a parametric approach plays a crucial role in solving circular problems.

3. The parametric method. In a parametric optimization problem some right-hand sides and some cost coefficients are given as linear functions of a single common parameter λ. The parametric method, introduced by Megiddo [Meg83] and improved for some special cases by Cole [Col87], solves a parametric optimization problem. The goal is to find an optimal value λ^* so that for the instance of the optimization problem where each linear function of λ is evaluated at λ^* the optimal solution has a maximum cost (among all values of λ).

The parametric shortest path problem and the parametric minimum cost network flow problem are defined on a graph $G = (V, E)$, where each arc $e \in E$ has a cost c_e that is a linear function of a common parameter λ. In the minimum cost network flow we are also given a demand vector d. The goal in the parametric shortest path problem is to compute a value λ^* for which the length of the shortest path between s and t is maximized. Similarly, the goal for the parametric MCNF problem is to compute a value λ^* that maximizes the optimal cost of the minimum cost network flow problem instance.

The methods of [Col87, Meg83] both use a parallel algorithm with $O(f(n))$ processors and $O(g(n))$ parallel time that solves the optimization problem for a single value of λ (the nonparametric problem). The method "simulates" the execution of this parallel algorithm for $\lambda = \lambda^*$ without the knowledge of λ^*. A rough sketch of the idea is as follows: In each parallel time unit a set of $O(f(n))$ comparisons needs to be answered. For a single comparison, Megiddo [Meg83] proposed to use an algorithm that solves the nonparametric problem in the breakpoint of the two linear functions (of λ) that we have to compare (this is done by a call to a serial algorithm with time complexity $T(n)$). Megiddo [Meg83] suggested using a binary search over these comparisons to evaluate the $O(f(n))$ comparisons of a single parallel time unit using only $O(\log f(n))$ comparisons. This results in time complexity of $O(f(n)g(n) + (\log f(n))g(n)T(n))$ for the parametric problem.

Cole [Col87] suggested that one should "slow down" the comparisons that this algorithm evaluates, and use a sorting network instead of the comparisons. Using Cole's method the number of comparisons that the algorithm evaluates is only $O(\log f(n) + g(n))$. Therefore, the total complexity is $O(f(n)g(n) + [\log f(n) + g(n)]T(n))$. Cole's improvement is suitable only for cases when the parallel algorithm is based on sorting. Cole explicitly stated that his improvement cannot be applied to a general parametric problem but only to special cases. Cole also designed an algorithm with time complexity of $O(n^3 \log n + (\log^2 n)T(n))$ for the parametric shortest path problem where $T(n)$ is the time complexity of a single comparison. His algorithm was designed for the minimum ratio cycle problem where a comparison is answered using a negative cycle detector. The result, however, holds also for other applications of the parametric shortest path problem.

Since Orlin's [Orl93] algorithm for the minimum cost network flow uses $O(n \log n)$ shortest path computations, the parametric MCNF problem can be solved in $O(n^4 \log^2 n + n \log^3 nT(n))$, where $T(n)$ is the time complexity of a single comparison; using Orlin's serial algorithm for the MCNF problem $T(n) = O(n \log n(m + n \log n))$. Therefore, the parametric MCNF problem can be solved in $O(n^4 \log^2 n + n^2 m \log^4 n)$ time.

We conclude this discussion with the following theorem.

THEOREM 3.1. *The parametric shortest path problem can be solved in* $O(n^3 \log n +$

$(\log^2 n)T(n))$ *time where* $T(n)$ *is the time complexity of a single comparison. The parametric MCNF problem can be solved in* $O(n^4 \log^2 n + n^2 m \log^4 n)$ *time.*

4. Extensions on the algorithm to solve the general linear objective function of (Circular). The algorithm of Bartholdi, Orlin, and Ratliff is based on treating the transformed circular problem, as explained in section 2, as a parametric MCNF problem. The algorithm of [BOR80] consists of a binary search for the optimal value of $\lambda = y_n$ where in each call a MCNF problem is solved for a specific value of $\lambda = y_n$. That algorithm is applicable to (Circular) with a general linear objective function, $\min \sum_{j=1}^{n} c_j x_j$.

The algorithm in [BOR80] is given for the pure covering problem. We show in the appendix how to extend it to the mixed packing and covering constraints. A second modification we propose is a strongly polynomial time algorithm based on the parametric network flow algorithm in Theorem 3.1.

With these modifications we establish the following theorem (proved in the appendix).

THEOREM 4.1. *There is an* $O(n^4 \log^2 n + n^2 m \log^4 n)$*-time algorithm that solves problem (Circular) with a general linear cost function.*

In the remainder of this paper we consider the unweighted objective function case $c_j = 1\ \forall j$, and present faster algorithms for this problem and its special cases.

5. Covering consecutive 1's constraints. For the pure covering problem with $A = A_{cover}$, only a minimization objective is meaningful since the maximization problem $\max \sum_i x_i$ is trivially unbounded.

Consider the minimum *longest path problem* in an acyclic graph. This is an optimization problem in which the objective is to find the smallest bound within which we can traverse every path in a network where each arc (i, j) has a cost, or distance, b_{ij}. The formulation of the longest path problem is precisely the transformed formulation of the problem (Consec) on pure covering constraints. In the transformed problem variable y_i corresponds to node i and b_{ij} represents the *distance* between nodes i and j. The objective function of the minimum *longest path problem* is to minimize y_n which is equivalent to the objective function $\sum_{j=1}^{n} x_j$ in the formulation of (Consec).

Construct a graph $G = (V, A)$ corresponding to the formulation with a node $i \in V$ for each variable y_i. For every constraint of the type $y_j - y_i \geq b_{ij}$ there is a directed arc $(i, j) \in A$ with weight b_{ij}. Arcs $(i-1, i)$ correspond to the variable $x_i = y_i - y_{i-1}$ and can have length 0 for every nonnegativity constraint, or other lower bound constraint. The graph is illustrated in Figure 1. Since for every arc (i, j), $i < j$ the graph G is necessarily acyclic, or directed acyclic graph (DAG).

To demonstrate that our problem is indeed the longest path problem on the DAG G, let P be any path from 0 to n in G. Consider the inequality which is the sum of the constraints that correspond to the arcs of P. This inequality has a left-hand side $y_n - y_0$, and its right-hand side equals the length of P. Since $y_0 = 0$, we conclude that y_n is at least as large as the total length of P. Since P is an arbitrary path from 0 to n, y_n is at least the length of the longest path from 0 to n. Because our objective is to minimize y_n, the optimal value of y_n equals the length of the longest path from 0 to n.

The longest path problem is solvable in polynomial time on a DAG using dynamic programming (DP). Let the distance labels $\underline{y}_j^*$ be the length of the longest path from 0 to j. We compute lower bounds $\underline{y}_j^*$ on the partial sum variables y_j.

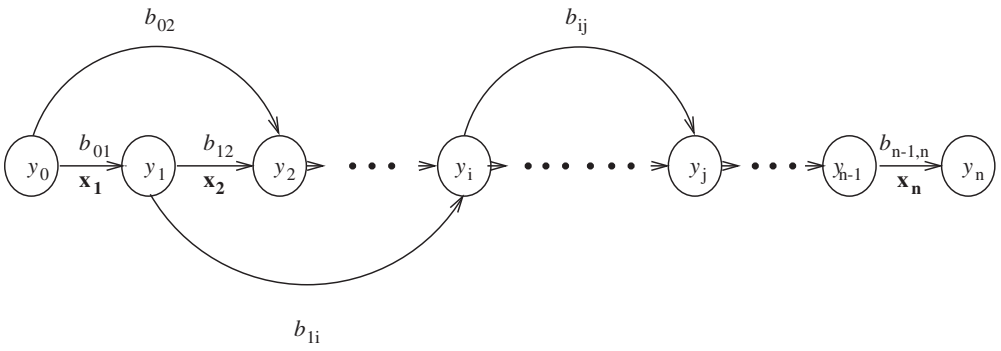


FIG. 1. *Directed Acyclic Graph* (DAG).

Let $\underline{y}_0^* = 0$. Then, once $\underline{y}_1^*, \underline{y}_2^*, \ldots, \underline{y}_{j-1}^*$ has been computed, we evaluate $\underline{y}_j^*$ using the following forward recursion:

Forward recursion	$\underline{y}_j^* = \max_{(i,j)\in A}\{\underline{y}_i^* + b_{ij}\}$

Every feasible solution with partial sums vector $\mathbf{y}$ obviously must satisfy $y_j \geq \underline{y}_j^*$ for $j = 1, \ldots, n$. Since the solution $x_j = \underline{y}_j^* - \underline{y}_{j-1}^*$ is feasible with objective value $\underline{y}_n^*$, it is an optimal solution. The validity of this DP recursion implies the following theorem.

THEOREM 5.1. *Problem (Consec) with pure covering constraints is solved in* $O(m+n)$ *time.*

6. Packing consecutive 1's constraints. For the pure packing problem $A = A_{pack}$, only a maximization objective is meaningful as the minimization problem $\min \sum_j x_j$ is trivially solved by $x_j = 0\ \forall j$.

For the transformed problem there is a corresponding graph $G = (V, A)$ with one node i corresponding to each variable y_i and an arc (i, j) of weight b_{ij} corresponding to each constraint $y_j - y_i \leq b_{ij}$. A nonnegativity constraint $x_j \geq 0$ is transformed into $y_j - y_{j-1} \geq 0$, and is represented by a backward zero arc directed from j to $j-1$.

The graph G contains cycles but the lengths of arcs are all nonnegative.

Our problem is the shortest path problem on G from 0 to n. To see this let P be any path from 0 to n in G. Consider the inequality which is the sum of the constraints that correspond to the arcs of P. This inequality has a left-hand side $y_n - y_0$, and its right-hand side equals the length of P. Since $y_0 = 0$, we conclude that y_n is at most as large as the total length of P. Since P is an arbitrary path from 0 to n, y_n is at most the length of the shortest path from 0 to n. Because our objective is to maximize y_n, the optimal value of y_n equals the length of the shortest path from 0 to n.

Since the lengths are nonnegative we can apply Dijkstra's algorithm [Dij59] to find the shortest path from 0 to n in G:

THEOREM 6.1. *Problem (Consec) with pure packing constraints is solved in* $O(m + n \log n)$ *time.*

7. The relation between the pure packing problem and the pure covering problem. In this section we consider problems (Consec) and (Circular) with pure covering constraints or pure packing constraints. We note that for the pure packing problem the meaningful objective function is $\max \sum_{j=1}^{n} x_j$ as a minimization

problem is trivially solved by the zero vector, and for the pure covering problem the meaningful objective function is $\min \sum_{j=1}^{n} x_j$ as a maximization problem is trivially unbounded.

Given a "pure" problem we let $M \geq \sum_{i=1}^{m} b_i$, and consider the transformation of the variables $(x_1, \ldots, x_n)$ to $(x'_1, \ldots, x'_n)$ defined by $x'_j = M - x_j$ for all j.

A covering constraint $\sum_{j \in S} x_j \geq b$ is transformed to $\sum_{j \in S}(M - x'_j) \geq b$ which is equivalent to the packing constraint $\sum_{j \in S} x'_j \leq M|S| - b$. Similarly, a packing constraint $\sum_{j \in S} x_j \leq b$ is transformed to the covering constraint $\sum_{j \in S} x'_j \geq M|S| - b$.

This transformation maps a consecutive (circular) constraint to a consecutive (circular) constraint. The objective function $\max \sum_{j=1}^{n} x_j$ is mapped to $nM + \max \sum_{j=1}^{n} (-x'_j)$ which is equivalent to $\min \sum_{j=1}^{n} x'_j$. Similarly, the objective function $\min \sum_{j=1}^{n} x_j$ is mapped to $nM + \min \sum_{j=1}^{n}(-x'_j)$ which is equivalent to $\max \sum_{j=1}^{n} x'_j$.

This transformation maps a pure covering problem into a pure packing problem. However, a pure packing problem is mapped into a pure covering problem with additional upper bounds constraints. These upper bounds result from the nonnegativity constraints in the pure packing formulation. This also explains the different time complexities for solving the pure covering (Consec) problem and the pure packing (Consec) problem.

This simple transformation proves the following theorem.

THEOREM 7.1. *Let $T(n, m)$ be a function that grows at least at a linear rate, $T(n, m) = \Omega(n)$. Then,*

1. *If there is an algorithm of complexity $T(n, m)$ that solves the pure packing (Circular) problem, then there is an algorithm of complexity $T(n, m)$ that solves the pure covering (Circular) problem.*
2. *There is an algorithm of complexity $T(n, m)$ that solves the mixed (Consec) problem with a maximization objective if and only if there is an algorithm of complexity $T(n, m)$ that solves the mixed (Consec) problem with a minimization objective.*
3. *There is an algorithm of complexity $T(n, m)$ that solves the mixed (Circular) problem with a maximization objective if and only if there is an algorithm of complexity $T(n, m)$ that solves the mixed (Circular) problem with a minimization objective.*

8. The mixed (Consec) problem. In this section we present an $O(mn)$ time algorithm for problem (Consec) with both covering and packing constraints.

We begin by applying Veinott and Wagner's [VW62] transformation, getting a transformed constraint of the type $y_q - y_p \leq b_{pq}$ for each packing constraint $(p, q) \in A_{pack}$, and $y_q - y_p \geq b_{pq}$ for each covering constraint $(p, q) \in A_{cover}$ (or nonnegativity constraint). By multiplying each covering constraint by -1, $y_p - y_q \leq -b_{pq}$, all constraints become uniformly of packing type, albeit no longer with positive right-hand sides.

We claim that the sum of variables at the optimum is the length of the *shortest path* from node 0 to n in the corresponding network, which is possibly cyclic and contains negative arc lengths. The network $G = (V, A)$ is composed of a node set $V = \{0, 1, \ldots, n\}$. The set of arcs A consists of three types of arcs: For every packing constraint $(p, q) \in A_{pack}$ there is an arc (p, q) of length b_{pq}; for every covering constraint $(p, q) \in A_{cover}$ there is an arc (q, p) of length $-b_{pq}$ (note that such an arc is in the reverse direction and has a negative length), and for every nonnegativity constraint $y_j - y_{j-1} \geq 0$ there is an arc $(j, j-1)$ of zero length.

To see that our problem is the shortest path from 0 to n in the resulting network

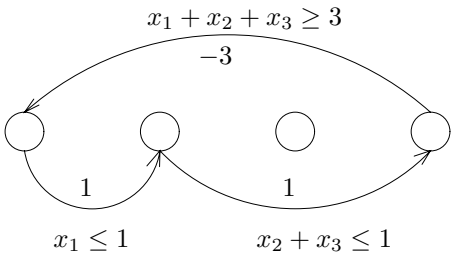


FIG. 2. *The graph G with a negative length cycle that corresponds to the set of constraints $x_1 \leq 1$, $x_2 + x_3 \leq 1$, and $x_1 + x_2 + x_3 \geq 3$.*

we use an argument similar to the one in section 5. Let $P = [j_o, j_1, j_2, \ldots, j_{k-1}, j_k]$, where $j_0 = 0$ and $j_k = n$ be an arbitrary path from 0 to n in G. Consider the inequality which is the sum of the constraints that correspond to the arcs of P,

$$(y_{j_1} - y_0) + (y_{j_2} - y_{j_1}) + \cdots + (y_n - y_{j_{k-1}}) \leq \sum_{\ell=1}^{n} b_{j_{\ell-1} j_\ell} = B_P.$$

This constraint has a left-hand side $y_n - y_0$, and its right-hand side equals the length of P, B_P. Since $y_0 = 0$, we conclude that y_n is at most as large as the total length of P. Since P is an arbitrary path from 0 to n, y_n is at most the length of a shortest path from 0 to n, $y_n \leq \min_P B_P$. Because our objective is to maximize y_n, the optimal value of y_n equals the length of a shortest path from 0 to n.

We now argue that the existence of a negative length cycle C in G implies that the original (Consec) problem is infeasible (as illustrated in an example in Figure 2). Summing up the constraints that correspond to the arcs of C we get 0 on the left-hand side, and the length of C on the right-hand side. That is, the aggregate constraint is $0 \leq$ a negative number. Therefore, the original problem is infeasible.

To solve problem (Consec) we can therefore use the Bellman–Ford algorithm for computing the shortest path from 0 to n in G. If there is a negative length cycle, then (Consec) is provably infeasible. Otherwise, the shortest path distance from node 0 to i is the value of the variable y_i in the optimal solution. All these distances are computed by the Bellman–Ford's algorithm in $O(mn)$ time.

THEOREM 8.1. *There is an $O(mn)$-time algorithm that solves mixed (Consec) problem.*

Remark 8.1. Under a restriction about the relative values of the coefficients of the objective function, the algorithm above can also solve (Consec) with nonuniform coefficients. The transformed constraints of (Consec) are monotone as in each row of the constraint matrix there is at most one positive coefficient and at most one negative coefficient. It is well known [HN94] that the feasible integral solution set on monotone constraints forms a lattice. In this lattice there is a least element and a largest element. The least element solves the problem of minimizing $\sum_j \bar{c}_j y_j$ for all nonnegative $\bar{c}$. The largest element of the lattice solves the problem of maximizing $\sum_j \bar{c}_j y_j$ for all nonnegative $\bar{c}$. The transformed problem (Consec) has nonnegative cost coefficients if the original cost coefficients are monotone nondecreasing. In this case we now show that our solution for problem (Consec) finds the least or largest element of the lattice. For the pure covering (Consec) problem our algorithm finds the least element of the lattice. To see this note that for each i y_i is the length of the longest path in G from 0 to i. Therefore, it is the optimal solution cost of $\min y_i$ subject to

the same transformed constraint matrix. This shows that the solution returned by the algorithm is the least element of the lattice. For pure packing (Consec) and mixed (Consec) our algorithms find the largest element of the lattice. To see this note that in both cases y_i is set to the length of a shortest path from 0 to i in G (or $\tilde{G}$). Therefore, it is the optimal solution value of $\max y_i$ subject to the same transformed constraint matrix. This shows that the solution returned by the algorithm is the largest element of the lattice.

If the cost coefficients are monotone nonincreasing we can reverse the order of the variables x_j (i.e., let $x'_j = x_{n+1-j}\ \forall j$), and then use the result in the previous paragraph.

9. Pure packing and pure covering (Circular). In this section we consider the pure cases of (Circular) with either all packing constraints or all consecutive constraints. We present an $O(n^3 \log n + mn \log^2 n)$ time algorithm for the pure packing problem, and show how to improve it to an $O(mn + n^2 \log n)$-time algorithm. For the pure covering problem the same algorithms are derived analogously.

Applying Veinott and Wagner's [VW62] transformation we get a problem where the objective function is $\max y_n$, and there are three types of constraints:

1. $y_j - y_k \leq b_{kj}$ $(n \geq j > k)$ for each consecutive ones constraint.
2. $y_k - y_j + y_n \leq b_{kj}$ $(n > j > k)$ for each circular ones constraint. Here we move y_n to the right-hand side to get a constraint $y_k - y_j \leq b_{kj} - y_n$ $(n > j > k)$.
3. $y_j - y_{j-1} \geq 0$ for each nonnegativity constraint. Multiplying by -1 the constraint is $y_{j-1} - y_j \leq 0$.

The resulting formulation is of the shortest path from 0 to n in a graph $\tilde{G} = (V, A)$, with nodes corresponding to the n variables and one additional node corresponding to $y_0 = 0$, and the set of arcs has all the consecutive constraints represented by *forward* arcs, (i, j) for $i < j$, each of cost b_{ij} (the circular constraints are represented by *backward* arcs (p, q) for $p > q$ of cost $b_{qp} - y_n$ and the nonnegativity constraints, are represented by zero cost *backward* arcs $(j, j-1)$). A graph of this type is displayed in Figure 3. Thus $\tilde{G}$ contains cycles, and parameterized arc costs with $\lambda = y_n$ the parameter. For values of y_n that are large enough there could be negative cost arcs.

Consider applying Cole's algorithm for computing the parametric shortest path from 0 to n in a parametric network in order to compute the optimal value of the parameter y_n^* for which the shortest path is maximum (see section 3). To apply this algorithm, we need a serial algorithm that resolves, for a given parameter value λ, whether y_n^* is greater than λ, equal to λ, or smaller than λ.

The selected serial algorithm is the Bellman–Ford algorithm computing a shortest path from 0 to n for the parameter value λ in $T(n) = O(mn)$ time. Note that decreasing λ increases all the arc lengths, and the length of the shortest path can thus only increase. Respectively, increasing λ only decreases the arc lengths, and the length of the shortest path can only decrease. Therefore, when applying a parametric shortest path algorithm in the network $\tilde{G}$ for a parameter λ, only the following outcomes are possible.

LEMMA 9.1. *Either $y_n^* \leq \lambda$ if there is a negative cycle or the shortest path is at most λ; or $y_n^* > \lambda$ if the shortest path is longer than λ.*

Using Theorem 3.1 with $T(n) = O(mn)$ we conclude that the resulting algorithm for computing y_n^* has a total time complexity of $O(n^3 \log n + \log^2 n(mn))$.

THEOREM 9.1. *Problem (Circular) where all constraints are packing constraints can be solved in $O(n^3 \log n + mn \log^2 n)$ time.*

COROLLARY 9.1. *Problem (Circular) where all constraints are covering con-*

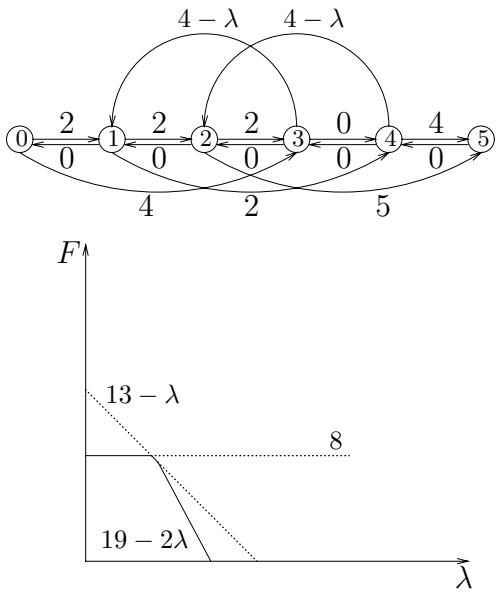


FIG. 3. *A graph $\tilde{G}$ with the corresponding function $F(\lambda) = \min\{8, 13 - \lambda, 19 - 2\lambda\}$.*

straints can be solved in $O(n^3 \log n + mn \log^2 n)$ time.

Proof. By Theorems 7.1 and 9.1. □

If there were a faster single-source shortest path algorithm that can run in poly-log parallel time, we could have gotten a faster algorithm for this case. However, we are not aware of such an algorithm.

An alternative algorithm with better run time to solve the parametric shortest path from 0 to n in $\tilde{G}$ is shown next.

We observe that the length of any path P in $\tilde{G}$ from 0 to n is a linear function of y_n with a slope $-k$ where k is the number of circular arcs in P. The key idea of the improved algorithm is to compute the *entire* lower envelope of the 0 to n shortest path length as a function of the parameter $\lambda = y_n$. For that we apply the algorithm of Young, Tarjan, and Orlin [YTO91], (YTO-algorithm) which computes the entire lower envelope $F(\lambda)$ of the 0 to n shortest path length when each arc has length that is a linear function of a common parameter λ with slope that is either 0 or -1. The time complexity of YTO-algorithm is $O(mn + n^2 \log n)$.

The lower envelope $F(\lambda)$ is a list of $O(n)$ linear functions of λ each representing the length of a shortest path from 0 to n in $\tilde{G}$ using k circular arcs (for some $k \in \{0, 1, 2, \ldots, n-1\}$). In Figure 3 we present a graph $\tilde{G}$ and the corresponding $F(\lambda)$. In this example the shortest path from 0 to 5 that uses only consecutive arcs is 8, with one circular arc is $13 - \lambda$, and two circular arcs is $19 - 2\lambda$.

By Lemma 9.1 and since $F(\lambda)$ is continuous monotone nonincreasing, then if a fixed point solution to the equation $y_n^* = F(y_n^*)$ exists, it is the correct optimal fractional solution to (Circular). Therefore, in order to find an optimal fractional solution, we seek the value y_n^* such that $y_n^* = F(y_n^*)$. It remains to consider the case in which there is no fixed point solution.

A *feasible value* of λ is a value for which the network $\tilde{G}$ does not contain a negative length cycle. Suppose that a fixed point solution y_n^* does not exist. For $\lambda = 0$, $F(0) \geq 0$ since for $\lambda = 0$ the network does not contain a negative length

arc. Since there is no fixed point solution, for all feasible values of λ, $F(\lambda) > \lambda$. Then, by Lemma 9.1, we can resolve comparisons, and therefore conclude that the optimal fractional solution is the maximum value for which the resulting network has no negative cycle. We denote this value by y_n^* as well.

The value of y_n^* can be deduced from the lower envelope in $O(n)$ time as follows: We traverse the list of intervals forming the linear sections of the lower envelope, seeking whether each intersects with the linear function λ. If such intersection exists, then this is the value y_n^*. Otherwise, y_n^* does not belong to this interval. For each interval we spend $O(1)$ time, and since there are $O(n)$ intervals, the total time complexity of this procedure is $O(n)$. If we traverse the entire lower envelope $F(\lambda)$ without reaching to a fixed point solution, then such a fixed point does not exist and we can compute the maximum value for which the resulting network has no negative cycle.

We next use the value of y_n^* to compute for each u, the shortest length of a path from 0 to u in $\tilde{G}$ where the length of the arcs are set using the parameter value $\lambda = y_n^*$. This can be done by using a single application of the Bellman–Ford algorithm on the graph where the values of λ are substituted by y_n^*, in $O(mn)$ time. Denote the resulting distance vector by $\mathbf{y}^* = (y_1^*, y_2^*, \ldots, y_n^*)$.

If y_n^* is an integer, then $\mathbf{y}^*$ is an optimal integral solution. To see this note that when we substitute $\lambda = y_n^*$ we obtain a right-hand side that is an integral vector and the constraint matrix is totally unimodular, and therefore the resulting solution is integral. This solution is obtained by the Bellman–Ford algorithm, and therefore it equals $\mathbf{y}^*$. This solution has a value y_n^* that is the optimal value.

Otherwise, $\mathbf{y}^*$ is an optimal fractional solution which is an optimal solution for the linear programming relaxation of the problem. Reference [BOR80] proved for the unweighted case (as we have) and pure covering problem, that an optimal integral solution is obtained from an optimal fractional solution by rounding-up all the elements of the solution vector $\mathbf{y}$. Using the relation between pure covering problems and pure packing problems, the optimal solution for our problem is obtained by rounding-down $\mathbf{y}^*$; i.e., the optimal solution is $(\lfloor y_1^* \rfloor, \lfloor y_2^* \rfloor, \ldots, \lfloor y_n^* \rfloor)$. The total time complexity of the algorithm is therefore dominated by the complexity of finding $F(\lambda)$, $O(mn + n^2 \log n)$.

THEOREM 9.2. *Problem (Circular) where all constraints are packing constraints that can be solved in $O(mn + n^2 \log n)$ time.*

COROLLARY 9.2. *Problem (Circular) where all constraints are covering constraints that can be solved in $O(mn + n^2 \log n)$ time.*

Proof. By Theorems 7.1 and 9.2. □

10. An O(n min{mn, n² log n + m log² n}) time algorithm for the mixed (Circular) problem. In this section we show an $O(n \min\{mn, n^2 \log n + m \log^2 n\})$-time algorithm that solves problem mixed (Circular). We first present an $O(n^3 \log n + mn \log^2 n)$-time algorithm based on the parametric method, and then an alternative $O(mn^2)$-time algorithm. Together these yield the stated running time.

10.1. An O(n³ log n + mn log² n) time algorithm for mixed (Circular). Consider a covering constraint with circular 1's $\sum_{j=1}^{q'} x_j + \sum_{j=p'+1}^{n} x_j \geq b_{q'p'}$. Let y_n^* be the value of $\sum_{j=1}^{n} x_j$ in an optimal solution. Suppose we knew what y_n^* is, then this constraint can be written as

$$\sum_{j=q'+1}^{p'} x_j \leq y_n^* - b_{q'p'}.$$

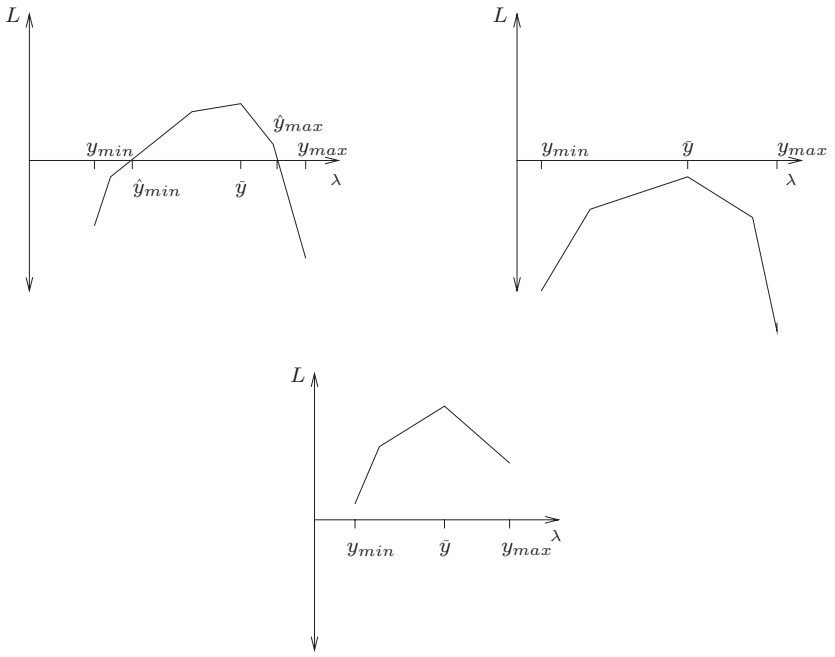


FIG. 4. *Examples of* $L(\lambda)$.

Similarly, a packing constraint with circular 1's $\sum_{j=1}^{q'} x_j + \sum_{j=p'+1}^{n} x_j \leq b_{q'p'}$ can be written as

$$\sum_{j=q'+1}^{p'} x_j \geq y_n^* - b_{q'p'}.$$

Thus the problem becomes a parameterized instance of (Consec) with mixed packing and covering constraints.

Let the right-hand side of a circular constraint (either covering or packing) be denoted by $\bar{b}_{q'p'} = y_n^* - b_{q'p'}$ and for a consecutive constraint $\bar{b}_{q'p'} = b_{q'p'}$. The right-hand sides in this formulation are linear functions of y_n^* with slope 1 or 0. We now convert the covering constraints (and the nonnegativity constraints) into packing constraints, as in section 8, by multiplying these constraints by -1. This result in a formulation of the shortest path from 0 to n where the arcs' lengths are linear functions of the parameter $\lambda = y_n^*$ with slopes -1 or 0 or 1. Note that the YTO-algorithm [YTO91] cannot be applied to solve this parametric shortest path problem as their algorithm is suitable only to the cases where the slopes of λ are either 0 or -1 (by changing the definition of λ it can work also for cases where the slopes are either 0 or $+1$ but not for networks where arcs' lengths are linear functions of the parameter with slopes in $\{-1, 0, +1\}$).

Let the length of a shortest path from 0 to n as a function of $\lambda = y_n$ be as before, $F(\lambda)$. This function is piecewise linear with integer slopes in the range $[-n+1, n-1]$. Since it is a lower envelope of the shortest paths functions of λ, $F(\lambda)$ is a concave function. $F(\lambda)$ is defined for an interval $I = [y_{min}, y_{max}]$, where for $\lambda \in I$ the network does not contain a negative length cycle. In I we define the function $L(\lambda) = F(\lambda) - \lambda$

which is also a piecewise linear concave function with integer slopes in the range $[-n, n-2]$; Figure 4 shows typical cases of $L(\lambda)$.

THEOREM 10.1. *For any $\lambda \in I$, $L(\lambda) \geq 0$ if and only if there is a feasible (fractional) solution vector $\mathbf{y}(\lambda)$ to problem (Circular) whose objective value is λ.*

Proof. Let G_λ be the network where the arcs lengths are set to the parameter value λ. The constraints in (Circular) for $\lambda \in I$ are equivalent to requiring for each arc (u, v) in G_λ that $y_v - y_u$ is at most the length of (u, v).

Assume first that $L(\lambda) \geq 0$. We add a constraint of the form $y_n \leq \lambda$, and find a feasible solution to this augmented set of constraints. Let G'_λ be the corresponding network resulting from adding to G_λ an arc from 0 to n whose length is λ.

Let $\mathbf{y}(\lambda)$ be the shortest paths vector in G'_λ. Then, $y(\lambda)_n \leq \lambda$ because the new arc is a possible path from 0 to n. However, since $L(\lambda) \geq 0$, the length of the shortest path from 0 to n in G_λ is at least λ. Therefore, the length of the shortest path from 0 to n in G'_λ is exactly λ, and $y(\lambda)_n = \lambda$. Thus $\mathbf{y}(\lambda)$ is a feasible solution to problem (Circular) whose cost is exactly λ.

Now suppose that there is a feasible solution $\mathbf{y}(\lambda)$ to problem (Circular) whose cost is exactly λ, so $y(\lambda)_n = \lambda$. Therefore, the shortest path in the network G_λ from 0 to n is of length at least λ, and $F(\lambda) \geq \lambda$. Thus $L(\lambda) \geq 0$. ∎

As a concave function, $L(\lambda)$ satisfies the following lemma.

LEMMA 10.1. *$L()$ has a single maximizer $\tilde{y} \in I$ and at most two zeros $z_1 \leq z_2$. Also, $L(\lambda)$ is increasing for $\lambda \in [y_{min}, \tilde{y})$, and $L(\lambda)$ is decreasing for $\lambda \in (\tilde{y}, y_{max}]$.*

If $L(\tilde{y}) \geq 0$, we let $\hat{y}_{min} = z_1$. We let $\hat{y}_{max}$ be the maximum value in I for which $L(\lambda) \geq 0$, $\hat{y}_{max} = \max\{\lambda \in I | L(\lambda) \geq 0\}$.

If $L(\tilde{y}) < 0$, then by Theorem 10.1, problem (Circular) is infeasible. In this case we do not define $\hat{y}_{min}$ and $\hat{y}_{max}$.

If $L(\lambda) > 0$ for all $\lambda \in I$ such that $\lambda \geq \tilde{y}$, then we set $\hat{y}_{max} = y_{max}$. It follows then from Theorem 10.1 that $\hat{y}_{max}$ is well-defined for all feasible instances of (Circular).

Since our objective is $\max y_n$, an optimal solution is $\hat{y}_{max} = \max\{\lambda \in I | L(\lambda) \geq 0\}$. This is derived by first finding a fractional value $\hat{y}_{max}$ that is an optimal fractional solution value of the linear programming relaxation of (Circular), and then generating from it an optimal integral solution.

We now show how to use Cole's method for computing the parametric shortest path from 0 to n in order to find $\hat{y}_{max}$, which is the solution of the linear programming relaxation of (Circular). The use of this algorithm requires the resolution of $O(\log^2 n)$ questions of the type, "is $y^*_n \leq \lambda_c$?". In order to resolve such a comparison we apply the Bellman–Ford algorithm to find the 0 to n shortest path where the arcs lengths are evaluated for $\lambda = \lambda_c$. The time complexity of this algorithm is $O(mn)$. The result obtained from the Bellman–Ford algorithm is either a negative length cycle C in the network or a shortest path P from 0 to n if a negative length cycle does not exist.

We now show how to resolve the comparison using the output of the Bellman–Ford algorithm.

1. If the Bellman–Ford algorithm finds a negative length cycle C, then we compute the linear function $f_C(\lambda)$ that defines its length. Since this is a negative cycle, we know that $f_C(\lambda_c) < 0$. We compute the slope (derivative) of $f_C(\lambda)$ at λ_c, $f'_C(\lambda_c)$.
   - If $f'_C(\lambda_c) \geq 0$, then for every $\lambda \leq \lambda_c$ the length of C remains negative, and therefore if (Circular) is feasible, then $y^*_n > \lambda_c$.
   - If $f'_C(\lambda_c) < 0$, then for every $\lambda \geq \lambda_c$ the length of C remains negative, and therefore if (Circular) is feasible, then $y^*_n < \lambda_c$.

- If $f'_C(\lambda_c) = 0$, then for all λ the length of C remains negative and therefore (Circular) is infeasible (because the maximum of $F(\lambda)$ is attained at λ_c).

2. If the Bellman–Ford algorithm finds a shortest path P from 0 to n, then there are no negative cycles and $\lambda_c \in I$. We compute the linear function $f_P(\lambda)$ for the length of P. Since P is a shortest path for $\lambda = \lambda_c$, there is a sufficiently small interval that contains λ_c such that P is the shortest path for λ in that interval, and $f_P(\lambda) - \lambda$ and $L(\lambda)$ coincide in that interval. The slope of $L(\lambda)$ in this interval is $L'(\lambda) = f'_P(\lambda) - 1$.
   - If $f'_P(\lambda_c) - 1 > 0$, then L is increasing in λ_c, and therefore by Lemma 10.1, $\lambda_c \in [y_{min}, \tilde{y})$. Therefore, if (Circular) is feasible, then $y^*_n \geq \tilde{y}$, and thus $y^*_n > \lambda_c$.
     If (Circular) is infeasible, then $I \neq \emptyset$ as for λ_c the network does not contain negative length cycles so $\lambda_c \in I$. Hence, from Theorem 10.1, $L(\lambda) < 0\ \forall \lambda \in I$. In this case it is correct to resolve the comparison in an arbitrary way as all possibilities will be infeasible. Therefore, it is correct to resolve the comparison with $y^*_n > \lambda_c$. Therefore, if the slope of $f_P(\lambda) - \lambda = L(\lambda)$ is positive at $\lambda = \lambda_c$ we conclude that $y^*_n > \lambda_c$.
   - If $f'_P(\lambda_c) - 1 \leq 0$, then $L(\lambda)$ is nonincreasing at λ_c.
     - If $f_P(\lambda_c) > \lambda_c$, then $L(\lambda_c) > 0$, and therefore $\lambda_c \in [\tilde{y}, \hat{y}_{max})$. Hence $y^*_n > \lambda_c$.
     - If $f_P(\lambda_c) = \lambda_c$, then $L(\lambda_c) = 0$, and therefore $y^*_n = \lambda_c$.
     - If $f_P(\lambda_c) < \lambda_c$, then $L(\lambda_c) < 0$. Hence $\lambda_c \in (\hat{y}_{max}, y_{max}]$, and therefore if (Circular) is feasible, then $y^*_n < \lambda_c$. Even if (Circular) is infeasible we may resolve the comparison as $y^*_n < \lambda_c$ (because in this case we can resolve any comparison in an arbitrary way).

We thus showed a $T(n) = O(mn)$-time algorithm that resolves a single comparison. By Theorem 3.1 it then follows that $\hat{y}_{max}$ is found in $O(n^3 \log n + mn \log^2 n)$ time. This value may be fractional as it is an optimal solution to the linear programming relaxation.

Using the optimal solution $\hat{y}_{max}$, if fractional, we get an optimal integer solution by rounding down the solution cost to $\lfloor \hat{y}_{max} \rfloor$. This is the optimal value of an integral solution, but we still need to find a corresponding integral solution. We substitute the value of $y^*_n = \lfloor \hat{y}_{max} \rfloor$ in the circular constraints: $\sum_{j=q'+1}^{p'} x_j \leq y^*_n - b_{q'p'}$ and $\sum_{j=q'+1}^{p'} x_j \geq y^*_n - b_{q'p'}$. The resulting constraints are exactly the constraints of mixed (Consec) which we solve with the algorithm of section 8. This computation is done in $O(mn)$ time and results in an integer solution. Therefore, we have the following theorem.

THEOREM 10.2. *There is an $O(n^3 \log n + mn \log^2 n)$-time algorithm that solves problem (Circular) with both covering and packing constraints.*

10.2. An $O(mn^2)$ time algorithm for mixed (Circular). We use the following facts obtained in the previous subsection.

- If the value of y^*_n is given, then an optimal solution can be found in $O(mn)$ time using the Bellman–Ford algorithm in a network where the arc lengths depend on y^*_n.
- For a given value of $\lambda = y_n$, one can check in $O(mn)$ time if (Circular) is feasible for this parameter value, and if so obtain a feasible solution of this value.

We next show how to implement the algorithm of the previous subsection for all values of y_n simultaneously in $O(mn^2)$ time. The idea is to construct the entire lower envelope of the shortest path lengths as a function of the parameter value λ. This is similar to the idea used by Young, Tarjan, and Orlin [YTO91] except that their algorithm constructs solutions for a sequence of different values of the parameter one after the other, whereas our algorithm constructs the entire lower envelope simultaneously. Our algorithm is based on implementing the Bellman–Ford algorithm for all values of λ simultaneously with an increase of run time compared to the nonparametric case of factor $O(n)$. We first present the nonparametric Bellman–Ford algorithm for arc costs c_{ij}, and then its adjustment to the parametric case.

Bellman–Ford algorithm:
Input: A graph $G = (\{0, 1, 2, \ldots, n\}, E)$ with arc lengths c_{ij}.
Output: Either a certificate of a negative length cycle, or for each i, the length of a shortest path from 0 to i in G.

Initialization: $u_0^1 = 0$, $u_i^1 = c_{0i}$ $\forall i$.
For $m = 2$ **to** $n + 1$ **do**
For $i = 0$ **to** n **do**
$u_i^m = \min\{u_i^{m-1}, \min_{j \neq i}\{u_j^{m-1} + c_{ji}\}\}$.

Negative cycle detector: If there is i such that $u_i^{n+1} < u_i^n$, then **return** G has a negative length cycle.
Otherwise, **return** $(u_i^n)_{i=1}^n$.

For the parametric problem each arc cost is of the form $\bar{b}_{i,j} = b_{i,j} + c\lambda$ for $c \in \{-1, 0, 1\}$. The length of the shortest path from 0 to i is therefore a concave piecewise linear lower envelope with up to $2n + 1$ linear functions with integer slopes in $[-n, n]$. We retain these functions as an *array* $\mathbf{u}$ of $2n + 1$ entries. Each entry $j \in [-n, n]$ has a linear function $a + j\lambda$ with the least constant value a among all functions of the same slope j, as only the one with the least constant can be on the lower envelope. Note that the lower envelope does not necessarily contain all the functions in the array, but possibly only a strict subset of them. The adjustments made in the parametric version of the Bellman–Ford algorithm are:

- *Initialization* $\mathbf{u}_0^1 = 0$, $\mathbf{u}_i^1 = \bar{b}_{0,i}$ $\forall i$*:* This is implemented in the same time complexity as the initialization step in the Bellman–Ford algorithm.
- *Computing* $\mathbf{u}_j^{m-1} + \bar{b}_{j,i}$*:* This requires computing the piecewise linear function array obtained from another function array by adding a linear function $c\lambda + d$ with $c \in \{-1, 0, 1\}$ and $d = \bar{b}_{j,i}$. This is done by adding d to all the constants of the linear functions of the array and then shift the array functions by one position to the right if $c = 1$, or to the left if $c = -1$, in $O(n)$ time. A faster way to implement this step is by using dynamic trees data structure in $O(\log n)$. This, however, is not going to affect the overall complexity, and thus is not discussed in detail.
- *Minimum of a pair of piecewise linear functions:* This operation is used in the line $\mathbf{u}_i^m = \min\{\mathbf{u}_i^{m-1}, \min_{j \neq i}\{\mathbf{u}_j^{m-1} + \bar{b}_{j,i}\}\}$ of the Bellman–Ford algorithm. For the parametric version this is done by comparing two entries of the same slope in the two arrays and take the one with the lower constant term. If only one of the arrays has a linear function, then that function becomes the one corresponding entry of the minimum array. This operation is implemented in $O(n)$ time.

So the main part of the parametric algorithm has complexity $O(mn^2)$. We now need to address the negative cycle detection. A negative cycle is identified, when the piecewise linear concave lower envelope U_i^{n+1} is strictly below U_i^n for some i.

To find the representation of the lower envelope of the linear functions as a sequence of $O(n)$ breakpoints and the slopes of the lines between them, we scan the array progressing from the largest slope linear function to lower slope functions. Having evaluated the jth lower envelope including the functions with slopes from n to j, we find the intersection of the jth lower envelope with the next linear function of slope $j-1$ or less. If the breakpoint of the intersection of the linear function of slope $j-1$ with the rightmost linear function in the envelope is left of the previously evaluated rightmost breakpoint br, then br and the slope of the line adjacent to br's right are omitted from the jth lower envelope and the intersection step of the rightmost linear function with the linear function of slope $j-1$ or less is repeated. If the intersection is to the right of all previously evaluated breakpoints it is added with the new linear function to the $j-1$st lower envelope. Each step involves finding the intersection of two linear functions in $O(1)$ and is associated either with proceeding to the next iteration, or else in the elimination of a breakpoint previously evaluated. Since there are at most $O(n)$ breakpoints, the total number of operations is $O(n)$.

We thus find the lower envelopes of U_i^n and U_i^{n+1} in $O(n^2)$ operations for all i. Note that $U_i^{n+1}() \leq U_i^n()$. For values of λ such that $U_i^{n+1}(\lambda) < U_i^n(\lambda)$, there is a negative length cycle that contains i. If (Circular) is feasible, then there is an interval $[y_{min}^i, y_{max}^i]$ such that for $y_n \in [y_{min}^i, y_{max}^i]$, the network does not contain a negative length cycle that contains i. If there exists i such that $U_i^{n+1}(\lambda) < U_i^n(\lambda)$ for all λ, then (Circular) is infeasible and we can identify such cases in $O(n)$ time (for each i) by computing the (empty) intersection of U_i^n and U_i^{n+1}. Therefore, the two functions U_i^n and U_i^{n+1} intersect in at most two points $y_{min}^i < y_{max}^i$, and we can compute these points in $O(n)$ time for each i.

Let $[y_{min}, y_{max}] = \bigcap_i [y_{min}^i, y_{max}^i]$. To find this interval of parameter values for which the network does not contain a negative cycle, we initialize the interval $[y_{min}, y_{max}]$ to $[0, \infty)$ and intersect it for each i with the interval $[y_{min}^i, y_{max}^i]$. This requires additional run time not exceeding $O(n^2)$.

Therefore, the total complexity of this procedure is $O(mn^2)$. At termination we have an interval $[y_{min}, y_{max}]$, such that for each integer value in this interval there is a feasible solution to (Circular) with this cost. We set the value of y_n to be $\lfloor y_{max} \rfloor$, and then apply the algorithm for mixed (Consec) problem in $O(mn)$-time complexity using Theorem 8.1. This algorithm outputs the solution with cost $\lfloor y_{max} \rfloor$.

THEOREM 10.3. *There is an $O(mn^2)$-time algorithm that solves problem (Circular) with both covering and packing constraints.*

11. Conclusions. We address here covering and packing problems on circular 1's constraints. We show how to solve such problems efficiently and in strongly polynomial time, thereby improving on the method of solving these problems as a parametric dual of minimum cost network flow.

Appendix: Extensions of Bartholdi, Orlin, and Ratliff.

Extension to mixed problems. In applying the transformation of Veinott and Wagner to the constraint matrix of the problem, let $\left[\begin{array}{c|c} \overline{A_{cover}} & \overline{A_{\cdot,n_{cover}}} \\ \hline \overline{A_{pack}} & \overline{A_{\cdot,n_{pack}}} \end{array}\right]$ be the resulting transformed constraint matrix where $\left[\begin{array}{c} \overline{A_{\cdot,n_{cover}}} \\ \hline \overline{A_{\cdot,n_{pack}}} \end{array}\right]$ is the transformed column that

corresponds to y_n. Let $\begin{bmatrix} b_{cover} \\ b_{pack} \end{bmatrix}$ denote the right-hand side vector. For $\overline{c_i} = \sum_{j=1}^{i} c_j$, $[\overline{c}, \overline{c_n}]$ is the transformed cost coefficient vector where $\overline{c_n}$ is the coefficient of y_n. Let $\overline{y} = [y_1, y_2, \ldots, y_{n-1}]$, then, the transformed problem is

$$\begin{aligned} \min \quad & \overline{cy} + \overline{c_n} y_n \\ \text{subject to:} & \\ \overline{A_{cover}}\overline{y} + \overline{A_{\cdot,n_{cover}}} y_n &\geq b_{cover} \\ \overline{A_{pack}}\overline{y} + \overline{A_{\cdot,n_{pack}}} y_n &\leq b_{pack} \\ \overline{y}, y_n \quad & \text{unrestricted integer.} \end{aligned}$$

For a specified integer value of y_n we get the problem

$$\begin{aligned} P(y_n) &= \overline{c_n} y_n + \min \overline{cy} \\ \text{subject to:} & \\ \overline{A_{cover}}\overline{y} &\geq b_{cover} - \overline{A_{\cdot,n_{cover}}} y_n \\ \overline{A_{pack}}\overline{y} &\leq b_{pack} - \overline{A_{\cdot,n_{pack}}} y_n \\ \overline{y} \quad & \text{unrestricted.} \end{aligned}$$

Although the integrality constraints for $\overline{y}$ are dropped here, it is shown in what follows that it does not affect the algorithm.

Multiplying the packing constraints by -1 we get

$$\begin{aligned} P(y_n) &= \overline{c_n} y_n + \min \overline{cy} \\ \text{subject to:} & \\ \overline{A_{cover}}\overline{y} &\geq b_{cover} - \overline{A_{\cdot,n_{cover}}} y_n \\ (-\overline{A_{pack}})\overline{y} &\geq \overline{A_{\cdot,n_{pack}}} y_n - b_{pack} \\ \overline{y} \quad & \text{unrestricted.} \end{aligned}$$

The dual of this problem is

$$\begin{aligned} D(y_n) = \overline{c_n} y_n &+ \max \lambda_{cover}(b_{cover} - \overline{A_{\cdot,n_{cover}}} y_n) \\ &+ \lambda_{pack}(\overline{A_{\cdot,n_{pack}}} y_n - b_{pack}) \\ \text{subject to:} & \\ \lambda_{cover}\overline{A_{cover}} - \lambda_{pack}\overline{A_{pack}} &= \overline{c} \\ \lambda_{cover}, \lambda_{pack} &\geq 0. \end{aligned}$$

$D(y_n)$ is a MCNF problem, where the arcs' cost are parameterized by a common parameter y_n. Then, as in Lemmas 1.1 and 1.2 in [BOR80], the optimal solution cost is a convex function of y_n and the optimal solution satisfies $y_n^* \leq \sum_i b_i$. One can thus apply binary search using $O(log \sum_i b_i)$ applications of a MCNF algorithm. The total complexity of the resulting algorithm when we use Orlin's [Orl93] algorithm for the MCNF is $O(\log(\sum_i b_i)[n \log n(m + n \log n)])$.

It remains to show that when we find the optimal integer value y_n^*, we can find a feasible integral solution $\overline{y}$ with this cost. We use Orlin's algorithm for solving $D(y_n^*)$. Orlin's algorithm produces both primal solution $(\lambda_{cover}, \lambda_{pack})$ (i.e., a solution to $D(y_n^*)$) and a dual solution $\overline{y}$ (i.e., a solution to $P(y_n^*)$). For an integer value of y_n^*, both the right-hand side and the objective function of $D(y_n^*)$ are integral. Therefore,

Orlin's algorithm produces an integer solution for both $P(y_n^*)$ and $D(y_n^*)$, and the above algorithm produces an integer optimal solution.

A modified strongly polynomial time algorithm. From the previous subsection it follows that problem (Circular) can be solved by solving a parametric MCNF problem where the costs are parameterized by the single common parameter $\lambda = y_n^*$. The use of the parametric method described in section 3 results in an $O(n^4 \log^2 n + n^2 m \log^4 n)$-time algorithm. Thus we show the following theorem.

Theorem. There is an $O(n^4 \log^2 n + n^2 m \log^4 n)$-time algorithm that solves problem (Circular) with a general linear cost function.

Acknowledgement. The authors wish to express their gratitude to A. Tamir for discussing an earlier version of this paper and to anonymous referees whose comments and suggestions improved and simplified the presentation of the results in this paper.

REFERENCES

[AS87] E. M. Arkin and E. B. Silverberg, *Scheduling jobs with fixed start and end times*, Discrete Appl. Math., 18 (1987), pp. 1–8.

[BOR80] J. J. Bartholdi, J. B. Orlin, and H. D. Ratliff, *Cyclic scheduling via integer programs with circular ones*, Oper. Res., 28 (1980), pp. 1074–1085.

[Boo75] K. S. Booth, *PQ-tree Algorithms*, Ph.D. thesis, University of California (Available as UCRL 51953, Lawrence Livermore Labs, Livermore, CA), 1975.

[BL76] K. S. Booth and G. S. Lueker, *Testing for the consecutive ones property, interval graphs, and graph planarity using PQ trees*, J. Comput. Syst. Sci., 13 (1976), pp. 335–379.

[Col87] R. Cole, *Slowing down sorting networks to obtain faster sorting algorithms*, J. Assoc. Comput. Math., 34 (1987), pp. 200–208.

[Dij59] E. W. Dijkstra, *A note on two problems in connexion with graphs*, Numer. Math., 1 (1959), pp. 269–271.

[Gol80] M. C. Golumbic, *Algorithmic Graph Theory and Perfect Graphs*, Academic Press, New York, 1980.

[GT86] A. V. Goldberg and R. E. Tarjan, *A new approach to the maximum-flow problem*, J. Assoc. Comput. Mach., 35 (1988), pp. 921–940.

[HN94] D. S. Hochbaum and J. Naor, *Simple and fast algorithms for linear and integer programs with two variables per inequality*, SIAM J. Comput., 23 (1994), pp. 1179–1192.

[Hoc93] D. S. Hochbaum, *Polynomial Algorithms for Convex Network Optimization*, in Network Optimization Problems: Algorithms, Complexity and Applications, D. Du and P. M. Pardalos eds., World Scientific, London, 1993, pp. 63–92.

[Hoc94] D. S. Hochbaum, *Lower and upper bounds for allocation problems and other nonlinear optimization problems*, Math. Oper. Res., 19 (1994), pp. 390–409.

[HL03] D. S. Hochbaum and A. Levin, *Cyclical scheduling and multi-shift scheduling: complexity and approximation algorithms*, Discrete Optim., accepted, 2006.

[Joh77] D. B. Johnson, *Efficient algorithms for shortest paths in sparse networks*, J. Assoc. Comput. Math., 24 (1977), pp. 1–13.

[KO81] R. M. Karp and J. B. Orlin, *Parametric shortest path algorithms with an application to cyclic staffing*, Discrete Appl. Math., 3 (1981), pp. 37–45.

[Meg83] N. Megiddo, *Applying parallel computation algorithms in the design of serial algorithms*, J. Assoc. Comput. Math., 30 (1983), pp. 852–865.

[Orl93] J. B. Orlin, *A faster strongly polynomial minimum cost flow algorithm*, Oper. Res., 41 (1993), pp. 338–350.

[Seg74] M. Segal, *The operator-scheduling problem: a network-flow approach*, Oper. Res., 22 (1974), pp. 803–824.

[SF02] R. Shah and M. Farach-Colton, *Undiscretized dynamic programming: faster algorithms for facility location and related problems on trees*, Proceedings of the 13th annual ACM-SIAM Symposium on Discrete Algorithms, San Francisco, 2002, pp. 108–115.

[Ta03] A. Tamir, *Improved algorithms for minimum multiple coverage and weighted maximum dispersion problems on trees*, manuscript, 2003.

[Tar86] E. Tardos, *A strongly polynomial algorithm to solve combinatorial linear programs*, Oper. Res., 34 (1986), pp. 250–256.

[VW62] A. F. Veinott and H. M. Wagner, *Optimal capacity scheduling: Parts* I *and* II, Oper. Res., 10 (1962), pp. 518–547.

[YTO91] N. E. Young, R. E. Tarjan, and J. B. Orlin, *Faster parametric shortest path and minimum-balance algorithms*, Networks, 21 (1991), pp. 205–221.

SIAM J. OPTIM.
Vol. 17, No. 2, pp. 331–352
© 2006 Society for Industrial and Applied Mathematics

APPROXIMATION ALGORITHM FOR THE MIXED FRACTIONAL PACKING AND COVERING PROBLEM*

KLAUS JANSEN†

Abstract. We propose an approximation algorithm based on the Lagrangian or price-directive decomposition method to compute an ϵ-approximate solution of the mixed fractional packing and covering problem: find $x \in B$ such that $f(x) \leq (1+\epsilon)a$, $g(x) \geq (1-\epsilon)b$, where $f(x), g(x)$ are vectors with M nonnegative convex and concave functions, a and b are M-dimensional nonnegative vectors, and B is a convex set that can be queried by an optimization oracle. We propose an algorithm that needs only $O(M\epsilon^{-2}\ln(M\epsilon^{-1}))$ iterations or calls to the oracle. The main contribution is that the algorithm solves the general mixed fractional packing and covering problem (in contrast to pure fractional packing and covering problems and to the special mixed packing and covering problem with $B = \mathbb{R}^N_+$) and runs in time independent of the so-called width of the problem.

Key words. linear and convex optimization, approximation algorithms

AMS subject classifications. 90C05, 90C25, 68Q25

DOI. 10.1137/030601570

1. Introduction.

We study mixed fractional packing and covering problems (MPC_ϵ) of the following form: Given a vector $f : B \to \mathbb{R}^M_+$ of M nonnegative continuous convex functions and a vector $g : B \to \mathbb{R}^M_+$ of M nonnegative continuous concave functions, two M-dimensional nonnegative vectors a, b, a nonempty convex compact set B, and a relative tolerance $\epsilon \in (0, 1)$, find an approximately feasible vector $x \in B$ such that $f(x) \leq (1+\epsilon)a$ and $g(x) \geq (1-\epsilon)b$ or find a proof that no vector is feasible (that satisfies $x \in B$, $f(x) \leq a$, and $g(x) \geq b$). Without loss of generality we may assume that a and b are equal to the vector e of all ones.

The fractional packing problem with convex constraints, i.e., to find $x \in B$ such that $f(x) \leq (1+\epsilon)a$, is solved in [3, 4, 7] by the Lagrangian decomposition method in $O(M(\epsilon^{-2} + \ln M))$ iterations where each iteration requires a call to an approximate block solver $ABS(p, t)$ of the form: find $\hat{x} \in B$ such that $p^T f(\hat{x}) \leq (1+t)\Lambda(p)$ where $\Lambda(p) = \min_{x \in B} p^T f(x)$. Furthermore, Grigoriadis et al. [5] also proposed an approximation algorithm for the fractional covering problem with concave constraints, i.e., to find $x \in B$ such that $g(x) \geq (1-\epsilon)b$, within $O(M(\epsilon^{-2}+\ln M))$ iterations where each iteration requires here a call to an approximate block solver $ABS(q, t)$ of the form: find $\hat{x} \in B$ such that $q^T g(\hat{x}) \geq (1-t)\Lambda(q)$, where $\Lambda(q) = \max_{x \in B} q^T g(x)$. Both algorithms also solve the corresponding min-max and max-min optimization variants within the same number of iterations. Furthermore, the algorithms can be generalized to the case where the block solver has arbitrary approximation ratio [6, 7, 8].

*Received by the editors October 31, 2003; accepted for publication (in revised form) December 22, 2005; published electronically May 19, 2006. A preliminary version of the paper has appeared in the *Proceedings of the 3rd IFIP Conference on Theoretical Computer Science* (TCS 2004), Toulouse, France, 2004, pp. 223–236.

http://www.siam.org/journals/siopt/17-2/60157.html

†Institut für Informatik und Praktische Mathematik, Universität zu Kiel, Kiel, Germany (kj@informatik.uni-kiel.de). The author's research was supported in part by EU Thematic Network APPOL, Approximation and Online Algorithms, IST-2001-30012, by EU Project CRESCCO, Critical Resource Sharing for Cooperation in Complex Systems, IST-2001-33135, and by DFG Project, Entwicklung und Analyse von Approximativen Algorithmen für Gemischte und Verallgemeinerte Packungs- und Überdeckungsprobleme, JA 612/10-1. Part of this work was done while the author was visiting the Department of Computer Science at ETH Zürich.

Further interesting algorithms for the fractional packing and fractional covering problem with linear constraints were developed by Plotkin, Shmoys, and Tardos [10] and Young [11]. These algorithms have a running time that depends linearly on the width—an unbounded function of the input instance. Several relatively complicated techniques were proposed to reduce this dependence. Garg and Könemann [2] and Könemann [9] described a nice algorithm for the fractional packing problem with linear constraints that needs only $O(M\epsilon^{-2}\ln M)$ iterations. On the other hand, the algorithm by Grigoriadis et al. [5] is the only known algorithm that solves the fractional covering problem with a number of iterations independently of the width.

For the mixed packing and covering problem (with linear constraints), Plotkin, Shmoys, and Tardos [10] also proposed approximation algorithms where the running time depends on the width. Young [12] described an approximation algorithm for a special mixed packing and covering problem with linear constraints and special convex set $B = \mathbb{R}^N_+$. The algorithm has a running time of $O(M^2\epsilon^{-2}\ln M)$. Recently, Fleischer [1] gave an approximation scheme for the optimization variant (minimizing $c^T x$ such that $Cx \geq b$, $Px \leq a$, and $x \geq 0$ where a, b, and c are nonnegative integer vectors and P and C are nonnegative integer matrices). Young [12] posed the following interesting open problem: find an efficient width-independent Lagrangian-relaxation algorithm for the abstract mixed packing and covering problem, find $x \in B$ such that $Px \leq (1+\epsilon)a$, $Cx \geq (1-\epsilon)b$, where P, C are nonnegative matrices, a, b are nonnegative vectors, and B is a polytope that can be queried by an optimization oracle (given a vector c, return $x \in B$ minimizing $c^T x$) or some other suitable oracle.

New result. Our contribution here is to present an efficient width-independent Lagrangian-relaxation algorithm for the mixed packing and covering problem that uses a suitable optimization oracle of the form: given two vectors c, d, return $x \in B$, $c^T x \geq 1$, minimizing $d^T x$. Interestingly, it is also sufficient to use a feasibility oracle of the form: given two vectors c, d, return $x \in B$ such that $c^T x \geq 1$ and $d^T x \leq 1$. This solves the open problem by Young [12]. Interestingly, our algorithm also works for a more general problem with a convex set B and nonnegative convex packing and concave covering constraints.

The algorithm uses a variant of the Lagrangian or price-directive decomposition method. This is an iterative strategy that solves (MPC_ϵ) by computing a sequence of triples (p, q, x) as follows. A coordinator uses the current vector $x \in B$ to compute two price vectors $p = p(x) \in \mathbb{R}^M_+$ and $q = q(x) \in \mathbb{R}^M_+$ with $\sum_{m=1}^M p_m + q_m = 1$. Then the coordinator calls an optimization oracle to compute a solution $\hat{x} \in B$ of the block problem (BP)

$$\Lambda(p,q) = \min\left\{ p^T f(y) | y \in B, q^T g(y) \geq \sum_{m=1}^M q_m \right\},$$

and makes a move from x to $(1-\tau)x + \tau\hat{x}$ with an appropriate step length $\tau \in (0,1)$. Such an iteration is called a coordination step. For our algorithm, we require only an approximate block solver (ABS) that solves the underlying block problem to a given relative tolerance $t \in (0,1)$:

$$\begin{aligned} ABS(p,q,t): \qquad \text{compute} \qquad & \hat{x} = \hat{x}(p,q) \in B \text{ such that} \\ & p^T f(\hat{x}) \leq (1+t)\Lambda(p,q), \\ & q^T g(\hat{x}) \geq \tfrac{1}{1+t} \sum_{m=1}^M q_m. \end{aligned}$$

Our main result is the following theorem.

THEOREM 1.1. *There is an approximation algorithm that for any given accuracy* $\epsilon \in (0,1)$ *solves the mixed fractional packing and covering problem* (MPC_ϵ) *within*

$$N = O(M\epsilon^{-2}\ln(M\epsilon^{-1}))$$

iterations or coordination steps, each of which requires a call to $ABS(p,q,\Theta(\epsilon))$ *and a coordination overhead of* $O(M\ln(M\epsilon^{-1}))$ *arithmetic operations.*

Alternatively, instead of using the approximate block solver, an approximate feasibility oracle of the form compute $\hat{x} \in B$ such that $p^T f(\hat{x}) \le (1+t)\sum_{m=1}^M p_m$ and $q^T g(\hat{x}) \ge \frac{1}{1+t}\sum_{m=1}^M q_m$ is also sufficient.

Main ideas. If the mixed packing and covering problem has a solution, then there is a vector $y \in B$ with $f(y) \le e$ and $g(y) \ge e$. This vector satisfies $q^T g(y) = \sum_{m=1}^M q_m g_m(y) \ge \sum_{m=1}^M q_m$ and $p^T f(y) = \sum_{m=1}^M p_m f_m(y) \le \sum_{m=1}^M p_m$. This implies that the block problem has a solution of value at most $\sum_{m=1}^M p_m$. Furthermore if there is a feasible solution, then the objective value $\Lambda(p,q) \le \sum_{m=1}^M p_m$ and $p^T f(\hat{x}) \le (1+t)\sum_{m=1}^M p_m$. In other words, if $p^T f(\hat{x}) > (1+t)\sum_{m=1}^M p_m$, then we can conclude that there is no solution of the mixed packing and covering problem.

Suppose now that there is a feasible solution of our mixed packing and covering problem. For a given vector $x \in B$, the objective value can be defined by

$$\lambda(x) = \max\{f_1(x), \ldots, f_M(x), 1/g_1(x), \ldots, 1/g_M(x)\}.$$

If $g_m(x) = 0$ for one component $m \in \{1,\ldots,M\}$, then we define $\lambda(x) = \infty$.

One of the main ideas is to combine two different potential functions that were proposed for pure fractional packing and covering problems [4, 5]. We associate here with the packing and covering constraints $f(x) \le \lambda e$ and $g(x) \ge (1/\lambda)e$ the following potential function:

$$\begin{aligned}\Phi'_t(\theta,x) &= (2+t)\ln\theta - \tfrac{t}{M}\sum_{m=1}^M \ln(\theta - f_m(x)) - \tfrac{t}{M}\sum_{m=1}^M \ln(g_m(x)\theta - 1)\\ &= 2\ln\theta - \tfrac{t}{M}\sum_{m=1}^M \ln(\theta - f_m(x)) - \tfrac{t}{M}\sum_{m=1}^M \ln(g_m(x) - \tfrac{1}{\theta}),\end{aligned}$$

where $\theta \in \mathbb{R}_+$ and $t > 0$ is a tolerance that depends on ϵ and is used in the approximate block solver. The function Φ' can be extremely small, since there is no upper bound on the function values $g_m(x)$. Let A be a nonempty subset of $\mathcal{M} = \{1,\ldots,M\}$. To control the values of the covering functions $g_m(x)$ and to have a lower bound for the potential function, we eliminate functions g_m (and the corresponding index in A) when the function value $g_m(x)$ is larger than a prespecified threshold value T and modify the potential function. Let $A(x)$ denote the index set corresponding to a given vector $x \in B$. Then the modified potential function has the form

$$\begin{aligned}\Phi_t(\theta,x,A(x)) &= 2\ln\theta - \tfrac{t}{M}\sum_{m=1}^M \ln(\theta - f_m(x)) - \tfrac{t}{M}\sum_{m\in A(x)} \ln(g_m(x) - \tfrac{1}{\theta})\\ &\quad - \tfrac{t}{M}\sum_{m\notin A(x)} \ln(T).\end{aligned}$$

The potential function Φ_t has a unique minimum $\theta_{A(x)}(x)$ that approximates the objective value $\lambda_{A(x)}(x) = \max(\max_{m\in\mathcal{M}} f_m(x), \max_{m\in A(x)} 1/g_m(x))$. This potential

function Φ_t and the minimizer $\theta_{A(x)}(x)$ are used to define the price vectors $p = p(x)$ and $q = q(x)$ for the current vector $x \in B$ and to optimize in the correct direction. Another important parameter for the convergence of the algorithm is the reduced potential value $\phi_t(x, A(x)) = \Phi_t(\theta_{A(x)}(x), x, A(x))$ for $x \in B$ and $A(x) \subset \{1, \ldots, M\}$ (see the discussion below). Since we cannot control the values of eliminated functions g_m for $m \notin A(x)$ (after the elimination), at the end of each phase s we take a convex combination over different computed vectors.

The step length τ is defined carefully in dependence on the minimizer $\theta_{A(x)}(x)$ of the potential function. In the general case, the coordinator moves from solution x to $(1-\tau)x + \tau\hat{x}$ and sets the index set $A(x') = \{m \in A(x) | g_m(x') < T\}$. In the case where $\max_{m \in A} g_m(x)(1-\tau) + g_m(\hat{x})\tau > T$, we reduce the step length from τ to $\bar{\tau}$ and use as next vector $x' = (1-\bar{\tau})x + \bar{\tau}\hat{x}$. This is important for the convergence analysis.

Our algorithm computes solutions within different phases. Starting with an initial solution $x^{(0)}$ with objective value $\lambda(x^{(0)}) = O(M)$, we compute in phase s a solution $x^{(s)}$ with objective value $\lambda(x^{(s)}) \leq 1/(1-\epsilon_s)$, where $\epsilon_1 = 1/2$ and $\epsilon_s = \epsilon_{s-1}/2$ for $s \geq 2$. In the potential function, the parameters t and T are replaced by parameters t_s and $T(s)$ (that depend on phase s), respectively. We stepwise decrease the objective values until $\epsilon_s \leq \epsilon/2$. The solution $x^{(s)}$ in the last phase satisfies $f_m(x^{(s)}) \leq 1/(1-\epsilon/2) \leq 1+\epsilon$ and $1/g_m(x^{(s)}) \leq 1/(1-\epsilon/2)$ or, equivalently, $g_m(x^{(s)}) \geq 1-\epsilon/2 > 1-\epsilon$.

The main argument in the proof of the convergence is to show that the reduced potential values $\phi_t(x, A(x))$ (that approximate the objective values $\lambda_{A(x)}(x)$) are monotone decreasing. If we do not eliminate a covering function g_m (i.e., $A(x) = A(x')$), then the difference

$$\phi_t(x, A(x)) - \phi_t(x', A(x')) \geq t^3/(4M).$$

In the case where we eliminate a covering function g_m (i.e., $A(x) \neq A(x')$), the difference

$$\phi_t(x, A(x)) - \phi_t(x', A(x')) \geq 0.$$

To show this inequality, we use the fact that the step length τ is reduced to $\bar{\tau}$ when $\max_{m \in A} g_m(x)(1-\tau) + g_m(\hat{x})\tau > T$. This enables us to prove that

$$\phi_t(y, A(y)) - \phi_t(\bar{y}, A(\bar{y})) \geq \Theta(t^3/M)(N_s - M - 1),$$

where N_s is the number of iterations, y is the initial solution, and $\bar{y}$ is the solution after $N_s - 1$ iterations in phase s. Using the modification of the potential function and the implied lower bound for $\phi_t(\bar{y}, A(\bar{y}))$, the difference $\phi_t(y, A(y)) - \phi_t(\bar{y}, A(\bar{y}))$ is at most $O(t \ln(M/t))$ (where $t = t_s = \Theta(\epsilon_s)$ depends on phase s). This gives us an upper bound for the number N_s of iterations in phase s and the total number N of iterations.

The paper is organized as follows. In section 2 we show some properties of the used potential function and define the price vectors $p(x)$ and $q(x)$. In section 3 we describe our approximation algorithm. After giving the algorithm we describe the main techniques in detail. We show how to compute an initial solution in subsection 3.1. In subsection 3.2 we describe the stopping rules and prove some properties. In subsections 3.3 and 3.4 we define the step lengths depending on the cases described above, and we prove that the decrease in the reduced potential is sufficiently large. After that in subsection 3.5 we show how to get a good solution for all constraints using the convex combination. Finally in section 4 we determine the total number of iterations and show how to approximate the minimizer of the potential function.

2. Potential function and price vectors. Let A be a nonempty subset of $\mathcal{M} = \{1, \ldots, M\}$. During a phase, we eliminate a concave function g_m (and the corresponding index in A) when the function value $g_m(x) \geq T$. Let $A(x)$ denote the index set corresponding to a given vector $x \in B$.

2.1. Potential function. The used potential function has the form

$$\begin{aligned}\Phi_t(\theta, x, A(x)) \;=\; & 2\ln\theta - \tfrac{t}{M}\sum_{m=1}^{M}\ln(\theta - f_m(x)) - \tfrac{t}{M}\sum_{m\in A(x)}\ln(g_m(x) - \tfrac{1}{\theta}) \\ & -\tfrac{t}{M}\sum_{m\notin A(x)}\ln(T).\end{aligned}$$

For simplicity we use $A = A(x)$ (if the dependence is clear). The potential function Φ_t is well defined for $\lambda_A(x) < \theta < \infty$, where

$$\lambda_A(x) = \max\left(\max_{1\leq m\leq M} f_m(x), \max_{m\in A}\frac{1}{g_m(x)}\right).$$

If $g_m(x) = 0$ for at least one index $m \in A$, then we define $\lambda_A(x) = \infty$. Furthermore, Φ_t has the barrier property (i.e., $\Phi_t(\theta, x, A) \to \infty$ for $\theta \to \infty$ and for $\theta \to \lambda_A(x)$). We define the reduced potential function $\phi_t(x, A)$ as the minimum value $\Phi_t(\theta, x, A)$ over $\theta \in (\lambda_A(x), \infty)$ for a given $x \in B$. The minimizer $\theta_A(x)$ can be determined from the first-order optimality condition:

$$\frac{t\theta}{M}\sum_{m=1}^{M}\frac{1}{\theta - f_m(x)} + \frac{t}{M\theta}\sum_{m\in A}\frac{1}{g_m(x) - 1/\theta} = 2. \tag{2.1}$$

Consider the function $h(\theta) = \frac{t}{M}(\sum_{m=1}^{M}\frac{\theta}{\theta - f_m(x)} + \frac{1}{\theta}\sum_{m\in A}\frac{1}{g_m(x)-1/\theta})$. Notice that $h(\theta) \to \infty$ for $\theta \to \lambda_A(x)$ and $h(\theta) \to t < 1$ for $\theta \to \infty$. Since $\frac{\theta}{\theta - f_m(x)}$ and $\frac{1}{g_m(x)\theta - 1}$ are decreasing in θ, the function $h(\theta)$ is also decreasing for $\theta \in (\lambda_A(x), \infty)$. Therefore, we have a unique minimum $\theta_A(x)$. The implicit function $\theta_A(x)$ approximates $\lambda_A(x)$. This is important for the further analysis.

LEMMA 2.1.

$$\frac{\theta_A(x)}{(1 + t/(2M))} \geq \lambda_A(x) \geq \theta_A(x)\left(1 - \frac{t}{2} - \frac{t|A|}{2M}\right) \geq \theta_A(x)(1 - t).$$

Proof. First, we consider a function value $f_m(x)$. Since $f_m(x) \leq \lambda_A(x)$, we have $\theta_A(x) - f_m(x) \geq \theta_A(x) - \lambda_A(x)$. This implies

$$\frac{\theta_A(x)}{\theta_A(x) - f_m(x)} \leq \frac{\theta_A(x)}{\theta_A(x) - \lambda_A(x)}.$$

Furthermore, $g_m(x) \geq 1/\lambda_A(x)$ for each $m \in A$. This gives

$$g_m(x) - \frac{1}{\theta_A(x)} \geq \frac{\theta_A(x) - \lambda_A(x)}{\lambda_A(x)\theta_A(x)}$$

or, equivalently,

$$\frac{\frac{1}{\theta_A(x)}}{g_m(x) - \frac{1}{\theta_A(x)}} \leq \frac{\lambda_A(x)}{\theta_A(x) - \lambda_A(x)} < \frac{\theta_A(x)}{\theta_A(x) - \lambda_A(x)}.$$

Combining both inequalities we obtain

$$\begin{aligned} 1 &= \tfrac{t}{2M}\sum_{m=1}^{M}\tfrac{\theta_A(x)}{\theta_A(x)-f_m(x)} + \tfrac{t}{2M}\sum_{m\in A}\tfrac{1/\theta_A(x)}{g_m(x)-1/\theta_A(x)} \\ &\leq \tfrac{t}{2M}(M+|A|)\tfrac{\theta_A(x)}{\theta_A(x)-\lambda_A(x)}. \end{aligned}$$

This implies now

$$\theta_A(x) - \lambda_A(x) \leq \frac{t(M+|A|)}{2M}\theta_A(x)$$

or

$$\lambda_A(x) \geq \theta_A(x)\left(1 - \frac{t}{2} - \frac{t|A|}{2M}\right) \geq \theta_A(x)(1-t).$$

On the other hand, (using the definition of $\lambda_A(x)$), there is an index $m \in \{1,\ldots,M\}$ with $\lambda_A(x) = f_m(x)$ or an index $m \in A$ with $\lambda_A(x) = 1/g_m(x)$.

Case 1. $\lambda_A(x) = f_m(x)$. In this case $\frac{t\theta_A(x)}{2M(\theta_A(x)-\lambda_A(x))} \leq 1$, which is equivalent to

$$\lambda_A(x) \leq \theta_A(x)\left(1 - \frac{t}{2M}\right).$$

Case 2. $\lambda_A(x) = 1/g_m(x)$. Here we have $\frac{t/\theta_A(x)}{2M(1/\lambda_A(x)-1/\theta_A(x))} \leq 1$. This implies

$$\lambda_A(x) \leq \theta_A(x)/(1+\tfrac{t}{2M}).$$

Notice that $(1-t/(2M)) \leq 1/(1+t/(2M))$ for any $t \geq 0$. Therefore, $\lambda_A(x)$ can be bounded in both cases by $\leq \theta_A(x)/(1+t/(2M))$. □

Lemma 2.1 shows that the value $\theta_A(x)$ approximates the objective value $\lambda_A(x)$ for small t. Interestingly, the reduced potential function $\phi_t(x,A)$ also can be bounded in terms of $\theta_A(x)$.

LEMMA 2.2. *If $g_m(x) \leq T$ for each $m \in A$, then*

$$\phi_t(x,A) \geq (2-t)\ln\theta_A(x) - t\ln T.$$

Furthermore, if $T > 1/\lambda_A(x)$, then

$$\phi_t(x,A) \leq 2\ln\theta_A(x) + 2t\ln\left(\frac{2M}{t}\right) + t\ln\left(1+\frac{t}{2M}\right).$$

Proof.

$$\begin{aligned} 2\ln\theta_A(x) &= \phi_t(x,A) + \tfrac{t}{M}\sum_{m=1}^{M}\ln(\theta_A(x)-f_m(x)) + \tfrac{t}{M}\sum_{m\in A}\ln(g_m(x)-1/\theta_A(x)) \\ &\quad + \tfrac{t}{M}\sum_{m\notin A}\ln(T) \\ &\leq \phi_t(x,A) + \tfrac{t}{M}\sum_{m=1}^{M}\ln\theta_A(x) + \tfrac{t}{M}\sum_{m=1}^{M}\ln(T) \\ &= \phi_t(x,A) + t\ln\theta_A(x) + t\ln(T). \end{aligned}$$

This implies the lower bound. On the other hand, $f_m(x) \le \lambda_A(x)$ for each $m \in \{1,\dots,M\}$ and $g_m(x) \ge 1/\lambda_A(x)$ for each $m \in A$. This gives $2\ln\theta_A(x) \ge \phi_t(x,A) + \frac{t}{M}\sum_{m=1}^{M}\ln(\theta_A(x)-\lambda_A(x)) + \frac{t}{M}\sum_{m\in A}\ln(1/\lambda_A(x)-1/\theta_A(x)) + \frac{t}{M}\sum_{m\notin A}\ln(T)$. For $T > 1/\lambda_A(x)$, this sum is at least $\phi_t(x,A)+2t\ln(\theta_A(x)-\lambda_A(x))+t\ln(1/(\lambda_A(x)\theta_A(x)))$.

Now we use the upper bound $\lambda_A(x) \le \theta_A(x)/(1+t/(2M))$ or, equivalently, $\theta_A(x)-\lambda_A(x) \ge \theta_A(x)(1-1/(1+t/(2M))) = \frac{t/(2M)\theta_A(x)}{1+t/(2M)}$. In addition $1/(\lambda_A(x)\theta_A(x)) \ge (1+t/(2M))/\theta_A(x)^2$. As a consequence we get

$$2\ln\theta_A(x) \ge \phi_t(x,A) + 2t\ln\left(\frac{t}{2M}\cdot\frac{\theta_A(x)}{1+t/(2M)}\right) + t\ln\left(\frac{1+t/(2M)}{\theta_A(x)^2}\right).$$

Using $\ln(\frac{1+t/(2M)}{\theta_A(x)^2}) = \ln(1+t/(2M)) - 2\ln\theta_A(x)$,

$$\begin{aligned} 2\ln\theta_A(x) &\ge \phi_t(x,A) + 2t\ln\left(\tfrac{t}{2M}\cdot\tfrac{1}{1+t/(2M)}\right) + t\ln\left(1+\tfrac{t}{2M}\right) \\ &= \phi_t(x,A) + 2t\ln\left(\tfrac{t}{2M}\right) - t\ln\left(1+\tfrac{t}{2M}\right). \end{aligned}$$

This gives the desired upper bound

$$\phi_t(A,x) \le 2\ln\theta_A(x) + 2t\ln\left(\frac{2M}{t}\right) + t\ln\left(1+\frac{t}{2M}\right). \qquad \square$$

2.2. Price vectors. Given an $x \in B$ and a subset $A \subset \{1,\dots,M\}$, the price vector $p(x,A)$ is defined by

$$p_m(x,A) = \frac{t}{2M}\frac{\theta_A(x)}{\theta_A(x)-f_m(x)} \tag{2.2}$$

and the price vector $q(x,A)$ is given by

$$q_m(x,A) = \begin{cases} \frac{t}{2M}\frac{1}{g_m(x)\theta_A(x)-1}, & m \in A, \\ 0 & \text{otherwise.} \end{cases} \tag{2.3}$$

Using the first-order condition, $\sum_{m=1}^{M} p_m(x,A) + \sum_{m=1}^{M} q_m(x,A) = 1$ and each component $p_m(x,A), q_m(x,A)$ is nonnegative.

LEMMA 2.3.

(a) $p(x,A)^T f(x) = \theta_A(x)(\sum_{m=1}^{M} p_m(x,A) - t/2) \le \theta_A(x)(1-t/2)$,

(b) $q(x,A)^T g(x) = (\sum_{m\in A} q_m(x,A) + t|A|/(2M))/\theta_A(x) \le (\sum_{m\in A} q_m(x,A) + t/2)/\theta_A(x) \le (1+t/2)/\theta_A(x)$.

Proof. The (in-)equalities corresponding to $p(x,A)^T f(x)$ follow from

$$\begin{aligned} p(x,A)^T f(x) &= \tfrac{t\theta_A(x)}{2M}\sum_{m=1}^{M}\tfrac{f_m(x)}{\theta_A(x)-f_m(x)} \\ &= \tfrac{t\theta_A(x)}{2M}\sum_{m=1}^{M}\left(-1+\tfrac{\theta_A(x)}{\theta_A(x)-f_m(x)}\right) \\ &= -\tfrac{t\theta_A(x)}{2} + \theta_A(x)\sum_{m=1}^{M} p_m(x,A) \\ &= \theta_A(x)\left(\sum_{m=1}^{M} p_m(x,A) - \tfrac{t}{2}\right) \\ &\le \theta_A(x)\left(1-\tfrac{t}{2}\right). \end{aligned}$$

For $q(x,A)^T g(x)$ we argue as follows:

$$\begin{aligned}
q(x,A)^T g(x) &= \tfrac{t}{2M} \sum_{m\in A} \tfrac{g_m(x)/\theta_A(x)}{g_m(x)-1/\theta_A(x)} \\
&= \tfrac{t}{2M\theta_A(x)} \sum_{m\in A} \left(1 + \tfrac{1/\theta_A(x)}{g_m(x)-1/\theta_A(x)}\right) \\
&= \tfrac{t|A|}{2M\theta_A(x)} + \tfrac{t}{2M\theta_A(x)} \sum_{m\in A} \tfrac{1/\theta_A(x)}{g_m(x)-1/\theta_A(x)} \\
&= \left(\sum_{m\in A} q_m(x,A) + \tfrac{t|A|}{2M}\right)/\theta_A(x) \\
&\le \left(\sum_{m\in A} q_m(x,A) + t/2\right)/\theta_A(x) \\
&\le (1+t/2)/\theta_A(x). \qquad \square
\end{aligned}$$

Notice that Lemma 2.3(a) implies that $\sum_{m=1}^{M} p_m(x,A) \ge t/2$ (using $p(x,A)^T f(x) \ge 0$). Let $\bar{p}(x) = \sum_{m=1}^{M} p_m(x,A)$ and use $\bar{p} = \bar{p}(x)$ if the dependence on x is clear.

3. Our approximation algorithm. In this section we describe the approximation algorithm for the mixed fractional packing and covering problem. First we suppose that there exists a feasible solution $x \in B$ with $f(x) \le e$ and $g(x) \ge e$. Then the approximation algorithm works as follows:

(1) compute initial solution $x^{(0)}$, $s := 0$, $\epsilon_0 := 1$;
(2) **repeat** {scaling phase}
  (2.1) $s := s+1$; $\epsilon_s := \epsilon_{s-1}/2$; $x := x^{(s-1)}$; $T(s) := 2112(M^3\epsilon_s^{-2})/\lambda_{\mathcal{M}}(x)$;
    $A := \{m \in \{1,\dots,M\} | g_m(x) < T(s)\}$; $finished := false$; $k := 0$;
  (2.2) **if** $A \neq \{1,\dots,M\}$ **then begin** $k := k+1$; $x_k := x$ **end**;
  (2.3) **if** stopping rule 1 is satisfied for x **then** $finished := true$; $y := x$ **end**;
  (2.4) **while** $not(finished)$ **do begin**
    (2.4.1) compute $\theta_A(x)$, $p(x,A)$ and $q(x,A)$;
    (2.4.2) $\hat{x} := ABS(p(x,A), q(x,A), \epsilon_s/32)$;
    (2.4.3) **if** one of the stopping rules is satisfied
      then begin $finished := true$; $y := x$ **end**
      else begin
      (2.4.3.1) compute step length τ and $x' := (1-\tau)x + \tau\hat{x}$;
      (2.4.3.2) **if** $\max_{m\in A} g_m(x)(1-\tau) + g_m(\hat{x})\tau > T(s)$ **then** reduce
        τ to $\bar{\tau}$ and $x' := (1-\bar{\tau})x + \bar{\tau}\hat{x}$;
      (2.4.3.3) $A' := A \setminus \{m | g_m(x') \ge T(s)\}$; $x := x'$;
      (2.4.3.4) **if** $A \neq A'$ **then begin** $k := k+1$; $x_k = x'$;
        $A := A'$ **end**
    end
  end;
  (2.5) compute convex combination of $x_1,\dots,x_k,y$ to get $x^{(s)}$;
  (2.6) **until** $\epsilon_s \le \epsilon/2$ or $\lambda(x^{(s)}) \le 1+\epsilon$;
(3) return$(x^{(s)})$.

The details of the algorithm are described later in this section (how to compute an initial solution, the stopping rules, the choice of the step length, and the reduction of the step length).

For the case where the set of feasible solutions $\{x \in B | f(x) \leq e, g(x) \geq e\}$ is empty, we have to modify the program above. If an inequality $p(x,A)^T f(\hat{x}) > (1+t)\sum p_m(x,A)$ holds for a block solution $\hat{x}$, then we can conclude that there is no feasible solution.

3.1. Initial solution. For each $m \in \{1,\dots,M\}$, we consider the block problem (B_m) with price vectors $p = (1/M,\dots,1/M)$ and $q = e_m$, where e_m is the unit vector with all zero coordinates except for its mth component which is 1:

$$(B_m) \qquad \begin{aligned} \min \quad & \tfrac{1}{M}\sum_{\ell=1}^{M} f_\ell(x), \\ & g_m(x) \geq 1, \\ & x \in B. \end{aligned}$$

If there is a solution $\bar{x} \in B$ with $f(\bar{x}) \leq e$ and $g(\bar{x}) \geq e$, then this solution satisfies $(1/M)\sum_{\ell=1}^{M} f_\ell(\bar{x}) \leq 1$ and $g_m(\bar{x}) \geq 1$. Let $\hat{x}^{[m]} \in B$ be an approximate solution of the block problem (B_m) with tolerance $t = 1/2$, and let $x^{(0)} = (1/M)\sum_{m=1}^{M} \hat{x}^{[m]}$. Using the convexity of B, $x^{(0)} \in B$. If the approximate solution satisfies $(1/M)\sum_{\ell=1}^{M} f_\ell(\hat{x}^{[m]}) > 1+t$, then we can conclude that the solution set of the mixed problem is empty.

LEMMA 3.1. *If there exists a feasible solution of the mixed packing and covering problem, then* $\lambda(x^{(0)}) \leq 3M/2$.

Proof. If there is a feasible solution, then $(1/M)\sum_{\ell=1}^{M} f_\ell(\hat{x}^{[m]}) \leq (1+t) = 3/2$ and $g_m(x^{[m]}) \geq 1/(1+t) = 2/3$ (using the approximate block solver $ABS(p,q,1/2)$ above). Then using the concavity and nonnegativity of g_m,

$$g_m(x^{(0)}) = g_m\left(\frac{1}{M}\sum_{\ell=1}^{M}\hat{x}^{[\ell]}\right) \geq \frac{1}{M}\sum_{\ell=1}^{M} g_m(\hat{x}^{[\ell]}) \geq \frac{1}{M} g_m(\hat{x}^{[m]}) \geq \frac{2}{3M}.$$

Using the convexity and nonnegativity of f_m,

$$f_m(x^{(0)}) = f_m\left(\frac{1}{M}\sum_{\ell=1}^{M}\hat{x}^{[\ell]}\right) \leq \frac{1}{M}\sum_{\ell=1}^{M} f_m(\hat{x}^{[\ell]}) \leq \frac{1}{M}\sum_{m=1}^{M}\sum_{\ell=1}^{M} f_m(\hat{x}^{[\ell]}) \leq \frac{3M}{2}.$$

Combining both inequalities, $\lambda(x^{(0)}) \leq 3M/2$. ∎

3.2. Stopping rules. In the algorithm we stepwise decrease the objective value λ from $3M/2$ to $1/(1-\epsilon/2)$. In the first phase we decrease $3M/2$ to $\epsilon_1 = 1/2$. After that we set $\epsilon_s = \epsilon_{s-1}/2$. The goal in phase s is to obtain a solution $x^{(s)}$ with $\lambda(x^{(s)}) \leq 1/(1-\epsilon_s)$. In order to get such a solution we need at the end of phase s a solution y with $\lambda_A(y) \leq 1/(1-\epsilon_s/4)$.

To obtain the solution and to show the convergence we use three stopping rules. For the first rule we simply test whether

$$\lambda_A(x) \leq 1 + \epsilon_s/4 \tag{3.1}$$

for the current solution x. For this rule we immediately get the following lemma.

LEMMA 3.2. *If $\lambda_A(x) \leq 1+\epsilon_s/4$, then $f_m(x) \leq 1+\epsilon_s/4 \leq 1/(1-\epsilon_s/4)$ for each $m \in \{1,\ldots,M\}$, and $g_m(x) \geq 1/(1+\epsilon_s/4) \geq 1-\epsilon_s/4$ for each $m \in A$.*

For the second rule we define a parameter ν that depends on the current iterate x and the approximate block solution $\hat{x}$ as follows:

$$\nu = \nu(x,\hat{x}) = \frac{(p^T f(x) - p^T f(\hat{x}))/\theta + \theta(q^T g(\hat{x}) - q^T g(x))}{(p^T f(x) + p^T f(\hat{x}))/\theta + \theta(q^T g(\hat{x}) + q^T g(x))}, \tag{3.2}$$

where $p = p(x,A)$, $q = q(x,A)$, and $\theta = \theta_A(x)$. Clearly, $\nu(x,\hat{x}) \leq 1$. The lemma below states that x is an approximate solution of the phase s corresponding to subset A, when ν is bounded by $t_s = \Theta(\epsilon_s)$. Therefore, for the second rule we test whether

$$\nu(x,\hat{x}) \leq t \tag{3.3}$$

for the current solution x and the block solution $\hat{x}$.

LEMMA 3.3. *Suppose $\epsilon_s \in (0,4)$ and $t_s = \epsilon_s/32$. For a given $x \in B$, let p,q be computed by (2.2), (2.3) and $\hat{x}$ computed by $ABS(p,q,t_s)$. If $\nu(x,\hat{x}) \leq t_s$, then $f_m(x) \leq 1+\epsilon_s/4 \leq 1/(1-\epsilon_s/4)$ for each $m \in \{1,\ldots,M\}$ and $g_m(x) \geq 1/(1+\epsilon_s/4) \geq (1-\epsilon_s/4)$ for each $m \in A$.*

Proof. Use (3.2) to rewrite $\nu(x,\hat{x}) \leq t = t_s$ as follows:

$$(p^T f(x) - p^T f(\hat{x}))/\theta + \theta(q^T g(\hat{x}) - q^T g(x)) \leq t[(p^T f(x) + p^T f(\hat{x}))/\theta + \theta(q^T g(\hat{x}) + q^T g(x))].$$

Since $p^T f(x) = \theta(\bar{p} - t/2)$, $q^T g(x) = (1 - \bar{p} + t|A|/(2M))/\theta \leq (1 - \bar{p} + t/2)/\theta$, $p^T f(\hat{x}) \leq (1+t)\bar{p}$, and $q^T g(\hat{x}) \geq (1-\bar{p})/(1+t)$, we obtain $\theta\frac{(1-\bar{p})(1-t)}{(1+t)} \leq (1-\bar{p})(1+t) + t + \bar{p}(t + t/\theta) + \bar{p}(1/\theta - 1) + t\bar{p}/\theta(1+t)$.

In the case $\theta \leq 1+8t$ we can prove the lemma directly as follows. Using $\lambda_A(x) < \theta$ and $t = t_s$, we obtain

$$\lambda_A(x) < (1+8t_s) = (1+\epsilon_s/4)$$

for $t_s = \epsilon_s/32$. This implies $f_m(x) \leq 1+\epsilon_s/4 \leq 1/(1-\epsilon_s/4)$ for each $m \in \{1,\ldots,M\}$ and $g_m(x) \geq 1/(1+\epsilon_s/4) \geq (1-\epsilon_s/4)$ for each $m \in A$.

Now suppose that $\theta > 1+8t \geq 1$. Then using the inequality above we obtain

$$\theta\frac{(1-\bar{p})(1-t)}{(1+t)} \leq (1-\bar{p})(1+t) + t + 3\bar{p}t - \bar{p}\frac{8t}{1+8t}.$$

Now we get (using $t < 1/8$) $3\bar{p}t - \frac{8t}{1+8t}\bar{p} = \bar{p}t\frac{-5+24t}{1+8t} < -\bar{p}t$. This implies

$$\theta\frac{(1-\bar{p})(1-t)}{(1+t)} < (1-\bar{p})(1+t) + t - t\bar{p}.$$

If $\bar{p} = 1$, we obtain with $0 < 0$ a contradiction. If $\bar{p} \neq 1$, then we get

$$\theta \leq \frac{(1+t)^2}{(1-t)} + \frac{(t - t\bar{p})(1+t)}{(1-\bar{p})(1-t)} = \frac{(1+2t)(1+t)}{(1-t)}.$$

The right-hand side can be bounded by $(1+2t)(1+3t) \leq (1+6t)$ for $t \leq 1/6$. But again this is a contradiction. ∎

The third stopping rule is used to control the number of iterations during one phase. Here we use a parameter ω_s that depends on the phase s:

$$\omega_s = \begin{cases} \frac{2}{3M(1-\epsilon_1/4)}, & s = 1, \\ \frac{1-\epsilon_{s-1}}{1-\epsilon_s/4}, & s > 1. \end{cases}$$

Then the third rule is defined by

$$\lambda_A(x) \le \omega_s \lambda_{\mathcal{M}}(x^{(s-1)}), \tag{3.4}$$

where $x^{(s-1)}$ is the solution of phase $s-1$ that satisfies $\lambda(x^{(s-1)}) \le 1/(1-\epsilon_{s-1})$.

LEMMA 3.4. *Let $x^{(s-1)}$ be the initial solution and x be a vector in phase $s \ge 1$ with $\lambda_A(x) \le \omega_s \lambda(x^{(s-1)})$ for $A \subset \mathcal{M}$. If*

$$\lambda_A(x^{(s-1)}) \le \begin{cases} 3M/2 & \text{for } s = 1, \\ 1/(1-\epsilon_{s-1}) & \text{for } s \ge 2, \end{cases}$$

then we get

$$\lambda_A(x) \le 1/(1-\epsilon_s/4).$$

Proof. For $s = 1$ we obtain

$$\lambda_A(x) \le \frac{2}{3M(1-\epsilon_1/4)} \cdot \lambda(x^{(0)}) \le \frac{2}{3M(1-\epsilon_1/4)} \cdot \frac{3M}{2} = \frac{1}{(1-\epsilon_1/4)}$$

and for $s \ge 2$ we get

$$\lambda_A(x) \le \omega_s \lambda(x^{(s-1)}) \le \frac{1-\epsilon_{s-1}}{1-\epsilon_s/4} \cdot \frac{1}{1-\epsilon_{s-1}} = \frac{1}{1-\epsilon_s/4}. \qquad \square$$

Notice that in both cases of the proof ($s = 1$ and $s \ge 2$), $f_m(x) \le 1/(1-\epsilon_s/4)$ for each $m \in \mathcal{M}$, and $g_m(x) \ge (1-\epsilon_s/4)$ for each $m \in A$.

3.3. Choice of the step length. In this subsection we describe the choice of the step length τ. We suppose that we have computed a vector x and an approximate block solution $\hat{x}$ in a phase s such that $\nu(x,\hat{x}) > t$, $p^T f(\hat{x}) \le (1+t)\sum_{m=1}^{M} p_m$, and $q^T g(\hat{x}) \ge \frac{1}{1+t}\sum_{m=1}^{M} q_m$ (where $t = t_s$, $p = p(x, A(x))$, and $q = q(x, A(x))$). Let $x' = (1-\tau)x + \tau\hat{x}$. First we focus on the case where $g_m(x') < T = T(s)$ for each $m \in A(x)$. In this case we do not eliminate a component (i.e., $A(x') = A(x)$). The other case will be discussed later (in some cases we also have to reduce the step length).

For simplification we use $\theta = \theta_{A(x)}(x)$.

Since each function f_m is convex, we get independently of the choice of τ the following inequality:

$$\begin{aligned} \theta - f_m(x') &\ge \theta - (1-\tau)f_m(x) - \tau f_m(\hat{x}) \\ &= (\theta - f_m(x))\left(1 + \tau\tfrac{f_m(x) - f_m(\hat{x})}{\theta - f_m(x)}\right) \\ &= (\theta - f_m(x))\left(1 + \tfrac{2\tau M}{t\theta} p_m (f_m(x) - f_m(\hat{x}))\right) \end{aligned}$$

for each $m \in \mathcal{M}$. Since each function g_m is concave, we obtain

$$\begin{aligned} g_m(x') - \tfrac{1}{\theta} &\ge (1-\tau)g_m(x) + \tau g_m(\hat{x}) - \tfrac{1}{\theta} \\ &= \left(g_m(x) - \tfrac{1}{\theta}\right)\left(1 + \tau\tfrac{g_m(\hat{x}) - g_m(x)}{g_m(x) - 1/\theta}\right) \\ &= \left(g_m(x) - \tfrac{1}{\theta}\right)\left(1 + \tfrac{2\tau M\theta}{t} q_m (g_m(\hat{x}) - g_m(x))\right) \end{aligned}$$

for each $m \in A$. We call a step length τ *feasible* if $\tau \in (0,1)$ and if the following bound is satisfied:

(3.5)
$$\max\left(\max_{m\in\mathcal{M}}\left|\frac{2\tau M}{t\theta}p_m(f_m(x)-f_m(\hat{x}))\right|,\max_{m\in A(x)}\left|\frac{2\tau M\theta}{t}q_m(g_m(\hat{x})-g_m(x))\right|\right)\le 1/2.$$

Suppose from now on that τ is a feasible step length. Later we will specify a step length τ with $\tau \in (0,1)$ to obtain the bound (3.5). Then using $\theta - f_m(x) > 0$ and $g_m(x) - 1/\theta > 0$ we obtain $\theta - f_m(x') > 0$ and $g_m(x') - 1/\theta > 0$ for the next computed vector $x' \in B$. This implies that the objective value $\lambda_{A(x')}(x')$ for the next vector x' is at most $\theta_{A(x)}(x)$, where here $A(x') = A(x)$.

LEMMA 3.5. *For any two consecutive iterations in a phase with computed vectors x, x' and $A(x') = A(x)$ and any feasible step length τ, the difference $\phi_t(x, A(x)) - \phi_t(x', A(x'))$ is at least*

$$\begin{aligned}&+2\tau[(p^Tf(x)-p^Tf(\hat{x}))/\theta+(q^Tg(\hat{x})-q^Tg(x))\theta]\\&-\tfrac{4M\tau^2}{t}[(p^Tf(x)+p^Tf(\hat{x}))/\theta+(q^Tg(\hat{x})+q^Tg(x))\theta]^2,\end{aligned}$$

where $\theta = \theta_{A(x)}(x)$, $p = p(x, A(x))$, and $q = q(x, A(x))$.

Proof. Using the definition of the reduced potential function and $\lambda_{A(x')}(x') \le \theta_{A(x)}(x)$, we get $\phi_t(x', A(x')) = \min_{\lambda_{A(x')}(x')\le\xi} \Phi_t(\xi, x', A(x')) \le \Phi_t(\theta, x', A(x))$. The inequality above implies the following upper bound for $\phi_t(x', A(x'))$:

$$\begin{aligned}&\phi_t(x',A(x'))\le\Phi_t(\theta,x',A(x))\\&=2\ln\theta-\tfrac{t}{M}\sum_{m=1}^{M}\ln(\theta-f_m(x'))-\tfrac{t}{M}\sum_{m\in A(x)}\ln\left(g_m(x')-\tfrac{1}{\theta}\right)-\tfrac{t}{M}\sum_{m\notin A(x)}\ln(T)\\&\le 2\ln\theta-\tfrac{t}{M}\sum_{m=1}^{M}\ln(\theta-f_m(x))-\tfrac{t}{M}\sum_{m=1}^{M}\ln\left(1+\tfrac{2\tau M}{t\theta}p_m(f_m(x)-f_m(\hat{x}))\right)\\&\quad-\tfrac{t}{M}\sum_{m\in A(x)}\ln\left(g_m(x)-\tfrac{1}{\theta}\right)-\tfrac{t}{M}\sum_{m\in A(x)}\ln\left(1+\tfrac{2\tau M\theta}{t}q_m(g_m(\hat{x})-g_m(x))\right)\\&\quad-\tfrac{t}{M}\sum_{m\notin A(x)}\ln(T)\\&=\phi_t(x,A(x))-\tfrac{t}{M}\sum_{m=1}^{M}\ln\left(1+\tfrac{2\tau M}{t\theta}p_m(f_m(x)-f_m(\hat{x}))\right)\\&\quad-\tfrac{t}{M}\sum_{m\in A(x)}\ln\left(1+\tfrac{2\tau M\theta}{t}q_m(g_m(\hat{x})-g_m(x))\right).\end{aligned}$$

Above we have used the lower bounds for $\theta - f_m(x')$ and $g_m(x') - 1/\theta$. The

calculation above shows that the difference $\phi_t(x, A(x)) - \phi_t(x', A(x'))$ is at least

$$\begin{aligned}&\tfrac{t}{M}\sum_{m=1}^{M}\ln\left(1+\tfrac{2\tau M}{t\theta}p_m(f_m(x)-f_m(\hat{x}))\right)\\&\quad+\tfrac{t}{M}\sum_{m\in A(x)}\ln\left(1+\tfrac{2\tau M\theta}{t}q_m(g_m(\hat{x})-g_m(x))\right).\end{aligned}$$

Now we can use the inequality $\ln(1+z) \geq z - z^2$ for $z \geq -1/2$:

$$\ln\left(1+\tfrac{2\tau M}{t\theta}p_m(f_m(x)-f_m(\hat{x}))\right) \geq \tfrac{2\tau M}{t\theta}p_m(f_m(x)-f_m(\hat{x}))-\left(\tfrac{2\tau M}{t\theta}p_m(f_m(x)-f_m(\hat{x}))\right)^2$$

and

$$\ln(1+\tfrac{2\tau M\theta}{t}q_m(g_m(\hat{x})-g_m(x))) \geq \tfrac{2\tau M\theta}{t}q_m(g_m(\hat{x})-g_m(x))-(\tfrac{2\tau M\theta}{t}q_m(g_m(\hat{x})-g_m(x)))^2.$$

Using both inequalities above and $q_m = 0$ for $m \notin A(x)$, the difference $\phi_t(x, A(x)) - \phi_t(x', A(x'))$ is at least

$$\begin{aligned}\geq\;&+\tfrac{2\tau}{\theta}(p^Tf(x)-p^Tf(\hat{x}))+2\tau\theta(q^Tg(\hat{x})-q^Tg(x))\\&-\tfrac{4M\tau^2}{t\theta^2}(p^Tf(x)+p^Tf(\hat{x}))^2+\tfrac{4M\tau^2\theta^2}{t}(q^Tg(\hat{x})+q^Tg(x))^2\\=\;&+2\tau[(p^Tf(x)-p^Tf(\hat{x}))/\theta+(q^Tg(\hat{x})-q^Tg(x))\theta]\\&-\tfrac{4M\tau^2}{t}[(p^Tf(x)+p^Tf(\hat{x}))^2/\theta^2+(q^Tg(\hat{x})+q^Tg(x))^2\theta^2]\\\geq\;&+2\tau[(p^Tf(x)-p^Tf(\hat{x}))/\theta+(q^Tg(\hat{x})-q^Tg(x))\theta]\\&-\tfrac{4M\tau^2}{t}[(p^Tf(x)+p^Tf(\hat{x}))/\theta+(q^Tg(\hat{x})+q^Tg(x))\theta]^2.\qquad\square\end{aligned}$$

In our algorithm we use the following feasible step length.

LEMMA 3.6. *The following step length is feasible for any $t < 1$:*

$$\tau \;=\; \frac{t^2}{4M[(p^Tf(x)+p^Tf(\hat{x}))/\theta+(q^Tg(\hat{x})+q^Tg(x))\theta]}$$

where $\theta = \theta_A(x)$ and $\nu = \nu(x, \hat{x})$.

Proof. Since $p^Tf(x)/\theta+\theta q^Tg(x) = 1-t/2+t|A|/(2M) > 1-t/2 \geq 1/2$, the step length τ is at most $t^2/(2M) < 1$. Furthermore,

$$\begin{aligned}\left|\tfrac{2\tau M}{t\theta}p_m(f_m(x)-f_m(\hat{x}))\right| &\leq \tfrac{2\tau M}{t\theta}(p^Tf(x)+p^Tf(\hat{x}))\\&= \tfrac{2M}{t}\tfrac{t^2}{4M}\frac{(p^Tf(x)+p^Tf(\hat{x}))/\theta}{[(p^Tf(x)+p^Tf(\hat{x}))/\theta+(q^Tg(\hat{x})+q^Tg(x))\theta]}\\&\leq \tfrac{t}{2} < \tfrac{1}{2}\end{aligned}$$

and

$$\begin{aligned}\left|\tfrac{2\tau M\theta}{t}q_m(g_m(\hat{x})-g_m(x))\right| &\leq \tfrac{2\tau M\theta}{t}(q^Tg(\hat{x})+q^Tg(x))\\&= \tfrac{2M}{t}\tfrac{t^2}{4M}\frac{(q^Tg(\hat{x})+q^Tg(x))\theta}{[(p^Tf(x)+p^Tf(\hat{x}))/\theta+(q^Tg(\hat{x})+q^Tg(x))\theta]}\\&\leq \tfrac{t}{2} < \tfrac{1}{2}.\qquad\square\end{aligned}$$

The main goal now is to prove the following result.

THEOREM 3.7. *For any two consecutive iterations in a phase with computed vectors x, x' and $A(x) = A(x')$ we obtain*

$$\phi_t(x, A(x)) - \phi_t(x', A(x')) \geq \frac{t^3}{4M}.$$

Proof. Since the second stopping rule is not satisfied we have $\nu(x, \hat{x}) > t$. This implies the following inequality:

$$\begin{aligned}(3.6)\qquad &[(p^T f(x) - p^T f(\hat{x}))/\theta + (q^T g(\hat{x}) - q^T g(x))\theta] \\ &\quad \geq t[(p^T f(x) + p^T f(\hat{x}))/\theta + (q^T g(\hat{x}) + q^T g(x))\theta].\end{aligned}$$

Then we obtain for the difference $\phi_t(x, A(x)) - \phi_t(x', A(x'))$ of the potential values the following lower bound:

$$\begin{aligned}&2\tau[(p^T f(x) - p^T f(\hat{x}))/\theta + (q^T g(\hat{x}) - q^T g(x))\theta] \\ &\quad - \tfrac{4M\tau^2}{t}[(p^T f(x) + p^T f(\hat{x}))/\theta + (q^T g(\hat{x}) + q^T g(x))\theta]^2 \\ &\qquad \geq 2\tau t[(p^T f(x) + p^T f(\hat{x}))/\theta + (q^T g(\hat{x}) + q^T g(x))\theta] \\ &\qquad\quad - \tau t[(p^T f(x) + p^T f(\hat{x}))/\theta + (q^T g(\hat{x}) + q^T g(x))\theta] \\ &\qquad = \tau t[(p^T f(x) + p^T f(\hat{x}))/\theta + (q^T g(\hat{x}) + q^T g(x))\theta] \\ &\qquad = \tfrac{t^3}{4M}.\end{aligned}$$

We used above the inequality (3.6) and inserted the step length τ. ∎

3.4. Reducing the step length. Let $x' = (1-\tau)x + \tau\hat{x}$, where x is the current vector, $\hat{x}$ is the block solution, and τ is the step length as used in the previous subsection. Consider a phase s with threshold value $T(s)$. For simplicity we use $T = T(s)$. If $g_m(x') \leq T$ for each $m \in A(x)$, then we use x' as the next iterate and set $A(x') = \{m \in A(x) | g_m(x') < T\}$. In this case some components may be eliminated, but we use the original step length. Now we consider the case that $g_m(x') > T$ for at least one coordinate $m \in A(x)$. Let

$$\gamma(\tilde{\tau}) = \max_{m \in A(x)} g_m(x)(1 - \tilde{\tau}) + g_m(\hat{x})\tilde{\tau}$$

for $0 \leq \tilde{\tau} \leq 1$. If $\gamma(\tau) > T$, then we reduce the step length τ. In this case we compute $\bar{\tau} < \tau$ such that $\gamma(\bar{\tau}) = T$. Using $g_m(x) < T$ for each $m \in A(x)$ and $\gamma(\tau) > T$, there is at least one component $m \in A(x)$ such that $g_m(\hat{x}) > T$. In addition, the value $\bar{\tau}$ is unique and can be computed in $O(M)$ time. We use here $x' = x(1 - \bar{\tau}) + \hat{x}\bar{\tau}$ as next iterate and set $A(x') = \{m \in A(x) | g_m(x') < T\}$. If $\gamma(\tau) \leq T$, then we do not have to reduce the step length τ and use again $x' = x(1 - \tau) + \hat{x}\tau$. But we eliminate as above all components $m \in A(x)$ with $g_m(x') \geq T$. Notice that the case with $g_m(x') > T > g_m(x)(1-\tau) + g_m(\hat{x})\tau$ is possible (since the functions g_m are concave).

For each $m \in A(x')$ we have $g_m(x') < T$. If we use a reduced step length $\bar{\tau} < \tau$, then $A(x) \neq A(x')$. But $A(x) \neq A(x')$ also can happen when $\gamma(\tau) < T$ or $g_m(x') \leq T$ for each $m \in A(x)$. The new potential value for x' and $A(x')$ is

$$\begin{aligned}\phi_t(x', A(x')) = 2\ln\theta' &- \frac{t}{M}\sum_{m=1}^{M} \ln(\theta' - f_m(x')) \\ &- \frac{t}{M}\sum_{m \in A(x')} \ln(g_m(x') - 1/\theta') - \frac{t}{M}\sum_{m \notin A(x')} \ln(T),\end{aligned}$$

where $\theta' = \theta_{A(x')}(x')$.

Now we consider two cases depending on whether we use the original step length τ or the reduced step length $\bar{\tau}$.

THEOREM 3.8. *For any two consecutive iterations with computed vectors x, x', index sets $A(x) \neq A(x')$, and $\max_{m\in A(x)} g_m(x)(1-\tau) + g_m(\hat{x})\tau \leq T$, we have*

$$\phi_t(x, A(x)) - \phi_t(x', A(x')) \geq \frac{t^3}{4M}.$$

Proof. If $g_m(x)(1-\tau) + g_m(\hat{x})\tau \leq T$ for each $m \in A(x)$, then $-\ln(T) \leq -\ln(g_m(x)(1-\tau)+g_m(\hat{x})\tau - 1/\theta)$, where $\theta = \theta_{A(x)}(x)$. Furthermore, for each feasible choice for τ, we have $\theta - f_m(x') > 0$ for each $m \in \{1, \ldots, M\}$ and $g_m(x') - 1/\theta > 0$ for each $m \in A(x)$. This implies $\lambda_{A(x')}(x') \leq \lambda_{A(x)}(x') < \theta_{A(x)}(x)$. Then using the definition of the potential function and $g_m(x') - 1/\theta \geq g_m(x)(1-\tau) + g_m(\hat{x})\tau - 1/\theta$ (concavity of functions g_m),

$$\begin{aligned}
\phi_t(x', A(x')) &= \min_{\lambda_{A(x')}(x') \leq \xi} \Phi_t(\xi, x', A(x')) \leq \Phi_t(\theta, x', A(x')) \\
&= 2\ln\theta - \tfrac{t}{M}\sum_{m=1}^{M} \ln(\theta - f_m(x')) - \tfrac{t}{M}\sum_{m\in A(x')} \ln(g_m(x') - \tfrac{1}{\theta}) \\
&\quad - \tfrac{t}{M}\sum_{m\notin A(x')} \ln(T) \\
&\leq 2\ln\theta - \tfrac{t}{M}\sum_{m=1}^{M} \ln(\theta - f_m(x')) \\
&\quad - \tfrac{t}{M}\sum_{m\in A(x)} \ln(g_m(x)(1-\tau) + g_m(\hat{x})\tau - \tfrac{1}{\theta}) \\
&\quad - \tfrac{t}{M}\sum_{m\notin A(x)} \ln(T) \\
&\leq 2\ln\theta - \tfrac{t}{M}\sum_{m=1}^{M} \ln\left((\theta - f_m(x))\left(1 + \tau\tfrac{f_m(x) - f_m(\hat{x})}{\theta - f_m(x)}\right)\right) \\
&\quad - \tfrac{t}{M}\sum_{m\in A(x)} \ln\left(\left(g_m(x) - \tfrac{1}{\theta}\right)\left(1 + \tau\tfrac{g_m(\hat{x}) - g_m(x)}{g_m(x) - 1/\theta}\right)\right) \\
&\quad - \tfrac{t}{M}\sum_{m\notin A(x)} \ln(T).
\end{aligned}$$

Using the same arguments as in Lemma 3.5, the difference of the potential values $\phi_t(x, A(x)) - \phi_t(x', A(x'))$ is at least

$$\begin{aligned}
&2\tau[(p^T f(x) - p^T f(\hat{x}))/\theta + (q^T g(\hat{x}) - q^T g(x))\theta] \\
&-\tfrac{4M\tau^2}{t}[(p^T f(x) + p^T f(\hat{x}))/\theta + (q^T g(\hat{x}) + q^T g(x))\theta]^2.
\end{aligned}$$

The remaining analysis goes in the same way as for $A(x) = A(x')$. ∎

THEOREM 3.9. *For any two consecutive iterations with computed vectors x, x', index sets $A(x) \neq A(x')$, and $\max_{m\in A(x)} g_m(x)(1-\tau) + g_m(\hat{x})\tau > T$, we have*

$$\phi_t(x, A(x)) - \phi_t(x', A(x')) \geq 0.$$

Proof. If $\max_{m\in A(x)} g_m(x)(1-\tau)+g_m(\hat{x})\tau > T$, then we use the reduced step length $\bar{\tau}$. In this case $g_m(x)(1-\bar{\tau})+g_m(\hat{x})\bar{\tau}$ is also bounded by T for each $m \in A(x)$. This implies $-\ln(T) \leq -\ln(g_m(x)(1-\bar{\tau})+g_m(\hat{x})\bar{\tau}-1/\theta_A(x))$. Since $\bar{\tau}$ is also a feasible choice, $\lambda_{A(x')}(x') < \theta_A(x)$. Using the definition of the potential function and an analysis similar to the one above, we get with $\theta = \theta_A(x)$ the following:

$$\begin{aligned}
\phi_t(x', A(x')) \;\leq\; & 2\ln\theta - \tfrac{t}{M}\sum_{m=1}^{M} \ln\left((\theta - f_m(x))\left(1+\bar{\tau}\tfrac{f_m(x)-f_m(\hat{x})}{\theta-f_m(x)}\right)\right) \\
& -\tfrac{t}{M}\sum_{m\in A(x)} \ln\left(\left(g_m(x)-\tfrac{1}{\theta}\right)\left(1+\bar{\tau}\tfrac{g_m(\hat{x})-g_m(x)}{g_m(x)-1/\theta}\right)\right) \\
& -\tfrac{t}{M}\sum_{m\notin A(x)} \ln(T) \\
\leq\; & \phi_t(x, A(x)) - \tfrac{t}{M}\sum_{m=1}^{M} \ln(1+\tfrac{2\bar{\tau}M}{t\theta}p_m(f_m(x)-f_m(\hat{x}))) \\
& -\tfrac{t}{M}\sum_{m\in A(x)} \ln\left(1+\tfrac{2\bar{\tau}M\theta}{t}q_m(g_m(\hat{x})-g_m(x))\right).
\end{aligned}$$

Again, since $\bar{\tau} < \tau$ is a feasible choice,

$$\begin{aligned}
\phi_t(x, A(x)) - \phi_t(x', A(x')) \;\geq\; & 2\bar{\tau}[(p^T f(x) - p^T f(\hat{x}))/\theta + (q^T g(\hat{x}) - q^T g(x))\theta] \\
& -\tfrac{4M\bar{\tau}^2}{t}[(p^T f(x) + p^T f(\hat{x}))/\theta + (q^T g(\hat{x}) + q^T g(x))\theta]^2 \\
\geq\; & 2\bar{\tau}[(p^T f(x) - p^T f(\hat{x}))/\theta + (q^T g(\hat{x}) - q^T g(x))\theta] \\
& -\tfrac{4M\bar{\tau}\tau}{t}[(p^T f(x) + p^T f(\hat{x}))/\theta + (q^T g(\hat{x}) + q^T g(x))\theta]^2 \\
=\; & 2\bar{\tau}[[(p^T f(x) - p^T f(\hat{x}))/\theta + (q^T g(\hat{x}) - q^T g(x))\theta] \\
& -\tfrac{2M\tau}{t}[(p^T f(x) + p^T f(\hat{x}))/\theta + (q^T g(\hat{x}) + q^T g(x))\theta]^2].
\end{aligned}$$

Now we use the inequality (3.6), insert step length τ, and get $\phi_t(x, A(x)) - \phi_t(x', A(x')) \geq \bar{\tau}t[(p^T f(x) + p^T f(\hat{x}))/\theta + (q^T g(\hat{x}) + q^T g(x))\theta] \geq 0$. ∎

3.5. Convex combination of different vectors. First we prove an upper bound for the packing constraints.

LEMMA 3.10. *For any iteration of the phase s with computed vector x, $\lambda_f(x) = \max_{1\leq m\leq M} f_m(x)$ is bounded by $4M/t_s$.*

Proof. For the initial solution, $f_m(x^{(0)}) \leq 3M/2$ for $m = 1, \ldots, M$. This implies $\lambda_f(x^{(0)}) \leq 3M/2 \leq 3M/(2t_s)$ (for any $t_s \in (0,1)$). Note that $t_s < t_{s-1}$ and $t_1 < 1$. For each block solution $\hat{x}$ in phase s we have

$$p(x, A)^T f(\hat{x}) \leq (1+t_s)\sum_{m=1}^{M} p_m(x, A) < 2\sum_{m=1}^{M} p_m(x, A(x)) \leq 2,$$

where $A = A(x)$. Suppose that $\lambda_f(\hat{x}) > 4M/t_s$. Then, there is an index $m \in \{1, \ldots, M\}$ such that $f_m(\hat{x}) = \lambda_f(\hat{x})$. We note that

$$p_m(x, A) = \frac{t_s}{2M}\frac{\theta_A(x)}{\theta_A(x) - f_m(x)} > \frac{t_s}{2M}.$$

Then we obtain a contradiction using

$$2 = \frac{t_s}{2M}\frac{4M}{t_s} < p_m(x,A)f_m(\hat{x}) \le p(x,A)^T f(\hat{x}) \le 2.$$

In other words $f_m(\hat{x}) \le 4M/t_s$ for each $m \in \{1,\dots,M\}$. Since each vector x' computed by the algorithm is the linear combination of the previous vector x and a block solution $\hat{x}$, and using the convexity of function f_m,

$$f_m(x') = f_m((1-\tau)x + \tau\hat{x}) \le (1-\tau)f_m(x) + \tau f_m(\hat{x}) \le \max\{f_m(x), f_m(\hat{x})\}.$$

Then, the lemma follows by induction on the number of iterations. ∎

Lemma 3.10 shows that the values $f_m(x)$ are not arbitrarily large in the algorithm. Notice that this is independent from the chosen step length $\tau \in (0,1)$. We use this bound for the convex combination below. Notice that, in addition, the components $p_m(x,A(x))$ of the price vector $p(x,A(x))$ are not arbitrarily small (i.e., $p_m(x,A(x)) > t_s/(2M)$).

At the end of phase s we have computed a vector $y \in B$ with $\lambda_{A(y)}(y) \le 1/(1-\epsilon_s/4)$. This implies $f_m(y) \le 1/(1-\epsilon_s/4)$ for each $m \in \{1,\dots,M\}$, and $g_m(y) \ge 1-\epsilon_s/4$ for each $m \in A(y)$. The goal is now to compute a vector $x^{(s)} \in B$ with $\lambda_{\mathcal{M}}(x^{(s)}) \le 1/(1-\epsilon_s)$. The key idea is to use a convex combination over several vectors computed during the phase. Let $x_1,\dots,x_k$ be the vectors in phase s where at least one function g_m is eliminated (i.e., where $g_m(x_i) \ge T(s)$). Clearly, $k \le M$. We have $x_1 = x^{(s-1)}$ if $g_m(x^{(s-1)}) \ge T(s)$ for at least one $m \in \mathcal{M}$ (here $x^{(s-1)}$ is the solution of the previous phase).

Take the following convex combination:

$$x^{(s)} \;=\; \sum_{i=1}^{k} \tfrac{\epsilon_s^2}{264M^2}\, x_i + \left(1 - \tfrac{k\epsilon_s^2}{264M^2}\right) y.$$

Since the set B is convex and $x_1,\dots,x_k,y \in B$, we obtain $x^{(s)} \in B$. Since the functions g_m are concave and the functions f_m are convex,

$$g_m(x^{(s)}) \;\ge\; \tfrac{\epsilon_s^2}{264M^2}\sum_{i=1}^{k} g_m(x_i) + \left(1 - \tfrac{k\epsilon_s^2}{264M^2}\right) g_m(y),$$

$$f_m(x^{(s)}) \;\le\; \tfrac{\epsilon_s^2}{264M^2}\sum_{i=1}^{k} f_m(x_i) + \left(1 - \tfrac{k\epsilon_s^2}{264M^2}\right) f_m(y).$$

Our threshold value $T(s)$ is given by

$$T(s) = 2112\left(\frac{M^3}{\epsilon_s^2}\right)\cdot\frac{1}{\lambda_{\mathcal{M}}(x^{(s-1)})}.$$

Notice that $T(s) \le 2112M^3/\epsilon_s^2$, since $\lambda_{\mathcal{M}}(x^{(s-1)}) \ge 1$ (otherwise we are done).

LEMMA 3.11. *The computed solution $x^{(s)}$ satisfies $\lambda_{\mathcal{M}}(x^{(s)}) \le 1/(1-\epsilon_s)$.*

Proof. First we consider the concave functions g_m. For each eliminated function g_m we have

$$g_m(x^{(s)}) \ge \frac{\epsilon_s^2}{264M^2} g_m(x_i) \ge \frac{\epsilon_s^2}{264M^2} T(s) = \frac{8M}{\lambda_{\mathcal{M}}(x^{(s-1)})},$$

since all other functions are nonnegative. This implies for $s > 1$ that

$$g_m(x^{(s)}) \geq 8(1-\epsilon_{s-1}) = 8(1-2\epsilon_s) \geq 1-\epsilon_s$$

for $8 - 16\epsilon_s \geq 1 - \epsilon_s$ or, equivalently, $\epsilon_s \leq 7/15 < 1/2$. In addition we get for $s = 1$ that

$$g_m(x^{(1)}) \geq 4M/(3M) \geq 1-\epsilon_s.$$

For each remaining component $m \in A(y)$ we have

$$g_m(x^{(s)}) \geq \left(1 - \frac{k\epsilon_s^2}{264M^2}\right) g_m(y) \geq \left(1 - \frac{\epsilon_s^2}{264M}\right)\left(1 - \frac{\epsilon_s}{4}\right) > 1-\epsilon_s.$$

The last inequality holds for each $\epsilon_s \leq 1$.

Next we consider the convex functions f_m. Here we use Lemma 3.10. For each iterate x we have $\lambda_f(x) \leq 4M/t_s$. This implies the following upper bound:

$$f_m(x^{(s)}) \leq \frac{\epsilon_s^2}{264M^2} \cdot \frac{4kM}{t_s} + \left(1 - \frac{k\epsilon_s^2}{264M^2}\right) f_m(y).$$

Since $t_s = \epsilon_s/32$, $k \leq M$, and $f_m(y) \leq 1/(1-\epsilon_s/4)$, we get

$$f_m(x^{(s)}) \leq \frac{32\epsilon_s}{66} + \left(1 - \frac{k\epsilon_s^2}{264M^2}\right) \frac{1}{1-\epsilon_s/4}.$$

The first term is at most $\epsilon_s/2$, and the second term can be bounded by $1 + \epsilon_s/2$ (for $\epsilon_s \leq 2$). This gives $f_m(x^{(s)}) \leq 1 + \epsilon_s$. All three cases together imply that $\lambda_{\mathcal{M}}(x^{(s)}) \leq \max\{1+\epsilon_s, 1/(1-\epsilon_s)\} = 1/(1-\epsilon_s)$. □

4. Analysis of the approximation algorithm.

4.1. Number of iterations. In this subsection we determine the total number of iterations of our algorithm. To do this we first calculate the number of iterations N_s in a single phase s. Let $y, \tilde{y}$ denote the initial and final iterate of phase s. Furthermore, let $\bar{y}$ be the solution after $\bar{N}_s = N_s - 1$ iterations. For consecutive iterations with computed vectors x, x' in a phase and $A(x) = A(x')$, the difference in the potential values $\phi_t(x, A(x)) - \phi_t(x', A(x')) \geq \frac{ct^3}{M}$, where c is a positive constant and $t = t_s = \epsilon_s/32$. In addition, there are at most M iterations with consecutive vectors x, x' and different subsets $A(x) \neq A(x')$ (i.e., in these iterations at least one component is eliminated). In these cases, we have $\phi_t(x, A(x)) - \phi_t(x', A(x')) \geq 0$. Therefore,

$$\phi_t(y, A(y)) - \phi_t(\bar{y}, A(\bar{y})) \geq \frac{ct^3}{M}(\bar{N}_s - M).$$

Next we determine an upper bound for the difference $\phi_t(y, A(y)) - \phi_t(\bar{y}, A(\bar{y}))$. Using Lemma 2.2,

$$\begin{aligned}
\phi_t(y, A(y)) &\leq 2\ln\theta_{A(y)}(y) + 2t\ln(2M/t) + t\ln(1 + t/(2M)), \\
\phi_t(\bar{y}, A(\bar{y})) &\geq (2-t)\ln\theta_{A(\bar{y})}(\bar{y}) - t\ln(T),
\end{aligned}$$

where $T = T(s)$.

This gives

$$\begin{aligned}\phi_t(y, A(y)) - \phi_t(\bar{y}, A(\bar{y})) \;\le\; & (2-t)\ln\tfrac{\theta_{A(y)}(y)}{\theta_{A(\bar{y})}(\bar{y})} + t\ln\theta_{A(y)}(y) + t\ln(T) \\ & + 2t\ln\left(\tfrac{2M}{t}\right) + t\ln\left(1+\tfrac{t}{2M}\right).\end{aligned}$$

Notice that $\theta_{A(y)}(y) \le \lambda_{A(y)}(y)/(1-t) \le O(M)$ for the initial solution y of each phase. This together with the definition of $T = T(s)$ and $t = t_s$ implies that $t\ln\theta_{A(y)}(y) + 2t\ln(2M/t) + t\ln(T) + t\ln(1+t/(2M))$ is bounded by $O(t_s\ln(M/t_s))$.

Now we need a bound for $\ln(\theta_{A(y)}(y)/\theta_{A(\bar{y})}(\bar{y}))$. Using $\theta_{A(y)}(y) < \lambda_{A(y)}(y)/(1-t)$ and $\theta_{A(\bar{y})}(\bar{y}) > \lambda_{A(\bar{y})}(\bar{y})(1+t/(2M)) > \lambda_{A(\bar{y})}(\bar{y})$, we get

$$\ln\frac{\theta_{A(y)}(y)}{\theta_{A(\bar{y})}(\bar{y})} < \ln\frac{\lambda_{A(y)}(y)}{\lambda_{A(\bar{y})}(\bar{y})} + \ln\frac{1}{1-t} \le \ln\frac{\lambda_{A(y)}(y)}{\lambda_{A(\bar{y})}(\bar{y})} + \ln(1+2t).$$

Notice that $\lambda_{A(\bar{y})}(\bar{y}) > \omega_s\lambda_{\mathcal{M}}(y) \ge \omega_s\lambda_{A(y)}(y)$, since $\bar{y}$ does not satisfy the second stopping rule. Therefore,

$$\phi_t(y, A(y)) - \phi_t(\bar{y}, A(\bar{y})) \le (2-t)\ln\left(\frac{1}{\omega_s}\right) + O\left(t_s\ln\left(\frac{M}{t_s}\right)\right).$$

The term $\ln(\frac{1}{\omega_s})$ depends on the scaling phase s. For $s = 1$, $\ln(\frac{1}{\omega_s}) = O(\ln M)$. For the other phases, $\ln(\frac{1}{\omega_s}) \le \ln(1+4\epsilon_s) \le O(\epsilon_s)$ (using $\epsilon_s \le 1/4$). This implies that the difference in the potential function is at most $O(\ln M)$ for $s = 1$ and at most $O(\epsilon_s\ln(M/\epsilon_s))$ for $s \ge 2$. The lower and upper bounds for $\phi_t(y, A(y)) - \phi_t(\bar{y}, A(\bar{y}))$ imply

$$\bar{N}_s \le O(M\epsilon_s^{-2}\ln(M\epsilon_s^{-1})).$$

Summing over all phases, the total number of iterations (calls to the block solver) is

$$O\left(M\ln(M\epsilon^{-1})\sum_{k=0}^{\lceil\log(1/\epsilon)\rceil}(2^k)^2\right).$$

Since $\sum_{k=0}^{\lceil\log(1/\epsilon)\rceil}(2^k)^2 \le O(\epsilon^{-2})$, the total number of iterations is

$$O(M\epsilon^{-2}\ln(M\epsilon^{-1})).$$

4.2. How to compute the minimizer $\boldsymbol{\theta_A(x)}$. We assumed in the sections above that the price vectors $p = p(x, A)$ and $q = q(x, A)$ are computed exactly from (2.2), (2.3) in each iteration. But this is impractical since (2.2), (2.3) requires the root $\theta_A(x)$ of (2.1). However, an approximation of the price vectors is sufficient. Suppose that $\tilde{p}, \tilde{q}$ is an approximation of the exact vectors p, q with relative accuracy $\delta \in (0, 1/2)$, i.e.,

$$(1-\delta)p \;\le\; \tilde{p} \;\le\; (1+\delta)p,$$

$$(1-\delta)q \;\le\; \tilde{q} \;\le\; (1+\delta)q.$$

Let $\hat{x} \in B$ be computed by $ABS(\tilde{p}, \tilde{q}, t)$ with $\tilde{p}^T f(\hat{x}) \leq (1+t) \sum_{m=1}^{M} \tilde{p}_m$ and $\tilde{q}^T g(\hat{x}) \geq \frac{1}{(1+t)} \sum_{m=1}^{M} \tilde{q}_m$. This implies for the original price vectors that

$$p^T f(\hat{x}) \leq \tfrac{(1+t)(1+\delta)}{(1-\delta)} \sum_{m=1}^{M} p_m \leq (1+t+4\delta) \sum_{m=1}^{M} p_m,$$

$$q^T g(\hat{x}) \geq \tfrac{(1-\delta)}{(1+\delta)(1+t)} \sum_{m=1}^{M} q_m \geq \tfrac{1}{1+t+4\delta} \sum_{m=1}^{M} q_m$$

for $t, \delta \leq 1/3$. Using $\delta = t/4 = \Theta(\epsilon)$, we obtain a relative accuracy of $2t$. This shows that (2.2), (2.3) need only the price components p_m, q_m up to a relative accuracy of $\delta = \Theta(\epsilon)$.

Notice that the minimizer $\theta_A(x)[f, g]$ (that also depends on the functions f and g) satisfies the condition $\theta_A(x)[f \cdot s, g/s] = s\, \theta_A(x)[f, g]$ where s is a positive scalar. In addition the price components are independent of a positive scalar s; i.e., $p_m(x, A)[f \cdot s] = p_m(x, A)[f]$ and $q_m(x, A)[g/s] = q_m(x, A)[g]$. This helps to prescale locally the f and g vectors such that $\lambda_A(x)[f, g] = 1$. Suppose that $\lambda_A(x)[f, g] \neq 1$. Then define $\bar{f}_m(x) = f_m(x)/\lambda_A(x)[f, g]$ and $\bar{g}_m(x) = g_m(x)\lambda_A(x)[f, g]$. This gives $\lambda_A(x)[\bar{f}, \bar{g}] = \max(\max_{m \in \mathcal{M}} \bar{f}_m(x), \max_{m \in A} 1/\bar{g}_m(x)) = 1$, and the price vectors for $(\bar{f}, \bar{g})$ are the same. Using this prescaling, we can suppose that the minimizer $\theta = \theta_A(x)$ lies in the interval $[1 + t/(2M), 1 + 2t]$.

Now we estimate the influence of a small absolute error in θ for the price vectors p, q. Suppose that $\tilde{\theta}$ is an approximation of the correct minimizer θ, i.e., $|\tilde{\theta} - \theta| \leq \Delta$. The modified price vectors $\tilde{p}, \tilde{q}$ are given by

$$\begin{aligned} \tilde{p}_m &= \tfrac{t}{2M} \tfrac{\tilde{\theta}}{\tilde{\theta} - f_m(x)}, \\ \tilde{q}_m &= \tfrac{t}{2M} \tfrac{1/\tilde{\theta}}{g_m(x) - 1/\tilde{\theta}}. \end{aligned}$$

These price vectors have to satisfy the following inequalities in order to obtain a relative accuracy of δ:

$$\begin{aligned} |\tfrac{\tilde{p}_m - p_m}{p_m}| &\leq \delta, \\ |\tfrac{\tilde{q}_m - q_m}{q_m}| &\leq \delta. \end{aligned}$$

These are equivalent to

$$\begin{aligned} \left| \frac{\tilde{\theta}}{\theta} \frac{\theta - f_m(x)}{\tilde{\theta} - f_m(x)} - 1 \right| &\leq \delta, \\ \left| \frac{g_m(x)\theta - 1}{g_m(x)\tilde{\theta} - 1} - 1 \right| &\leq \delta. \end{aligned}$$

These inequalities hold if we can prove the following stronger inequalities:

$$\begin{aligned}
\frac{(\theta+\Delta)(\theta-f_m(x))}{\theta(\theta-\Delta-f_m(x))}-1 &\leq \delta,\\
1-\frac{(\theta-\Delta)(\theta-f_m(x))}{\theta(\theta+\Delta-f_m(x))} &\leq \delta,\\
\frac{g_m(x)\theta-1}{g_m(x)(\theta-\Delta)-1}-1 &\leq \delta,\\
1-\frac{g_m(x)\theta-1}{g_m(x)(\theta+\Delta)-1} &\leq \delta.
\end{aligned}$$

The first type of inequality is equivalent to

$$\Delta \leq \frac{\delta\theta(\theta-f_m(x))}{2\theta-f_m(x)+\delta\theta}.$$

Since $\theta \geq 1$, $f_m(x) \leq \lambda_A(x) = 1$, $\theta \leq 1+2t$, and $t, \delta < 1/6$, we have $\delta\theta(\theta-f_m(x)) \geq \delta(\theta-f_m(x)) \geq \delta(\theta-1)$ and $2\theta-f_m(x)+\delta\theta \leq (2+\delta)(1+2t) \leq 3$. Using $\theta \geq 1+t/(2M)$, this gives

$$\frac{\delta\theta(\theta-f_m(x))}{2\theta-f_m(x)+\delta\theta} \geq \frac{\delta}{3}(\theta-1) \geq \frac{\delta t}{6M}.$$

In other words, $\Delta \leq \frac{\delta}{6M}$ is a sufficient condition that the first type of inequality is satisfied. The same holds for the inequalities of the second type. Since δ and t are of order $\Theta(\epsilon)$, the absolute error Δ should be $\Theta(\frac{\epsilon^2}{M})$.

The third type of inequality is equivalent to

$$\Delta \leq \frac{\delta}{1+\delta}\frac{g_m(x)\theta-1}{g_m(x)} = \frac{\delta}{1+\delta}\left(\theta-\frac{1}{g_m(x)}\right).$$

Since $1/g_m(x) \leq \lambda_A(x) = 1$, $\delta \leq 1$, and $\theta \geq 1+t/(2M)$, the right-hand side is at least $\frac{\delta}{1+\delta}(\theta-1) \geq \frac{\delta t}{4M}$. This implies that $\Delta \leq \frac{\delta t}{4M}$ is a sufficient condition that the third type of inequality is satisfied. The same holds for the inequalities of the fourth type. This analysis shows that an absolute error $\Delta = \Theta(\frac{\epsilon^2}{M})$ in the computation of $\theta_A(x)[f,g]$ results in a relative error of $\delta = \Theta(\epsilon)$ in the value of each p_m and q_m.

To compute the value $\theta_A(x)(f,g) \in [1+t/(2M), 1+2t]$ approximately with an absolute error of Δ, we can use binary search. Since the length of the interval is $\Theta(\epsilon)$, there are at most $O(\ln(M/\epsilon))$ steps necessary. In each step, we have to compute the sum

$$\frac{t\theta}{M}\sum_{m=1}^{M}\frac{1}{\theta-f_m(x)} + \frac{t}{M\theta}\sum_{m\in A(x)}\frac{1}{g_m(x)-1/\theta}$$

for a candidate $\theta \in [1+t/(2M), 1+2t]$. Therefore, at most $O(M\ln(M\epsilon^{-1}))$ arithmetic operations per iteration are necessary. Using Newton's method this probably can be improved to $O(M\ln\ln(M\epsilon^{-1}))$ (see the analysis in [4, 5]).

Acknowledgment. The author thanks Florian Diedrich for many helpful discussions.

REFERENCES

[1] L. Fleischer, *A fast approximation scheme for fractional covering problems with variable upper bounds*, in Proceedings of the 15th ACM-SIAM Symposium on Discrete Algorithms (SODA), New Orleans, 2004, pp. 994–1003.

[2] N. Garg and J. Könemann, *Fast and simpler algorithms for multicommodity flow and other fractional packing problems*, in Proceedings of the 39th IEEE Annual Symposium on Foundations of Computer Science (FOCS), 1998, pp. 300–309.

[3] M. D. Grigoriadis and L. G. Khachiyan, *Fast approximation schemes for convex programs with many blocks and coupling constraints*, SIAM J. Optim., 4 (1994), pp. 86–107.

[4] M. D. Grigoriadis and L. G. Khachiyan, *Coordination complexity of parallel price-directive decomposition*, Math. Oper. Res., 2 (1996), pp. 321–340.

[5] M. D. Grigoriadis, L. G. Khachiyan, L. Porkolab, and J. Villavicencio, *Approximate max-min resource sharing for structured concave optimization*, SIAM J. Optim., 11 (2001), pp. 1081–1091.

[6] K. Jansen and L. Porkolab, *On preemptive resource constrained scheduling: Polynomial-time approximation schemes*, SIAM J. Discrete Math., to appear.

[7] K. Jansen and H. Zhang, *Approximation algorithms for general packing problems with modified logarithmic potential function*, in Proceedings of the 2nd IFIP International Conference on Theoretical Computer Science (TCS), Foundations of Information Technology in the Era of Network and Mobile Computing, Kluwer Academic Publishers, Norwell, MA, 2002, pp. 255–266.

[8] K. Jansen, *Approximation algorithms for the general max-min resource sharing problem*, Math. Program., 106 (2006), pp. 547–566.

[9] J. Könemann, *Fast Combinatorial Algorithms for Packing and Covering Problems*, Diploma Thesis, Max-Planck-Institute for Computer Science, Saarbrücken, Germany, 2000.

[10] S. A. Plotkin, D. B. Shmoys, and E. Tardos, *Fast approximation algorithms for fractional packing and covering problems*, Math. Oper. Res., 20 (1995), pp. 257–301.

[11] N. E. Young, *Randomized rounding without solving the linear program*, in Proceedings of the 6th ACM-SIAM Symposium on Discrete Algorithms (SODA), San Francisco, 1995, pp. 170–178.

[12] N. E. Young, *Sequential and parallel algorithms for mixed packing and covering*, in Proceedings of the 42nd Annual IEEE Symposium on Foundations of Computer Science (FOCS), 2001, pp. 538–546.

SIAM J. OPTIM.
Vol. 17, No. 2, pp. 353–371
© 2006 Society for Industrial and Applied Mathematics

CALMNESS AND ERROR BOUNDS FOR CONVEX CONSTRAINT SYSTEMS*

WEN SONG†

Abstract. In the paper, we first compare the concepts of calmness, local error bound, and locally linear regularity. Second, we present some dual sufficient conditions for local error bound and local linear regularity. By using a characterization of the locally linear regularity for a collection of finite nonempty closed and convex subsets, we prove that the sufficient conditions provided in [R. Henrion and J. Outrata, *J. Math. Anal. Appl.*, 258 (2001), pp. 110–130; R. Henrion and A. Jourani, *SIAM J. Optim.*, 13 (2002), pp. 520–534; R. Henrion, A. Jourani, and J. Outrata, *SIAM J. Optim.*, 13 (2002), pp. 603–618] imply not only the calmness but also the existence of local error bounds for the convex constraint system in an infinite-dimensional setting.

Key words. calmness, error bounds, convex systems, local linear regularity, set-valued mapping

AMS subject classifications. 90C31, 26E25, 49J52

DOI. 10.1137/S1052623403430361

1. Introduction.

The concept of calmness or error bound plays an important role in various branches of the theory of constrained optimization such as nondegenerate multiplier rules (e.g., [7, 14, 28]), exactly penalty functions, and stability of constraint systems (e.g., [13, 27, 21, 5, 19]).

We recall (see [1, 28]) that a set-valued mapping $M\colon Y \rightrightarrows X$ between metric spaces Y, X is called pseudo-Lipschitz (or has Aubin property) 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 $\gamma > 0$ such that

$$d(x, M(y')) \leq \gamma d(y, y') \ \ \forall y, y' \in \mathcal{V}\ , \forall x \in M(y) \cap \mathcal{U},$$

or equivalently,

$$M(y) \cap \mathcal{U} \subset M(y') + \gamma d(y, y') B_X \ \forall y, y' \in \mathcal{V},$$

where B_X denotes the closed unit ball of X. When fixing $y' = \bar{y}$ in the above definition, M is called calm at $(\bar{y}, \bar{x})$ (see [28, 9]), i.e., there exist neighborhoods $\mathcal{V}$,$\mathcal{U}$ of $\bar{y}, \bar{x}$, respectively, and some $\gamma > 0$ such that

$$d(x, M(\bar{y})) \leq \gamma d(y, \bar{y}) \ \ \forall x \in M(y) \cap \mathcal{U}, \forall y \in \mathcal{V},$$

or equivalently,

$$M(y) \cap \mathcal{U} \subset M(\bar{y}) + \gamma d(y, \bar{y}) B_X \ \ \forall y \in \mathcal{V}.$$

When $\mathcal{U} = X$, the calmness reduces to the upper Lipschitz property of set-valued mapping introduced by Robinson [27]. Obviously, either the pseudo-Lipschitz property or the upper Lipschitz property implies calmness, but the converse is not true in

*Received by the editors June 26, 2003; accepted for publication (in revised form) January 3, 2006; published electronically May 19, 2006. This work was partially supported by the National Natural Sciences Grant (10471032) and the Excellent Young Teachers Program of MOE, P.R.C.
http://www.siam.org/journals/siopt/17-2/43036.html
†Department of Mathematics, Harbin Normal University, Harbin 150080, China (wsong218@yahoo.com.cn).

general (see [19, 11]). The pseudo-Lipschitz property and the upper Lipschitz property for a set-valued mapping have been extensively studied in the literature (see [1, 14, 23, 24, 15, 17, 19, 8]).

Sufficient conditions for calmness in a finite-dimensional setting have been derived in [9, 11] for set-valued maps of the type

$$M(y) = S(y) \cap C, \tag{1}$$

where $S\colon Y \rightrightarrows X$ is a set-valued map with closed graph and $C \subset X$ is closed. For instance, $S(y)$ may be a solution set of a constraint system

$$S(y) = \{x \in X \mid f(x) \in y - K\}, \tag{2}$$

where a function $f\colon X \to Y$ and a nonempty subset $K \subset Y$ are given. In particular, when $Y = \mathbb{R} \cup \{+\infty\}$ and $K = \mathbb{R}_+$

$$S(y) = \{x \in X \mid f(x) \leq y\}. \tag{2'}$$

In the convex case, a subdifferential condition for calmness of the constraint system (1) and (2′) has been given by Henrion and Jourani [10] in infinite-dimensional spaces.

We say that M has a local error bound at $(\bar{y}, \bar{x})$ if there exist $\delta, \gamma > 0$ such that

$$d(x, M(\bar{y})) \leq \gamma \max\{d(x, C), d(\bar{y}, S^{-1}(x))\} \ \ \forall x \in B(\bar{x}, \delta).$$

Clearly, when S is given by (2) or (2′), M has a local error bound $(\bar{y}, \bar{x})$ if and only if there exist positive scalars γ and δ such that

$$d(x, M(\bar{y})) \leq \gamma \max\{d(x, C), d(f(x), \bar{y} - K)\} \ \ \forall x \in B(\bar{x}, \delta) \tag{3}$$

or

$$d(x, M(\bar{y})) \leq \gamma \max\{d(x, C), (f(x) - \bar{y})^+\} \ \ \forall x \in B(\bar{x}, \delta), \tag{3'}$$

where $r^+ = \max\{r, 0\}$ for $r \in \mathbb{R}$.

Existence of error bounds for constraint systems has been extensively studied by many authors (see [13, 27, 20, 21, 26]).

The collection of closed convex sets $\{C_1, \ldots, C_n\}$ is called locally linearly regular around $\bar{x} \in C = \cap_{i=1}^n C_i$ if there exist $\delta, \gamma > 0$ such that

$$d(x, C) \leq \gamma \max\{d(x, C_1), \ldots, d(x, C_n)\} \ \ \forall x \in B(\bar{x}, \delta).$$

This notion is a special case of so-called bounded linear regularity introduced in [2]. It is known that bounded linear regularity is very useful in various branches of convex optimization, such as convex feasibility problem, constrained approximation, Fenchel duality, systems of convex inequalities and associated with error bounds, and subdifferential calculus (see [2, 3, 10, 13, 29, 15]).

In this paper, we first compare the concepts of calmness, local error bound, and local linear regularity. Second, we present some dual sufficient conditions for local error bound and local linear regularity. By using a characterization of the local linear regularity, we prove that the sufficient conditions provided in [9, 10, 11] imply not only the calmness but also the existence of local error bounds for the convex constraint system in an infinite-dimensional setting.

2. Notation and basic results. Let X be a normed space with topological dual space X^*. Denote by B_X and B_{X^*} the closed unit ball of X and X^*, respectively. We write $B(x,\delta)$ for $x+\delta B_X$, where $x\in X$ and $\delta>0$. Let A be a nonempty subset of X. Denote by $\bar{A}$ and $\mathrm{int}A$, respectively, the closure and the interior of A. The cone generated by A is $\mathrm{cone}(A)=\cup_{\lambda\geq 0}\lambda A$. For a nonempty closed convex subset C of X and $\bar{x}\in C$, define the tangent cone to C at $\bar{x}$ by $T_C(\bar{x})=\overline{\mathrm{cone}}(C-\bar{x})$ and the normal cone to C at $\bar{x}$ by

$$N_C(\bar{x})=(T_C(\bar{x}))^0=\{x^*\in X^*\mid \langle x^*,x-\bar{x}\rangle\leq 0\ \ \forall x\in C\}.$$

Let $f\colon X\to\mathbb{R}\cup\{+\infty\}$ be a lower semicontinuous proper convex function. By $\mathrm{epi}f$, $\partial f(\bar{x})$, and $\partial^\infty f(\bar{x})$ we denote the epigraph, the usual, and the singular subdifferentials of f, respectively, in the sense of convex analysis. It is well known that

$$\partial f(\bar{x})=\{x^*\in X^*\mid (x^*,-1)\in N_{\mathrm{epi}f}(\bar{x},f(\bar{x}))\},$$

$$\partial^\infty f(\bar{x})=\{x^*\in X^*\mid (x^*,0)\in N_{\mathrm{epi}f}(\bar{x},f(\bar{x}))\}.$$

Let $I_C(\cdot)$ be the indicator function of C, i.e.,

$$I_C(x)=\begin{cases}0 & \text{if } x\in C,\\ +\infty & \text{if } x\notin C.\end{cases}$$

It is obvious that $N_C(x)=\partial I_C(x)$.

Let us recall some notation and basic results on linear regularity of sets in X (see [3, 29]).

DEFINITION 2.1. *Suppose that $C_1,\dots,C_n$ are subsets of a normed space X with $C=\cap_{i=1}^n C_i\neq\emptyset$.*

(i) *The collection $\{C_1,\dots,C_n\}$ is linearly regular if there exists $k>0$ such that $d(x,C)\leq k\max\{d(x,C_1),\dots,d(x,C_n)\}$ for every $x\in X$.*

(ii) *The collection $\{C_1,\dots,C_n\}$ is boundedly linearly regular if for every bounded set S, there exists $k_S>0$ such that*

$$d(x,C)\leq k_S\max\{d(x,C_1),\dots,d(x,C_n)\}\ \ \forall x\in S.$$

(iii) *The collection $\{C_1,\dots,C_n\}$ is locally linearly regular around some point $\bar{x}\in C$ if there exist $\delta,\gamma>0$ such that $d(x,C)\leq\gamma\max\{d(x,C_1),\dots,d(x,C_n)\}$ for every $x\in B(\bar{x},\delta)$.*

It is trivial that linear regularity implies bounded linear regularity, which, in turn, implies locally linear regularity around every point of C.

In the following we assume that X is a Banach space. We recall the following result, which can be proved by Lemma 1.1 and Proposition 1.3 in [26]; for completeness we give its proof here.

LEMMA 2.1. *Let S be a closed convex subset of X and $\bar{x}\in S$. Then for every $\delta>0$, $0<\sigma<1$ and every $x\in B(\bar{x},\delta/3)$, there exists $z\in \mathrm{bd}S\cap B(\bar{x},\delta)$ such that*

$$\sigma d(x,S)\leq d(x-z,T_S(z)).$$

Proof. Let $x\in B(\bar{x},\delta/3)\setminus S$, let $d=d(x,S)$, and take $\epsilon>0$ such that $\sigma\leq\frac{d-\sqrt{\epsilon}(d+\epsilon+\sqrt{\epsilon})}{(1+\sqrt{\epsilon})(d+\epsilon+\sqrt{\epsilon})}$ and $\epsilon+\sqrt{\epsilon}\leq\delta/3$. By the separation theorem, there exists $x_0^*\in X^*$ with $\|x_0^*\|=1$ such that

$$\sup_{u\in S}\langle x_0^*,u\rangle=\inf_{u\in B_X}\langle x_0^*,x+du\rangle=\langle x_0^*,x\rangle-d.$$

Take $x_0 \in S$ such that $\|x - x_0\| \le d + \epsilon$. Then

$$\sup_{u\in S}\langle x_0^*, u\rangle = \langle x_0^*, x\rangle - d \le \|x_0^*\|\|x - x_0\| + \langle x_0^*, x_0\rangle - d \le \epsilon + \langle x_0^*, x_0\rangle.$$

By Phelps, Brøndsted, and Rockafellar's lemma (see [12]), there exist $z \in \mathrm{bd}S$, $x^* \in X^*$ such that

$$\sup_{u\in S}\langle x^*, u\rangle = \langle x^*, z\rangle \text{ (equivalently, } x^* \in N_S(z))$$

and

$$\|z - x_0\| \le \sqrt{\epsilon}, \ \ \|x^* - x_0^*\| \le \sqrt{\epsilon}.$$

This implies that

$$\|x^*\| \le 1 + \sqrt{\epsilon}, \ \ \|z - x\| \le d + \epsilon + \sqrt{\epsilon}$$

and

$$\langle x^*, z - x\rangle = \langle x^* - x_0^*, z - x\rangle + \langle x_0^*, z - x\rangle \le \sqrt{\epsilon}(d + \epsilon + \sqrt{\epsilon}) - d.$$

It follows that

$$\langle x^*, x - z\rangle \ge d - \sqrt{\epsilon}(d + \epsilon + \sqrt{\epsilon}) \ge \frac{d - \sqrt{\epsilon}(d + \epsilon + \sqrt{\epsilon})}{(1 + \sqrt{\epsilon})(d + \epsilon + \sqrt{\epsilon})}\|x^*\|\|z - x\| \ge \sigma\|x^*\|d(x, S).$$

Note that $T_S(z) \subset \{u \in X \mid \langle x^*, u\rangle \le 0\} =: D$. Hence, we can deduce that

$$\sigma d(x, S) \le \frac{\langle x^*, x - z\rangle}{\|x^*\|} = d(x - z, D) \le d(x - z, T_S(z)).$$

It is clear that

$$\|z - \bar{x}\| \le \|z - x\| + \|x - \bar{x}\| < d + \epsilon + \sqrt{\epsilon} + \frac{\delta}{3} < \delta,$$

i.e., $z \in B(\bar{x}, \delta)$. □

THEOREM 2.1. *Let $C_1, \ldots, C_n$ be closed convex subsets of X with $C = \cap_i C_i \ne \emptyset$ and let $\bar{x} \in C$. The following statements are equivalent:*

(i) *There exist $\delta, \gamma > 0$ such that for every $u \in C \cap B(\bar{x}, \delta)$*

$$N_C(u) \cap B_{X^*} \subset \gamma \sum_i (N_{C_i}(u) \cap B_{X^*}).$$

(ii) *There exist $\delta, \gamma > 0$ such that*

$$d(h, T_C(u)) \le \gamma \max\{d(h, T_{C_1}(u)), \ldots, d(h, T_{C_n}(u))\} \ \ \forall h \in X, u \in B(\bar{x}, \delta) \cap C.$$

(iii) *There exist $\hat{\delta}, \hat{\gamma} > 0$ such that*

$$d(x, C) \le \hat{\gamma} \max\{d(x, C_1), \ldots, d(x, C_n)\} \ \ \forall x \in B(\bar{x}, \hat{\delta}).$$

Proof. (i) $\Rightarrow$ (ii) For every $u \in C \cap B(\bar{x}, \delta)$, since $N_C(u) \cap B_{X^*} \subset \gamma \sum_i (N_{C_i}(u) \cap B_{X^*})$, we have that $N_C(u) \subset \sum_i N_{C_i}(u)$. The converse inclusion is obvious. Hence $N_C(u) = \sum_i N_{C_i}(u)$ and hence $\sum_i N_{C_i}(u)$ is weakly star closed. By Jameson's Theorem 2.1 in [16] (see also Proposition 6 in [3]), we have

$$d(h, T_C(u)) \leq \gamma \max\{d(h, T_{C_i}(u)), \ldots, d(h, T_{C_n}(u))\} \ \ \forall h \in X.$$

(ii) $\Rightarrow$ (i) Condition (ii) implies that for every $u \in C \cap B(\bar{x}, \delta)$, $T_C(u) = \cap_{i=1}^n T_{C_i}(u))$. By the bipolar theorem, we have $N_C(u) = \mathrm{cl}^*(\sum_i N_{C_i}(u))$. Again, by Jameson's Theorem 2.1 in [16] (see also Proposition 6 in [3]), we see that $\sum_i N_{C_i}(u)$ is weakly star closed and $N_C(u) \cap B_{X^*} \subset \gamma \sum_i (N_{C_i}(u) \cap B_{X^*})$. Hence, we obtain (i).

(ii) $\Rightarrow$ (iii) If $x \in C$, then the assertion is obvious. Suppose that $x \in B(\bar{x}, \hat{\delta}) \setminus C$, where $\hat{\delta} = \frac{\delta}{3}$. From Corollary 2 in [6], i.e., for every closed convex subset D in X and $u \in D$, $d^{'}(\cdot, D)(u, h) = d(h, T_D(u))$ for all $h \in X$, one has $d^{'}(\cdot, C_i)(u, h) = d(h, T_{C_i}(u)$ for all $h \in X$, $i = 1, \ldots, n$. By the convexity of the distance function, we have

$$d'(\cdot, C_i)(u, x - u) \leq d(x, C_i) - d(u, C_i) = d(x, C_i) \ \ \forall u \in C.$$

Consequently,

$$d(x - u, T_{C_i}(u)) \leq d(x, C_i) \ \ \forall u \in C. \tag{4}$$

For $0 < \sigma < 1$, by Lemma 2.1, there exists $u \in \mathrm{bd}C \cap B(\bar{x}, \delta)$ such that

$$\sigma d(x, C) \leq d(x - u, T_C(u)). \tag{5}$$

It follows from (4), (5), and (ii) that

$$\sigma d(x, C) \leq \gamma \max\{d(x, C_1), \ldots, d(x, C_n)\}.$$

Let σ tend to 1. We have

$$d(x, C) \leq \gamma \max\{d(x, C_1), \ldots, d(x, C_n)\}.$$

(iii) $\Rightarrow$ (ii) Let $\delta = \hat{\delta}/2, \gamma = \hat{\gamma}$, and let $u \in \mathrm{bd}C \cap B(\bar{x}, \delta)$ and $h \notin T_C(u)$ with $\|h\| = 1$. Then $u + th \notin C$ for all $t > 0$. Since, $\|u + th - \bar{x}\| \leq \|u - \bar{x}\| + t\|h\| \leq \hat{\delta}$ for all $t \in (0, \delta)$, we have

$$d(u + th, C) \leq \gamma \max\{d(u + th, C_1), \ldots, d(u + th, C_n)\}.$$

Noticing that $u \in C = \cap_{i=1}^n C_i$, it follows from taking the limit that

$$d'(\cdot, C)(u, h) \leq \gamma \max\{d'(\cdot, C_1)(u, h), \ldots, d^{'}(\cdot, C_n)(u, h)\}.$$

That is

$$d(h, T_C(u)) \leq \gamma \max\{d(h, T_{C_1}(u)), \ldots, d(h, T_{C_n}(u))\}.$$

For the other case, the conclusion is obviously true. □

Remark 1. Condition (i) is equivalent to each of the following conditions:

(i$^{'}$) There exist $\delta, \gamma > 0$ such that for every $u \in C \cap B(\bar{x}, \delta)$

$$N_C(u) = \sum_i N_{C_i}(u), \ \ \left(\sum_i N_{C_i}(u)\right) \cap B_{X^*} \subset \gamma \sum_i (N_{C_i}(u) \cap B_{X^*}).$$

(i$''$) There exist $\delta, \gamma > 0$ such that for every $u \in C \cap B(\bar{x}, \delta)$

$$T_C(u) = \cap_i T_{C_i}(u), \quad \left(\sum_i N_{C_i}(u)\right) \cap B_{X^*} \subset \gamma \sum_i (N_{C_i}(u) \cap B_{X^*}).$$

Indeed, (i) implies that $N_C(u) \subset \sum_i N_{C_i}(u)$, and the converse inclusion is obviously true. Hence (i$'$) is true. It clear that (i$'$) implies (i$''$). By the bipolar theorem, $T_C(u) = \cap_i T_{C_i}(u)$ implies that $N_C(u) = \mathrm{cl}^*(\sum_i N_{C_i}(u))$. By Proposition 5 in [16], $(\sum_i N_{C_i}(u)) \cap B_{X^*} \subset \gamma \sum_i (N_{C_i}(u) \cap B_{X^*})$ implies that $\sum_i N_{C_i}(u)$ is weakly star closed. Hence $N_C(u) = \sum_i N_{C_i}(u)$ and (i) is true.

The equivalence of (i$''$), (ii), and (iii) has been proved in [29]. From the proof of Theorem 2.1, we can see that the following result concerning the linear regularity is also true.

COROLLARY 2.1. *Let $C_1, \ldots, C_n$ be closed convex subsets of X with $C = \cap_i C_i \neq \emptyset$. The following statements are equivalent:*

(i) *There exist $\gamma > 0$ such that for every $u \in C$*

$$N_C(u) \cap B_{X^*} \subset \gamma \sum_i (N_{C_i}(u) \cap B_{X^*}).$$

(i$'$) *There exists $\gamma > 0$ such that for every $u \in C$*

$$N_C(u) = \sum_i N_{C_i}(u), \quad \left(\sum_i N_{C_i}(u)\right) \cap B_{X^*} \subset \gamma \sum_i (N_{C_i}(u) \cap B_{X^*}).$$

(i$''$) *There exists $\gamma > 0$ such that for every $u \in C$*

$$T_C(u) = \cap_i T_{C_i}(u), \quad \left(\sum_i N_{C_i}(u)\right) \cap B_{X^*} \subset \gamma \sum_i (N_{C_i}(u) \cap B_{X^*}).$$

(ii) *There exists $\gamma > 0$ such that*

$$d(h, T_C(u)) \leq \gamma \max\{d(h, T_{C_1}(u)), \ldots, d(h, T_{C_n}(u))\} \ \ \forall h \in X, u \in C.$$

(iii) *There exists $\hat{\gamma} > 0$ such that*

$$d(x, C) \leq \hat{\gamma} \max\{d(x, C_1), \ldots, d(x, C_n)\} \ \ \forall x \in X.$$

The equivalence of (i$'$), (ii), and (iii) has been proved in Theorem 2.6 in [26].

3. Calmness and error bounds. In the following we assume that X and Y are Banach spaces. Consider now a set-valued mapping $M\colon Y \rightrightarrows X$ defined as the intersection $M(y) = S(y) \cap C$, where $S\colon Y \rightrightarrows X$ is a set-valued mapping and C is a subset of X.

THEOREM 3.1. *Let the set-valued mapping M defined as above and let $(\bar{y}, \bar{x}) \in \mathrm{gr}M$. Consider the following conditions:*

(i) *The set-valued mapping M is calm at $(\bar{y}, \bar{x})$.*

(ii) *There exist $\delta, \gamma > 0$ such that*

$$d(x, M(\bar{y})) \leq \gamma d(\bar{y}, S^{-1}(x)) \ \ \forall x \in C \cap B(\bar{x}, \delta).$$

(iii) *There exist* $\delta, \gamma > 0$ *such that*

$$d(x, M(\bar{y})) \le \gamma \max\{d(x, C), d(\bar{y}, S^{-1}(x))\} \ \ \forall x \in B(\bar{x}, \delta).$$

(iv) *There exist* $\delta, \gamma > 0$ *such that*

$$d(x, M(\bar{y})) \le \gamma \max\{d((x, \bar{y}), C \times \{\bar{y}\}), d((x, \bar{y}), \mathrm{gr}S^{-1})\} \ \ \forall x \in B(\bar{x}, \delta).$$

Then

$$\text{(iv)} \Longrightarrow \text{(iii)} \Longrightarrow \text{(ii)} \Longleftrightarrow \text{(i)}.$$

When $C = X$*, all the conditions are equivalent.*

Proof. It is obvious that

$$\text{(iv)} \Longrightarrow \text{(iii)} \Longrightarrow \text{(ii)} \Longrightarrow \text{(i)}.$$

(i) $\Rightarrow$ (ii) By the calmness of M at $(\bar{y}, \bar{x})$, fix some positive numbers $\delta, \hat{\gamma}$ $(\hat{\gamma} \ge 1)$ such that

$$d(x, M(\bar{y})) \le \hat{\gamma}\|\bar{y} - y\| \ \ \forall y \in B(\bar{y}, \delta), \ x \in S(y) \cap C \cap B(\bar{x}, \delta).$$

Let $x \in B(\bar{x}, \delta) \cap C$, $y \in S^{-1}(x)$. If $y \in B(\bar{y}, \delta)$ as well, then $d(x, M(\bar{y})) \le \hat{\gamma}\|\bar{y} - y\|$. Otherwise we get the desired estimate that

$$d(x, M(\bar{y})) \le \|x - \bar{x}\| \le \delta \le \hat{\gamma}\|\bar{y} - y\|.$$

Consequently, we have

$$d(x, M(\bar{y})) \le \hat{\gamma} d(\bar{y}, S^{-1}(x)) \ \ \forall x \in B(\bar{x}, \delta) \cap C.$$

Suppose $C = X$. It is trivial that (iii) $\Longleftrightarrow$ (ii). For the proof of (iii) $\Longrightarrow$ (iv), see Proposition 4 in [15]. For the completeness, we include the proof. Fix $x \in B(\bar{x}, \delta)$ and consider any $(y', x') \in \mathrm{gr}S$ with $y' \in B(\bar{y}, \delta)$, $x' \in B(\bar{x}, \delta)$. Then

$$\begin{aligned} d(x, M(\bar{y})) &\le \|x - x'\| + d(x', M(\bar{y})) \\ &\le \|x - x'\| + \gamma\|\bar{y} - y'\| \\ &\le \max\{1, \gamma\}(\|x - x'\| + \|\bar{y} - y'\|). \end{aligned}$$

So, putting $\hat{\gamma} = \max\{1, \gamma\}$ we obtain

$$d(x, M(\bar{y})) \le \hat{\gamma} \inf\{\|x - x'\| + \|\bar{y} - y'\| \mid (x', y') \in (\mathrm{gr}S^{-1}) \cap B(\bar{x}, \delta) \times B(\bar{y}, \delta)\}.$$

Since for any $(x', y') \in X \times Y$ with $\|x' - \bar{x}\| \ge \delta$ and $\|y' - \bar{y}\| \ge \delta$ we have

$$\|x - x'\| + \|\bar{y} - y'\| \ge \|x' - \bar{x}\| - \|x - \bar{x}\| + \|y' - \bar{y}\| \ge 2\delta - \delta \ge d((x, \bar{y}), \mathrm{gr}S^{-1}).$$

Hence

$$d(x, M(\bar{y})) \le \hat{\gamma} d((x, \bar{y}), \mathrm{gr}S^{-1}) \ \ \forall x \in B(\bar{x}, \delta). \qquad \square$$

Remark 2. (a) The equivalence between (i) and (ii) shows that there is no need at all to restrict y to a neighborhood V of $\bar{y}$ in the description of calmness (this equivalence was also proved in [8] under some additional condition). Moreover, the

equivalence between (i) and (ii) also shows that the calmness of M at $(\bar{y},\bar{x})$ amounts to the existence of a local error bound of M at the same point whenever $C = X$.

(b) Condition (ii) implies (iii) whenever S^{-1} is pseudo-Lipschitz at $(\bar{x},\bar{y})$.

Indeed, if $S^{-1}(x)$ is pseudo-Lipschitz at $(\bar{x},\bar{y})$, then the function $\rho(x) := d(\bar{y}, S^{-1}(x))$ is locally Lipschitz at $\bar{x}$. Without loss of generality, we may assume that

$$|\rho(x_1) - \rho(x_2)| \le L\|x_1 - x_2\| \ \ \forall x_1, x_2 \in B(\bar{x},\delta)$$

for some $\delta > 0$ and $L > 0$. By the nonexpansivity of distance function and (ii), for every $x_1 \in C \cap B(\bar{x},\delta)$, $x \in B(\bar{x},\delta)$, we have

$$0 \le \gamma d(\bar{y}, S^{-1}(x_1)) - d(x_1, M(\bar{y})) \le \gamma d(\bar{y}, S^{-1}(x)) - d(x, M(\bar{y})) + (\gamma L + 1)\|x - x_1\|.$$

This implies that

$$d(x, M(\bar{y})) \le \gamma d(\bar{y}, S^{-1}(x)) + (\gamma L + 1) d(x, C \cap B(\bar{x},\delta)).$$

Hence,

$$d(x, M(\bar{y})) \le \hat{\gamma} \max\{d(\bar{y}, S^{-1}(x)), d(x, C)\} \ \ \forall x \in B(\bar{x},\delta/3),$$

where $\hat{\gamma} = \max\{\gamma L + 1, \gamma\}$.

(c) When $C = X$, the equivalence between (i) and (iv) implies that if S is calm at $(\bar{y},\bar{x})$ and $\bar{x}$ is a locally minimizer of a locally Lipschitz function $f\colon X \to \mathbb{R}$ on $S(\bar{y})$, then $\bar{x}$ is a free local minimizer of $p(x) := f(x) + \alpha p_S(x)$ whenever α is sufficiently large, where $p_S(x) = d((x,\bar{y}), \mathrm{gr}S^{-1}) = d((\bar{y},x), \mathrm{gr}S)$. This result has been proved by Klatte and Kummer in [19] by constructing a locally upper Lipschitz submapping.

PROPOSITION 3.1. *The condition* (iv) *in Theorem* 3.1 *holds if and only if* $\{C \times \{\bar{y}\}, \mathrm{gr}S^{-1}\}$ *is locally linear regular around* $(\bar{x},\bar{y})$.

Proof. The implication "$\Leftarrow$" is trivial. Suppose (iv) holds, i.e., there exist $\delta, \gamma > 0$ such that

$$d(x, M(\bar{y})) \le \gamma \max\{d((x,\bar{y}), C \times \{\bar{y}\}), d((x,\bar{y}), \mathrm{gr}S^{-1})\} \ \ \forall x \in B(\bar{x},\delta).$$

This implies that

$$d((x,y), (C \times \{\bar{y}\}) \cap \mathrm{gr}S^{-1}) \le \gamma d((x,y), \mathrm{gr}S^{-1}) \ \ \forall (x,y) \in (C \times \{\bar{y}\}) \cap B((\bar{x},\bar{y}),\delta).$$

By the nonexpansivity of the distance function, we can deduce that for all $(x,y) \in B((\bar{x},\bar{y}),\delta/3)$

$$d((x,y), (C \times \{\bar{y}\}) \cap \mathrm{gr}S^{-1}) \le (2\gamma + 1) \max\{d((x,y), C \times \{\bar{y}\}), d((x,y), \mathrm{gr}S^{-1})\}$$

(see Proposition 6 in [29] and also Proposition 2.43 in [7]). Hence $\{C \times \{\bar{y}\}, \mathrm{gr}S^{-1}\}$ is locally linear regular around $(\bar{x},\bar{y})$. □

We recall the following notions of normal cones (see [23, 24]). Let $\Omega \subset X$ and $\epsilon \ge 0$. Given $x \in \mathrm{cl}\Omega$, the nonempty set

$$\hat{N}_\epsilon(x,\Omega) := \left\{ x^* \in X^* \ \middle|\ \limsup_{u \xrightarrow{\Omega} x} \frac{\langle x^*, u - x\rangle}{\|u - x\|} \le \epsilon \right\}$$

is called the set of (Fréchet) ϵ-normal cones to Ω at x. When $\epsilon = 0$, the set $\hat{N}_0(x,\Omega)$ is called the Fréchet normal cone to Ω at x and is denoted by $\hat{N}(x,\Omega)$. If $x \notin \mathrm{cl}\Omega$, we set $\hat{N}_\epsilon(x,\Omega) = \emptyset$ for all $\epsilon \ge 0$.

The nonempty cone

$$N(x,\Omega) := \limsup_{x\to\bar{x},\epsilon\searrow 0} \hat{N}_\epsilon(x,\Omega)$$

is called the normal cone to Ω at x, where

$$\limsup_{x\to\bar{x},\epsilon\searrow 0} \hat{N}_\epsilon(x,\Omega) = \{x^* \in X^* \mid \exists \text{ sequences } x_k \to \bar{x}, \epsilon_k \searrow 0$$
$$\text{and } x_k^* \overset{w^*}{\to} x^* \text{ with } x_k^* \in \hat{N}_{\epsilon_k}(x_k,\Omega)\}.$$

We set $N(x,\Omega) = \emptyset$ for $\bar{x} \notin \mathrm{cl}\Omega$.

It is well known that for a convex set Ω, both the Fréchet normal cone $\hat{N}(x,\Omega)$ and normal cone $N(x,\Omega)$ coincide with the normal cone $N_\Omega(x)$ in the sense of convex analysis, i.e.,

$$N_\Omega(x) = \{x^* \in X^* \mid \langle x^*, u-x\rangle \le 0 \ \ \forall u \in \Omega\}.$$

A closed set Ω in X is called normally compact around $\bar{x} \in \Omega$ (see [22, 23]) if there exist positive numbers γ, σ and a compact subset K of X such that

$$\hat{N}(x,\Omega) \subset \left\{x^* \in X^* \mid \sigma\|x^*\| \le \max_{z\in K} |\langle x^*, z\rangle|\right\} \ \ \forall x \in B_\gamma(\bar{x}) \cap \Omega.$$

It is clear that if Ω is convex and normally compact around $\bar{x} \in \Omega$, then for any net

$$x_i^* \in N_\Omega(x_i) \text{ with } x_i \to \bar{x} \text{ and } x_i^* \overset{w^*}{\to} 0$$

one has $x_i^* \to 0$ in the norm topology of X^* (see [22, 25]).

Observe that each closed set Ω in a finite-dimensional space is normally compact around every point $\bar{x} \in \Omega$. Loewen [22] shows that Ω is normally compact around $\bar{x} \in \Omega$ if Ω is compactly epi-Lipschitzian at $\bar{x} \in \Omega$ in the sense of Borwein and Strojwas [4]. The latter means that there exist a neighborhood N_x of x, a neighborhood U of the origin, a positive number ϵ, and a compact set K such that

$$C \cap N_x + \lambda U \subset C + \lambda K \ \ \forall 0 < \lambda < \epsilon.$$

For a set-valued map $S\colon X \to Y$ and some point $(x,y) \in \mathrm{gr}S$, the coderivative $D^*S(x,y)\colon Y^* \to X^*$ is defined by

$$D^*S(x,y)(u^*) := \{v^* \in X^* \mid (v^*, -u^*) \in N((x,y), \mathrm{gr}S\}.$$

If S is single-valued, we simply write $D^*S(x)$ instead of $D^*S(x, S(x))$.

A set-valued mapping $S\colon X \to Y$ is said to be partially ∂-coderivatively compact at $(\bar{x},\bar{y}) \in \mathrm{gr}S$ (see [18]) if for every net $\{(x_i,y_i)\} \subset \mathrm{gr}S$ converging to $(\bar{x},\bar{y})$ and every net $\{(x_i^*, y_i^*)\}$ satisfying $x_i^* \in D^*S(x_i,y_i)(y_i^*)$, $\|x_i^*\| \to 0$, and $y_i^* \overset{w^*}{\to} 0$ we have $\|y_i^*\| \to 0$.

Observe that the above property always holds when Y is finite-dimensional. It has been proved that $S\colon X \to Y$ is partially ∂-coderivatively compact at $(\bar{x},\bar{y})$ if the graph of S is closed and convex and $\bar{y} \in \mathrm{int}S(X)$ (see [18]) or if S is partially compactly epi-Lipschitz at $(\bar{x},\bar{y})$ in the sense of Jourani and Thibault (see [18]) or if the graph of S is closed and S partially normally compact with respect to y around $(\bar{x},\bar{y})$ in the sense of Mordukhovich and Shao (see [25]).

LEMMA 3.1. *Let $S\colon X \to Y$ be a set-valued mapping with closed and convex graph, let C be a closed convex subset of X, let D be a closed convex subset of Y, and let $(\bar{x},\bar{y}) \in \mathrm{gr}S \cap (C \times D)$. Suppose that one of the following conditions holds:*

(i) *Either* $C \times D$ *or* $\mathrm{gr}S$ *is normally compact around* $(\bar{x}, \bar{y})$.
(ii) C *is normally compact around* $\bar{x}$ *and* S *is partially* ∂*-coderivatively compact at* $(\bar{x}, \bar{y})$.

Then $N_{\mathrm{gr}S}((\bar{x}, \bar{y})) \cap (-N_{C\times D}((\bar{x}, \bar{y}))) = \{(0, 0)\}$ *if and only if* $(0, 0) \in \mathrm{int}[\mathrm{gr}S - (C \times D)]$, *which implies that there exist* $\delta, \gamma > 0$ *such that*

$$d((x, y), (C \times D) \cap \mathrm{gr}S) \leq \gamma \max\{d((x, y), C \times D), d((x, y), \mathrm{gr}S\} \ \ \forall (x, y) \in B((\bar{x}, \bar{y}), \delta).$$

Proof. "$\Longrightarrow$" Denote by E the product space $X \times Y$ with the norm $\|u\| = \|x\| + \|y\|$ for every $u = (x, y) \in E$. Assume that $(0, 0) \notin \mathrm{int}[\mathrm{gr}S - (C \times D)]$. For an arbitrary positive sequence $\epsilon_k \to 0$ as $k \to \infty$, we can find sequence $\{\bar{u}_k\} = \{(\bar{x}_k, \bar{y}_k)\} \in E$ with $\|\bar{u}_k\| \leq \epsilon_k^2$ such that $\bar{u}_k \notin \mathrm{gr}S - (C \times D)$. Set $A = \mathrm{gr}S - \bar{u}_k$, $B = C \times D$, and $\bar{u} = (\bar{x}, \bar{y})$, and equip E with the norm $\|(u, v)\| = \|u\| + \|v\|$ for every $u, v \in E$. Applying Ekeland's variational principle to the function $f(u, v) := \|u - v\|$ on the complete metric space $A \times B$, since

$$f(\bar{u} - \bar{u}_k, \bar{u}) = \|\bar{u}_k\| \leq \inf\{f(u, v) \mid (u, v) \in A \times B\} + \epsilon_k^2,$$

there exists $(u_k, v_k) \in A \times B$ such that

$$\|u_k - \bar{u} + \bar{u}_k\| + \|v_k - \bar{u}\| \leq \epsilon_k \tag{6}$$

and

$$f(u_k, v_k) \leq f(u, v) + \epsilon_k(\|u - u_k\| + \|v - v_k\|) \ \ \forall (u, v) \in A \times B.$$

This implies that

$$(0, 0) \in \partial f(u_k, v_k) + \epsilon_k B_{E^*} \times B_{E^*} + N_A(u_k) \times N_B(v_k).$$

Hence there exist $w_k^* \in E^*$ with $\|w_k^*\| = 1$ and $\langle w_k^*, u_k - v_k\rangle = \|u_k - v_k\|$ such that

$$(0, 0) \in (w_k^*, -w_k^*) + \epsilon_k B_{E^*} \times B_{E^*} + N_A(u_k) \times N_B(v_k).$$

It follows that there exist $\{w_{ik}^*\} \subset E^*$, $i = 1, 2$, with $w_{1k}^* \in N_A(u_k)$, $w_{2k}^* \in N_B(v_k)$ such that

$$1 - \epsilon_k \leq \|w_{ik}^*\| \leq 1 + \epsilon_k, \ i = 1, 2, \qquad \text{and} \qquad \|w_{1k}^* + w_{2k}^*\| \leq 2\epsilon_k. \tag{7}$$

Let $\{(x_{1k}, y_{1k})\} \subset \mathrm{gr}S$, $\{(x_{2k}, y_{2k})\} \subset C$ be such that $u_k = (x_{1k}, y_{1k}) - \bar{u}_k$, $v_k = (x_{2k}, y_{2k})$ and let $\{(x_{ik}^*, y_{ik}^*)\} \subset E^*$ be such that $w_{ik}^* = (x_{ik}^*, y_{ik}^*)$, $i = 1, 2$. Then $(x_{1k}^*, y_{1k}^*) \in N_A(u_k) = N_{\mathrm{gr}S}((x_{1k}, y_{1k}))$ and $(x_{2k}^*, y_{2k}^*) \in N_C((x_{2k}, y_{2k}))$, and by (6) and (7), we get that

$$(x_{ik}, y_{ik}) \to (\bar{x}, \bar{y}) \ (i = 1, 2) \qquad \text{and} \qquad (x_{1k}^* + x_{2k}^*, y_{1k}^* + y_{2k}^*) \to (0, 0) \text{ as } k \to \infty. \tag{8}$$

Since $\|(x_{ik}^*, y_{ik}^*)\| \to 1$ as $k \to \infty$, $i = 1, 2$, and by taking into account that the closed unit ball B_{E^*} is weak-star compact, without of loss of generality we may assume that

$$(x_{ik}^*, y_{ik}^*) \xrightarrow{w^*} (x_i^*, y_i^*) \ \text{ as } k \to \infty, \quad i = 1, 2.$$

Thus

$$(x_1^*, y_1^*) \in N_{\mathrm{gr}S}((\bar{x}, \bar{y})), \ (x_2^*, y_2^*) \in N_{C\times D}((\bar{x}, \bar{y})), \text{ and } (x_1^* + x_2^*, y_1^* + y_2^*) = (0, 0).$$

Let us denote $(x^*, y^*) = (x_1^*, y_1^*) = -(x_2^*, y_2^*)$. To finish the proof it remains to show that $(x^*, y^*) \neq (0, 0)$. It is clear that this is true if (i) holds since $\|(x_{ik}^*, y_{ik}^*)\| \to 1$ as $k \to \infty$, $i = 1, 2$. Suppose that (ii) holds. If $\|x_{2k}^*\| \not\to 0$, then by the normal compactness of C around $\bar{x}$, we have $x^* = -x_2^* \neq 0$; if $\|x_{2k}^*\| \to 0$, noticing that $\|(x_{2k}^*, y_{2k}^*)\| \to 1$, then $\|y_{2k}^*\| \to 1$. It follows from (8) that $\|x_{1k}^*\| \to 0$ and $\|y_{1k}^*\| \to 1$. Since S is partially ∂-coderivatively compact at $(\bar{x}, \bar{y})$, we obtain $y^* = y_1^* \neq 0$.

"$\Longleftarrow$" Choose arbitrary $(x^*, y^*) \in N_{\mathrm{gr}S}(\bar{x}, \bar{y}) \cap (-N_{C\times D}(\bar{x}, \bar{y}))$. Then

$$\langle (x^*, y^*), (x_1 - \bar{x}, y_1 - \bar{y})\rangle \leq 0 \ \ \forall (x_1, y_1) \in \mathrm{gr}S,$$

$$\langle (x^*, y^*), (x_2 - \bar{x}, y_2 - \bar{y})\rangle \geq 0 \ \ \forall (x_2, y_2) \in C \times D.$$

In other words,

$$\langle (x^*, y^*), (x_1 - x_2, y_1 - y_2)\rangle \leq 0 \ \ \forall (x_1, y_1) \in \mathrm{gr}S, (x_2, y_2) \in C \times D.$$

However, by assumption $(0, 0) \in \mathrm{int}[\mathrm{gr}S - (C \times D)]$, one has that $(x^*, y^*) = (0, 0)$.

The last implication is well known. For instance, see Theorem 4.1 in [10] or Theorem 1 in [29]. □

The conclusion of Lemma 3.1 was proved by Henrion and Jourani [10] under the assumption that either $C \times D$ or $\mathrm{gr}S$ is compactly epi-Lipschitzian at $(\bar{x}, \bar{y})$.

THEOREM 3.2. *Consider the set-valued mapping $M\colon Y \rightrightarrows X$ defined as $M(y) = S(y) \cap C$, where $S\colon Y \rightrightarrows X$ is a set-valued mapping with closed convex graph and C is a closed convex set of X. Let $(\bar{y}, \bar{x}) \in \mathrm{gr}M$. Then the following statements hold:*

(i) *If there exist positive numbers γ, δ such that*

$$B_{X^*} \cap N_{M(\bar{y})}(u) \subset \gamma[B_{X^*} \cap N_C(u) + B_{X^*} \cap D^*S^{-1}(u, \bar{y})(B_{Y^*})] \tag{9}$$

for all $u \in B(\bar{x}, \delta) \cap M(\bar{y})$, then M has a local error bound at $(\bar{y}, \bar{x})$.

(ii) *If either C is normally compact around $\bar{x}$ and S^{-1} is partially ∂-coderivatively compact at $(\bar{x}, \bar{y})$ or $\mathrm{gr}S^{-1}$ is normally compact around $(\bar{x}, \bar{y})$, and if*

$$D^*S^{-1}(\bar{x}, \bar{y})(y^*) \cap (-\mathrm{bd}N_C(\bar{x})) = \begin{cases} \emptyset & \text{or} \\ \{0\} & \text{if } y^* = 0, \end{cases} \tag{10}$$

then $\{C \times \{\bar{y}\}, \mathrm{gr}S^{-1}\}$ is locally linearly regular around $(\bar{x}, \bar{y})$.

Proof. (i) Suppose (9) holds. Let $x \in B(\bar{x}, \frac{\delta}{3}) \setminus M(\bar{y})$ and $\sigma \in (0, 1)$. From the proof of Lemma 2.1, one sees that there exist $z \in \mathrm{bd}M(\bar{y}) \cap B(\bar{x}, \delta)$ and $x^* \in N_{M(\bar{y})}(z)$ with $\|x^*\| = 1$ such that

$$\sigma\|x - z\| \leq \langle x^*, x - z\rangle.$$

According to (9), there exist $u^* \in B_{X^*} \cap N_C(z)$, $y^* \in B_{Y^*}$, $v^* \in B_{X^*} \cap D^*S^{-1}(z, \bar{y})(y^*)$ such that $x^* = \gamma(u^* + v^*)$. It follows that

$$\sigma\|x - z\| \leq \gamma(\langle u^*, x - z\rangle + \langle v^*, x - z\rangle).$$

Noticing that $u^* \in \partial d(\cdot, C)(z)$ and $(v^*, -y^*) \in N_{\mathrm{gr}S^{-1}}(z, \bar{y})$, one has that

$$\langle u^*, x - z\rangle \leq d(x, C)$$

and

$$\langle v^*, x - z\rangle \le \langle y^*, y - \bar{y}\rangle \le \|y - \bar{y}\| \ \ \forall y \in S^{-1}(x).$$

Hence, we get that

$$\sigma d(x, M(\bar{y})) \le 2\gamma \max\{d(x, C), d(\bar{y}, S^{-1}(x))\}.$$

Letting $\sigma \to 1$, we obtain the desired result.

(ii) If $0 \notin -\mathrm{bd}N_C(\bar{x}))$, then $N_C(\bar{x}) = X^*$, which implies that $C = \{\bar{x}\}$. In this case, the conclusion is obvious. We assume that $0 \in -\mathrm{bd}N_C(\bar{x}))$. This, together with (10), implies that $D^*S^{-1}(\bar{x}, \bar{y})(0) \cap (-\mathrm{bd}N_C(\bar{x})) = \{0\}$ and that

$$N_{\mathrm{gr}S^{-1}}((\bar{x}, \bar{y})) \cap (-\mathrm{bd}N_{C\times\{\bar{y}\}}((\bar{x}, \bar{y}))) = \{(0, 0)\}.$$

If

$$N_{\mathrm{gr}S^{-1}}((\bar{x}, \bar{y})) \cap -\mathrm{int}(N_{C\times\{\bar{y}\}}((\bar{x}, \bar{y}))) \ne \emptyset, \tag{11}$$

then there exists $(x^*, y^*) \in N_{\mathrm{gr}S^{-1}}((\bar{x}, \bar{y}))$ such that $-B((x^*, y^*), \delta) \subset N_{C\times\{\bar{y}\}}((\bar{x}, \bar{y}))$ for some $\delta > 0$. In other words,

$$\langle -x^*, x - \bar{x}\rangle + \delta\|x - \bar{x}\| \le 0 \ \ \forall x \in C,$$

$$\langle x^*, x - \bar{x}\rangle + \langle y^*, y - \bar{y}\rangle \le 0 \ \ \forall (x, y) \in \mathrm{gr}S^{-1}.$$

Hence for any $x \in M(\bar{y})$, i.e., $(x, \bar{y}) \in \mathrm{gr}S^{-1} \cap C \times \{\bar{y}\}$, we have

$$\delta\|x - \bar{x}\| \le \langle -x^*, x - \bar{x}\rangle + \delta\|x - \bar{x}\| \le 0,$$

and hence $x = \bar{x}$. This shows that $M(\bar{y}) = \{\bar{x}\}$. Hence

$$N_{M(\bar{y})\times\{\bar{y}\}}((\bar{x}, \bar{y})) = X^* \times Y^*.$$

On the other hand, formula (11) implies

$$N_{\mathrm{gr}S^{-1}}((\bar{x}, \bar{y})) + N_{C\times\{\bar{y}\}}((\bar{x}, \bar{y})) = X^* \times Y^*.$$

Hence,

$$N_{\mathrm{gr}S^{-1}}((u, v)) + N_{C\times\{\bar{y}\}}((u, v)) = X^* \times Y^* = N_{M(\bar{y})\times\{\bar{y}\}}((u, v)) \tag{12}$$

for all $(u, v) \in (M(\bar{y}) \times \{\bar{y}\}) \cap B((\bar{x}, \bar{y}), \delta) = \{(\bar{x}, \bar{y})\}$, where δ is an arbitrary positive number.

By Proposition 4 in [16], (12) implies that there exists some $\gamma > 0$ such that

$$N_{M(\bar{y})\times\{\bar{y}\}}((u, v)) \cap B_{X^*\times Y^*} \subset \gamma(N_{C\times\{\bar{y}\}}((u, v)) \cap B_{X^*\times Y^*} + N_{\mathrm{gr}S^{-1}}((u, v)) \cap B_{X^*\times Y^*})$$

for all $(u, v) \in (M(\bar{y}) \times \{\bar{y}\}) \cap B((\bar{x}, \bar{y}), \delta)$. It follows from Theorem 2.1 that there exists some $\hat{\delta} > 0$ such that

$$d(x, M(\bar{y}) \le \gamma \max\{d((x, \bar{y}), C \times \{\bar{y}\}), d((x, \bar{y}), \mathrm{gr}S^{-1})\} \ \ \forall x \in B(\bar{x}, \hat{\delta}).$$

Hence, the conclusion follows from Proposition 3.1.

Otherwise, we have

$$N_{\mathrm{gr}S^{-1}}((\bar{x},\bar{y}))\cap -N_{C\times\{\bar{y}\}}((\bar{x},\bar{y}))=\{(0,0)\}.$$

Then the conclusion follows from Lemma 3.1. □

Remark 3. From the proof, it is easy to see that the hypotheses of (ii) imply (9). The second part of Theorem 3.2 generalizes Theorem 3.2 in [11] to the infinite-dimensional space in the case when the set C and the graph of the set-valued map S are convex.

When $C=X$, obviously, C is normally compact around $\bar{x}$ and condition (10) reduces to

$$\ker D^*S^{-1}(\bar{x},\bar{y})=\{0\},$$

or equivalently,

$$D^*S(\bar{y},\bar{x})(0)=\{0\}. \tag{13}$$

It has been proved that (see [18]) S is pseudo-Lipschitz at $(\bar{y},\bar{x})$ under the assumption of (ii) with $C=X$, which, in turn, implies the conclusion of (ii); when X and Y are finite-dimensional spaces, condition (13) is equivalent to the pseudo-Lipschitz property of S at $(\bar{y},\bar{x})$. The above-mentioned results are valid even in the nonconvex case.

Let $K\subset Y$ be a closed convex cone. A single-valued map $f\colon X\to Y$ is said to be K-convex if its epigraph

$$\mathrm{epi}_K f=\{(x,y)\mid f(x)\in y-K\}$$

is a convex set in $X\times Y$.

COROLLARY 3.1. *Let X, Y be Asplund spaces, let C be a closed convex subset of X, and let K be a closed convex cone of Y. Consider the set-valued maps $S,M\colon Y\rightrightarrows X$ defined as*

$$S(y)=\{x\in X\mid f(x)\in y-K\}$$

and $M(y)=C\cap S(y)$, where $f\colon X\to Y$ is a continuous K-convex map. Suppose that $(0,\bar{x})\in \mathrm{gr}M$ and that K is normally compact around $-f(\bar{x})$ (especially $\mathrm{int}K\neq\emptyset$). Then the following statements hold:

(i) *If there exist positive numbers γ,δ such that*

$$B_{X^*}\cap N_{M(0)}(u)\subset\gamma[B_{X^*}\cap N_C(u)+B_{X^*}\cap D^*(f+K)(u,0)(B_{Y^*})]$$

for all $u\in B(\bar{x},\delta)\cap M(0)$, then M has a local error bound at $(0,\bar{x})$.

(ii) *If either C is normally compact around $\bar{x}$ and $0\in\mathrm{int}(f(X)+K)$ or $\mathrm{epi}f$ is normally compact around $(\bar{x},f(\bar{x}))$, and if for every $y^*\in N_{-K}(f(\bar{x}))$,*

$$D^*f(\bar{x})(y^*)\cap(-\mathrm{bd}N_C(\bar{x}))=\begin{cases}\emptyset & \text{or}\\ \{0\} & \text{if } y^*=0,\end{cases}$$

then $\{C\times\{0\},\mathrm{gr}S^{-1}\}$ is locally linearly regular around $(\bar{x},0)$ (where the coderivative $D^*f(\bar{x})(y^*)$ is defined in the sense of limit Fréchet).

Proof. It is easy to see that $S^{-1}(x) = f(x) + K$ is a set-valued map with closed convex graph. (i) follows from Theorem 3.2 directly.

(ii) The case in which $0 \notin \mathrm{bd}N_C(\bar{x})$ is trivial, so assume that $0 \in \mathrm{bd}N_C(\bar{x})$. Define a mapping $g\colon X \times Y \to Y$ by

$$g(x, y) = f(x) - y.$$

Then $\mathrm{gr}S^{-1} = g^{-1}(-K)$. Applying the coderivative sum rule (see Theorem 3.5 in [24]), we get

$$D^*g(\bar{x}, 0)(y^*) = \{(x^*, -y^*) \mid x^* \in D^*f(\bar{x})(y^*)\}$$

(where the coderivative is defined in the sense of limit Fréchet). It is clear that

$$N_{-K}(g(\bar{x}, 0)) = N_{-K}(f(\bar{x})), \quad \ker D^*g(\bar{x}, 0) = \{0\}$$

and so that $N_{-K}(g(\bar{x}, 0)) \cap \ker D^*g(\bar{x}, 0) = \{0\}$. By Corollary 6.9 of [23], under the hypotheses of Corollary 3.1, we have

$$\begin{aligned} N_{\mathrm{gr}S^{-1}}(\bar{x}, 0) &\subset \bigcup \big\{D^*g(\bar{x}, 0))(y^*) \mid y^* \in N_{-K}(g(\bar{x}, 0)\big\} \\ &= \bigcup \big\{\{(x^*, -y^*) \mid x^* \in D^*f(\bar{x})(y^*)\} \mid y^* \in N_{-K}(f(\bar{x})\big\}. \end{aligned}$$

It follows that

$$D^*S^{-1}(\bar{x}, 0)(y^*) \subset \begin{cases} D^*f(\bar{x})(y^*) & \text{if } y^* \in N_{-K}(f(\bar{x})), \\ \emptyset & \text{otherwise.} \end{cases}$$

Thus, the conclusion follows from Theorem 3.2. □

COROLLARY 3.2. *Let X, Y be Banach spaces, let C be a closed convex subset of X, and let K be a closed convex cone of Y. Consider the set-valued maps $S, M\colon Y \rightrightarrows X$ defined as*

$$S(y) = \{x \in X \mid T(x) \in y - K\}$$

and $M(y) = C \cap S(y)$, where $T\colon X \to Y$ is a continuous linear operator. Let $(0, \bar{x}) \in \mathrm{gr}M$. Then the following statements hold:

(i) *M has a local error bound at $(0, \bar{x})$ if and only if there exist positive numbers γ, δ such that*

$$B_{X^*} \cap N_{M(0)}(u) \subset \gamma[B_{X^*} \cap N_C(u) + T^*(B_{Y^*} \cap N_{-K}(T(\bar{x})))] \tag{14}$$

for all $u \in B(\bar{x}, \delta) \cap M(0)$.

(ii) *If C is normally compact around $\bar{x}$ and $0 \in \mathrm{int}(T(X) + K)$, and if*

$$\left.\begin{array}{c} T^*(y^*) \in -\mathrm{bd}N_C(\bar{x}) \\ y^* \in N_{-K}(f(\bar{x})) \end{array}\right\} \Longrightarrow y^* = 0,$$

then $\{C \times \{\bar{y}\}, \mathrm{gr}S^{-1}\}$ is locally linearly regular around $(\bar{x}, 0)$.

Proof. (i) Noticing that $S^{-1}(x) = T(x) + K$, we have that $\mathrm{gr}S^{-1} = \mathrm{gr}T + \{0\} \times K$. It is easy to verify that

$$N_{\mathrm{gr}S^{-1}}(u, 0) = \bigcup \big\{(-T^*(y^*), y^*) \mid y^* \in N_K(-T(u))\big\} \ \ \forall u \in S^{-1}(0).$$

It follows that

$$D^*S^{-1}(u,0)(y^*) = \begin{cases} T^*(y^*) & \text{if } y^* \in N_{-K}(T(u)), \\ \emptyset & \text{otherwise.} \end{cases}$$

If condition (14) is satisfied, then

$$B_{X^*} \cap N_{M(0)}(u) \subset \tau(\|T^*\|+1)[B_{X^*} \cap N_C(u) + B_{X^*} \cap D^*S^{-1}(u,0)(B_{Y^*})]$$

for all $u \in B(\bar{x},\delta) \cap M(0)$. By Theorem 3.2, M has a local error bound at $(0,\bar{x})$. Conversely, suppose that M has a local error bound at $(0,\bar{x})$. Then there exist positive scalars γ and δ_1 such that

$$d(x, M(0)) \le \gamma(d(x,C) + d(T(x), -K)) \ \ \forall x \in B(\bar{x},\delta_1).$$

Hence

$$d(x, M(0)) \le \gamma(d(x,C) + d(T(x), -K)) + I_{\text{int}B(\bar{x},\delta_1)}(x) \ \ \forall x \in X.$$

As both functions on the two sides of the above inequality are 0 at each $u \in M(0) \cap \text{int}B(\bar{x},\delta_1)$, we obtain that

$$\begin{aligned} N_{M(0)}(u) \cap B_{X^*} &= \partial d(\cdot, M(0))(u) \subset \partial[\gamma(d(\cdot,C) + d(T(\cdot),-K)) + I_{\text{int}B(\bar{x},\delta_1)}(\cdot)](u) \\ &= \gamma[\partial d(\cdot,C)(u) + \partial d(T(\cdot),-K)(u)] + \partial I_{\text{int}B(\bar{x},\delta_1)}(u) \\ &= \gamma[B_{X^*} \cap N_C(u) + T^*(B_{Y^*} \cap N_{-K}(T(\bar{x})))]. \end{aligned}$$

By taking $\delta > 0$ such that $B(\bar{x},\delta) \subset \text{int}B(\bar{x},\delta_1)$, we get (14).

The assertion of (ii) follows directly from Theorem 3.2 and the calculation of $D^*S^{-1}(u,0)(y^*)$. □

Consider the set-valued map S defined by a convex inequality system, i.e.,

$$S(y) := \{x \in X \mid f(x) \le y\},$$

where $f\colon X \to \mathbb{R} \cup \{+\infty\}$ is a lower semicontinuous proper convex function. Let $\bar{x} \in M(0) = C \cap S(0)$. If $f(\bar{x}) < 0$, then it is easy to show that M is calm at $(0,\bar{x})$ by the Robinson–Ursescu theorem (see [27, 10]). In this special case, the conditions of Theorem 3.2 can be refined.

THEOREM 3.3. *Consider the set-valued maps $S, M\colon Y \rightrightarrows X$ defined as*

$$S(y) = \{x \in X \mid f(x) \le y\}$$

and $M(y) = C \cap S(y)$, where C is a closed convex subset of X and $f\colon X \to \mathbb{R} \cup \{+\infty\}$ is a lower semicontinuous proper convex function. Let $\bar{x} \in C$ be a given point such that $f(\bar{x}) = 0$. Consider the following conditions:

(a) *M has a local error bound at $(0,\bar{x})$.*

(b) *$\{C \times \{0\}, \text{epi}f\}$ is locally linearly regular around $(\bar{x},0)$.*

(c) *There exist $\delta,\gamma > 0$ such that for all $h \in X$, $u \in B(\bar{x},\delta) \cap M(0)$,*

$$d(h, T_{M(0)}(u)) \le \gamma \max\{d(h, T_C(u)), d((h,0), T_{\text{epi}f}((u,0)))\}.$$

(d) *There exist $\delta,\gamma > 0$ such that for all $u \in M(0) \cap B(\bar{x},\delta)$,*

$$(N_{M(0)}(u) \cap B_{X^*}) \times B_{\mathbb{R}} \subset \gamma[(N_C(u) \cap B_{X^*}) \times B_{\mathbb{R}} + N_{\text{epi}f}((u,0)) \cap B_{X^* \times \mathbb{R}}].$$

(e) *Either* (i) *or* (ii) *holds:*
(i) $\mathrm{bd}\partial f(\bar{x}) \cap -\mathrm{bd}N_C(\bar{x}) \neq \partial f(\bar{x}) \cap -N_C(\bar{x})$*;*
(ii) $\mathrm{bd}\partial f(\bar{x}) \cap (-\mathrm{bd}N_C(\bar{x})) = \emptyset$*,* $\partial^\infty f(\bar{x}) \cap -N_C(\bar{x}) = \{0\}$ *and either* C *is normally compact around* $\bar{x}$ *or* $\mathrm{epi}f$ *is normally compact around* $(\bar{x}, 0)$.

Then (e) $\Longrightarrow$ (d) $\Longleftrightarrow$ (c) $\Longleftrightarrow$ (b) $\Longrightarrow$ (a).

Proof. It is clear that conditions (c) and (d) can be rewritten as

$$\begin{aligned} d(h, T_{M(0)}(u)) &= d((h,0), T_{(C\times\{0\})\cap \mathrm{epi}f}((u,0))) \\ &\leq \gamma \max\{d((h,0), T_{C\times\{0\}}((u,0))), d((h,0), T_{\mathrm{epi}f}((u,0)))\} \end{aligned} \tag{15}$$

and

$$N_{M(0)\times\{0\}}((u,0)) \cap B_{X^*\times\mathbb{R}} \subset \gamma(N_{C\times\{0\}}((u,0)) \cap B_{X^*\times\mathbb{R}} + N_{\mathrm{epi}f}((u,0)) \cap B_{X^*\times\mathbb{R}}). \tag{16}$$

We can see from Theorem 2.1 that conditions (15) and (16) are equivalent to the condition that there exist $\delta, \gamma > 0$ such that

$$d((x,0), C\times\{0\} \cap \mathrm{epi}f) \leq \gamma \max\{d(x,C), d((x,0), \mathrm{epi}f)\} \ \ \forall x \in B(\bar{x}, \delta),$$

which is equivalent to (b) and implies (a) by Proposition 3.1 and Theorem 3.1.

In the following we shall prove that (e) implies (d). Suppose that (i) of (e) is satisfied. Then, since both $\partial f(\bar{x})$ and $N_C(\bar{x})$ are strongly closed in X^*, it follows that

$$\mathrm{int}\partial f(\bar{x}) \cap -N_C(\bar{x}) \neq \emptyset \ \text{ or } \partial f(\bar{x}) \cap -\mathrm{int}N_C(\bar{x}) \neq \emptyset. \tag{17}$$

We now show that $M(0) = \{\bar{x}\}$, i.e., $C\times\{0\} \cap \mathrm{epi}f = \{(\bar{x}, 0)\}$. Indeed, if the first condition of (17) holds, then there exist $x^* \in \mathrm{int}\partial f(\bar{x}) \cap -N_C(\bar{x})$ and $\delta > 0$ such that $B(x^*, \delta) \subset \partial f(\bar{x})$. In other words,

$$\begin{gathered} \delta\|x - \bar{x}\| + \langle x^*, x - \bar{x}\rangle \leq f(x) - f(\bar{x}) \ \ \forall x \in X, \\ \langle x^*, x - \bar{x}\rangle \geq 0 \ \ \forall x \in C. \end{gathered}$$

Hence for any x with $(x,0) \in C\times\{0\} \cap \mathrm{epi}f$, we have

$$\delta\|x - \bar{x}\| \leq \delta\|x - \bar{x}\| + \langle x^*, x - \bar{x}\rangle \leq f(x) - f(\bar{x}) \leq 0,$$

and hence $x = \bar{x}$. This shows our claim. If the second condition of (17) holds, we can prove our claim similarly. Therefore, $M(0) \cap B(\bar{x}, \delta) = \{\bar{x}\}$. On the other hand, the condition (17) yields that

$$0 \in \mathrm{int}(\partial f(\bar{x}) + N_C(\bar{x})).$$

Hence, there exists $\varepsilon > 0$ such that for any $x^* \in \varepsilon B_{X^*}$ there exist $x_1^* \in \partial f(\bar{x})$, $x_2^* \in N_C(\bar{x})$ such that $x^* = x_1^* + x_2^*$, and hence for every $r \in \mathbb{R}$ with $|r| \leq \varepsilon$,

$$(x^*, r) = (x_1^*, -1) + (x_2^*, r+1) \in N_{\mathrm{epi}f}(\bar{x}, 0) + N_{C\times\{0\}}(\bar{x}, 0).$$

This implies that

$$N_{\mathrm{epi}f}(\bar{x}, 0) + N_{C\times\{0\}}(\bar{x}, 0) = X^*\times\mathbb{R}.$$

Therefore, for every $u \in M(0) \cap B(\bar{x}, \delta)$, we have

$$N_{\mathrm{epi}f}(u, 0) + N_C(u)\times\mathbb{R} = X^*\times\mathbb{R} = N_{M(0)}(u)\times\mathbb{R}. \tag{18}$$

By Proposition 4 in [16], (18) implies that there exists some $\gamma > 0$ such that

$$(N_{M(0)}(u) \cap B_{X^*}) \times B_{\mathbb{R}} \subset \gamma((N_C(u) \cap B_{X^*}) \times B_{\mathbb{R}} + N_{\mathrm{epi}f}((u,0)) \cap B_{X^* \times \mathbb{R}}).$$

Finally, assume that (ii) of (e) holds. If $0 \in \mathrm{int}\partial f(\bar{x})$, then $0 \in \mathrm{int}\partial f(\bar{x}) \cap -N_C(\bar{x})$. This means that the first case of (17) is satisfied and that the conclusion follows. Suppose that $0 \in \mathrm{bd}\partial f(\bar{x})$. If $0 \in -\mathrm{int}N_C(\bar{x})$, the case reduces to the second case of (17); if $0 \in -\mathrm{bd}N_C(\bar{x})$, then it leads to a contradiction. It remains to check the case of $0 \notin \partial f(\bar{x})$. Then one has

$$\partial f(\bar{x}) \cap -N_C(\bar{x}) = \emptyset \ \text{ or } \partial f(\bar{x}) \subset -\mathrm{int}N_C(\bar{x}). \tag{19}$$

Suppose that the second case of (19) holds. If $\partial f(\bar{x}) = \emptyset$, then we are back to the first case of (19). Hence, assume that $\partial f(\bar{x}) \neq \emptyset$. Then the second case of (19), along with (ii), yields (i).

We now consider the first case of (19). We claim, in this case, that

$$N_{\mathrm{epi}f}((\bar{x},0)) \cap (-N_{C\times\{0\}}((\bar{x},0))) = \{(0,0)\}. \tag{20}$$

Indeed, $\partial^\infty f(\bar{x}) \cap (-N_C(\bar{x})) = \{0\}$ is equivalent to $N_{\mathrm{epi}f}(\bar{x},0) \cap (-N_C(\bar{x}) \times \{0\}) = \{(0,0)\}$; $\partial f(\bar{x}) \cap (-N_C(\bar{x})) = \emptyset$ is equivalent to $N_{\mathrm{epi}f}(\bar{x},0) \cap -(N_C(\bar{x}) \times (0,+\infty)) = \emptyset$. Let $(x^*, r) \in N_{\mathrm{epi}f}(\bar{x},0) \cap (-N_{C\times\{0\}}(\bar{x},0))$. Then $x^* \in -N_C(\bar{x})$ and $r \geq 0$, and then

$$\langle x^*, x - \bar{x}\rangle + r(f(x) + \varepsilon) \leq 0 \ \ \forall x \in \mathrm{dom}f, \varepsilon > 0.$$

Taking $x = \bar{x}$ in the above inequality, one deduces that $r = 0$. This, together with $\partial^\infty f(\bar{x}) \cap -N_C(\bar{x}) = \{0\}$, implies that $x^* = 0$. Hence, our claim is proved and hence (d) is true by Lemma 3.1. □

It has been proved by Henrion and Jourani [10, Theorem 3.3] that M is calm at $(0, \bar{x})$ if condition (e) is satisfied.

The following example in [10] will show that (d) does not imply (e), in general.

Example. Let $X = C = \mathbb{R}$, $\bar{x} = 0$, and $f(x) = \max\{x, 0\}$. Then

$$M(y) = \begin{cases} \emptyset & \text{if } y < 0, \\ (-\infty, y] & \text{if } y \geq 0. \end{cases}$$

Hence $\bar{x} \in M(0) = -\mathbb{R}_+$. It is obvious that

$$N_{M(0)}(\bar{x}) = \mathbb{R}_+, \quad N_C(\bar{x}) = \{0\}, \quad \text{and} \quad N_{\mathrm{epi}f}((\bar{x},0)) = \{(r_1, r_2) \mid r_2 \leq -r_1, r_1 \geq 0\}.$$

It is easy to verify that

$$N_{M(0)}(\bar{x}) \times \mathbb{R} = N_C(\bar{x}) \times \mathbb{R} + N_{\mathrm{epi}f}((\bar{x},0)) = \mathbb{R}_+ \times \mathbb{R}$$

and

$$(N_{M(0)}(\bar{x}) \cap [-1,1]) \times [-1,1] \subset 2((N_C(\bar{x}) \cap [-1,1]) \times [-1,1] + N_{\mathrm{epi}f}((\bar{x},0)) \cap ([-1,1] \times [-1,1])).$$

Let δ be an arbitrary positive number. For every $u \in M(0) \cap B(\bar{x}, \delta)$ with $u \neq \bar{x}$, we have

$$N_{M(0)}(u) = \{0\}, \quad N_C(u) = \{0\}, \quad \text{and} \quad N_{\mathrm{epi}f}((u,0)) = \{(0, r_2) \mid r_2 \leq 0\}.$$

Hence

$$N_{M(0)}(u) \times \mathbb{R} = N_C(u) \times \mathbb{R} + N_{\mathrm{epi}f}((u,0)) = \{0\} \times \mathbb{R}$$

and

$$(N_{M(0)}(u) \cap [-1,1]) \times [-1,1] \subset ((N_C(u) \cap [-1,1]) \times [-1,1] + N_{\mathrm{epi}f}((u,0)) \cap ([-1,1] \times [-1,1])).$$

Therefore condition (d) in Theorem 3.3 holds true. On the other hand, since $\partial f(\bar{x}) = [0,1]$ and $N_C(\bar{x}) = \{0\}$, we have $\mathrm{bd}\partial f(\bar{x}) \cap -\mathrm{bd}N_C(\bar{x}) = \partial f(\bar{x}) \cap -N_C(\bar{x}) = \{0\}$. This shows that conditions (i) and (ii) of (e) do not hold.

Acknowledgment. The author would like to thank professor L. Thibault and an anonymous referee for their helpful comments on improving the first version of this paper.

REFERENCES

[1] J. P. Aubin and I. Ekeland, *Applied Nonlinear Analysis*, Wiley-Interscience, New York, 1984.
[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, J. M. Borwein, and W. Li, *Strong conical hull intersection property, bounded linear regularity, Jameson's property (G), and error bounds in convex optimization*, Math. Program. Ser. A, 86 (1999), pp. 135–160.
[4] J. M. Borwein and H. M. Strojwas, *Tangential approximations*, Nonlinear Anal., 9 (1985), pp. 1347–1366.
[5] J. V. Burke, *Calmness and exact penalization*, SIAM J. Control Optim., 29 (1991), pp. 493–497.
[6] J. Burke, M. C. Ferris, and M. Qian, *On the Clarke subdifferential of the distance function of a closed set*, J. Math. Anal. Appl., 166 (1992), pp. 199–213.
[7] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.
[8] A. L. Dontchev and R. T. Rockafellar, *Regularity and conditioning of solution mappings in variational analysis*, Set-Valued Anal., 12 (2004), pp. 79–109.
[9] R. Henrion and J. Outrata, *A subdifferential condition for calmness of multifunctions*, J. Math. Anal. Appl., 258 (2001), pp. 110–130.
[10] R. Henrion and A. Jourani, *Subdifferential conditions for calmness of convex constraints*, SIAM J. Optim., 13 (2002), pp. 520–534.
[11] R. Henrion, A. Jourani, and J. Outrata, *On the calmness of a class of multifunctions*, SIAM J. Optim., 13 (2002), pp. 603–618.
[12] R. B. Holmes, *Geometric Functional Analysis and Its Applications*, Springer-Verlag, New York, 1975.
[13] A. Ioffe, *Regular points of Lipschitz functions*, Trans. Amer. Math. Soc., 251 (1975), pp. 61–69.
[14] A. D. Ioffe, *Necessary and sufficient conditions for a local minimum,* I: *A reduction theorem and first order conditions*, SIAM J. Control Optim., 17 (1979), pp. 245–250.
[15] A. D. Ioffe, *Metric regularity and subdifferential calculus*, Russian Math. Surveys, 55 (2000), pp. 501–558.
[16] G. J. O. Jameson, *The duality of pairs of wedges*, Proc. London Math. Soc., 249 (1972), pp. 531–547.
[17] A. Jourani, *Intersection formulae and the marginal function in Banach spaces*, J. Math. Anal. Appl., 192 (1995), pp. 867–891.
[18] A. Jourani and L. Thibault, *Coderivatives of multivalued mappings, locally compact cones and metric regularity*, Nonlinear Anal., 35 (1999), pp. 925–945.
[19] D. Klatte and B. Kummer, *Constrained minima and Lipschitzian penalties in metric spaces*, SIAM J. Optim., 13 (2002), pp. 619–633.
[20] A. S. Lewis and J. S. Pang, *Error bounds for convex inequality systems*, in Proceedings of the 5th International Symposium on Generalized Convexity, Luminy, 1996, J.-P. Crouzex, J.-E. Martinez-Legaz, and M. Volle, eds., Kluwer Academic, Dordrecht, The Netherlands, 1998, pp. 75–110.

[21] W. Li and I. Singer, *Global error bounds for convex multifunctions and applications*, Math. Oper. Res., 23 (1998), pp. 443–462.
[22] P. D. Loewen, *Limits of Fréchet normals in nonsmooth analysis*, in Optimization and Nonlinear Analysis, A. D. Ioffe et al., eds., Pitman Res. Notes Math. Ser. 244, Longman Scientific & Technical, Harlow, UK, 1992, pp. 178–188.
[23] B. S. Mordukhovich and Y. Shao, *Nonsmooth sequential analysis in Asplund spaces*, Trans. Amer. Math. Soc., 348 (1996), pp. 1235–1280.
[24] B. S. Mordukhovich and Y. Shao, *Nonconvex differential calculus for infinite-dimensional multifunctions*, Set-Valued Anal., 4 (1996), pp. 205–236.
[25] 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.
[26] K. F. Ng and W. H. Yang, *Error bounds for abstract linear inequality systems*, SIAM J. Optim., 13 (2002), pp. 24–43.
[27] S. Robinson, *Regularity and stability for convex multivalued functions*, Math. Oper. Res., 1 (1976), pp. 130–143.
[28] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
[29] W. Song and R. Zang, *Bounded linear regularity of convex sets in Banach spaces*, Math. Program., 106 (2006), pp. 59–79.

SIAM J. OPTIM.
Vol. 17, No. 2, pp. 372–400
© 2006 Society for Industrial and Applied Mathematics

CONVERGENT LAGRANGIAN AND CONTOUR CUT METHOD FOR NONLINEAR INTEGER PROGRAMMING WITH A QUADRATIC OBJECTIVE FUNCTION*

D. LI†, X. L. SUN‡, AND F. L. WANG‡

Abstract. In this paper we present an efficient exact solution method for solving nonlinear separable integer programming problems with a quadratic objective function. The proposed method combines the Lagrangian dual method with a duality reduction scheme using contour cut. At each iteration of the algorithm, lower and upper bounds of the problem are determined by the Lagrangian dual search. To eliminate the duality gap, a novel cut-and-partition scheme is derived by exploring the special structure of the quadratic contour. The method finds an exact solution of the problem in a finite number of iterations. Computational results are reported for problems with up to 2000 integer variables. Comparison results with other methods are also presented.

Key words. nonlinear integer programming, quadratic integer programming, Lagrangian relaxation, duality theory, objective contour cut

AMS subject classifications. 90C10, 90C26, 90C46

DOI. 10.1137/040606193

1. Introduction.

Consider the following nonlinear integer programming problem with a quadratic objective function:

$$
\begin{aligned}
(P) \qquad \min\ & q(x) = \sum_{j=1}^{n} \left(\frac{1}{2} c_j x_j^2 + d_j x_j \right) \\
\text{s.t.}\ & g_i(x) = \sum_{j=1}^{n} g_{ij}(x_j) \le b_i,\ i = 1, \ldots, m, \\
& x \in X = \{ x \mid l_j \le x_j \le u_j,\ x_j \text{ integer},\ j = 1, \ldots, n \},
\end{aligned}
$$

where g_{ij}'s are continuous functions and l_j and u_j are integer lower and upper bounds of x_j for $j = 1, \ldots, n$. Two cases of quadratic objective functions are considered first in this paper: (a) $q(x)$ is a convex function, i.e., $c_j > 0$ for $j = 1, \ldots, n$, and (b) $q(x)$ is a concave function, i.e., $c_j < 0$ for $j = 1, \ldots, n$. Problems with an indefinite quadratic objective function will be considered later in the paper as an extension.

Integer programming models with a convex quadratic objective function have various applications, including capital budgeting [27], [36], capacity planning [9], and optimization problems from graph theory [2], [26]. An important class of applications of problem (P) arises in portfolio selection models with discrete features (see [1], [6], [23], [31]). It was shown in [41], [42] that the Markowitz mean-variance model [33] can be simplified to a separable problem formulation of (P) by using market indices

*Received by the editors April 2, 2004; accepted for publication (in revised form) January 19, 2006; published electronically May 19, 2006. This work was supported by Research Grants Council of Hong Kong grant CUHK 4214/01E and National Natural Science Foundation of China grant 10571116.

http://www.siam.org/journals/siopt/17-2/60619.html

†Corresponding author. Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, N. T., Hong Kong (dli@se.cuhk.edu.hk).

‡Department of Mathematics, Shanghai University, Baoshan 200444, Shanghai, China (xlsun@staff.shu.edu.cn, flwang@graduate.shu.edu.cn).

together with some additional variables and constraints. A method for reformulating general nonlinear programs to separable forms is discussed in [37].

Concave quadratic cost functions are often encountered in real-world integer programming models involving economies of scale (see [19], [39]), which corresponds to the economic phenomenon of "decreasing marginal cost." The continuous version of problem (P) with $q(x)$ being concave and $g_i(x)$ linear or convex quadratic has been extensively studied (see, e.g., [7], [14], [39], [24], [43], [48]) and has been considered the standard test problem in concave minimization. Its methods exploit the special structures of quadratic functions and the extreme point property of concave programming in which the minimum of a concave function over a polyhedron is always achieved at one of its extreme points. There are, however, few methods in the literature for concave integer programming. Branch-and-bound methods based on continuous relaxation and convex underestimating were proposed in [4], [5], [8], [9], [11] for solving concave integer problems over a polyhedron.

Existing methods in the literature for nonlinear separable integer programming problems are mainly dynamic programming–based methods and continuous relaxation–based branch-and-bound methods. When an integer programming problem is separable, dynamic programming can be used to find its optimal solution (see [12], [13], [25]). Dynamic programming, however, suffers from the curse of dimensionality when m is large. Marsten and Morin [34] proposed a method that uses a dynamic programming technique to generate efficient feasible solutions and prunes nonpromising incomplete feasible solutions by a branch-and-bound strategy. Various branch-and-bound methods and their combination with dynamic programming were proposed for nonlinear knapsack-type problems (see [9], [10], [22], [35], [46])—in particular, convex quadratic knapsack problems [15], [36]. To guarantee the global optimality, the continuous relaxation–based branch-and-bound methods often require the convexity of all q and g_i's.

The Lagrangian dual method has been a powerful method in dealing with discrete optimization problems. The separability of the primal problem, when it exists, allows one to solve the Lagrangian relaxation by decomposition, thus searching efficiently for the dual optimal solution. The optimal dual value provides a lower bound for the optimal value of the objective function of the primal problem. The Lagrangian dual method, however, does not provide an optimal solution or even a feasible solution to the primal problem, in most situations, due to the existence of a duality gap (see [3], [17], [18], [20], [40]). Nonlinear Lagrangian formulations and convexification methods were proposed as attempts to eliminate the duality gap [29], [32], [44], [45]. In spite of a theoretical advantage of achieving a zero duality gap, certain computational issues in the implementation of the nonlinear Lagrangian dual methods remain unsolved due to the destruction of the separability in the primal problem when adopting a nonlinear Lagrangian formulation.

In this paper we develop a new exact method for problem (P). The framework of the proposed algorithm is a combination of the Lagrangian dual method and an objective contour cut-and-partition approach. The key motivation behind our method is an observation that cutting certain integer boxes inside and outside the ellipsoids formed by the objective contours still retains the optimal solution to (P) in the revised domain. This in turn leads to successive reductions of the duality gap of the primal problem. At each iteration, the algorithm first finds a lower bound of the problem using the Lagrangian dual search. To reduce the duality gap, a novel contour cut method is used to remove certain integer boxes that do not contain any feasible solution better than the incumbent. The revised integer domain is then partitioned

into a union of integer subboxes to facilitate the dual search on the revised integer domain in the next iteration. Numerical results show that the proposed algorithm is efficient and robust in solving large-scale instances of (P) with up to 2000 integer variables.

The paper is organized as follows. In section 2, we first introduce some preliminary results in Lagrangian duality theory for general singly constrained integer programming. New solution properties of the Lagrangian relaxation problem are presented. An exact dual search procedure is also described in section 2. In section 3, the quadratic contour cuts and the partition scheme are derived. We motivate the algorithm in section 4 by a small numerical example. The algorithm for singly constrained problem (P) is then formally described. Extensions to problems with multiple constraints and problems with an indefinite quadratic objective function are discussed in sections 5 and 6, respectively. Computational results are reported in section 7 for problems with different types of objective functions and constraint functions. Comparison results with other existing methods are also presented in section 7. Finally, a short concluding remark is given in section 8.

2. Lagrangian duality and dual search. In this section, we present some basic properties of the Lagrangian dual for general singly constrained integer programming. The relationship between the perturbation function and the Lagrangian dual is established. Solution properties of the Lagrangian relaxation problem are also derived. Furthermore, we will describe an exact dual search scheme specifically for singly constrained integer programming.

2.1. Lagrangian dual. Consider the following singly constrained integer program:

$$(P_1) \quad \min_{x\in X} \{f(x) \mid g(x) \le b\}, \tag{2.1}$$

where f and g are continuous functions on $\mathbb{R}^n$ and X is an arbitrary finite integer set. The Lagrangian relaxation of (P_1) is

$$(L_\lambda) \quad d(\lambda) = \min_{x\in X} L(x,\lambda), \tag{2.2}$$

where

$$L(x,\lambda) = f(x) + \lambda(g(x) - b)), \quad \lambda \ge 0. \tag{2.3}$$

Let

$$S = \{x \in X \mid g(x) \le b\},$$
$$f^* = \min_{x\in S} f(x).$$

The following weak duality holds:

$$d(\lambda) \le f(x) \quad \forall x \in S, \quad \lambda \ge 0. \tag{2.4}$$

Therefore $d(\lambda)$ always provides a lower bound for f^*. The Lagrangian dual problem of (P_1) is

$$(D) \quad \max_{\lambda\ge 0} d(\lambda). \tag{2.5}$$

Let λ^* be the optimal solution to (D). The nonnegative constant $f^* - d(\lambda^*)$ is called the *duality gap* of the problem and $f(x) - d(\lambda^*)$ is called a *duality bound* for any $x \in S$.

We assume in the following that $S \neq \emptyset$. Define the perturbation function of (P_1) as

$$w(y) = \min_{x \in X} \{f(x) \mid g(x) \leq y\}, \quad y \in \mathbb{R}. \tag{2.6}$$

The domain of w is $Y = [\tau, +\infty)$, where $\tau = \min_{x \in X} g(x)$. It is easy to see that w is a nonincreasing function on Y and it is continuous from the right. Since X is finite, there exists a finite number of points $a_i \in [\tau, +\infty)$ $(i = 1, \ldots, K)$ such that w can be expressed as

$$w(y) = f_i \quad \text{for } a_i \leq y < a_{i+1}, \quad i = 1, \ldots, K, \tag{2.7}$$

where $\tau = a_1 < a_2 < \cdots < a_K < a_{K+1} = +\infty$, and $f_1 > f_2 > \cdots > f_K$. By the assumption that $S \neq \emptyset$, we have $a_1 \leq b$. If we further assume that $X \setminus S \neq \emptyset$ and $\min_{x \in X} f(x) < f^*$, we then have $b < a_K$. Let

$$\Phi = \{(a_i, f_i) \mid i = 1, \ldots, K\}.$$

A point in Φ is called a *corner point* of $w(y)$.

Define the convex envelope function of w to be the maximum convex function underestimating w:

$$\psi(y) = \max\{h(y) \mid h \text{ is convex on } Y,\ h(\tilde{y}) \leq w(\tilde{y}) \quad \forall \tilde{y} \in Y\}. \tag{2.8}$$

It is easy to see that ψ is a nonincreasing piecewise linear functions on Y. We have

$$\psi(y) = \max_{\lambda, r \in \mathbb{R}} \{\lambda y + r \mid \lambda \tilde{y} + r \leq w(\tilde{y}) \quad \forall \tilde{y} \in Y\},$$

or equivalently,

$$\begin{aligned} \psi(y) = \max_{\lambda \in \mathbb{R}_-, r \in \mathbb{R}} & (\lambda y + r) \\ \text{s.t. } & \lambda a_i + r \leq f_i,\ i = 1, \ldots, K. \end{aligned} \tag{2.9}$$

For any fixed $y \in Y$, a dual variable $\mu_i \geq 0$ is introduced for each constraint $\lambda a_i + r \leq f_i,\ i = 1, \ldots, K$. Dualizing the linear program (2.9) yields

$$\psi(y) = \min \left\{ \sum_{i=1}^{K} \mu_i f_i \,\middle|\, \sum_{i=1}^{K} \mu_i a_i \leq y,\ \sum_{i=1}^{K} \mu_i = 1,\ \mu_i \geq 0,\ i = 1, \ldots, K \right\}. \tag{2.10}$$

The perturbation function characterizes the duality by the following theorem.

THEOREM 1. *Let $(-\lambda^*, r^*)$ and μ^* be optimal solutions to* (2.9) *and* (2.10) *with $y = b$, respectively. Then*

(i) *λ^* is an optimal solution to the dual problem (D) and*

$$\psi(b) = \max_{\lambda \geq 0} d(\lambda) = d(\lambda^*). \tag{2.11}$$

(ii) *for each i with $\mu_i^* > 0$, any $\bar{x} \in X$ satisfying $(g(\bar{x}), f(\bar{x})) = (a_i, f_i)$ is an optimal solution to the Lagrangian problem (L_{λ^*}).*

Proof. See [28], [32], and [30]. □

The following results investigate the primal feasibility and infeasibility of the solutions to the Lagrangian relaxation.

THEOREM 2. *Assume that $S \neq \emptyset$ and $X \setminus S \neq \emptyset$.*

(i) *Let λ^* be an optimal solution to (D). Then, there exists an $\tilde{x} \in S$ that solves (L_{λ^*}). Moreover, if $d(\lambda^*) < f^*$, then there exists a $\tilde{y} \in X \setminus S$ that solves (L_{λ^*}).*

(ii) *If for some $\lambda^* \geq 0$ there exist an $\tilde{x} \in S$ and a $\tilde{y} \in X \setminus S$ that both solve (L_{λ^*}), then λ^* must be an optimal solution to the dual problem (D).*

Proof. (i) By (2.10), there exist $\mu_i^* \geq 0$, $i = 1, \ldots, K$, that solve the following problem:

$$\begin{aligned} \psi(b) = \min & \sum_{i=1}^{K} \mu_i f_i, \\ \text{s.t.} & \sum_{i=1}^{K} \mu_i a_i \leq b, \\ & \sum_{i=1}^{K} \mu_i = 1, \quad \mu_i \geq 0, \quad i = 1, \ldots, K. \end{aligned} \tag{2.12}$$

Let $I = \{i \mid \mu_i^* > 0\}$. It follows from (2.12) that

$$\sum_{i \in I} \mu_i^* (a_i - b) \leq 0. \tag{2.13}$$

This implies that there exists at least one $i \in I$ such that $a_i \leq b$. Let $\tilde{x}$ be such that $(g(\tilde{x}), f(\tilde{x})) = (a_i, f_i)$. Then $\tilde{x} \in S$, and by Theorem 1(ii), $\tilde{x}$ solves (L_{λ^*}).

Suppose that $a_i \leq b_i$ for all $i \in I$. If $\sum_{i=1}^{K} \mu_i^* a_i = b$, then $a_i = b$ for all $i \in I$. Hence I is a singleton and $\mu_i^* = 1$, $i \in I$. By Theorem 1(i) and (2.12), we have

$$d(\lambda^*) = \psi(b) = f_i = w(a_i) = w(b) = f^*, \quad i \in I,$$

which contradicts the assumption of $d(\lambda^*) < f^*$. If $\sum_{i=1}^{K} \mu_i^* a_i < b$, then we claim that there do not exist $k, l \in I$ such that $f_k \neq f_l$. Otherwise, suppose that $f_k > f_l$, $k \neq l$. Define $\tilde{\mu} = (\tilde{\mu}_1, \ldots, \tilde{\mu}_K)$ as follows: $\tilde{\mu}_i = \mu_i^*$ if $i \neq k$ and $i \neq l$; $\tilde{\mu}_k = \mu_k^* - \epsilon$, $\tilde{\mu}_l = \mu_l^* + \epsilon$, where $\epsilon > 0$. We can always choose a sufficiently small $\epsilon > 0$ such that $\tilde{\mu}_k > 0$ and $\sum_{i \in I} \tilde{\mu}_i a_i < b$. Thus $\tilde{\mu}$ is feasible for problem (2.12). However, we have

$$\sum_{i \in I} \tilde{\mu}_i f_i = \sum_{i \in I} \mu_i^* f_i + \epsilon(f_l - f_k) < \sum_{i \in I} \mu_i^* f_i,$$

which contradicts that μ^* is an optimal solution to (2.12). Therefore, $f_k = f_l$ for any $k, l \in I$. It then follows that $\psi(b) = f_i$ for any $i \in I$. Also, since $a_i \leq b$ for $i \in I$, $w(a_i) \geq w(b)$. Thus, by Theorem 1(i), we have

$$d(\lambda^*) = \psi(b) = f_i = w(a_i) \geq w(b) = f^*, \quad i \in I,$$

which contradicts the assumption of nonzero duality gap.

The above arguments conclude that there exists an $i \in I$ such that $a_i > b$. Let $\tilde{y}$ be such that $(g(\tilde{y}), f(\tilde{y})) = (a_i, f_i)$. Then $\tilde{y}$ is infeasible, and by Theorem 1(ii), $\tilde{y}$ solves (L_{λ^*}).

(ii) For any $\lambda \geq 0$, if $\lambda^* > \lambda$, then

$$d(\lambda) \leq L(\tilde{y}, \lambda) = f(\tilde{y}) + \lambda(g(\tilde{y}) - b) < f(\tilde{y}) + \lambda^*(g(\tilde{y}) - b) = L(\tilde{y}, \lambda^*) = d(\lambda^*).$$

If $\lambda^* < \lambda$, then

$$d(\lambda) \leq L(\tilde{x}, \lambda) = f(\tilde{x}) + \lambda(g(\tilde{x}) - b) \leq f(\tilde{x}) + \lambda^*(g(\tilde{x}) - b) = L(\tilde{x}, \lambda^*) = d(\lambda^*).$$

Thus λ^* is an optimal solution to (D). □

THEOREM 3. *If the dual optimal solution $\lambda^* = 0$, then any feasible solution to (L_{λ^*}) is an optimal solution to (P_1) and $f^* = d(\lambda^*)$. Conversely, if there is a feasible solution x^* in the optimal solution set of (L_λ) with $\lambda = 0$, then $\lambda = 0$ is an optimal solution to (D) and x^* is an optimal solution to (P_1) with $f(x^*) = d(\lambda^*)$.*

Proof. Let x^* be a feasible solution to (L_{λ^*}) with $\lambda^* = 0$. Since

$$f(x^*) = \min_{x \in X} L(x, 0) = \min_{x \in X} f(x) \leq \min_{x \in S} f(x) = f^*, \tag{2.14}$$

we imply that x^* is optimal to (P_1) and $f(x^*) = f^* = d(\lambda^*)$. Conversely, if x^* solves (L_λ) with $\lambda = 0$ and is feasible for (P_1), then x^* must be optimal to (P_1). Moreover, by weak duality, we have $d(\lambda) \leq f(x^*) = d(0)$ for all $\lambda \geq 0$. Thus $\lambda = 0$ is the dual optimal solution. □

2.2. Dual search scheme. Motivated by the relationship between the convex envelope of the perturbation function and the optimal dual value, we can derive an exact dual search solution procedure for (D). Geometrically, the procedure visits the corner points of the perturbation function $w(y)$ at each iteration and eventually determines the optimal Lagrangian multiplier λ^*, where $-\lambda^*$ is exactly the slope of the line segment in the graph of $\psi(y)$ that intersects line $y = b$. The procedure starts by finding the corner point (a_1, f_1) and the corner point (a_K, f_K) in Φ. At each iteration, the Lagrangian relaxation (L_λ) is solved, where $-\lambda$ is the slope of the line connecting two corner points that correspond to the feasible incumbent solution and an infeasible solution with the least violation of the constraint up to the current stage, respectively. The distance between the feasible incumbent solution and the infeasible solution with the least constraint violation reduces monotonically in the iteration process. The algorithm terminates when a feasible optimal solution and an infeasible optimal solution to (L_λ) are found simultaneously or when $\lambda = 0$. By Theorems 2 and 3, an optimal dual solution is achieved when the algorithm terminates.

PROCEDURE 1 (dual search procedure).

Step 1. Calculate

$$x^0 = \arg\min_{x \in X} \ g(x), \ y^0 = \arg\min_{x \in X} \ f(x).$$

(i) If $g(x^0) > b$, stop. Problem (P_1) has no feasible solution.
(ii) If $g(y^0) \leq b$, stop. We conclude that y^0 is an optimal solution to (P_1) and $\lambda^* = 0$ is the optimal solution to (D).
(iii) Let $f_0^- = f(x^0)$, $g_0^- = g(x^0)$, $f_0^+ = f(y^0)$, $g_0^+ = g(y^0)$. Set $k = 0$.

Step 2. Compute

$$\lambda_k = -\frac{f_k^+ - f_k^-}{g_k^+ - g_k^-}.$$

Step 3. Solve (L_{λ_k}). Let x^k and y^k be the optimal solutions to (L_{λ_k}) with a minimum value of g and a maximum value of g, respectively.

(i) If $g(x^k) \le b < g(y^k)$, set $\tilde{x} = x^k$ and $\tilde{y} = y^k$, stop. $\lambda^* = \lambda^k$ is the optimal solution to the dual problem (D).

(ii) If $g(y^k) \le b$, then set

$$f_{k+1}^- = f(y^k),\ f_{k+1}^+ = f_k^+,$$
$$g_{k+1}^- = g(y^k),\ g_{k+1}^+ = g_k^+.$$

(iii) If $g(x^k) > b$, then set

$$f_{k+1}^- = f_k^-,\ f_{k+1}^+ = f(x^k),$$
$$g_{k+1}^- = g_k^-,\ g_{k+1}^+ = g(x^k).$$

Set $k := k + 1$. Return to Step 2.

THEOREM 4. *Procedure* 1 *stops at an optimal solution to* (D) *within a finite number of iterations.*

Proof. The proof of the finite termination of the procedure is similar to the one in [29]. The optimality of λ^* in Step 1(ii) and Step 3(i) follows from Theorems 2 and 3. □

It is clear that if Procedure 1 does not stop at Step 1, then the procedure stops at Step 3(i) and generates a lower bound $d(\lambda^*)$ to (P_1), a feasible solution $\tilde{x}$, and an infeasible solution $\tilde{y}$, where both $\tilde{x}$ and $\tilde{y}$ solve (L_{λ^*}).

The efficiency of the dual search procedure depends on whether or not the Lagrangian relaxation problem (L_λ) can be easily solved. Let $q_j(x_j) = (1/2)c_j x_j^2 + d_j x_j$ for $j = 1, \ldots, n$. Consider the singly constrained case of (P):

$$(P_s) \qquad \min\ q(x) = \sum_{j=1}^{n} q_j(x_j)$$
$$\text{s.t.}\ \ g(x) = \sum_{j=1}^{n} g_j(x_j) \le b,$$
$$x \in X.$$

A subproblem (SP) of (P_s) is formed by replacing X with a subset $\widetilde{X} \subseteq X$. The Lagrangian relaxation problem of (SP) can be written as

$$d(\lambda) = \min_{x \in \widetilde{X}} L(x, \lambda) = -\lambda b + \sum_{j=1}^{n} \min_{x_j \in \widetilde{X}_j} L_j(x_j, \lambda), \tag{2.15}$$

where $L_j(x_j, \lambda) = q_j(x_j) + \lambda g_j(x_j)$, $\widetilde{X}_j = \{x_j \mid \tilde{l}_j \le x_j \le \tilde{u}_j,\ x_j \text{ integer}\}$. The complexity of a total enumeration of evaluating $d(\lambda)$ is $O(\sum_{j=1}^{n}(\tilde{u}_j - \tilde{l}_j + 1))$. For problems with convex $q(x)$ and linear constraints, it is possible to derive a more efficient procedure for computing $d(\lambda)$ by exploiting the convex hull of q_j on $\widetilde{X}_j$ (see [16]). For nonlinear constraints, consider the following two cases: (a) $q_j(x_j)$ and $g_j(x_j)$ are convex; (b) $q_j(x_j)$ and $g_j(x_j)$ are concave. For case (a), the optimal solution of the one-dimensional integer problem in (2.15) can be obtained by comparing the values of L_j on two neighboring integer points of the continuous optimal solution of L_j on $[\tilde{l}_j, \tilde{u}_j]$. For case (b), since L_j is a concave function of x_j, the integer minimum of L_j over $[\tilde{l}_j, \tilde{u}_j]$ is either $\tilde{l}_j$ or $\tilde{u}_j$.

Therefore, the Lagrangian relaxation problems of the subproblems (SP) can be solved efficiently by decomposition. Moreover, the solutions x^k and y^k in Step 3 of Procedure 1 can also be easily obtained by computing the minimum and maximum values of g_j, respectively, when solving the one-dimensional problem in (2.15).

3. Quadratic contour cut. In this section, we will establish a cut-and-partition scheme by exploiting the geometry of the quadratic contour of the objective function $q(x)$. The cut-and-partition technique will be used later on to develop an exact solution method for solving (P).

Let $\alpha, \beta \in \mathbb{Z}^n$, where $\mathbb{Z}^n$ denotes the set of integer points in $\mathbb{R}^n$. Denote by $[\alpha, \beta]$ the box (hyperrectangle) formed by α and β, $[\alpha, \beta] = \{x \mid \alpha_j \le x_j \le \beta_j, j = 1, \ldots, n\}$. Let $\langle \alpha, \beta \rangle$ denote the set of integer points in $[\alpha, \beta]$,

$$\langle \alpha, \beta \rangle = \Pi_{j=1}^n \langle \alpha_j, \beta_j \rangle = \langle \alpha_1, \beta_1 \rangle \times \langle \alpha_2, \beta_2 \rangle \times \cdots \times \langle \alpha_n, \beta_n \rangle.$$

The set $\langle \alpha, \beta \rangle$ is called an *integer box*. For convenience, we define $[\alpha, \beta] = \langle \alpha, \beta \rangle = \emptyset$ if $\alpha \not\le \beta$.

3.1. Ellipse of quadratic contour. Let $q(x)$ be the quadratic function defined in (P). Let $\tau = -\sum_{j=1}^n d_j^2/(2c_j)$. Consider the ellipse contour of $q(x)$,

$$\sum_{j=1}^{n} [(1/2)c_j x_j^2 + d_j x_j] = v, \tag{3.1}$$

where $v \ge \tau$ when $c_j > 0$ $(j = 1, \ldots, n)$ and $v \le \tau$ when $c_j < 0$ $(j = 1, \ldots, n)$. The center of ellipse (3.1) is

$$o = (-d_1/c_1, \ldots, -d_n/c_n)^T. \tag{3.2}$$

The length of the jth axis of ellipse (3.1) is

$$2r_j = 2\sqrt{|2(v - \tau)/c_j|}. \tag{3.3}$$

Let $E(v)$ denote the ellipsoid formed by the contour (3.1). Then

$$E(v) = \begin{cases} \{x \in \mathbb{R}^n \mid q(x) \le v\} & \text{if } q(x) \text{ is convex,} \\ \{x \in \mathbb{R}^n \mid q(x) \ge v\} & \text{if } q(x) \text{ is concave.} \end{cases} \tag{3.4}$$

The minimum rectangle that encloses the ellipsoid $E(v)$ is $[a, b]$ with

$$a = (o_1 - r_1, \ldots, o_n - r_n)^T,$$
$$b = (o_1 + r_1, \ldots, o_n + r_n)^T,$$

where o is defined in (3.2) and r_j is defined in (3.3). Let $\lfloor t \rfloor$ denote the maximum integer less than or equal to t and $\lceil t \rceil$ the minimum integer greater than or equal to t. Then the minimum integer box containing all the integer points in the ellipsoid $E(v)$ can be expressed as $M(v) = \langle \alpha, \beta \rangle$, where

$$\alpha = (\lceil o_1 - r_1 \rceil, \ldots, \lceil o_n - r_n \rceil)^T, \tag{3.5}$$
$$\beta = (\lfloor o_1 + r_1 \rfloor, \ldots, \lfloor o_n + r_n \rfloor)^T. \tag{3.6}$$

Let $\tilde{x}$ be an integer point inside the ellipsoid $E(v)$. Let $N(\tilde{x})$ denote the integer subbox inside $E(v)$ with $\tilde{x}$ being one of its corner point. By the symmetry of $E(v)$, we have $N(\tilde{x}) = \langle \gamma, \delta \rangle$, where

$$\gamma = (\lceil o_1 - |\tilde{x}_1 - o_1| \rceil, \ldots, \lceil o_n - |\tilde{x}_n - o_n| \rceil)^T, \tag{3.7}$$

$$\delta = (\lfloor o_1 + |\tilde{x}_1 - o_1| \rfloor, \ldots, \lfloor o_n + |\tilde{x}_n - o_n| \rfloor)^T. \tag{3.8}$$

Notice that if $q(\tilde{x}) = v$, then $\langle \gamma, \delta \rangle$ is the maximum integer box inside $E(v)$ that passes through $\tilde{x}$.

3.2. Contour cuts of quadratic function. Consider the subproblem (SP) of the singly constrained problem (P_s) in subsection 2.2. Assume that $\widetilde{X} \cap S \neq \emptyset$ and $\widetilde{X} \setminus S \neq \emptyset$, where S is the feasible region of (P_s). Let q_s denote the optimal value of (SP). Let $\lambda^* > 0$ be the dual optimal solution to (SP). Suppose that the duality gap of (SP) is nonzero, i.e., $d(\lambda^*) < q_s$. By Theorem 2, the dual search procedure described in subsection 2.2 can find two optimal solutions, $\tilde{x} \in S$ and $\tilde{y} \in \widetilde{X} \setminus S$, to the Lagrangian relaxation problem (L_{λ^*}). The following always holds:

$$q(\tilde{y}) < d(\lambda^*) < q_s \leq q(\tilde{x}). \tag{3.9}$$

In the following we will show that cutting certain integer boxes from $\widetilde{X}$ will not remove any optimal solution of (SP) after recording $\tilde{x}$. We consider the contour cut for the two cases, where $q(x)$ is either convex or concave.

Case (a). $q(x)$ is convex, i.e., $c_j > 0$, $j = 1, \ldots, n$. Let $v_1 = q(\tilde{x})$ and $v_2 = d(\lambda^*)$. By (3.9) and the convexity of q, either $\tilde{x}$ is the optimal solution of (SP) or the optimal solution still lies in the set

$$\Omega = (\widetilde{X} \cap E(v_1)) \setminus E(v_2), \tag{3.10}$$

where $E(v_1)$ and $E(v_2)$ are defined by (3.4). In other words, removing sets $\widetilde{X} \setminus E(v_1)$ and $E(v_2)$ from $\widetilde{X}$ will not miss any optimal solution to (SP) after we record $\tilde{x}$. Since both $E(v_1)$ and $E(v_2)$ are ellipsoids, it is difficult to calculate Ω in (3.10). We instead outerapproximate Ω using integer boxes. More specifically, we consider a union of boxes of which Ω is a subset. Note that set Ω is a finite set containing only integer points. It is true that

$$\widetilde{X} \cap M(v_1) \supset \widetilde{X} \cap E(v_1), \tag{3.11}$$

where $M(v_1)$ is the minimum integer box enclosing all the integer points in $E(v_1)$. Let $B(v_1) = \widetilde{X} \cap M(v_1)$. Then $B(v_1) = \langle \bar{\alpha}, \bar{\beta} \rangle$, where

$$\bar{\alpha} = (\max(\tilde{l}_1, \alpha_1), \ldots, \max(\tilde{l}_n, \alpha_n))^T, \tag{3.12}$$

$$\bar{\beta} = (\min(\tilde{u}_1, \beta_1), \ldots, \min(\tilde{u}_n, \beta_n))^T, \tag{3.13}$$

with α and β defined in (3.5) and (3.6), respectively.

By (3.9), the infeasible point $\tilde{y}$ is contained in the ellipsoid $E(v_2)$. Thus, the integer box $N(\tilde{y}) = \langle \gamma, \delta \rangle$ is also contained in $E(v_2)$, where γ and δ can be found by using (3.7)–(3.8). This, combined with (3.11), implies that

$$B(v_1) \setminus N(\tilde{y}) \supset \Omega. \tag{3.14}$$

We further would like to cut $\tilde{x}$ from $\widetilde{X}$ if $\tilde{x} \in B(v_1)$ after recording $\tilde{x}$. Let $T(\tilde{x}) = \langle \tilde{\alpha}, \tilde{\beta} \rangle$ be the integer box with (i) $\tilde{x}$ being one of its corner points and (ii) all edges starting

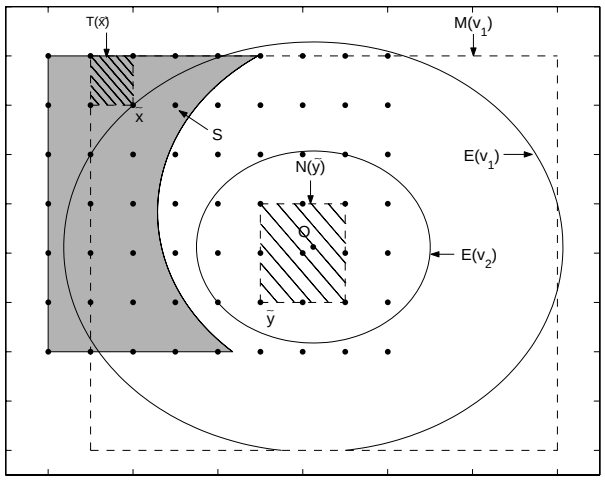


FIG. 3.1. *Contour cuts for Case* (a).

from $\tilde{x}$ leaving the ellipsoid $E(v_1)$ and moving towards the boundaries of $B(v_1)$. Specifically, $T(\tilde{x})$ can be determined by

$$\tilde{\alpha}_j = \begin{cases} \min(\tilde{x}_j, \bar{\alpha}_j), & \tilde{x}_j \le o_j, \\ \min(\tilde{x}_j, \bar{\beta}_j), & \tilde{x}_j > o_j, \end{cases} \tag{3.15}$$

$$\tilde{\beta}_j = \begin{cases} \max(\tilde{x}_j, \bar{\alpha}_j), & \tilde{x}_j \le o_j, \\ \max(\tilde{x}_j, \bar{\beta}_j), & \tilde{x}_j > o_j, \end{cases} \tag{3.16}$$

where o is defined in (3.2) and $\bar{\alpha}$ and $\bar{\beta}$ are defined in (3.12) and (3.13), respectively. Since $\tilde{x}$ is on the boundary of $E(v_1)$, we can cut $T(\tilde{x})$ from $B(v_1)$. We have

$$\tilde{\Omega} = [B(v_1) \setminus N(\tilde{y})] \setminus T(\tilde{x}) \supset \Omega \setminus \{\tilde{x}\}. \tag{3.17}$$

Figure 3.1 illustrates the contour cut process for Case (a).

Case (b). $q(x)$ is concave, i.e., $c_j < 0$, $j = 1, \ldots, n$. Let $v_1 = d(\lambda^*)$ and $v_2 = q(\tilde{x})$. Then, by (3.9) and the concavity of q, the optimal solution of (SP) must lie in the set Ω defined in (3.10). Similar to Case (a), we have

$$B(v_1) \setminus N(\tilde{x}) \supset \Omega. \tag{3.18}$$

Since $q(\tilde{y}) < d(\lambda^*) = v_1$, $\tilde{y}$ is outside the ellipsoid $E(v_1)$. If $\tilde{y}$ is contained in $B(v_1)$, then we can cut $T(\tilde{y})$ from $B(v_1)$, where $T(\tilde{y}) = \langle \tilde{\alpha}, \tilde{\beta} \rangle$; $\tilde{\alpha}$ and $\tilde{\beta}$ are defined in (3.15)–(3.16), with $\tilde{x}$ replaced by $\tilde{y}$. Therefore, we have

$$\tilde{\Omega} = [B(v_1) \setminus N(\tilde{x})] \setminus T(\tilde{y}) \supset \Omega. \tag{3.19}$$

Figure 3.2 illustrates the contour cut process for Case (b).

One clear conclusion is that after recording the feasible solution $\tilde{x}$, we can reduce the domain of (SP) from $\widetilde{X}$ to $\tilde{\Omega}$ without missing any optimal solution to (SP). This domain reduction process will improve the quality of the dual search, as witnessed in the following sections.

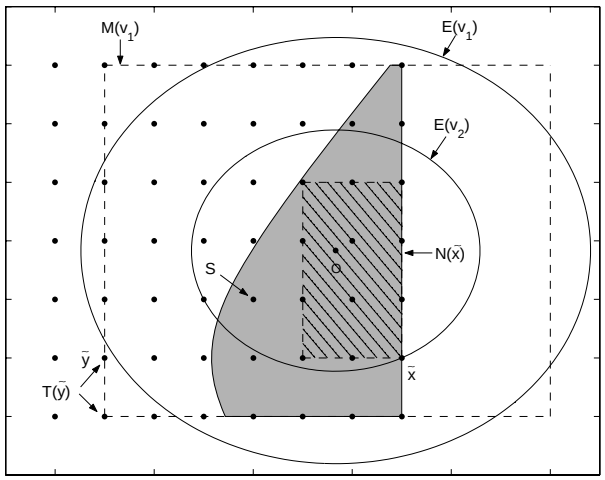


FIG. 3.2. *Contour cuts for Case* (b).

4. Convergent Lagrangian and contour cut method for singly constrained problems. In this section, we develop a convergent Lagrangian and contour cut method for the singly constrained problem (P_s). The method will be extended in section 5 to handle multiple constraints. We first motivate the method by an example and then describe the method formally.

4.1. Motivation. To motivate the method, let us consider a two-dimensional example with a concave quadratic objective function.

Example 1.

$$\begin{aligned} \min\ & q(x) = -1.5x_1^2 + 2x_1 - 2x_2^2 + 8x_2 \\ \text{s.t.}\ & g(x) = 3x_1^2 - 2x_1 + 2x_2^2 - 6x_2 \le 35, \\ & x \in X = \{x \mid -1 \le x_1 \le 5,\ 0 \le x_2 \le 6 \text{ integer},\ j = 1, 2\}. \end{aligned}$$

The optimal solution of this problem is $x^* = (-1, 5)^T$ with $q(x^*) = -13.5$. The perturbation function of the example is illustrated in Figure 4.1. It can be observed from Figure 4.1 that the point C that corresponds to the optimal solution x^* is "hidden" above the convex envelope of the perturbation function, and thus the traditional Lagrangian dual method will fail to find the optimal solution x^*.

Solving the dual problem of the example, we obtain the optimal multiplier $\lambda^0 = 0.6667$ with $d(\lambda^0) = -23.5$. The optimal solutions to (L_{λ^0}) are $x^0 = (-1, 0)^T$ and $y^0 = (-1, 6)^T$. The current duality bound is $q(x^0) - d(\lambda^0) = -3.5 + 23.5 = 20$.

Now, let $v_1^0 = -23.5$, $v_2^0 = -3.5$. Applying the contour cut scheme in section 3 to the example by using (3.19), we obtain a revised domain

$$X^1 = [B(v_1^0) \setminus N(x^0)] \setminus T(y^0),$$

where

$$B(v_1^0) = X \cap M(v_1^0) = \langle (-1, 0)^T, (5, 6)^T \rangle \cap \langle (-3, -2)^T, (5, 6)^T \rangle = \langle (-1, 0)^T, (5, 6)^T \rangle,$$
$$N(x^0) = \langle (-1, 0)^T, (2, 4)^T \rangle,\ T(y^0) = \{(-1, 6)^T\}.$$

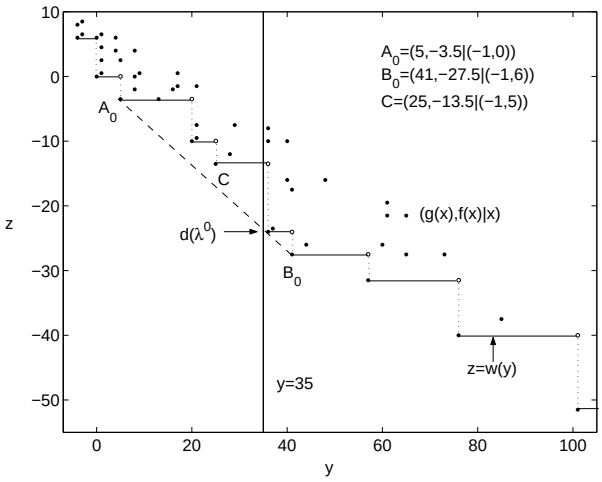


FIG. 4.1. *Perturbation function of Example* 1.

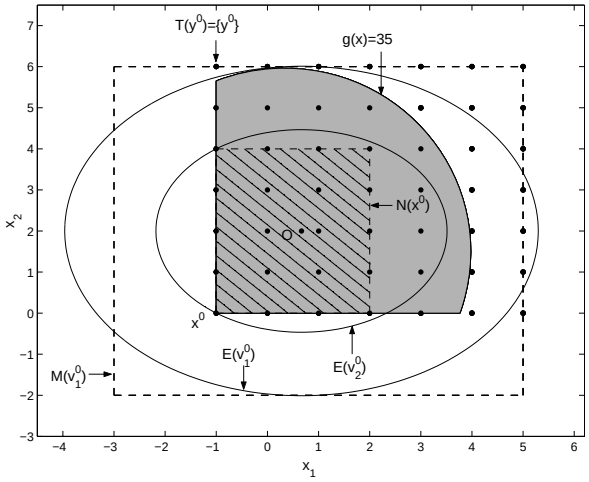


FIG. 4.2. *Domain X and the objective contour cuts.*

The ellipsoids $E(v_1^0)$, $E(v_2^0)$ and the integer boxes $M(v_1^0)$, $N(x^0)$, and $T(y^0)$ are illustrated in Figure 4.2. It can be seen from Figures 4.1 and 4.2 that cutting sets $N(x^0)$ and $T(y^0)$ from the domain X will remove the corner points A_0 and B_0 in the plot of the perturbation function, and thus will raise the dual value. The revised domain X^1 and the corresponding perturbation function are shown in Figures 4.3 and 4.4, respectively. The optimal dual value of the revised problem is $d(\lambda^1) = -23.125$ and the feasible and infeasible solutions of (L_{λ^1}) are $x^1 = (0,5)^T$, $y^1 = (0,6)^T$. The dual bound is reduced to $q(x^1) - d(\lambda^1) = -10 + 23.125 = 13.125$. Let $v_1^1 = -23.125$ and $v_2^1 = -10$. The ellipsoids $E(v_1^1)$, $E(v_2^1)$ and the integer boxes $M(v_1^1)$, $N(x^1)$, and $T(y^1)$ are illustrated in Figure 4.3.

The above discussion reveals that the contour cut scheme described in section 3 will reduce the duality bound, and thus the duality gap, and will eventually expose

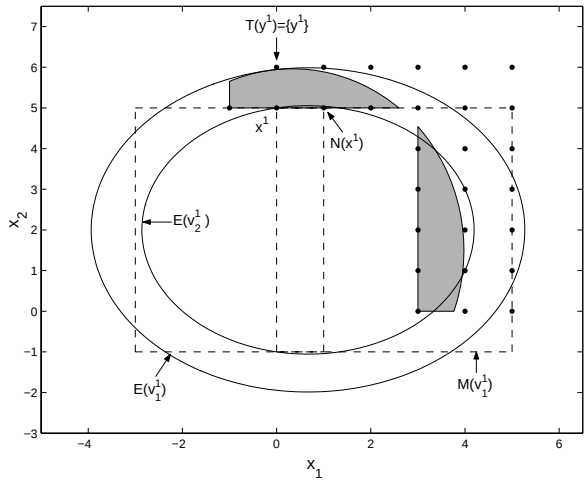


FIG. 4.3. *The revised domain* X^1.

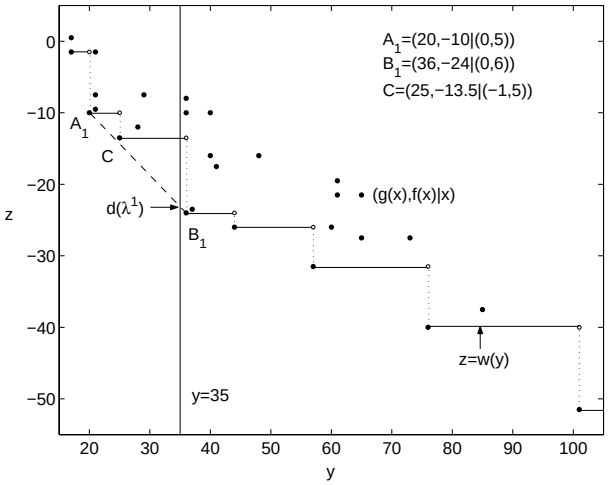


FIG. 4.4. *Perturbation function of the revised problem on* X^1.

the "hidden" optimal point to the convex envelope of the perturbation function. In fact, as we can foresee from Figure 4.4, one more contour cut will make the point C lie on the convex envelope of the revised perturbation function, thus enabling the dual search to find the optimal solution x^*.

4.2. Partition of a nonrectangular integer set. A key issue in the proposed Lagrangian dual and contour cut method is how to partition the sets in the right-hand sides of (3.17) and (3.19) into a union of integer boxes so that Lagrangian relaxation and dual search can still be applied to the revised problems after a cutting process. We have the following result.

LEMMA 1. *Let $A = \langle \alpha, \beta \rangle$ and $B = \langle \gamma, \delta \rangle$, where α, β, γ, $\delta \in \mathbb{Z}^n$, and $\alpha \le \gamma \le \delta \le \beta$. Then $A \setminus B$ can be partitioned into at most $2n$ integer boxes:*

$$A \setminus B = \big\{ \cup_{j=1}^{n} \big(\Pi_{i=1}^{j-1} \langle \alpha_i, \delta_i \rangle \times \langle \delta_j + 1, \beta_j \rangle \times \Pi_{i=j+1}^{n} \langle \alpha_i, \beta_i \rangle \big) \big\} \cup \big\{ \cup_{j=1}^{n} \big(\Pi_{i=1}^{j-1} \langle \gamma_i, \delta_i \rangle \times \langle \alpha_j, \gamma_j - 1 \rangle \times \Pi_{i=j+1}^{n} \langle \alpha_i, \delta_i \rangle \big) \big\}. \tag{4.1}$$

Proof. As illustrated in Figure 4.5, $A \setminus B$ can be expressed as

$$A \setminus B = \langle \alpha, \beta \rangle \setminus \langle \gamma, \delta \rangle = (\langle \alpha, \beta \rangle \setminus \langle \alpha, \delta \rangle) \cup (\langle \alpha, \delta \rangle \setminus \langle \gamma, \delta \rangle). \tag{4.2}$$

Let $C = \langle \alpha, \delta \rangle$. Then, by (4.2), we have

$$A \setminus B = (A \setminus C) \cup (C \setminus B). \tag{4.3}$$

For $j = 0, 1, \ldots, n-1$, define

$$A_j = \Pi_{i=j+1}^{n} \langle \alpha_i, \beta_i \rangle,$$
$$C_j = \Pi_{i=j+1}^{n} \langle \alpha_i, \delta_i \rangle.$$

Then

$$\begin{aligned}
&A_{j-1} \setminus C_{j-1} \\
&= \Pi_{i=j}^{n} \langle \alpha_i, \beta_i \rangle \setminus \Pi_{i=j}^{n} \langle \alpha_i, \delta_i \rangle \\
&= \big\{ (\langle \alpha_j, \delta_j \rangle \times \Pi_{i=j+1}^{n} \langle \alpha_i, \beta_i \rangle) \cup (\langle \delta_j + 1, \beta_j \rangle \times \Pi_{i=j+1}^{n} \langle \alpha_i, \beta_i \rangle) \big\} \setminus \Pi_{i=j}^{n} \langle \alpha_i, \delta_i \rangle \\
&= \big\{ (\langle \alpha_j, \delta_j \rangle \times \Pi_{i=j+1}^{n} \langle \alpha_i, \beta_i \rangle) \setminus \Pi_{i=j}^{n} \langle \alpha_i, \delta_i \rangle \big\} \cup (\langle \delta_j + 1, \beta_j \rangle \times \Pi_{i=j+1}^{n} \langle \alpha_i, \beta_i \rangle) \\
&= \big\{ \langle \alpha_j, \delta_j \rangle \times (\Pi_{i=j+1}^{n} \langle \alpha_i, \beta_i \rangle \setminus \Pi_{i=j+1}^{n} \langle \alpha_i, \delta_i \rangle) \big\} \cup (\langle \delta_j + 1, \beta_j \rangle \times \Pi_{i=j+1}^{n} \langle \alpha_i, \beta_i \rangle) \\
&= \{ \langle \alpha_j, \delta_j \rangle \times (A_j \setminus C_j) \} \cup (\langle \delta_j + 1, \beta_j \rangle \times \Pi_{i=j+1}^{n} \langle \alpha_i, \beta_i \rangle).
\end{aligned} \tag{4.4}$$

Using the partition formulation (4.4) recursively for $j = 1, \ldots, n-1$, and noting that $A = A_0$, $C = C_0$, $A_{n-1} \setminus C_{n-1} = \langle \alpha_n, \beta_n \rangle \setminus \langle \alpha_n, \delta_n \rangle = \langle \delta_n + 1, \beta_n \rangle$, we get

$$A \setminus C = \cup_{j=1}^{n} \big(\Pi_{i=1}^{j-1} \langle \alpha_i, \delta_i \rangle \times \langle \delta_j + 1, \beta_j \rangle \times \Pi_{i=j+1}^{n} \langle \alpha_i, \beta_i \rangle \big). \tag{4.5}$$

Similarly, we have

$$C \setminus B = \cup_{j=1}^{n} \big(\Pi_{i=1}^{j-1} \langle \gamma_i, \delta_i \rangle \times \langle \alpha_j, \gamma_j - 1 \rangle \times \Pi_{i=j+1}^{n} \langle \alpha_i, \delta_i \rangle \big). \tag{4.6}$$

Combining (4.3) with (4.5) and (4.6), we obtain (4.1). ∎

As an example, let us consider the X^1 in Figure 4.3. By using Lemma 1, X^1 can be partitioned into three integer boxes:

$$\begin{aligned}
X^1 &= [\langle (-1, 0), (5, 6)^T \rangle \setminus \langle (-1, 0)^T, (2, 4)^T \rangle] \setminus \langle (-1, 6)^T, (-1, 6)^T \rangle \\
&= [\langle (3, 0)^T, (5, 6)^T \rangle \cup \langle (-1, 5)^T, (2, 6)^T \rangle] \setminus \langle (-1, 6)^T, (-1, 6)^T \rangle \\
&= \langle (3, 0)^T, (5, 6)^T \rangle \cup \langle (0, 5)^T, (2, 6)^T \rangle \cup \langle (-1, 5)^T, (-1, 5)^T \rangle.
\end{aligned}$$

Lemma 1 shows that the revised domain $\tilde{\Omega}$ in (3.17) or (3.19) can be partitioned into a union of integer subboxes. The Lagrangian relaxation problem on the revised domain (e.g., X^1 in Example 1) can be easily solved by using decomposition. We will refer to the cutting process in (3.17) or (3.19) and the partition of the complement set

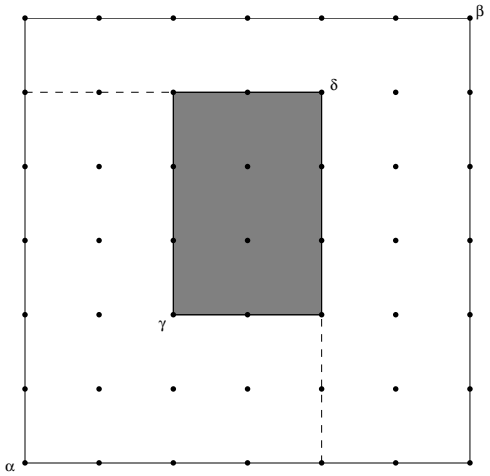


FIG. 4.5. *Partition of* $A \setminus B$.

by formula (4.1) as the *cut-and-partition* scheme. It is easy to see that the number of the new integer subboxes generated by partitioning set $\tilde{\Omega}$ is at most $3n-1$.

In implementation, instead of performing dual search on the revised domain as a whole, we apply the dual search procedure separately on each new integer subbox. This will yield a better lower bound of the revised problem as proved in the following lemma.

LEMMA 2. *Let*

$$\begin{aligned}
\hat{X} &= \cup_{i=1}^{t} \hat{X}_i, \\
\hat{S} &= \{x \in \hat{X} \mid g(x) \le b\}, \\
\hat{S}_i &= \{x \in \hat{X}_i \mid g(x) \le b\},\ i = 1, \ldots, t, \\
\hat{d}(\lambda) &= \min_{x \in \hat{X}} L(x, \lambda), \\
\hat{d}_i(\lambda) &= \min_{x \in \hat{X}_i} L(x, \lambda),\ i = 1, \ldots, t.
\end{aligned}$$

Further let $\hat{\lambda}^*$ *be the optimal solution of* $\max_{\lambda \ge 0} \hat{d}(\lambda)$ *and* $\hat{\lambda}_i^*$ *be the optimal solution of* $\max_{\lambda \ge 0} \hat{d}_i(\lambda)$, $i = 1, \ldots, t$. *Then,*

$$\hat{d}(\hat{\lambda}^*) \le \min_{1 \le i \le t} \hat{d}_i(\hat{\lambda}_i^*) \le \min_{x \in \hat{S}} f(x). \tag{4.7}$$

Proof. Since $\hat{X}_i \subseteq \hat{X}$, $\hat{d}(\lambda) \le \hat{d}_i(\lambda)$ for all $\lambda \ge 0$ and $i = 1, \ldots, t$. We thus have $\hat{d}(\hat{\lambda}^*) \le \hat{d}_i(\hat{\lambda}_i^*)$ for $i = 1, \ldots, t$. This further leads to the first inequality in (4.7). On the other hand, from the weak duality, we have $\hat{d}_i(\hat{\lambda}_i^*) \le \min_{x \in \hat{S}_i} f(x)$. This further yields $\min_{1 \le i \le t} \hat{d}_i(\hat{\lambda}_i^*) \le \min_{1 \le i \le t} \min_{x \in \hat{S}_i} f(x) = \min_{x \in \hat{S}} f(x)$, which is the second inequality in (4.7). □

It is clear from the above lemma that the minimum from among the dual values of all integer subboxes provides a Lagrangian lower bound higher than the dual value of the entire revised domain.

4.3. The main algorithm. Based on the above discussion, a convergent Lagrangian and contour cut algorithm can be developed by combining the Lagrangian relaxation with the cut-and-partition scheme. Let $X^0 = \{X\}$. Initially, a dual search procedure is applied to (P_s) to produce an optimal dual value $d(\lambda^0)$ together with a feasible optimal solution x^0 and an infeasible optimal solution y^0 to (L_{λ^0}). The optimal dual value $d(\lambda^0)$ gives a lower bound of the problem and x^0 is set to be the incumbent. At the kth iteration, the integer subbox with the minimum dual value is selected from X^k. The cut-and-partition scheme is then applied to that integer subbox. For each newly generated integer subbox, Procedure 1 is applied to determine its dual value together with a feasible solution and an infeasible solution. The current best feasible solution is recorded as the incumbent solution and all integer subboxes whose dual value is greater than or equal to the objective function value of the incumbent are removed. The process repeats until there is no integer subbox in X^k and the incumbent solution is the optimal solution to (P_s) when the algorithm terminates.

We now formally present the algorithm.

ALGORITHM 1 (convergent Lagrangian and contour cut algorithm for (P_s)).

Step 0 (initialization). Apply the dual search procedure to (P_s) and obtain the dual value $d(\lambda^0)$, a feasible solution x^0, and an infeasible solution y^0. Set $LB = d(\lambda^0)$ as the lower bound, $x_{opt} = x^0$, $f_{opt} = q(x_{opt})$, $X^0 = X$, $k = 0$.

Step 1. Select the integer subbox $\langle \alpha^k, \beta^k \rangle$ from X^k that yields the minimum lower bound LB. Let $x^k,\ y^k \in \langle \alpha^k, \beta^k \rangle$ be the feasible and infeasible solutions generated by Procedure 1, respectively.

Step 2 (contour cut and partition).

Case (a). q is a convex function. Set $v_1 = q(x^k)$, $v_2 = LB$; calculate integer boxes $B(v_1)$, $N(y^k)$, and $T(x^k)$. Use (4.1) to partition the set

$$Y^{k+1} = [B(v_1) \setminus N(y^k)] \setminus T(x^k). \tag{4.8}$$

Case (b). q is a concave function. Set $v_1 = LB$, $v_2 = q(x^k)$; calculate integer boxes $B(v_1)$, $N(x^k)$, and $T(y^k)$. Use (4.1) to partition the set

$$Y^{k+1} = [B(v_1) \setminus N(x^k)] \setminus T(y^k). \tag{4.9}$$

Step 3 (dual search).

(i) Apply Procedure 1 to each integer subbox $\langle \alpha, \beta \rangle \in Y^{k+1}$ with X replaced by $\langle \alpha, \beta \rangle$. Let

$$\tilde{x}^0 \in \arg \min_{x \in \langle \alpha, \beta \rangle} g(x),\ \tilde{y}^0 \in \arg \min_{x \in \langle \alpha, \beta \rangle} q(x).$$

One of the following three cases happens: (a) If $g(\tilde{x}^0) > b$, then remove $\langle \alpha, \beta \rangle$ from Y^{k+1}; (b) if $g(\tilde{y}^0) \le b$, then set $x_{opt} = \tilde{y}^0$ and $f_{opt} = q(\tilde{y}^0)$ if $f(\tilde{y}^0) < f_{opt}$, and remove $\langle \alpha, \beta \rangle$ from Y^{k+1}; (c) if $g(\tilde{x}^0) \le b$ and $g(\tilde{y}^0) > b$, then Procedure 1 generates a dual value on the integer box, a feasible solution, and an infeasible solution. If the dual value is greater than or equal to f_{opt}, then remove $\langle \alpha, \beta \rangle$ from Y^{k+1}. Compute the objective function value of the feasible solution and update x_{opt} and f_{opt} if necessary.

(ii) Set $X^{k+1} = Y^{k+1} \cup (X^k \setminus \{\langle \alpha^k, \beta^k \rangle\})$.

Step 4 (termination). If X^{k+1} is empty, stop. x_{opt} is an optimal solution to (P_s). Otherwise, set $k := k + 1$, goto Step 1.

THEOREM 5. *Algorithm* 1 *stops within a finite number of iterations with either an optimal solution to* (P_s) *being found or an infeasibility of* (P_s) *being reported.*

Proof. The finite convergence is obvious by noting that X is a finite integer set, and at each iteration, x^k and y^k are cut from X^k in Step 2 and are not included in X^{k+1}. From the discussion in section 3, no feasible solution better than x^k will be cut from X^k in Step 2. Also, by weak duality, no feasible solution better than x^k will be cut from X^k in Step 3. Thus, at each iteration, either x_{opt} is already the optimal solution or there is an optimal solution in X^k. Therefore, x_{opt} must be an optimal solution to the original problem when the algorithm stops at Step 4. □

5. Extension to problems with multiple constraints. The algorithm developed in section 4 can be extended to deal with multiply constrained cases of (P). Consider a subproblem (SP) of (P) with X replaced by an integer subbox $\widetilde{X} \subseteq X$. The Lagrangian dual of (SP) is

$$\max_{\lambda \in \mathbb{R}^m_+} d(\lambda), \tag{5.1}$$

where

$$d(\lambda) := \min_{x \in \widetilde{X}} \left[q(x) + \sum_{i=1}^{m} \lambda_i (g_i(x) - b_i) \right]. \tag{5.2}$$

From the weak duality, $d(\lambda) \le q(x)$ for any feasible solution $x \in \widetilde{X}$. Therefore, $d(\lambda)$ provides a lower bound of the optimal value of (SP). Let λ^* be an optimal solution to (5.1). Then, $LB = d(\lambda^*)$ is the best lower bound generated by the Lagrangian relaxation (5.2).

Since $d(\lambda)$ is a concave piecewise linear function, the subgradient method is an efficient method for computing an approximate solution to (5.1). Alternatively, we can use the outer Lagrangian linearization method (see [38]) to compute an exact solution to (5.1) when an initial feasible solution to (P) is available.

Consider the following surrogate constraint problem:

$$\begin{aligned} \min \ & q(x) = \sum_{j=1}^{n} \left(\frac{1}{2} c_j x_j^2 + d_j x_j \right) \\ \text{s.t.} \ & g_{\lambda^*}(x) = \sum_{i=1}^{m} \lambda_i^* g_i(x) \le \sum_{i=1}^{m} \lambda_i^* b_i, \\ & x \in \tilde{X}. \end{aligned} \tag{5.3}$$

Let $b_{\lambda^*} = \sum_{i=1}^{m} \lambda_i^* b_i$. Denote by $\underline{g}_{\lambda^*}$ and $\overline{g}_{\lambda^*}$ the minimum value and maximum value of $g_{\lambda^*}(x)$ over $\widetilde{X}$, respectively. Without loss of generality, we can assume that

$$\underline{g}_{\lambda^*} \le b_{\lambda^*} < \overline{g}_{\lambda^*}. \tag{5.4}$$

Suppose that λ^* is an exact solution to (5.1). It is easy to see that (5.3) and (SP) have the same dual value and that the optimal solution to the dual problem of problem (5.3) is 1. Moreover, by Theorem 2, there exist a feasible solution $\tilde{x}$ and an infeasible $\tilde{y}$ to problem (5.3) that solve the Lagrangian relaxation (5.2) with $\lambda = \lambda^*$.

If $\tilde{\lambda}$ is an approximate solution to (5.1), then we can apply Procedure 1 to search for an exact dual solution μ^* to problem (5.3) with λ^* replaced by $\tilde{\lambda}$. Set $\lambda^* = \mu^* \tilde{\lambda}$. Again, by Theorem 2, there exist a feasible solution $\tilde{x}$ and an infeasible $\tilde{y}$ to problem (5.3) that solve the Lagrangian relaxation (5.2) with $\lambda = \lambda^*$.

Now we are ready to extend Algorithm 1 to the multiply constrained case of (P). It is noticed that

$$q(\tilde{y}) < d(\lambda^*) \leq q(\tilde{x}). \tag{5.5}$$

Moreover, $\tilde{y}$ is infeasible to (P) while $\tilde{x}$ is not necessarily feasible to (P). Therefore, the contour cutting process in Step 2 of Algorithm 1 has to be modified for situations in which $\tilde{x}$ is infeasible to (P). More specifically, we need the following modifications in Algorithm 1.

Step $2'$. Case (a). q is a convex function. If $\tilde{x}$ is feasible to (P), set $v_1 = q(\tilde{x})$ and compute Y^{k+1} by

$$Y^{k+1} = [B(v_1) \setminus T(\tilde{x})] \setminus N(\tilde{y}).$$

Otherwise, if $\tilde{x}$ is infeasible to (P), then compute Y^{k+1} by

$$Y^{k+1} = [\langle \tilde{l}, \tilde{u} \rangle \setminus \{\tilde{x}\}] \setminus N(\tilde{y}).$$

Case (b). q is a concave function. Set $v_1 = LB$. If $\tilde{x}$ is feasible to (P), then compute Y^{k+1} by

$$Y^{k+1} = [B(v_1) \setminus N(\tilde{x})] \setminus T(\tilde{y}).$$

Otherwise, if $\tilde{x}$ is infeasible to (P), compute Y^{k+1} by

$$Y^{k+1} = [B(v_1) \setminus \{\tilde{x}\}] \setminus T(\tilde{y}).$$

We also need to replace the dual search procedure used in Step 0 and Step 3(i) of Algorithm 1 with an exact dual search method or an approximate method for (5.1). When the dual problems (5.1) in Step 0 and Step 3(i) are solved approximately, Procedure 1 is applied to the surrogate problem (5.3) to search for the lower bound, together with a feasible solution and infeasible solution for (5.3). Finally, two special cases have to be considered in the algorithm when (5.4) does not hold. If $\underline{g}_{\lambda^*} > b_{\lambda^*}$, then there is no feasible solution in $\widetilde{X}$, and $\widetilde{X}$ can be removed from further consideration. If $\overline{g}_{\lambda^*} \leq b_{\lambda^*}$, then solving (5.3) using the dual search will yield a zero dual solution and an optimal solution $\tilde{x}$ which is feasible for (5.3). If $\tilde{x}$ is also feasible for (P), discard $\tilde{X}$ from further consideration after updating x_{opt} and f_{opt} if $q(\tilde{x}) < f_{opt}$. Otherwise, remove $\tilde{x}$ from $\tilde{X}$.

The finite convergence of the extended algorithm for multiply constrained problems and the optimality of x_{opt} when the algorithm stops can be proved similarly as in Theorem 5.

An important observation from Step $2'$ is that in multiply constrained situations, we are not always able to find a feasible solution to the primal problem during the dual search procedure, which constitutes a major difference between multiply constrained problems and singly constrained problems. The unavailability of feasible solutions to the primal problem affects the efficiency of the contour cut algorithm for multiply constrained problems, as witnessed from our computational experiences. Specifically, a guaranteed two-direction cutting process (cutting the outside of a bigger ellipse and the inside of a smaller ellipse) in singly constrained situations often becomes a one-direction cutting process in multiply constrained situations when a feasible solution is not available. Nevertheless, in some situations, certain heuristics can be used to

search for a feasible solution which does not necessarily solve problem (5.2). This may improve the efficiency of the contour cutting process. For example, if the constraint functions are nondecreasing, as is the case in nonlinear knapsack problems, then the lower bound point $\tilde{l}$ of $\widetilde{X}$ is always feasible for (SP) in nontrivial cases.

We now illustrate the extended algorithm for multiply constrained problems by a two-dimensional example with a concave quadratic objective function, a convex constraint, and a nonconvex constraint.

Example 2.

$$\begin{aligned} \min\ & q(x) = -1.5x_1^2 + 2x_1 - 2x_2^2 + 8x_2 \\ \text{s.t.}\ & g_1(x) = 3x_1^2 - 2x_1 + 2x_2^2 - 6x_2 \le 66, \\ & g_2(x) = -x_1^2 - x_1 + x_2^2 - 2x_2 \le -3.5, \\ & x \in X = \{x \mid -1 \le x_1 \le 5,\ 0 \le x_2 \le 6,\ x_i \text{ integer}\}. \end{aligned}$$

The optimal solution is $x^* = (5,0)^T$ with $q(x^*) = -27.5$.

For this example, we use the subgradient method to solve the dual problem (5.1). The iterative process is described as follows.

Iteration 0.

Step 0. Solving (5.1) with $\widetilde{X} = X$, we get $\lambda^* = (0.5145, 0.2284)^T$. Applying Procedure 1 to the surrogate constraint problem (5.3), we obtain the dual value $LB = -34.0771$ and two optimal solutions $x^0 = (-1,6)^T$ and $y^0 = (5,6)^T$. An initial feasible solution $(5,0)^T$ is also obtained during the dual search. Set $x_{opt} = (5,0)^T$ and $f_{opt} = q(x_{opt}) = -27.5$. Notice that both x^0 and y^0 are infeasible for (P). Set $X^0 = X$ and $k = 0$.

Iteration 1.

Step 1. Select X to generate new integer boxes.

Step 2. Set $v_1 = LB = -34.0771$. We have

$$B(v_1) = M(v_1) \cap X = \langle (-4,-2)^T, (6,6)^T \rangle \cap X = X$$

and

$$Z^1 = B(v_1) \setminus \{x^0\} = \langle (0,0)^T, (5,6)^T \rangle \cup \langle (-1,0)^T, (-1,5)^T \rangle = Z_1^1 \cup Z_2^1.$$

Since the dual value on Z_2^1 is $-13.5 > -27.5 = f_{opt}$, we can remove Z_2^1 from Z^1. We have $T(y^0) = \langle (5,6)^T, (5,6)^T \rangle$. Thus,

$$Y^1 = Z^1 \setminus T(y^0) = \langle (0,0)^T, (4,6)^T \rangle \cup \langle (5,0)^T, (5,5)^T \rangle = Y_1^1 \cup Y_2^1.$$

For Y_1^1, the dual value is -33.1476 with two solutions $(0,6)^T$ and $(4,6)^T$; for Y_2^1, the dual value is -32.2875 with two solutions $(5,0)^T$ and $(5,5)^T$.

Step 3. Set $X^1 = Y^1$, $k = 1$.

Iteration 2.

Step 1. Select Y_1^1 from X^1 to generate new integer boxes. Set $x^1 = (0,6)^T$ and $y^1 = (4,6)^T$. Notice that x^1 is infeasible to (P).

Step 2. Set $v_1 = -33.1476$. Calculate $B(v_1) = M(v_1) \cap Y_1^1 = \langle (-4,-2)^T, (5,6)^T \rangle \cap Y_1^1 = Y_1^1$. We have

$$Z^2 = B(v_1) \setminus \{x^1\} = \langle (1,0)^T, (4,6)^T \rangle \cup \langle (0,0)^T, (0,5)^T \rangle = Z_1^2 \cup Z_2^2.$$

Since the dual value on Z_2^2 is $-1.5966 > -27.5 = f_{opt}$, we can remove Z_2^2 from Z^2. We have $T(y^1) = \langle (4,6)^T, (4,6)^T \rangle$. Thus

$$Y^2 = Z^2 \setminus T(y^1) = \langle (1,0)^T, (3,6)^T \rangle \cup \langle (4,0)^T, (4,5)^T \rangle = Y_1^2 \cup Y_2^2.$$

The dual value on Y_1^2 is -20.7748 and the dual value on Y_2^2 is -26.0. Since both of them are greater than f_{opt}, we can remove Y_1^2 and Y_2^2 from Y^2.

Step 3. Set $X^2 = \{Y_2^1\}$, $k = 2$.

Iteration 3.

Step 1. Select Y_2^1 to generate the new integer subboxes. Set $x^2 = (5,0)^T$ and $y^2 = (5,5)^T$. Note that x^2 is feasible to (P).

Step 2. Set $v_1 = -32.2875$. Calculate $B(v_1) = M(v_1) \cap Y_2^1 = \langle (-4,-2)^T, (5,6)^T \rangle \cap Y_2^1 = Y_2^1$ and $N(x^2) = \langle (5,0)^T, (5,4)^T \rangle$. We have

$$Z^2 = B(v_1) \setminus N(x^2) = \{(5,5)^T\}.$$

Thus

$$Y^2 = Z^2 \setminus \{y^2\} = \emptyset.$$

Step 3. $X^3 = \emptyset$.

Step 4. Stop. $x_{opt} = (5,0)^T$ is an optimal solution to the example.

6. Extension to problems with indefinite q. The contour cut method developed in the previous sections can be extended to handle problems with an indefinite quadratic objective function. We describe the main idea of this extension in this section. Let's first consider the singly constrained problem (P_s), where some c_j coefficients are positive and some others are negative.

We can always express $q(x)$ as the sum of a convex quadratic function and a concave quadratic function: $q(x) = q_1(x) + q_2(x)$ with $q_1(x) = \sum_{j=1}^n (\frac{1}{2} c_j^1 x_j^2 + d_j x_j)$ and $q_2(x) = -\sum_{j=1}^n \frac{1}{2} c_j^2 x_j^2$, where all c_j^1 and c_j^2, $j = 1, 2, \ldots, n$, are positive. Note that the expression of $q(x)$ is not unique. The subproblem (SP) of problem (P_s) can be expressed as

$$\begin{aligned} \min\ & q(x) = q_1(x) + q_2(x) \\ \text{s.t.}\ & g(x) = \sum_{j=1}^n g_j(x_j) \le b, \\ & x \in \widetilde{X} = \{x \mid \tilde{l}_j \le x_j \le \tilde{u}_j,\ x_j \text{ integer},\ j = 1, \ldots, n\}, \end{aligned}$$

where $\widetilde{X} \subseteq X$. Consider the following two problems associated with (SP):

$$\begin{aligned} (SP^1) \qquad \min\ & q_1(x) = \sum_{j=1}^n \left(\frac{1}{2} c_j^1 x_j^2 + d_j x_j \right) \\ \text{s.t.}\ & g(x) = \sum_{j=1}^n g_j(x_j) \le b, \\ & x \in \widetilde{X} = \{x \mid \tilde{l}_j \le x_j \le \tilde{u}_j,\ x_j \text{ integer},\ j = 1, \ldots, n\} \end{aligned}$$

and

$$(SP^2) \quad \min\ q_2(x) = -\sum_{j=1}^{n} \frac{1}{2} c_j^2 x_j^2$$
$$\text{s.t.}\ \ g(x) = \sum_{j=1}^{n} g_j(x_j) \le b,$$
$$x \in \widetilde{X} = \{x \mid \tilde{l}_j \le x_j \le \tilde{u}_j,\ x_j \text{ integer},\ j = 1, \ldots, n\}.$$

Obviously, (SP^1) and (SP^2) are nonlinear integer programming problems with a convex quadratic objective function and a concave quadratic objective function, respectively. Let $f_i^* = \min_{x \in S \cap \widetilde{X}} q_i(x)$, $i = 1, 2$, where S is the feasible region of (P_s). Further define the following Lagrangian relaxation for (SP^1) and (SP^2), respectively, for $\lambda \ge 0$:

$$(L_\lambda^i) \quad d_i(\lambda) = \min_{x \in \widetilde{X}} q_i(x) + \lambda(g(x) - b)), \qquad i = 1, 2.$$

Let λ_i^* be the optimal solutions to the dual problems of $\max_{\lambda \ge 0} d_i(\lambda)$ for $i = 1, 2$, respectively. Let $\tilde{x} \in S \cap \widetilde{X}$. By the weak duality, we have

$$d_1(\lambda_1^*) + d_2(\lambda_2^*) \le f_1^* + f_2^* \le f^* \le q_1(\tilde{x}) + q_2(\tilde{x}). \tag{6.1}$$

Let

$$C_1 = \{x \in \widetilde{X} \mid q_i(x) < d_i(\lambda_i^*), \quad i = 1, 2\},$$
$$C_2(\tilde{x}) = \{x \in \widetilde{X} \mid q_i(x) \ge q_i(\tilde{x}), \quad i = 1, 2\}.$$

It is easy to see from (6.1) and the weak duality that sets C_1 and $C_2(\tilde{x})$ can be cut off from $\widetilde{X}$ without removing the optimal solution after recording $\tilde{x}$. Let $\tilde{x}_i$ and $\tilde{y}_i$ be the feasible and infeasible optimal solutions to $(L_{\lambda_i^*}^i)$ $(i = 1, 2)$, respectively. Notice that $q_i(\tilde{y}_i) \le d_i(\lambda_i^*)$, $i = 1, 2$. Let $v_i = q_1(\tilde{x}_i)$, $i = 1, 2$, and $w = d_2(\lambda_2^*)$. Similar to section 3, we define sets $B_i(\cdot)$ and $N_i(\cdot)$ for functions q_i, $i = 1, 2$, respectively. Then we have

$$Q_1 = N_1(\tilde{y}_1) \cap [\widetilde{X} \setminus B_2(w)] \subseteq C_1 \cap \widetilde{X},$$
$$Q_2(\tilde{x}_i) = [\widetilde{X} \setminus B_1(v_i)] \cap N_2(\tilde{x}_i) \subseteq C_2(\tilde{x}_i) \cap \widetilde{X}, \qquad i = 1, 2.$$

Thus, cutting both Q_1 and $Q_2(\tilde{x}_i)$ $(i = 1, 2)$ from $\widetilde{X}$ will not remove any optimal solution to the primal problem after recording the current best feasible solution as the incumbent. Note Q_1 and/or $Q_2(\tilde{x}_i)$ could be empty in certain circumstances. In the cutting process, points $\tilde{x}_i$, $i = 1, 2$, will be removed from $\widetilde{X}$ after updating the incumbent.

Replacing Step 2 of Algorithm 1 with the above contour cutting process, we can then deal with (P_s) with an indefinite quadratic objective function. Similar to section 5, we can further extend the algorithm to solve the multiply constrained case of (P) with an indefinite objective function.

Now, let's demonstrate the above solution idea by an illustrative example.

Example 3.

$$\min\ q(x) = -1.75x_1^2 - 1.75x_1 + x_2^2 - 12x_2$$
$$\text{s.t.}\ \ g(x) = 4(x_1 - 1)^2 + 9(x_2 - 2.5)^2 \le 10,$$
$$x \in X = \{x \mid 0 \le x_i \le 4,\ x_i \text{ integer},\ i = 1, 2\}.$$

The optimal solution of the example is $x^* = (2, 3)^T$ with $q(x^*) = -37.5$.

Decompose the above example into the following two associated problems, of which the first has a convex quadratic objective function and the second has a concave quadratic objective function:

$$\begin{aligned} \min\ & q_1(x) = 0.25x_1^2 - 1.75x_1 + 3x_2^2 - 12x_2 \\ \text{s.t.}\ & g(x) = 4(x_1-1)^2 + 9(x_2-2.5)^2 \le 10, \\ & x \in X = \{x \mid 0 \le x_i \le 4,\ x_i \text{ integer},\ i = 1,2\} \end{aligned} \tag{6.2}$$

and

$$\begin{aligned} \min\ & q_2(x) = -2x_1^2 - 2x_2^2 \\ \text{s.t.}\ & g(x) = 4(x_1-1)^2 + 9(x_2-2.5)^2 \le 10, \\ & x \in X = \{x \mid 0 \le x_i \le 4,\ x_i \text{ integer},\ i = 1,2\}. \end{aligned} \tag{6.3}$$

Iteration 0.

Step 0. Solving the dual problem of (6.2) yields a dual value, $d_1 = -14.6563$, and two solutions, $\tilde{x}_1 = (2,2)^T$ and $\tilde{y}_1 = (3,2)^T$. Solving the dual problem of (6.3) yields a dual value, $d_2 = -29.1250$, and two solutions, $\tilde{x}_2 = (2,3)^T$ and $\tilde{y}_2 = (3,3)^T$. Thus, the lower bound is $LB = d_1 + d_2 = -14.6563 - 29.1250 = -43.7813$ and the incumbent is $x_{opt} = (2,3)^T$ with $f_{opt} = q((2,3)^T) = -37.5$. Set $X^0 = \{\langle (0,0)^T, (4,4)^T \rangle\}$ and $k = 0$.

Iteration 1.

Step 1. Select the unique integer subbox in X^0.

Steps 2 *and* 3. Since $N_1(\tilde{y}_1) = \langle (3,2)^T, (4,2)^T \rangle$ and $B_2(d_2) = \langle (0,0)^T, (3,3)^T \rangle$, we have $Q_1 = N_1(\tilde{y}_1) \cap [X \setminus B_2(d_2)] = \{(4,2)^T\}$. Thus

$$Z^1 = X \setminus Q_1 = \langle (0,3)^T, (4,4)^T \rangle \cup \langle (0,0)^T, (3,2)^T \rangle \cup \langle (4,0)^T, (4,1)^T \rangle = Z_1^1 \cup Z_2^1 \cup Z_3^1.$$

For Z_1^1, we have $d_1 + d_2 = -11.6563 - 29.1250 = -40.7813 < -37.5 = f_{opt}$. For Z_2^1, we have $d_1 + d_2 = -14.6563 - 19.1250 = -33.7813 > -37.5 = f_{opt}$. Thus Z_2^1 is removed from Z^1. Since there is no feasible solution in Z_3^1, Z_3^1 is also removed from Z^1. Set $Y^1 = \{Z_1^1\}$. Figure 6.1 illustrates the set $Z^1 = X \setminus Q_1$.

Let $v_1 = q_1(\tilde{x}_1) = -14.5$. Since $N_2(\tilde{x}_1) = \langle (0,0)^T, (2,2)^T \rangle$ and $B_1(v_1) = \langle (2,2)^T, (4,2)^T \rangle$, we have

$$Q_2(\tilde{x}_1) = [X \setminus B_1(v_1)] \cap N_2(\tilde{x}_1) = N_2(\tilde{x}_1) \setminus \{(2,2)^T\} = \langle (0,0)^T, (2,2)^T \rangle \setminus \{(2,2)^T\}.$$

Notice $Q_2(\tilde{x}_2)$ is an empty set. Since $Q_2(\tilde{x}_1) \cap Y^1 = \emptyset$, a revised domain X^1 is generated from cutting $\tilde{x}_2$ from Y^1 (see Figure 6.2). Decompose X^1 as

$$X^1 = \langle (3,3)^T, (4,4)^T \rangle \cup \langle (0,4)^T, (2,4)^T \rangle \cup \langle (0,3)^T, (1,3)^T \rangle = X_1^1 \cup X_2^1 \cup X_3^1.$$

Since there is no feasible solution in X_1^1 and X_2^1, they can be removed from X^1. For X_3^1, we have $d_1 + d_2 = -10.5 - 20.0 = -30.5 > f_{opt}$, and thus X_3^1 is also removed. Therefore, $X^2 = \emptyset$.

Step 4. Stop. $x_{opt} = (2,3)^T$ is an optimal solution.

In computational implementation with an indefinite q, we can also solve the dual problem of (SP) directly to obtain a dual value $d(\lambda^*)$ and use $\max\{d(\lambda^*), d_1(\lambda_1^*) + d_2(\lambda_2^*)\}$ as the lower bound to identify unpromising subboxes to be fathomed. Let $\tilde{x}$ and $\tilde{y}$ be the feasible and infeasible optimal solutions to the Lagrangian relaxation problem of (SP) with λ set at λ^*, respectively. Instead of cutting $Q_2(\tilde{x}_i)$, $i = 1, 2$, we cut $Q_2(\tilde{x})$ in the algorithm.

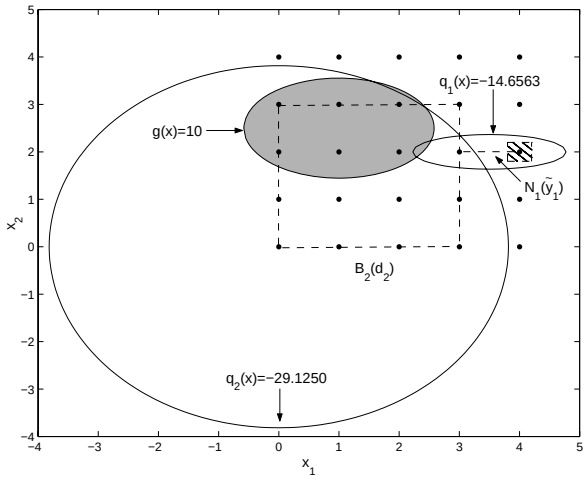


FIG. 6.1. *Set* $Z^1 = X \setminus Q_1$.

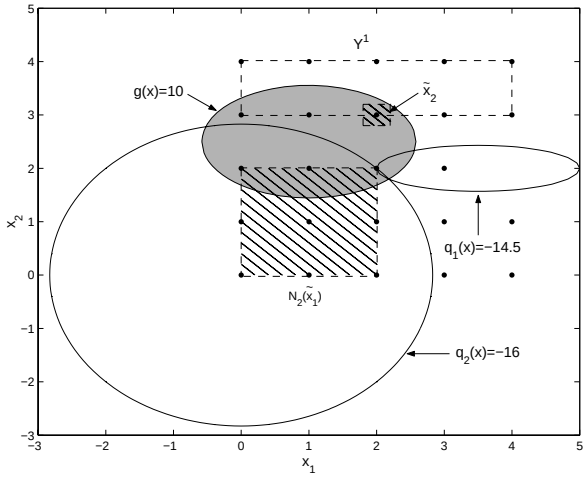


FIG. 6.2. *Sets* $X^1 = Y^1 \setminus \{\tilde{x}_2\}$ *and* $Q_2(\tilde{x}_1)$.

7. Computational experiment. In this section, we will present computational results of Algorithm 1 developed in section 4 and its extensions in sections 5 and 6. The algorithms were programmed in FORTRAN 90 and run on a SUN Workstation (Blade 2000). Comparison results with other methods in the literature will also be reported.

7.1. Test problems. Two sets of test problems are considered in our computational experiments. The first set of test problems consists of 12 problems with different types of objective functions and constraint functions. The second set of test problems is a class of convex quadratic integer programming problems arising in portfolio optimization. All the coefficients in the test problems are randomly generated from uniform distributions.

TABLE 7.1
Coefficients in the test problems.

Type	Single constraint			Multiple constraints		
	α_{1j}	β_{1j}	γ_{1j}	α_{ij}	β_{ij}	γ_{ij}
1	$[-10, 40]$	0	0	$[1, 40]$	0	0
2	$[-10, 30]$	$[1, 20]$	0	$[10, 50]$	$[1, 10]$	0
3	$[100, 200]$	$[-20, -1]$	0	$[-10\beta_{ij}, -10\beta_{ij} + 5]$	$[-15, -5]$	0
4	$[-10, 20]$	$[5, 25]$	$[-2, 8]$	$[10, 50]$	$[1, 10]$	$[1, 5]$

In the first set of test problems, three types of objective functions in the form of $q(x) = \sum_{j=1}^{n}(\frac{1}{2}c_j x_j^2 + d_j x_j)$ are generated using the following data:

- $q(x)$ is convex quadratic with $c_j \in [2, 20]$ and $d_j \in [-100, -50]$;
- $q(x)$ is concave quadratic with $c_j \in [-20, -2]$ and $d_j \in [-10, 40]$;
- $q(x)$ is indefinite quadratic with $c_j \in [-10, 10]$ and $d_j \in [-40, 10]$.

The constraint functions in the test problems are in the following form:

$$g_i(x) = \sum_{j=1}^{n}(\alpha_{ij}x_j + \beta_{ij}x_j^2 + \gamma_{ij}x_j^3), \quad i = 1, \ldots, m.$$

Table 7.1 describes the ranges of coefficients in g_i's for singly constrained test problems and multiply constrained test problems, where Type 1 denotes the linear constraints, Type 2 the convex quadratic constraints, Type 3 the concave quadratic constraints and Type 4 the third polynomial constraints.

In the first set of test problems, we take $l_j = 1$ and $u_j = 5$, $j = 1, \ldots, n$, and the right-hand side of b is taken as $b = g_{min} + r \cdot (g_{max} - g_{min})$, where g_{min} and g_{max} are the minimum and maximum values of $g(x)$ over X, respectively, and $r \in (0, 1)$.

The second set of problems arises from portfolio optimization. It has been shown in [41], [42] that the Markowitz mean-variance portfolio selection model can be reformulated as a simplified model which is a separable convex quadratic programming problem with linear constraints. The discrete version of the simplified portfolio selection problem [47] can be expressed as

$$\begin{aligned}(SMV) \qquad \min\ & q(x) = \sum_{j=1}^{n}\left(\frac{1}{2}c_j x_j^2 + d_j x_j\right) \\ \text{s.t.}\ & Ax \le b, \\ & x \in X = \{x \mid l_j \le x_j \le u_j,\ x_j \text{ integer},\ j = 1, \ldots, n\},\end{aligned}$$

where $c_j > 0$ for all j and $A = (a_{ij})$ is an $m \times n$ matrix. Obviously, problem (SMV) is a special case of (P). In our testing, the data in (SMV) are taken as the same as in [47], where additional dependency relationships are considered. The ranges of coefficients in (SMV) are $c_j \in [10, 50]$, $d_j \in [-3000, -1000]$, $a_{ij} \in [1, 5]$, $l_j \in [0, 40]$, and $u_j = l_j + 5$. The right-hand side b is taken as $b = A \cdot [l + r \cdot (u - l)]$, where $l = (l_1, \ldots, l_n)^T$, $u = (u_1, \ldots, u_n)^T$ and $r \in (0, 1)$.

7.2. Computational results. The computational results of our proposed solution algorithms for the first set of test problems are summarized in Tables 7.2–7.4. The following notations are used in the numerical results:

- n = number of variables;
- m = number of constraints;
- N_{iter} = average number of iterations of the algorithm for 20 test problems;

TABLE 7.2
Numerical results for problems with convex $q(x)$ ($r = 0.6$).

Type of constraint	Single constraint				Multiple constraints				
	n	N_{iter}	N_{box}	T_{cpu}	n	m	N_{iter}	N_{box}	T_{cpu}
Linear	500	162	42928	52.9	30	10	598	10732	74.9
	1000	253	123197	332.0	40	10	1074	23765	211.0
	1500	538	404297	1716.6	50	10	5313	129684	1901.4
Convex quadratic	500	155	46853	86.1	30	10	123	2426	8.0
	1000	242	138212	521.5	40	10	204	4952	34.3
	1500	354	286512	1746.4	50	10	432	13568	80.9
Concave quadratic	500	186	43213	76.9	30	10	161	1875	4.8
	1000	508	189918	669.0	50	10	340	6521	27.6
	1500	555	338348	2075.4	70	10	674	16870	107.4
3rd polynomial	500	111	31924	77.2	30	10	45	927	2.4
	1000	155	80996	427.9	50	10	140	4314	16.8
	1500	199	156222	1301.4	70	10	344	14296	78.9

TABLE 7.3
Numerical results for problems with concave $q(x)$ ($r = 0.6$).

Type of constraint	Single constraint				Multiple constraints				
	n	N_{iter}	N_{box}	T_{cpu}	n	m	N_{iter}	N_{box}	T_{cpu}
Linear	500	32	8748	13.4	50	10	32	765	2.3
	1000	53	27622	97.1	100	10	112	4580	24.8
	2000	58	60408	464.1	150	10	268	17602	173.8
Convex quadratic	500	43	9388	20.0	100	10	33	1407	5.9
	1000	57	23858	114.1	150	10	77	4973	31.2
	2000	149	120776	1294.3	200	10	110	9739	79.5
Concave quadratic	500	18	4334	9.8	100	10	26	1268	4.5
	1000	70	34094	163.0	150	10	65	4659	24.4
	2000	108	105606	1085.8	200	10	237	21910	169.1
3rd polynomial	500	47	8943	27.6	100	10	53	2302	9.9
	1000	76	30419	196.6	150	10	113	6701	45.7
	2000	104	75337	1080.5	200	10	215	17852	188.2

- N_{box} = average number of the total integer boxes examined during the algorithm for 20 test problems;
- T_{cpu} = average CPU seconds measured on a SUN Workstation (Blade 2000) for 20 test problems.

In our implementation of the algorithm for multiply constrained problems, the outer Lagrangian linearization method is used to solve the dual problem (5.1). The results in Tables 7.2–7.4 show that the proposed algorithm is efficient and robust for solving large-scale quadratic integer problems with convex, concave, and indefinite objective functions and different types of constraint functions. Comparing results in Table 7.2–Table 7.4, we can see that the algorithm is more efficient for problems with a concave objective function. We can also see that the efficiency of the algorithm is not sensitive to the convexity of the constraint functions. This is partially due to the fact that the cut-and-partition scheme does not depend on the property of the constraints.

The computational results for portfolio selection problems (SMV) are presented in Table 7.5, where N_{iter}, N_{box}, and T_{cpu} are obtained by running the code for 10 test problems. We see from Table 7.5 that the problem becomes more difficult as the ratio of right-hand side, r, decreases.

TABLE 7.4
Numerical results for problems with indefinite $q(x)$ ($r = 0.6$).

Type of constraint	Single constraint				Multiple constraints				
	n	N_{iter}	N_{box}	T_{cpu}	n	m	N_{iter}	N_{box}	T_{cpu}
Linear	200	183	23743	4.2	30	10	121	4081	29.7
	600	447	184493	166.1	40	10	179	8188	73.9
	1000	744	482480	997.1	50	10	601	38933	414.3
Convex quadratic	200	288	36056	7.6	30	10	65	2403	15.9
	600	896	349573	174.2	50	10	103	6107	58.7
	1000	1145	739706	572.7	70	10	238	20093	249.8
Concave quadratic	200	191	22386	5.0	30	10	42	1469	9.3
	600	776	272840	157.1	50	10	74	4184	36.9
	1000	2386	1376007	1806.0	70	10	126	9958	114.1
3rd polynomial	200	121	17283	5.9	30	10	39	1566	12.9
	600	836	341505	246.9	50	10	72	4713	55.1
	1000	1059	756976	857.0	70	10	75	7570	101.7

TABLE 7.5
Numerical results for problem (SMV).

r	n	m	N_{iter}	N_{box}	T_{cpu}
0.5	30	5	243	3274	9.6
	50	5	2191	46787	345.5
	80	5	4265	128091	860.0
0.6	30	5	326	4653	11.5
	50	5	1437	32564	143.0
	80	5	7888	232647	1292.2
0.7	30	5	95	1523	3.6
	50	5	361	7889	32.7
	80	5	596	22140	106.4

7.3. Comparison with other methods. To compare our algorithm with other existing methods in the literature, we implemented two exact methods in the literature which are applicable to (P):

- branch-and-bound method of Bretthauer and Shetty (see [9]).
- hybrid method of Marsten and Morin (see [34]).

Classical branch-and-bound methods using continuous relaxation can solve (P) when both $q(x)$ and $g(x)$ are convex functions. Gupta and Ravindran [21] reported computational results of the branch-and-bound method for general convex integer programming, where the generalized reduced gradient method was used as a solver for the continuous subproblems. Bretthauer and Shetty [9] proposed a special branch-and-bound method for singly constrained convex nonlinear separable integer programming problems (see also [10]). This method is based on solving the continuous relaxation subproblems by manipulating the KKT conditions of the subproblems and uses the standard branch rule to generate nodes in the search tree.

The hybrid method proposed in [34] is applicable to general separable integer programming problems, including (P). The method is a combination of the dynamic programming approach and the branch-and-bound method. The basic idea of the method is to recursively generate the efficient feasible solutions of the problem and to remove the inefficient feasible solutions by dominance rules. The branch-and-bound strategy is employed to remove unpromising incomplete solutions during the recursion.

TABLE 7.6
Comparison results for convex problems.

n	Our method	Branch-and-bound method	Hybrid method
	T_{cpu}	T_{cpu}	T_{cpu}
50	0.10	0.32	8.0
100	0.88	16.5	152.1
150	2.0	485.1	833.6

TABLE 7.7
Comparison results for nonconvex problems.

n	Our method	Hybrid method
	T_{cpu}	T_{cpu}
100	0.4	26.6
150	2.0	131.0
200	1.6	397.0

We have implemented the above two methods by FORTRAN 90 and tested for two sets of test problems for comparison. The first set of test problems is a convex instance of (P) with a single linear constraint. Both the branch-and-bound method and hybrid method are applicable to this set of test problems. The ranges of the parameters of $q(x)$ are $c_j \in [1, 10]$ and $d_j \in [-100, -300]$. The linear constraint is $g(x) = \sum_{j=1}^{n} \alpha_j x_j$ with $\alpha_j \in [1, 50]$. The ratio of the right-hand side b is taken as $r = 0.7$, and $l_j = 1$, $u_j = 5$, $j = 1, \ldots, n$. Table 7.6 summarizes the average CPU time of our proposed method, the branch-and-bound method, and the hybrid method for 20 randomly generated test problems in the first set.

The second set of test problems for comparison is a concave instance of (P) with a single linear constraint. Note that only the hybrid method in the literature is applicable to this kind of nonconvex problem. The ranges of the parameters of $q(x)$ are $c_j \in [-10, -1]$ and $d_j \in [-50, -1]$. The ranges of the coefficients in the linear constraint are $\alpha_j \in [1, 50]$. The ratio of the right-hand side b is taken as $r = 0.7$, and $l_j = 1$, $u_j = 5$, $j = 1, \ldots, n$. The comparison results for test problems with different n are reported in Table 7.7, where the average CPU time is obtained by running the algorithms for 20 randomly generated test problems.

The average CPU time in Tables 7.6 and 7.7 indicates that our algorithm is much more efficient than the branch-and-bound method and the hybrid method for both convex and nonconvex problems. Part of the theoretical reason for the outperformance of the proposed method over the continuous relaxation-based branch-and-bound method is that the Lagrangian bound of a convex integer programming problem is at least as good as the continuous bound. Moreover, cutting certain integer boxes from the domain at each iteration in the cut-and-partition scheme of our algorithm speeds up the convergence of the algorithm significantly. We also notice that it is difficult for dynamic programming in the hybrid method to exploit the special structure of the problems in generating efficient feasible solutions and it is thus not efficient to find an exact solution of the original problem.

8. Concluding remarks. We have presented in this paper an efficient exact algorithm for solving nonlinear separable integer programming problems with convex, concave, and indefinite quadratic objective functions and general constraints. The algorithm exploits the special structure of the ellipsoid contour in order to eliminate the duality bound, and thus the duality gap. A prominent feature of the proposed

algorithm is that at each iteration, the algorithm cuts certain unpromising integer subboxes by using quadratic contour cuts. This greatly speeds up the convergence of the Lagrangian dual algorithm. The computational results for large-scale test problems are promising.

Acknowledgment. The authors would like to thank Dr. Jun Wang for his help during this research.

REFERENCES

[1] J. E. Beasley, N. Meade, and T.-J. Chang, *An evolutionary heuristic for the index tracking problem*, European J. Oper. Res., 148 (2003), pp. 621–643.
[2] A. Beck and M. Teboulle, *Global optimality conditions for quadratic optimization problems with binary constraints*, SIAM J. Optim., 11 (2000), pp. 179–188.
[3] D. E. Bell and J. F. Shapiro, *A convergent duality theory for integer programming*, Oper. Res., 25 (1997), pp. 419–434.
[4] H. P. Benson and S. S. Erenguc, *An algorithm for concave integer minimization over a polyhedron*, Naval Res. Logist., 37 (1990), pp. 515–525.
[5] H. P. Benson, S. S. Erengue, and R. Horst, *A note on adopting methods for continuous global optimization to the discrete case*, Ann. Oper. Res., 25 (1990), pp. 243–252.
[6] D. Bienstock, *A computational study of a family of mixed-integer quadratic programming problems*, Math. Program., 74 (1996), pp. 121–140.
[7] I. M. Bomze and G. Danninger, *A global optimization algorithm for concave quadratic programming problems*, SIAM J. Optim., 3 (1993), pp. 826–842.
[8] K. M. Bretthauer, A. V. Cabot, and M. A. Venkataramanan, *An algorithm and new penalties for concave integer minimization over a polyhedron*, Naval Res. Logist., 41 (1994), pp. 435–454.
[9] K. M. Bretthauer and B. Shetty, *The nonlinear resource allocation problem*, Oper. Res., 43 (1995), pp. 670–683.
[10] K. M. Bretthauer and B. Shetty, *A pegging algorithm for the nonlinear resource allocation problem*, Comput. Oper. Res., 29 (2002), pp. 505–527.
[11] A. V. Cabot and S. S. Erengue, *A branch and bound algorithm for solving a class of nonlinear integer programming problems*, Naval Res. Logist., 33 (1986), pp. 559–567.
[12] M. W. Cooper, *The use of dynamic programming for the solution of a class of nonlinear programming problems*, Nav. Res. Logist. Quart., 27 (1980), pp. 89–95.
[13] M. W. Cooper, *Survey of methods of pure nonlinear integer programming*, Manage. Sci., 27 (1981), pp. 353–361.
[14] G. Danninger and I. M. Bomze, *Using copositivity for global optimality criteria in concave quadratic programming problems*, Math. Program., 62 (1993), pp. 575–580.
[15] M. Djerdjour, K. Mathur, and H. M. Salkin, *A surrogate relaxation based algorithm for a general class quadratic multidimensional knapsack problem*, Oper. Res. Lett., 7 (1988), pp. 253–258.
[16] M. E. Dyer and J. Walker, *Solving the subproblem in the Lagrangian dual of separable discrete programs with linear constraints*, Math. Program., 24 (1982), pp. 107–112.
[17] M. L. Fisher, *The Lagrangian relaxation method for solving integer programming problems*, Manage. Sci., 27 (1981), pp. 1–18.
[18] M. L. Fisher and J. F. Shapiro, *Constructive duality in integer programming*, SIAM J. Appl. Math., 27 (1974), pp. 31–52.
[19] C. Floudas and V. Visweswaran, *Quadratic optimization*, in Handbook of Global Optimization, R. Horst and P. M. Pardalos, eds., Kluwer, Dordrecht, 1995, pp. 217–270.
[20] A. M. Geoffrion, *Lagrangean relaxation for integer programming*, Math. Program. Stud., 2 (1974), pp. 82–114.
[21] O. K. Gupta and A. Ravindran, *Branch-and-bound experiments in convex nonlinear integer programming*, Manage. Sci., 31 (1985), pp. 1533–1546.
[22] D. Hochbaum, *A nonlinear knapsack problem*, Oper. Res. Lett., 17 (1995), pp. 103–110.
[23] N. J. Jobst, M. D. Horniman, C. A. Lucas, and G. Mitra, *Computational aspects of alternative portfolio selection models in the presence of discrete asset choice constraints*, Quant. Finance, 1 (2001), pp. 489–501.
[24] B. Kalantari and J. B. Rosen, *An algorithm for global minimization of linearly constrained concave quadratic functions*, Math. Oper. Res., 12 (1987), pp. 544–561.

[25] F. Korner, *A hybrid method for solving nonlinear knapsack problems*, European J. Oper. Res., 38 (1989), pp. 238–241.
[26] J. B. Lasserre, *An explicit equivalent positive semidefinite program for nonlinear* 0-1 *programs*, SIAM J. Optim., 12 (2002), pp. 756–769.
[27] D. J. Laughhunn, *Quadratic binary programming with application to capital-budgeting problems*, Oper. Res., 18 (1970), pp. 454–461.
[28] C. Lemaŕechal and A. Renaud, *A geometric study of duality gaps, with applications*, Math. Program., 90 (2001), pp. 399–427.
[29] D. Li and X. L. Sun, *Success guarantee of dual search in integer programming: p-th power Lagrangian method*, J. Global Optim., 18 (2000), pp. 235–254.
[30] D. Li and X. L. Sun, *Nonlinear Integer Programming*, Springer, New York, 2006.
[31] D. Li, X. L. Sun, and J. Wang, *Optimal lot solution to cardinality constrained mean-variance formulation for portfolio selection*, Math. Finance, (2006), pp. 83–101.
[32] D. Li and D. J. White, *pth power Lagrangian method for integer programming*, Ann. Oper. Res., 98 (2000), pp. 151–170.
[33] H. M. Markowitz, *Portfolio Selection: Efficient Diversification of Investment*, John Wiley & Sons, New York, 1959.
[34] R. E. Marsten and T. L. Morin, *A hybrid approach to discrete mathematical programming*, Math. Program., 14 (1978), pp. 21–40.
[35] K. Mathur, H. M. Salkin, and B. B. Mohanty, *A note on a general non-linear knapsack problems*, Oper. Res. Lett., 5 (1986), pp. 79–81.
[36] K. Mathur, H. M. Salkin, and S. Morito, *A branch and search algorithm for a class of nonlinear knapsack problems*, Oper. Res. Lett., 2 (1983), pp. 55–60.
[37] G. Mitra, K. Darby-Dowman, C. A. Lucas, and J. Yadegar, *Linear, integer separable and fuzzy programming problems: A unified approach towards reformulatoin*, J. Oper. Res. Soc., 39 (1988), pp. 161–171.
[38] R. G. Parker and R. L. Rardin, *Discrete Optimization*, Academic Press, Boston, 1988.
[39] J. B. Rosen and P. M. Pardalos, *Global minimization of large-scale constrained concave quadratic problems by separable programming*, Math. Program., 34 (1986), pp. 163–174.
[40] J. F. Shapiro, *A survey of Lagrangian techniques for discrete optimization*, Ann. Discrete Math., 5 (1979), pp. 113–138.
[41] W. F. Sharpe, *A simplified model for portfolio analysis*, Manage. Sci., 9 (1963), pp. 277–293.
[42] W. F. Sharpe, *Portfolio Theory and Capital Markets*, McGraw–Hill, New York, 1970.
[43] J. P. Shectman and N. V. Sahinidis, *A finite algorithm for global minimization of separable concave programs*, J. Global Optim., 12 (1998), pp. 1–36.
[44] X. L. Sun and D. Li, *Asymptotic strong duality for bounded integer programming: A logarithmic-exponential dual formulation*, Math. Oper. Res., 25 (2000), pp. 625–644.
[45] X. L. Sun and D. Li, *New dual formulations in constrained integer programming*, in Progress in Optimization, X. Q. Yang, ed., Kluwer, Dordrecht, 2000, pp. 79–91.
[46] X. L. Sun and D. Li, *Optimality condition and branch and bound algorithm for constrained redundancy optimization in series systems*, Optim. Eng., 3 (2002), pp. 53–65.
[47] S. S. Syam, *A dual ascent method for the portfolio selection problem with multiple constraints and linked proposals*, European J. Oper. Res., 108 (1998), pp. 196–207.
[48] V. Visweswaran and C. A. Floudas, *New properties and computational improvement of the GOP algorithm for problems with quadratic objective function and constraints*, J. Global Optim., 3 (1993), pp. 439–462.

SIAM J. OPTIM.
Vol. 17, No. 2, pp. 401–429
© 2006 Society for Industrial and Applied Mathematics

A FEASIBLE ACTIVE SET QP-FREE METHOD FOR NONLINEAR PROGRAMMING*

LIFENG CHEN†, YONGLI WANG‡, AND GUOPING HE§

Abstract. We propose a monotone descent active set QP-free method for inequality constrained optimization that ensures the feasibility of all iterates and allows for iterates on the boundary of the feasible set. The study is motivated by the Facchinei–Fischer–Kanzow active set identification technique for nonlinear programming and variational inequalities [F. Facchinei, A. Fischer, and C. Kanzow, *SIAM J. Optim.*, 9 (1999), pp. 14–32]. Distinguishing features of the proposed method compared with existing QP-free methods include lower subproblem costs and a fast convergence rate under milder assumptions. Specifically, four reduced linear systems with a common coefficient matrix involving only constraints in a working set are solved at each iteration. To determine the working set, the method makes use of multipliers from the last iteration, eliminating the need to compute a new estimate, and no additional linear systems are solved to select linearly independent constraint gradients. A new technique is presented to avoid possible ill-conditioned Newton systems caused by dual degeneracy. It is shown that the method converges globally to KKT points under the linear independence constraint qualification (LICQ), and the asymptotic rate of convergence is Q-superlinear under an additional strong second-order sufficient condition (SSOSC) without strict complementarity.

Key words. constrained optimization, nonlinear programming, QP-free methods, global convergence, superlinear convergence, strict complementarity

AMS subject classifications. 90C30, 65K10

DOI. 10.1137/040605904

1. Introduction.

This paper is concerned with finding a solution of the inequality constrained optimization problem

$$\text{(P)} \qquad \begin{aligned} &\min\ f(x) \\ &\text{s.t. } x \in \mathcal{F} = \{x \in \Re^n | c_i(x) \le 0,\ i = 1, \dots, m\}, \end{aligned}$$

where $f : \Re^n \to \Re$ and $c : \Re^n \to \Re^m$ are assumed to be real valued and twice continuously differentiable on $\mathcal{F}$. A pair $(x, \lambda) \in \Re^{n+m}$ with $x \in \mathcal{F}$ is called a stationary point of problem (P) if it satisfies

$$\begin{aligned} &\nabla_x \mathcal{L}(x, \lambda) = 0, \\ &\lambda_i c_i(x) = 0,\ i = 1, \dots, m, \end{aligned} \tag{1.1}$$

*Received by the editors March 29, 2004; accepted for publication (in revised form) January 27, 2006; published electronically June 9, 2006.

http://www.siam.org/journals/siopt/17-2/60590.html

†IEOR Department, Columbia University, New York, NY 10027 (lifeng.chen@columbia.edu). The research of this author was supported by the Presidential Fellowship of Columbia University. This work was started while the author was attending the School of Information Science and Engineering, Shandong University of Science and Technology, China.

‡Department of Mathematics, Shanghai Jiaotong University, Shanghai, 200240, China (wylyb@sjtu.edu.cn). Current address: School of Information Science and Engineering, Shandong University of Science and Technology, 579 Qian Wan Gang Road, Qingdao, 266510, China.

§School of Information Science and Engineering, Shandong University of Science and Technology, 579 Qian Wan Gang Road, Qingdao, 266510, China (hegp@sdust.edu.cn). The research of this author was supported in part by China National Natural Science Foundation grant 10571109 and Tianyuan Foundation of China.

where $\mathcal{L}(x,\lambda) = f(x) + \sum_{i=1}^{m} \lambda_i c_i(x)$ is the Lagrange function of (P). If furthermore $\lambda \geq 0$, (x,λ) is called a KKT point. We also call $x \in \mathcal{F}$ a stationary point or a KKT point of (P) if there exists $\lambda \in \Re^m$ such that (x,λ) is a stationary point or a KKT point of (P).

QP-free methods, which at each iteration require only the solution of a few linear systems usually with common coefficient matrices, were developed to address some computational issues in traditional sequential quadratic programming methods (SQPs). For example, the QP subproblems may be infeasible and the cost for finding their exact solutions can become prohibitive in the absence of a QP truncating scheme. A number of numerical results (see, e.g., [7, 9, 19, 28]) have shown the promise of QP-free methods as an alternative to SQPs for a class of primal nondegenerate problems, such as bound constrained optimization. Under certain regularity assumptions, QP-free methods enjoy global convergence as well as local superlinear/quadratic convergence. For the advantages of QP-free methods, see Facchinei and Lucidi [8].

QP-free methods can be classified into type-1 methods (see, e.g., [18, 13, 19]), which are based on applying Newton's method to the KKT systems (1.1), and type-2 methods (see, e.g., [7, 9, 20, 28]), which are derived from the alternative formulation of the first-order necessary optimality conditions

$$\begin{aligned} &\nabla f(x) + \sum_{i \in I(x)} \lambda_i \nabla c_i(x) = 0, \\ &c_i(x) = 0,\ \lambda_i \geq 0,\ i \in I(x);\ \lambda_i = 0,\ i \in I \backslash I(x), \end{aligned} \tag{1.2}$$

where $I = \{1, \ldots, m\}$ and, for $x \in \mathcal{F}$, $I(x)$ denotes the active set, i.e.,

$$I(x) = \{i | c_i(x) = 0, \quad i \in I\}.$$

1.1. Related work. Panier, Tit, and Herskovits [18] proposed a feasible QP-free method for problem (P) which forms the basic framework of type-1 QP-free methods. Their method directly applies the Newton method to (1.1) and at each iteration first calculates a descent direction $d^{k,0}$ by solving the following Newton equations:

$$\begin{bmatrix} H_k & \nabla c(x^k) \\ \operatorname{diag}(\mu^k)\nabla c(x^k)^\top & \operatorname{diag}(c(x^k)) \end{bmatrix} \begin{bmatrix} d \\ \lambda \end{bmatrix} = \begin{bmatrix} -\nabla f(x^k) \\ 0 \end{bmatrix}, \tag{1.3}$$

where $H_k \in \Re^{n \times n}$ is an estimate of the Lagrangian Hessian, (x^k, μ^k) is an approximation to a KKT point, and, for a vector $\mu \in \Re^m$, $\operatorname{diag}(\mu)$ denotes the diagonal matrix whose ith diagonal element is μ_i. In order to maintain the feasibility of the next iterate, the method calculates a second direction $d^{k,1}$ pointing toward the feasible region by solving a perturbed system of (1.3),

$$\begin{bmatrix} H_k & \nabla c(x^k) \\ \operatorname{diag}(\mu^k)\nabla c(x^k)^\top & \operatorname{diag}(c(x^k)) \end{bmatrix} \begin{bmatrix} d \\ \lambda \end{bmatrix} = \begin{bmatrix} -\nabla f(x^k) \\ -\|d^{k,0}\|^\nu \operatorname{diag}(\mu^k) e \end{bmatrix}, \tag{1.4}$$

where $\nu > 2$ and e is the vector of all ones of an appropriate dimension (here $e \in \Re^m$). Globally, the search direction is a convex combination of these two directions, namely,

$$d^k = (1 - \rho_k) d^{k,0} + \rho_k d^{k,1},$$

where ρ_k is calculated explicitly. To avoid the Maratos effect, locally the search direction is slightly bent by a relatively small amount using an arc search, which is

obtained through a second-order correction. It is shown that any accumulation point of the iterates generated by the algorithm is a stationary point of problem (P). Under the assumption that any stationary point is isolated, this point is shown to be a KKT point. The algorithm was later improved by Gao, He, and Wu [13] in the sense that every accumulation point of the iterates is a KKT point. To achieve this, they solve an extra linear system obtained from (1.3) by slightly perturbing the right-hand side of (1.3). As pointed out in [18, 19], the Panier–Tits–Herskovits framework is very sensitive to the parameters chosen due to dual degeneracy. Specifically, in spite of the presence of the linear independence constraint qualification (LICQ), the linear systems (1.3) and (1.4) may become very ill-conditioned when some multiplier μ_i corresponding to a nearly active constraint c_i becomes very small. This may occur when strict complementarity does not hold at the solution of problem (P). In this case the multiplier approximation sequence can diverge, and thus global convergence fails.

To avoid the ill-conditioning, Qi and Qi [19] presented a new type-1 QP-free method for problem (P), based on a nonsmooth equation reformulation of the KKT system (1.1), by using the Fischer–Burmeister function that is often used in nonlinear complementarity problems (see, e.g., [5, 16]). The coefficient matrix of their Newton equations has the form

$$V_k = \begin{bmatrix} H_k & \nabla c(x^k) \\ \operatorname{diag}(\xi^k)\nabla c(x^k)^\top & \Theta_k \end{bmatrix}, \tag{1.5}$$

where

$$\xi_i^k = \frac{c_i(x^k)}{\sqrt{c_i^2(x^k) + (\mu_i^k)^2}} + 1, \quad \Theta_k = -\sqrt{2}\operatorname{diag}(\theta^k), \quad \theta_i^k = \left(1 - \frac{\mu_i^k}{\sqrt{c_i^2(x^k) + (\mu_i^k)^2}}\right)^{1/2}.$$

It is shown that under the LICQ, (1.5) is uniformly nonsingular and well conditioned even if strict complementarity does not hold at accumulation points. We note that the methods in [18, 13, 19] require the sequence of Hessian estimates $\{H_k\}$ to be uniformly positive definite. This may interfere with the fast local convergence of their methods since for a general nonlinear programming problem, the Lagrangian Hessian at a second-order stationary point is usually only positive definite on the null space of the active constraint gradients. This issue has been recently circumvented in [1, 22] in the context of primal-dual interior-point methods.

From both a practical and a theoretical point of view, there are still some shortcomings of type-1 QP-free methods. For example, the methods in [18, 13, 19] need all inequality constraints to be involved in each subproblem computation. However, since QP-free methods do not follow the spirit of interior-point methods, whose iterates are forced to be away from the boundary of the feasible region, intuitively some active set strategies should be incorporated into these methods to ignore redundant constraints and save subproblem costs. Moreover, to guarantee global convergence, the methods in [18, 13, 19, 1, 22] must bound the multiplier approximation sequence $\{\mu^k\}$ by a preselected threshold value $\bar{\mu}$. However, to achieve fast local convergence, they have to assume that this truncation does not affect the convergence of the approximate multipliers, i.e., all multipliers are less than or equal to $\bar{\mu}$. Finally, all type-1 QP-free methods need strict complementarity for proving superlinear convergence.

Type-2 QP-free methods were first studied as locally convergent Newton methods for the KKT system (1.2) (see, e.g., [2, 8]). If the active set at a solution is available,

(1.2) reduces to a nonlinear system of equations involving only active constraints, which can be handled by standard Newton methods. Unfortunately, knowledge of the active set becomes available only when the iterates are close to the solution. In order to ensure global convergence, when the iterates are far from the solution, type-2 QP-free methods try to guess a so-called working set A_k for approximating the active set. Consequently, the Newton equations for these methods usually have the following coefficient matrices:

$$V_k = \begin{bmatrix} H_k & \nabla c_{A_k}(x^k) \\ \nabla c_{A_k}(x^k)^\top & 0 \end{bmatrix}, \tag{1.6}$$

where H_k is an estimate of the Lagrangian Hessian and $\nabla c_{A_k}(x^k)^\top$ is the Jacobian corresponding to the constraints in A_k. It is easy to see from (1.6) that V_k is nonsingular if and only if the constraint gradients in A_k are linearly independent, provided H_k is positive definite on the null space of $\nabla c_{A_k}(x^k)^\top$. The convergence properties of type-2 QP-free methods are very similar to that of type-1 QP-free methods as they eventually generate almost identical search directions. However, in practice, their behavior can be rather different in that type-2 QP-free methods, as active set methods, can be relatively sensitive to changes in the working set, while they do not suffer the pitfall triggered by dual degeneracy, as do some type-1 QP-free methods.

A careful examination of existing type-2 QP-free methods, however, reveals two shortcomings. The first one concerns the computation of multiplier estimates. Since the performance of type-2 QP-free methods largely depends on how well the active set can be identified, these methods require that the working set be able to approximate the active set quickly and accurately. To this end, methods in [7, 20, 28] resort to the following continuous multiplier function to obtain an ideal estimate of the multipliers:

$$\lambda(x) = -W(x)^{-1}\nabla c(x)^\top \nabla f(x),$$

where

$$W(x) = \nabla c(x)^\top \nabla c(x) + C(x)^2$$

and $C(x) = \mathrm{diag}(c(x))$. This involves solving a product form linear system of size $m \times m$, which is very likely to be fully dense. Second, type-2 QP-free methods require an expensive procedure to select linearly independent constraint gradients. In [20, 28] this is done iteratively by (i) checking the rank of the constraint gradients in an ε-working set; (ii) if they are rank deficient, reducing ε until they have full rank. Some do this by iteratively computing matrix determinant, which provides a criterion for singularity. Generally, these operations require computing factorizations for a sequence of down-dated matrices. An exception is [7], which relies on a restrictive assumption that, for a fixed ε, the constraint gradients in the ε-working set are linearly independent.

QP-free methods have many desirable local convergence properties. In [2, 8] some classes of local QP-free methods were shown to be superlinearly convergent without strict complementarity. This feature was extended to SC^1 problems in [6]. Moreover, local convergence properties of Newton or quasi-Newton methods have been widely studied under the Mangasarian–Fromovitz constraint qualification (MFCQ) or even without constraint qualifications (see, e.g., [10, 11, 15, 25, 26, 27]). Unfortunately, these nice local properties were rarely achieved in a globally convergent framework.

1.2. Basic results and notation. In this paper we propose a new feasible descent active set QP-free method for solving problem (P) that combines the best of type-1 and type-2 QP-free methods while overcoming several problematic aspects of them. The method, compared with existing QP-free methods, enjoys some theoretical advantages in saving computational cost and achieving fast local convergence. Specifically, two ε-working sets are maintained throughout the algorithm. One aims at identifying the final active set, while the other tries to identify the final strong active set. Our working set strategy is based on the Facchinei–Fischer–Kanzow active set identification technique [4], which can be accurate even without the LICQ or strict complementarity. To determine the working sets, we use multiplier information from the previous iteration, eliminating the need to compute a new estimate. Moreover, to avoid the expensive procedure of selecting linearly independent constraint gradients, we use as our Newton system (see (2.3)) a modified version of (1.3). At each iteration four reduced linear systems with a common coefficient matrix involving only constraints in the working set identifying the active set are computed. It is shown that under the LICQ, our method converges globally to KKT points of problem (P). To achieve fast local convergence, our method always employs the exact Lagrangian Hessian to compute the step if it is positive definite on the null space of some (nearly) active constraint gradients. This is particularly so when a KKT point satisfying the strong second-order sufficient condition (SSOSC) is approached. A new technique is introduced to avoid the ill-conditioning caused by dual degeneracy in some type-1 QP-free methods. In particular, since some active multipliers may vanish when strict complementarity fails, we control the multiplier estimates μ^k in a way that the active ones converge to the true multipliers plus a positive parameter δ_*, namely, δ_*-drifted multipliers. It will be shown that this technique does not affect the global and local convergence analysis. Furthermore, in order to guarantee the uniform nonsingularity of our Newton systems, we do not simply bound μ^k by a preselected parameter as in [1, 13, 18, 19, 22]. Instead, we adaptively increase the bound estimate χ of μ^k and decrease the active set parameter ε until a better estimate of the active set is obtained. It will be shown that under the LICQ, χ and ε are updated at most finitely many times.

Another contribution of this paper is that our method achieves fast local convergence without assuming strict complementarity. This property benefits a lot from the identification of the strong active set and a new system of inequalities introduced in the paper, through which better choices of search directions become possible. At each iteration, a "fast" direction, which can generate superlinear convergence when a solution is approached, is always computed first and is accepted as the search direction if the system of inequalities is met. However, if the system is violated, a steeper descent direction is computed to ensure global convergence. A key observation for our local analysis is that the "fast" direction enjoys both descent and feasible features in the vicinity of a solution and can be eventually accepted in spite of dual degeneracy.

The paper is organized as follows. In section 2 we present our algorithm and show that it is well defined. The global convergence of the method is established in section 3. Local superlinear convergence is proved in section 4. Concluding remarks are given in section 5.

A few words concerning the notation: The Euclidean vector norm or its associated matrix norm is denoted by $\|\cdot\|$. Given $h : \Re^n \to \Re^m$ 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 $\nabla h_A(x)^\top$ the transposed Jacobian of $h_A(x)$. For a positive integer p, $e \in \Re^p$ is the vector of all ones and $E \in \Re^{p\times p}$ is the identity matrix. Given two vectors x and y of the same

dimension l, we say $x \geq (>)\ y$ if and only if $x_i \geq (>)\ y_i$ for all $i = 1, \ldots, l$, and $\min(x, y)$ is a vector whose ith element is $\min(x_i, y_i)$. For two symmetric matrices $\mathcal{A}$ and $\mathcal{B}$ of the same dimension, $\mathcal{A} \succ \mathcal{B}$ means $\mathcal{A} - \mathcal{B}$ is positive definite. We denote by $\emptyset$ the empty set.

2. Algorithm. In the remainder of the paper, we let (x^*, λ^*) denote a KKT point of problem (P). Our algorithm makes use of the following identification function $\varphi : \Re^{n+m} \to \Re$ proposed in [4]:

$$\varphi(x, \lambda) = \sqrt{\|\Phi(x, \lambda)\|}, \tag{2.1}$$

where the operator $\Phi : \Re^{n+m} \to \Re^{n+m}$ is defined by

$$\Phi(x, \lambda) = \begin{pmatrix} \nabla_x \mathcal{L}(x, \lambda) \\ \min\{-c(x), \lambda\} \end{pmatrix}. \tag{2.2}$$

It follows from [4, Theorem 3.15] that φ is nonnegative and continuous with $\varphi(x, \lambda) = 0$ if and only if (x, λ) is a KKT point of problem (P). The identification function $\varphi(x, \lambda)$ plays two roles in our algorithm. First, it is used to determine an ε-working set at each iteration,

$$A_{\varepsilon, \varphi_{\max}}(x, \lambda) = \{i \in I | c_i(x) \geq -\varepsilon \min\{\varphi(x, \lambda), \varphi_{\max}\}\},$$

which is an estimate of the final active set $I(x^*)$. When (x, λ) is sufficiently close to (x^*, λ^*), the estimate is accurate, provided both the MFCQ and the SSOSC hold at (x^*, λ^*); see [4, Theorem 2.3]. Second, in the algorithm we also need to "guess" the strong active set, i.e., the set of active constraints with positive multipliers,

$$I^+(x^*) = \{i \in I(x^*) | \lambda_i^* > 0\}.$$

To do so, we again employ function $\varphi(x, \lambda)$ to measure the multiplier estimate. Specifically, we set

$$\Lambda_{\varepsilon, \varphi_{\max}}(x, \lambda) = \{i \in A_{\varepsilon, \varphi_{\max}} | \lambda_i \geq \varepsilon \min\{\varphi(x, \lambda), \varphi_{\max}\}\}.$$

If a strict MFCQ[1] and the SSOSC hold at (x^*, λ^*), $\Lambda_{\varepsilon, \varphi_{\max}}(x, \lambda)$ eventually identifies the strong active set $I^+(x^*)$; see [4, Theorem 2.4]. Note that the strict MFCQ is implied by the LICQ. To simplify the presentation, for the kth iteration ($k = 1, 2, \ldots$) we let

$$I_k = A_{\varepsilon_k, \varphi_{\max}}(x^k, \lambda^{k-1,0}) \quad \text{and} \quad L_k = \Lambda_{\varepsilon_k, \varphi_{\max}}(x^k, \lambda^{k-1,0}),$$

where $\lambda^{k-1,0}$ and ε_k will be specified in the algorithm.

In order to avoid computing linearly independent constraint gradients, the coefficient matrices of our Newton equations follow the form of (1.3) but involving only constraints in the working set I_k,

$$M_k = \begin{bmatrix} H_k & \nabla c_{I_k}(x^k) \\ U_k \nabla c_{I_k}(x^k)^\top & C_{I_k}(x^k) \end{bmatrix}, \tag{2.3}$$

where $U_k = \operatorname{diag}(\mu_{I_k}^k)$ and $C_{I_k}(x^k) = \operatorname{diag}(c_{I_k}(x^k))$. Note that M_k is slightly different from the coefficient matrix of (1.3) in that μ^k in (2.3) are controlled to be componentwise bounded below over zero under the LICQ and the SSOSC. In particular, it

[1] This nomenclature is from [4].

will be shown that μ^k actually converges to some δ_*-drifted ($\delta_* > 0$) multipliers with respect to the active set, i.e.,

$$\mu_i^k \to \delta_* + \lambda_i^* \quad \text{for} \quad i \in I(x^*).$$

This ensures that M_k are uniformly well conditioned even if λ^* is degenerate or nearly degenerate.

We are now ready to state the algorithm.

ALGORITHM 2.1.

Step 0 (INITIALIZATION).

Parameters: $\beta \in (0,1)$, $\sigma \in (0,1/2)$, $\eta \in (2,3)$, $\tau \in (0,1)$, $\nu \in (2,3)$, $\theta \in (0,1)$, $\gamma > 2$, $\delta > 0$, $\varphi_{\max} > 0$, $\vartheta \in (0,1)$.

Data: $x^1 \in \mathcal{F}$, $\lambda^{0,0} \geq 0$. If $L_1 \neq \emptyset$ and $\varphi(x^1, \lambda^{0,0}) > 0$, $\delta_1 = \vartheta \min\{\lambda_i^{0,0}, i \in L_1\}$; otherwise, $\delta_1 = \delta$. $\mu^1 = \lambda^{0,0} + \delta_1 e$, $\varepsilon_1 > 0$, $\chi_1 \gg \varphi_{\max}$, $H_1 = \nabla^2_{xx}\mathcal{L}(x^1, \lambda^{0,0})$.

Set $k = 1$.

Step 1 (COMPUTATION OF SEARCH DIRECTION).

(i) Modify H_k, if necessary, so that

$$d^\top \hat{H}_k d > 0 \ \forall d \in \mathcal{T}(x^k)\backslash\{0\}, \tag{2.4}$$

where $\mathcal{T}(x) = \{y \in \Re^n | \nabla c_i(x)^\top y = 0, \ i \in I(x)\}$ for $x \in \mathcal{F}$ and

$$\hat{H}_k = H_k - \sum_{i \in I_k \backslash I(x^k)} \frac{\mu_i^k}{c_i(x^k)} \nabla c_i(x^k) \nabla c_i(x^k)^\top. \tag{2.5}$$

Compute $(d^{k,0}, \lambda_{I_k}^{k,0})$ by solving the following linear system in (d, λ):

$$M_k \begin{bmatrix} d \\ \lambda \end{bmatrix} = \begin{bmatrix} -\nabla f(x^k) \\ 0 \end{bmatrix}. \tag{2.6}$$

If $d^{k,0} = 0$ and $\lambda_{I_k}^{k,0} \geq 0$, stop.

(ii) Compute $(d^{k,1}, \lambda_{I_k}^{k,1})$ by solving the following linear system in (d, λ):

$$M_k \begin{bmatrix} d \\ \lambda \end{bmatrix} = \begin{bmatrix} -\nabla f(x^k) \\ -\delta_k c_{I_k}(x^k) \end{bmatrix}. \tag{2.7}$$

If $(d^{k,1}, \lambda_{I_k}^{k,1})$ satisfies

$$\begin{cases} \nabla f(x^k)^\top d^{k,1} \leq -\|d^{k,1}\|^\gamma, \\ |\lambda_{-,i}^{k,1}| \leq \|d^{k,1}\| \ \ \forall i \in I_k, \end{cases} \tag{2.8}$$

set FAST = TRUE; else set FAST = FALSE, where $\lambda_{-,i}^{k,1} = \min\{\lambda_i^{k,1}, \ 0\}$, $i \in I_k$.

(iii) Compute $(d^{k,2}, \lambda_{I_k}^{k,2})$ by solving the following linear system in (d, λ):

$$M_k \begin{bmatrix} d \\ \lambda \end{bmatrix} = \begin{bmatrix} -\nabla f(x^k) \\ v^k \end{bmatrix}, \tag{2.9}$$

where if FAST = TRUE, then

$$v_i^k = -\delta_k c_i(x^k) - \mu_i^k \rho_k, \ i \in I_k, \tag{2.10}$$

with

$$\rho_k = \frac{(\theta-1)\nabla f(x^k)^\top d^{k,1}}{1+\sum_{i\in I_k}|\lambda_i^{k,0}|\|d^{k,1}\|^\eta}\|d^{k,1}\|^\eta; \tag{2.11}$$

else

$$v_i^k = \lambda_{-,i}^{k,0} = \min\{\lambda_i^{k,0}, 0\}, \quad i \in I_k. \tag{2.12}$$

(iv) Compute $(d^{k,3}, \lambda_{I_k}^{k,3})$ by solving the following linear system in (d, λ):

$$M_k \begin{bmatrix} d \\ \lambda \end{bmatrix} = \begin{bmatrix} -\nabla f(x^k) \\ \varpi^k \end{bmatrix}, \tag{2.13}$$

where if FAST = TRUE, then

$$\varpi_i^k = v_i^k - \mu_i^k c_i(x^k + d^{k,2}) - \pi_k, \quad i \in I_k, \tag{2.14}$$

with

$$\pi_k = \max\left\{\|d^{k,2}\|^\nu, \max_{i\in I_k}\left\{\left|1 - \frac{\mu_i^k}{\max\{\frac{\delta_k}{2}, \delta_k + \lambda_i^{k,2}\}}\right|^\tau \|d^{k,2}\|^2\right\}\right\}; \tag{2.15}$$

else

$$\varpi_i^k = v_i^k - \mu_i^k \rho_k, \quad i \in I_k, \tag{2.16}$$

with

$$\rho_k = \frac{(\theta-1)\nabla f(x^k)^\top d^{k,2}}{1+\sum_{i\in I_k}|\lambda_i^{k,0}|\|d^{k,2}\|^2}\|d^{k,2}\|^2. \tag{2.17}$$

Step 2 (Arc search).

Set $\lambda_{I\setminus I_k}^{k,j} = 0$ $(j = 0, 1, 2, 3)$. If FAST = TRUE, then set $(d^k, \lambda^k) = (d^{k,2}, \lambda^{k,2})$ and

$$(\hat{d}^k, \hat{\lambda}^k) = \begin{cases} (d^{k,3}, \lambda^{k,3}) & \text{if } \|d^{k,3} - d^{k,2}\| \le \|d^{k,2}\|, \\ (d^{k,2}, \lambda^{k,2}) & \text{else.} \end{cases}$$

If FAST = FALSE, then set $(d^k, \lambda^k) = (\hat{d}^k, \hat{\lambda}^k) = (d^{k,3}, \lambda^{k,3})$.

Compute t_k, the first number t of the sequence $\{1, \beta, \beta^2, \dots\}$ satisfying

$$\begin{cases} f(x^k + td^k + t^2(\hat{d}^k - d^k)) - f(x^k) \le \sigma t \nabla f(x^k)^\top d^k, \\ c_i(x^k + td^k + t^2(\hat{d}^k - d^k)) \le 0, \ i \in I. \end{cases} \tag{2.18}$$

Step 3 (Update).

Set $x^{k+1} = x^k + t_k d^k + t_k^2(\hat{d}^k - d^k)$. If $\|\lambda^{k,0}\|_\infty > \chi_k$, set $\varepsilon_{k+1} = \frac{1}{2}\varepsilon_k$ and $\chi_{k+1} = 2\chi_k$; else set $(\varepsilon_{k+1}, \chi_{k+1}) = (\varepsilon_k, \chi_k)$. If $L_{k+1} \neq \emptyset$ and $\varphi(x^{k+1}, \lambda^{k,0}) > 0$, set $\delta_{k+1} = \vartheta \min\{\lambda_i^{k,0}, i \in L_{k+1}\}$; otherwise, set $\delta_{k+1} = \delta$. Set

$$\mu_i^{k+1} = \begin{cases} \delta_{k+1} + \max\{\lambda_i^{k,0}, 0\} & \text{if } i \in I_k, \\ \delta_{k+1} & \text{if } i \in I\setminus I_k. \end{cases} \tag{2.19}$$

Set $H_{k+1} = \nabla^2_{xx}\mathcal{L}(x^{k+1}, \mu^{k+1} - \delta_{k+1}e)$. Set $k = k + 1$ and go to Step 1.

Remark 2.1. Clearly, $\mu^k > 0$ at each iteration. Hence, by Lemma 6.1 in the appendix, condition (2.4) and the LICQ can guarantee the nonsingularity of M_k. To motivate fast local convergence, in Algorithm 2.1 the exact Lagrangian Hessian is used to compute the step if it satisfies (2.4). This is the case for iterates in a neighborhood of a KKT point of problem (P) satisfying the SSOSC; see section 4. However, outside such a neighborhood, it may be necessary to modify H_k so that (2.4) holds. There are two often used ways to do so. One way is to repeatedly add multiples of the identity to H_k until (2.4) holds; e.g., see [23, 24]. The other way is based on the inertia-controlling factorization of the matrix obtained by symmetrizing M_k that determines a positive quantity and those diagonal elements of H_k to which the positive quantity should be added so that the resulting $\hat{H}_k$ satisfies (2.4); e.g., see [12].

Remark 2.2. The role of linear equations (2.6) is to provide a descent direction $d^{k,0}$ of f and an approximate multiplier vector $\lambda^{k,0}$, which is used to compute the perturbation value ρ_k in (2.11) and (2.17). A system of inequalities (2.8) is introduced in the algorithm through which we can determine whether direction $d^{k,1}$, a "fast" local direction, is acceptable as an ideal search direction. The first inequality of (2.8) checks whether $d^{k,1}$ can provide a sufficient reduction in f, while the second inequality measures the convergence progress of $\lambda^{k,1}$ toward the optimal multipliers. If (2.8) holds (i.e., FAST = TRUE), another linear equation (2.9), a slight perturbation of (2.7), is solved to keep the feasibility of the next iterate. To avoid the Maratos effect, second-order correction is carried out by solving (2.13), a slight perturbation of (2.9). If (2.8) is violated, a steeper descent direction is gotten by (2.9) and feasibility of the next iterate is ensured by solving (2.13). Consequently, linear equations (2.9) and (2.13) may serve for different purposes depending on the value of "FAST." Note that the inequality system (2.8) actually provides an efficient framework for balancing well the global and local behavior of our algorithm. We believe that without much modification, this framework can be readily extended to other active set–based algorithms for solving problem (P).

The following assumptions guarantee that our algorithm is well defined.

ASSUMPTION A1. *$\mathcal{F}$ is nonempty.*

ASSUMPTION A2. *The LICQ holds on $\mathcal{F}$, i.e., vectors $\{\nabla c_i(x),\ i \in I(x)\}$ are linearly independent for any $x \in \mathcal{F}$.*

Assumption A1 ensures the existence of a feasible starting point x^1. Assumption A2 is a common assumption for both type-1 and type-2 QP-free methods. The rest of the section is devoted to showing that Algorithm 2.1 is well defined. It has been analyzed in Remark 2.1 that under Assumption A2, M_k is nonsingular for every k. Hence, $(d^{k,j}, \lambda^{k,j})$ $(j = 0, 1, 2, 3)$ are all well defined. In the following we continue to show that if the algorithm terminates at Step 1(i), the current iterate is a KKT point of problem (P); otherwise, the arc search in Step 2 is executable and the algorithm generates the next iterate.

LEMMA 2.1. *Under Assumptions A1 and A2, Algorithm* 2.1 *terminates at Step* 1(i), *i.e., $d^{k,0} = 0$ and $\lambda^{k,0} \geq 0$, if and only if x^k is a KKT point of problem* (P).

Proof. Suppose $d^{k,0} = 0$ and $\lambda^{k,0} \geq 0$ at the kth iteration. Then it follows from linear equations (2.6) that

$$\nabla f(x^k) + \sum_{i \in I_k} \lambda_i^{k,0} \nabla c_i(x^k) = 0, \quad \lambda_i^{k,0} c_i(x^k) = 0, \quad i \in I_k.$$

This implies that $(x^k, \lambda^{k,0})$ is a KKT point of problem (P) as $\lambda^{k,0}_{I \setminus I_k} = 0$.

Suppose x^k is a KKT point of problem (P) and $\bar{\lambda}$ is the corresponding Lagrangian multiplier vector. By Assumption A2 and the nonsingularity of M_k, we know that $(0, \bar{\lambda}_{I_k})$ is the unique solution of linear equations (2.6). Hence, we conclude that $d^{k,0} = 0$ and $\lambda^{k,0} = \bar{\lambda}$. □

In the remainder of the paper, we assume that Algorithm 2.1 generates an infinite sequence of iterates, i.e., it does not stop at Step 1(i) with a KKT point of problem (P). Before proceeding further, we need some basic relations that are useful for our analysis. To simplify the presentation, in the following algebra we omit superscript k, subscript k, and function variables. For example, C, ∇c, and ∇f will denote $C_{I_k}(x^k)$, $\nabla c_{I_k}(x^k)$, and $\nabla f(x^k)$, respectively. Consider the following linear equations:

$$\begin{cases} Hd + \nabla c\lambda = -\nabla f, \\ U\nabla c^\top d + C\lambda = w. \end{cases} \tag{2.20}$$

It follows from Lemma 6.1 that (2.20) is nonsingular and has a unique solution. Letting $a = I_k \backslash I(x^k)$ and $b = I(x^k)$, we have

$$\lambda_a = -C_a^{-1}U_a\nabla c_a^\top d + C_a^{-1}w_a, \tag{2.21}$$

$$\kappa\nabla c_b U_b \nabla c_b^\top d = \kappa\nabla c_b w_b, \tag{2.22}$$

where κ is a positive scalar. Substituting (2.21) into the first block of equations in (2.20) and adding both sides of (2.22) to the first block of equations in (2.20) gives

$$\bar{H}d = -\nabla f - \nabla c_a C_a^{-1} w_a + \kappa\nabla c_b w_b - \nabla c_b \lambda_b,$$

where

$$\bar{H} = H - \nabla c_a C_a^{-1} U_a \nabla c_a^\top + \kappa\nabla c_b U_b \nabla c_b^\top.$$

Since (2.4) holds and $U \succ 0$ in view of Step 3 of Algorithm 2.1, by Lemma 6.2 in the appendix, we can pick κ large enough so that $\bar{H} \succ 0$. Thus,

$$d = -\bar{H}^{-1}\nabla f - \bar{H}^{-1}\nabla c_a C_a^{-1} w_a + \kappa\bar{H}^{-1}\nabla c_b w_b - \bar{H}^{-1}\nabla c_b \lambda_b. \tag{2.23}$$

Substitute (2.23) into the second block of equations in (2.20) and we get

$$\lambda_b = -D^{-1}\nabla c_b^\top \bar{H}^{-1}\nabla f - D^{-1}\nabla c_b^\top \bar{H}^{-1}\nabla c_a C_a^{-1} w_a + \kappa w_b - D^{-1}U_b^{-1} w_b, \tag{2.24}$$

where $D = \nabla c_b^\top \bar{H}^{-1}\nabla c_b$. Note that Assumption A2 ensures the nonsingularity of D. Let (d^0, λ^0) be the solution of (2.20) when $w = 0$, i.e., the solution of (2.6). We have

$$\lambda_a^0 = (C_a^{-1}U_a\nabla c_a^\top \bar{H}^{-1} - C_a^{-1}U_a\nabla c_a^\top \bar{H}^{-1}\nabla c_b D^{-1}\nabla c_b^\top \bar{H}^{-1})\nabla f, \tag{2.25}$$

$$\lambda_b^0 = -D^{-1}\nabla c_b^\top \bar{H}^{-1}\nabla f, \tag{2.26}$$

and

$$\begin{aligned} d = d^0 &+ (\bar{H}^{-1}\nabla c_b D^{-1}\nabla c_b^\top \bar{H}^{-1}\nabla c_a C_a^{-1} - \bar{H}^{-1}\nabla c_a C_a^{-1}) w_a \\ &+ \bar{H}^{-1}\nabla c_b D^{-1} U_b^{-1} w_b. \end{aligned} \tag{2.27}$$

Consequently, we obtain from (2.25), (2.26), and (2.27) that

$$\nabla f^\top d = \nabla f^\top d^0 - (\lambda_a^0)^\top U_a^{-1} w_a - (\lambda_b^0)^\top U_b^{-1} w_b = \nabla f^\top d^0 - (\lambda^0)^\top U^{-1} w. \tag{2.28}$$

In the following lemmas, (2.28) is used to show the descent and feasible properties of search direction d^k. To this end, we first show that $d^{k,0}$ is a descent direction of f. Since $I(x^k) \subseteq I_k$, it follows from (2.6) that

$$\mu_i^k \nabla c_i(x^k)^\top d^{k,0} = -c_i(x^k)\lambda_i^{k,0} = 0 \qquad \forall i \in I(x^k). \tag{2.29}$$

Hence, $d^{k,0} \in \mathcal{T}(x^k)$ as $\mu^k > 0$ by Step 3 of Algorithm 2.1. Hence, from (2.6), (2.4), and (2.5), we have

$$\begin{aligned} &\nabla f(x^k)^\top d^{k,0} \\ &= -(d^{k,0})^\top \left(H_k - \sum_{i \in I_k \setminus I(x^k)} \frac{\mu_i^k}{c_i(x^k)} \nabla c_i(x^k) \nabla c_i(x^k)^\top \right) d^{k,0} \\ &\quad - \sum_{i \in I(x^k)} \lambda_i^{k,0} \nabla c_i(x^k)^\top d^{k,0} \\ &= -(d^{k,0})^\top \hat{H}_k d^{k,0} \le 0. \end{aligned} \tag{2.30}$$

LEMMA 2.2. *Suppose Assumptions A1 and A2 hold and* FAST = TRUE *at the kth iteration. Then* $\nabla f(x^k)^\top d^{k,1} \neq 0$.

Proof. Suppose $\nabla f(x^k)^\top d^{k,1} = 0$. Since FAST = TRUE, it follows from (2.8) that $d^{k,1} = 0$ and $\lambda_{I_k}^{k,1} \ge 0$. Hence, from linear equations (2.7) we obtain that

$$\begin{aligned} &\nabla f(x^k) + \sum_{i \in I_k} \lambda_i^{k,1} \nabla c_i(x^k) = 0, \\ &0 \le -\delta_k c_i(x^k) = \lambda_i^{k,1} c_i(x^k) \le 0, \quad i \in I_k. \end{aligned}$$

Since $I(x^k) \subseteq I_k$ and $\lambda_{I \setminus I_k}^{k,1} = 0$, we conclude that $(x^k, \lambda^{k,1})$ is a KKT point of problem (P). Hence, Lemma 2.1 implies that Algorithm 2.1 should have terminated at Step 1(i), a contradiction. □

LEMMA 2.3. *Under Assumptions A1 and A2, if* FAST = TRUE *at the kth iteration, then*

(i) $\nabla f(x^k)^\top d^k \le \theta \nabla f(x^k)^\top d^{k,1} < 0$;

(ii) $\nabla c_i(x^k)^\top d^k = -\rho_k < 0$ *for all* $i \in I(x^k)$.

Proof. Since FAST = TRUE, we know that (2.8) holds, $d^k = d^{k,2}$, and ρ_k is defined by (2.11). Since we have assumed that Algorithm 2.1 generates an infinite iterate sequence, it follows from Lemma 2.1 that x^k is not a KKT point of (P). Thus, it follows from Lemma 2.2 and (2.8) that $\nabla f(x^k)^\top d^{k,1} < 0$. Hence, from (2.10), (2.11), and (2.28), we obtain that

$$\begin{aligned} &\nabla f(x^k)^\top d^k \\ &= \nabla f(x^k)^\top d^{k,1} + (\lambda_{I_k}^{k,0})^\top U_k^{-1} U_k(\rho_k e) \\ &= \nabla f(x^k)^\top d^{k,1} + \frac{(\theta - 1)\nabla f(x^k)^\top d^{k,1}}{1 + \sum_{i \in I_k} |\lambda_i^{k,0}| \|d^{k,1}\|^\eta} \|d^{k,1}\|^\eta \sum_{i \in I_k} \lambda_i^{k,0} \\ &\le \theta \nabla f(x^k)^\top d^{k,1} < 0. \end{aligned} \tag{2.31}$$

Notice that $I(x^k) \subseteq I_k$. By (2.9) and (2.10), we have that $\nabla c_i(x^k)^\top d^k = -\rho_k < 0$ for $i \in I(x^k)$. □

LEMMA 2.4. *Under Assumptions A1 and A2, if* FAST = FALSE *at the kth iteration, then*

(i) $\nabla f(x^k)^\top d^{k,2} = \nabla f(x^k)^\top d^{k,0} - \sum_{i\in I_k}(\lambda^{k,0}_{-,i})^2/\mu^k_i < 0$;
(ii) $\nabla f(x^k)^\top d^k \le \theta \nabla f(x^k)^\top d^{k,2}$;
(iii) $\nabla c_i(x^k)^\top d^k \le -\rho_k < 0$ *for* $i \in I(x^k)$.

Proof. Since FAST = FALSE, we have that $d^k = d^{k,3}$ and v^k and ρ_k are computed by (2.12) and (2.17), respectively. It follows from (2.28) and (2.30) that

$$\begin{aligned} &\nabla f(x^k)^\top d^{k,2} \\ &= \nabla f(x^k)^\top d^{k,0} - (\lambda^{k,0}_{I_k})^\top U_k^{-1} v^k \\ &= \nabla f(x^k)^\top d^{k,0} - \sum_{i\in I_k}(\lambda^{k,0}_{-,i})^2/\mu^k_i \\ &= -(d^{k,0})^\top \hat{H}_k d^{k,0} - \sum_{i\in I_k}(\lambda^{k,0}_{-,i})^2/\mu^k_i. \end{aligned} \tag{2.32}$$

Since Algorithm 2.1 generates an infinite iterate sequence, x^k is not a KKT point of (P), i.e., either $d^{k,0} = 0$ or $\lambda^{k,0}_{I_k} \ge 0$ is violated in view of Lemma 2.1. Moreover, since $I(x^k) \subseteq I_k$, we have from (2.6) that (2.29) holds. Hence, $d^{k,0} \in \mathcal{T}(x^k)$ as $\mu^k > 0$. Consequently, we know from (2.4) and (2.32) that $\nabla f(x^k)^\top d^{k,2} < 0$. This establishes (i). From (2.16), (2.17), (2.28), and (2.13), we obtain that

$$\begin{aligned} &\nabla f(x^k)^\top d^k \\ &= \nabla f(x^k)^\top d^{k,2} + (\lambda^{k,0}_{I_k})^\top U_k^{-1} U_k(\rho_k e) \\ &= \nabla f(x^k)^\top d^{k,2} + \frac{(\theta-1)\nabla f(x^k)^\top d^{k,2}}{1+\sum_{i\in I_k}|\lambda^{k,0}_i|\|d^{k,2}\|^2}\|d^{k,2}\|^2 \sum_{i\in I_k}\lambda^{k,0}_i \\ &\le \theta \nabla f(x^k)^\top d^{k,2}. \end{aligned} \tag{2.33}$$

This establishes (ii). Moreover, it follows from (2.12), (2.16), and (2.13) that for $i \in I(x^k)$, $\nabla c_i(x^k)^\top d^k \le -\rho_k < 0$. ∎

PROPOSITION 2.5. *Under Assumptions A1 and A2, Algorithm 2.1 is well defined.*

Proof. Assumption A1 ensures that Algorithm 2.1 can start properly. At each iteration, $\mu^k > 0$ by Step 3 of Algorithm 2.1, and it is always possible to modify H_k, if necessary, so that (2.4) holds (e.g., it suffices to add a sufficiently large multiple of the identity to H_k). Since (2.4) and Assumption A2 guarantee the nonsingularity of M_k by Lemma 6.1, every linear system involved in Algorithm 2.1 has a unique solution. By Lemma 2.1, if Algorithm 2.1 terminates at an iteration k, x^k is a KKT point of problem (P). To finish our analysis, we only need to show that the arc search in Step 2 of Algorithm 2.1 is well defined.

Suppose Algorithm 2.1 does not terminate at x^k. Then by Lemmas 2.3 and 2.4, $\nabla f(x^k)^\top d^k < 0$ and $\nabla c_i(x^k)^\top d^k < 0$ for all $i \in I(x^k)$. Let $\tilde{d}^k = \hat{d}^k - d^k$. Since f and c are twice continuously differentiable, we have

$$\begin{aligned} &f(x^k + td^k + t^2\tilde{d}^k) = f(x^k) + t\nabla f(x^k)^\top d^k + O(t^2), \\ &c_i(x^k + td^k + t^2\tilde{d}^k) = c_i(x^k) + t\nabla c_i(x^k)^\top d^k + O(t^2),\ i \in I. \end{aligned}$$

Hence, it follows that for all sufficiently small t ($t \in (0,1)$),

$$f(x^k + td^k + t^2\tilde{d}^k) - f(x^k) \le \sigma t \nabla f(x^k)^\top d^k$$

as $\sigma \in (0, 1/2)$ and

$$c_i(x^k + td^k + t^2\tilde{d}^k) \le c_i(x^k) = 0 \quad \forall i \in I(x^k).$$

Moreover, since $c_i(x^k) < 0$ for $i \in I \backslash I(x^k)$, it follows that for such i and all sufficiently small t, $c_i(x^k + td^k + t^2\tilde{d}^k) < 0$. Therefore, we conclude that Step 2 of Algorithm 2.1 is well defined. Thus, Algorithm 2.1 can proceed to the next iterate. ∎

3. Global convergence. In this section we show that Algorithm 2.1 is globally convergent to KKT points of problem (P). To this end, we further assume the following.

ASSUMPTION A3. *There exist $\varrho_1, \varrho_2 > 0$ such that for all k,*

$$d^\top \hat{H}_k d \geq \varrho_1 \|d\|^2 \quad \forall d \in \mathcal{T}(x^k),$$

where $\hat{H}_k$ is defined by (2.5)*, and $\|H_k\| \leq \varrho_2$.*

ASSUMPTION A4. *The set $\mathcal{F} \cap \{x | f(x) \leq f(x^1)\}$ is compact.*

Assumption A3 is weaker than the uniform positive definiteness assumption on H_k made in [13, 18, 19, 28]. It holds with the exact Hessian in the neighborhood of any solution of problem (P) at which the SSOSC holds. Assumption A4 ensures that our descent algorithm generates accumulated iterates.

LEMMA 3.1. *Under Assumptions A1–A4, sequences $\{\chi_k\}$ and $\{\varepsilon_k\}$ are changed at most finitely many times.*

Proof. The proof is by contradiction. Suppose that $\{\chi_k\}$ and $\{\varepsilon_k\}$ are changed infinitely many times, i.e., there exists an infinite index set $\mathcal{K}$ such that $\chi_{k+1} = 2\chi_k$ and $\varepsilon_{k+1} = \frac{1}{2}\varepsilon_k$ for all $k \in \mathcal{K}$. Then we have $\{\chi_k\} \to +\infty$ and $\varepsilon_k \to 0^+$ as $k \to \infty$. Due to the finiteness of set I, we can assume without loss of generality that I_k are identical for all $k \in \mathcal{K}$ and let $I_{\mathcal{K}} = I_k$. Moreover, by Assumption A4 we can assume $x^k \to \bar{x}$ as $k \in \mathcal{K} \to \infty$. We get from the definition of I_k that $I_{\mathcal{K}} \subseteq I(\bar{x})$ since $\{\varepsilon_k \min\{\varphi(x^k, \lambda^{k-1,0}), \varphi_{\max}\}\} \to 0$ as $k \to \infty$.

In addition, the criteria that trigger updating of $\{\chi_k\}$ and $\{\varepsilon_k\}$ must be satisfied for all $k \in \mathcal{K}$, i.e., $\|\lambda^{k,0}\|_\infty > \chi_k$. This implies that $\|\lambda^{k,0}\|_\infty \to \infty$. Consequently, the sequence $\{\alpha_k\}$, with

$$\alpha_k = \max\{\|d^{k,0}\|,\ \|\lambda^{k,0}_{I_{\mathcal{K}}}\|_\infty,\ 1\},$$

tends to infinity on $\mathcal{K}$. Define

$$\bar{d}^k = \alpha_k^{-1} d^{k,0} \quad \text{and} \quad \bar{\lambda}^k_{I_{\mathcal{K}}} = \alpha_k^{-1} \lambda^{k,0}_{I_{\mathcal{K}}}$$

for $k \in \mathcal{K}$. By construction we have $\max\{\|\bar{d}^k\|, \|\bar{\lambda}^k_{I_{\mathcal{K}}}\|_\infty\} = 1$ for all $k \in \mathcal{K}$ large enough. Hence, there exists an infinite index set $\mathcal{K}_1 \subseteq \mathcal{K}$ and nonzero vector $(\bar{d}, \bar{\lambda}_{I_{\mathcal{K}}})$ such that $\bar{d}^k \to \bar{d}$ and $\bar{\lambda}^k_{I_{\mathcal{K}}} \to \bar{\lambda}_{I_{\mathcal{K}}}$ as $k \in \mathcal{K}_1 \to \infty$. From (2.30) and Assumption A3, we have

$$\nabla f(x^k)^\top d^{k,0} \leq -\varrho_1 \|d^{k,0}\|^2. \tag{3.1}$$

Dividing both sides of (3.1) by α_k^2 and letting $k \in \mathcal{K}_1 \to \infty$ yields $\bar{d} = 0$. Thus, $\bar{\lambda}_{I_{\mathcal{K}}}$ is nonzero. Besides, it follows from (2.6) that for $k \in \mathcal{K}$,

$$H_k d^{k,0} + \nabla c_{I_{\mathcal{K}}}(x^k) \lambda^{k,0}_{I_{\mathcal{K}}} = -\nabla f(x^k). \tag{3.2}$$

Dividing both sides of (3.2) by α_k and letting $k \in \mathcal{K}_1 \to \infty$ gives us

$$\nabla c_{I_{\mathcal{K}}}(\bar{x}) \bar{\lambda}_{I_{\mathcal{K}}} = 0.$$

This contradicts Assumption A2, as we have proved that $I_{\mathcal{K}} \subseteq I(\bar{x})$. ∎

LEMMA 3.2. *Under Assumptions A1–A4, sequences* $\{\lambda^{k,0}\}$, $\{\delta_k\}$, *and* $\{\mu^k\}$ *are bounded.*

Proof. Lemma 3.1 gives that $\{\chi_k\}$ has an upper bound, and thus $\{\lambda^{k,0}\}$ is bounded by Step 3 of Algorithm 2.1. The boundedness of $\{\delta_k\}$ and $\{\mu^k\}$ follows directly from their definitions and the boundedness of $\{\lambda^{k,0}\}$. ∎

LEMMA 3.3. *Under Assumptions A1–A4, if* $\{\delta_k\}_{\mathcal{K}} \to 0$, *then any accumulation point of* $\{(x^k, \lambda^{k-1,0})\}_{\mathcal{K}}$ *is a KKT point of problem* (P), *where* $\mathcal{K}$ *is an infinite index set.*

Proof. First note that $\{(x^k, \lambda^{k-1,0})\}_{\mathcal{K}}$ is bounded by Assumption A4 and Lemma 3.2. Since $\{\delta_k\}_{\mathcal{K}} \to 0$, the definitions of δ_k and L_k indicate that $\{\varphi(x^k, \lambda^{k-1,0})\}_{\mathcal{K}} \to 0$. Since $\varphi(x, \lambda) = 0$ if and only if (x, λ) is a KKT point of problem (P), the result follows from continuity. ∎

LEMMA 3.4. *Suppose Assumptions A1–A4 hold. If* $\{\delta_k\}_{\mathcal{K}}$ *is bounded below over zero, sequence* $\{\|M_k^{-1}\|\}_{\mathcal{K}}$ *is uniformly bounded, where* $\mathcal{K}$ *is an infinite index set.*

Proof. Since $\mu^k > 0$ by Algorithm 2.1, we know from Lemma 6.1 that (2.4) and Assumption A2 guarantee the nonsingularity of M_k for each k. Now assume to the contrary that there exists an infinite index set $\mathcal{K}' \subseteq \mathcal{K}$ such that $\|M_k^{-1}\| \to \infty$ as $k \in \mathcal{K}' \to \infty$. Due to Lemma 3.2 and Assumptions A3 and A4, we can assume that $\delta_k \to \bar{\delta} > 0$, $\mu^k \to \bar{\mu} > 0$, $x^k \to \bar{x} \in \mathcal{F}$, and $H_k \to H_*$ as $k \in \mathcal{K}' \to \infty$. Moreover, since set I is finite, we can further assume that I_k, L_k, and $I(x^k)$ are, respectively, identical for all $k \in \mathcal{K}'$ and let $I_{\mathcal{K}'} = I_k$, $L_{\mathcal{K}'} = L_k$, and $\bar{I}_{\mathcal{K}'} = I(x^k)$. Putting all the limits together, we have

$$\{M_k\}_{\mathcal{K}'} \to \bar{M} = \begin{bmatrix} H_* & \nabla c_{I_{\mathcal{K}'}}(\bar{x}) \\ \operatorname{diag}(\bar{\mu}_{I_{\mathcal{K}'}})\nabla c_{I_{\mathcal{K}'}}(\bar{x})^\top & \operatorname{diag}(c_{I_{\mathcal{K}'}}(\bar{x})) \end{bmatrix}.$$

Pick any $y \in \mathcal{T}(\bar{x})\backslash\{0\}$. Assumption A2 ensures that $\nabla c_{I(\bar{x})}(x^k)^\top$ has full row rank for all $k \in \mathcal{K}'$ large enough. So we can let

$$y^k = (E - \nabla c_{I(\bar{x})}(x^k)(\nabla c_{I(\bar{x})}(x^k)^\top \nabla c_{I(\bar{x})}(x^k))^{-1}\nabla c_{I(\bar{x})}(x^k)^\top)y$$

for sufficiently large $k \in \mathcal{K}'$. Obviously, we have $\nabla c_{I(\bar{x})}(x^k)^\top y^k = 0$ for all large $k \in \mathcal{K}'$ and $y^k \to y$ as $k \in \mathcal{K}' \to \infty$. This implies that $y^k \in \mathcal{T}(x^k)$ since $\bar{I}_{\mathcal{K}'} \subseteq I(\bar{x})$ when x^k is close to $\bar{x}$. Therefore, we obtain by Assumption A3 that for $k \in \mathcal{K}'$ large enough

$$\begin{aligned} &(y^k)^\top \left(H_k - \sum_{I_{\mathcal{K}'}\backslash I(\bar{x})} \frac{\mu_i^k}{c_i(x^k)} \nabla c_i(x^k)\nabla c_i(x^k)^\top \right) y^k \\ &= (y^k)^\top \left(H_k - \sum_{I_{\mathcal{K}'}\backslash \bar{I}_{\mathcal{K}'}} \frac{\mu_i^k}{c_i(x^k)} \nabla c_i(x^k)\nabla c_i(x^k)^\top \right) y^k \\ &= (y^k)^\top \hat{H}_k y^k \geq \varrho_1 \|y^k\|^2. \end{aligned}$$

Thus, letting $k \in \mathcal{K}' \to \infty$ yields that

$$y^\top \left(H_* - \sum_{I_{\mathcal{K}'}\backslash I(\bar{x})} \frac{\bar{\mu}_i}{c_i(\bar{x})} \nabla c_i(\bar{x})\nabla c_i(\bar{x})^\top \right) y > 0.$$

Now we can use Lemma 6.1 to show that $\bar{M}$ is nonsingular, a contradiction. ∎

The following corollary is a direct consequence of Lemmas 3.2 and 3.4.

COROLLARY 3.5. *Suppose Assumptions A1–A4 hold. If $\{\delta_k\}_{\mathcal{K}}$ is bounded below over zero, sequences $\{d^{k,0}\}$, $\{d^{k,j}, \lambda^{k,j}\}$ $(j = 1, 2, 3)$ are all bounded, where $\mathcal{K}$ is an infinite index set.*

Proof. Since the matrix sequence $\{M_k^{-1}\}_{\mathcal{K}}$ is uniformly bounded by Lemma 3.4 and $\{x^k\}$ is bounded due to Assumption A4, we obtain the boundedness of $\{d^{k,0}\}$ and $\{(d^{k,1}, \lambda^{k,1})\}_{\mathcal{K}}$ from (2.6) and (2.7), respectively. This together with Lemma 3.2 implies the boundedness of $\{v^k\}_{\mathcal{K}}$ in view of (2.10), (2.11), and (2.12). Hence, $\{(d^{k,2}, \lambda^{k,2})\}_{\mathcal{K}}$ is bounded, and this further implies the boundedness of $\{\varpi^k\}_{\mathcal{K}}$ in view of (2.14)–(2.17). Hence, $\{(d^{k,3}, \lambda^{k,3})\}_{\mathcal{K}}$ is also bounded. ∎

LEMMA 3.6. *Suppose Assumptions A1–A4 hold. If $\{\delta_k\}_{\mathcal{K}}$ is bounded below over zero and there exists a constant $\alpha > 0$ such that $\nabla f(x^k)^\top d^k \le -\alpha$ for all $k \in \mathcal{K}$, where $\mathcal{K}$ is an infinite index set, then there exists a constant $\bar{\alpha} > 0$ such that $\rho_k \ge \bar{\alpha}$ for all $k \in \mathcal{K}$ large enough.*

Proof. Assume to the contrary that there exists an infinite subset $\mathcal{K}' \subseteq \mathcal{K}$ such that $\{\rho_k\} \to 0$ as $k \in \mathcal{K}' \to \infty$. There are two cases.

Case 1. There exists an infinite subset $\mathcal{K}_1 \subseteq \mathcal{K}'$ such that FAST = TRUE for all $k \in \mathcal{K}_1$. In this case, $d^k = d^{k,2}$ and ρ_k is defined by (2.11). Since ρ_k tends to zero as $k \in \mathcal{K}' \to \infty$, it follows from (2.11) that $\{\nabla f(x^k)^\top d^{k,1}\}_{\mathcal{K}_1} \to 0$. Considering the finiteness of I, we can assume that I_k are identical for all $k \in \mathcal{K}_1$ and let $I_{\mathcal{K}_1} = I_k$. Moreover, due to boundedness we can further assume that $\{\delta_k\} \to \bar{\delta} > 0$, $\{M_k\} \to \bar{M}$, $\{(d^{k,1}, \lambda^{k,1})\} \to (\bar{d}, \bar{\lambda})$, and $\{(d^k, \lambda^k)\} \to (\bar{d}', \bar{\lambda}')$ as $k \in \mathcal{K}_1 \to \infty$. Now letting $k \in \mathcal{K}_1 \to \infty$ in linear systems (2.7), (2.9), and (2.10) yields that

$$\bar{M} \begin{bmatrix} \bar{d} \\ \bar{\lambda}_{I_{\mathcal{K}_1}} \end{bmatrix} = \bar{M} \begin{bmatrix} \bar{d}' \\ \bar{\lambda}'_{I_{\mathcal{K}_1}} \end{bmatrix} = \begin{bmatrix} -\nabla f(\bar{x}) \\ -\bar{\delta} c_{I_{\mathcal{K}_1}}(\bar{x}) \end{bmatrix}.$$

This implies that $(\bar{d}, \bar{\lambda}) = (\bar{d}', \bar{\lambda}')$ as $\bar{M}$ is nonsingular by Lemma 3.4. Hence, we have

$$\{\nabla f(x^k)^\top d^k\}_{\mathcal{K}_1} \to \nabla f(\bar{x})^\top \bar{d}' = \nabla f(\bar{x})^\top \bar{d} = 0.$$

This contradicts the condition that $\nabla f(x^k)^\top d^k \le -\alpha$ for all $k \in \mathcal{K}$.

Case 2. There exists an infinite subset $\mathcal{K}_2 \subseteq \mathcal{K}'$ such that FAST = FALSE for all $k \in \mathcal{K}_2$. In this case, $d^k = d^{k,3}$ and ρ_k is defined by (2.17). Since $\{\rho_k\} \to 0$ as $k \in \mathcal{K}' \to \infty$, (2.17) gives that $\{\nabla f(x^k)^\top d^{k,2}\}_{\mathcal{K}_2} \to 0$. Without loss of generality, assume that $\{d^{k,2}\}_{\mathcal{K}_2} \to \hat{d}$ and $\{d^k\}_{\mathcal{K}_2} \to \hat{d}'$. By following a similar argument to that of Case 1, we can show that $\hat{d} = \hat{d}'$ and thus $\{\nabla f(x^k)^\top d^k\}_{\mathcal{K}_2} \to 0$, a contradiction.

Since either Case 1 or Case 2 happens, the result follows immediately. ∎

LEMMA 3.7. *Suppose Assumptions A1–A4 hold, $\{x^k\}_{\mathcal{K}} \to \bar{x}$, and $\{\delta_k\}_{\mathcal{K}}$ is bounded below over zero, where $\mathcal{K}$ is an infinite index set. If $I(\bar{x}) \subseteq I_k$ for all $k \in \mathcal{K}$ large enough, then $\{\nabla f(x^k)^\top d^k\} \to 0$ as $k \in \mathcal{K} \to \infty$.*

Proof. Assume to the contrary that there exist an infinite index set $\mathcal{K}' \subseteq \mathcal{K}$ and a constant $\alpha > 0$ such that

$$\nabla f(x^k)^\top d^k \le -\alpha \quad \forall k \in \mathcal{K}'. \tag{3.3}$$

Then it follows from Lemma 3.6 that there exists a constant $\bar{\alpha} > 0$ such that $\rho_k \ge \bar{\alpha}$ for all $k \in \mathcal{K}'$. Since I is finite, we can assume that I_k are identical for all $k \in \mathcal{K}'$ and let $I_{\mathcal{K}'} = I_k$. Due to the boundedness of $\{d^k\}$, we can also assume that $\{d^k\} \to \bar{d} \ne 0$ as $k \in \mathcal{K}' \to \infty$. It follows from linear equations (2.9) and (2.13) that for each $i \in I_{\mathcal{K}'}$,

$$\mu_i^k \nabla c_i(x^k)^\top d^k + \lambda_i^k c_i(x^k) = \begin{cases} -\delta_k c_i(x^k) - \mu_i^k \rho_k & \text{if FAST = TRUE;} \\ \lambda_{-,i}^{k,0} - \mu_i^k \rho_k & \text{if FAST = FALSE.} \end{cases} \tag{3.4}$$

Since $I(\bar{x}) \subseteq I_{\mathcal{K}'}$, dividing both sides of (3.4) by μ_i^k and letting $k \in \mathcal{K}' \to \infty$ yields

$$\nabla c_i(\bar{x})^\top \bar{d} \leq -\inf_{k\in\mathcal{K}'}\{\rho_k\} \leq -\bar{\alpha} \quad \forall i \in I(\bar{x}).$$

This implies that for all $k \in \mathcal{K}'$ large enough,

$$\nabla c_i(x^k)^\top d^k \leq -\frac{1}{2}\bar{\alpha} \quad \forall i \in I(\bar{x}). \tag{3.5}$$

From (3.3) and the proof of [18, Lemma 3.9], we have the following basic relations (because of the differentiability of the functions) for the functions f, c and $k \in \mathcal{K}'$ large enough:

$$\begin{aligned} & f(x^k + td^k + t^2(\hat{d}^k - d^k)) - f(x^k) - \sigma t \nabla f(x^k)^\top d^k \\ \leq\ & t\Big\{ \sup_{\xi\in[0,1]} \|\nabla f(x^k + t\xi d^k + t^2\xi^2(\hat{d}^k - d^k)) - \nabla f(x^k)\|\|d^k\| \\ & +2t\ \sup_{\xi\in[0,1]} \|\nabla f(x^k + t\xi d^k + t^2\xi^2(\hat{d}^k - d^k))\|\|\hat{d}^k - d^k\| \\ & -(1-\sigma)\alpha\Big\}, \end{aligned} \tag{3.6}$$

$$\begin{aligned} & c_i(x^k + td^k + t^2(\hat{d}^k - d^k)) - c_i(x^k) \\ \leq\ & t\Big\{ \sup_{\xi\in[0,1]} \|\nabla c_i(x^k + t\xi d^k + t^2\xi^2(\hat{d}^k - d^k)) - \nabla c_i(x^k)\|\|d^k\| \\ & +2t\ \sup_{\xi\in[0,1]} \|\nabla c_i(x^k + t\xi d^k + t^2\xi^2(\hat{d}^k - d^k))\|\|\hat{d}^k - d^k\| \\ & +\nabla c_i(x^k)^\top d^k\Big\}, \quad i \in I. \end{aligned} \tag{3.7}$$

Note that $\{\hat{d}^k - d^k\}$ is also bounded in view of the relation $\|\hat{d}^k - d^k\| \leq \|d^k\|$. Hence, (3.6) implies that there exists $t_f > 0$, independent of k, such that, for all $k \in \mathcal{K}'$ large enough,

$$f(x^k + td^k + t^2(\hat{d}^k - d^k)) - f(x^k) - \sigma t \nabla f(x^k)^\top d^k \leq 0 \quad \forall t \in (0, t_f]. \tag{3.8}$$

Moreover, (3.5) and (3.7) imply that there exists $t_c > 0$, independent of k, such that for $t \in (0, t_c]$ and $k \in \mathcal{K}'$ large enough,

$$c_i(x^k + td^k + t^2(\hat{d}^k - d^k)) < 0 \quad \forall i \in I(\bar{x}).$$

For all $i \in I\backslash I(\bar{x})$, there exists $\hat{\alpha} > 0$ such that $c_i(x^k) \leq -\hat{\alpha}$ for all $k \in \mathcal{K}'$ large enough. Since d^k and $\hat{d}^k$ are bounded, there exists $\bar{t}_c > 0$, independent of k, such that

$$c_i(x^k + td^k + t^2(\hat{d}^k - d^k)) < 0 \quad \forall i \in I\backslash I(\bar{x})$$

for all $t \in (0, \bar{t}_c]$ and $k \in \mathcal{K}'$ large enough.

Let $\bar{t} = \min\{t_f, \bar{t}_c, t_c\}$. By the arc search step of Algorithm 2.1, we have $t_k \geq \beta\bar{t}$ for all $k \in \mathcal{K}'$ large enough. Thus, it follows from (3.3) and (3.8) that

$$f(x^k + t_k d^k + t_k^2(\hat{d}^k - d^k)) - f(x^k) \leq -\sigma\beta\bar{t}\alpha,$$

which implies $f(x^k) \to -\infty$, a contradiction to Assumption A4 and the assumption that f is real valued and continuously differentiable on $\mathcal{F}$. ∎

LEMMA 3.8. *Suppose Assumptions A1–A4 hold,* $\{x^k, \lambda^{k,0}\}_{\mathcal{K}} \to (\bar{x}, \bar{\lambda})$, *and* $\{\delta_k\}_{\mathcal{K}}$ *is bounded below over zero, where* $\mathcal{K}$ *is an infinite index set. If* $\{\nabla f(x^k)^\top d^k\}_{\mathcal{K}} \to 0$, *then* $(\bar{x}, \bar{\lambda})$ *is a KKT point of problem* (P).

Proof. It is obvious that $\bar{x} \in \mathcal{F}$. There are two cases.

Case 1. There exists an infinite subset $\mathcal{K}_1 \subseteq \mathcal{K}$ such that FAST = TRUE for all $k \in \mathcal{K}_1$. In this case, $d^k = d^{k,2}$ and $(d^{k,1}, \lambda^{k,1}_{I_k})$ satisfies system (2.8). Since I is finite, we can assume that I_k are identical for all $k \in \mathcal{K}_1$ and let $I_{\mathcal{K}_1} = I_k$. Since $\nabla f(x^k)^\top d^k \le \theta \nabla f(x^k)^\top d^{k,1}$ by Lemma 2.3(i), it follows that $\{\nabla f(x^k)^\top d^{k,1}\} \to 0$ as $k \in \mathcal{K}_1 \to \infty$. Without loss of generality, suppose $\{\delta_k\}_{\mathcal{K}_1} \to \bar{\delta} > 0$ and $\{\lambda^{k,1}\}_{\mathcal{K}_1} \to \lambda'$. Letting $k \in \mathcal{K}_1 \to \infty$ in (2.8) yields that $\{d^{k,1}\}_{\mathcal{K}_1} \to 0$ and $\lambda'_{I_{\mathcal{K}_1}} \ge 0$. Moreover, we obtain from linear equations (2.7) that

$$\nabla f(\bar{x}) + \sum_{i \in I_{\mathcal{K}_1}} \lambda'_i \nabla c_i(\bar{x}) = 0 \quad \text{and} \quad c_i(\bar{x}) = 0 \quad \forall i \in I_{\mathcal{K}_1}.$$

Since $\lambda'_{I \setminus I_{\mathcal{K}_1}} = 0$, we conclude that $(\bar{x}, \lambda')$ is a KKT point of problem (P). Furthermore, since $\{d^{k,0}\}$ and all elements of M_k are bounded, we can assume that $\{d^{k,0}\} \to \bar{d}$ and $\{M_k\} \to \bar{M}$ as $k \in \mathcal{K}_1 \to \infty$. Due to the nonsingularity of $\bar{M}$, letting $k \in \mathcal{K}_1 \to \infty$ in linear equations (2.6) and (2.7) yields that $\bar{d} = 0$ and $\bar{\lambda}_{I_{\mathcal{K}_1}} = \lambda'_{I_{\mathcal{K}_1}}$. Hence, $(\bar{x}, \bar{\lambda})$ is a KKT point of problem (P).

Case 2. There exists an infinite subset $\mathcal{K}_2 \subseteq \mathcal{K}$ such that FAST = FALSE for all $k \in \mathcal{K}_2$. In this case, $d^k = d^{k,3}$. Similarly to Case 1, we assume that I_k are identical for all $k \in \mathcal{K}_2$ and let $I_{\mathcal{K}_2} = I_k$. Since $\nabla f(x^k)^\top d^k \le \theta \nabla f(x^k)^\top d^{k,2}$ by Lemma 2.4(ii), it follows that $\{\nabla f(x^k)^\top d^{k,2}\} \to 0$ as $k \in \mathcal{K}_2 \to \infty$. Notice that $\{\mu^k\}_{\mathcal{K}_2}$ is bounded and componentwise bounded below over zero. Therefore, we obtain from (2.30) and Lemma 2.4(i) that $\{d^{k,0}\}_{\mathcal{K}_2} \to 0$ and $\bar{\lambda}_{I_{\mathcal{K}_2}} \ge 0$. Furthermore, letting $k \in \mathcal{K}_2 \to \infty$ in linear equations (2.6) yields that

$$\nabla f(\bar{x}) + \sum_{i \in I_{\mathcal{K}_2}} \bar{\lambda}_i \nabla c_i(\bar{x}) = 0 \quad \text{and} \quad \bar{\lambda}_i c_i(\bar{x}) = 0 \quad \forall i \in I_{\mathcal{K}_2}.$$

Since $\bar{\lambda}_{I \setminus I_{\mathcal{K}_2}} = 0$, we conclude that $(\bar{x}, \bar{\lambda})$ is a KKT point of problem (P).

Since either Case 1 or Case 2 happens, the result follows immediately. ∎

Now we are in a position to establish the global convergence of Algorithm 2.1. By Lemma 2.1 we have known that if Algorithm 2.1 terminates in a finite number of iterations, a KKT point of problem (P) is found. The following theorem deals only with the case of infinite iterations.

THEOREM 3.9. *Suppose Assumptions A1–A4 hold. Let* $(\bar{x}, \hat{\lambda}, \bar{\lambda})$ *be an accumulation point of sequence* $\{x^k, \lambda^{k-1,0}, \lambda^{k,0}\}$ *generated by Algorithm* 2.1, *and let* $\mathcal{K}$ *be the corresponding index set such that* $\{x^k, \lambda^{k-1,0}, \lambda^{k,0}\}_{\mathcal{K}} \to (\bar{x}, \hat{\lambda}, \bar{\lambda})$. *Then either* $(\bar{x}, \hat{\lambda})$ *or* $(\bar{x}, \bar{\lambda})$ *is a KKT point of problem* (P).

Proof. First note that Lemma 3.2 and Assumption A4 ensure the existence of an accumulation point of $\{x^k, \lambda^{k-1,0}, \lambda^{k,0}\}$. Lemma 3.3 states that if there exists an infinite subset $\mathcal{K}' \subseteq \mathcal{K}$ such that $\{\delta_k\}_{\mathcal{K}'} \to 0$, then $(\bar{x}, \hat{\lambda})$ is a KKT point of problem (P). Therefore, we only need to consider the case that $\{\delta_k\}_{\mathcal{K}}$ is bounded below over zero. Assume to the contrary that neither $(\bar{x}, \hat{\lambda})$ nor $(\bar{x}, \bar{\lambda})$ is a KKT point of problem (P). Then there exists a constant $\alpha > 0$ such that $\varphi(x^k, \lambda^{k-1,0}) \ge \alpha$ for all $k \in \mathcal{K}$ large enough. Moreover, Lemma 3.1 shows that the sequence $\{\varepsilon_k\}$ has a lower bound $\bar{\varepsilon} > 0$. Therefore, the definition of working set indicates that $I(\bar{x}) \subseteq I_k$ for all $k \in \mathcal{K}$ large enough. Then by Lemma 3.7 we have that $\{\nabla f(x^k)^\top d^k\} \to 0$ as $k \in \mathcal{K} \to \infty$.

Thus, it follows from Lemma 3.8 that $(\bar{x}, \bar{\lambda})$ is a KKT point of problem (P). This contradicts the assumption that $(\bar{x}, \bar{\lambda})$ is not a KKT point. ∎

4. Fast local convergence. Suppose $(x^*, \hat{\lambda}, \bar{\lambda})$ is an accumulation point of the sequence $\{(x^k, \lambda^{k-1,0}, \lambda^{k,0})\}$ generated by Algorithm 2.1. It follows from Theorem 3.9 that either $(x^*, \hat{\lambda})$ or $(x^*, \bar{\lambda})$ is a KKT point of problem (P). Since by Assumption A2 there exists a unique multiplier vector corresponding to x^*, we have either $\hat{\lambda} = \lambda^*$ or $\bar{\lambda} = \lambda^*$. To begin our analysis, we further assume that $f,\ c_i,\ i \in I$, are twice continuously differentiable and $\nabla^2 f,\ \nabla^2 c_i,\ i \in I$, are locally Lipschitz continuous in a neighborhood of x^*. The following assumption is used to guarantee that x^* is an isolated accumulation point of $\{x^k\}$.

ASSUMPTION A5. *The SSOSC holds at (x^*, λ^*), i.e., the Lagrangian Hessian $\nabla^2_{xx}\mathcal{L}(x^*, \lambda^*)$ is positive definite on the null space*

$$\{y \in \Re^n | \nabla c_i(x^*)^\top y = 0 \quad \forall i \in I^+(x^*)\}.$$

From Lemma 3.2 we know that ε_k are changed at most finitely many times. Therefore, we can assume $\varepsilon_k = \bar{\varepsilon}$ for all k large enough. The following lemma follows directly from [4, Theorems 2.3, 2.4, and 3.7].

LEMMA 4.1. *Under Assumptions A2 and A5, there exists a neighborhood of (x^*, λ^*) such that, for each (x, λ) in this neighborhood, $A_{\bar{\varepsilon},\varphi_{\max}}(x, \lambda) = I(x^*)$ and $\Lambda_{\bar{\varepsilon},\varphi_{\max}}(x, \lambda) = I^+(x^*)$.*

LEMMA 4.2. *Suppose Assumptions A1–A5 hold and that $\{(x^k, \lambda^{k-1,0}, \lambda^{k,0})\}_{\mathcal{K}} \to (x^*, \hat{\lambda}, \bar{\lambda})$. Then $I(x^*) \subseteq I_k$ for all $k \in \mathcal{K}$ large enough and $\bar{\lambda} = \lambda^*$.*

Proof. If $\hat{\lambda} = \lambda^*$, it follows from Theorem 3.9 and Lemma 4.1 that for all $k \in \mathcal{K}$ large enough, $I_k = I(x^*)$. If $\hat{\lambda} \neq \lambda^*$, then $(x^*, \hat{\lambda})$ is not a KKT point of problem (P). The properties of φ imply that there exists $\alpha > 0$ such that $\varphi(x^k, \lambda^{k-1,0}) \geq \alpha$ for all $k \in \mathcal{K}$ large enough. Hence, $I(x^*) \subseteq I_k$ for all $k \in \mathcal{K}$ large enough. Now we show that $\bar{\lambda} = \lambda^*$. Since $I(x^*) \subseteq I_k$ for all sufficiently large $k \in \mathcal{K}$, it follows from Lemma 3.7 that $\{\nabla f(x^k)^\top d^k\} \to 0$ as $k \in \mathcal{K} \to \infty$. Hence, we know from Lemma 3.8 that $(x^*, \bar{\lambda})$ is a KKT point of problem (P). Thus, by Assumption A2, we have $\bar{\lambda} = \lambda^*$. ∎

LEMMA 4.3. *Suppose Assumptions A1–A5 hold and that $\{(x^k, \lambda^{k-1,0}, \lambda^{k,0})\}_{\mathcal{K}} \to (x^*, \hat{\lambda}, \bar{\lambda})$. Then $\{d^k\}_{\mathcal{K}} \to 0$.*

Proof. Consider any subsequence $\{\lambda^{k-1,0}\}_{\bar{\mathcal{K}}}$ with $\bar{\mathcal{K}} \subseteq \mathcal{K}$. If it converges to λ^*, by Lemma 4.1, we have $L_k = I^+(x^*)$ for $k \in \bar{\mathcal{K}}$ large enough. Hence, if $I^+(x^*) \neq \emptyset$, then $\delta_k > \frac{1}{2}\min\{\lambda_i^*,\ i \in I^+(x^*)\}$ for large $k \in \bar{\mathcal{K}}$; otherwise, $\delta_k = \delta > 0$ due to Step 3 of Algorithm 2.1. If $\{\lambda^{k-1,0}\}_{\bar{\mathcal{K}}}$ does not converge to λ^*, then $\{\varphi(x^k, \lambda^{k-1,0})\}_{\bar{\mathcal{K}}}$ is bounded below over zero and so is $\{\delta_k\}_{\bar{\mathcal{K}}}$ by the definition of δ_k. Hence, we conclude that $\{\delta_k\}_{\mathcal{K}}$ is uniformly bounded below over zero and then by Lemma 3.4 that $\{\|M_k^{-1}\|\}_{\mathcal{K}}$ is uniformly bounded.

From Lemma 4.2 we know that $I(x^*) \subseteq I_k$ for all $k \in \mathcal{K}$ large enough. Hence it follows from Lemma 3.7 that $\{\nabla f(x^k)^\top d^k\} \to 0$ as $k \in \mathcal{K} \to \infty$. Now consider any infinite index set $\mathcal{K}' \subseteq \mathcal{K}$ in which $\{d^k\} \to \bar{d}$. There are two cases.

Case 1. There exists an infinite subset $\mathcal{K}_1 \subseteq \mathcal{K}'$ such that FAST = TRUE for all $k \in \mathcal{K}_1$. In this case, $d^k = d^{k,2}$ and ρ_k is defined by (2.11). Without loss of generality, we can assume that $\{\lambda^{k,1}\}_{\mathcal{K}_1} \to \tilde{\lambda}$, $\{\lambda^{k,2}\}_{\mathcal{K}_1} \to \breve{\lambda}$ and $\{\delta_k\}_{\mathcal{K}_1} \to \delta_* > 0$. Moreover, since I is finite, we can assume that I_k are identical for all $k \in \mathcal{K}_1$ and let $I_{\mathcal{K}_1} = I_k$. Since all elements of M_k are bounded, we can further assume that $\{M_k\}_{\mathcal{K}_1} \to \bar{M}$. By Lemma 2.3(i) we have that $\{\nabla f(x^k)^\top d^{k,1}\}_{\mathcal{K}_1} \to 0$ as $\{\nabla f(x^k)^\top d^k\}_{\mathcal{K}_1} \to 0$. Thus, (2.8) and (2.11) imply that $\{d^{k,1}\}_{\mathcal{K}_1} \to 0$ and $\{\rho_k\}_{\mathcal{K}_1} \to 0$. Letting $k \in \mathcal{K}_1 \to \infty$ in

linear equations (2.7) and (2.9) yields that

$$\bar{M}\begin{bmatrix} 0 \\ \tilde{\lambda}_{I_{\mathcal{K}_1}} \end{bmatrix} = \bar{M}\begin{bmatrix} \bar{d} \\ \breve{\lambda}_{I_{\mathcal{K}_1}} \end{bmatrix} = \begin{bmatrix} -\nabla f(x^*) \\ -\delta_* c_{I_{\mathcal{K}_1}}(x^*) \end{bmatrix}.$$

Since we have proved that $\{\|M_k^{-1}\|\}_{\mathcal{K}_1}$ is bounded, i.e., $\bar{M}$ is nonsingular, we get that $\bar{d} = 0$.

Case 2. There exists an infinite subset $\mathcal{K}_2 \subseteq \mathcal{K}'$ such that FAST = FALSE for all $k \in \mathcal{K}_2$. In this case, $d^k = d^{k,3}$, and v^k, ϖ^k, and ρ_k are defined by (2.12), (2.16), and (2.17), respectively. Similarly to the analysis in Case 1, we can assume, without loss of generality, that $\{\lambda^{k,3}\}_{\mathcal{K}_2} \to \lambda'$, $\{M_k\}_{\mathcal{K}_2} \to \hat{M}$ and that I_k are identical for all $k \in \mathcal{K}_2$ and let $I_{\mathcal{K}_2} = I_k$. Lemma 2.4(ii) implies that $\{\nabla f(x^k)^\top d^{k,2}\}_{\mathcal{K}_2} \to 0$ as $\{\nabla f(x^k)^\top d^k\}_{\mathcal{K}} \to 0$. Thus, we know from (2.17), Lemma 2.4(i), (2.30), and Assumption A3 that $\{\rho_k\}_{\mathcal{K}_2} \to 0$ and $\{d^{k,0}\}_{\mathcal{K}_2} \to 0$. Hence, it follows from (2.12) and (2.16) that $\{\varpi^k\}_{\mathcal{K}_2} \to 0$. Letting $k \in \mathcal{K}_2 \to \infty$ in linear equations (2.6) and (2.13) yields that

$$\hat{M}\begin{bmatrix} 0 \\ \bar{\lambda}_{I_{\mathcal{K}_2}} \end{bmatrix} = \hat{M}\begin{bmatrix} \bar{d} \\ \lambda'_{I_{\mathcal{K}_2}} \end{bmatrix} = \begin{bmatrix} -\nabla f(x^*) \\ 0 \end{bmatrix}.$$

Again, the nonsingularity of $\hat{M}$ implies that $\bar{d} = 0$.

Since either Case 1 or Case 2 happens, the result follows immediately. ∎

The original version of the next result is due to Moré and Sorensen [17]; here we cite a slightly different version of the result from [16, Proposition 5.4].

PROPOSITION 4.4. *Assume that $\omega^* \in \Re^t$ is an isolated accumulation point of a sequence $\{\omega^k\} \subset \Re^t$ such that for every subsequence $\{\omega^k\}_{\mathcal{K}}$ converging to ω^*, there is an infinite subset $\bar{\mathcal{K}} \subseteq \mathcal{K}$ such that $\{\|\omega^{k+1} - \omega^k\|\}_{\bar{\mathcal{K}}} \to 0$. Then the whole sequence $\{\omega^k\}$ converges to ω^*.*

LEMMA 4.5. *Under Assumptions A1–A5, the whole sequence $\{(x^k, \lambda^{k,0})\}$ converges to (x^*, λ^*).*

Proof. The SSOSC and the LICQ guarantee that x^* is an isolated accumulation point of $\{x^k\}$; see [21]. Let $\{x^k\}_{\mathcal{K}}$ be a subsequence of iterates converging to x^*. It follows from Lemma 4.3 that there exists a subsequence $\mathcal{K}' \subseteq \mathcal{K}$ such that $\{d^k\}_{\mathcal{K}'} \to 0$. Since

$$\|x^{k+1} - x^k\| \le \|d^k\| + \|\hat{d}^k\| \le 2\|d^k\|,$$

we have that $\{\|x^{k+1} - x^k\|\}_{\mathcal{K}'} \to 0$. Hence, by Proposition 4.4 we conclude that the whole sequence $\{x^k\}$ converges to x^*. Furthermore, Lemma 4.2 implies that the whole sequence $\{\lambda^{k,0}\}$ converges to λ^*. ∎

COROLLARY 4.6. *Under Assumptions A1–A5, $I_k = I(x^*)$, $L_k = I^+(x^*)$ for all k large enough, and the sequences generated by Algorithm* 2.1 *satisfy*

(i) *$\delta_k \to \delta_* > 0$ and $\{\|M_k^{-1}\|\}$ is bounded, where*

$$\delta_* = \begin{cases} \vartheta \min\{\lambda_i^*,\ i \in I^+(x^*)\} & \text{if } I^+(x^*) \neq \emptyset, \\ \delta & \text{if } I^+(x^*) = \emptyset; \end{cases}$$

(ii) *$\mu_i^k \to \mu_i^* = \delta_* + \lambda_i^*$ for all $i \in I(x^*)$;*
(iii) *$d^k \to 0$, $d^{k,0} \to 0$, $d^{k,1} \to 0$;*
(iv) *$\lambda^{k,1} \to \lambda^*$, $\lambda^k \to \lambda^*$.*

Proof. Since Lemma 4.5 gives that the whole sequence $\{(x^k, \lambda^{k,0})\}$ converges to (x^*, λ^*), we have from Lemma 4.1 that $I_k = I(x^*)$ and $L_k = I^+(x^*)$ for all k large enough. Hence, result (i) is a direct consequence of the definition of δ_k and Lemma 3.4. Moreover, (2.19) implies result (ii). Lemma 4.3 gives that $d^k \to 0$. Letting $k \to \infty$ in linear equations (2.6) and (2.7) yields that $d^{k,0} \to 0$, $d^{k,1} \to 0$, $\lambda^{k,1} \to \lambda^*$. This further implies that $\rho_k \to 0$, $v^k \to 0$, and $\varpi^k \to 0$. Finally, letting $k \to \infty$ in linear equations (2.9) and (2.13) yields that $\lambda^k \to \lambda^*$. □

The next result shows that under the SSOSC, the exact Lagrangian Hessian is eventually accepted by Algorithm 2.1 without any modification, and Assumption A3 holds correctly.

LEMMA 4.7. *Suppose $\{(x^k, \lambda^{k,0})\} \to (x^*, \lambda^*)$ and Assumptions A2 and A5 hold. Then there exists $\bar{\varrho} > 0$ such that for all k large enough and any $d \in \mathcal{T}(x^k)$,*

$$d^\top \left(\nabla^2_{xx}\mathcal{L}(x^k, \mu^k - \delta_k e) - \sum_{i \in L_k \setminus I(x^k)} \frac{\mu_i^k}{2c_i(x^k)} \nabla c_i(x^k) \nabla c_i(x^k)^\top \right) d \geq \bar{\varrho}\|d\|^2, \tag{4.1}$$

and $H_k = \nabla^2_{xx}\mathcal{L}(x^k, \mu^k - \delta_k e)$ eventually.

Proof. Since the SSOSC holds, Lemma 6.2 implies that for each $\bar{I} \subseteq I^+(x^*)$, there exist $\kappa(\bar{I}) > 0$ and $\varrho(\bar{I}) > 0$ such that

$$d^\top \left(\nabla^2_{xx}\mathcal{L}(x^*, \lambda^*) + \sum_{i \in I^+(x^*) \setminus \bar{I}} \kappa(\bar{I}) \nabla c_i(x^*) \nabla c_i(x^*)^\top \right) d \geq \varrho(\bar{I})\|d\|^2 \tag{4.2}$$

for any $d \in \{y \in \Re^n | \nabla c_i(x^*)^\top y = 0, i \in \bar{I}\}$. Let $\tilde{\kappa} = \max\{\kappa(\bar{I}), \bar{I} \subseteq I^+(x^*)\}$ and $\tilde{\varrho} = \min\{\varrho(\bar{I}), \bar{I} \subseteq I^+(x^*)\}$.

Suppose (4.1) is not true for $\bar{\varrho} = \frac{1}{2}\tilde{\varrho}$. Then there exists an infinite sequence $\{y^k\}_{\mathcal{K}}$, $y^k \in \{y \in \mathcal{T}(x^k) | \|y\| = 1\}$, such that

$$(y^k)^\top \left(\nabla^2_{xx}\mathcal{L}(x^k, \mu^k - \delta_k e) - \sum_{i \in L_k \setminus I(x^k)} \frac{\mu_i^k}{2c_i(x^k)} \nabla c_i(x^k) \nabla c_i(x^k)^\top \right) y^k < \frac{1}{2}\tilde{\varrho}\|y^k\|^2.$$

Since $\{y^k\}_{\mathcal{K}}$ is bounded, we can assume $\{y^k\}_{\mathcal{K}} \to \bar{y}$ with $\|\bar{y}\| = 1$. Since I is finite, we can assume that $I(x^k)$ are identical for all $k \in \mathcal{K}$ and let $\bar{I}_{\mathcal{K}} = I(x^k)$. It is obvious that $\bar{y} \in \{y \in \Re^n | \nabla c_i(x^*)^\top y = 0, i \in I^+(x^*) \cap \bar{I}_{\mathcal{K}}\}$. Since $L_k = I^+(x^*)$ and $-\mu_i^k / c_i(x^k) \to \infty$ for $i \in I^+(x^*) \setminus \bar{I}_{\mathcal{K}}$, we obtain by continuity that for sufficiently large $k \in \mathcal{K}$,

$$\begin{aligned}
&\bar{y}^\top \left(\nabla^2_{xx}\mathcal{L}(x^*, \lambda^*) + \sum_{i \in I^+(x^*) \setminus \bar{I}_{\mathcal{K}}} \tilde{\kappa} \nabla c_i(x^*) \nabla c_i(x^*)^\top \right) \bar{y} \\
&\leq \frac{3}{2}(y^k)^\top \left(\nabla^2_{xx}\mathcal{L}(x^k, \mu^k - \delta_k e) - \sum_{i \in L_k \setminus I(x^k)} \frac{\mu_i^k}{2c_i(x^k)} \nabla c_i(x^k) \nabla c_i(x^k)^\top \right) y^k \\
&< \frac{3}{4}\tilde{\varrho}\|y^k\|^2 \leq \frac{7}{8}\varrho(I^+(x^*) \cap \bar{I}_{\mathcal{K}})\|\bar{y}\|^2.
\end{aligned}$$

This contradicts (4.2). Hence (4.1) holds, which implies that for all k large enough

$$d^\top \left(\nabla^2_{xx}\mathcal{L}(x^k, \mu^k - \delta_k e) - \sum_{i \in I_k \setminus I(x^k)} \frac{\mu_i^k}{c_i(x^k)} \nabla c_i(x^k) \nabla c_i(x^k)^\top \right) d > 0 \tag{4.3}$$

for any $d \in \mathcal{T}(x^k)\backslash\{0\}$. Therefore, we know from Step 1(i) and Step 3 of Algorithm 2.1 that $H_k = \nabla^2_{xx}\mathcal{L}(x^k, \mu^k - \delta_k e)$ eventually. ∎

LEMMA 4.8. *Under Assumptions A1–A5, the sequences generated by Algorithm 2.1 satisfy*

(i) $\|x^k + d^{k,1} - x^*\| = o(\|x^k - x^*\|)$;

(ii) $|\lambda_i^{k,1} - \lambda_i^*| = o(\|d^{k,1}\|)$ *for all* $i \in I(x^*)$.

Proof. It follows from linear equations (2.7) and Corollary 4.6 that for each $i \in I(x^*)$,

$$\begin{aligned} c_i(x^k) &= -\frac{\mu_i^k}{\delta_k + \lambda_i^{k,1}} \nabla c_i(x^k)^\top d^{k,1} \\ &= -\nabla c_i(x^k)^\top d^{k,1} + \left(1 - \frac{\mu_i^k}{\delta_k + \lambda_i^{k,1}}\right) \nabla c_i(x^k)^\top d^{k,1} \\ &= -\nabla c_i(x^k)^\top d^{k,1} + o(\|d^{k,1}\|). \end{aligned} \tag{4.4}$$

Hence, for all k large enough, we can rewrite (2.7) as follows:

$$\begin{bmatrix} H_k & \nabla c_{I(x^*)}(x^k) \\ \nabla c_{I(x^*)}(x^k)^\top & 0 \end{bmatrix} \begin{bmatrix} d^{k,1} \\ \lambda^{k,1}_{I(x^*)} \end{bmatrix} = - \begin{bmatrix} \nabla f(x^k) \\ c_{I(x^*)}(x^k) + o(\|d^{k,1}\|) \end{bmatrix}. \tag{4.5}$$

Thus, we obtain that

$$\begin{bmatrix} H_k & \nabla c_{I(x^*)}(x^k) \\ \nabla c_{I(x^*)}(x^k)^\top & 0 \end{bmatrix} \begin{bmatrix} x^k + d^{k,1} - x^* \\ \lambda^{k,1}_{I(x^*)} - \lambda^*_{I(x^*)} \end{bmatrix} = \mathcal{W}_{\lambda^*}(x^k), \tag{4.6}$$

where

$$\mathcal{W}_{\lambda^*}(x^k) = \begin{bmatrix} -\nabla f(x^k) - \nabla c_{I(x^*)}(x^k)\lambda^*_{I(x^*)} + H_k(x^k - x^*) \\ -c_{I(x^*)}(x^k) + \nabla c_{I(x^*)}(x^k)^\top (x^k - x^*) + o(\|d^{k,1}\|) \end{bmatrix}. \tag{4.7}$$

By the medium value theorem, we have

$$\begin{aligned} \nabla f(x^*) &= \nabla f(x^k) - \nabla^2 f(x^k)(x^k - x^*) \\ &\quad - \int_0^1 [\nabla^2 f(x^k + t(x^* - x^k)) - \nabla^2 f(x^k)](x^k - x^*)dt, \\ \nabla c_i(x^*) &= \nabla c_i(x^k) - \nabla^2 c_i(x^k)(x^k - x^*) \\ &\quad - \int_0^1 [\nabla^2 c_i(x^k + t(x^* - x^k)) - \nabla^2 c_i(x^k)](x^k - x^*)dt,\ i \in I. \end{aligned}$$

Since $\nabla^2 f$ and $\nabla^2 c_i$ $(i \in I)$ are locally Lipschitz continuous in a neighborhood of

x^*, we can assume that L_f and L_i $(i \in I)$ are their Lipschitz constants, respectively. Thus, it follows that for all k large enough,

$$\begin{aligned} &\|\nabla f(x^*) - \nabla f(x^k) - \nabla^2 f(x^k)(x^* - x^k)\| \le \frac{L_f}{2}\|x^k - x^*\|^2, \\ &\|\nabla c_i(x^*) - \nabla c_i(x^k) - \nabla^2 c_i(x^k)(x^* - x^k)\| \le \frac{L_i}{2}\|x^k - x^*\|^2,\ i \in I. \end{aligned} \tag{4.8}$$

Moreover, we have from Corollary 4.6 and Lemma 4.7 that

$$\|[H_k - \nabla^2_{xx}\mathcal{L}(x^*, \lambda^*)](x^k - x^*)\| = o(\|x^k - x^*\|). \tag{4.9}$$

By (4.8) and (4.9), we further obtain that

$$\begin{aligned} &- \nabla f(x^k) - \nabla c_{I(x^*)}(x^k)\lambda^*_{I(x^*)} + H_k(x^k - x^*) \\ &= [\nabla f(x^*) - \nabla f(x^k) - \nabla^2 f(x^k)(x^* - x^k)] \\ &\quad + \sum_{i \in I(x^*)} \lambda_i^*[\nabla c_i(x^*) - \nabla c_i(x^k) - \nabla^2 c_i(x^k)(x^* - x^k)] \\ &\quad + [H_k - \nabla^2_{xx}\mathcal{L}(x^*, \lambda^*)](x^k - x^*) \\ &\quad + \left[\nabla^2 f(x^*) - \nabla^2 f(x^k) + \sum_{i \in I(x^*)} \lambda_i^*(\nabla^2 c_i(x^*) - \nabla^2 c_i(x^k))\right](x^k - x^*) \\ &= o(\|x^k - x^*\|), \end{aligned}$$

$$\begin{aligned} &- c_{I(x^*)}(x^k) + \nabla c_{I(x^*)}(x^k)^\top (x^k - x^*) \\ &= c_{I(x^*)}(x^*) - c_{I(x^*)}(x^k) + \nabla c_{I(x^*)}(x^k)^\top (x^k - x^*) \\ &= O(\|x^k - x^*\|^2). \end{aligned}$$

This together with (4.7) implies that $\mathcal{W}_{\lambda^*}(x^k) = o(\|x^k - x^*\|) + o(\|d^{k,1}\|)$. By following an identical proof to that of [8, Proposition 3.1], we know that under Assumptions A2 and A5, the coefficient matrices of (4.6) are uniformly nonsingular for all k large enough. Hence, we know from (4.6) that $\|x^k + d^{k,1} - x^*\| = o(\|x^k - x^*\|) + o(\|d^{k,1}\|)$, which implies that $\|x^k - x^*\| = O(\|d^{k,1}\|)$. Thus, results (i) and (ii) again follow from (4.6). ∎

LEMMA 4.9. *Under Assumptions A1–A5, we eventually have* FAST = TRUE *in Step* 1(ii) *of Algorithm* 2.1*, i.e.,* (2.8) *holds for all k large enough.*

Proof. Since $\lambda^{k,1} \to \lambda^*$, it follows from Lemma 4.8 that for all sufficiently large k,

$$\begin{aligned} &\lambda_i^{k,1} > 0 && \forall i \in I^+(x^*); \\ &\lambda_i^{k,1} = \lambda_i^{k,1} - \lambda_i^* = o(\|d^{k,1}\|) && \forall i \in I(x^*) \backslash I^+(x^*). \end{aligned} \tag{4.10}$$

From the fact that $I(x^k) \subseteq I_k$ and (2.7), we have

$$\mu_i^k \nabla c_i(x^k)^\top d^{k,1} = (-\delta_k - \lambda_i^{k,1}) c_i(x^k) = 0 \quad \forall i \in I(x^k),$$

which implies that $d^{k,1} \in \mathcal{T}(x^k)$ since $\mu^k > 0$. Moreover, by Corollary 4.6 we have $\delta_k < \lambda_i^{k,1}$ for all $i \in I^+(x^*)$ and large enough k. Hence, we can obtain from (2.7),

Corollary 4.6, Lemma 4.7, and (4.10) that for all k large enough,

$$
\begin{aligned}
&\nabla f(x^k)^\top d^{k,1}\\
&= -(d^{k,1})^\top H_k d^{k,1} - \sum_{i\in I(x^*)} \lambda_i^{k,1}\nabla c_i(x^k)^\top d^{k,1}\\
&= -(d^{k,1})^\top H_k d^{k,1} + \sum_{i\in I^+(x^*)\setminus I(x^k)} \frac{\mu_i^k}{2c_i(x^k)}(d^{k,1})^\top \nabla c_i(x^k)\nabla c_i(x^k)^\top d^{k,1}\\
&\quad + \sum_{i\in I^+(x^*)\setminus I(x^k)} \left(\frac{\mu_i^k}{2c_i(x^k)}\nabla c_i(x^k)^\top d^{k,1} + \delta_k\right)\nabla c_i(x^k)^\top d^{k,1}\\
&\quad - \sum_{i\in I(x^*)\setminus I^+(x^*)} \lambda_i^{k,1}\nabla c_i(x^k)^\top d^{k,1}\\
&= -(d^{k,1})^\top \left(H_k - \sum_{i\in I^+(x^*)\setminus I(x^k)} \frac{\mu_i^k}{2c_i(x^k)}\nabla c_i(x^k)\nabla c_i(x^k)^\top\right) d^{k,1}\\
&\quad + \sum_{i\in I^+(x^*)\setminus I(x^k)} \frac{c_i(x^k)}{2\mu_i^k}((\lambda_i^{k,1})^2 - \delta_k^2) + o(\|d^{k,1}\|^2)\\
&\le -\bar{\varrho}\|d^{k,1}\|^2 + o(\|d^{k,1}\|^2) \le -\|d^{k,1}\|^\gamma
\end{aligned}
\tag{4.11}
$$

as $\gamma > 2$, where $\bar{\varrho} > 0$ is defined in Lemma 4.7. Formulas (4.10) and (4.11) imply that (2.8) eventually holds. □

LEMMA 4.10. *Under Assumptions A1–A5, the sequences generated by Algorithm 2.1 eventually satisfy*

(i) $\|d^k - d^{k,1}\| = o(\|d^{k,1}\|^2)$, $\|\lambda^k - \lambda^{k,1}\| = o(\|d^{k,1}\|^2)$;

(ii) $(\hat{d}^k, \hat{\lambda}^k) = (d^{k,3}, \lambda^{k,3})$ *and*

$$
\|\hat{d}^k - d^k\| = O\left(\max\left\{\|d^k\|^2, \max_{i\in I(x^*)}\left\{\left|1 - \frac{\mu_i^k}{\delta_k + \lambda_i^k}\right| \|d^k\|\right\}\right\}\right) = o(\|d^k\|),
$$

$$
\|\hat{\lambda}^k - \lambda^k\| = O\left(\max\left\{\|d^k\|^2, \max_{i\in I(x^*)}\left\{\left|1 - \frac{\mu_i^k}{\delta_k + \lambda_i^k}\right| \|d^k\|\right\}\right\}\right) = o(\|d^k\|).
$$

Proof. From (2.11) we have $\rho_k = o(\|d^{k,1}\|^2)$. Subtracting both sides of (2.7) by the corresponding sides of (2.9) yields that

$$
\begin{aligned}
&\|d^k - d^{k,1}\| = O(\rho_k) = o(\|d^{k,1}\|^2);\\
&|\lambda_i^k - \lambda_i^{k,1}| = O(\rho_k) = o(\|d^{k,1}\|^2) \quad \forall i \in I(x^*).
\end{aligned}
\tag{4.12}
$$

Since $\lambda_i^{k,1} = \lambda_i^k = 0$ for $i \in I\setminus I(x^*)$, result (i) follows from (4.12). Now subtracting both sides of (2.13) by the corresponding sides of (2.9) yields that

$$
M_k \begin{bmatrix} d^{k,3} - d^k \\ \lambda^{k,3}_{I(x^*)} - \lambda^k_{I(x^*)} \end{bmatrix} = \begin{bmatrix} 0 \\ \varpi^k - v^k \end{bmatrix}. \tag{4.13}
$$

From (2.15) we have $\pi_k = o(\|d^k\|^2)$. Hence, it follows from (2.14), (4.4), and (4.12)

that

$$
\begin{aligned}
&\max_{i\in I(x^*)}\{|\varpi_i^k - v_i^k|\}\\
&= \max_{i\in I(x^*)}\{|-\mu_i^k c_i(x^k+d^k)-\pi_k|\}\\
&= \max_{i\in I(x^*)}\{|-\mu_i^k(c_i(x^k)+\nabla c_i(x^k)^\top d^{k,1})\\
&\quad -\mu_i^k\nabla c_i(x^k)^\top(d^k-d^{k,1})-\pi_k+O(\|d^k\|^2)|\}\\
&= \max_{i\in I(x^*)}\left\{\left|-\mu_i^k\left(1-\frac{\mu_i^k}{\delta_k+\lambda_i^{k,1}}\right)\nabla c_i(x^k)^\top d^{k,1}+O(\|d^k\|^2)\right|\right\}\\
&= \max_{i\in I(x^*)}\left\{O\left(\left|1-\frac{\mu_i^k}{\delta_k+\lambda_i^{k,1}}\right|\|d^{k,1}\|\right)+O(\|d^k\|^2)\right\}\\
&= O\left(\max\left\{\|d^k\|^2,\max_{i\in I(x^*)}\left\{\left|1-\frac{\mu_i^k}{\delta_k+\lambda_i^k}\right|\|d^k\|\right\}\right\}\right),
\end{aligned}
\tag{4.14}
$$

where $\lambda_i^{k,1}$ and $d^{k,1}$ are directly replaced by λ_i^k and d^k due to (4.12). Since $\{\|M_k^{-1}\|\}$ is bounded by Corollary 4.6 and $\lambda_i^{k,3}=\lambda_i^k=0$ for $i\in I\backslash I(x^*)$, it follows that for all sufficiently large k, $\|d^{k,3}-d^k\|<\|d^k\|$, which indicates $(\hat{d}^k,\hat{\lambda}^k)=(d^{k,3},\lambda^{k,3})$. Thus, (ii) follows immediately from (4.13) and (4.14). □

LEMMA 4.11. *Under Assumptions A1–A5, the arc search in Step 2 of Algorithm 2.1 eventually accepts a full step of one, i.e., $t_k=1$ for all k large enough.*

Proof. First, we show that for all k large enough,

$$
c_i(x^k+\hat{d}^k)\le 0 \quad \forall i\in I. \tag{4.15}
$$

Lemma 4.10 implies that $\hat{d}^k\to 0$. Hence for all k large enough, $c_i(x^k+\hat{d}^k)<0$ for all $i\in I\backslash I(x^*)$. For $i\in I(x^*)$, by using Taylor expansion we can obtain that for all sufficiently large k,

$$
\begin{aligned}
&\mu_i^k c_i(x^k+\hat{d}^k)\\
&=\mu_i^k c_i(x^k+d^k+\hat{d}^k-d^k)\\
&=\mu_i^k\{c_i(x^k+d^k)+\nabla c_i(x^k+d^k)^\top(\hat{d}^k-d^k)+O(\|\hat{d}^k-d^k\|^2)\}\\
&=\mu_i^k\{c_i(x^k+d^k)+\nabla c_i(x^k)^\top(\hat{d}^k-d^k)+O(\|d^k\|\cdot\|\hat{d}^k-d^k\|)\}\\
&=-(\hat{\lambda}_i^k-\lambda_i^k)c_i(x^k)-\pi_k+O(\|d^k\|\cdot\|\hat{d}^k-d^k\|)\\
&=\frac{\mu_i^k}{\delta_k+\lambda_i^{k,1}}(\hat{\lambda}_i^k-\lambda_i^k)\nabla c_i(x^k)^\top d^{k,1}-\pi_k+O(\|d^k\|\cdot\|\hat{d}^k-d^k\|)\\
&=-\pi_k+O(\|d^k\|\cdot|\hat{\lambda}_i^k-\lambda_i^k|)+O(\|d^k\|\cdot\|\hat{d}^k-d^k\|)\\
&=-\pi_k+O\left(\max\left\{\|d^k\|^3,\max_{i\in I(x^*)}\left\{\left|1-\frac{\mu_i^k}{\delta_k+\lambda_i^k}\right|\|d^k\|^2\right\}\right\}\right)\\
&<-\frac{1}{2}\pi_k<0,
\end{aligned}
\tag{4.16}
$$

where the fourth equality follows from (2.13) and (2.14), the fifth equality follows from (4.4), the last equality is given by Lemma 4.10(ii), and the last line of inequalities

follows from (2.15) since $\nu \in (2,3)$ and $\tau \in (0,1)$. Consequently, it follows from (2.15) and (4.16) that for each $i \in I(x^*)$

$$|c_i(x^k + \hat{d}^k)| = O(\pi_k) = o(\|d^k\|^2). \tag{4.17}$$

Now we show that $t_k = 1$ provides a sufficient reduction in f, i.e.,

$$f(x^k + \hat{d}^k) \le f(x^k) + \sigma \nabla f(x^k)^\top d^k. \tag{4.18}$$

By using Taylor expansion, we get from Lemma 4.10(ii) that for each $i \in I(x^*)$

$$\begin{aligned}
&c_i(x^k + \hat{d}^k)\\
&= c_i(x^k) + \nabla c_i(x^k)^\top \hat{d}^k + \frac{1}{2}(\hat{d}^k)^\top \nabla^2 c_i(x^k)\hat{d}^k + o(\|\hat{d}^k\|^2)\\
&= c_i(x^k) + \nabla c_i(x^k)^\top d^k + \nabla c_i(x^k)^\top (\hat{d}^k - d^k) + \frac{1}{2}(d^k)^\top \nabla^2 c_i(x^k) d^k + o(\|d^k\|^2).
\end{aligned}$$

Thus, by (4.4), (4.17), and Lemma 4.10 we can obtain that

$$\begin{aligned}
&-\frac{1}{2}\lambda_i^k \nabla c_i(x^k)^\top d^k - \lambda_i^k \nabla c_i(x^k)^\top (\hat{d}^k - d^k)\\
&= \lambda_i^k c_i(x^k) + \frac{1}{2}\lambda_i^k \nabla c_i(x^k)^\top d^k + \frac{1}{2}\lambda_i^k (d^k)^\top \nabla^2 c_i(x^k) d^k + o(\|d^k\|^2)\\
&= \lambda_i^k c_i(x^k) + \frac{1}{2}\lambda_i^k \nabla c_i(x^k)^\top d^{k,1} + \frac{1}{2}\lambda_i^k \nabla c_i(x^k)^\top (d^k - d^{k,1})\\
&\quad + \frac{1}{2}\lambda_i^k (d^k)^\top \nabla^2 c_i(x^k) d^k + o(\|d^k\|^2)\\
&= \lambda_i^k c_i(x^k) - \lambda_i^k \left(\frac{\delta_k + \lambda_i^{k,1}}{2\mu_i^k}\right) c_i(x^k) + \frac{1}{2}\lambda_i^k (d^k)^\top \nabla^2 c_i(x^k) d^k + o(\|d^k\|^2)\\
&= \left(1 - \frac{\delta_k + \lambda_i^{k,1}}{2\mu_i^k}\right) \lambda_i^k c_i(x^k) + \frac{1}{2}\lambda_i^k (d^k)^\top \nabla^2 c_i(x^k) d^k + o(\|d^k\|^2).
\end{aligned} \tag{4.19}$$

Since $\frac{\delta_k + \lambda_i^{k,1}}{\mu_i^k} \to 1$ by Corollary 4.6, it follows that for all sufficiently large k, $1 - \frac{\delta_k + \lambda_i^{k,1}}{2\mu_i^k} > 0$. Hence, we get from (4.19) that for $i \in I^+(x^*)$ and large enough k,

$$\begin{aligned}
&-\frac{1}{2}\lambda_i^k \nabla c_i(x^k)^\top d^k - \lambda_i^k \nabla c_i(x^k)^\top (\hat{d}^k - d^k)\\
&\le \frac{1}{2}\lambda_i^k (d^k)^\top \nabla^2 c_i(x^k) d^k + o(\|d^k\|^2).
\end{aligned} \tag{4.20}$$

On the other hand, it follows from (4.4), (4.19), and Lemmas 4.8(ii) and 4.10(i) that for $i \in I(x^*) \backslash I^+(x^*)$,

$$\begin{aligned}
&-\frac{1}{2}\lambda_i^k \nabla c_i(x^k)^\top d^k - \lambda_i^k \nabla c_i(x^k)^\top (\hat{d}^k - d^k)\\
&= (\lambda_i^k - \lambda_i^*)\left(1 - \frac{\delta_k + \lambda_i^{k,1}}{2\mu_i^k}\right)(-\nabla c_i(x^k)^\top d^{k,1} + o(\|d^{k,1}\|))\\
&\quad + \frac{1}{2}\lambda_i^k (d^k)^\top \nabla^2 c_i(x^k) d^k + o(\|d^k\|^2)\\
&= \frac{1}{2}\lambda_i^k (d^k)^\top \nabla^2 c_i(x^k) d^k + o(\|d^k\|^2).
\end{aligned} \tag{4.21}$$

Now we are prepared to derive the following key relation, which also uses Taylor expansion:

$$
\begin{aligned}
& f(x^k+\hat{d}^k) \\
&= f(x^k)+\nabla f(x^k)^\top \hat{d}^k+\frac{1}{2}(\hat{d}^k)^\top \nabla^2 f(x^k)\hat{d}^k+o(\|\hat{d}^k\|^2) \\
&= f(x^k)+\nabla f(x^k)^\top d^k+\nabla f(x^k)^\top(\hat{d}^k-d^k) \\
&\quad +\frac{1}{2}(d^k)^\top \nabla^2 f(x^k)d^k+o(\|d^k\|^2) \\
&= f(x^k)+\frac{1}{2}\nabla f(x^k)^\top d^k-\frac{1}{2}(d^k)^\top H_k d^k-\frac{1}{2}\sum_{i\in I(x^*)}\lambda_i^k \nabla c_i(x^k)^\top d^k \\
&\quad -(\hat{d}^k-d^k)^\top H_k d^k-\sum_{i\in I(x^*)}\lambda_i^k \nabla c_i(x^k)^\top(\hat{d}^k-d^k) \\
&\quad +\frac{1}{2}(d^k)^\top \nabla^2 f(x^k)d^k+o(\|d^k\|^2) \\
&= f(x^k)+\frac{1}{2}\nabla f(x^k)^\top d^k+\frac{1}{2}(d^k)^\top(\nabla^2 f(x^k)-H_k)d^k \\
&\quad -\sum_{i\in I(x^*)}\left(\frac{1}{2}\lambda_i^k \nabla c_i(x^k)^\top d^k+\lambda_i^k \nabla c_i(x^k)^\top(\hat{d}^k-d^k)\right)+o(\|d^k\|^2) \\
&\le f(x^k)+\frac{1}{2}\nabla f(x^k)^\top d^k+\frac{1}{2}(d^k)^\top(\nabla^2 f(x^k)-H_k)d^k \\
&\quad +\frac{1}{2}\sum_{i\in I(x^*)}\lambda_i^k (d^k)^\top \nabla^2 c_i(x^k)d^k+o(\|d^k\|^2) \\
&= f(x^k)+\frac{1}{2}\nabla f(x^k)^\top d^k+\frac{1}{2}(d^k)^\top(\nabla^2_{xx}\mathcal{L}(x^k,\lambda^k)-H_k)d^k+o(\|d^k\|^2) \\
&= f(x^k)+\sigma\nabla f(x^k)^\top d^k+\left(\frac{1}{2}-\sigma\right)\nabla f(x^k)^\top d^k+o(\|d^k\|^2) \\
&\le f(x^k)+\sigma\nabla f(x^k)^\top d^k-\bar{\varrho}\theta\left(\frac{1}{2}-\sigma\right)\|d^k\|^2+o(\|d^k\|^2) \\
&\le f(x^k)+\sigma\nabla f(x^k)^\top d^k,
\end{aligned}
$$

where the second equality is implied by Lemma 4.10, the third equality follows from (2.9), the first inequality is obtained by combining (4.20) and (4.21), the last equality follows from Lemma 4.7, and the second inequality follows from Lemma 2.3(i), (4.11) with $\bar{\varrho}$ defined in Lemma 4.7, Lemma 4.10(i), and the fact that $\sigma<\frac{1}{2}$.

Consequently, we conclude that both (4.15) and (4.18) hold for all sufficiently large k, and thus the unit step size is eventually accepted. □

Putting the results of Lemmas 4.8, 4.9, 4.10, and 4.11 together, we directly obtain the Q-superlinear convergence of Algorithm 2.1.

THEOREM 4.12. *Under Assumptions A1–A5, the sequence* $\{x^k\}$ *generated by Algorithm* 2.1 *converges Q-superlinearly, i.e.,*

$$\|x^{k+1}-x^*\|=o(\|x^k-x^*\|).$$

5. Concluding remarks. In this paper we have presented a new feasible active set QP-free method for inequality constrained optimization. Under very general assumptions (the existence of a feasible initial point, smoothness of the objective

and constraint functions, LICQ and SSOSC, etc.), we have proved that our QP-free method is globally convergent and locally superlinearly convergent to a KKT point of the problem. Noticeably, the superlinear convergence is achieved without assuming that strict complementarity holds or that all optimal multipliers are less than a preselected parameter. Moreover, our method avoids several computational issues in existing active set QP-free methods. For example, our method does not have to compute a new multiplier estimate and select linearly independent constraint gradients to determine a working set at each iteration. A new technique based on so-called δ_*-drifted multipliers is introduced in the method to avoid the possible ill-conditioning of the Newton system caused by dual degeneracy.

Although the main interest of this paper is theoretical, we are also interested in testing the practical efficiency of our method. By implementing it on some small problems from [14], we found that it worked well even for problems failing strict complementarity. However, we feel that it is premature to report numerical results at this stage because of two facts that largely limit the test set for which our method is applicable. First, the current method only tackles problems with inequality constraints. Second, a feasible starting point is required. To deal with equality constraints, we plan to incorporate the ℓ_2-penalty technique introduced in [3] in our QP-free method. Our strategy for the second issue is to transform some inequality constraints to equality constraints by adding slack variables and again, make use of the ℓ_2-penalty technique. We will report on the effectiveness of our approach in due course.

6. Appendix.

LEMMA 6.1. *Suppose Assumption A2 holds,* $H \in \Re^{n\times n}$, $x \in \mathcal{F}$, $I(x) \subseteq \tilde{I} \subseteq I$, *and* $\mu_i > 0$ *for all* $i \in \tilde{I}$. *If*

$$y^\top \left(H - \sum_{i\in \tilde{I}\setminus I(x)} \frac{\mu_i}{c_i(x)} \nabla c_i(x) \nabla c_i(x)^\top \right) y > 0 \tag{6.1}$$

for all $y \in \mathcal{T}(x)\setminus\{0\}$, *then the following matrix is nonsingular:*

$$M = \begin{bmatrix} H & \nabla c_{\tilde{I}}(x) \\ U\nabla c_{\tilde{I}}(x)^\top & C_{\tilde{I}}(x) \end{bmatrix},$$

where $U = \operatorname{diag}(\mu_{\tilde{I}})$ *and* $C_{\tilde{I}}(x) = \operatorname{diag}(c_{\tilde{I}}(x))$.

Proof. Suppose (d, λ) is a solution of the following linear equations:

$$M \begin{bmatrix} d \\ \lambda \end{bmatrix} = 0. \tag{6.2}$$

It follows that

$$\lambda_i = -\frac{\mu_i}{c_i(x)} \nabla c_i(x)^\top d,\ i \in \tilde{I}\setminus I(x), \quad \text{and} \quad \nabla c_i(x)^\top d = 0,\ i \in I(x).$$

Substituting this into (6.2) yields that

$$d^\top \left(H - \sum_{\tilde{I}\setminus I(x)} \frac{\mu_i}{c_i(x)} \nabla c_i(x) \nabla c_i(x)^\top \right) d = 0,$$

which implies that $d = 0$. Moreover, it follows from (6.2) that $\nabla c_{\tilde{I}}(x)\lambda = 0$ and $C_{\tilde{I}}(x)\lambda = 0$. Hence, $\nabla c_{I(x)}\lambda_{I(x)} = 0$ and $\lambda_{\tilde{I}\setminus I(x)} = 0$. Since Assumption A2 implies $\lambda_{I(x)} = 0$, zero is the unique solution of (6.2), i.e., M is nonsingular. ∎

LEMMA 6.2. *Suppose $H \in \Re^{n\times n}$, $J \in \Re^{n\times q}$, and $J = [J_1|J_2]$. If H is positive definite on $S(J)$, where $S(J) = \{y \in \Re^n | J^\top y = 0\}$, then there exists $\bar{r} > 0$ such that for any $r \geq \bar{r}$, $H + rJ_1J_1^\top$ is positive definite on $S(J_2)$, where $S(J_2) = \{y \in \Re^n | J_2^\top y = 0\}$ and in case that $J = J_1$, $S(J_2) = \Re^n$.*

Proof. It suffices to show that there exists $\bar{r} > 0$ such that for any $r \geq \bar{r}$,

$$d^\top (H + rJ_1J_1^\top)d > 0$$

for all $d \in B = \{y \in S(J_2) | \|y\| = 1\}$. Let $N = B \cap \{y \in \Re^n | y^\top Hy \leq 0\}$. Obviously, B and N are both closed and compact. Hence, there exist r_1 and r_2 such that

$$r_1 = \min_{d\in B} d^\top Hd \quad \text{and} \quad r_2 = \min_{d\in N} d^\top J_1J_1^\top d.$$

Since $J_1^\top d \neq 0$ for any $d \in N$, we have $r_2 > 0$. Set $\bar{r} = \max\{-\frac{r_1}{r_2} + 1, 1\}$ and consider any $r \geq \bar{r}$. For any $d \in B$, if $d \in N$, then

$$d^\top (H + rJ_1J_1^\top)d \geq r_1 + rr_2 \geq r_2 > 0;$$

otherwise, $d \in B\backslash N$ and

$$d^\top (H + rJ_1J_1^\top)d \geq d^\top Hd > 0.$$

Hence, we conclude that $H + rJ_1J_1^\top$ is positive definite on B, and hence on $S(J_2)$. ∎

Acknowledgments. We thank Profs. André L. Tits, Andreas Fischer, and Christian Kanzow for their valuable suggestions and comments on a early version of this paper. We are grateful to Prof. Donald Goldfarb for carefully reading the manuscript and improving its clarity and to Drs. Wanmo Kang and Wotao Yin for many helpful discussions. We also wish to thank two anonymous referees and the associate editor Prof. Shuzhong Zhang for their insightful comments, which helped to improve the paper considerably.

REFERENCES

[1] S. BAKHTIARI AND A. L. TITS, *A simple primal-dual feasible interior-point method for nonlinear programming with monotone descent*, Comput. Optim. Appl., 25 (2003), pp. 17–38.

[2] J. F. BONNANS, *Local analysis of Newton type methods for variational inequalities and nonlinear programming*, Appl. Math. Optim., 29 (1994), pp. 161–186.

[3] L. CHEN AND D. GOLDFARB, *Interior-point ℓ_2-penalty methods for nonlinear programming with strong global convergence properties*, Math. Program., to appear.

[4] F. FACCHINEI, A. FISCHER, AND C. KANZOW, *On the accurate identification of active constraints*, SIAM J. Optim., 9 (1998), pp. 14–32.

[5] 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.

[6] F. FACCHINEI AND C. LAZZARI, *Local feasible QP-free algorithms for the constrained minimization of SC^1 functions*, J. Optim. Theory Appl., 119 (2003), pp. 281–316.

[7] F. FACCHINEI, G. LIUZZI, AND S. LUCIDI, *A truncated Newton method for the solution of large-scale inequality constrained minimization problems*, Comput. Optim. Appl., 25 (2003), pp. 85–122.

[8] 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.

[9] 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.

[10] A. Fischer, *Modified Wilson method for nonlinear programs with nonunique multipliers*, Math. Oper. Res., 24 (1999), pp. 699–727.
[11] A. Fischer, *Local behavior of an iterative framework for generalized equations with nonisolated solutions*, Math. Programming, 94 (2002), pp. 91–124.
[12] A. Forsgren, *Inertia-controlling factorizations for optimization algorithms*, Appl. Numer. Math., 43 (2002), pp. 91–107.
[13] Z. Gao, G. He, and F. Wu, *Sequential systems of linear equations algorithm for nonlinear optimization problems with general constraints*, J. Optim. Theory Appl., 95 (1997), pp. 371–397.
[14] W. Hock and K. Schittkowski, *Test Examples for Nonlinear Programming Codes*, Lecture Notes in Econom. and Math. Systems 187, Springer-Verlag, Berlin, 1981.
[15] A. F. Izmailov and M. V. Solodov, *Newton-type Methods for Optimization Problems without Constraint Qualifications*, Technical report, IMPA, Brazil, 2003.
[16] C. Kanzow and H. Qi, *A QP-free constrained Newton-type method for variational inequality problems*, Math. Programming, 85 (1999), pp. 81–106.
[17] J. J. Moré and D. C. Sorensen, *Computing a trust region step*, SIAM J. Sci. Statist. Comput., 4 (1983), pp. 553–572.
[18] 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.
[19] 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.
[20] L. Qi and Y. Yang, *Globally and superlinearly convergent QP-free algorithm for nonlinear constrained optimization*, J. Optim. Theory Appl., 113 (2002), pp. 297–323.
[21] S. M. Robinson, *Strongly regular generalized equations*, Math. Oper. Res., 5 (1980), pp. 43–62.
[22] A. L. Tits, A. Wächter, S. Bakhtiari, T. J. Urban, and C. T. Lawrence, *A primal-dual interior-point method for nonlinear programming with strong global and local convergence properties*, SIAM J. Optim., 14 (2003), pp. 173–199.
[23] R. J. Vanderbei and D. F. Shanno, *An interior-point algorithm for nonconvex nonlinear programming*, Comput. Optim. Appl., 13 (1999), pp. 231–252.
[24] A. Wächter and L. T. Biegler, *On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming*, Math. Program., 106 (2006), pp. 25–57.
[25] S. J. Wright, *Superlinear convergence of a stabilized SQP method to a degenerate solution*, Comput. Optim. Appl., 11 (1998), pp. 253–275.
[26] S. J. Wright, *Modifying SQP for degenerate problems*, SIAM J. Optim., 13 (2002), pp. 470–497.
[27] S. J. Wright, *An Algorithm for Degenerate Nonlinear Programming with Rapid Local Convergence*, Optimal TR 03-02, Computer Sciences Department, University of Wisconsin-Madison, Madison, WI, 2003.
[28] Y.-F. Yang, D.-H. Li, and L. Qi, *A feasible sequential linear equation method for inequality constrained optimization*, SIAM J. Optim., 13 (2003), pp. 1222–1244.

SIAM J. OPTIM.
Vol. 17, No. 2, pp. 430–458
© 2006 Society for Industrial and Applied Mathematics

RAPID SOURCE INVERSION FOR CHEMICAL/BIOLOGICAL ATTACKS, PART 1: THE STEADY-STATE CASE*

PAUL T. BOGGS†, KEVIN R. LONG†, STEPHEN B. MARGOLIS†, AND PATRICIA A. HOWARD‡

Abstract. A critical first step in responding to an airborne chemical or biological attack is determining the location of the source of the toxin. We have formulated the mathematical description of source location as an inverse problem constrained by the partial differential equation (PDE) that describes the toxin's transport. This transport is advection-dominated, but takes place in a flow field that in realistic settings will be turbulent, thereby inducing an effective diffusivity tensor in the transport model. We model the turbulent flow using a Reynolds-averaged Navier–Stokes (RANS) approach, which can be solved offline for an arbitrary building of interest. The inversion problem then consists of finding the (regularized) source distribution that best reproduces a set of sensor measurements, subject to the transport model constraint relating the source to the concentration at the sensor positions. Though individual toxin sources are likely to be point sources, we cannot make any assumptions about the number of such sources. Hence, because multiple sources are a possibility, we assume a spatially continuous source distribution, thus eliminiating any need to impose assumptions about the number and nature of the sources. The operational context for this problem implies certain practical requirements. In particular, it is critical to reduce the time for inversion and the number of sensors required for an accurate determination of the source field. A particular focus of this paper is the exploration of the degree to which we can economize on computational effort through adaptive mesh coarsening tailored to preserve the essential features of the flow field. We have found that location of multiple sources is well accommodated by this method, and have shown that it is possible to reduce significantly the computational time through flow-tailored mesh adaptation without adverse impact on the accuracy of the source location. Finally, we have done a preliminary study of the number of sensors required for useful inversion. These conclusions will be of considerable use in developing sensor deployment strategies.

Key words. source inversion, PDE-constrained optimization, software environments, adaptive meshes, Sundance, O3D, Split, Trilinos, TSF

AMS subject classifications. 49N45, 49K20, 49M05, 65K10, 76F10

DOI. 10.1137/040603036

1. Introduction.

The key to responding to a chemical or biological attack in a building, e.g., an airport, is speed. In an attack using an airborne toxin, emergency officials need to know as soon as possible what the toxin is and where its source is located. This will allow them to take appropriate action to minimize the impact of the attack, including moving people to safety, confining or venting the release, and possibly even determining who is responsible for the attack.

Two things are needed to determine the type and source of a biological attack: First, one needs sensors that can reliably determine the nature of the toxin and its local concentration; and second, one needs to have a mathematical model of the

*Received by the editors January 9, 2004; accepted for publication (in revised form) January 29, 2006; published electronically June 9, 2006. This work was supported by the U.S. Department of Energy under contract DE-AC04-94AL85000. The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. Copyright is owned by SIAM to the extent not limited by these rights.

http://www.siam.org/journals/siopt/17-2/60303.html

†Computational Science and Mathematics Research Department, Sandia National Laboratories, Livermore, CA 94551 (ptboggs@sandia.gov, krlong@sandia.gov, margoli@sandia.gov).

‡Computational and Applied Mathematics Department, Rice University, Houston, TX 77251 (pahoward@caam.rice.edu).

toxin dispersion in the air flow in the building that can be inverted to provide the location of the source. Here we assume that appropriate sensors have been placed in the building. Such sensors are being actively developed in several laboratories, including Sandia National Laboratories (see, e.g., [13]); in section 6 we comment on some related work designed to improve these sensors. The location of the sensors within the building is clearly an important issue in reconstructing the source field. We do not consider the optimal location of the sensors in this paper, but, again, in section 6 we discuss the issues involved. In this paper we concentrate on providing an optimization model that simultaneously predicts both the number of sources in the attack and the location of each, given the concentration data from the sensors and the air flow field. We refer to this problem as the *source-inversion problem.*

Source inversion in this context is an example of optimization problems that are constrained by partial differential equations (PDE). Due to the dramatic increase in computational power afforded by massively parallel computers, the substantial gains in our optimization technology, and the major improvements in our understanding and ability to precondition the linear systems that arise in these problems, PDE-constrained optimization has emerged as an important research area. (See [26] for an introduction and the papers in [2] for more extensive coverage.) Source inversion has been considered by [1] and excellent work in preconditioning is given in [4].

The source-inversion problem that we have briefly described above has important features that have motivated our work. In particular, speed of solution is much more important than accuracy. Indeed, if we can correctly predict the number of sources and their locations to within a few meters, emergency personnel can easily find the release devices. As stated above, we need to know the flow field in the building to determine the locations quickly. The air flow in a modern building is determined by the heating, ventilating, and air-conditioning (HVAC) system. It is our understanding that the flows in such a building do not change much over the course of a day. Thus we can assume that a small number of flow fields can be computed in advance and the appropriate one can be selected based on the time of the attack. We can also assume that there will be moderate computing capability in the building. Indeed, given the expense of security in general, the purchase of a small cluster of processors seems quite reasonable.

Since one of our main motivations is speed, we also investigate some strategies for decreasing the time for the calculations. In particular, we will require a relatively fine mesh to solve for the flow field within a given building. The question arises, however, of whether the estimation of the source, or sources, can be done on a coarser mesh. We show that, in fact, a much coarser mesh will suffice and that using such a coarse mesh reduces the computing time by factors of 40–100. This rather surprising result indicates that a practical system using these strategies may be possible. Thus, novel contributions in this paper include the model that allows for the simultaneous determination of both the number and the locations of the source(s) and the fact that very coarse meshes can be used to accelerate the inversion algorithms significantly.

Finally, the problem is clearly time-dependent, but we think that much can be learned initially from considering the simpler steady-state case. We demonstrate in this paper that consideration of a two-dimensional, steady-state model can provide much insight that will guide our efforts in the general three-dimensional, time-dependent case. We do not claim that we have answered all of the questions for this case, but we can do the additional development and experiments in the context of the more general case that will be considered in part 2 of this work. In addition, the

steady-state case is of interest in its own right; one can use the techniques developed here to find, for example, a persistent leak in a complex facility.

In conducting such a project, it was crucial to have a powerful software environment in which to develop and test ideas rapidly. This work would have been far more difficult without the tools that have been developed at Sandia National Laboratories by many people. Most important for this work is Sundance, a code that takes a symbolic description of PDEs and then efficiently creates finite element (mass) matrices and associated right-hand side vectors. The details of how we formulate the flow-field problems and the optimization problems are dictated by our use of Sundance. An optimization code, Split/O3D, has been built to work directly with Sundance operators.

The paper is organized as follows. Given the importance of the software environment, and its impact on some of our analyses and decisions, we describe it first in section 2. We next describe (in section 3) the flow problem that we formulate and solve. In creating the flow field, we tried to specify realistic models that take into account the fact that the air flow in a building such as an airport is subject to many perturbations, including the opening and closing of doors and the movement of people, luggage, and equipment. We then provide (in section 4) the appropriate source dispersion model and the resulting optimization problems that we consider for locating the sources. In section 5, we give a simple two-dimensional example of a building and the resulting flow field. We then give the results of our numerical tests based on the optimization approaches in section 4. We also develop our coarse mesh approximations and show the results using these. In section 6 we provide some discussion about the work that needs to be done to extend this approach to the time-dependent case and to the question of optimal sensor location.

2. Software environment. In this section we briefly describe the software environment that was used for all of the computations in this paper. We emphasize that the development of and experimentation with the models and approaches described here were greatly aided by this powerful set of software tools. We also point out that most of the tools described here are publically available.

The most important tool for our work with PDE-constrained optimization models is Sundance (see [18]). Sundance is a system for specifying, building, and applying finite element approximations to general PDEs. Sundance consists of user-callable components written in C++ that allow the user to specify the PDE and associated boundary conditions in weak form using operator overloading on a family of symbolic objects. The Sundance symbolic objects and operators can be used to assemble virtually any PDE. Each test or unknown function in a Sundance problem is constructed with a specifier of its finite element basis, and any integral can be given a specifier for the type and order of quadrature rule to be used. Stabilization terms can be added at the symbolic level; typically, these involve the mesh size h, so a special symbolic object, `CellDiameterExpr`, has been created which, when evaluated, refers to the mesh to obtain a numerical value of h on each element. The ability to specify basis, quadrature, and optional stabilization terms gives the Sundance user fine control over the discretization process.

Since it is easy to change the equation and/or the boundary conditions, it is easy to experiment with different models by making a small number of changes to a code that uses Sundance. This ability to modify models and solution procedures at a high level of abstraction was key to rapidly creating the flow-field models to be described in section 3. In fact, the code for the Reynolds-averaged Navier–Stokes (RANS) equations were created and solved in fewer than 200 lines of Sundance code. The

symbolic problem setup capability of Sundance is useful not only for rapid development of forward simulators such as our flow model, but even more important, for making possible the concurrent specification of gradient and/or adjoint equations, greatly facilitating the application of gradient-based optimization methods.

Sundance does the work of assembling matrices and vectors from a problem specification; computations on those mathematical objects are then done using the Trilinos family of solver components (see [17]). Trilinos includes a high-performance, low-level matrix/vector library (EPetra), incomplete factorization preconditioners (Ifpack), algebraic multilevel solvers and preconditioners (ML), Krylov solvers (belos), and nonlinear solvers (nox). Trilinos also provides a set of abstract interfaces allowing interoperability with other solver libraries.

Finally, we use a particular optimization method that was also built using Trilinos components and is therefore directly compatible with Sundance. This method, called O3D (see [5]), is an interior-point quadratic programming solver. It has been used as the basis for a sequential quadratic programming (SQP) algorithm that has been successfully applied to a number of problems. (See [7] and [6].) O3D requires the solution of linear systems that may conveniently be expressed mathematically as block matrices. Trilinos allows such linear operators to be created and manipulated easily. It also allowed us to experiment with different formulations and solution strategies quickly.

3. Flow-field computation. As noted above, there are only a few flow fields that need to be calculated to solve the source-inversion problem. Since all of these can be computed in advance, there is no need to be overly concerned with efficiency at this stage. We first describe and justify the turbulent flow model that we have adopted. We next show the manipulations necessary to create the appropriate weak form required by Sundance. This leads naturally to the means of handling the boundary conditions. We then discuss the use of the eikonal equation to calculate a distance variable that appears in our turbulence model. We conclude this section with a discussion of a pressure-stabilization technique that was useful in formulating a stable and well-posed computational algorithm.

3.1. The turbulence model. The time-dependent continuity and momentum (Navier–Stokes) equations for an incompressible fluid are given, in the absence of body forces, by

$$\frac{\partial u_i}{\partial x_i} = 0\,, \tag{3.1}$$

$$\frac{\partial u_i}{\partial t} + u_j \frac{\partial u_i}{\partial x_j} = -\frac{1}{\rho}\frac{\partial p}{\partial x_i} + \nu \frac{\partial^2 u_i}{\partial x_j \partial x_j}\,, \tag{3.2}$$

where repeated unhatted indices denote summation. Here, the dependent variables u_i and p are, respectively, the ith component of the fluid velocity and pressure, the independent variables x (with components x_i in a domain Ω) and t are the spatial and temporal coordinates, and ρ and ν are the density and kinematic viscosity, both of which are assumed constant.

For air flow in large rooms, typical values of the dimensionless Reynolds number $Re = UL/\nu$, where U is a characteristic flow velocity ($\sim 1\,m/s$) and L is a characteristic length scale ($\sim 10m$), are quite large. In particular, the kinematic viscosity of air at standard temperature and pressure is $\nu \sim 10^{-6}\,m^2/s$, which gives a value for the Reynolds number of $Re \sim 10^7$. In this regime, the flow is inherently unstable and

thus certain to exhibit turbulent behavior in the vicinity of walls and other obstacles. Hence, (3.1) and (3.2), short of direct numerical simulation, must be supplemented by an appropriate turbulence model.

Our approach to the problem follows the classical methodology of first decomposing the dependent variables into steady and fluctuating quantities, where the latter are not necessarily small. In particular, we define time-averaged (mean) and fluctuating quantities as $u_i = U_i + u_i'$ and $p = P + p'$, where $U_i = \overline{u}_i \equiv \lim_{\tau\to\infty} \int_{t_0}^{t_0+\tau} u_i\, dt/\tau$ and similarly for P. Substituting the definitions for u_i and p into (3.1) and (3.2) and performing the above time-averaging operation yields the well-known RANS equations given by

$$\frac{\partial U_i}{\partial x_i} = 0\,, \tag{3.3}$$

$$U_j \frac{\partial U_i}{\partial x_j} = -\frac{1}{\rho}\frac{\partial P}{\partial x_i} + \nu \frac{\partial^2 U_i}{\partial x_j \partial x_j} + \frac{1}{\rho}\frac{\partial \tau_{ij}}{\partial x_j}\,. \tag{3.4}$$

Here, the components τ_{ij} of the Reynolds stress tensor, which arise from the nonlinear terms in (3.2), are given by $\tau_{ij} = -\rho\,\overline{u_i' u_j'}$ (cf. [23]).

The closure of (3.3)–(3.4) for the mean quantities U_i and P require that additional relations be specified for the quadratic averages τ_{ij}. Although equations can be developed for these quantities, such equations generally introduce higher-order correlations (e.g., triple averages), thus giving rise to the well-known closure difficulty associated with turbulent flow modeling (cf. [23], [11], [27]). Alternatively, physically motivated assumptions regarding the relationship of τ_{ij} to the mean-flow variables can be introduced, and it is a variant of this approach that we use here. In particular, algebraic relationships (so-called zero-equation turbulence models) are introduced for the dominant components of the turbulent stress tensor based on the classical eddy-viscosity/mixing-length concept for turbulent shear- and boundary-layer flows.

For an essentially two-dimensional parallel mean flow [e.g., $U_2 \sim 0$, $U_1 \sim U_1(x_2)$], the standard analogy with laminar flow, originally postulated by Boussinesq [10], is to assume that $\tau_{12} = \tau_{21}$ can be expressed as

$$\tau_{12} = \tau_{21} = \rho\nu_t \frac{\partial U_1}{\partial x_2}\,, \tag{3.5}$$

where ν_t is the kinematic eddy viscosity. In contrast to the kinematic laminar viscosity ν, however, it is generally an unacceptably poor approximation to simply regard ν_t as a constant parameter (cf. [21]). For example, it is clear by definition that τ_{12} should approach zero at a no-slip and/or no-penetration boundary, but this is necessarily true only if ν_t vanishes there. Hence, it is customary to appeal to a mixing-length argument originally put forth by Prandtl [22] and to further express ν_t in terms of a mixing length ℓ as

$$\nu_t = \ell^2 \left|\frac{\partial U_1}{\partial x_2}\right|\,, \tag{3.6}$$

where ℓ is also a function of position that must approach zero near a wall.

For boundary-layer flows, an analysis of the turbulent boundary layer indicates that its structure consists of an inner viscous sublayer in which laminar viscous effects are important and the Reynolds stress is negligible, an outer defect layer in which the effects of ν are small but those due to ν_t are significant, and an intermediate overlap,

or matching, region in which both kinematic viscosities ν_t and ν are nonnegligible. It is in the latter regime that the mean-flow velocity varies logarithmically with distance x_2 from the wall, giving rise to the well-known "law of the wall." One modern-day uniformly valid expression for ℓ is given by

$$\frac{\ell}{\delta} = \lambda \left[1 - \exp\left(-\frac{x_2}{A}\right)\right] \tanh\left(\frac{\kappa}{\lambda} \cdot \frac{x_2}{\delta}\right) \tag{3.7}$$

(see [20]), where δ is the local turbulent boundary-layer thickness, A is a damping constant, and κ and λ are fitted constants which, for smooth walls and no mass transfer, are found to have the approximate values $\kappa = 0.41$, $\lambda = 0.085$, and $A = 26\nu/u_\tau$, where $u_\tau = (\tau_w/\rho)^{1/2}$ is the friction velocity expressed in terms of a specified wall stress τ_w. It is observed that ℓ/δ approaches zero in either a linear or quadratic fashion (depending on whether or not the damping factor is present) as $x_2 \to 0$, and approaches a constant value as x_2 exceeds the boundary-layer thickness.

In the present work, we propose an extension of (3.7), applicable to multidimensional wall-bounded flows, as follows. However, we first remark that since the boundaries in the present problem are not necessarily smooth (due to carpeting, acoustic tiles, structural protrusions, chairs, desks, counters, people, etc.), neither the laminar viscosity ν nor the above quoted turbulence parameters are regarded as given. Rather, they should ideally be fitted to actual data. Second, while the functional form given by (3.7) is retained to describe the turbulent boundary layers, an appropriate generalization must be given to account for multiple boundaries (e.g., both a floor and a ceiling) and the fact that the local mean flow is not necessarily parallel.

In generalizing the classical mixing-length model to the present two-dimensional problem, we assume that the dominant effects of turbulence are felt in the effective turbulent boundary layers in the vicinity of walls, floors, and ceilings. Consequently, it is assumed that the shear components of the Reynolds stress tensor, τ_{ij}, $i \neq j$, dominate the normal components $\tau_{\hat{i}\hat{i}}$. Hence, we define the shear components of the stress tensor $\tau^s_{ij} = \tau_{ij}(1 - \delta_{\hat{i}\hat{j}})$, where δ_{ij} is the Kronecker delta and hatted indices imply no summation, and approximate τ_{ij} with τ^s_{ij} in (3.4). Restricting further consideration to two dimensions, we then generalize (3.5) and (3.6) by representing the only shear component $\tau_{12} = \tau_{21} = -\rho\overline{u_1{}'u_2{}'}$ as

$$\frac{1}{\rho}\tau_{12} = \nu^t_{12}\frac{\partial U_1}{\partial x_2} + \nu^t_{21}\frac{\partial U_2}{\partial x_1} = \nu^t_{ij}\frac{\partial U_i}{\partial x_j}(1 - \delta_{ij})\,, \tag{3.8}$$

where

$$\nu^t_{ij} = \ell^2_{\hat{j}}\left|\frac{\partial U_i}{\partial x_j}\right|, \quad i \neq j\,, \tag{3.9}$$

with the mixing length $\ell_{\hat{j}} = \ell_{\hat{j}}(x_j)$ defined in a manner consistent with (3.7). In particular, we adopt the functional form of (3.7) for each $\ell_{\hat{j}}$, replacing the coordinate x_2 with a distance variable ℓ^* that represents the distance from the nearest *wall* boundary. The latter is calculated by first solving the eikonal equation, as described in section 3.3. For simplicity, we take the turbulence parameters δ, λ, κ, and A as constants independent of $\hat{j}$, so that in fact

$$\ell_1 = \ell_2 \equiv \ell = \delta\lambda\left[1 - \exp(-\ell^*/A)\right]\tanh\left[(\kappa/\delta\lambda)\ell^*\right]\,. \tag{3.10}$$

We note that other multidimensional generalizations are possible (cf. [14], [27]); however, the present flow-decomposition approach is appealing in that it facilitates direct use of generally accepted functional forms for parallel flows, such as (3.7).

3.2. Implementation of the flow-field model in Sundance. The use of the symbolic PDE package Sundance requires a weak form of the problem, which is then solved iteratively from an appropriately defined linearized form to determine the solution of the nonlinear flow field. While Sundance can generate such a scheme internally, we choose instead to specify this aspect of the solution procedure explicitly. We thus first derive the exact weak form of the nonlinear problem, and then prescribe an efficient functional iteration scheme for solving, via Sundance, a convergent sequence of appropriately linearized approximations.

The weak formulation of the flow-field problem is derived as follows. We first rewrite (3.3) and (3.4) in vector form as

$$\overrightarrow{\nabla} \cdot \overrightarrow{U} = 0\,, \tag{3.11}$$

$$\big(\overrightarrow{U} \cdot \overrightarrow{\nabla}\big)\overrightarrow{U} + \frac{1}{\rho}\overrightarrow{\nabla} P - \nu\nabla^2\overrightarrow{U} - \frac{1}{\rho}\big(\underline{\underline{\tau}} \cdot \overleftarrow{\nabla}\big) = 0\,, \tag{3.12}$$

where a double underlined quantity denotes a tensor (e.g., the Reynolds stress tensor $\underline{\underline{\tau}}$) and the arrow over the gradient operator denotes the location (left or right) of the object on which the corresponding operation is to be applied. Introducing test functions q and $\overrightarrow{v}$, we multiply (3.11) by q, (3.12) by $\overrightarrow{v}$, and integrate over the space Ω to obtain the weak forms

$$\int_\Omega q\overrightarrow{\nabla} \cdot \overrightarrow{U}\, d\Omega = 0\,, \tag{3.13}$$

$$\int_\Omega \overrightarrow{v} \cdot \Big[\big(\overrightarrow{U} \cdot \overrightarrow{\nabla}\big)\overrightarrow{U} + \rho^{-1}\overrightarrow{\nabla} P - \nu\nabla^2\overrightarrow{U} - \rho^{-1}\big(\underline{\underline{\tau}} \cdot \overleftarrow{\nabla}\big)\Big]\, d\Omega = 0\,. \tag{3.14}$$

Equation (3.13) is satisfactory in its present state, but in order to eliminate second derivatives (Sundance requires at most first derivatives of the variables and test functions in the integrands) and to otherwise simplify (using boundary conditions) the weak form (3.14), we manipulate the various terms in (3.14) as follows. First, though not essential, we write

$$\begin{aligned}\int_\Omega \big(\overrightarrow{v} \cdot \overrightarrow{\nabla}\big) P\, d\Omega &= \int_\Omega \Big[\overrightarrow{\nabla} \cdot \big(\overrightarrow{v} P\big) - P\big(\overrightarrow{\nabla} \cdot \overrightarrow{v}\big)\Big]\, d\Omega \\ &= \int_{\partial\Omega} P\big(\overrightarrow{v} \cdot \overrightarrow{n}\big)\, d(\partial\Omega) - \int_\Omega P\big(\overrightarrow{\nabla} \cdot \overrightarrow{v}\big)\, d\Omega,\end{aligned} \tag{3.15}$$

where $\overrightarrow{n}$ is the unit normal and the last term introduces a symmetry with respect to (3.11). Second, we remove explicit second derivatives according to

$$\begin{aligned}\int_\Omega \overrightarrow{v} \cdot \big(\nabla^2\overrightarrow{U}\big)\, d\Omega &= \int_\Omega v_i\partial_j\partial_j U_i\, d\Omega \\ &= \int_\Omega \big[\partial_j(v_i\partial_j U_i) - (\partial_j v_i)(\partial_j U_i)\big]\, d\Omega \\ &= \int_{\partial\Omega} v_i\big(\overrightarrow{\nabla} U_i\big) \cdot \overrightarrow{n}\, d(\partial\Omega) - \int_\Omega \big(\overrightarrow{\nabla} v_i\big) \cdot \big(\overrightarrow{\nabla} U_i\big)\, d\Omega \\ &= \int_{\partial\Omega} \overrightarrow{v} \cdot \big(\overrightarrow{U}\overleftarrow{\nabla}\big) \cdot \overrightarrow{n}\, d(\partial\Omega) - \int_\Omega \big(\overrightarrow{v}\overleftarrow{\nabla}\big) : \big(\overrightarrow{U}\overleftarrow{\nabla}\big)\, d\Omega,\end{aligned} \tag{3.16}$$

where repeated indices imply summation, $\partial_j \equiv \partial/\partial x_j$, and the tensor contraction operator is defined by $\underline{\underline{a}} : \underline{\underline{b}} = a_{ij}b_{ij}$. Finally, since the Reynolds stress $\underline{\underline{\tau}}$ involves first

derivatives of the flow variables, which would thus lead to the appearance of second derivatives in (3.14), we rewrite the last term in (3.14) as

$$\begin{aligned}\int_\Omega \overrightarrow{v}\cdot\left(\underline{\underline{\tau}}\cdot\overleftarrow{\nabla}\right)d\Omega &= \int_\Omega v_i\partial_j\tau_{ij}\,d\Omega \\ &= \int_\Omega \left[\partial_j(v_i\tau_{ij}) - \tau_{ij}(\partial_j v_i)\right]d\Omega \\ &= \int_{\partial\Omega}\overrightarrow{v}\cdot\underline{\underline{\tau}}\cdot\overrightarrow{n}\,d(\partial\Omega) - \int_\Omega \underline{\underline{\tau}} : \left(\overrightarrow{v}\overleftarrow{\nabla}\right)d\Omega.\end{aligned} \tag{3.17}$$

Substituting (3.15)–(3.17) into (3.14), the weak form of momentum conservation thus becomes

$$\begin{aligned}\int_\Omega \Big[\overrightarrow{v}\cdot\left(\overrightarrow{U}\cdot\overleftarrow{\nabla}\right)\overrightarrow{U} - \rho^{-1}P\left(\overleftarrow{\nabla}\cdot\overrightarrow{v}\right) + \nu\left(\overrightarrow{v}\overleftarrow{\nabla}\right) : \left(\overrightarrow{U}\overleftarrow{\nabla}\right) + \rho^{-1}\,\underline{\underline{\tau}} : \left(\overrightarrow{v}\overleftarrow{\nabla}\right)\Big]\,d\Omega \\ + \int_{\partial\Omega}\left[\rho^{-1}P\left(\overrightarrow{v}\cdot\overrightarrow{n}\right) - \nu\,\overrightarrow{v}\cdot\left(\overrightarrow{U}\overleftarrow{\nabla}\right)\cdot\overrightarrow{n} - \rho^{-1}\,\overrightarrow{v}\cdot\underline{\underline{\tau}}\cdot\overrightarrow{n}\right]d(\partial\Omega) = 0.\end{aligned} \tag{3.18}$$

The boundary integral in (3.18) can be simplified further by considering the boundary conditions (at inlet(s), outlet(s), and walls (including floor and ceilings)) for the problem of interest. At outlets, denoted by $\partial\Omega_o$, we assume no forcing, which is expressed as

$$\left[\rho^{-1}P\overrightarrow{n} - \left(\overrightarrow{U}\overleftarrow{\nabla}\right)\cdot\overrightarrow{n} - \rho^{-1}\,\underline{\underline{\tau}}\cdot\overrightarrow{n}\right]_{\partial\Omega_o} = 0\,. \tag{3.19}$$

Consequently, condition (3.19) implies that the boundary integral in (3.18) vanishes on an outlet boundary $\partial\Omega_o$. At an inlet, denoted by $\partial\Omega_I$, or at walls, denoted by $\partial\Omega_w$, the boundary conditions for the present problem are given by $\overrightarrow{U} = \overrightarrow{U}_I$ and $\overrightarrow{U} = \overrightarrow{U}_w = 0$, respectively, where the latter corresponds to a no-slip, no-penetration condition at $\partial\Omega_w$. (It is also true, by definition, that the Reynolds stress tensor $\underline{\underline{\tau}} = 0$ on $\partial\Omega_w$.) However, since (3.19) involves only gradients of $\overrightarrow{U}$, we approximate these conditions as

$$-\left(\overrightarrow{U}\overleftarrow{\nabla}\right)\cdot\overrightarrow{n}\Big|_{\partial\Omega_{I,w}} = \epsilon^{-1}\left(\overrightarrow{U} - \overrightarrow{U}_{I,w}\right)_{\partial\Omega_{I,w}}, \quad \epsilon \ll 1\,, \tag{3.20}$$

which implies $\left|\left|\overrightarrow{U} - \overrightarrow{U}_{I,w}\right|\right| \to 0$ on $\partial\Omega_{I,w}$ as $\epsilon \to 0$. Substituting (3.19) and (3.20) into (3.18) and taking the limit $\epsilon \to 0$ thus allows us, after invoking the assumption that the discrete basis functions used to represent the variables are nodal [19], to replace the boundary integral in the latter with a simple penalty term according to

$$\begin{aligned}\int_{\partial\Omega}\left[\rho^{-1}P\left(\overrightarrow{v}\cdot\overrightarrow{n}\right) - \nu\,\overrightarrow{v}\cdot\left(\overrightarrow{U}\overleftarrow{\nabla}\right)\cdot\overrightarrow{n} - \rho^{-1}\,\overrightarrow{v}\cdot\underline{\underline{\tau}}\cdot\overrightarrow{n}\right]d(\partial\Omega) \\ \sim \int_{\partial\Omega}\overrightarrow{v}\cdot\left(\overrightarrow{U} - \overrightarrow{U}_{I,w}\right)d(\partial\Omega).\end{aligned} \tag{3.21}$$

Hence, the final weak form of momentum conservation to be used in the present study is given by

$$\begin{aligned}\int_\Omega \Big[\overrightarrow{v}\cdot\left(\overrightarrow{U}\cdot\overleftarrow{\nabla}\right)\overrightarrow{U} - \rho^{-1}P\left(\overleftarrow{\nabla}\cdot\overrightarrow{v}\right) + \nu\left(\overrightarrow{v}\overleftarrow{\nabla}\right) : \left(\overrightarrow{U}\overleftarrow{\nabla}\right) + \rho^{-1}\,\underline{\underline{\tau}} : \left(\overrightarrow{v}\overleftarrow{\nabla}\right)\Big]\,d\Omega \\ + \int_{\partial\Omega_{I,w}}\overrightarrow{v}\cdot\left(\overrightarrow{U} - \overrightarrow{U}_{I,w}\right)d(\partial\Omega) = 0,\end{aligned} \tag{3.22}$$

where the (shear) components of $\underline{\underline{\tau}} \approx \underline{\underline{\tau}}^s$ are given by (3.8)–(3.10).

We observe that (3.22) is nonlinear with respect to $\overrightarrow{U}$ through the appearance of the factors $\big(\overrightarrow{U}\cdot\overrightarrow{\nabla}\big)\overrightarrow{U}$ and the Reynolds stress tensor $\underline{\underline{\tau}}$, where, from (3.8) and (3.9), the latter consists of the shear components $\tau_{12}=\tau_{21}=\rho\ell_1^2\big|\partial U_2/\partial x_1\big|(\partial U_2/\partial x_1)+\rho\ell_2^2|\partial U_1/\partial x_2|(\partial U_1/\partial x_2)$. Consequently, since we ultimately require a weak form that is linear with respect to the dependent variables (as well as the test functions), we solve the coupled system (3.13) and (3.22) by an iterative approach. In particular, the nonlinear terms are formally linearized and functional iteration on the resultant linear problem is employed to obtain the solution to the original nonlinear problem. Thus, if the solution to the kth iterate $\overrightarrow{U}_k$ is known, the next iterate $\overrightarrow{U}_{k+1}$ is determined by solving a linear problem obtained from the original nonlinear equations by linearizing nonlinear terms about $\overrightarrow{U}_k$. In particular, making the a priori assumption that $|\overrightarrow{U}_{k+1}-\overrightarrow{U}_k|$ is sufficiently small, we have

$$
\begin{aligned}
(3.23)\qquad \big(\overrightarrow{U}_{k+1}\cdot\overrightarrow{\nabla}\big)\overrightarrow{U}_{k+1} &= \Big\{\big[\overrightarrow{U}_k+\big(\overrightarrow{U}_{k+1}-\overrightarrow{U}_k\big)\big]\cdot\overrightarrow{\nabla}\Big\}\big[\overrightarrow{U}_k+\big(\overrightarrow{U}_{k+1}-\overrightarrow{U}_k\big)\big]\\
&\approx \big(\overrightarrow{U}_k\cdot\overrightarrow{\nabla}\big)\overrightarrow{U}_k+\Big[\big(\overrightarrow{U}_{k+1}-\overrightarrow{U}_k\big)\cdot\overrightarrow{\nabla}\Big]\overrightarrow{U}_k\\
&\quad+\big(\overrightarrow{U}_k\cdot\overrightarrow{\nabla}\big)\big(\overrightarrow{U}_{k+1}-\overrightarrow{U}_k\big)\\
&= \big(\overrightarrow{U}_{k+1}\cdot\overrightarrow{\nabla}\big)\overrightarrow{U}_k+\big(\overrightarrow{U}_k\cdot\overrightarrow{\nabla}\big)\overrightarrow{U}_{k+1}-\big(\overrightarrow{U}_k\cdot\overrightarrow{\nabla}\big)\overrightarrow{U}_k
\end{aligned}
$$

and, denoting $\partial U_i/\partial x_j$ by $U_{i,j}$,

$$
\begin{aligned}
(3.24)\ \big|U_{i,j}^{(k+1)}\big|\,U_{i,j}^{(k+1)} &= \Big[U_{i,j}^{(k+1)}U_{i,j}^{(k+1)}\Big]^{1/2}U_{i,j}^{(k+1)}\\
&= \Big\{\big[U_{i,j}^{(k)}+\big(U_{i,j}^{(k+1)}-U_{i,j}^{(k)}\big)\big]\big[U_{i,j}^{(k)}+\big(U_{i,j}^{(k+1)}-U_{i,j}^{(k)}\big)\big]\Big\}^{1/2}\\
&\qquad\times\big[U_{i,j}^{(k)}+\big(U_{i,j}^{(k+1)}-U_{i,j}^{(k)}\big)\big]\\
&= \Big\{U_{i,j}^{(k)}U_{i,j}^{(k)}+2\big(U_{i,j}^{(k+1)}-U_{i,j}^{(k)}\big)U_{i,j}^{(k)}+\big[U_{i,j}^{(k+1)}-U_{i,j}^{(k)}\big]^2\Big\}^{1/2}\\
&\qquad\times\big[U_{i,j}^{(k)}+\big(U_{i,j}^{(k+1)}-U_{i,j}^{(k)}\big)\big]\\
&\approx \Big[U_{i,j}^{(k)}U_{i,j}^{(k)}\Big]^{1/2}\Big\{1+\big[U_{i,j}^{(k)}\big]^{-1}\big[U_{i,j}^{(k+1)}-U_{i,j}^{(k)}\big]\Big\}\\
&\qquad\times\Big\{U_{i,j}^{(k)}+\big[U_{i,j}^{(k+1)}-U_{i,j}^{(k)}\big]\Big\}\\
&\approx \Big[U_{i,j}^{(k)}U_{i,j}^{(k)}\Big]^{1/2}\Big[2U_{i,j}^{(k+1)}-U_{i,j}^{(k)}\Big].
\end{aligned}
$$

Using these linearized representations in (3.22), we thus arrive at a functional iteration scheme that computes the successive approximations $\overrightarrow{U}_{k+1}$ and P_{k+1} in terms of initial guesses $\overrightarrow{U}_k$ and P_k according to

$$
(3.25)\qquad \int_\Omega q\,\overrightarrow{\nabla}\cdot\overrightarrow{U}_{k+1}\,d\Omega=0\,,
$$

$$
\begin{aligned}
(3.26)\qquad \int_\Omega\Big\{\overrightarrow{v}\cdot\big[\big(\overrightarrow{U}_{k+1}\cdot\overrightarrow{\nabla}\big)\overrightarrow{U}_k+\big(\overrightarrow{U}_k\cdot\overrightarrow{\nabla}\big)\overrightarrow{U}_{k+1}-\big(\overrightarrow{U}_k\cdot\overrightarrow{\nabla}\big)\overrightarrow{U}_k\big]\\
-\rho^{-1}P_{k+1}\big(\overrightarrow{\nabla}\cdot\overrightarrow{v}\big)+\nu\big(\overrightarrow{v}\overleftarrow{\nabla}\big):\big(\overrightarrow{U}_{k+1}\overleftarrow{\nabla}\big)+\rho^{-1}\,\underline{\underline{\tau}}^{\,l}:\big(\overrightarrow{v}\overleftarrow{\nabla}\big)\Big\}\,d\Omega\\
+\int_{\partial\Omega_{I,w}}\overrightarrow{v}\cdot\big(\overrightarrow{U}-\overrightarrow{U}_{I,w}\big)\,d(\partial\Omega)\;=\;0,
\end{aligned}
$$

where $\underline{\underline{\tau}}^l$, whose components are given by $\tau^l_{ij} = \rho\big[U^{(k)}_{i,j}U^{(k)}_{i,j}\big]^{1/2}\big[2U^{(k+1)}_{i,j} - U^{(k)}_{i,j}\big](1 - \delta_{\hat{i}\hat{j}})$, is the linearized form of $\underline{\underline{\tau}}$ based on (3.24). Equation (3.26) is linear with respect to the unknown function $\overrightarrow{U}_{k+1}$ and can be handled directly by Sundance.

The above sequence of approximations is expected to be convergent, provided the kth iterate is a sufficiently good approximation to the actual solution of the nonlinear problem. In practice, this generally requires that the iteration scheme defined by (3.25) and (3.26) be embedded in an outer iterative loop with respect to increasing values of a suitably defined Reynolds number U^*L^*/ν, where L^* and U^* are appropriately defined length and time scales. Indeed, for a sufficiently small initial value of the Reynolds number, the linear terms dominate and convergence is therefore more likely to be achieved in that case. The resulting converged solution for such a given Reynolds number then provides a good starting guess for the next sequence of iterations at a modestly larger value of that parameter, and so forth. It is also found, consistent with the need to resolve boundary layers and other finer aspects of the flow field, that the ability to achieve convergence at a given Reynolds number depends on the discretization (mesh). In particular, larger Reynolds numbers are generally found to require finer discretizations, especially in the vicinity of boundaries, independent of the goodness of the previous iterate with respect to the true solution.

3.3. The eikonal equation and its regularization. The turbulence model described in section 3.1 requires knowing the wall distance ℓ^* introduced in the expression (3.10) for the mixing length ℓ. The wall distance function can, in principle, be computed exactly given knowledge of the geometry of the walls; however, this computation is tedious and must be repeated for each new geometry considered. To simplify and automate the computation of ℓ^*, we use the eikonal equation of geometrical optics (e.g., [9]) which, with appropriate boundary conditions, has the wall distance ℓ^* as a solution. We can thus bypass tedious computational geometry and obtain the wall function as the solution of a PDE. This additional PDE, with regularization for numerical stability, is easily incorporated into our numerical algorithm and is described in detail in the appendix.

3.4. Pressure stabilization. It is well known that representing velocity and pressure with basis functions of the same order tends to be unstable (cf. [15]). Consequently, various stabilization schemes have been proposed to allow these variables to be represented with the same set of basis functions. One popular approach is to augment the continuity equation with a (small) term proportional to the Laplacian of pressure according to

$$\overrightarrow{\nabla} \cdot \overrightarrow{U} = \beta h^2 \nabla^2 P\,, \tag{3.27}$$

where h is the linear dimension (diameter) of the finite element discretization and β is an optimally chosen small parameter (cf. [15]). Aside from allowing equal-order interpolants, the added term has the positive effect of removing the indefiniteness of the discretized linear system and smoothing out numerical oscillations in the pressure variable.

Multiplication of (3.27) by the test function q followed by an integration over the domain Ω then leads to the weak form

$$\int_\Omega \Big[q\overrightarrow{\nabla}\cdot\overrightarrow{U} + \beta h^2(\overrightarrow{\nabla}q)\cdot(\overrightarrow{\nabla}P)\Big]\,d\Omega - \int_{\partial\Omega}\beta h^2(q\overrightarrow{\nabla}P)\cdot\overrightarrow{n}\;d(\partial\Omega) = 0\,, \tag{3.28}$$

where we have used the identity $q\nabla^2 P = \overrightarrow{\nabla}\cdot(q\overrightarrow{\nabla}P) - (\overrightarrow{\nabla}q)\cdot(\overrightarrow{\nabla}P)$ and applied the divergence theorem to obtain the form (3.28). Dropping the boundary term, which

implies that we are actually only regularizing the weak form (3.28) rather than (3.27), thus leads to

$$\int_\Omega \left[q\overrightarrow{\nabla}\cdot\overrightarrow{U} + \beta h^2 (\overrightarrow{\nabla} q)\cdot(\overrightarrow{\nabla} P)\right] d\Omega = 0 \tag{3.29}$$

and

$$\int_\Omega \left[q\overrightarrow{\nabla}\cdot\overrightarrow{U}_{k+1} + \beta h^2 (\overrightarrow{\nabla} q)\cdot(\overrightarrow{\nabla} P_{k+1})\right] d\Omega = 0 \tag{3.30}$$

in place of (3.13) and (3.25), respectively.

4. The optimization problem. In this section we consider the steady-state transport of the toxin in the flow field developed in the preceding section. We thus make the implicit assumption that the toxin concentration is insufficient to affect the flow field itself, but that the transport influences of the flow clearly play a key role in determining the distribution of the toxin. After first formulating the toxin-transport model, the optimization problem is then constructed so as to recover the source location(s) from specific concentration data, where the latter consist of readings obtained from sensors placed strategically throughout the building. Prior to that phase of the calculation, we consider several possible optimization formulations before settling on the one used in our numerical experiments.

4.1. The toxin-transport model. Given the time-averaged flow field U calculated in section 3, we can write the corresponding time-averaged continuity equation describing the transport of the toxin as

$$k\nabla^2 c - \big(U\cdot\overrightarrow{\nabla}\big)c + \overrightarrow{\nabla}\cdot\overrightarrow{\mathcal{J}} + s = 0\,, \tag{4.1}$$

where $c(\overrightarrow{x})$ and $s(\overrightarrow{x})$ are, respectively, the (time-averaged) concentration and source fields at a point $\overrightarrow{x}\in\Omega$, k is the binary mass diffusivity (assumed constant) of the toxin with respect to the mixture, and $\overrightarrow{\mathcal{J}}$ is the turbulent mass-flux vector whose components are defined as $\mathcal{J}_i = -\overline{u_i' c'}$. The derivation of an expression for $\mathcal{J}$ in terms of time-averaged quantities follows in an analogous fashion the derivation of the expression for the Reynolds stress tensor in section 3.1. In particular, by analogy with Fick's law of diffusion, it is logical to write, at least for nearly parallel flows $\overrightarrow{U}\sim\big(U_1(x_2),0\big)$,

$$\overrightarrow{\mathcal{J}}\sim(0,\mathcal{J}_2)\,,\quad \mathcal{J}_2 = k^t\frac{\partial c}{\partial x_2}\,,\quad k^t = \ell^2\left|\frac{\partial U_1}{\partial x_2}\right|\,, \tag{4.2}$$

where k^t is the turbulent, or eddy, diffusivity and ℓ is the same mixing length that was defined by (3.6) and (3.7) (cf. [3]). This expression may be extended to multi-dimensional flows using reasoning somewhat similar to that which led to (3.8) and (3.9). Hence, we define

$$\mathcal{J}_j = k^t_{\hat{j}}\frac{\partial c}{\partial x_j}\,,\quad k^t_{\hat{j}} = \ell^2_{\hat{j}}\left|\frac{\partial U_i}{\partial x_j}\right|\,,\quad i\neq j\,, \tag{4.3}$$

where the $\ell_{\hat{j}}$ are the same as those introduced in (3.9) and, for our present purposes, given by the single expression for $\ell_{\hat{j}}\equiv\ell$ in (3.10).

As in the flow-field problem, a corresponding reduction of the weak form of the toxin continuity equation (4.1) is obtained by first multiplying that equation by a scalar test function r and integrating. This gives

$$\int_\Omega r\left[(\overrightarrow{U}\cdot\overrightarrow{\nabla})c - k\nabla^2 c - \overrightarrow{\nabla}\cdot\overrightarrow{\mathcal{J}} - s\right]\,d\Omega = 0\,, \tag{4.4}$$

where the components $\mathcal{J}_j$ of $\overrightarrow{\mathcal{J}}$ are given in terms of $\partial c/\partial x_j$ according to (4.3). As in the case of the Navier–Stokes equations, we manipulate the individual terms appearing in (4.4) so as to eliminate second derivatives. In particular, by analogy with (3.16), we have

$$\begin{aligned}\int_\Omega r\nabla^2 c\,d\Omega &= \int_\Omega r\partial_j\partial_j c\,d\Omega\\ &= \int_\Omega \left[\partial_j(r\partial_j c) - (\partial_j r)(\partial_j c)\right]\,d\Omega\\ &= \int_{\partial\Omega} r\left(\overrightarrow{\nabla}c\right)\cdot\overrightarrow{n}\;d(\partial\Omega) - \int_\Omega \left(\overrightarrow{\nabla}r\right)\cdot\left(\overrightarrow{\nabla}c\right)d\Omega\,,\end{aligned} \tag{4.5}$$

and, since $\overrightarrow{\mathcal{J}}^i$ depends on derivatives of c, we also write, analogous to (3.17),

$$\begin{aligned}\int_\Omega r\overrightarrow{\nabla}\cdot\overrightarrow{\mathcal{J}}\,d\Omega &= \int_\Omega \left[\overrightarrow{\nabla}\cdot\left(r\overrightarrow{\mathcal{J}}\right) - \overrightarrow{\mathcal{J}}\cdot\left(\overrightarrow{\nabla}r\right)\right]\,d\Omega\\ &= \int_{\partial\Omega} r\overrightarrow{\mathcal{J}}\cdot\overrightarrow{n}\;d(\partial\Omega) - \int_\Omega \overrightarrow{\mathcal{J}}\cdot\left(\overrightarrow{\nabla}r\right)d\Omega.\end{aligned} \tag{4.6}$$

Finally, since $\overrightarrow{\nabla}\cdot\overrightarrow{U} = 0$,

$$\begin{aligned}\int_\Omega r\left(\overrightarrow{U}\cdot\overrightarrow{\nabla}\right)c\,d\Omega &= \int_\Omega r\overrightarrow{\nabla}\cdot\left(c\overrightarrow{U}\right)d\Omega\\ &= \int_\Omega \left[\overrightarrow{\nabla}\cdot\left(rc\overrightarrow{U}\right) - c\overrightarrow{U}\cdot\left(\overrightarrow{\nabla}r\right)\right]\,d\Omega\\ &= \int_{\partial\Omega}\left(rc\overrightarrow{U}\right)\cdot\overrightarrow{n}\;d(\partial\Omega) - \int_\Omega c\overrightarrow{U}\cdot\left(\overrightarrow{\nabla}r\right)d\Omega.\end{aligned} \tag{4.7}$$

Thus, in place of (4.4), the weak form of toxin mass conservation may be written as

$$\begin{aligned}&\int_\Omega \left\{\left(\overrightarrow{\nabla}r\right)\cdot\left[k\overrightarrow{\nabla}c - c\overrightarrow{U} + \overrightarrow{\mathcal{J}}\right] - rs\right\}\,d\Omega\\ &\qquad - \int_{\partial\Omega} r\left(k\overrightarrow{\nabla}c - c\overrightarrow{U} + \overrightarrow{\mathcal{J}}\right)\cdot\overrightarrow{n}\;d(\partial\Omega) = 0.\end{aligned} \tag{4.8}$$

As before, the boundary integral in (4.8) can be simplified by applying the boundary conditions. In particular, the assumption of no mass transport across the walls implies that the integrand in the boundary integral vanishes on $\partial\Omega_w$ since the mixing lengths ℓ_j, and hence $\overrightarrow{\mathcal{J}}$, also vanish there. At the inlet and outlet boundaries $\partial\Omega_I$ and $\partial\Omega_o$, we neglect the contribution of $\overrightarrow{\mathcal{J}}$ because the mixing length ℓ is likely to be small along most of such boundaries. In addition, at the inlet, we specify the incoming flux fraction (or concentration) of toxin α, which implies the set of conditions

$$\left[\left(c\overrightarrow{U}_I - k\overrightarrow{\nabla}c\right)\cdot\overrightarrow{n}\right]_{\partial\Omega_I} = -\alpha\overrightarrow{U}_I\cdot\overrightarrow{n}\,, \tag{4.9}$$

where $\overrightarrow{n}$ always denotes the outward-facing normal at the boundary. At the outflow boundary, we represent the boundary condition in the general form

$$\left[\left(\overrightarrow{\nabla} c\right) \cdot \overrightarrow{n} + \beta c\right]_{\partial\Omega_o} = 0\,, \tag{4.10}$$

where $\beta > 0$ is a loss coefficient; the choice $\beta = 0$ would correspond to a finite-boundary approximation if the actual outlet were at infinity. Applying these conditions to (4.8) then yields the final weak form of species mass conservation as

$$\begin{aligned}\int_\Omega \left\{\left(\overrightarrow{\nabla} r\right) \cdot \left[k\overrightarrow{\nabla} c - c\overrightarrow{U} + \overrightarrow{\mathcal{J}}\right] - rs\right\} d\Omega - \int_{\partial\Omega_I} r\alpha_i \overrightarrow{U}_I \cdot \overrightarrow{n}\; d(\partial\Omega) \\ + \int_{\partial\Omega_o} rc\left(\overrightarrow{U} \cdot \overrightarrow{n} + k\beta\right) d(\partial\Omega) = 0.\end{aligned} \tag{4.11}$$

We note that in the event that there is no incoming flux of toxin, $\alpha = 0$ and the first boundary integral in (4.11) vanishes.

4.2. The source-inversion optimization problem. We can now form the preliminary version of the optimization problem. Suppose we have data c_i^* that is the concentration reading from sensor i. Then we seek the source $s(x)$ that creates a calculated concentration field that best matches this data. That is, we define

$$f_p(s, c) = \frac{1}{2}\sum_{i=1}^{N_s} (c(x_i) - c_i^*)^2,$$

where N_s is the number of sources. This is the L_2 measure of the discrepancy. We then write the first version of the optimization problem as

$$\begin{aligned} &\underset{c,\,s}{\text{minimize}} \quad f_p(c, s) \\ &\text{subject to (4.11).}\end{aligned} \tag{4.12}$$

There are several points that need to be addressed before we have a problem that we can attempt to solve.

First, there are several ways to handle the source term in (4.11). One way is to assume a model for the sources. For example, we could assume a Gaussian model for source i, i.e.,

$$s_i(x) = \alpha_i e^{\beta_i (x - x_i)^2},$$

where α_i and β_i are constants and $x_i \in \Omega$ is the location. Then the source field is

$$s(x) = \sum_{i=1}^{S} s_i(x),$$

where S is the number of sources. Alternatively, we could consider source i to be a δ-function with a specified strength. In both of these cases, the number of sources is not known in advance and must be determined as part of the source inversion problem. Also, it is clear that some of the parameters enter the source term nonlinearly. The advantage of these models, however, is that there is a small number of parameters to estimate.

Another way to proceed is to leave $s(x)$ as simply a function over all of Ω to be determined. The two major advantages of this approach are that the function s enters (4.11) linearly and there is no need to know in advance the number of sources. The disadvantage is that the number of parameters to be estimated is the number of nodes, a potentially large number. We believe, however, that the advantages of this approach outweigh the disadvantages and we thus choose this form for the source term.

Second, given the form of the source term, there is a need to regularize the problem. As posed, the problem is underdetermined; there are many source functions that will make $f_p = 0$, but they will be excessively oscillatory. We therefore need to smooth the solution since we expect s to be essentially zero except where real sources exist. Thus we use standard Tikhonov regularization, i.e., we modify the objective function f_p to be of the form

$$f(c, s) = f_p + \frac{1}{2}\sigma \int_\Omega (\nabla s)^2.$$

Our modified form of the optimization problem becomes

$$\begin{aligned} &\underset{c,\, s}{\text{minimize}} \quad f(c, s) \\ &\text{subject to (4.11).} \end{aligned} \tag{4.13}$$

As posed, problem (4.13) is an equality-constrained quadratic programming problem. It is well known that, when discretized, such quadratic programs (QPs) can be solved by solving the linear system that arises from writing the first order necessary, or KKT, conditions. We point out that Sundance allows for the creation of discrete functions defined on the mesh. Thus we can easily define the objective function $f(c, s)$ and then create the Lagrangian

$$L(c, s, \lambda) = f(c, s) + \int_\Omega \lambda h(c, s),$$

where λ is the Lagrange multiplier and $h(c, s)$ is the left hand side of (4.11). In Sundance we can define an expression that is the variation of $L(c, s, \lambda)$ with respect to its arguments, i.e., the first order conditions. We can then create a Sundance problem, specify a solver (and a preconditioner), and solve the system. This procedure takes fewer than 20 lines of code.

We need to consider, however, the possibility of adding additional constraints to problem (4.13). In particular, we know that both the source field and the concentration field must be nonnegative. In the case of (4.11), it is easy to see that if $s \geq 0$, then it follows that $c \geq 0$. Thus, the final form of the optimization problem that we consider is

$$\begin{aligned} &\underset{c,\, s}{\text{minimize}} \quad f(c, s) \\ &\text{subject to} \quad \begin{aligned} &(4.11) \\ &s \geq 0. \end{aligned} \end{aligned} \tag{4.14}$$

Sundance can also be used to create this problem. Let

$$a^{\mathrm{t}} x + \frac{1}{2} x^{\mathrm{t}} Q x \tag{4.15}$$

be the general form of a quadratic objective function, where $a \in \mathcal{R}^n$ and Q is an $(n \times n)$ symmetric matrix. We create a Sundance problem for the function given by $f(c, s)$. Then by getting the linear operator for this problem, we have the matrix Q, and by getting the right-hand side we have a. Similarly, we can then get the matrix and right-hand side associated with the Sundance problem corresponding to (4.11), thus obtaining the discretized linear equality constraints. It is then trivial in Trilinos to create the identity matrix (never actually formed) and vector of zeros corresponding to the constraints $s \geq 0$. These matrices and vectors can be passed to any appropriate QP solver.

We now make the following remarks about problem (4.14) that motivate our choice of QP solver:

1. Recall that we are not interested in high accuracy solutions.
2. With this formulation, there will be many near-active constraints, i.e., many points with very small, but positive, function values; such problems often create severe computational difficulties.
3. We expect that our formulation, which uses the Tikhonov regularization term, will tend to smear the constraints. Thus, we are not concerned with getting the correct active set.
4. Given that we do not care about getting the correct active set, we are certainly not interested in getting the Lagrange multipliers.
5. Since the equality constraint given by (4.11) will be a discretized PDE in the QP, we do not need to worry about satisfying these constraints to high accuracy.

The above observations lead us to consider the use of O3D to solve the inequality-constrained problem (4.14). O3D is a primal, interior-point method that can be controlled so that estimates of the multipliers are not computed. It operates on a QP with the objective function given by (4.15) and with only inequality constraints of the form

$$Ax + b \leq 0.$$

In essence, O3D, which stands for optimizing over three-dimensional subspaces, is an iterative method that at each iteration creates a three-dimensional subspace, solves the QP restricted to that subspace, and moves 99% of the distance to the boundary in the direction implied by the solution. The three directions, p_i, all satisfy a linear system of the form

$$\left(A^{\mathsf{t}} D^2 A + Q/\gamma\right) p_i = t_i, \qquad i = 1, 2, 3,$$

where

$$D = \operatorname{diag} 1/\left(Ax + b\right)_i,$$

γ is a scalar computed at each iteration, and t_i is an appropriate right-hand side. There are, of course, many other details; see [5] for a more complete description.

In the tests reported in the next section, we tried several formulations of the problem in an attempt to find an efficient strategy. The original strategy for handling equality constraints in O3D is to convert them into a pair of inequality constraints and then to form a "big M phase 1 problem." This problem has an artificial variable that is forced to zero as the solution draws near. In the limit, this forces the two inequalities corresponding to an equality constraint to become equal, thus satisfying

the equality constraint. Although this has often worked well in practice, it was not efficient enough for our purposes here; we thus looked for a more efficient formulation. Although not often recommended, the best formulation that we found was to create a penalty form of (4.14) so that we have a problem with only inequality constraints. This formulation allows us to control the tolerance to which the equality constraints will be satisfied. As noted above, it does not make sense to require satisfaction of the equality constraint to high accuracy since it is a finite element approximation to (4.11). The penalty form that we consider is given by

$$\begin{aligned} &\underset{c,\,s}{\text{minimize}} \quad f(c,s) + \frac{1}{2}\rho \left\| h(c,s) \right\|^2 \\ &\text{subject to } s \geq 0. \end{aligned} \tag{4.16}$$

We choose an increasing sequence ρ_i and solve (4.16) for $\rho_i,\ i = 1, \ldots,$ for a relaxed set of convergence criteria. After each solution, we check the value of $\|h(c,s)\|$. If the value is less than ϵ, where ϵ is a given tolerance, then we fix ρ at the current value and solve (4.16) to a tighter set of convergence criteria. We note that in this form it is easy to get a good interior starting approximation by solving problem (4.13) for s (by solving one linear system) and then modifying any value of $s \leq 0$ to be some small positive value. As reported in the next section, this allowed convergence to acceptable accuracy in a very small number of iterations.

5. Numerical results. To conduct our numerical tests, it was first necessary to solve the flow model developed in section 3. We therefore first describe a simple two-dimensional room with one inlet and one outlet and display the resulting flow field. After commenting on some of the realistic features of this field, we then show the source-inversion results that were obtained when sources and sensors were placed in the room.

5.1. Sample two-dimensional problem geometry and flow field. As indicated above, a simple two-dimensional room was constructed with an inlet and an outlet, as shown by the domain in Figure 1. The main part of the room is 15×20 units, with two small (2×2) inlet/outlet sections on the left and right, respectively. Various triangular meshes for this geometry were created with Shewchuk's mesh-generating package, Triangle [24], [25], with two layers of additional nodes specified along the boundaries to better resolve the (turbulent) boundary-layer structure in the vicinity of the walls. In all, the flow field used in the numerical calculations described below was generated by Sundance on a roughly 15,000-node mesh with approximately 31,000 triangles. This degree of mesh resolution enabled us to calculate, using the iterative procedure described in section 3, a flow field corresponding to fairly high Reynolds numbers (the source-inversion calculations described below corresponded to $Re = 10{,}000$). Generally speaking, the ability to iterate to increasingly large Reynolds numbers required increasingly finer meshes.

As suggested in section 3.2, the specific procedure we successfully employed, given a sufficiently fine mesh, was to start with zero initial data for the velocity and pressure fields and iterate to a converged solution for $Re = 100$. That solution was then used as initial data for the $Re = 200$ problem, and so forth, in increments of 100 in the Reynolds number. Since (3.24) represents a formal linearization of the flow-field model based on Taylor expansions of the nonlinear terms, roughly quadratic convergence was achieved at each step in the Reynolds-number loop. Typically, only four or five iterations were required to meet our fairly stringent convergence criteria, although

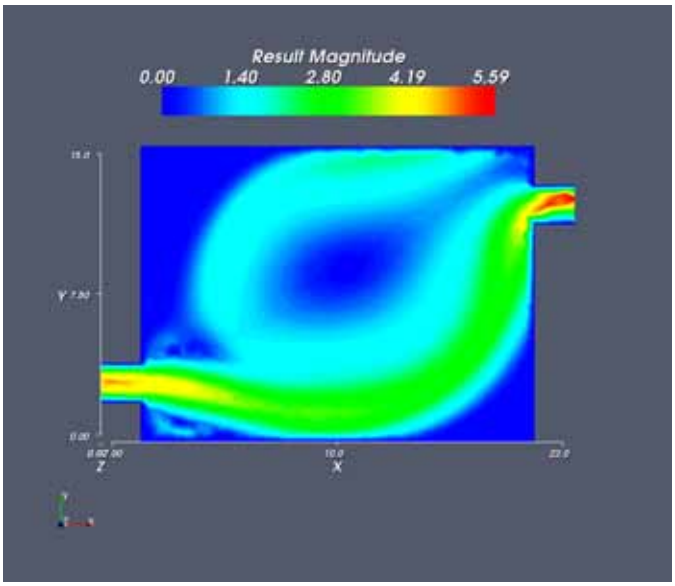


FIG. 1. *Calculated vector flow field for* $Re = 10{,}000$.

somewhat more (about 10) were usually needed on the first $Re = 100$ step due to our starting the iteration from zero initial data. A convergence failure at a particular Reynolds number generally implied an insufficiently fine mesh and could be remedied by recalculating on a finer mesh.

The inlet velocity profile was specified to be a Poiseuille flow with the peak inlet flow velocity at the center of the profile taken to be 4. The resulting vector flow field for $Re = 10{,}000$ is shown in Figure 1. It is readily observed that the main flow direction is from the inlet to the outlet. A more careful look, however, reveals several recirculation zones and other structures that are indicative of a realistic flow field.

5.2. Source-inversion tests. To test our source-inversion procedure we first generate some data on the two-dimensional example above. We do this by assuming a Gaussian source as described in section 4 and solving the forward problem, i.e., we solve (4.11). In particular, we take our sources to be

$$s(x) = \alpha_i e^{\beta_i (x - x_i)^2},$$

where $\alpha_i = 20$ and $\beta_i = 2$ for all i. We then specify the location for a set of sensors in the room, get the concentration c at each sensor location, and modify it by a random value of $\pm 5\%$. (The value of 5% was chosen in consultation with one of our Sandia National Laboratory colleagues who is working on the design of the sensors [12].) We investigated two patterns of the locations for the sensors. The first was a regular pattern of 9 sensor locations; the second was a hand-selected pattern of 30 locations, shown in Figure 2. This second pattern will be discussed further later.

The results for the 9-sensor source-inversion problems are as follows. First, it was sometimes possible to achieve reasonable predictions, even without using the inequality constraints, but it was very easy to pick source locations for which the prediction of both the number and location of the sources was poor. Without using inequality constraints, we often observed negative values of the source field, but even

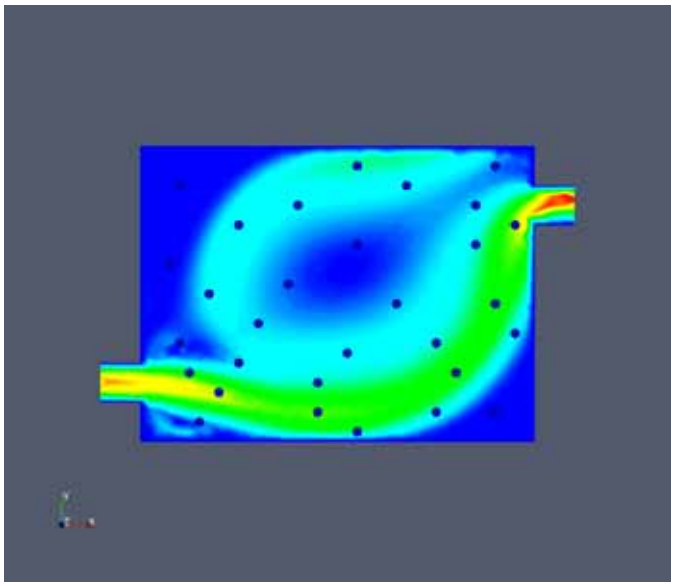


FIG. 2. *Flow field with* 30 *irregularly spaced sensors.*

when we used the inequality constraints, the results were still poor. We conclude from this that 9 sensors on a regular grid are not sufficient to reconstruct the sources for many source locations.

To determine an effective sensor arrangement, it is necessary to consider how the toxin is transported throughout the building and how this information is represented by the sensor readings. Here, we will look at two different cases. First, suppose a source is located in a region of the building in which the air flow is minimal, e.g., a corner. Over time the toxin will accumulate and any sensor in that region will exhibit a high concentration reading while all other sensors return comparatively low readings, assuming only one source exists. Thus, given at least one sensor in this region, any solution would exhibit a source in the proper area of the building. Since any sensor in this location will eventually have a comparatively high concentration reading, we require only one sensor to adequately determine sources in these types of regions. For the second case, suppose we have a source in the main stream of the flow. Instead of accumulating around the source location, the toxin will be distributed throughout the building as dictated by the flow field. Thus, if the main stream of the flow contains too few sensors, as in the 9-sensor arrangement, many different source locations could result in the same set of concentration readings. Therefore, in the main stream of the flow it is necessary to place sensors based on both the direction and the magnitude of the flow.

Using this knowledge, we picked sensor locations by hand by conducting many experiments. We make no claims, however, that these are the optimal sensor locations, a topic to which we return in section 6. With these locations, we were able to locate any pair of sources with acceptable accuracy.

We can draw several conclusions from these results:

1. Thirty sensors usually get acceptable accuracy.
2. Adding inequality constraints sometimes helps significantly by reducing nonphysical oscillations of the source field and is not computationally too

expensive. In fact, it generally took fewer than 10 iterations of O3D to solve the problem to acceptable accuracy.
3. Experience and knowledge of the flow field helps to predict locations of the sources.
4. As expected, because of our use of the Tikhonov regularization, the sources are smeared, especially for sources placed in the main stream of the flow. Nevertheless, the peaks in the computed source field correspond reasonably well with the true source locations.
5. We could do better with more sensors, but we want to restrict the number due to practical considerations: In a real three-dimensional building, we will not be able to have a high concentration of sensors with complete freedom of placement. On the other hand, the sources cannot be placed arbitrarily either.
6. The reconstructed concentration field is much better than the source field; this implies that we have an ill-conditioned problem.
7. We also ran these problems with 10% and 15% error with progressively worse, but not horrible, results.

5.3. Coarse meshes. In the results reported above, we showed that we were, indeed, able to reconstruct the source or sources with acceptable accuracy. In this section, we consider the issue of doing this rapidly. In particular, we discuss the possibility of using a coarser mesh to reconstruct the source field than was necessary to compute the flow field. Obviously, if a coarser meshes suffices, the time to solve the required linear systems will decrease.

Two possibilities present themselves for creating a coarser mesh. As noted above, we solved for the flow field on a fine, uniform mesh that was necessary to achieve convergence for the Reynolds number we desired. Thus, the first way to generate a coarser mesh is simply to regenerate the mesh with a larger value of the parameter that controls the size of the elements. Having done this, we can then interpolate the flow field from the solution on the fine mesh onto this coarser mesh and proceed with the source inversion. The second way to proceed is to try to generate a coarser mesh that is adapted to the flow field itself so that larger elements will appear in regions where the flow field is not changing rapidly and smaller elements where the flow field is rapidly varying. Fortunately, the meshing tool that we are using allows us to do this relatively easily. A few words about this tool are in order.

We have installed and used the bidirectional anisotropic mesh generator (BAMG) that is being developed at INRIA (see [16]). This code allows the generation of meshes over a two-dimensional domain by specifying a number of parameters. These parameters control properties such as maximum and minimun edge length, the maximum number of triangles, etc., to generate uniform meshes. More interestingly, it also allows the specification of a "metric" field and will attempt to adapt the mesh to that field. Thus, if we compute the flow field on a fine mesh and use this as the metric field, BAMG will produce a mesh that is adapted to this field. The other parameters are also used so that by specifying the minimum and maximum edge lengths along with this metric file, BAMG will produce a coarser mesh that is adapted to the flow field. BAMG will then compute a Lagrangian interpolation of the fine flow-field values onto the coarse meshes. BAMG is currently being extended to three-dimensions.

Using BAMG we were easily able to generate both the uniform and adapted coarse meshes necessary to run our source-inversion method on these problems. We were (pleasantly) surprised by the amount of coarsening that was possible without noticing

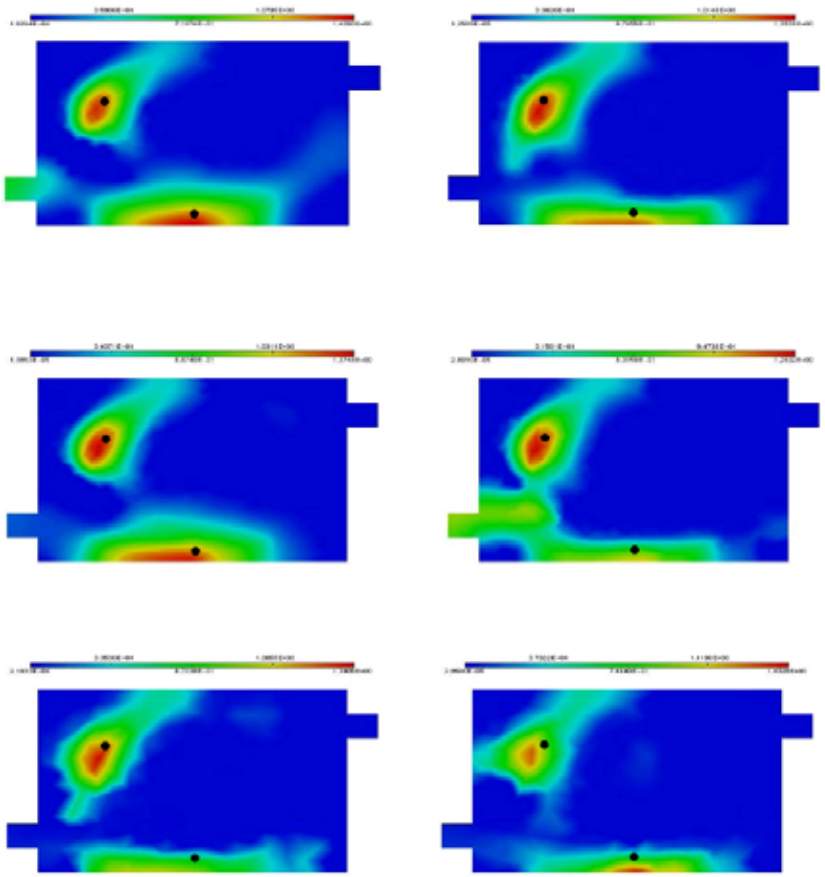


FIG. 3. *Source inversion with adaptive (left) and uniform (right) meshes. Top row,* 1471 *and* 1488 *triangles; middle row,* 1172 *and* 1196 *triangles; bottom row,* 474 *and* 476 *triangles. The true source locations are the points* $(10.0, 1.0)$ *and* $(4.5, 10.0)$.

any degradation in the accuracy of the reconstructed sources. Indeed, we could not see any difference in the predicted source locations by reducing the number of triangles from the original 31,000 to approximately 1500, a factor of over 20! Recall that a grid of 1500 triangles would not be nearly fine enough to compute the flow field. Some representative results are shown in Figures 3–7.

These figures show the results of using ≈ 1500 triangles, ≈ 1200 triangles, and ≈ 500 triangles for both adaptive and uniform meshes. The cases pictured generally show situations where there were some problems. In many of the cases not pictured here, we were able to obtain acceptable results to the lowest number of triangles. Overall, our experience suggests that the adaptive meshes yield better results for the coarsest meshes. There were certainly cases in which the adaptive meshes did not perform better. For example, Figure 5 shows that the adaptive mesh predicts two sources at the coarsest level, whereas the uniform mesh correctly predicts only one.

For some source configurations, it was not possible to get acceptable predictions at the coarsest levels for either mesh strategy; see, e.g., Figure 7. In this figure, we note that the adaptive mesh was able to achieve acceptable results using approximately

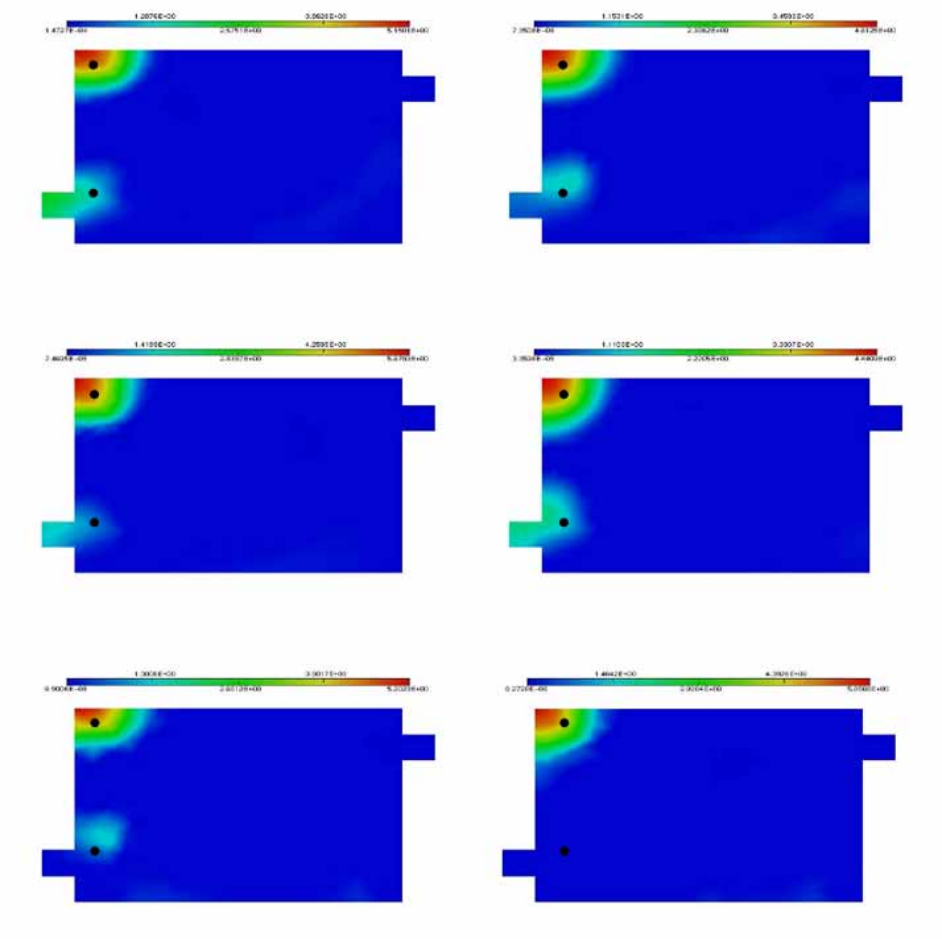


FIG. 4. *Source inversion with adaptive (left) and uniform (right) meshes. Top row,* 1471 *and* 1488 *triangles; middle row,* 1172 *and* 1196 *triangles; bottom row,* 474 *and* 476 *triangles. The true source locations are the points* $(1.0, 14.0)$ *and* $(1.0, 4.0)$.

1200 triangles, whereas the uniform mesh incorrectly predicts three sources. In Figure 6 we also see that the adaptive mesh performs better at the coarser levels.

We observe that sources placed in the main flow are much harder to locate due to the rapid dispersal of the agent in this region of the flow field. It is also true that the predicted strength in these regions is also much less than the true value. Again, we think that that the use of the Tikhonov regularization accounts for some of this.

Finally, we point out that the times required for doing the inversion are, as expected, substantially reduced by using these coarser meshes. Table 1 gives the the number of O3D iterations necessary to achieve the required accuracy and the number of seconds it took to run. Clearly the reduction of run-times by a factor between 40 and 100 will be significant.

6. Discussion and conclusions. In this paper we have developed a formulation for the source-inversion problem that we have shown to be effective for locating sources in a steady-state environment. Although we have not stressed computational efficiency in our numerical experiments, we have been able to solve these problems relatively

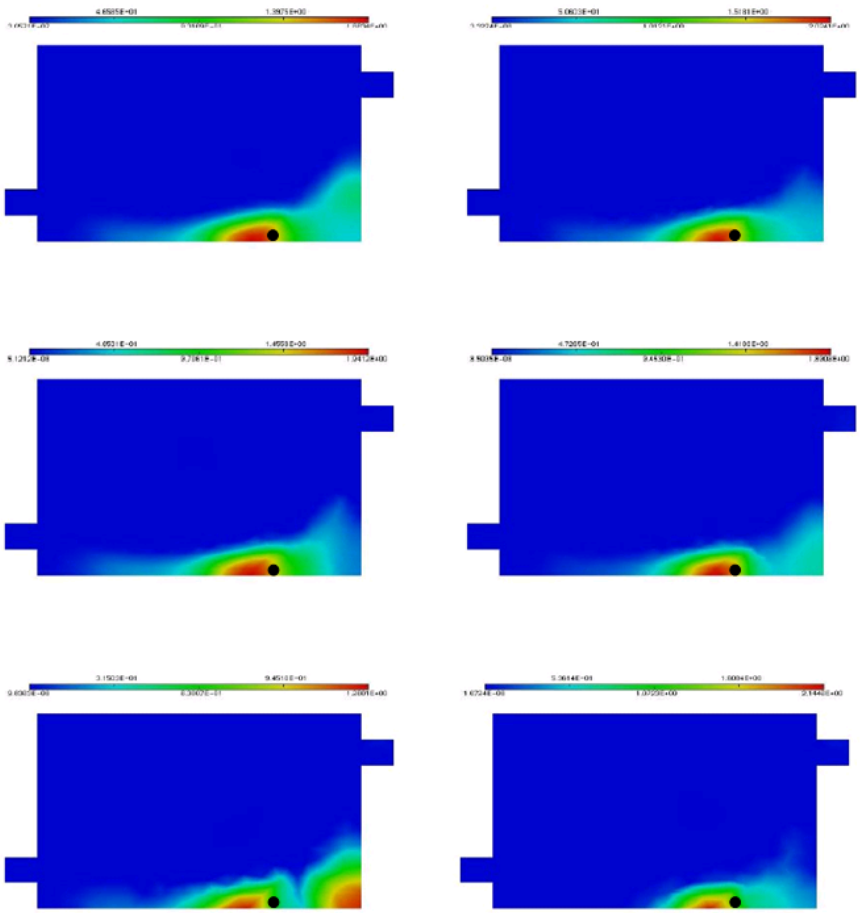


FIG. 5. *Source inversion with adaptive (left) and uniform (right) meshes. Top row,* 1471 *and* 1488 *triangles; middle row,* 1172 *and* 1196 *triangles; bottom row,* 474 *and* 476 *triangles. The true source location is the point* $(14.5, 0.5)$.

quickly, in terms of the number of iterations of O3D. Much work remains, however, to improve the efficiency of the linear solvers, the main source of computational costs in our runs. We have also shown that Sundance and Trilinos are powerful tools for the rapid prototyping and testing of various ideas and strategies.

In addition to continuing to work on the linear solver strategies, both in serial and in parallel, we need to investigate several other topics as follows:

1. As noted above, problem (4.14) uses a standard Tikhonov regularization scheme. Our computational results indicate that this tends to smear the source. Other possible regularizations can be considered, as in [1].
2. The problem of the optimal location of sensors is difficult. It is not at all clear what the objective should be. Our understanding is that different toxins dictate different attack strategies. Thus, we must take that into account when determining the location of the sensors. We are pursuing ideas related to hierarchical control (see, e.g., [8]); this will be the subject of a future paper. We are also considering ideas related to the mesh strategies noted

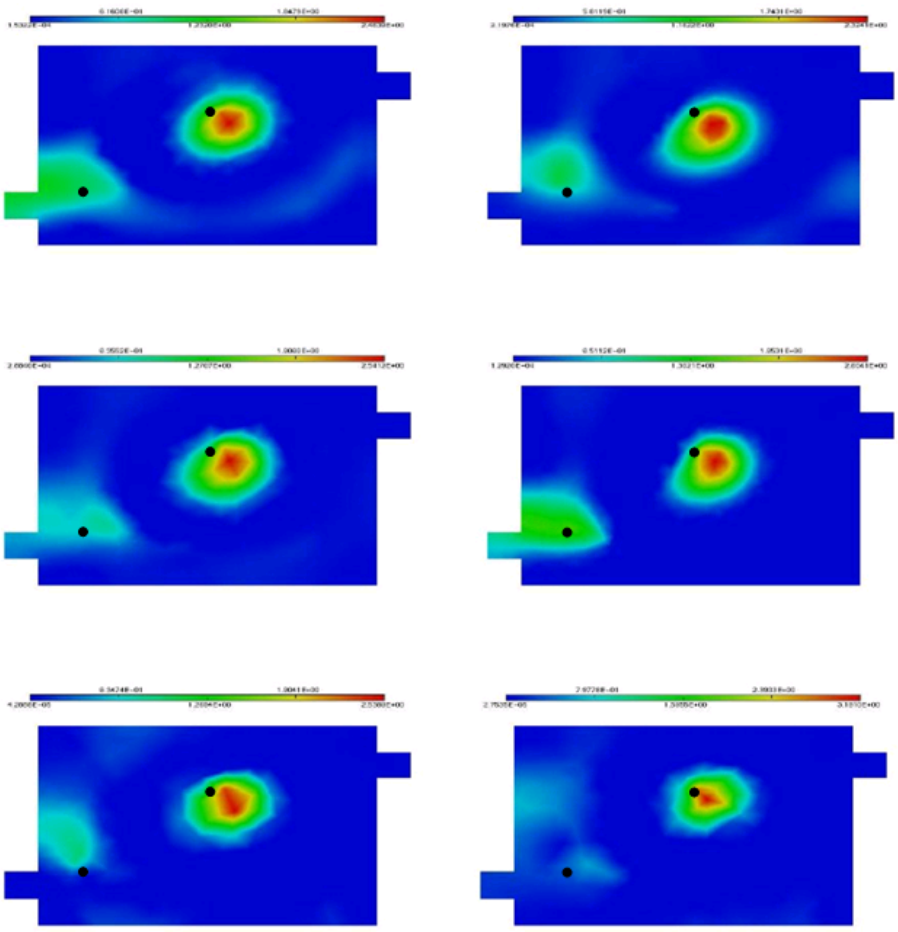


FIG. 6. *Source inversion with adaptive (left) and uniform (right) meshes. Top row,* 1471 *and* 1488 *triangles; middle row,* 1172 *and* 1196 *triangles; bottom row,* 474 *and* 476 *triangles. The true source locations are the points* $(3.0, 4.0)$ *and* $(10.0, 8.0)$.

above, where sensors could be located in regions of rapid change in the flow field.

We note in closing that PDE-constrained optimization problems arise in the context of improving the sensors themselves. We have been studying the problem of improving the shape and the topology of the channels in a microfluidic sensor. Here we want to improve certain aspects of the movement of the contaminant within the sensor subject to the flow within the channels.

In part 2 of this paper we will consider the time-dependent source-inversion problem. Although the problem will be much larger, we will also have more information. That is, we will have the time history of the transport, or, put another way, each sensor will provide much more that one measurement. In our preliminary analysis of the time-dependent problem, we have observed that the linear systems that we will have to solve will have a structure that can be exploited by a direct solver. In addition, we will consider many of the ideas enumerated above concerning the mesh and the regularization strategy in the time-dependent case.

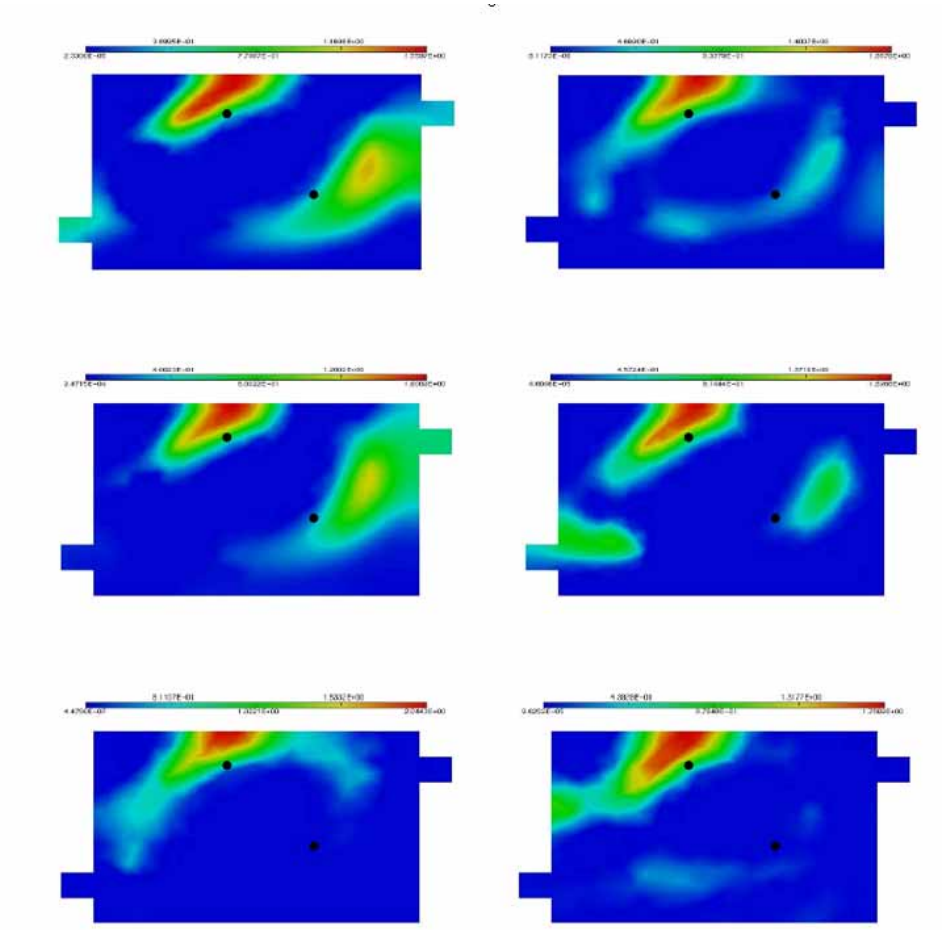


FIG. 7. *Source inversion with adaptive (left) and uniform (right) meshes. Top row,* 1471 *and* 1488 *triangles; middle row,* 1172 *and* 1196 *triangles; bottom row,* 474 *and* 476 *triangles. The true source locations are the points* $(6.0, 12.0)$ *and* $(16.0, 6.0)$.

TABLE 1
Table giving run-times and the number of O3D *iterations at various levels of mesh coarsening.*

Approximate no. of triangles	Uniform mesh		Adaptive mesh	
	O3D iterations	Time	O3D iterations	Time
31,000	6.27	385.77	–	–
1500	7.33	9.21	5.78	8.52
1200	6.88	7.13	5.84	6.58
500	7.78	3.00	8.09	3.05

Appendix. The regularized eikonal equation. We motivate and derive the weak form of the regularized eikonal equation used in the main body of the paper as follows. First, we illustrate ideas by considering a couple of (essentially) one-dimensional analytical examples; then, a weak form of the fully multidimensional problem is derived; finally, an iterative procedure is specified and illustrated using the flow geometry prescribed in section 5.

We begin by considering the one-dimensional eikonal equation

$$\left(\frac{d\ell^*}{dx}\right)^2 = 1\,, \tag{A.1}$$

subject to $\ell^* = 0$ at the boundaries $x = 0$ and $x = 1$. Solutions to (A.1) are given by $\ell^* = \pm x$, which cannot solve the boundary conditions without introducing discontinuities in either ℓ^* or its derivative at $x = 1/2$. For example, the generalized function

$$\ell^* = \begin{cases} x, & 0 \le x \le 1/2, \\ 1-x, & 1/2 \le x \le 1, \end{cases} \tag{A.2}$$

which clearly satisfies (A.1) except at $x = 1/2$, measures the distance from the nearest boundary point ($x = 0$ or $1/2$). For stable solution by finite elements in an enclosed region, the eikonal equation must be regularized. If we add a regularization term $-\epsilon d^2\ell^*/dx^2$, where ϵ is a small positive quantity, to the left-hand side of (A.1), the problem becomes

$$\left(\frac{d\ell^*}{dx}\right)^2 - \epsilon\frac{d^2\ell^*}{dx^2} = 1\,, \quad \ell^*(0) = \ell^*(1) = 0\,. \tag{A.3}$$

The general solution of the differential equation for ℓ^* is obtained by first solving the first order equation for $d\ell^*/dx$ and then performing a second integration to give $\ell^* = d - \epsilon \ln\cosh[(x-c)/\epsilon]$, where c and d are constants of integration. Applying the boundary conditions then determines ℓ^* uniquely as

$$\ell^*(x) = \epsilon \ln\left\{\frac{\cosh\left[-1/(2\epsilon)\right]}{\cosh\left[(2x-1)/(2\epsilon)\right]}\right\} = -\epsilon\ln\left[\frac{e^{(2x-1)/(2\epsilon)} + e^{-(2x-1)/(2\epsilon)}}{e^{-1/(2\epsilon)} + e^{1/(2\epsilon)}}\right]. \tag{A.4}$$

It is readily deduced that for $0 < \epsilon \ll 1$, ℓ^* has the asymptotic behavior $\ell^* \sim x$ for $0 \le x < 1/2$ and $\ell^* \sim 1-x$ for $1/2 < x \le 1$. In the neighborhood of $x = 1/2$, we define the stretched coordinate $\eta = \epsilon^{-1}(x - 1/2)$, in terms of which $\ell^* \sim 1/2 - \epsilon\ln[2\cosh\eta]$. In the asymptotic context, a composite approximation spanning both the outer regions ($0 \le x < 1/2$ and $1/2 < x \le 1$) and the inner zone $|x - 1/2| \sim \mathrm{O}(\epsilon)$ is in fact given by the inner approximation written in terms of x. Hence, an asymptotic approximation for (A.4), valid in the limit of small ϵ, is given by

$$\ell^*(x) \sim \frac{1}{2} - \epsilon\ln\left[2\cosh\left(\frac{2x-1}{2\epsilon}\right)\right] = \frac{1}{2} - \epsilon\ln\left[e^{(2x-1)/(2\epsilon)} + e^{-(2x-1)/(2\epsilon)}\right]. \tag{A.5}$$

We note that the maximum deviation in ℓ^* from $\sup\{x, 1-x\}$ occurs at $x = 1/2$, where $\ell^*(1/2) \sim 1/2 - \epsilon\ln 2$. The exact solution (A.4) and its approximation (A.5) are illustrated in Figure 8.

The straightforward multidimensional generalization of (A.3) used in the present work is given, on a domain Ω, by

$$\left(\overrightarrow{\nabla}\ell^*\right)\cdot\left(\overrightarrow{\nabla}\ell^*\right) - \epsilon\nabla^2\ell^* = 1\,, \quad \ell^* = 0 \text{ on } \partial\Omega_w\,, \quad \overrightarrow{\nabla}\ell^* = 0 \text{ on } \partial\Omega_{I,o}\,. \tag{A.6}$$

The solution of this problem yields a tent-like structure, with the aforementioned distance-representation property for ℓ^*, over the domain Ω. For example, if the domain is the rectangle $0 \le x \le 1$, $0 \le y \le L$, with $\ell^* = 0$ on $x = 0$ and $x = 1$, and

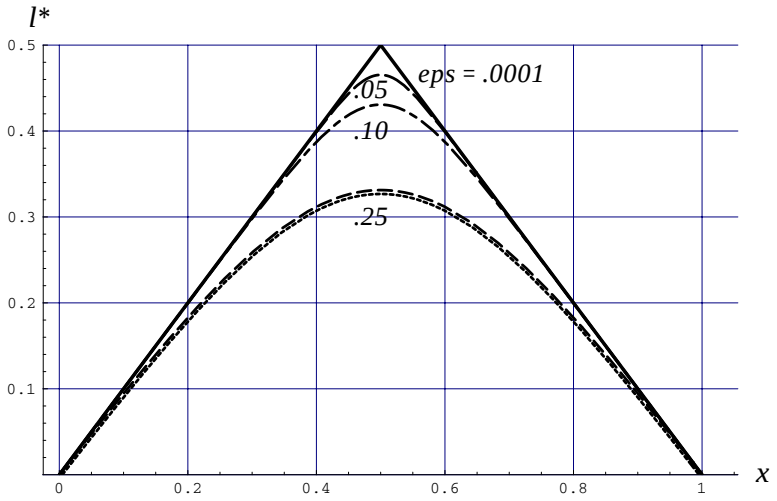


FIG. 8. *Solution of* (A.3) *for several values of* ϵ *("eps"). The two curves (dashed and dotted) for* $\epsilon = .25$ *correspond to the exact solution* (A.4) *and its asymptotic representation* (A.5)*, respectively; the two expressions are virtually indistinguishable for smaller values of* ϵ*.*

$\partial \ell^*/\partial y = 0$ on $y = 0$ and $y = L$, the solution is simply (A.4), independent of y. A more nontrivial example is the solution of (A.6) in polar coordinates (r, ϑ) when $\partial\Omega_w$ is the ring $0 < r_i \le r \le r_0$ and the boundary conditions are $\ell^*(r_i) = \ell^*(r_0) = 0$. Due to angular symmetry, (A.6) reduces to solving $(d\ell^*/dr)^2 - \epsilon r^{-1}(d/dr)(r\, d\ell^*/dr) = 1$, ultimately giving the final result

$$\ell^* = \epsilon \ln \left\{ \frac{I_0(r_i/\epsilon)K_0(r_0/\epsilon) - I_0(r_0/\epsilon)K_0(r_i/\epsilon)}{I_0(r/\epsilon)\big[K_0(r_0/\epsilon) - K_0(r_i/\epsilon)\big] - K_0(r/\epsilon)\big[I_0(r_0/\epsilon) - I_0(r_i/\epsilon)\big]} \right\}, \tag{A.7}$$

where I_0 and K_0 are the zero order modified Bessel functions. This solution is illustrated in Figure 9; its physical appearance in three dimensions is obtained by rotating the profiles about the ℓ^*-axis to yield the upper part of a torus-like structure.

As with the Navier–Stokes equations, the use of Sundance requires a weak form of the eikonal problem (A.6). We begin by writing

$$\int_\Omega r \left[\big(\overrightarrow{\nabla}\ell^*\big) \cdot \big(\overrightarrow{\nabla}\ell^*\big) - \epsilon \nabla^2 \ell^* - 1 \right] d\Omega = 0\,, \tag{A.8}$$

where r is the test function. Use of the identity $r\nabla^2\ell^* = \overrightarrow{\nabla} \cdot \big(r\overrightarrow{\nabla}\ell^*\big) - \big(\overrightarrow{\nabla} r\big) \cdot \big(\overrightarrow{\nabla}\ell^*\big)$ and application of the divergence theorem allows (A.8) to be expressed in terms of first derivatives as

$$\begin{aligned} \int_\Omega \Big\{ r \Big[\big(\overrightarrow{\nabla}\ell^*\big) \cdot \big(\overrightarrow{\nabla}\ell^*\big) - 1 \Big] + \epsilon \big(\overrightarrow{\nabla} r\big) \cdot \big(\overrightarrow{\nabla}\ell^*\big) \Big\}\, d\Omega \\ - \epsilon \int_{\partial\Omega} r \big(\overrightarrow{\nabla}\ell^*\big) \cdot \overrightarrow{n}\, d(\partial\Omega) = 0\,, \end{aligned} \tag{A.9}$$

where $\overrightarrow{n}$ is the unit normal to $\partial\Omega$.

FIG. 9. *Solution of* (A.7) *for* $r_i = 1/2$, $r_0 = 1$, *and several values of* ϵ (*"eps"*).

Based on the boundary conditions expressed in (A.6), the surface integral vanishes on the inlet and outlet portions of $\partial\Omega$. Approximating the wall boundary condition by

$$-\big(\overrightarrow{\nabla}\ell^*\big)\cdot\overrightarrow{n}\Big|_{\partial\Omega_w} = \hat{\epsilon}^{-1}\ell^*\,, \quad \hat{\epsilon}\ll 1\,, \tag{A.10}$$

and taking the limit $\hat{\epsilon}\to 0$ (independent of the ϵ in (A.9)), we obtain, by the same argument advanced in section 3.2, a replacement for the surface term according to

$$\int_\Omega \Big\{ r\Big[\big(\overrightarrow{\nabla}\ell^*\big)\cdot\big(\overrightarrow{\nabla}\ell^*\big)-1\Big] + \epsilon\big(\overrightarrow{\nabla}r\big)\cdot\big(\overrightarrow{\nabla}\ell^*\big)\Big\}\, d\Omega + \epsilon\int_{\partial\Omega_w} r\ell^*\, d(\partial\Omega) = 0\,. \tag{A.11}$$

Equation (A.11) is thus the final weak form of problem (A.6). We note that (A.11) is decoupled from the larger problem, although the reverse is obviously not true since ℓ^* enters into (3.22) through the mixing length $\ell_{\hat{j}}$ that appears in the expression for the turbulent stress tensor $\underline{\underline{\tau}}$.

In order to actually solve (A.11), which is nonlinear in the unknown function ℓ^*, it is necessary to adopt an iterative approach as discussed in section 3.2. In particular, since we ultimately require a weak form that is linear with respect to the unknown function ℓ^* (as well as the test function r), we formally linearize the nonlinear term $\big(\overrightarrow{\nabla}\ell^*\big)\cdot\big(\overrightarrow{\nabla}\ell^*\big)$ about an initial guess ℓ_0^* by writing $\ell^* = \ell_0^* + \ell_1^*$, where the correction ℓ_1^* is presumed to be small. Consequently,

$$\begin{aligned}
\big(\overrightarrow{\nabla}\ell^*\big)\cdot\big(\overrightarrow{\nabla}\ell^*\big) &= \big(\overrightarrow{\nabla}\ell_0^*\big)\cdot\big(\overrightarrow{\nabla}\ell_0^*\big) + 2\big(\overrightarrow{\nabla}\ell_1^*\big)\cdot\big(\overrightarrow{\nabla}\ell_0^*\big) + \big(\overrightarrow{\nabla}\ell_1^*\big)\cdot\big(\overrightarrow{\nabla}\ell_1^*\big)\\
&\approx \big(\overrightarrow{\nabla}\ell_0^*\big)\cdot\big(\overrightarrow{\nabla}\ell_0^*\big) + 2\big(\overrightarrow{\nabla}\ell_1^*\big)\cdot\big(\overrightarrow{\nabla}\ell_0^*\big)\\
&= \big(\overrightarrow{\nabla}\ell_0^*\big)\cdot\big(\overrightarrow{\nabla}\ell_0^*\big) + 2\big[\overrightarrow{\nabla}(\ell^*-\ell_0^*)\big]\cdot\big(\overrightarrow{\nabla}\ell_0^*\big)\\
&= \big(\overrightarrow{\nabla}\ell_0^*\big)\cdot\big(2\overrightarrow{\nabla}\ell^* - \overrightarrow{\nabla}\ell_0^*\big).
\end{aligned} \tag{A.12}$$

Using this linearized representation for $\big(\overrightarrow{\nabla}\ell^*\big)\cdot\big(\overrightarrow{\nabla}\ell^*\big)$ in (A.11), we thus arrive at a functional iteration scheme that computes the successive approximation ℓ_{i+1}^* in terms

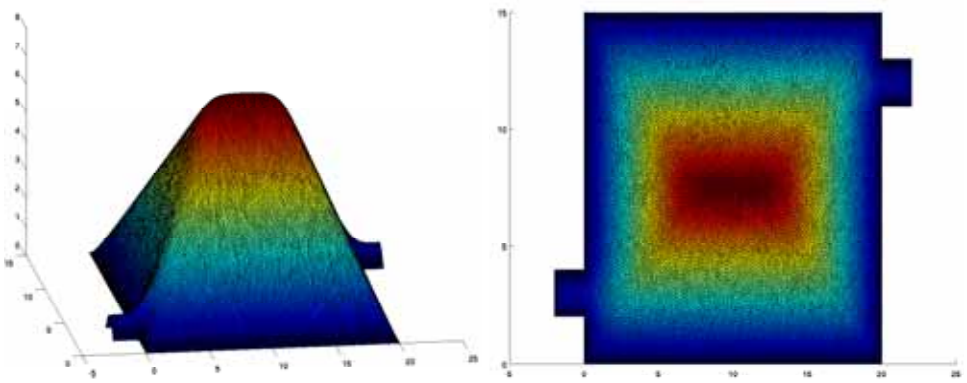


FIG. 10. *Angle and top views of the converged iterative solution of* (A.13) *for* $\epsilon = .05$ *on a sample domain* Ω *with inlet (lower left) and outlet (upper right) boundaries.*

of an initial guess ℓ_i^* according to

$$\int_\Omega \left\{ r\left[\left(\vec{\nabla}\ell_i^*\right)\cdot\left(2\vec{\nabla}\ell_{i+1}^* - \vec{\nabla}\ell_i^*\right) - 1\right] + \epsilon\left(\vec{\nabla}r\right)\cdot\left(\vec{\nabla}\ell_{i+1}^*\right)\right\} d\Omega + \epsilon\int_{\partial\Omega} r\ell_{i+1}^*\, d(\partial\Omega) = 0\,. \tag{A.13}$$

Equation (A.13) is linear with respect to the unknown function ℓ_{i+1}^* and can be handled directly by Sundance. The sequence of approximations is expected to be convergent given even a crude starting guess ℓ_0^* for sufficiently large ϵ, since in that limit the original nonlinearity becomes a perturbation of an otherwise linear problem. An outer iteration scheme can then be used to generate sufficiently good starting guesses ℓ_0^* for the recursive algorithm (A.13) with successively smaller values of ϵ. An example of such a calculation is illustrated in Figure 10.

REFERENCES

[1] V. Akçelik, G. Biros, O. Ghattas, K. Long, and B. van Bloemen Waanders, *A variational finite element method for source inversion for convective-diffusive transport*, Finite Elem. Anal. Des., 39 (2003), pp. 683–705.

[2] L. T. Biegler, O. Ghattas, M. Heinkenschloss, and B. van Bloemen Waanders, eds., *Large-Scale PDE-Constrained Optimization*, Lecture Notes in Comput. Sci. Eng. 30, Springer-Verlag, Berlin, 2003.

[3] R. B. Bird, W. E. Stewart, and E. N. Lightfoot, *Transport Phenomena*, John Wiley & Sons, New York, 1960.

[4] G. Biros and O. Ghattas, *Parallel Lagrange–Newton–Krylov–Shur methods for PDE-constrained optimization. Part* I: *The Krylov–Shur solver*, SIAM J. Sci. Comput., 27 (2005), pp. 687–713.

[5] P. T. Boggs, P. D. Domich, and J. E. Rogers, *An interior-point method for general large scale quadratic programming problems*, Ann. Oper. Res., 62 (1996), pp. 419–437.

[6] P. T. Boggs, A. J. Kearsley, and J. W. Tolle, *A global convergence analysis of an algorithm for large-scale nonlinear optimization problems*, SIAM J. Optim., 9 (1999), pp. 833–862.

[7] P. T. Boggs, A. J. Kearsley, and J. W. Tolle, *A practical algorithm for general large scale nonlinear optimization problems*, SIAM J. Optim., 9 (1999), pp. 755–778.

[8] A. J. Kearsley, P. T. Boggs, and J. W. Tolle, *Hierarchical control of a linear diffusion equation*, Large-scale PDE-Constrained Optimization, L. T. Biegle, et al., eds., Lecture Notes in Comput. Sci. Eng. 30, Springer, Berlin, 2003, pp. 236–249.

[9] M. Born and E. Wolf, *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light*, Cambridge University Press, Cambridge, UK, 1999.
[10] J. Boussinesq, *Essai sur la théorie des eaux courantes*, Memoirs Acad. des Sciences, 23 (1872).
[11] T. Cebeci and J. Cousteix, *Modeling and Computation of Boundary-Layer Flows*, Horizons Publishing/Springer-Verlag, Long Beach/Berlin, 1999.
[12] E. Cummings, *private communication*, 2003.
[13] E. B. Cummings, S. K. Griffiths, and R. H. Nilson, *Applied Microfluidic Physics LDRD Final Report*, Tech. Report SAND2002-8018, Sandia National Laboratories, Livermore, CA, 2002.
[14] T. B. Gatski, *Turbulent flows: Model equations and solution methodology*, in Handbook of Computational Fluid Mechanics, R. Peyret, ed., Academic Press, San Diego, CA, 1996, pp. 339–415.
[15] P. M. Gresho and R. L. Sani, *Incompressible Flow and the Finite Element Method*, John Wiley & Sons, New York, 2000.
[16] F. Hecht, *BAMG: Bidimensional Anisotropic Mesh Generator*, INRIA, Le Chesnay, France, 1998. Available online at http://www-rocq1.inria.fr/gamma/cdrom/www/bamg/eng.htm.
[17] M. Heroux, R. Bartlett, V. Howle, R. Heokstra, J. Hu, T. Kolda, R. Lehoucq, K. Long, R. Pawlowski, E. Phipps, A. Salinger, H. Thornquist, R. Tuminaro, J. Willenbring, and A. Williams, *An Overview of Trilinos*, Tech. Report SAND2002-2729, Sandia National Laboratories, Livermore, CA, 2003.
[18] K. R. Long, *Sundance rapid prototyping tool for parallel pde optimization*, in Large-Scale PDE-Constrained Optimization, L. Biegler, et al., eds., Lecture Notes in Comput. Sci. Eng. 30, Springer, Berlin, 2003, pp. 331–341.
[19] K. R. Long, *Sundance User's Manual*, Tech. Report SAND2004-4793, Computer Science and Mathematics Research Department, Sandia National Laboratories, Livermore, CA, 2004.
[20] R. Michel, C. Quémard, and R. Durant, *Application d'un schéma de longueur de mélange à l'études des couches limites d'équilibre*, ONERA Note Technique 154, ONERA, 1969.
[21] S. B. Pope, *Turbulent Flows*, Cambridge University Press, Cambridge, UK, 2000.
[22] L. Prandtl, *Bericht über untersuchungen zur ausgebildeten turbulenz*, ZAMM. Z. Angew. Math. Mech., 5 (1925), pp. 136–139.
[23] H. Schlichting and K. Gersten, *Boundary Layer Theory*, Springer-Verlag, Berlin, 2000.
[24] J. R. Shewchuk, *Triangle: Engineering a* $2d$ *quality mesh generator and delaunay triangulator*, in Proceedings of the First Workshop on Applied Computational Geometry: Towards Geometric Engineering, M. C. Lin and D. Manocha, eds., Lecture Notes in Comput. Sci. 1148, Springer-Verlag, 1996, pp. 203–222.
[25] J. R. Shewchuk, *Delaunay refinement algorithms for triangular mesh generation*, Comput. Geom. Theory Appl., 22 (2002), pp. 21–74.
[26] B. van Bloemen Waanders, R. Bartlett, K. Long, P. Boggs, and A. Salinger, *Large Scale Non-linear Programming for PDE Constrained Optimization*, Tech. Report SAND2002-3198, Sandia National Laboratories, Livermore, CA, 2002.
[27] D. C. Wilcox, *Turbulence Modeling for CFD*, DCW Industries, La Cañada, CA, 2000.

SIAM J. OPTIM.
Vol. 17, No. 2, pp. 459–484
© 2006 Society for Industrial and Applied Mathematics

AN ACTIVE SET METHOD FOR SINGLE-CONE SECOND-ORDER CONE PROGRAMS*

E. ERDOĞAN† AND G. IYENGAR†

Abstract. We develop an active set method for solving second-order cone programs that may have an arbitrary number of linear constraints but are restricted to having only one second-order cone constraint. Problems of this form arise in the context of robust optimization and trust region methods. The proposed active set method exploits the fact that a second-order cone program with only one second-order cone constraint and no inequality constraints can be solved in closed form.

Key words. second-order cone programs, single-cone, active set method, robust optimization, uncertain linear program, duality

AMS subject classifications. 90C25, 90C30, 90C20, 49N15

DOI. 10.1137/040612592

1. Introduction.

In this paper we are concerned with the following special case of a second-order cone program (SOCP):

$$\begin{array}{rl} \min & \mathbf{f}^T\mathbf{x} \\ \text{subject to} & \mathbf{H}\mathbf{x} = \mathbf{g}, \\ & \mathbf{E}\mathbf{x} \geq \mathbf{0}, \\ & \mathbf{D}\mathbf{x} \succeq \mathbf{0}, \end{array} \tag{1.1}$$

where $\mathbf{x} \in \mathbf{R}^n$, $\mathbf{f} \in \mathbf{R}^n$, $\mathbf{H} \in \mathbf{R}^{m\times n}$, $\mathbf{g} \in \mathbf{R}^m$, $\mathbf{E} \in \mathbf{R}^{l\times n}$, $\mathbf{D} \in \mathbf{R}^{p\times n}$, and $\succeq$ denotes the partial order with respect to the standard conic quadratic cone $\mathcal{Q} = \{(z_0, \bar{\mathbf{z}})^T \in \mathbf{R}^p : z_0 \geq \sqrt{\bar{\mathbf{z}}^T\bar{\mathbf{z}}}\} \subset \mathbf{R}^p$. We shall call the optimization problem (1.1) a single-cone SOCP since it is restricted to having only *one* second-order cone constraint.

Our interest in single-cone SOCPs stems from the fact that they arise as the robust counterpart of uncertain linear programs (LPs). Many decision problems in engineering and operations research can be formulated as LPs of the form

$$\begin{array}{rl} \min & \mathbf{c}^T\mathbf{x} \\ \text{subject to} & \mathbf{A}\mathbf{x} = \mathbf{b}, \\ & \mathbf{x} \geq \mathbf{0}. \end{array}$$

Solution techniques for LPs compute a solution assuming that the parameters $(\mathbf{A}, \mathbf{b}, \mathbf{c})$ are known exactly. However, in practice, these parameters are typically the result of some measurement or estimation process and are therefore never certain. LPs whose parameters are not known exactly are called *uncertain* LPs. Several strategies have been proposed to address parameter uncertainty in optimization problems. One approach is to solve the LP for a nominal set of parameters $(\mathbf{A}_0, \mathbf{b}_0, \mathbf{c}_0)$ and then analyze the quality of the solution using a postoptimization tool such as sensitivity

*Received by the editors July 30, 2004; accepted for publication (in revised form) January 5, 2006; published electronically July 17, 2006.
http://www.siam.org/journals/siopt/17-2/61259.html
†IEOR Department, Columbia University, New York, NY 10027 (ee168@columbia.edu, garud@ieor.columbia.edu). The research of the first author was partially supported by NSF grant CCR-00-09972. The research of the second author was partially supported by NSF grants CCR-00-09972 and DMS-01-04282, DOE grant DE-FG02-92ER25216, and ONR grant N000140310514.

analysis [5]. This approach is particularly attractive when the uncertainty is "small" in an appropriate sense. In the stochastic programming approach, the uncertainty is assumed to be random with a known distribution, and samples from this known distribution are used to compute good solutions [16]. However, identifying appropriate distributions for the parameters is not straightforward. Also, as the dimension of the problem grows, the complexity of the stochastic program quickly becomes prohibitive. Recently Ben-Tal and Nemirovski [2, 3, 4] proposed *robust* optimization as another approach to address data uncertainty. In this approach, the uncertain parameters $(\mathbf{A}, \mathbf{b}, \mathbf{c})$ are assumed to belong to a bounded uncertainty set $\mathcal{U}$ and the goal of the *robust* counterpart is to compute a minimax optimal solution. The results in [2, 3, 4] establish that when $\mathcal{U}$ satisfies some regularity properties, the robust counterpart can be reformulated as an SOCP and therefore can be solved efficiently both in theory [14] and in practice [17].

The robust counterpart of an uncertain LP where the parameters $(\mathbf{A}, \mathbf{b})$ are completely known and the uncertain cost vector $\mathbf{c}$ belongs to an ellipsoidal uncertainty set can be reformulated as a single-cone SOCP [3] (see also section 5). In many engineering applications the constraints in the LP are given by design considerations and are therefore fixed and certain. For example, in routing problems arising in the context of road or air traffic control and communication networks, the capacities are determined at the network design stage; therefore, the constraints in the problem, namely, the flow balance equations and capacity constraints, are completely known when the routing problem is to be solved. However, the "cost" of an arc is typically a nonlinear function of the capacity and flows in the network, and measuring this cost is often difficult and expensive [8]. The "cost" of a feasible flow can often be modeled as an uncertain linear function with an ellipsoidal uncertainty set by using the so-called delta method [12]. Production planning is another natural example where the constraints are fixed and only the costs are uncertain. Here the vector $\mathbf{c}$ denotes the vector of future expected market prices for the various raw materials and is, typically, estimated from historical prices via linear regression. Since the confidence regions associated with linear regression are ellipsoidal [10, 12], the resulting robust counterpart is a single-cone SOCP.

From the equivalence

$$\|\mathbf{Pu}\| \leq 1 \quad \Leftrightarrow \quad u_0 = 1, \ \begin{bmatrix} u_0 \\ \mathbf{Pu} \end{bmatrix} \succeq \mathbf{0},$$

it follows that the trust region problem is a special case of a single-cone SOCP. This provides another motivation for developing active set methods for single-cone SOCPs. Note that formulating the trust region problem as a single-cone SOCP allows one to consider hyperbolic and parabolic trust regions.

Alizadeh and Goldfarb [1] showed that under appropriate regularity conditions, the optimal solution of a single-cone SOCP with no inequality constraints can be computed in closed form. We use this result to explicitly compute the value of the Lagrangian obtained by dualizing the nonnegativity constraints $\mathbf{Ex} \geq \mathbf{0}$. We compute the optimal dual multipliers using an active set method, and then recover an optimal primal solution using the results in [1]. The formulation of the appropriate dual problem is discussed in section 2, the active set method is detailed in section 3, and section 4 details how to recover an optimal solution of (1.1).

Clearly, any algorithm for solving general SOCPs can be used to solve a single-cone SOCP. All the known codes for solving general SOCPs, e.g., SeDuMi [17] and

MOSEK, are based on interior point methods. Our efforts in developing an active set method for the single-cone SOCP were motivated, in part, by the observation that active set methods are known to solve convex quadratic programs efficiently. Our goal was to investigate whether a simple active set algorithm outperforms general purpose SOCP codes at least for certain problem classes. We report the results of our computational experiments in section 5.

Finally, Muramatsu [13] proposed a simplex-type algorithm for solving a special case of the single-cone SOCP (1.1). His algorithm is a generalization of the simplex method used for solving LPs [7], which is different than our approach.

2. Formulation of the Lagrangian dual. In this section we formulate a Lagrangian dual for the single-cone SOCP (1.1). We assume that $\mathbf{H} \in \mathbf{R}^{m\times n}$ has full row rank and the following constraint qualification holds.

ASSUMPTION 2.1. *There exists* $\bar{\mathbf{x}} \in \mathbf{R}^n$ *such that* $\mathbf{H}\bar{\mathbf{x}} = \mathbf{g}$, $\mathbf{E}\bar{\mathbf{x}} \succeq \mathbf{0}$, *and* $\mathbf{D}\bar{\mathbf{x}} \succ \mathbf{0}$.

The active set algorithm proposed in this paper exploits the following result from [1].

LEMMA 2.2. *Suppose the pair of primal-dual SOCPs*

$$\begin{array}{rlrl} \min & \mathbf{c}^T\mathbf{x} & \max & \mathbf{b}^T\mathbf{y} \\ \textit{subject to} & \mathbf{A}\mathbf{x} = \mathbf{b}, & \textit{subject to} & \mathbf{A}^T\mathbf{y} + \mathbf{z} = \mathbf{c}, \\ & \mathbf{x} \succeq \mathbf{0}, & & \mathbf{z} \succeq \mathbf{0} \end{array} \tag{2.1}$$

are both *strictly feasible. Then the optimal solution of the primal SOCP is given by*

$$\mathbf{x}^* = \left(\sqrt{\frac{-\mathbf{b}^T(\mathbf{A}\mathbf{R}\mathbf{A}^T)^{-1}\mathbf{b}}{\mathbf{c}^T\mathbf{P}_R\mathbf{c}}} \right) \mathbf{P}_R\mathbf{c} + \mathbf{R}\mathbf{A}^T(\mathbf{A}\mathbf{R}\mathbf{A}^T)^{-1}\mathbf{b}, \tag{2.2}$$

where

$$\mathbf{R} = \begin{bmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & -\mathbf{I} \end{bmatrix}, \quad \mathbf{P}_R = \mathbf{R} - \mathbf{R}\mathbf{A}^T(\mathbf{A}\mathbf{R}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{R},$$

and $\mathbf{I}$ *denotes an identity matrix.*

REMARK 2.3. *In Lemma* 2.2 *we have implicitly assumed that* $\mathbf{A}\mathbf{R}\mathbf{A}^T$ *is nonsingular. A similar result holds when* $\mathbf{A}\mathbf{R}\mathbf{A}^T$ *is singular. See* [1] *for details.*

In order to reformulate (1.1) into a form similar to the primal SOCP in (2.1), we dualize the nonnegativity constraints to obtain the Lagrangian

$$\begin{array}{rlrl} q(\boldsymbol{\lambda}) \equiv & & \min & (\mathbf{f} - \mathbf{E}^T\boldsymbol{\lambda})^T\mathbf{x} \\ & & \text{subject to} & \mathbf{H}\mathbf{x} = \mathbf{g}, \\ & & & \mathbf{D}\mathbf{x} \succeq \mathbf{0}, \end{array} \tag{2.3}$$

where $\boldsymbol{\lambda} \in \mathbf{R}^l_+$ denotes the Lagrange multipliers for the inequality constraints. Note that the result in [1] applies only when the primal and the dual SOCPs are *both* strictly feasible. For SOCPs, feasibility is a subtle issue; e.g., the fact that the primal is bounded does not imply that the dual is feasible [4]; therefore, one has to be careful in applying the results in [1]. Elementary properties of convex duality [6] implies the following claim.

CLAIM 2.4. *Let* $q(\boldsymbol{\lambda})$ *denote the Lagrangian defined in* (2.3). *Let* $\mathbf{x}^*$ *and* v^* *denote, respectively, any optimal solution and the optimum value of* (1.1). *Then*

(a) $v^* = \max\{q(\boldsymbol{\lambda}) : \boldsymbol{\lambda} \geq \mathbf{0},\ \boldsymbol{\lambda} \in \mathcal{D}_q\}$, *where* $\mathcal{D}_q = \{\boldsymbol{\lambda} : q(\boldsymbol{\lambda}) > -\infty\}$,
(b) $\mathbf{x}^* \in \operatorname{argmin}\{(\mathbf{f} - \mathbf{E}^T\boldsymbol{\lambda}^*)^T\mathbf{x} : \mathbf{H}\mathbf{x} = \mathbf{g},\ \mathbf{D}\mathbf{x} \succeq \mathbf{0}\}$, *where* $\boldsymbol{\lambda}^* \in \operatorname{argmax}\{q(\boldsymbol{\lambda}) : \boldsymbol{\lambda} \geq 0,\ \boldsymbol{\lambda} \in \mathcal{D}_q\}$.

Thus, an optimal solution to (1.1) can be obtained by first computing an optimal multiplier $\boldsymbol{\lambda}^* \in \operatorname{argmax}\{q(\boldsymbol{\lambda}) : \boldsymbol{\lambda} \geq \mathbf{0},\ \boldsymbol{\lambda} \in \mathcal{D}_q\}$ and then computing an optimal $\mathbf{x}^*$ by solving $q(\boldsymbol{\lambda}^*)$. In section 2.1 we show how to compute the value of the Lagrange dual function $q(\boldsymbol{\lambda})$ for a fixed value of $\boldsymbol{\lambda} \in \mathcal{D}_q$, in section 3 we describe an active set algorithm to solve for the optimal dual multipliers $\boldsymbol{\lambda}^*$, and in section 4 we show how to recover the optimal primal solution $\mathbf{x}^*$.

2.1. Computing the Lagrangian $q(\boldsymbol{\lambda})$. Claim 2.4 allows us to restrict ourselves to $\boldsymbol{\lambda} \geq \mathbf{0}$ such that $q(\boldsymbol{\lambda}) > -\infty$, i.e., $\boldsymbol{\lambda} \in \mathcal{D}_q \cap \mathbf{R}^l_+$, without any loss of generality. Fix $\mathbf{y} \succeq \mathbf{0}$ and consider the optimization problem in $\mathbf{x}$:

$$q(\boldsymbol{\lambda}, \mathbf{y}) \equiv \begin{array}{rl} \min & (\mathbf{f} - \mathbf{E}^T\boldsymbol{\lambda})^T\mathbf{x} \\ \text{subject to} & \mathbf{H}\mathbf{x} = \mathbf{g}, \\ & \mathbf{D}\mathbf{x} = \mathbf{y}. \end{array} \tag{2.4}$$

Note that $q(\boldsymbol{\lambda}) > -\infty$ if and only if $q(\boldsymbol{\lambda}, \mathbf{y}) > -\infty$ for all $\mathbf{y} \succeq \mathbf{0}$. Since $\mathbf{H}$ has full row rank, $\mathbf{H}\mathbf{x} = \mathbf{g}$ if and only if $\mathbf{x} = \mathbf{x}_0 + \mathbf{B}\mathbf{z}$, where $\mathbf{x}_0 = \mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}\mathbf{g} \in \mathbf{R}^n$, $\mathbf{B} \in \mathbf{R}^{n\times(n-m)}$ is any orthonormal basis for the nullspace $\mathcal{N}(\mathbf{H})$ of $\mathbf{H}$, and $\mathbf{z} \in \mathbf{R}^{n-m}$. Thus, we have that

$$q(\boldsymbol{\lambda}, \mathbf{y}) = (\mathbf{f} - \mathbf{E}^T\boldsymbol{\lambda})^T\mathbf{x}_0 + \begin{array}{rl} \min & (\mathbf{f} - \mathbf{E}^T\boldsymbol{\lambda})^T\mathbf{B}\mathbf{z} \\ \text{subject to} & \mathbf{D}\mathbf{B}\mathbf{z} = \mathbf{y} - \mathbf{D}\mathbf{x}_0. \end{array} \tag{2.5}$$

Since $\mathbf{DB} \in \mathbf{R}^{p\times(n-m)}$ the following three cases exhaust all possibilities:

(i) $\mathbf{rank}(\mathbf{DB}) = r < \min\{p, n-m\}$. In this case, a singular value decomposition (SVD) of $\mathbf{DB}$ has the form

$$\mathbf{DB} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T = \begin{bmatrix}\mathbf{U}_0 & \mathbf{U}_1\end{bmatrix}\begin{bmatrix}\boldsymbol{\Sigma}_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{V}_0^T \\ \mathbf{V}_1^T\end{bmatrix} = \mathbf{U}_0\boldsymbol{\Sigma}_0\mathbf{V}_0^T,$$

where $\mathbf{U}_0 \in \mathbf{R}^{p\times r}$, $\mathbf{U}_1 \in \mathbf{R}^{p\times(p-r)}$, $\mathbf{V}_0 \in \mathbf{R}^{(n-m)\times r}$, $\mathbf{V}_1 \in \mathbf{R}^{(n-m)\times(n-m-r)}$, and $\boldsymbol{\Sigma}_0 \in \mathbf{R}^{r\times r}$ is a diagonal matrix. Consequently, $\mathbf{U}_1^T(\mathbf{y} - \mathbf{D}\mathbf{x}_0) = \mathbf{0}$, and $\mathbf{z} = \mathbf{V}_0\boldsymbol{\Sigma}_0^{-1}\mathbf{U}_0^T(\mathbf{y} - \mathbf{D}\mathbf{x}_0) + \mathbf{V}_1\mathbf{t}$, where $\mathbf{t} \in \mathbf{R}^{n-m-r}$. Thus,

$$q(\boldsymbol{\lambda}, \mathbf{y}) = (\mathbf{f} - \mathbf{E}^T\boldsymbol{\lambda})^T\mathbf{z}_0 + \boldsymbol{\xi}^T\mathbf{y} + \min_{\mathbf{t}}\left\{(\mathbf{f} - \mathbf{E}^T\boldsymbol{\lambda})^T\mathbf{B}\mathbf{V}_1\mathbf{t}\right\}, \tag{2.6}$$

where $\mathbf{z}_0 = (\mathbf{I} - \mathbf{B}\mathbf{V}_0\boldsymbol{\Sigma}_0^{-1}\mathbf{U}_0^T\mathbf{D})\mathbf{x}_0$ and $\boldsymbol{\xi} = \mathbf{U}_0\boldsymbol{\Sigma}_0^{-1}\mathbf{V}_0^T\mathbf{B}^T(\mathbf{f} - \mathbf{E}^T\boldsymbol{\lambda})$. From (2.6), we have

$$\mathbf{q}(\boldsymbol{\lambda}, \mathbf{y}) > -\infty \quad \Leftrightarrow \quad \mathbf{V}_1^T\mathbf{B}^T(\mathbf{f} - \mathbf{E}^T\boldsymbol{\lambda}) = \mathbf{0}, \tag{2.7}$$

and in that case

$$q(\boldsymbol{\lambda}) = (\mathbf{f} - \mathbf{E}^T\boldsymbol{\lambda})^T\mathbf{z}_0 + \bar{q}(\boldsymbol{\xi}), \tag{2.8}$$

where

$$\bar{q}(\boldsymbol{\xi}) = \begin{array}{rl} \min & \boldsymbol{\xi}^T\mathbf{y} \\ \text{subject to} & \mathbf{A}\mathbf{y} = \mathbf{b}, \\ & \mathbf{y} \succeq \mathbf{0} \end{array} \tag{2.9}$$

and

$$\begin{aligned} \mathbf{z}_0 &= (\mathbf{I} - \mathbf{B}\mathbf{V}_0\boldsymbol{\Sigma}_0^{-1}\mathbf{U}_0^T\mathbf{D})\mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}\mathbf{g}, \\ \boldsymbol{\xi} &= \mathbf{U}_0\boldsymbol{\Sigma}_0^{-1}\mathbf{V}_0^T\mathbf{B}^T(\mathbf{f} - \mathbf{E}^T\boldsymbol{\lambda}), \\ \mathbf{b} &= \mathbf{U}_1^T\mathbf{D}\mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}\mathbf{g}, \\ \mathbf{A} &= \mathbf{U}_1^T. \end{aligned} \tag{2.10}$$

(ii) $\mathbf{rank}(\mathbf{DB}) = n - m < p$. In this case, we have

$$\mathbf{DB} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T = \begin{bmatrix}\mathbf{U}_0 & \mathbf{U}_1\end{bmatrix}\begin{bmatrix}\boldsymbol{\Sigma}_0 \\ \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{V}_0^T\end{bmatrix} = \mathbf{U}_0\boldsymbol{\Sigma}_0\mathbf{V}_0^T,$$

where $\mathbf{U}_0 \in \mathbf{R}^{p\times(n-m)}$, $\mathbf{U}_1 \in \mathbf{R}^{p\times(p-n+m)}$, $\mathbf{V}_0 \in \mathbf{R}^{(n-m)\times(n-m)}$, and $\boldsymbol{\Sigma}_0 \in \mathbf{R}^{(n-m)\times(n-m)}$ is a diagonal matrix. Thus, (2.5) is feasible if and only if

$$\mathbf{U}_1^T(\mathbf{y} - \mathbf{D}\mathbf{x}_0) = \mathbf{0}. \tag{2.11}$$

Since $\mathbf{V}_0$ has full rank, it follows that when equation (2.11) holds we have $\mathbf{z} = \mathbf{V}_0\boldsymbol{\Sigma}_0^{-1}\mathbf{U}_0^T(\mathbf{y} - \mathbf{D}\mathbf{x}_0)$. Consequently,

$$q(\boldsymbol{\lambda}, \mathbf{y}) = (\mathbf{f} - \mathbf{E}^T\boldsymbol{\lambda})^T\mathbf{z}_0 + \boldsymbol{\xi}^T\mathbf{y}, \tag{2.12}$$

where $\mathbf{z}_0 = (\mathbf{I} - \mathbf{B}\mathbf{V}_0\boldsymbol{\Sigma}_0^{-1}\mathbf{U}_0^T\mathbf{D})\mathbf{x}_0$ and $\boldsymbol{\xi} = \mathbf{U}_0\boldsymbol{\Sigma}_0^{-1}\mathbf{V}_0^T\mathbf{B}^T(\mathbf{f} - \mathbf{E}^T\boldsymbol{\lambda})$. Thus, (2.8), (2.9), and (2.10) remain valid in this case.

(iii) $\mathbf{rank}(\mathbf{DB}) = p < n - m$. An SVD of $\mathbf{DB}$ is given by

$$\mathbf{DB} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T = \begin{bmatrix}\mathbf{U}_0\end{bmatrix}\begin{bmatrix}\boldsymbol{\Sigma}_0 & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{V}_0^T \\ \mathbf{V}_1^T\end{bmatrix} = \mathbf{U}_0\boldsymbol{\Sigma}_0\mathbf{V}_0^T,$$

where $\mathbf{U}_0 \in \mathbf{R}^{p\times p}$, $\mathbf{V}_0 \in \mathbf{R}^{(n-m)\times p}$, $\mathbf{V}_1 \in \mathbf{R}^{(n-m)\times(n-m-p)}$, and $\boldsymbol{\Sigma}_0 \in \mathbf{R}^{p\times p}$ is a diagonal matrix. Since $\mathbf{U}_0$ has full rank, (2.5) is always feasible. Thus, $\mathbf{z} = \mathbf{V}_0\boldsymbol{\Sigma}_0^{-1}\mathbf{U}_0^T(\mathbf{y} - \mathbf{D}\mathbf{x}_0) + \mathbf{V}_1\mathbf{t}$, where $\mathbf{t} \in \mathbf{R}^{n-m-p}$. Consequently,

$$q(\boldsymbol{\lambda}, \mathbf{y}) = (\mathbf{f} - \mathbf{E}^T\boldsymbol{\lambda})^T\mathbf{z}_0 + \boldsymbol{\xi}^T\mathbf{y} + \min_{\mathbf{t}}\left\{(\mathbf{f} - \mathbf{E}^T\boldsymbol{\lambda})^T\mathbf{B}\mathbf{V}_1\mathbf{t}\right\}, \tag{2.13}$$

where $\mathbf{z}_0 = (\mathbf{I} - \mathbf{B}\mathbf{V}_0\boldsymbol{\Sigma}_0^{-1}\mathbf{U}_0^T\mathbf{D})\mathbf{x}_0$ and $\boldsymbol{\xi} = \mathbf{U}_0\boldsymbol{\Sigma}_0^{-1}\mathbf{V}_0^T\mathbf{B}^T(\mathbf{f} - \mathbf{E}^T\boldsymbol{\lambda})$. From (2.13) we have

$$\mathbf{q}(\boldsymbol{\lambda}, \mathbf{y}) > -\infty \quad \Leftrightarrow \quad \mathbf{V}_1^T\mathbf{B}^T(\mathbf{f} - \mathbf{E}^T\boldsymbol{\lambda}) = \mathbf{0}. \tag{2.14}$$

Thus,

$$q(\boldsymbol{\lambda}) = (\mathbf{f} - \mathbf{E}^T\boldsymbol{\lambda})^T\mathbf{z}_0 + \hat{q}(\boldsymbol{\xi}), \tag{2.15}$$

where

$$\begin{aligned} \hat{q}(\boldsymbol{\xi}) \;=\; & \min \quad \boldsymbol{\xi}^T\mathbf{y} \\ & \text{subject to} \quad \mathbf{y} \succeq \mathbf{0} \end{aligned} \tag{2.16}$$

and

$$\begin{aligned} \mathbf{z}_0 &= (\mathbf{I} - \mathbf{B}\mathbf{V}_0\boldsymbol{\Sigma}_0^{-1}\mathbf{U}_0^T\mathbf{D})\mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}\mathbf{g}, \\ \boldsymbol{\xi} &= \mathbf{U}_0\boldsymbol{\Sigma}_0^{-1}\mathbf{V}_0^T\mathbf{B}^T(\mathbf{f} - \mathbf{E}^T\boldsymbol{\lambda}). \end{aligned} \tag{2.17}$$

Since the structures of the optimization problems (2.9) and (2.16), although similar, are not identical, the corresponding active set methods are also similar but not identical. In the paper we focus on developing an active set method for optimizing the Lagrangian defined in (2.8). The active set method for optimizing the Lagrangian defined in (2.15) is in Appendix B.

LEMMA 2.5. *Let $\bar{q} : \mathbf{R}^p \mapsto \mathbf{R}$ denote the function defined in* (2.9). *Then the domain $\mathcal{D}_{\bar{q}} = \{\boldsymbol{\xi} : \bar{q}(\boldsymbol{\xi}) > -\infty\}$ is given by*

$$\mathcal{D}_{\bar{q}} = \begin{cases} \mathbf{R}^p, & \gamma < 0, \\ \left\{\boldsymbol{\xi} : \mathbf{e}^T\mathbf{P}\boldsymbol{\xi} \geq 0, (\mathbf{e}^T\mathbf{P}\boldsymbol{\xi})^2 - \gamma\|\mathbf{P}\boldsymbol{\xi}\|^2 \geq 0\right\}, & \gamma \geq 0, \end{cases} \tag{2.18}$$

where $\mathbf{e} = (1, \mathbf{0}^T)^T$, $\mathbf{P} = \mathbf{I} - \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}$, $\mathbf{a} = \mathbf{A}\mathbf{e}$, *and* $\gamma = \frac{1}{2} - \mathbf{a}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{a}$. *For all* $\boldsymbol{\xi} \in \mathcal{D}_{\bar{q}}$,

$$\bar{q}(\boldsymbol{\xi}) = \mathbf{v}^T\boldsymbol{\xi} + f(\mathbf{P}\boldsymbol{\xi}),$$

where

$$\mathbf{v} = \begin{cases} \mathbf{R}\mathbf{A}^T(\mathbf{A}\mathbf{R}\mathbf{A}^T)^{-1}\mathbf{b}, & \gamma \neq 0, \\ \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{b}, & \gamma = 0, \end{cases} \tag{2.19}$$

$$f(\mathbf{u}) = \begin{cases} \frac{\sqrt{-\gamma(\mathbf{b}^T(\mathbf{A}\mathbf{R}\mathbf{A}^T)^{-1}\mathbf{b})}}{\gamma}\sqrt{(\mathbf{e}^T\mathbf{u})^2 - \gamma\|\mathbf{u}\|^2}, & \gamma \neq 0, \\ \left(\frac{\|\mathbf{y}_0\|^2 - 2(\mathbf{e}^T\mathbf{y}_0)^2}{2\mathbf{e}^T\mathbf{y}_0}\right)\mathbf{e}^T\mathbf{u} - \mathbf{e}^T\mathbf{y}_0\left(\frac{\|\mathbf{u}\|^2 - 2(\mathbf{e}^T\mathbf{u})^2}{2\mathbf{e}^T\mathbf{u}}\right), & \gamma = 0, \end{cases} \tag{2.20}$$

and $\mathbf{y}_0 = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{b}$.

The proof of this result is fairly straightforward and is therefore relegated to Appendix A.

3. Active set algorithm for the Lagrangian dual problem. Note that from (2.10) we have that $\mathbf{A}\boldsymbol{\xi} = \mathbf{U}_1^T\mathbf{U}_0\boldsymbol{\Sigma}_0^{-1}\mathbf{V}_0^T\mathbf{B}^T(\mathbf{f} - \mathbf{E}^T\boldsymbol{\lambda}) = \mathbf{0}$, i.e., $\boldsymbol{\xi} = \mathbf{P}\boldsymbol{\xi}$. Thus, (2.7), (2.8), (2.9), (2.10), and Lemma 2.5 imply that the Lagrangian dual problem is given by

$$\begin{aligned} \max \quad & (\mathbf{f} - \mathbf{E}^T\boldsymbol{\lambda})^T\mathbf{z}_0 + \mathbf{v}^T\boldsymbol{\xi} + f(\boldsymbol{\xi}) \\ \text{subject to} \quad & \mathbf{L}\boldsymbol{\lambda} + \boldsymbol{\xi} = \mathbf{h}, \\ & \mathbf{M}\boldsymbol{\lambda} \quad = \mathbf{p}, \\ & \boldsymbol{\lambda} \quad \geq \mathbf{0}, \\ & \boldsymbol{\xi} \in \mathcal{K}, \end{aligned} \tag{3.1}$$

where

$$\begin{aligned} \mathbf{L} &= \mathbf{U}_0\boldsymbol{\Sigma}_0^{-1}\mathbf{V}_0^T\mathbf{B}^T\mathbf{E}^T, \\ \mathbf{h} &= \mathbf{U}_0\boldsymbol{\Sigma}_0^{-1}\mathbf{V}_0^T\mathbf{B}^T\mathbf{f}, \\ \mathbf{M} &= \mathbf{V}_1^T\mathbf{B}^T\mathbf{E}^T, \\ \mathbf{p} &= \mathbf{V}_1^T\mathbf{B}^T\mathbf{f}, \end{aligned}$$

and

The LagrangeDual algorithm.
Input: Optimization problem (3.1).
Output: Optimal solution of (3.1).
set $\boldsymbol{\mu}^{(1)} \leftarrow \operatorname{argmax}\{-\mathbf{z}_0^T\mathbf{E}^T\boldsymbol{\lambda} : \mathcal{A}[\boldsymbol{\lambda}, \mathbf{0}, \mathbf{h}, \mathbf{p}] = \mathbf{0},\ \boldsymbol{\lambda} \geq \mathbf{0}\}$
set $(\alpha^{(0)}, \boldsymbol{\mu}^{(0)}) \leftarrow \operatorname{argmin}\{(3.2)\}$
if $(\alpha^{(0)}, \boldsymbol{\mu}^{(0)}) = \emptyset$ or $(\|\alpha^{(0)}\mathbf{h} - \mathbf{L}\boldsymbol{\mu}^{(0)}\|^2 > 1/\gamma)$
  set $\boldsymbol{\mu} \leftarrow \boldsymbol{\mu}^{(1)}$
else if $(\|\alpha^{(0)}\mathbf{h} - \mathbf{L}\boldsymbol{\mu}^{(0)}\|^2 < 1/\gamma)$
  if $(\alpha^{(0)} > 0)$ **set** $\boldsymbol{\mu}^{(2)} \leftarrow \frac{1}{\alpha^{(0)}}\boldsymbol{\mu}^{(0)}$
  else **if** $\boldsymbol{\mu}^{(1)} \neq \emptyset$ **set** $\boldsymbol{\mu} \leftarrow \boldsymbol{\mu}^{(1)}$; **else choose** $\boldsymbol{\mu} \in \{\boldsymbol{\lambda} : \mathbf{M}\boldsymbol{\lambda} = \mathbf{p}, \boldsymbol{\lambda} \geq \mathbf{0}\}$ **choose** $\hat{\omega}$ s.t. $\mathbf{h} - \mathbf{L}(\boldsymbol{\mu} + \hat{\omega}\boldsymbol{\mu}^{(0)}) \in \mathbf{int}(\mathcal{K})$
    set $\boldsymbol{\mu}^{(2)} \leftarrow \boldsymbol{\mu} + \hat{\omega}\boldsymbol{\mu}^{(0)}$
  end
  set $\boldsymbol{\mu} \leftarrow$ ActiveSet$(\boldsymbol{\mu}^{(1)}, \boldsymbol{\mu}^{(2)})$
else
  if $(\alpha^{(0)} > 0)$
    set $\boldsymbol{\mu}^{(2)} \leftarrow \frac{1}{\alpha^{(0)}}\boldsymbol{\mu}^{(0)}$
    set $(\omega, \boldsymbol{\mu}) \leftarrow \operatorname{argmin}\{(3.3)\}$
  else
    set $(\omega, \boldsymbol{\mu}) \leftarrow \operatorname{argmin}\{(3.4)\}$
  end
end
return $\boldsymbol{\mu}$

FIG. 3.1. *Lagrangian dual algorithm.*

$$\mathcal{K} = \begin{cases} \mathbf{R}^p, & \gamma < 0, \\ \left\{\mathbf{z} : \mathbf{e}^T\mathbf{z} \geq 0, (\mathbf{e}^T\mathbf{z})^2 - \gamma\|\mathbf{z}\|^2 \geq 0\right\}, & \gamma \geq 0, \end{cases}$$

where $\mathbf{v}$ and $f(\cdot)$ are as defined in (2.19) and (2.20), respectively. In the rest of the paper we denote the system of linear equalities in (3.1) by $\mathcal{A}[\boldsymbol{\lambda}, \boldsymbol{\xi}, \mathbf{h}, \mathbf{p}] = \mathbf{0}$.

When $\gamma \leq 0$, the constraints in (3.1) are linear, and hence (3.1) can be solved using any standard active set method for optimizing a concave function over a polytope. Moreover, γ is strictly positive for all single-cone SOCPs arising in the context of robust optimization. Therefore, in this paper we focus on constructing an active set algorithm for the case when γ is strictly positive. In the rest of this section we prove that the LagrangeDual algorithm displayed in Figure 3.1 computes an optimal solution of (3.1). We adopt the convention that a solution algorithm returns the empty set as a solution if and only if the problem is infeasible.

Let $\mathcal{C} = \left\{\boldsymbol{\xi} : \exists\boldsymbol{\lambda} \geq \mathbf{0} \text{ s.t. } \mathcal{A}[\boldsymbol{\lambda}, \boldsymbol{\xi}, \mathbf{h}, \mathbf{p}] = \mathbf{0}\right\}$. Then $\mathcal{C} \cap \mathcal{K} = \{\mathbf{h} - \mathbf{L}\boldsymbol{\lambda} : \mathbf{M}\boldsymbol{\lambda} = \mathbf{p}, \mathbf{h} - \mathbf{L}\boldsymbol{\lambda} \in \mathcal{K},\ \boldsymbol{\lambda} \geq \mathbf{0}\}$. We construct the active set algorithm by considering the following three mutually exclusive cases: $\mathcal{C} \cap \mathcal{K} = \emptyset$, $\mathcal{C} \cap \mathcal{K} \subset \partial\mathcal{K}$, and $\mathcal{C} \cap \mathbf{int}(\mathcal{K}) \neq \emptyset$. In order to distinguish between these three cases we "homogenize" the set $\mathcal{C} \cap \mathcal{K}$ and solve the following least squares problem:

$$\begin{aligned} \min \quad & \|\alpha\mathbf{h} - \mathbf{L}\boldsymbol{\lambda}\|^2 \\ \text{subject to} \quad & \alpha\mathbf{e}^T\mathbf{h} - \mathbf{e}^T\mathbf{L}\boldsymbol{\lambda} = 1, \\ & \alpha\mathbf{p} - \mathbf{M}\boldsymbol{\lambda} = \mathbf{0}, \\ & \alpha \geq 0, \\ & \boldsymbol{\lambda} \geq \mathbf{0}. \end{aligned} \tag{3.2}$$

Let $(\alpha^{(0)}, \boldsymbol{\mu}^{(0)})$ denote the optimal solution of (3.2). Then one of the following three mutually exclusive conditions holds:

```
The ActiveSet algorithm.
Input: Optimization problem (3.1), μ^(1), and μ^(2).
Output: Optimal solution of (3.1).
 quit ← 0    k ← 0
 if (μ^(1) ≠ ∅)
                (d_ξ, d_λ) ← FindDirection(∅)
                if (v^T d_ξ − z_0^T E^T d_λ + f(d_ξ) ≤ 0) return (0, μ^(1))
                else (ξ^(0), λ^(0)) ← FindStep((0, μ^(1)), (d_ξ, d_λ), ∞)
 else
                λ^(0) ← μ^(2)
 end if
 W^(k) ← Σ_{i:λ_i^(k)=0} e_i e_i^T
 /* e_i denotes the ith column of an identity matrix */
 while (∼quit)
                (d_ξ, d_λ, α_q) ← FindOpt(ξ^(k), λ^(k), W^(k))
                if ((d_ξ, d_λ) ≠ 0)
                        (ξ^(k+1), λ^(k+1)) ← FindStep((ξ^(k), λ^(k)), (d_ξ, d_λ), α_q)
                        k ← k + 1
                        W^(k) ← Σ_{i:λ_i^(k)=0} e_i e_i^T
                else    ρ ← FindMultipliers(W^(k), ξ^(k))
                        if (max_i{ρ_i} ≤ 0)
                                quit ← 1
                        else    (ξ^(k+1), λ^(k+1)) ← (ξ^(k), λ^(k))
                                choose j: ρ_j > 0
                                W^(k+1) ← W^(k) − e_j e_j^T
                                k ← k + 1
                        end
                end
 end
 return λ^(k)
```

FIG. 3.2. *Active set algorithm.*

(i) Either (3.2) is infeasible or $\|\alpha^{(0)}\mathbf{h}-\mathbf{L}\boldsymbol{\mu}^{(0)}\|^2 > \frac{1}{\gamma}$. Since (3.2) was constructed by "homogenizing" $\mathcal{C}\cap\mathcal{K}$, it follows that either $\mathcal{C}\cap\mathcal{K}=\emptyset$ or $\mathcal{C}\cap\mathcal{K}=\{\mathbf{0}\}$. The latter can be checked by solving an LP.

(ii) $\|\alpha^{(0)}\mathbf{h}-\mathbf{L}\boldsymbol{\mu}^{(0)}\|^2 < \frac{1}{\gamma}$. We have the following two possibilities:
   (a) $\alpha^{(0)} > 0$: $\boldsymbol{\mu}^{(2)} = \frac{1}{\alpha^{(0)}}\boldsymbol{\mu}^{(0)}$ satisfies $\mathbf{h}-\mathbf{L}\boldsymbol{\mu}^{(2)} \in \mathcal{C}\cap\mathbf{int}(\mathcal{K})$.
   (b) $\alpha^{(0)} = 0$: It is easy to check that $\boldsymbol{\mu}^{(0)}$ is a recession direction of the polytope $\mathcal{P} = \{\boldsymbol{\lambda} : \mathbf{M}\boldsymbol{\lambda}=\mathbf{p},\ \boldsymbol{\lambda}\geq\mathbf{0}\}$ and $-\mathbf{e}^T\mathbf{L}\boldsymbol{\mu}^{(0)} = 1$. Let $\hat{\boldsymbol{\lambda}}\in\mathcal{P}$. Then, by definition, $\boldsymbol{\lambda}_\omega = \hat{\boldsymbol{\lambda}} + \omega\boldsymbol{\mu}^{(0)} \in \mathcal{P}$ for all $\omega\geq 0$. Since $\mathbf{e}^T(\mathbf{h}-\mathbf{L}\boldsymbol{\lambda}_\omega) > 0$ for all large enough ω, and $\lim_{\omega\to\infty}\{\|\mathbf{h}-\mathbf{L}\boldsymbol{\lambda}_\omega\|/(\mathbf{e}^T(\mathbf{h}-\mathbf{L}\boldsymbol{\lambda}_\omega))\} < \frac{1}{\sqrt{\gamma}}$, it follows that there exists $\omega > 0$ such that $\boldsymbol{\mu}^{(2)} = \boldsymbol{\lambda}_\omega$ satisfies $\mathbf{h}-\mathbf{L}\boldsymbol{\mu}^{(2)} \in \mathcal{C}\cap\mathbf{int}(\mathcal{K})$.

   In this case, LagrangeDual completes the optimization by calling the ActiveSet algorithm displayed in Figure 3.2.

(iii) $\|\alpha^{(0)}\mathbf{h}-\mathbf{L}\boldsymbol{\mu}^{(0)}\|^2 = \frac{1}{\gamma}$. In this case $\mathcal{C}\cap\mathbf{int}(\mathcal{K}) = \emptyset$ and one has to consider the following two possibilities.
   (a) $\alpha^{(0)} > 0$: $\boldsymbol{\mu}^{(2)} = \frac{1}{\alpha^{(0)}}\boldsymbol{\mu}^{(0)}$ satisfies $\mathbf{h}-\mathbf{L}\boldsymbol{\mu}^{(2)} \in \mathcal{C}\cap\partial\mathcal{K}$. Since the optimal value of (3.2) is $\frac{1}{\gamma}$ and the Euclidean norm is a strictly convex function, it follows that $\mathcal{C}\cap\mathcal{K} = \{\omega(\mathbf{h}-\mathbf{L}\boldsymbol{\mu}^{(2)}) : \mathbf{h}-\mathbf{L}\boldsymbol{\lambda} = \omega(\mathbf{h}-\mathbf{L}\boldsymbol{\mu}^{(2)}),\ \mathbf{M}\boldsymbol{\lambda}=\mathbf{p},\ \boldsymbol{\lambda}\geq\mathbf{0},\ \omega\geq 0\}$. Since $f(\xi) = 0$ for all $\xi\in\mathcal{C}\cap\mathcal{K}$ (see (2.20)), it follows

that the optimization problem (3.1) reduces to the LP

$$\begin{array}{rl} \max & (\mathbf{f}-\mathbf{E}^T\boldsymbol{\lambda})^T\mathbf{z}_0+\mathbf{v}^T(\mathbf{h}-\mathbf{L}\boldsymbol{\mu}^{(2)})\omega \\ \text{subject to} & \mathbf{L}\boldsymbol{\lambda}+(\mathbf{h}-\mathbf{L}\boldsymbol{\mu}^{(2)})\omega=\mathbf{h}, \\ & \mathbf{M}\boldsymbol{\lambda} = \mathbf{p}, \\ & \boldsymbol{\lambda} \geq \mathbf{0}, \\ & \omega \geq 0. \end{array} \tag{3.3}$$

(b) $\alpha^{(0)}=0$: The recession direction $\boldsymbol{\mu}^{(0)}$ satisfies $-\mathbf{L}\boldsymbol{\mu}^{(0)}\in\partial\mathcal{K}$. An argument similar to the one in part (a) implies that (3.1) reduces to the LP

$$\begin{array}{rl} \max & (\mathbf{f}-\mathbf{E}^T\boldsymbol{\lambda})^T\mathbf{z}_0-\mathbf{v}^T(\mathbf{L}\boldsymbol{\mu}^{(0)})\omega \\ \text{subject to} & \mathbf{L}\boldsymbol{\lambda}-(\mathbf{L}\boldsymbol{\mu}^{(0)})\omega=\mathbf{h}, \\ & \mathbf{M}\boldsymbol{\lambda} = \mathbf{p}, \\ & \boldsymbol{\lambda} \geq \mathbf{0}, \\ & \omega \geq 0. \end{array} \tag{3.4}$$

Next, we establish the correctness of the procedure ActiveSet displayed in Figure 3.2. We begin by showing that for any optimal solution $(\boldsymbol{\xi}^*,\boldsymbol{\lambda}^*)$ of (3.1), either $\boldsymbol{\xi}^*=\mathbf{0}$ or $\boldsymbol{\xi}^*\in\mathcal{C}\cap\mathbf{int}(\mathcal{K})$, i.e., $\boldsymbol{\xi}^*\notin\mathcal{C}\cap\big(\partial\mathcal{K}\backslash\{\mathbf{0}\}\big)$.

Lemma 3.1. *Suppose* $\mathcal{C}\cap\mathbf{int}(\mathcal{K})\neq\emptyset$ *and let* $(\boldsymbol{\xi}^*,\boldsymbol{\lambda}^*)$ *denote any optimal solution of* (3.1). *Then* $\boldsymbol{\xi}^*\notin\mathcal{C}\cap\big(\partial\mathcal{K}\backslash\{\mathbf{0}\}\big)$.

Proof. Assume otherwise, i.e., $\boldsymbol{\xi}^*\in\mathcal{C}\cap\big(\partial\mathcal{K}\backslash\{\mathbf{0}\}\big)$ for some optimal solution $(\boldsymbol{\xi}^*,\boldsymbol{\lambda}^*)$. Let $(\boldsymbol{\xi}_0,\boldsymbol{\lambda}_0)$ denote any feasible solution of (3.1) with $\boldsymbol{\xi}_0\in\mathcal{C}\cap\mathbf{int}(\mathcal{K})$. For $\beta\in[0,1]$, let $(\boldsymbol{\xi}_\beta,\boldsymbol{\lambda}_\beta)$ denote the convex combination $(\boldsymbol{\xi}_\beta,\boldsymbol{\lambda}_\beta)=\beta(\boldsymbol{\xi}_0,\boldsymbol{\lambda}_0)+(1-\beta)(\boldsymbol{\xi}^*,\boldsymbol{\lambda}^*)$ and let $r(\beta)$ denote the objective value of (3.1) evaluated at $(\boldsymbol{\xi}_\beta,\boldsymbol{\lambda}_\beta)$. Then

$$\begin{aligned} r(\beta) &= \mathbf{v}^T\boldsymbol{\xi}_\beta+\mathbf{f}^T\mathbf{z}_0-\mathbf{z}_0^T\mathbf{E}^T\boldsymbol{\lambda}_\beta+f(\boldsymbol{\xi}_\beta) \\ &= \mathbf{v}^T\boldsymbol{\xi}_\beta+\mathbf{f}^T\mathbf{z}_0-\mathbf{z}_0^T\mathbf{E}^T\boldsymbol{\lambda}_\beta+\theta\sqrt{(\mathbf{e}^T\boldsymbol{\xi}_\beta)^2-\gamma\|\boldsymbol{\xi}_\beta\|^2} \\ &\geq \underbrace{\big(\mathbf{v}^T\boldsymbol{\xi}^*+\mathbf{f}^T\mathbf{z}_0-\mathbf{z}_0^T\mathbf{E}^T\boldsymbol{\lambda}^*\big)}_{=r(0)}+\beta\underbrace{\big(\mathbf{v}^T(\boldsymbol{\xi}_0-\boldsymbol{\xi}^*)-\mathbf{z}_0^T\mathbf{E}^T(\boldsymbol{\lambda}_0-\boldsymbol{\lambda}^*)\big)}_{\triangleq\delta} \\ &\quad+\theta\sqrt{\beta^2\left((\mathbf{e}^T\boldsymbol{\xi}_0)^2-\gamma\|\boldsymbol{\xi}_0\|^2\right)+2\beta(1-\beta)\left((\mathbf{e}^T\boldsymbol{\xi}_0)(\mathbf{e}^T\boldsymbol{\xi}^*)-\gamma\|\boldsymbol{\xi}_0\|\|\boldsymbol{\xi}^*\|\right)}, \end{aligned} \tag{3.5}$$

where $\theta=\frac{\sqrt{-\gamma\mathbf{b}^T(\mathbf{A}\mathbf{R}\mathbf{A}^T)^{-1}\mathbf{b}}}{\gamma}$ and the last inequality follows from the fact that $(\mathbf{e}^T\boldsymbol{\xi}^*)^2-\gamma\|\boldsymbol{\xi}^*\|^2=0$. Since $\boldsymbol{\xi}_0\in\mathcal{C}\cap\mathbf{int}(\mathcal{K})$ we have

$$\epsilon=\min\left\{(\mathbf{e}^T\boldsymbol{\xi}_0)^2-\gamma\|\boldsymbol{\xi}_0\|^2,(\mathbf{e}^T\boldsymbol{\xi}_0)(\mathbf{e}^T\boldsymbol{\xi}^*)-\gamma\|\boldsymbol{\xi}_0\|\|\boldsymbol{\xi}^*\|\right\}>0.$$

From (3.5) we have that $r(\beta)-r(0)\geq\theta\sqrt{\epsilon}\sqrt{2\beta-\beta^2}+\beta\delta$. Choose β_0 as follows:

$$\beta_0=\begin{cases}1, & \delta\geq 0,\\ 1+\frac{\delta}{\sqrt{\theta^2\epsilon+\delta^2}}, & \delta<0.\end{cases}$$

Then it follows that $\beta_0>0$ and $r(\beta_0)-r(0)>0$, a contradiction. ∎

The ActiveSet algorithm receives as input

(i) $\boldsymbol{\mu}^{(1)}=\text{argmax}\{-\mathbf{z}_0^T\mathbf{E}^T\boldsymbol{\lambda}:\mathcal{A}[\boldsymbol{\lambda},\mathbf{0},\mathbf{h},\mathbf{p}],\ \boldsymbol{\lambda}\geq\mathbf{0}\}$, and

(ii) a vector $\boldsymbol{\mu}^{(2)}$ such that $\mathbf{h}-\mathbf{L}\boldsymbol{\mu}^{(2)}\in\mathcal{C}\cap\mathbf{int}(\mathcal{K})$.

When $\boldsymbol{\mu}^{(1)} \neq \emptyset$, the algorithm calls the procedure FindDirection that returns an ascent direction at $(\boldsymbol{\xi}, \boldsymbol{\lambda}) = (\mathbf{0}, \boldsymbol{\mu}^{(1)})$, if it exists; otherwise it returns $(\mathbf{0}, \mathbf{0})$. If FindDirection returns $(\mathbf{0}, \mathbf{0})$, it follows that $(\mathbf{0}, \boldsymbol{\mu}^{(1)})$ is optimal and the algorithm terminates; otherwise ActiveSet calls the procedure FindStep$((\boldsymbol{\xi}, \boldsymbol{\lambda}), (\mathbf{d}_\xi, \mathbf{d}_\lambda), \alpha_q)$ that computes the iterate $(\boldsymbol{\xi}^{(0)}, \boldsymbol{\lambda}^{(0)})$ as follows:

$$(\boldsymbol{\xi}^{(0)}, \boldsymbol{\lambda}^{(0)}) = (\boldsymbol{\xi}, \boldsymbol{\lambda}) + \alpha_{\min}(\mathbf{d}_\xi, \mathbf{d}_\lambda), \quad \alpha_{\min} = \min\{\max\{\alpha : \boldsymbol{\lambda} + \alpha \mathbf{d}_\lambda \geq \mathbf{0}\}, \alpha_q\}.$$

Since $\alpha \mathbf{d}_\xi \in \mathcal{K}$ for all $\alpha \geq 0$, $\alpha_{\min}$ is only limited by the nonnegativity constraints on $\boldsymbol{\lambda}$. Note that the iterate $(\boldsymbol{\xi}^{(0)}, \boldsymbol{\lambda}^{(0)})$ satisfies $\boldsymbol{\xi}^{(0)} \in \mathbf{int}(\mathcal{K})$; therefore, the optimum solution $(\boldsymbol{\xi}^*, \boldsymbol{\lambda}^*)$ also satisfies $\boldsymbol{\xi}^* \in \mathbf{int}(\mathcal{K})$ by Lemma 3.1.

Next, we show that the procedure FindDirection can be implemented efficiently. The pair $(\mathbf{d}_\xi, \mathbf{d}_\lambda)$ is an ascent direction at $(\mathbf{0}, \boldsymbol{\mu}^{(1)})$ if and only if $(\mathbf{d}_\xi, \mathbf{d}_\lambda)$ is a recession direction for the set

$$\begin{aligned} -\mathbf{z}_0^T \mathbf{E}^T \mathbf{d}_\lambda + \mathbf{v}^T \mathbf{d}_\xi + \theta \sqrt{(\mathbf{e}^T \mathbf{d}_\xi)^2 - \gamma \|\mathbf{d}_\xi\|^2} > 0, \\ \mathcal{A}[\mathbf{d}_\lambda, \mathbf{d}_\xi, \mathbf{0}, \mathbf{0}] = \mathbf{0}, \\ \mathbf{d}_\xi \in \mathcal{K}, \end{aligned} \tag{3.6}$$

Lemma 3.2. *Let* $\mathcal{A}_{\mathbf{W}}[\mathbf{u}, \mathbf{v}, \boldsymbol{\gamma}, \boldsymbol{\nu}] = \mathbf{0}$ *denote the system of linear equalities*

$$\begin{bmatrix} \mathbf{L} & \mathbf{I} \\ \mathbf{M} & \mathbf{0} \\ \mathbf{W} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\gamma} \\ \boldsymbol{\nu} \\ \mathbf{0} \end{bmatrix},$$

where $(\mathbf{u}, \mathbf{v})$ *are variables, and* $(\boldsymbol{\gamma}, \boldsymbol{\nu}, \mathbf{W})$ *are parameters. Then a recession direction* $(\mathbf{d}_\xi, \mathbf{d}_\lambda)$ *for the set*

$$\begin{aligned} -\mathbf{z}_0^T \mathbf{E}^T \mathbf{d}_\lambda + \mathbf{v}^T \mathbf{d}_\xi + \theta \sqrt{(\mathbf{e}^T \mathbf{d}_\xi)^2 - \gamma \|\mathbf{d}_\xi\|^2} > 0, \\ \mathcal{A}_{\mathbf{W}}[\mathbf{d}_\lambda, \mathbf{d}_\xi, \mathbf{0}, \mathbf{0}] = \mathbf{0}, \\ \mathbf{d}_\xi \in \mathcal{K}, \end{aligned} \tag{3.7}$$

if it exists, can be computed by solving two systems of linear equalities.

Remark 3.3. *Although* FindDirection *computes an ascent direction of the set* (3.7) *for the special case* $\mathbf{W} = \mathbf{0}$*, we prove the result for general* $\mathbf{W}$ *since we need such a result at a later stage.*

Proof. The set in (3.7) has a recession direction if and only if the optimization problem

$$\begin{aligned} \max \quad & -\mathbf{z}_0^T \mathbf{E}^T \mathbf{d}_\lambda + \mathbf{v}^T \mathbf{d}_\xi + \theta \sqrt{(\mathbf{e}^T \mathbf{d}_\xi)^2 - \gamma \|\mathbf{d}_\xi\|^2} \\ \text{subject to} \quad & \mathcal{A}_{\mathbf{W}}[\mathbf{d}_\lambda, \mathbf{d}_\xi, \mathbf{0}, \mathbf{0}] = \mathbf{0}, \\ & \mathbf{d}_\xi \in \mathcal{K} \end{aligned} \tag{3.8}$$

is unbounded.

An argument similar to the one employed in the proof of Lemma 3.1 establishes that one can restrict one's attention to $(\mathbf{d}_\xi, \mathbf{d}_\lambda)$ satisfying $\mathbf{d}_\xi \in \mathbf{int}(\mathcal{K}) \cup \{\mathbf{0}\}$. The direction $(\mathbf{d}_\xi, \mathbf{d}_\lambda)$ can be computed by considering the following three cases:

(a) First consider positive recession directions of the form $(\mathbf{0}, \mathbf{d}_\lambda)$. It is easy to see that all such directions, modulo a positive multiple, are solutions of the following set of linear equalities:

$$\begin{aligned} -\mathbf{z}_0^T \mathbf{E}^T \mathbf{d}_\lambda = 1, \\ \mathcal{A}_{\mathbf{W}}[\mathbf{d}_\lambda, \mathbf{0}, \mathbf{0}, \mathbf{0}] = \mathbf{0}. \end{aligned} \tag{3.9}$$

(b) Next, suppose (3.9) is infeasible; there still exists, however, a positive recession direction for (3.8). Set $\mathbf{e}^T\mathbf{d}_\xi = 1$ in (3.8) to obtain

$$\begin{aligned} \max \quad & -\mathbf{z}_0^T\mathbf{E}^T\mathbf{d}_\lambda + \mathbf{v}^T\mathbf{d}_\xi + \theta\sqrt{1-\gamma\|\mathbf{d}_\xi\|^2} \\ \text{subject to} \quad & \mathcal{A}_{\mathbf{W}}[\mathbf{d}_\lambda, \mathbf{d}_\xi, \mathbf{0}, \mathbf{0}] = \mathbf{0}, \\ & \mathbf{e}^T\mathbf{d}_\xi = 1, \\ & \gamma\|\mathbf{d}_\xi\|^2 \le 1. \end{aligned} \tag{3.10}$$

Since (3.9) is assumed to be infeasible, (3.10) is bounded. Setting $\mathbf{d}_\xi = -\mathbf{L}\mathbf{d}_\lambda$, we get

$$\begin{aligned} \max \quad & -\left(\mathbf{E}\mathbf{z}_0 + \mathbf{L}^T\mathbf{v}\right)^T\mathbf{d}_\lambda + \theta\sqrt{1-\gamma\|\mathbf{L}\mathbf{d}_\lambda\|^2} \\ \text{subject to} \quad & \mathcal{A}_{\mathbf{W}}[\mathbf{d}_\lambda, -\mathbf{L}\mathbf{d}_\lambda, \mathbf{0}, \mathbf{0}] = \mathbf{0}, \\ & \mathbf{e}^T\mathbf{L}\mathbf{d}_\lambda = -1, \\ & \gamma\|\mathbf{L}\mathbf{d}_\lambda\|^2 \le 1. \end{aligned} \tag{3.11}$$

Since the optimal $\mathbf{d}_\xi^* \in \mathbf{int}(\mathcal{K})$, we have $\gamma\|\mathbf{L}\mathbf{d}_\lambda^*\|^2 < 1$, and therefore the optimal Lagrange multiplier corresponding to this constraint is zero. Thus, the Lagrangian $\mathcal{L}$ of (3.11) reduces to

$$\begin{aligned} \mathcal{L} = & -(\mathbf{E}\mathbf{z}_0 + \mathbf{L}^T\mathbf{v})^T\mathbf{d}_\lambda + \theta\sqrt{1-\gamma\|\mathbf{L}\mathbf{d}_\lambda\|^2} \\ & - \boldsymbol{\tau}^T\mathbf{M}\mathbf{d}_\lambda - \boldsymbol{\rho}^T\mathbf{W}\mathbf{d}_\lambda - \eta(\mathbf{e}^T\mathbf{L}\mathbf{d}_\lambda + 1) \end{aligned}$$

and the first-order optimality conditions are given by

$$\begin{aligned} & \tfrac{\theta\gamma}{\beta}\mathbf{L}^T\mathbf{L}\mathbf{d}_\lambda + \mathbf{M}^T\boldsymbol{\tau} + \mathbf{W}^T\boldsymbol{\rho} + \mathbf{L}^T\mathbf{e}\eta = -(\mathbf{E}\mathbf{z}_0 + \mathbf{L}^T\mathbf{v}), \\ & \mathbf{M}\mathbf{d}_\lambda = \mathbf{0}, \\ & \mathbf{W}\mathbf{d}_\lambda = \mathbf{0}, \\ & \mathbf{e}^T\mathbf{L}\mathbf{d}_\lambda = -1, \end{aligned} \tag{3.12}$$

where $\beta = \sqrt{1-\gamma\|\mathbf{L}\mathbf{d}_\lambda\|^2}$. Since we are looking for solutions $\mathbf{d}_\xi = -\mathbf{L}\mathbf{d}_\lambda \in \mathbf{int}(\mathcal{K})$, we are interested only in the solutions to (3.12) that satisfy $\beta > 0$. By setting $\bar{\boldsymbol{\rho}} = \beta\boldsymbol{\rho}$, $\bar{\boldsymbol{\tau}} = \beta\boldsymbol{\tau}$, and $\bar{\eta} = \beta\eta$, we see that (3.12) is equivalent to

$$\underbrace{\begin{bmatrix} \theta\gamma\mathbf{L}^T\mathbf{L} & \mathbf{M}^T & \mathbf{W}^T & \mathbf{L}^T\mathbf{e} \\ \mathbf{M} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{W} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{e}^T\mathbf{L} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}}_{\triangleq \mathbf{K}} \begin{bmatrix} \mathbf{d}_\lambda \\ \bar{\boldsymbol{\tau}} \\ \bar{\boldsymbol{\rho}} \\ \bar{\eta} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ -1 \end{bmatrix} - \beta \begin{bmatrix} \mathbf{E}\mathbf{z}_0 + \mathbf{L}^T\mathbf{v} \\ \mathbf{0} \\ \mathbf{0} \\ 0 \end{bmatrix}. \tag{3.13}$$

Suppose $\mathbf{K}$ is nonsingular. Let $\mathbf{w} = (\bar{\boldsymbol{\tau}}^T, \bar{\boldsymbol{\rho}}^T, \bar{\eta})^T$, $\mathbf{b}_1 = (\mathbf{0}^T, \mathbf{0}^T, -1)^T$, and $\mathbf{b}_2 = \mathbf{E}\mathbf{z}_0 + \mathbf{L}^T\mathbf{v}$. Partition $\mathbf{K}^{-1}$ into submatrices

$$\mathbf{K}^{-1} = \begin{bmatrix} \mathbf{K}_{11}^{-1} & \mathbf{K}_{12}^{-1} \\ \mathbf{K}_{12}^{-T} & \mathbf{K}_{22}^{-1} \end{bmatrix}$$

such that

$$\begin{bmatrix} \mathbf{d}_\lambda \\ \mathbf{w} \end{bmatrix} = \mathbf{K}^{-1}\begin{bmatrix} \mathbf{0} \\ \mathbf{b}_1 \end{bmatrix} - \beta\mathbf{K}^{-1}\begin{bmatrix} \mathbf{b}_2 \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{K}_{12}^{-1}\mathbf{b}_1 - \beta\mathbf{K}_{11}^{-1}\mathbf{b}_2 \\ \mathbf{K}_{22}^{-1}\mathbf{b}_1 - \beta\mathbf{K}_{12}^{-T}\mathbf{b}_2 \end{bmatrix}.$$

This partition implies that $\mathbf{K}_{12}^{-T}\mathbf{L}^T\mathbf{L}\mathbf{K}_{11}^{-1} = \mathbf{0}$. Therefore,

$$\begin{aligned}\beta^2 + \gamma\|\mathbf{L}\mathbf{d}_\lambda\|^2 - 1 &= \beta^2(1+\gamma\|\mathbf{L}\mathbf{K}_{11}^{-1}\mathbf{b}_2\|^2) - 2\beta\gamma(\mathbf{L}\mathbf{K}_{12}^{-1}\mathbf{b}_1)^T\mathbf{L}\mathbf{K}_{11}^{-1}\mathbf{b}_2 \\ &\quad + \gamma\|\mathbf{L}\mathbf{K}_{12}^{-1}\mathbf{b}_1\|^2 - 1 \\ &= \beta^2(1+\gamma\|\mathbf{L}\mathbf{K}_{11}^{-1}\mathbf{b}_2\|^2) + \gamma\|\mathbf{L}\mathbf{K}_{12}^{-1}\mathbf{b}_1\|^2 - 1.\end{aligned}$$

Consequently, the unique positive solution of the quadratic equation $\beta^2 = 1 - \gamma\|\mathbf{L}\mathbf{d}_\lambda\|^2$ is

$$\beta = \sqrt{\frac{1-\gamma\|\mathbf{L}\mathbf{K}_{12}^{-1}\mathbf{b}_1\|^2}{1+\gamma\|\mathbf{L}\mathbf{K}_{11}^{-1}\mathbf{b}_2\|^2}}.$$

Thus, (3.12) has a solution if and only if $1 - \gamma\|\mathbf{L}\mathbf{K}_{12}^{-1}\mathbf{b}_1\|^2 > 0$.

The case where $\mathbf{K}$ is singular can be handled by taking the SVD of $\mathbf{K}$ and working in the appropriate range spaces.

(c) In case one is not able to produce a solution in either (a) or (b), it follows that the optimal solution of (3.8) is 0, and $(\mathbf{d}_\xi, \mathbf{d}_\lambda) = (\mathbf{0}, \mathbf{0})$ achieves this value. ∎

When the ActiveSet algorithm enters the `while` loop, we are guaranteed that $\boldsymbol{\xi}^* \in \mathbf{int}(\mathcal{K})$. Within the loop, one has to compute the optimal value of

$$\begin{aligned} &\max && -\mathbf{z}_0^T\mathbf{E}^T\boldsymbol{\lambda} + \mathbf{v}^T\boldsymbol{\xi} + f(\boldsymbol{\xi}) \\ &\text{subject to} && \mathcal{A}_{\mathbf{W}}[\boldsymbol{\lambda}, \boldsymbol{\xi}, \mathbf{h}, \mathbf{p}] = \mathbf{0}, \\ &&& \boldsymbol{\xi} \in \mathcal{K}\backslash\{\mathbf{0}\}, \end{aligned} \tag{3.14}$$

where $\mathbf{W}$ denotes the current inactive set, i.e., $\mathbf{W} = \sum_{i:\lambda_i=0}\mathbf{e}_i\mathbf{e}_i^T$. At this stage we have already determined that $\boldsymbol{\xi} = \mathbf{0}$ is *not* optimal for (3.1); therefore, by Lemma 3.1 it follows that we can restrict ourselves to $\boldsymbol{\xi} \in \mathbf{int}(\mathcal{K})$. The procedure $(\mathbf{d}_\xi, \mathbf{d}_\lambda, \alpha_q) =$ FindOpt$(\boldsymbol{\xi}, \boldsymbol{\lambda}, \mathbf{W})$ takes as input the current iterate and the current $\mathbf{W}$, and returns an output $(\mathbf{d}_\xi, \mathbf{d}_\lambda, \alpha_q)$ that satisfies the following:

(i) When (3.14) is bounded, $(\mathbf{d}_\xi, \mathbf{d}_\lambda) = (\boldsymbol{\xi}^*, \boldsymbol{\lambda}^*) - (\boldsymbol{\xi}, \boldsymbol{\lambda})$, where $(\boldsymbol{\xi}^*, \boldsymbol{\lambda}^*)$ is the optimal solution of (3.14), and $\alpha_q = 1$.

(ii) When (3.14) is unbounded, $(\mathbf{d}_\xi, \mathbf{d}_\lambda)$ is any recession direction of the feasible set of (3.14) satisfying $-\mathbf{z}_0^T\mathbf{E}^T\mathbf{d}_\lambda + \mathbf{v}^T\mathbf{d}_\xi + f(\mathbf{d}_\xi) > 0$ and $\alpha_q = \infty$.

When $(\mathbf{d}_\xi, \mathbf{d}_\lambda) = (\mathbf{0}, \mathbf{0})$, the ActiveSet algorithm checks the Lagrange multipliers $\boldsymbol{\rho}$ corresponding to the constraints $\mathbf{W}\boldsymbol{\lambda} = \mathbf{0}$ by calling the procedure FindMultipliers that computes the solution of

$$\begin{bmatrix}\mathbf{W}^T & \mathbf{M}^T\end{bmatrix}\begin{bmatrix}\boldsymbol{\rho} \\ \boldsymbol{\tau}\end{bmatrix} = -\left[\mathbf{E}\mathbf{z}_0 + \mathbf{L}^T\mathbf{v} + \mathbf{L}^T\nabla f(\boldsymbol{\xi}^*)\right]. \tag{3.15}$$

If the signs of all the Lagrange multipliers are consistent with the KKT conditions, i.e., $\max_i\{\rho_i\} \le 0$, the algorithm terminates; otherwise, it drops one of the constraints with the incorrect sign. Lemma 3.5 establishes that ActiveSet terminates finitely. Thus, all that remains to be shown is that FindOpt can be implemented efficiently.

Lemma 3.4. *Suppose there exists a feasible* $(\bar{\boldsymbol{\xi}}, \bar{\boldsymbol{\lambda}})$ *for* (3.14) *such that* $\bar{\boldsymbol{\xi}} \in \mathbf{int}(\mathcal{K})$. *Then* (3.14) *can be solved in closed form by solving at most three systems of linear equations.*

Proof. Let $\mathbf{d}_\xi = \boldsymbol{\xi} - \bar{\boldsymbol{\xi}}$ and $\mathbf{d}_\lambda = \boldsymbol{\lambda} - \bar{\boldsymbol{\lambda}}$. Then (3.14) is equivalent to

$$\begin{aligned} &\max && -\mathbf{z}_0^T\mathbf{E}^T\mathbf{d}_\lambda + \mathbf{v}^T\mathbf{d}_\xi + \theta\sqrt{(\mathbf{e}^T(\bar{\boldsymbol{\xi}} + \mathbf{d}_\xi))^2 - \gamma\|\bar{\boldsymbol{\xi}} + \mathbf{d}_\xi\|^2} \\ &\text{subject to} && \mathcal{A}_{\mathbf{W}}[\mathbf{d}_\lambda, \mathbf{d}_\xi, \mathbf{0}, \mathbf{0}] = \mathbf{0}, \\ &&& \bar{\boldsymbol{\xi}} + \mathbf{d}_\xi \in \mathcal{K}\backslash\{\mathbf{0}\}. \end{aligned} \tag{3.16}$$

First, suppose (3.16) is unbounded, i.e., there exists $(\mathbf{d}_\xi, \mathbf{d}_\lambda)$ such that

$$\begin{aligned}&\lim_{t\to\infty}\left\{-t\mathbf{z}_0^T\mathbf{E}^T\mathbf{d}_\lambda + t\mathbf{v}^T\mathbf{d}_\xi + \theta\sqrt{(\mathbf{e}^T(\bar{\boldsymbol{\xi}}+t\mathbf{d}_\xi))^2 - \gamma\|\bar{\boldsymbol{\xi}}+t\mathbf{d}_\xi\|^2}\right\}\\ &= \lim_{t\to\infty}\left\{-t\mathbf{z}_0^T\mathbf{E}^T\mathbf{d}_\lambda + t\mathbf{v}^T\mathbf{d}_\xi + t\theta\sqrt{(\mathbf{e}^T(\bar{\boldsymbol{\xi}}/t+\mathbf{d}_\xi))^2 - \gamma\|\bar{\boldsymbol{\xi}}/t+\mathbf{d}_\xi\|^2}\right\} = +\infty.\end{aligned}$$

Since $\bar{\boldsymbol{\xi}}/t \to 0$, it follows that (3.16) is unbounded if and only if $(\mathbf{d}_\xi, \mathbf{d}_\lambda)$ is a recession direction for

$$\begin{aligned}&-\mathbf{z}_0^T\mathbf{E}^T\mathbf{d}_\lambda + \mathbf{v}^T\mathbf{d}_\xi + \theta\sqrt{(\mathbf{e}^T\mathbf{d}_\xi)^2 - \gamma\|\mathbf{d}_\xi\|^2} > 0,\\ &\mathcal{A}_{\mathbf{W}}[\mathbf{d}_\lambda, \mathbf{d}_\xi, \mathbf{0}, \mathbf{0}] = \mathbf{0},\\ &\mathbf{d}_\xi \in \mathcal{K}.\end{aligned} \tag{3.17}$$

Since (3.17) is the same as (3.7), it follows that a positive recession direction for (3.16), if it exists, can be computed by solving at most two systems of linear equations.

Next, suppose (3.16) is bounded. By introducing a scaling parameter α, (3.16) can be reformulated as

$$\begin{aligned}\max\quad & -(\mathbf{E}\mathbf{z}_0 + \mathbf{L}^T\mathbf{v})^T\mathbf{d}_\lambda + \theta\sqrt{1-\gamma\|\alpha\bar{\boldsymbol{\xi}} - \mathbf{L}\mathbf{d}_\lambda\|^2}\\ \text{subject to}\quad & \mathcal{A}_{\mathbf{W}}[\mathbf{d}_\lambda, -\mathbf{L}\mathbf{d}_\lambda, \mathbf{0}, \mathbf{0}] = \mathbf{0},\\ & \alpha \ge 0,\\ & -\mathbf{e}^T\mathbf{L}\mathbf{d}_\lambda + \mathbf{e}^T\bar{\boldsymbol{\xi}}\alpha = 1,\\ & \gamma\|\alpha\bar{\boldsymbol{\xi}} - \mathbf{L}\mathbf{d}_\lambda\|^2 \le 1.\end{aligned}$$

Since (3.16) is bounded, i.e., it does not have any positive recession direction, we have that $\alpha^* > 0$. Also, by Lemma 3.1 it follows that $\alpha^*\bar{\boldsymbol{\xi}} + \mathbf{d}_\xi^* \in \mathbf{int}(K)$, i.e., $\gamma\|\alpha^*\bar{\boldsymbol{\xi}} - \mathbf{L}\mathbf{d}_\lambda^*\|^2 < 1$; therefore the optimal Lagrange multiplier corresponding to this constraint is zero. Consequently, the Lagrangian $\mathcal{L}$ reduces to

$$\begin{aligned}\mathcal{L} = &-(\mathbf{E}\mathbf{z}_0 + \mathbf{L}^T\mathbf{v})^T\mathbf{d}_\lambda + \theta\sqrt{1-\gamma\|\alpha\bar{\boldsymbol{\xi}} - \mathbf{L}\mathbf{d}_\lambda\|^2} - \boldsymbol{\tau}^T\mathbf{M}\mathbf{d}_\lambda - \boldsymbol{\rho}^T\mathbf{W}\mathbf{d}_\lambda\\ &- \eta(\mathbf{e}^T\alpha\bar{\boldsymbol{\xi}} - \mathbf{e}^T\mathbf{L}\mathbf{d}_\lambda - 1).\end{aligned}$$

The first-order optimality conditions are given by

$$\begin{aligned}\tfrac{\theta\gamma}{\beta}\mathbf{L}^T\mathbf{L}\mathbf{d}_\lambda - \tfrac{\theta\gamma}{\beta}\mathbf{L}^T\bar{\boldsymbol{\xi}}\alpha - \mathbf{L}^T\mathbf{e}\eta + \mathbf{M}^T\boldsymbol{\tau} + \mathbf{W}^T\boldsymbol{\rho} &= -(\mathbf{E}\mathbf{z}_0 + \mathbf{L}^T\mathbf{v}),\\ -\tfrac{\theta\gamma}{\beta}\bar{\boldsymbol{\xi}}^T\mathbf{L}\mathbf{d}_\lambda + \tfrac{\theta\gamma}{\beta}\|\bar{\boldsymbol{\xi}}\|^2\alpha + \mathbf{e}^T\bar{\boldsymbol{\xi}}\eta &= 0,\\ -\mathbf{e}^T\mathbf{L}\mathbf{d}_\lambda + \mathbf{e}^T\bar{\boldsymbol{\xi}}\alpha &= 1,\\ \mathbf{M}\mathbf{d}_\lambda &= \mathbf{0},\\ \mathbf{W}\mathbf{d}_\lambda &= \mathbf{0},\end{aligned} \tag{3.18}$$

where $\beta = \sqrt{1-\gamma\|\alpha\bar{\boldsymbol{\xi}} - \mathbf{L}\mathbf{d}_\lambda\|^2}$. Set $\bar{\boldsymbol{\rho}} = \beta\boldsymbol{\rho}$, $\bar{\boldsymbol{\tau}} = \beta\boldsymbol{\tau}$, and $\bar{\eta} = \beta\eta$. Then (3.18) is equivalent to

$$\underbrace{\begin{bmatrix}\theta\gamma\mathbf{L}^T\mathbf{L} & -\theta\gamma\mathbf{L}^T\bar{\boldsymbol{\xi}} & -\mathbf{L}^T\mathbf{e} & \mathbf{M}^T & \mathbf{W}^T\\ -\theta\gamma\bar{\boldsymbol{\xi}}^T\mathbf{L} & \theta\gamma\|\bar{\boldsymbol{\xi}}\|^2 & \mathbf{e}^T\bar{\boldsymbol{\xi}} & \mathbf{0} & \mathbf{0}\\ -\mathbf{e}^T\mathbf{L} & \mathbf{e}^T\bar{\boldsymbol{\xi}} & \mathbf{0} & \mathbf{0} & \mathbf{0}\\ \mathbf{M} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0}\\ \mathbf{W} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0}\end{bmatrix}}_{\triangleq\mathbf{K}}\begin{bmatrix}\mathbf{d}_\lambda\\ \alpha\\ \bar{\eta}\\ \bar{\boldsymbol{\tau}}\\ \bar{\boldsymbol{\rho}}\end{bmatrix} = \begin{bmatrix}\mathbf{0}\\ 0\\ 1\\ \mathbf{0}\\ \mathbf{0}\end{bmatrix} - \beta\begin{bmatrix}\mathbf{E}\mathbf{z}_0 + \mathbf{L}^T\mathbf{v}\\ 0\\ 0\\ \mathbf{0}\\ \mathbf{0}\end{bmatrix}. \tag{3.19}$$

Suppose $\mathbf{K}$ is nonsingular. Let $\hat{\mathbf{d}} = (\mathbf{d}_\lambda^T, \alpha)^T$, $\mathbf{w} = (\bar{\eta}, \bar{\boldsymbol{\tau}}^T, \bar{\boldsymbol{\rho}}^T)^T$, $\mathbf{b}_1 = (1, \mathbf{0}^T, \mathbf{0}^T)^T$, and $\mathbf{b}_2 = ((\mathbf{E}\mathbf{z}_0 + \mathbf{L}^T\mathbf{v})^T, \mathbf{0}^T)^T$. Partition $\mathbf{K}^{-1}$ such that

$$\begin{bmatrix} \hat{\mathbf{d}} \\ \mathbf{w} \end{bmatrix} = \mathbf{K}^{-1} \begin{bmatrix} \mathbf{0} \\ \mathbf{b}_1 \end{bmatrix} - \beta \mathbf{K}^{-1} \begin{bmatrix} \mathbf{b}_2 \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{K}_{12}^{-1}\mathbf{b}_1 - \beta \mathbf{K}_{11}^{-1}\mathbf{b}_2 \\ \mathbf{K}_{22}^{-1}\mathbf{b}_1 - \beta \mathbf{K}_{12}^{-T}\mathbf{b}_2 \end{bmatrix}.$$

This partition implies that $\mathbf{K}_{12}^{-T}[-\mathbf{L}, \bar{\boldsymbol{\xi}}]^T[-\mathbf{L}, \bar{\boldsymbol{\xi}}]\mathbf{K}_{11}^{-1} = \mathbf{0}$. Therefore,

$$\begin{aligned} \beta^2 + \gamma\|[-\mathbf{L}, \bar{\boldsymbol{\xi}}]\hat{\mathbf{d}}\|^2 - 1 &= \beta^2(1 + \gamma\|[-\mathbf{L}, \bar{\boldsymbol{\xi}}]\mathbf{K}_{11}^{-1}\mathbf{b}_2\|^2) - 2\beta\gamma([-\mathbf{L}, \bar{\boldsymbol{\xi}}]\mathbf{K}_{12}^{-1}\mathbf{b}_1)^T \\ &\qquad [-\mathbf{L}, \bar{\boldsymbol{\xi}}]\mathbf{K}_{11}^{-1}\mathbf{b}_2 + \gamma\|[-\mathbf{L}, \bar{\boldsymbol{\xi}}]\mathbf{K}_{12}^{-1}\mathbf{b}_1\|^2 - 1 \\ &= \beta^2(1 + \gamma\|[-\mathbf{L}, \bar{\boldsymbol{\xi}}]\mathbf{K}_{11}^{-1}\mathbf{b}_2\|^2) + \gamma\|[-\mathbf{L}, \bar{\boldsymbol{\xi}}]\mathbf{K}_{12}^{-1}\mathbf{b}_1\|^2 - 1. \end{aligned}$$

Consequently, the unique positive solution of the quadratic equation $\beta^2 = 1 - \gamma\|[-\mathbf{L}, \bar{\boldsymbol{\xi}}]\hat{\mathbf{d}}\|^2$ is

$$\beta = \sqrt{\frac{1 - \gamma\|[-\mathbf{L}, \bar{\boldsymbol{\xi}}]\mathbf{K}_{12}^{-1}\mathbf{b}_1\|^2}{1 + \gamma\|[-\mathbf{L}, \bar{\boldsymbol{\xi}}]\mathbf{K}_{11}^{-1}\mathbf{b}_2\|^2}}.$$

The case where $\mathbf{K}$ is singular can be handled by taking the SVD of $\mathbf{K}$ and working in the appropriate range spaces. ∎

In our numerical experiments we found that solving (3.19) as a least squares problem was much faster than computing the inverse or the SVD of $\mathbf{K}$.

Lemma 3.4 implies that at each iteration of the ActiveSet algorithm, we have to solve at most three systems of linear equations, namely, (3.9), (3.13), and (3.19). Next we show that the special structures of these systems of linear equalities can be leveraged to solve them more efficiently. We will demonstrate our technique on the linear system (3.13). Extensions to (3.9) and (3.19) are straightforward.

The matrix $\mathbf{K}$ in (3.13) is an $(l + n - m - r + 1 + w)$-dimensional square matrix, where $r = \mathbf{rank}\,(\mathbf{DB})$ and w is the cardinality of the current inactive set, i.e., number of rows of $\mathbf{W}$. Only the matrix $\mathbf{W}$ changes from one iteration to the next—all the other elements of $\mathbf{K}$ remain fixed. This fact can be leveraged as follows:

1. The equality $\mathbf{W}\mathbf{d}_\lambda = \mathbf{0}$ sets the components of $\mathbf{d}_\lambda$ corresponding to the current inactive set to zero. Removing these variables and dropping the corresponding rows of $\mathbf{K}$ reduces the dimension of $\mathbf{K}$ to $l+n-m-r+1$. Thus, this simple operation ensures that the size of the linear equations remains independent of the cardinality of the inactive set.
2. Let

   (3.20)
   $$\tilde{\mathbf{d}} = \begin{bmatrix} \mathbf{d}_\lambda \\ \bar{\boldsymbol{\tau}} \\ \bar{\eta} \end{bmatrix}, \quad \tilde{\mathbf{K}} = \begin{bmatrix} \theta\gamma\mathbf{L}^T\mathbf{L} & \mathbf{M}^T & \mathbf{L}^T\mathbf{e} \\ \mathbf{M} & \mathbf{0} & \mathbf{0} \\ \mathbf{e}^T\mathbf{L} & \mathbf{0} & \mathbf{0} \end{bmatrix} \in \mathbf{R}^{(l+n-m-r+1)\times(l+n-m-r+1)},$$

   let $\mathbf{B}_1$ be any orthonormal basis for row space of $\tilde{\mathbf{K}}$, and let $\mathbf{B}_2$ be any orthonormal basis for the nullspace $\mathcal{N}(\tilde{\mathbf{K}})$. Then $\tilde{\mathbf{d}} = \mathbf{B}_1\boldsymbol{\mu} + \mathbf{B}_2\boldsymbol{\zeta}$, where $\boldsymbol{\mu} \in \mathbf{R}^{r_K}$, $\boldsymbol{\zeta} \in \mathbf{R}^{l+n-m-r+1-r_K}$, and $r_K = \mathbf{rank}(\tilde{\mathbf{K}})$. An SVD-based argument similar to the one in section 2.1 (detailed in Appendix C) shows that the dimension of $\mathbf{K}$ can be reduced to $l + n - m - r + 1 - r_K + w$.

These observations suggest that one can speed up FindOpt as follows: If $(l + n - m - r + 1) < (l + n - m - r + 1 - r_K + w)$, i.e., if $w > r_K$, solve (3.13) using the first dimension reduction technique; otherwise, use the second dimension reduction.

In each iteration either new rows are added to $\mathbf{W}$ or some of the rows of $\mathbf{W}$ are dropped. Since every row of $\mathbf{W}$ is a row of an identity matrix, one can suitably adapt the revised simplex method [5] to efficiently update the iterates. For example, adding a new row to $\mathbf{W}$ forces an entry of $\mathbf{d}_\lambda$ to be equal to zero, i.e., a variable leaves the basis, and introduces a new variable through $\bar{\boldsymbol{\rho}}$, i.e., a variable enters the basis. This process, although requiring a careful bookkeeping of variables and bases, is fairly straightforward.

We conclude this section with the following finite convergence result.

LEMMA 3.5. *The* ACTIVESET *algorithm terminates after a finite number of iterations.*

Proof. Let $\mathcal{A}_j$, $j \geq 1$, denote the active set on the jth call to the procedure FINDMULTIPLIERS. Since every iteration of ACTIVESET that does not call FINDMULTIPLIERS strictly improves the objective value of (3.1), it follows that $\mathcal{A}_{j_1} \neq \mathcal{A}_{j_2}$ for all $j_1 \neq j_2$. Since the size of the active set can only increase between successive calls to FINDMULTIPLIERS, it follows that ACTIVESET terminates after, at most, $l2^l$ iterations, where l is number of inequality constraints in the single-cone SOCP (1.1). ∎

4. Recovering an optimal solution. Let $\boldsymbol{\lambda}^*$ denote the solution returned by LAGRANGEDUAL, i.e., $\boldsymbol{\lambda}^*$ is optimal for (3.1). Set $\boldsymbol{\xi}^* = \mathbf{U}_0\boldsymbol{\Sigma}_0^{-1}\mathbf{V}_0^T\mathbf{B}^T(\mathbf{f} - \mathbf{E}^T\boldsymbol{\lambda}^*)$, and using Lemma (2.5) obtain the closed form optimal solution $\mathbf{y}^*$ to $\bar{q}(\xi^*)$ defined in (2.9). Then all $\mathbf{x}^*$ satisfying

$$\mathbf{x}^* = \mathbf{x}_0 + \mathbf{B}\mathbf{z}^* = \mathbf{x}_0 + \mathbf{B}\mathbf{V}_0\boldsymbol{\Sigma}_0^{-1}\mathbf{U}_0^T(\mathbf{y}^* - \mathbf{D}\mathbf{x}_0) + \mathbf{B}\mathbf{V}_1\mathbf{t},$$

where $\mathbf{x}_0 = \mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}\mathbf{g}$, $\mathbf{t} \in \mathbf{R}^{n-m-r}$, and $r = \mathbf{rank}(\mathbf{DB})$, are optimal for (1.1). Thus, if $\mathbf{V}_1 \neq \emptyset$, i.e., $\mathbf{rank}(\mathbf{DB}) \neq n - m$, the optimal solution is not unique; in fact, an entire affine space is optimal.

5. Computational experiments. In this section we discuss the computational performance of the LAGRANGEDUAL algorithm on special classes of single-cone SOCPs that arise in the context of robust optimization.

Consider the LP

$$\begin{aligned} &\min && \mathbf{c}^T\mathbf{z} \\ &\text{subject to} && \mathbf{A}\mathbf{z} = \mathbf{b}, \\ &&& \mathbf{z} \geq \mathbf{0}, \end{aligned} \tag{5.1}$$

where $\mathbf{c}, \mathbf{z} \in \mathbf{R}^{\bar{n}}$, $\mathbf{A} \in \mathbf{R}^{\bar{m}\times\bar{n}}$, and $\mathbf{b} \in \mathbf{R}^{\bar{m}}$. Suppose the constraint matrix $(\mathbf{A}, \mathbf{b})$ is known exactly; the cost vector $\mathbf{c}$, however, is uncertain and is only known to lie within an ellipsoidal uncertainty set $\mathcal{S}$ given by

$$\mathcal{S} = \{\mathbf{c} = \mathbf{c}_0 + \mathbf{P}^T\boldsymbol{\alpha} : \boldsymbol{\alpha} \in \mathbf{R}^s, \boldsymbol{\alpha}^T\boldsymbol{\alpha} \leq 1\}.$$

We will call (5.1) an LP with uncertain cost. Such an LP is a special case of a more general class of uncertain LPs where the constraints are also uncertain [2, 3].

Let $f(\mathbf{z}) = \max_{\mathbf{c}\in\mathcal{S}}\{\mathbf{c}^T\mathbf{z}\}$ denote the worst-case cost of the decision $\mathbf{z}$. Then we have that

$$f(\mathbf{z}) = \mathbf{c}_0^T\mathbf{z} + \max_{\{\boldsymbol{\alpha}:\boldsymbol{\alpha}^T\boldsymbol{\alpha}\leq 1\}}\left\{\boldsymbol{\alpha}^T\mathbf{P}\mathbf{z}\right\} = \mathbf{c}_0^T\mathbf{z} + \|\mathbf{P}\mathbf{z}\|.$$

The robust counterpart of the uncertain LP is defined as follows [2, 3]:

$$\begin{array}{rl} \min & \mathbf{c}_0^T\mathbf{z} + \|\mathbf{P}\mathbf{z}\| \\ \text{subject to} & \mathbf{A}\mathbf{z} = \mathbf{b}, \\ & \mathbf{z} \geq \mathbf{0}. \end{array} \tag{5.2}$$

By defining,

$$\mathbf{x} = \begin{bmatrix} \mathbf{z} \\ y_0 \\ \mathbf{y} \end{bmatrix}, \quad \mathbf{H} = \begin{bmatrix} \mathbf{A} & 0 & \mathbf{0} \\ \mathbf{P} & 0 & -\mathbf{I} \end{bmatrix}, \quad \mathbf{E} = \begin{bmatrix} \mathbf{I} & 0 & \mathbf{0} \end{bmatrix},$$
$$\mathbf{D} = \begin{bmatrix} \mathbf{0} & 1 & \mathbf{0} \\ \mathbf{0} & 0 & \mathbf{I} \end{bmatrix}, \quad \mathbf{g} = \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix}, \quad \mathbf{f} = \begin{bmatrix} \mathbf{c}_0 \\ 1 \\ \mathbf{0} \end{bmatrix},$$

it is easy to see that (5.2) can be reformulated as a single-cone SOCP. The constant γ for problems of the form (5.2) is given by $\gamma = 0.5$. Thus, we are in a position to use the LagrangeDual algorithm.

All the systems of linear equations encountered during the course of the LagrangeDual algorithm were solved using the MATLAB function `mldivide` and all the computations were carried out using MATLAB R13 on a PC with a Pentium M (1.50GHz) processor and 512 MB of RAM. For moderate values of $(\bar{n}, \bar{m})$ the LP that defines $\boldsymbol{\mu}^{(1)}$ (see Figure 3.1) was solved using SeDuMi 1.05 R5. For large $(\bar{n}, \bar{m})$, $\boldsymbol{\mu}^{(1)}$ was computed using the simplex algorithm.

In the first set of experiments, the LP instances were randomly generated. In particular, the entries of matrix $\mathbf{A}$ and the cost vector $\mathbf{c}_0$ were drawn independently at random according to the uniform distribution on the unit $[0, 1]$ interval. To ensure feasibility of (5.1), the vector $\mathbf{b}$ was set to $\mathbf{b} = \mathbf{A}\mathbf{w}$, where each component of the vector $\mathbf{w}$ was generated independently at random from the uniform distribution on $[0, 1]$. The matrix $\mathbf{P}$ defining the uncertainty set $\mathcal{S}$ was set equal to the $\bar{n}$-dimensional identity matrix, and for each $(\bar{n}, \bar{m})$ pair, we generated 50 random instances.

Table 5.1 compares the running times of LagrangeDual to those of SeDuMi 1.05 R5 on the randomly generated instances. Column 3 lists the average of the ratio of running time t_{sed} of SeDuMi 1.05 R5 to running time t_{alg} of LagrangeDual, and column 4 lists the average of the ratio of t_{sed} to the running time t_{act} of ActiveSet. Note that the running time of ActiveSet is equal to the difference between the running time of LagrangeDual and the time t_{init} required to compute the initial Lagrange multipliers $(\boldsymbol{\mu}^{(1)}, \boldsymbol{\mu}^{(2)})$. The time t_{init} is listed in column 6. Columns 5 and 7 list, respectively, the average running time t_{alg} of LagrangeDual and the average number of iterations of the `while` loop in ActiveSet.

From the results displayed in Table 5.1, it is clear that the performance of the LagrangeDual algorithm (including the time spent to obtain the initial Lagrange multipliers) is superior to the SeDuMi 1.05 R5 when either

(i) the number of variables $\bar{n}$ is small, and/or
(ii) the ratio of the number of constraints to the number of variables $\bar{m}/\bar{n} \leq 0.1$ or $\bar{m}/\bar{n} \geq 0.5$.

The data in column 4 of Table 5.1 implies that the performance of LagrangeDual algorithm is superior to that of SeDuMi 1.05 R5 when the time spent to obtain the initial Lagrange multipliers is excluded. This observation suggests that the performance of LagrangeDual is likely to improve if it is initialized using a more efficient LP-solver.

Since network flow problems are a natural class of LPs where the number of variables is large but the number of constraints is reasonably small, next we tested LagrangeDual on random instances of the uncertain min-cost flow problems. The

TABLE 5.1
Running times of SeDuMi 1.05 *R*5 *and the* LAGRANGEDUAL *algorithm.*

$\bar{n}$	$\bar{m}$	$t_{\text{sed}}/t_{\text{alg}}$	$t_{\text{sed}}/t_{\text{act}}$	t_{alg}	t_{init}	Iterations
100	20	2.5880	14.0623	0.2225	0.1892	6.0400
100	40	2.1039	10.7260	0.2767	0.2230	7.7000
100	60	2.1201	6.9539	0.3435	0.2247	6.6200
100	80	2.8624	9.4862	0.3019	0.2048	3.4600
200	20	2.0144	9.8526	0.8085	0.6560	10.4400
200	50	1.0479	2.0670	1.4371	0.7446	20.3200
200	80	1.4784	2.7523	1.5546	0.7540	22.2800
200	100	1.5296	2.4418	1.7406	0.6909	32.4500
200	125	1.6418	2.2056	1.9302	0.6854	37.5400
200	150	2.6301	4.1392	1.5328	0.6821	20.5200
200	175	2.9156	4.9983	1.3913	0.6639	12.4600
300	30	1.3012	7.8696	2.4029	2.0261	13.1200
300	60	0.7355	1.5558	4.3684	2.1425	26.5800
300	90	0.8926	1.8139	4.7719	2.1576	32.7200
300	120	1.0898	2.3107	5.1750	2.1817	40.3400
300	150	1.0190	1.8005	8.3897	2.0995	75.5400
300	180	1.4281	2.9468	6.4953	2.1301	54.7800
300	210	1.3864	2.3917	5.6592	2.0499	47.4200
300	240	1.8482	3.8423	6.5621	2.1846	46.0600
300	270	2.4261	6.3807	6.6542	2.3099	36.1400
500	50	1.2789	9.5464	10.6301	9.8295	16.2400
500	100	0.6193	1.3513	18.0874	8.9554	34.3800
500	200	0.8238	1.5742	23.3625	8.9827	56.9000
500	300	1.0865	1.9030	26.2689	8.8647	74.1200
500	400	1.4921	2.6069	29.8203	9.0223	76.0400
1000	100	1.1382	9.3120	88.6851	74.3340	40.2400
1000	250	0.6622	1.1354	199.0423	76.2742	84.2800
1000	500	1.0170	1.5283	249.0415	75.6579	121.1200
1000	750	1.5285	2.5135	214.9937	75.5053	92.3800
1500	150	1.1910	14.0390	369.9276	345.9236	46.1400
1500	500	0.7092	1.0651	867.7425	283.1938	138.8200
1500	1000	0.9616	1.2590	1259.9096	280.9712	230.9400
1500	1250	1.5668	2.1655	1024.6152	269.9120	158.5800
2000	200	1.1245	10.2186	604.3456	513.5234	55.2200
2000	500	∞	∞	3456.4591	2183.3089	208.3400
5000	500	∞	∞	4067.1300	2967.4054	405.1200

TABLE 5.2
Running times of SeDuMi 1.05 *R*5 *and the* LAGRANGEDUAL *algorithm on networks.*

$\bar{n}$	$\bar{m}$	$t_{\text{sed}}/t_{\text{alg}}$	t_{alg}	Iterations
1000	100	4.2342	20.5434	53.5000
1000	150	1.7765	37.9068	65.4000
1500	150	3.6572	55.1678	74.8000
1500	250	3.0398	64.8549	88.3000
2000	330	2.4105	817.8575	94.1000

random networks were generated using the network generator developed by Goldberg [9]. Results are averaged over 10 runs for each pair $(\bar{n}, \bar{m})$.

Table 5.2 displays the results for the randomly generated network matrices. In order to be consistent with the previous set of results, we continue to denote the number of variables by $\bar{n}$ and the number of constraints by $\bar{m}$. Thus, $\bar{n}$ and $\bar{m}$ denote, respectively, the number of *arcs* and the number of *nodes* in the network. As before, column 3 lists the average of the ratio of running time t_{sed} of SeDuMi 1.05 R5 to running time t_{alg} of LAGRANGEDUAL. Columns 4 and 5 list, respectively, the average

running time of LagrangeDual and the average number of iterations of the `while` loop in ActiveSet. Since the version of LagrangeDual that we implemented did not take advantage of sparsity, in this set of experiments we did not allow SeDuMi 1.05 R5 to leverage sparsity. From the results of our computational experiments it appears that LagrangeDual is faster than SeDuMi 1.05 R5 on relatively dense networks. Also, for large networks, $\bar{n} \approx 5000$, SeDuMi 1.05 R5 failed to solve the problem but LagrangeDual did not have any trouble converging.

We also compared the performance of LagrangeDual with that of SeDuMi 1.05 R5 on some of the small problems from the NETLIB LP [15] library. All the LP instances were converted to canonical form LPs (5.1). To define the uncertainty set $\mathcal{S}$, we took the nominal cost vector $\mathbf{c}_0$ as given by the NETLIB LP library, assumed that only the nonzero elements of $\mathbf{c}_0$ are uncertain, and then defined the matrix $\mathbf{P}$ accordingly. In these experiments the performance of SeDuMi 1.05 R5 was superior to that of LagrangeDual. This is not surprising given that for most of these small problems the ratio $\bar{m}/\bar{n}$ (after the problem was converted to the canonical form) was between 0.1 and 0.6. Note that in this set of experiments we are allowing SeDuMi to exploit sparsity.

Before concluding, we would like to mention that these experiments are biased in favor of SeDuMi. As mentioned in [17] (the version updated for SeDuMi 1.05) SeDuMi "takes full advantage of sparsity," which increases its speed considerably, and it uses a dense column factorization proposed in [11]. In addition, most of the subroutines of SeDuMi are written in C code. On the other hand, the LagrangeDual algorithm was implemented using only MATLAB functions, without any special treatment of sparsity and dense columns. Indeed, since the SVD steps will destroy any sparsity in the input matrices $\mathbf{H}$ and $\mathbf{D}$, it is not clear how one could exploit sparsity to improve the run time of the LagrangeDual algorithm. Recall that when comparing the performance of LagrangeDual on robust min-cost flow problems with that of SeDuMi we had not allowed SeDuMi to exploit sparsity. When SeDuMi is allowed to take advantage of sparsity, it outperforms LagrangeDual.

Appendix A. Proofs.

A.1. Structural results for single-cone SOCPs.

Lemma A.1. *Suppose* $\mathbf{A} \in \mathbf{R}^{m\times n}$ *has full row rank and there exists* $\mathbf{d} \succeq \mathbf{0}$, $\mathbf{d} \neq \mathbf{0}$ *such that* $\mathbf{Ad} = \mathbf{0}$. *Let* $\mathbf{a}$ *denote the first column of the matrix* $\mathbf{A}$. *Then*

(a) $\gamma = \frac{1}{2} - \mathbf{a}^T(\mathbf{AA}^T)^{-1}\mathbf{a} \in [0, 0.5]$;

(b) *for all* $\mathbf{d}$ *such that* $\mathbf{Ad} = \mathbf{0}$, *we have* $\|\mathbf{d}\|^2 \geq \frac{2}{1+2\gamma}(\mathbf{e}^T\mathbf{d})^2$;

(c) $\gamma > 0 \Leftrightarrow \exists \mathbf{d} \succ \mathbf{0} : \mathbf{Ad} = \mathbf{0}$;

(d) $\gamma = 0 \Leftrightarrow \{\mathbf{d} : \mathbf{Ad} = \mathbf{0}, \mathbf{d} \succeq \mathbf{0}\} = \{\beta(\mathbf{e} - \mathbf{A}^T(\mathbf{AA}^T)^{-1}\mathbf{a}) : \beta \geq 0\}$.

Proof. Partition the matrix $\mathbf{A}$ as $\mathbf{A} = [\mathbf{a}, \bar{\mathbf{A}}]$. By scaling, we can assume that $\mathbf{d} = (1; \bar{\mathbf{d}}) \succeq \mathbf{0}$. Since $\mathbf{d} \succeq \mathbf{0}$, $\|\bar{\mathbf{d}}\| \leq 1$. Then

$$\begin{aligned}
\mathbf{AA}^T - 2\mathbf{aa}^T &= \mathbf{aa}^T + \bar{\mathbf{A}}\bar{\mathbf{A}}^T - 2\mathbf{aa}^T \\
&= \bar{\mathbf{A}}\bar{\mathbf{A}}^T - \mathbf{aa}^T \\
&= \bar{\mathbf{A}}\bar{\mathbf{A}}^T - \bar{\mathbf{A}}\bar{\mathbf{d}}\bar{\mathbf{d}}^T\bar{\mathbf{A}}^T \qquad \text{(A.1)} \\
&= \bar{\mathbf{A}}(\mathbf{I} - \bar{\mathbf{d}}\bar{\mathbf{d}}^T)\bar{\mathbf{A}}^T \succeq \mathbf{0}, \qquad \text{(A.2)}
\end{aligned}$$

where (A.1) follows from the fact that $\mathbf{Ad} = \mathbf{a} + \bar{\mathbf{A}}\bar{\mathbf{d}} = \mathbf{0}$, and (A.2) follows from the fact that $\|\bar{\mathbf{d}}\| \leq 1$. Define

$$\mathbf{M} = \begin{bmatrix} \frac{1}{2} & \mathbf{a}^T \\ \mathbf{a} & \mathbf{AA}^T \end{bmatrix}.$$

Since $\frac{1}{2} > 0$ and the Schur complement of $\frac{1}{2}$ in $\mathbf{M}$ is $\mathbf{A}\mathbf{A}^T - 2\mathbf{a}\mathbf{a}^T \succeq \mathbf{0}$, it follows that $\mathbf{M} \succeq \mathbf{0}$. Since $\mathbf{A}$ has full row rank, it follows that $\mathbf{A}\mathbf{A}^T \succ \mathbf{0}$, and the matrix $\mathbf{M} \succeq \mathbf{0}$ if and only if the Schur complement of $\mathbf{A}\mathbf{A}^T$

$$\gamma = \frac{1}{2} - \mathbf{a}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{a} \geq 0.$$

Since $(\mathbf{A}\mathbf{A}^T)^{-1} \succ \mathbf{0}$, it follows that $\gamma = \frac{1}{2} - \mathbf{a}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{a} \leq \frac{1}{2}$. This establishes part (a).

To establish the other results, consider the minimum norm problem

$$\begin{array}{rl} \min & \|\mathbf{d}\|^2 \\ \text{subject to} & \mathbf{A}\mathbf{d} = \mathbf{0}, \\ & \mathbf{e}^T\mathbf{d} = 1. \end{array} \tag{A.3}$$

The optimal solution $\mathbf{d}^*$ and the optimal value v^* of (A.3) can be obtained easily via the Lagrange multipliers technique, and is given by

$$\mathbf{d}^* = \frac{2}{1+2\gamma}\Big(\mathbf{e} - \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{a}\Big), \quad v^* = \frac{2}{1+2\gamma}.$$

Thus, it follows that for all $\mathbf{d}$ such that $\mathbf{A}\mathbf{d} = \mathbf{0}$, we have $\|\mathbf{d}\|^2 \geq \frac{2}{1+2\gamma}(\mathbf{e}^T\mathbf{d})^2$.

Since there exists $\mathbf{d} = (1; \bar{\mathbf{d}}) \succeq \mathbf{0}$ with $\mathbf{A}\mathbf{d} = \mathbf{0}$, there exists a $\mathbf{d} \succ \mathbf{0}$ with $\mathbf{A}\mathbf{d} = \mathbf{0}$ if and only if $v^* < 2$, i.e., $\gamma > 0$. Moreover when $\gamma = 0$, $\big\{\mathbf{d} : \mathbf{A}\mathbf{d} = \mathbf{0}, \mathbf{d} \succeq \mathbf{0}\big\} = \{\beta\mathbf{d}^* : \beta \geq 0\}$. □

LEMMA A.2. *Suppose* $\mathbf{A} \in \mathbf{R}^{m\times n}$ *has full row rank and consider the SOCP*

$$\begin{array}{rl} \min & \boldsymbol{\xi}^T\mathbf{y} \\ \textit{subject to} & \mathbf{A}\mathbf{y} = \mathbf{b}, \\ & \mathbf{y} \succeq \mathbf{0}. \end{array} \tag{A.4}$$

Let $\mathbf{a}$ *denote the first column of the matrix* $\mathbf{A}$*, and let* $\gamma = \frac{1}{2} - \mathbf{a}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{a} \neq 0$*. Then we have the following:*

(i) *The dual of* (A.4) *is strictly feasible for all* $\gamma < 0$*.*

(ii) *When* $\gamma > 0$*, the dual of* (A.4) *is strictly feasible if and only if* $\mathbf{e}^T\mathbf{P}\boldsymbol{\xi} > 0$ *and* $(\mathbf{e}^T\mathbf{P}\boldsymbol{\xi})^2 - \gamma\|\mathbf{P}\boldsymbol{\xi}\|^2 > 0$*, where* $\mathbf{P} = \mathbf{I} - \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}$ *denotes the orthogonal projector operator onto* $\mathcal{N}(\mathbf{A})$*.*

Proof. The dual of (A.4) is given by

$$\begin{array}{rl} \max & \mathbf{b}^T\boldsymbol{\mu} \\ \text{subject to} & \boldsymbol{\xi} - \mathbf{A}^T\boldsymbol{\mu} \succeq \mathbf{0}. \end{array}$$

Since $\mathbf{A}$ has full row rank, $\boldsymbol{\xi}$ can be written as $\boldsymbol{\xi} = \mathbf{P}\boldsymbol{\xi} + \mathbf{A}^T\mathbf{w}$ for some $\mathbf{w} \in \mathbf{R}^m$. Thus, it follows that there exists a $\boldsymbol{\mu}$ such that $\boldsymbol{\xi} - \mathbf{A}^T\boldsymbol{\mu} \succ \mathbf{0}$ if and only if there exists a $\boldsymbol{\mu}$ such that $\mathbf{P}\boldsymbol{\xi} + \mathbf{A}^T\boldsymbol{\mu} \succ \mathbf{0}$.

From the definition of the Lorentz cone, it follows that there exists a $\boldsymbol{\mu}$ such that $\mathbf{P}\boldsymbol{\xi} + \mathbf{A}^T\boldsymbol{\mu} \succ \mathbf{0}$ if and only if the optimal value of

$$\begin{array}{rl} \min & \|\alpha\mathbf{P}\boldsymbol{\xi} + \mathbf{A}^T\boldsymbol{\mu}\|^2 \\ \text{subject to} & \alpha\mathbf{e}^T\mathbf{P}\boldsymbol{\xi} + \mathbf{a}^T\boldsymbol{\mu} = 1, \\ & \alpha \geq 0 \end{array} \tag{A.5}$$

is less than 2.

First consider the case $\gamma < 0$. Note that the solution $\alpha = 0$, $\boldsymbol{\mu} = \frac{(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{a}}{\mathbf{a}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{a}}$ is feasible to (A.5) with the objective function value $\frac{1}{\mathbf{a}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{a}} = \frac{2}{1-2\gamma} < 2$.

If $\gamma > 0$, then the first part of Lemma 2.5 shows that $\boldsymbol{\xi}$ has to satisfy $\mathbf{e}^T\mathbf{P}\boldsymbol{\xi} \geq 0$ and $(\mathbf{e}^T\mathbf{P}\boldsymbol{\xi})^2 - \gamma\|\mathbf{P}\boldsymbol{\xi}\|^2 \geq 0$. Otherwise, (A.4) becomes unbounded and therefore, by the weak duality lemma for SOCPs [1], its dual is infeasible.

The rest of the analysis is very similar to the one used in the proof of Lemma A.1 and is left to the reader. □

A.2. Proof of Lemma 2.5. By definition, $\boldsymbol{\xi} \in \mathcal{D}_{\bar{q}}$ if and only if (2.9) is bounded, or, equivalently, the optimal value of the homogeneous problem

$$\begin{array}{rl} \min & \boldsymbol{\xi}^T\mathbf{d} \\ \text{subject to} & \mathbf{A}\mathbf{d} = \mathbf{0}, \\ & \mathbf{d} \succeq \mathbf{0} \end{array} \tag{A.6}$$

is nonnegative. Without loss of generality, we assume that $\boldsymbol{\xi} \in \mathcal{N}(\mathbf{A})$. Otherwise, $\boldsymbol{\xi}$ can be decomposed as $\boldsymbol{\xi} = \mathbf{P}\boldsymbol{\xi} + \boldsymbol{\xi}_1$ where $\mathbf{P}\boldsymbol{\xi} \in \mathcal{N}(\mathbf{A})$ and $\boldsymbol{\xi}_1$ belongs to the row space of $\mathbf{A}$ (the space orthogonal to $\mathcal{N}(\mathbf{A})$). Since $\mathbf{A}\mathbf{d} = \mathbf{0}$ implies $\boldsymbol{\xi}_1^T\mathbf{d} = 0$, we can drop $\boldsymbol{\xi}_1$ from the objective.

Lemma A.1(b) in Appendix A.1 establishes that $\|\mathbf{d}\|^2 \geq \frac{2}{1+2\gamma}(\mathbf{e}^T\mathbf{d})^2$ for all $\mathbf{d}$ such that $\mathbf{A}\mathbf{d} = \mathbf{0}$. Since $\mathbf{d} \succeq \mathbf{0}$ implies that $2(\mathbf{e}^T\mathbf{d})^2 \geq \|\mathbf{d}\|^2$, it follows that $\mathbf{d} = \mathbf{0}$ is the only feasible solution to (A.6) when $\gamma < 0$. Hence, $\mathcal{D}_{\bar{q}} = \mathbf{R}^p$.

Next, suppose $\gamma \geq 0$. Then (A.6) is bounded if and only if

$$\begin{array}{rl} \min & \boldsymbol{\xi}^T\mathbf{d} \\ \text{subject to} & \mathbf{A}\mathbf{d} = \mathbf{0}, \\ & \mathbf{e}^T\mathbf{d} = 1, \\ & \mathbf{d}^T\mathbf{d} \leq 2 \end{array} \tag{A.7}$$

has a nonnegative optimal value.

The Lagrangian of (A.7) is given by

$$\mathcal{L} = \boldsymbol{\xi}^T\mathbf{d} - \hat{\boldsymbol{\tau}}^T\mathbf{A}\mathbf{d} - \hat{\delta}(\mathbf{e}^T\mathbf{d} - 1) + \hat{\beta}(\mathbf{d}^T\mathbf{d} - 2),$$

where $\hat{\beta} \geq 0$. Setting the derivative $\nabla\mathcal{L} = 0$ we get

$$\mathbf{d} = -\beta\boldsymbol{\xi} + \mathbf{A}^T\boldsymbol{\tau} + \delta\mathbf{e},$$

where β, $\boldsymbol{\tau}$, and δ are rescaled values of $\hat{\beta}$, $\hat{\boldsymbol{\tau}}$, and $\hat{\delta}$; however, $\beta \geq 0$ still holds. Since $\mathbf{A}\boldsymbol{\xi} = \mathbf{0}$, the constraint $\mathbf{A}\mathbf{d} = \mathbf{0}$ yields

$$\boldsymbol{\tau} = -\delta(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{e} = -\delta(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{a}.$$

Next, the constraint $\mathbf{e}^T\mathbf{d} = 1$ implies that

$$\begin{aligned} 1 &= -\beta\mathbf{e}^T\boldsymbol{\xi} + \delta\mathbf{e}^T(\mathbf{e} - \mathbf{A}^T(\mathbf{A}\mathbf{A})^{-1}\mathbf{a}) = -\beta\mathbf{e}^T\boldsymbol{\xi} + \delta(1 - \mathbf{a}^T(\mathbf{A}\mathbf{A})^{-1}\mathbf{a}) \\ &= -\beta\mathbf{e}^T\boldsymbol{\xi} + \delta\left(\frac{1+2\gamma}{2}\right). \end{aligned}$$

Thus,

$$\mathbf{d} = -\beta\boldsymbol{\xi} + \left(\frac{2}{1+2\gamma}\right)(1 + \beta\mathbf{e}^T\boldsymbol{\xi})\Big(\mathbf{e} - \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{a}\Big).$$

From Lemma A.1(a) we have $\gamma \in [0, 0.5]$, and therefore $\mathbf{d}$ is well-defined.

Since (A.7) has a linear objective and its feasible set is the intersection of an affine set with a Euclidean ball, there exists an optimal solution to (A.7) that satisfies $\mathbf{d}^T\mathbf{d} = 2$. It is easy to see this when the matrix $[\mathbf{A}; \mathbf{e}^T]$ does not have full column rank. When $[\mathbf{A}; \mathbf{e}^T]$ has full column rank, the system $\mathbf{Ad} = \mathbf{0}$, $\mathbf{e}^T\mathbf{d} = 1$ admits the unique solution $\tilde{\mathbf{d}} = \frac{2}{1+2\gamma}(\mathbf{e} - \mathbf{A}^T(\mathbf{AA}^T)^{-1}\mathbf{a})$, which implies $\{\mathbf{d} : \mathbf{Ad} = 0,\ \mathbf{d} \succeq \mathbf{0}\} = \{t\big(\mathbf{e} - \mathbf{A}^T(\mathbf{AA}^T)^{-1}\mathbf{a}\big) : t \geq 0\}$. Then, by Lemma A.1(d) in Appendix A.1, we have $\gamma = 0$. Therefore, $\tilde{\mathbf{d}}^T\tilde{\mathbf{d}} = \frac{2}{1+2\gamma} = 2$. Simplifying the constraint $\mathbf{d}^T\mathbf{d} = 2$, we get

$$\beta^2 \left(\|\boldsymbol{\xi}\|^2 - \frac{2}{1+2\gamma}(\mathbf{e}^T\boldsymbol{\xi})^2 \right) = \frac{4\gamma}{1+2\gamma}.$$

Since $\mathbf{A}\boldsymbol{\xi} = \mathbf{0}$, Lemma A.1(b) implies that $\|\boldsymbol{\xi}\|^2 \geq \frac{2}{1+2\gamma}(\mathbf{e}^T\boldsymbol{\xi})^2$. Therefore, we only have to consider the following two cases:

(i) $(\mathbf{e}^T\boldsymbol{\xi})^2 = \frac{1+2\gamma}{2}\|\boldsymbol{\xi}\|^2$.
Suppose that $\mathbf{e}^T\boldsymbol{\xi} = \sqrt{\frac{1+2\gamma}{2}}\|\boldsymbol{\xi}\|$. Then $\boldsymbol{\xi} \succeq \mathbf{0}$, and the optimal value of (A.7) is nonnegative, or, equivalently, that (A.6) is bounded. Next, suppose $\mathbf{e}^T\boldsymbol{\xi} = -\sqrt{\frac{1+2\gamma}{2}}\|\boldsymbol{\xi}\|$. Then $\mathbf{d} = -\boldsymbol{\xi} \succeq \mathbf{0}$, and $\mathbf{d}^T\boldsymbol{\xi} = -\|\boldsymbol{\xi}\|^2 < 0$. Therefore, (A.6) is unbounded.

(ii) $(\mathbf{e}^T\boldsymbol{\xi})^2 < \frac{1+2\gamma}{2}\|\boldsymbol{\xi}\|^2$.
In this case $\beta = \sqrt{\frac{4\gamma}{(1+2\gamma)\|\boldsymbol{\xi}\|^2 - 2(\mathbf{e}^T\boldsymbol{\xi})^2}}$, and (A.7) has a nonnegative optimal value if and only if

$$\begin{aligned} 0 &\leq \boldsymbol{\xi}^T\mathbf{d} \\ &= -\beta\|\boldsymbol{\xi}\|^2 + \left(\frac{2}{1+2\gamma}\right)(1 + \beta\mathbf{e}^T\boldsymbol{\xi})(\mathbf{e}^T\boldsymbol{\xi}) \\ &= -\frac{\beta}{1+2\gamma}\Big((1+2\gamma)\|\boldsymbol{\xi}\|^2 - 2(\mathbf{e}^T\boldsymbol{\xi})^2\Big) + \frac{2(\mathbf{e}^T\boldsymbol{\xi})}{1+2\gamma}. \end{aligned}$$

Substituting the value of β and simplifying we get

$$\mathbf{e}^T\boldsymbol{\xi} \geq 0, \quad (\mathbf{e}^T\boldsymbol{\xi})^2 \geq \gamma\|\boldsymbol{\xi}\|^2.$$

Since, as we discussed above, assuming $\boldsymbol{\xi} \in \mathcal{N}(\mathbf{A})$ is equivalent to replacing $\boldsymbol{\xi}$ by $\mathbf{P}\boldsymbol{\xi}$, the result follows.

For the second part of Lemma 2.5, first consider the case $\gamma \neq 0$, or, equivalently, $\mathbf{ARA}^T$ is nonsingular [1]. Using the results of the first part of this lemma, one can prove (see Lemma A.2 in Appendix A.1) that if $\gamma < 0$, then the dual of (A.6) is strictly feasible for any $\boldsymbol{\xi} \in \mathbf{R}^p$, and when $\gamma > 0$ the dual of (A.6) is strictly feasible if and only if $\mathbf{e}^T\mathbf{P}\boldsymbol{\xi} \geq 0$ and $(\mathbf{e}^T\mathbf{P}\boldsymbol{\xi})^2 - \gamma\|\mathbf{P}\boldsymbol{\xi}\|^2 > 0$; and from [1, section 5] it follows that when the dual is strictly feasible,

$$\bar{q}(\boldsymbol{\xi}) = \frac{\sqrt{-\gamma(\mathbf{b}^T(\mathbf{ARA}^T)^{-1}\mathbf{b})}}{\gamma}\sqrt{(\mathbf{e}^T\mathbf{P}\boldsymbol{\xi})^2 - \gamma\|\mathbf{P}\boldsymbol{\xi}\|^2} + \boldsymbol{\xi}^T\mathbf{RA}^T(\mathbf{ARA}^T)^{-1}\mathbf{b}.$$

When the dual is not strictly feasible, i.e., $(\mathbf{e}^T\mathbf{P}\boldsymbol{\xi})^2 = \gamma\|\mathbf{P}\boldsymbol{\xi}\|^2$, choose $\hat{\boldsymbol{\xi}} \in \mathcal{D}_{\bar{q}}$ such that the dual corresponding to $\hat{\boldsymbol{\xi}}$ is strictly feasible. For $0 < \epsilon \leq 1$, let $\boldsymbol{\xi}_\epsilon = (1-\epsilon)\boldsymbol{\xi} + \epsilon\hat{\boldsymbol{\xi}}$. Then we have two cases:

(i) $\mathbf{P}\boldsymbol{\xi} = \mathbf{0}$. In this case, $(\mathbf{e}^T\mathbf{P}\boldsymbol{\xi}_\epsilon)^2 - \gamma\|\mathbf{P}\boldsymbol{\xi}_\epsilon\|^2 = \epsilon^2\big((\mathbf{e}^T\mathbf{P}\hat{\boldsymbol{\xi}})^2 - \gamma\|\mathbf{P}\hat{\boldsymbol{\xi}}\|^2\big) > 0$.

(ii) $\mathbf{P}\boldsymbol{\xi} \neq \mathbf{0}$. In this case, $\gamma > 0$. Since $\mathbf{e}^T\mathbf{P}\boldsymbol{\xi} - \sqrt{\gamma}\|\mathbf{P}\boldsymbol{\xi}\|$ is a concave function of $\boldsymbol{\xi}$, it follows that

$$\begin{aligned}\mathbf{e}^T\mathbf{P}\boldsymbol{\xi}_\epsilon - \sqrt{\gamma}\|\mathbf{P}\boldsymbol{\xi}_\epsilon\| &\geq \epsilon\big(\mathbf{e}^T\mathbf{P}\hat{\boldsymbol{\xi}} - \sqrt{\gamma}\|\mathbf{P}\hat{\boldsymbol{\xi}}\|\big) + (1-\epsilon)\big(\mathbf{e}^T\mathbf{P}\boldsymbol{\xi} - \sqrt{\gamma}\|\mathbf{P}\boldsymbol{\xi}\|\big)\\ &= \epsilon\big(\mathbf{e}^T\mathbf{P}\hat{\boldsymbol{\xi}} - \sqrt{\gamma}\|\mathbf{P}\hat{\boldsymbol{\xi}}\|\big) > 0.\end{aligned}$$

Thus, the dual corresponding to $\boldsymbol{\xi}_\epsilon$ is always strictly feasible and

$$\bar{q}(\boldsymbol{\xi}_\epsilon) = f(\mathbf{P}\boldsymbol{\xi}_\epsilon) + \mathbf{v}^T\boldsymbol{\xi}_\epsilon.$$

Taking the limit as $\epsilon \downarrow 0$ establishes the result.

Next, consider the case $\gamma = 0$, or, equivalently, $\mathbf{A}\mathbf{R}\mathbf{A}^T$ is singular. Note that

$$\bar{q}(\boldsymbol{\xi}) = \boldsymbol{\xi}^T\mathbf{y}_0 + \hat{q}(\mathbf{P}\boldsymbol{\xi}),$$

where $\mathbf{y}_0 = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{b}$, and

$$\hat{q}(\mathbf{P}\boldsymbol{\xi}) = \begin{array}{rl}\min & (\mathbf{P}\boldsymbol{\xi})^T\mathbf{w}\\ \text{subject to} & \mathbf{A}\mathbf{w} = \mathbf{0},\\ & \mathbf{y}_0 + \mathbf{w} \succeq \mathbf{0}.\end{array} \tag{A.8}$$

The following are easy to check linear algebra facts:

(i) $\gamma = 0 \Rightarrow 2(\mathbf{e}^T\mathbf{y}_0)^2 \leq \|\mathbf{y}_0\|^2$.

(ii) $\gamma = 0 \Rightarrow 2(\mathbf{e}^T\mathbf{w})^2 \leq \|\mathbf{w}\|^2$ for all $\mathbf{w} \in \mathcal{N}(\mathbf{A})$. In particular, $\|\mathbf{P}\boldsymbol{\xi}\|^2 \geq 2(\mathbf{e}^T\mathbf{P}\boldsymbol{\xi})^2$.

We solve (A.8) by first scaling it to reduce it to a minimum norm problem and then optimizing over the scaling factor. Let $\mathbf{w}^*$ denote the optimal solution of (A.8) and let $\alpha^* = \mathbf{e}^T(\mathbf{y}_0 + \mathbf{w}^*)$. Then

$$\|\mathbf{w}^* + \mathbf{y}_0\|^2 = \|\mathbf{w}^*\|^2 + \|\mathbf{y}_0\|^2 \leq 2(\alpha^*)^2,$$

where the equality follows from the fact that $\mathbf{y}_0^T\mathbf{w}^* = \mathbf{b}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{w}^* = 0$. It follows that $\mathbf{w}^*$ is the optimal solution of

$$\begin{array}{rl}\min & (\mathbf{P}\boldsymbol{\xi})^T\mathbf{w}\\ \text{subject to} & \mathbf{A}\mathbf{w} = 0,\\ & \mathbf{e}^T\mathbf{w} = \alpha - \mathbf{e}^T\mathbf{y}_0,\\ & \|\mathbf{w}\|^2 \leq 2\alpha^2 - \|\mathbf{y}_0\|^2,\end{array} \tag{A.9}$$

with α set equal to α^*. Using Lagrange multipliers, the optimal value of (A.9) is given by

$$\begin{aligned}(\mathbf{P}\boldsymbol{\xi})^T\mathbf{w}_\alpha = &-\sqrt{\|\mathbf{P}\boldsymbol{\xi}\|^2 - 2(\mathbf{e}^T\mathbf{P}\boldsymbol{\xi})^2}\sqrt{4\alpha(\mathbf{e}^T\mathbf{y}_0) - \|\mathbf{y}_0\|^2 - 2(\mathbf{e}^T\mathbf{y}_0)^2}\\ &+ 2(\mathbf{e}^T\mathbf{P}\boldsymbol{\xi})(\alpha - \mathbf{e}^T\mathbf{y}_0).\end{aligned} \tag{A.10}$$

Differentiating this expression with respect to α we get

$$4\alpha^*(\mathbf{e}^T\mathbf{y}_0) - \|\mathbf{y}_0\|^2 - 2(\mathbf{e}^T\mathbf{y}_0)^2 = \left(\frac{\mathbf{e}^T\mathbf{y}_0}{\mathbf{e}^T\mathbf{P}\boldsymbol{\xi}}\right)^2\left(\|\mathbf{P}\boldsymbol{\xi}\|^2 - 2(\mathbf{e}^T\mathbf{P}\boldsymbol{\xi})^2\right).$$

It is easy to check that $2(\alpha^*)^2 \geq \|\mathbf{y}_0\|^2$. Substituting α^* into (A.10) and simplifying we get

$$\begin{aligned}\bar{q}(\boldsymbol{\xi}) &= \boldsymbol{\xi}^T\mathbf{y}_0 + (\mathbf{P}\boldsymbol{\xi})^T\mathbf{w}_{\alpha^*}\\ &= \boldsymbol{\xi}^T\mathbf{y}_0 + \left(\frac{\|\mathbf{y}_0\|^2 - 2(\mathbf{e}^T\mathbf{y}_0)^2}{2\mathbf{e}^T\mathbf{y}_0}\right)\mathbf{e}^T\mathbf{P}\boldsymbol{\xi} - \mathbf{e}^T\mathbf{y}_0\left(\frac{\|\mathbf{P}\boldsymbol{\xi}\|^2 - 2(\mathbf{e}^T\mathbf{P}\boldsymbol{\xi})^2}{2\mathbf{e}^T\mathbf{P}\boldsymbol{\xi}}\right).\end{aligned}$$

Appendix B. Analysis for the case rank(DB) = $p < n - m$. Note that in this case $\mathbf{A} = \mathbf{U}_1^T = \emptyset$, so $\gamma = \frac{1}{2}$ and $\mathbf{P} = \mathbf{I}$. The following lemma is easy to prove.

LEMMA B.1. *Let $\hat{q} : \mathbf{R}^p \mapsto \mathbf{R}$ denote the function defined in* (2.16). *Then the domain $\mathcal{D}_{\hat{q}} = \{\boldsymbol{\xi} : \hat{q}(\boldsymbol{\xi}) > -\infty\}$ is given by*

$$\mathcal{D}_{\hat{q}} = \Big\{\boldsymbol{\xi} : \mathbf{e}^T\boldsymbol{\xi} \geq 0, 2(\mathbf{e}^T\boldsymbol{\xi})^2 - \|\boldsymbol{\xi}\|^2 \geq 0\Big\}, \tag{B.1}$$

where $\mathbf{e} = (1, \mathbf{0}^T)^T$. For all $\boldsymbol{\xi} \in \mathcal{D}_{\hat{q}}$, we have $\hat{q}(\boldsymbol{\xi}) = 0$.

Then (2.14), (2.15), (2.16), (2.17), and Lemma B.1 imply that the Lagrangian dual problem is given by

$$\begin{aligned} \max \quad & (\mathbf{f} - \mathbf{E}^T\boldsymbol{\lambda})^T\mathbf{z}_0 \\ \text{subject to} \quad & \mathcal{A}[\boldsymbol{\lambda}, \boldsymbol{\xi}, \mathbf{h}, \mathbf{p}] = \mathbf{0}, \\ & \boldsymbol{\lambda} \geq \mathbf{0}, \\ & \boldsymbol{\xi} \in \mathcal{K}, \end{aligned} \tag{B.2}$$

where $\mathcal{A}[\boldsymbol{\lambda}, \boldsymbol{\xi}, \mathbf{h}, \mathbf{p}]$ denotes the set of linear equalities in (3.1) and $\mathcal{K} = \{\mathbf{z} : \mathbf{e}^T\mathbf{z} \geq 0, 2(\mathbf{e}^T\mathbf{z})^2 - \|\mathbf{z}\|^2 \geq 0\}$.

As in the case discussed in the paper, first set $\boldsymbol{\xi} = \mathbf{0}$ and solve (B.2). Let $\boldsymbol{\mu}^{(1)}$ be its optimal solution. A direction $(\mathbf{d}_\xi, \mathbf{d}_\lambda)$ is an ascent direction at $(\mathbf{0}, \boldsymbol{\mu}^{(1)})$ if and only if $(\mathbf{d}_\xi, \mathbf{d}_\lambda)$ is a recession direction of the set

$$\begin{aligned} -\mathbf{z}_0^T\mathbf{E}^T\mathbf{d}_\lambda > 0, \\ \mathcal{A}_{\mathbf{W}}[\mathbf{d}_\lambda, \mathbf{d}_\xi, \mathbf{0}, \mathbf{0}] = \mathbf{0}, \\ \mathbf{d}_\xi \in \mathcal{K}, \end{aligned} \tag{B.3}$$

with the matrix $\mathbf{W} = \mathbf{0}$.

LEMMA B.2. *A recession direction $(\mathbf{d}_\xi, \mathbf{d}_\lambda)$ of* (B.3)*, if it exists, can be computed by solving two systems of linear equations.*

Proof. We will find a recession direction of (B.3) by solving the following problem:

$$\begin{aligned} \max \quad & -\mathbf{z}_0^T\mathbf{E}^T\mathbf{d}_\lambda \\ \text{subject to} \quad & \mathcal{A}_{\mathbf{W}}[\mathbf{d}_\lambda, \mathbf{d}_\xi, \mathbf{0}, \mathbf{0}] = \mathbf{0}, \\ & \mathbf{d}_\xi \in \mathcal{K}. \end{aligned} \tag{B.4}$$

If the optimal value of this problem is positive, then (B.3) has a recession direction. The direction $(\mathbf{d}_\xi, \mathbf{d}_\lambda)$ can be computed by considering the following three cases:

(a) Suppose $\mathbf{d}_\xi = \mathbf{0}$. Then $(\mathbf{0}, \mathbf{d}_\lambda)$ is a recession direction for (B.3) if and only if $\mathbf{d}_\lambda$ solves

$$\begin{aligned} -\mathbf{z}_0^T\mathbf{E}^T\mathbf{d}_\lambda = 1, \\ \mathcal{A}_{\mathbf{W}}[\mathbf{d}_\lambda, \mathbf{0}, \mathbf{0}, \mathbf{0}] = \mathbf{0}. \end{aligned} \tag{B.5}$$

(b) Next, suppose (B.5) is infeasible; however, there exists a positive recession direction for (B.4). Set $\mathbf{e}^T\mathbf{d}_\xi = 1$ in (B.4) to obtain

$$\begin{aligned} \max \quad & -\mathbf{z}_0^T\mathbf{E}^T\mathbf{d}_\lambda \\ \text{subject to} \quad & \mathcal{A}_{\mathbf{W}}[\mathbf{d}_\lambda, \mathbf{d}_\xi, \mathbf{0}, \mathbf{0}] = \mathbf{0}, \\ & \mathbf{e}^T\mathbf{d}_\xi = 1, \\ & \|\mathbf{d}_\xi\|^2 \leq 2. \end{aligned} \tag{B.6}$$

Since (B.5) is assumed to be infeasible, (B.6) is bounded. Setting $\mathbf{d}_\xi = -\mathbf{L}\mathbf{d}_\lambda$, we get

$$\begin{aligned} \max \quad & -\big(\mathbf{E}\mathbf{z}_0\big)^T \mathbf{d}_\lambda \\ \text{subject to} \quad & \mathcal{A}_{\mathbf{W}}[\mathbf{d}_\lambda, -\mathbf{L}\mathbf{d}_\lambda, \mathbf{0}, \mathbf{0}] = \mathbf{0}, \\ & \mathbf{e}^T\mathbf{L}\mathbf{d}_\lambda = -1, \\ & \|\mathbf{L}\mathbf{d}_\lambda\|^2 \le 2. \end{aligned}$$

Since the objective function of this problem is linear, the optimal $\mathbf{d}_\lambda^*$ satisfies $\|\mathbf{L}\mathbf{d}_\lambda^*\|^2 = 2$ and the Lagrangian function $\mathcal{L}$ is given by

$$\mathcal{L} = -(\mathbf{E}\mathbf{z}_0)^T\mathbf{d}_\lambda - \boldsymbol{\tau}^T\mathbf{M}\mathbf{d}_\lambda - \boldsymbol{\rho}^T\mathbf{W}\mathbf{d}_\lambda - \eta(\mathbf{e}^T\mathbf{L}\mathbf{d}_\lambda + 1) - \beta(\|\mathbf{L}\mathbf{d}_\lambda\|^2 - 2),$$

where $\beta \ge 0$ and the first-order optimality conditions are given by

$$\begin{aligned} 2\beta\mathbf{L}^T\mathbf{L}\mathbf{d}_\lambda + \mathbf{M}^T\boldsymbol{\tau} + \mathbf{W}^T\boldsymbol{\rho} + \mathbf{L}^T\mathbf{e}\eta &= -\mathbf{E}\mathbf{z}_0, \\ \mathbf{M}\mathbf{d}_\lambda &= \mathbf{0}, \\ \mathbf{W}\mathbf{d}_\lambda &= \mathbf{0}, \\ \mathbf{e}^T\mathbf{L}\mathbf{d}_\lambda &= -1, \end{aligned} \tag{B.7}$$

and $\beta(\|\mathbf{L}\mathbf{d}_\lambda\|^2 - 2) = 0$. If $\beta = 0$, then (B.7) can be solved easily. Suppose $\beta > 0$. Then by setting $\bar{\boldsymbol{\rho}} = \frac{1}{\beta}\boldsymbol{\rho}$, $\bar{\boldsymbol{\tau}} = \frac{1}{\beta}\boldsymbol{\tau}$, and $\bar{\eta} = \frac{1}{\beta}\eta$, we see that (B.7) is equivalent to

$$\underbrace{\begin{bmatrix} 2\mathbf{L}^T\mathbf{L} & \mathbf{M}^T & \mathbf{W}^T & \mathbf{L}^T\mathbf{e} \\ \mathbf{M} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{W} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{e}^T\mathbf{L} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}}_{\triangleq \mathbf{K}} \begin{bmatrix} \mathbf{d}_\lambda \\ \bar{\boldsymbol{\tau}} \\ \bar{\boldsymbol{\rho}} \\ \bar{\eta} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ -1 \end{bmatrix} - \frac{1}{\beta}\begin{bmatrix} \mathbf{E}\mathbf{z}_0 \\ \mathbf{0} \\ \mathbf{0} \\ 0 \end{bmatrix}.$$

Suppose $\mathbf{K}$ is nonsingular. Let $\mathbf{w} = (\bar{\boldsymbol{\tau}}^T, \bar{\boldsymbol{\rho}}^T, \bar{\eta})^T$, $\mathbf{b}_1 = (\mathbf{0}^T, \mathbf{0}^T, -1)^T$, and $\mathbf{b}_2 = \mathbf{E}\mathbf{z}_0$. Partition $\mathbf{K}^{-1}$ into submatrices

$$\mathbf{K}^{-1} = \begin{bmatrix} \mathbf{K}_{11}^{-1} & \mathbf{K}_{12}^{-1} \\ \mathbf{K}_{12}^{-T} & \mathbf{K}_{22}^{-1} \end{bmatrix}$$

such that

$$\begin{bmatrix} \mathbf{d}_\lambda \\ \mathbf{w} \end{bmatrix} = \mathbf{K}^{-1}\begin{bmatrix} \mathbf{0} \\ \mathbf{b}_1 \end{bmatrix} - \frac{1}{\beta}\mathbf{K}^{-1}\begin{bmatrix} \mathbf{b}_2 \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{K}_{12}^{-1}\mathbf{b}_1 - \frac{1}{\beta}\mathbf{K}_{11}^{-1}\mathbf{b}_2 \\ \mathbf{K}_{22}^{-1}\mathbf{b}_1 - \frac{1}{\beta}\mathbf{K}_{12}^{-T}\mathbf{b}_2 \end{bmatrix}.$$

This partition implies that $\mathbf{K}_{12}^{-T}\mathbf{L}^T\mathbf{L}\mathbf{K}_{11}^{-1}$. Therefore, β is the unique positive root of

$$\begin{aligned} 2 &= \|\mathbf{L}\mathbf{d}_\lambda\|^2 \\ &= \|\mathbf{L}\mathbf{K}_{12}^{-1}\mathbf{b}_1\|^2 - 2\frac{1}{\beta}(\mathbf{L}\mathbf{K}_{12}^{-1}\mathbf{b}_1)^T\mathbf{L}\mathbf{K}_{11}^{-1}\mathbf{b}_2 + \frac{1}{\beta^2}\|\mathbf{L}\mathbf{K}_{11}^{-1}\mathbf{b}_2\|^2 \\ &= \|\mathbf{L}\mathbf{K}_{12}^{-1}\mathbf{b}_1\|^2 + \frac{1}{\beta^2}\|\mathbf{L}\mathbf{K}_{11}^{-1}\mathbf{b}_2\|^2. \end{aligned}$$

Consequently,

$$\beta = \sqrt{\frac{\|\mathbf{L}\mathbf{K}_{11}^{-1}\mathbf{b}_2\|^2}{2 - \|\mathbf{L}\mathbf{K}_{12}^{-1}\mathbf{b}_1\|^2}}.$$

Thus, (B.7) has a solution if and only if $2 - \|\mathbf{L}\mathbf{K}_{12}^{-1}\mathbf{b}_1\|^2 > 0$.
The case where $\mathbf{K}$ is singular can be handled by taking an SVD of $\mathbf{K}$ and working in the appropriate range spaces.

(c) In case one is not able to produce a solution in either (a) or (b), it follows that the optimal solution of (B.4) is 0, and $(\mathbf{d}_\xi, \mathbf{d}_\lambda) = (\mathbf{0}, \mathbf{0})$ achieves this value. □

In a typical step of the ActiveSet when $\mathbf{rank}(\mathbf{DB}) = p$, we have to solve the following problem:

$$\begin{aligned} \max \quad & -\mathbf{z}_0^T \mathbf{E}^T \boldsymbol{\lambda} \\ \text{subject to} \quad & \mathcal{A}_{\mathbf{W}}[\boldsymbol{\lambda}, \boldsymbol{\xi}, \mathbf{h}, \mathbf{p}] = \mathbf{0}, \\ & \boldsymbol{\xi} \in \mathcal{K}, \end{aligned} \tag{B.8}$$

where $\mathbf{W}$ denotes the current inactive set, i.e., $\mathbf{W} = \sum_{i:\lambda_i=0} \mathbf{e}_i \mathbf{e}_i^T$.

Lemma B.3. *Suppose there exists a feasible* $(\bar{\boldsymbol{\xi}}, \bar{\boldsymbol{\lambda}})$ *for* (B.8) *such that* $\bar{\boldsymbol{\xi}} \in \mathbf{int}(\mathcal{K})$. *Then* (B.8) *can be solved in closed form by solving at most three systems of linear equations.*

This result can be established using a combination of the techniques used to establish Lemmas B.2 and 3.4.

Appendix C. Decreasing dimensions of K. Consider the system of linear equalities

$$\begin{aligned} \tilde{\mathbf{K}}\tilde{\mathbf{d}} + \tilde{\mathbf{W}}^T \bar{\boldsymbol{\rho}} &= \mathbf{b}, \\ \tilde{\mathbf{W}}\tilde{\mathbf{d}} &= \mathbf{0}, \end{aligned} \tag{C.1}$$

where $\tilde{\mathbf{K}}$ and $\tilde{\mathbf{d}}$ are given in (3.20) and $\tilde{\mathbf{W}} = [\mathbf{W}, \mathbf{0}, \mathbf{0}]$. Let an SVD of $\tilde{\mathbf{K}}$ be given by

$$\tilde{\mathbf{K}} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T = \mathbf{U} \begin{bmatrix} \boldsymbol{\Sigma}_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{V}_0^T \\ \mathbf{V}_1^T \end{bmatrix},$$

where $\boldsymbol{\Sigma}_0 \in \mathbf{R}^{r_K \times r_K}$ is a diagonal matrix and $r_K = \mathbf{rank}(\tilde{\mathbf{K}})$. Decompose $\tilde{\mathbf{d}} = \mathbf{V}_0\boldsymbol{\mu} + \mathbf{V}_1\boldsymbol{\zeta}$. Then (C.1) is equivalent to

$$\begin{aligned} \mathbf{U} \begin{bmatrix} \boldsymbol{\Sigma}_0\boldsymbol{\mu} \\ \mathbf{0} \end{bmatrix} + \tilde{\mathbf{W}}^T \bar{\boldsymbol{\rho}} &= \mathbf{b}, \\ \tilde{\mathbf{W}}(\mathbf{V}_0\boldsymbol{\mu} + \mathbf{V}_1\boldsymbol{\zeta}) &= \mathbf{0}, \end{aligned}$$

which is equivalent to

$$\begin{aligned} \begin{bmatrix} \boldsymbol{\Sigma}_0\boldsymbol{\mu} \\ \mathbf{0} \end{bmatrix} + \mathbf{U}^T\tilde{\mathbf{W}}^T \bar{\boldsymbol{\rho}} &= \mathbf{U}^T\mathbf{b}, \\ \tilde{\mathbf{W}}(\mathbf{V}_0\boldsymbol{\mu} + \mathbf{V}_1\boldsymbol{\zeta}) &= \mathbf{0}. \end{aligned} \tag{C.2}$$

Let

$$\mathbf{U}^T = \begin{bmatrix} \mathbf{U}_0^T \\ \mathbf{U}_1^T \end{bmatrix}.$$

Then (C.2) is equivalent to

$$\begin{aligned} \boldsymbol{\Sigma}_0\boldsymbol{\mu} + \mathbf{U}_0^T\tilde{\mathbf{W}}^T \bar{\boldsymbol{\rho}} &= \mathbf{U}_0^T\mathbf{b}, \\ \mathbf{U}_1^T\tilde{\mathbf{W}}^T \bar{\boldsymbol{\rho}} &= \mathbf{U}_1^T\mathbf{b}, \\ \tilde{\mathbf{W}}\mathbf{V}_0\boldsymbol{\mu} + \tilde{\mathbf{W}}\mathbf{V}_1\boldsymbol{\zeta} &= \mathbf{0}. \end{aligned}$$

Setting $\boldsymbol{\mu} = \boldsymbol{\Sigma}_0^{-1}(\mathbf{U}_0^T\mathbf{b} - \mathbf{U}_0^T\tilde{\mathbf{W}}^T\bar{\boldsymbol{\rho}})$, we obtain the following system which has a smaller number of variables:

$$\mathbf{U}_1^T\tilde{\mathbf{W}}^T\bar{\boldsymbol{\rho}} = \mathbf{U}_1^T\mathbf{b},$$
$$\tilde{\mathbf{W}}\mathbf{V}_1\boldsymbol{\zeta} - \tilde{\mathbf{W}}\mathbf{V}_0\boldsymbol{\Sigma}_0^{-1}\mathbf{U}_0^T\tilde{\mathbf{W}}^T\bar{\boldsymbol{\rho}} = -\tilde{\mathbf{W}}\mathbf{V}_0\boldsymbol{\Sigma}_0^{-1}\mathbf{U}_0^T\mathbf{b}.$$

Acknowledgments. The authors would like to thank the anonymous referees and the associate editor, Michael J. Todd, for constructive comments and suggestions.

REFERENCES

[1] F. Alizadeh and D. Goldfarb, *Second-order cone programming*, Math. Program., 95 (2003), pp. 3–51.
[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*, SIAM, Philadelphia, 2001.
[5] D. Bertsimas and J. N. Tsitsiklis, *Introduction to Linear Optimization*, Athena Scientific, Nashua, NH, 1997.
[6] S. P. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004.
[7] G. Dantzig, *Linear Programming and Extensions*, Princeton University Press, Princeton, NJ, 1963.
[8] N. Gartner, C. J. Messer, and A. K. Rathi, *Traffic Flow Theory: A State of the Art Report*, technical report, Turner Fairbank Highway Research Center (TFHRC), McLean, VA, 1997; available online from http://www.tfhrc.gov/its/tft/tft.htm.
[9] A. Goldberg, *C-Code for a Random Network Generator*, available via anonymous ftp from ftp://dimacs.rutgers.edu/pub/netflow/generators/network/grid-on-torus/goto.c.
[10] D. Goldfarb and G. Iyengar, *Robust portfolio selection problems*, Math. Oper. Res., 28 (2003), pp. 1–37.
[11] D. Goldfarb and K. Scheinberg, *A product-form Cholesky factorization method for handling dense columns in interior point methods for linear programming*, Math. Program., 99 (2004), pp. 1–34.
[12] W. H. Greene, *Econometric Analysis*, Macmillan, New York, 1990.
[13] M. Muramatsu, *A pivoting procedure for a class of second-order cone programming*, Optim. Methods Softw., 21 (2006), pp. 295–315.
[14] Y. Nesterov and A. Nemirovski, *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM, Philadelphia, 1994.
[15] *NETLIB LP Test Problem Set*, available from http://cuter.rl.ac.uk/cuter-www/Problems/netlib.shtml.
[16] A. Ruszczynski and A. Shapiro, eds., *Stochastic Programming*, Handbooks Oper. Res. Management Sci. 10, Elsevier, Amsterdam, The Netherlands, 2003.
[17] J. Sturm, *Using SeDuMi* 1.02*, a MATLAB toolbox for optimization over symmetric cones*, Optim. Methods Softw., 11–12 (1999), pp. 625–653.

SIAM J. OPTIM.
Vol. 17, No. 2, pp. 485–499
© 2006 Society for Industrial and Applied Mathematics

ON SOLUTIONS TO THE MASS TRANSFER PROBLEM*

JUAN GONZÁLEZ-HERNÁNDEZ†, J. RIGOBERTO GABRIEL‡, AND ONÉSIMO HERNÁNDEZ-LERMA§

Abstract. This paper studies the Monge–Kantorovich mass transfer (MT) problem on metric spaces, with possibly unbounded "cost" function. Conditions are given under which the MT problem is solvable and, furthermore, an optimal solution can be obtained as the weak limit of a sequence of optimal solutions to suitably approximating MT problems.

Key words. mass transfer problem, Monge–Kantorovich problem, approximation of optimization problems

AMS subject classifications. 90C48, 90C31

DOI. 10.1137/050623991

1. Introduction.

In this paper we study the Monge–Kantorovich mass transfer (MT) problem (introduced in section 2; see (2.1), (2.2)) on metric spaces X and Y, and with a possibly unbounded real-valued "cost" function c on $X \times Y$. We propose *two* new approaches to show that MT is solvable. First in section 4 we show that MT is *equivalent* to a problem MTK on a set of randomized strategies, also known as stochastic kernels or Young measures [5, 13, 14]. Then in section 5 we prove that MTK is solvable and, therefore, so is MT. Moreover, in section 6, we introduce an *approximating approach* to MT by means of mass transfer problems MTi with marginals (see ν_1, ν_2 in (2.2)) concentrated on *finite sets*. We show that the sequence of optimal solutions to MTi and the sequence of the corresponding optimal values converge to an optimal solution to MT and to the optimal value, respectively.

The MT problem is among the oldest and most well-known problems in probability theory and its applications. It was originally introduced by Monge in 1781 [19], but it was posed as a mathematical programming problem by Kantorovich in 1942 [17]. Kantorovich considered the case of compact metric spaces $X = Y$ and cost $c(x, y) := d(x, y)$, the distance function on X. He proved that the problem is solvable in this case. In the 1970s, Levin and Milyutin (see [21, 22, 23] for earlier references) proved solvability for compact metric spaces and some classes of discontinuous cost functions. In fact, several authors have studied the MT problem with a *lower semicontinuous* (l.s.c.) cost function. In particular, Jiménez–Guerra and Rodríguez–Salinas [16] dealt with Hausdorff topological spaces X and Y, Radon measures, and an l.s.c. cost function. Hernández–Lerma and Gabriel [15] proved solvability of the MT problem in general metric spaces, and an *inf-compact* cost function, which means that the lower level set $\{(x, y) \in X \times Y \mid c(x, y) \le r\}$ is compact for each $r \in \mathbb{R}$. Other

*Received by the editors February 4, 2005; accepted for publication (in revised form) February 20, 2006; published electronically July 17, 2006. This research was partially supported by CONACYT grant 45693-F.

http://www.siam.org/journals/siopt/17-2/62399.html

†Departamento de Probabilidad y Estadística, IIMAS-UNAM, A. Postal 20-726, México D.F. 01000, México (juan@sigma.iimas.unam.mx).

‡Facultad de Matemáticas, UV, A. Postal 270, Xalapa, Ver. 91090, México (jgabriel@uv.mx).
This author's research was partially supported by PROMEP: UVER-EXB-01-01.
§Corresponding author. Departamento de Matemáticas, CINVESTAV-IPN, A. Postal 14-740, México D.F. 07000, México (ohernand@math.cinvestav.mx).

solvability results when X and Y are compact spaces and $c(x, y)$ is l.s.c. are discussed in [23, Chapter 4] and [2, Chapter 5].

In addition to solvability, several other aspects of the MT problem have been studied by many authors using different approaches. For instance, Wu [26] studied the structure of extreme points for the MT problem with X and Y compact Hausdorff spaces. Cuesta-Albertos and Tuero-Díaz [10] characterized the solution of the MT problem in terms of a so-called optimal coupling in the case that ν_2 has *finite* support and ν_1 verifies a continuity condition. A similar approach was used by Abdellaoui [1]. Cambanis, Simons, and Stout [7] and also McCann [18] found some explicit solutions to the MT problem on the real line, i.e., $X = Y = \mathbb{R}$. Ruzankin [25] also studied the form of optimal solutions on the real line. In [11], Gangbo used randomized strategies to formulate the MT problem as a variational one on $\mathbb{R}^d \times \mathbb{R}^d$.

In this paper we also use randomized strategies, but in general metric spaces. The idea is to introduce a suitable topology on the space of those strategies, and then use the Bauer extremum principle [8] to prove that MT (actually, the equivalent problem MTK) is solvable. Our approximation approach in section 6 is partly inspired by [3], although our setting and proof techniques are quite different.

Summarizing, in this paper we present *two* new, different, approaches to obtaining the solvability of MT. The first one is to give conditions for the *equivalent* problem MTK to be solvable. The second is an approximation approach by means of MT problems with marginals concentrated on finite sets.

After some technical preliminaries in sections 2 and 3, the plan for the remainder of the paper (sections 4, 5, and 6) is as sketched in the first paragraph of the introduction.

2. The MT problem. In the MT problem with which we are concerned, we are given the following data: (i) two metric spaces X and Y endowed with the corresponding Borel σ-algebras $\mathbb{B}(X)$ and $\mathbb{B}(Y)$; (ii) a nonnegative measurable function $c : X \times Y \to \mathbb{R}$; and (iii) a probability measure (p.m.) ν_1 on X, and a p.m. ν_2 on Y. Moreover, let $M(X \times Y)$ be the linear space of finite signed measures on $X \times Y$, endowed with the topology of weak convergence, and let $M^+(X \times Y)$ be the convex cone of nonnegative measures in $M(X \times Y)$. If μ is in $M(X \times Y)$, we denote by $\Pi_1\mu$ and $\Pi_2\mu$ the marginals (or projections) of μ on X and Y, respectively; that is, for all $A \in \mathbb{B}(X)$ and $B \in \mathbb{B}(Y)$

$$\Pi_1\mu(A) := \mu(A \times Y) \quad \text{and} \quad \Pi_2\mu(B) := \mu(X \times B).$$

Then, with $\langle \mu, c\rangle := \int c \, d\mu$, the MT problem can be stated as follows:

$$\textbf{MT} \qquad \text{Minimize} \quad \langle \mu, c\rangle \tag{2.1}$$

$$\text{subject to} \quad \Pi_1\mu = \nu_1, \quad \Pi_2\mu = \nu_2, \quad \mu \in M^+(X \times Y). \tag{2.2}$$

A measure $\mu \in M(X \times Y)$ is said to be a *feasible solution* for the MT problem if it satisfies (2.2) and $\int c \, d\mu$ is finite. The MT problem is called *consistent* if the set of feasible solutions is nonempty, in which case its (optimum) *value* is defined as

$$\inf(\text{MT}) := \inf\{\langle \mu, \text{c}\rangle | \mu \text{ is feasible for MT}\}.$$

It is said that the MT problem is *solvable* if there is a feasible solution μ^* that attains the optimum value. In this case, μ^* is called an *optimal solution* for the MT problem, and the value $\inf(\text{MT})$ is written as $\min(\text{MT}) = \langle \mu^*, c\rangle$.

Remark 2.1.

(a) Since ν_1 and ν_2 are p.m.'s, a feasible solution for the MT problem is necessarily a p.m.

(b) If c is a *bounded* function, then the product measure $\mu := \nu_1 \times \nu_2$ is feasible. This fact is not necessarily true if c is *unbounded*; see Example 1.2 in [12]. Even for unbounded c, however, mild assumptions ensure that the MT problem is consistent [12].

(c) If X and Y are compact metric spaces and c is l.s.c., then the set of feasible solutions is compact and, therefore, the MT problem is solvable. (See Theorem 2.2 and Remark 2.5 in [15].)

Now we will introduce an optimization problem on a family of stochastic kernels, which in section 4 is shown to be equivalent to the MT problem, when ν_2 has finite support.

Let $\mathbb{F}$ be the set of all (Borel) measurable functions from X to Y. Following the usage in control and game theory [6, 13, 14], a function in $\mathbb{F}$ will be referred to as a *deterministic strategy*. We also need the following more general concept of strategy.

DEFINITION 2.2. *A* randomized strategy (*also known as a* stochastic kernel) φ *from X to Y is a real-valued mapping $(x, B) \mapsto \varphi(B|x)$ on $X \times \mathbb{B}(Y)$ such that*

(a) *$\varphi(\cdot|x)$ is a p.m. on $\mathbb{B}(Y)$ for every fixed $x \in X$, and*

(b) *$\varphi(B|\cdot)$ is a measurable function on X for every fixed $B \in \mathbb{B}(Y)$.*

If (a) *is replaced with*

(a′) *$\varphi(\cdot|x)$ is a finite signed measure on $\mathbb{B}(Y)$ for every fixed $x \in X$,*

then φ is called a signed kernel *from X to Y. We shall denote by Φ the set* (*actually linear space*) *of all such kernels, and by Φ_1 the convex subset of randomized strategies.*

Given a deterministic strategy $f \in \mathbb{F}$, we may identify each $f(x) \in Y$ with the Dirac measure (or unit mass) $\varphi(\cdot|x) := \delta_{f(x)}(\cdot)$ concentrated at $f(x)$ for each $x \in X$. This of course defines a randomized strategy. With this identification we may write

$$\mathbb{F} \subset \Phi_1. \tag{2.3}$$

Let ν be a p.m. on X, and consider the following optimization problem:

$$\mathbb{P} \qquad \text{Minimize} \quad \int_X \int_Y c(x,y)\varphi(dy|x)\nu(dx), \tag{2.4}$$

$$\text{subject to} \quad \int_X \int_Y c_i(x,y)\varphi(dy|x)\nu(dx) \le b_i \quad \forall i = 1, \dots, n, \quad \varphi \in \Phi_1, \tag{2.5}$$

where $c_1, \dots, c_n$ are given measurable functions on $X \times Y$ and $b_1, \dots, b_n$ are given real numbers. This problem is a "randomized version" of the typical allocation problem in economics; see [4, 13, 27]. In section 6 below, we use a particular case of the MT problem that is of the form (2.4)–(2.5), and we will need to describe the extreme points $\varphi \in \Phi_1$ that satisfy (2.5). We describe these extreme points in Theorem 3.2.

Remark 2.3. An important variant of MT is the mass transshipment problem defined as [9, 23, 24]:

$$\text{Minimize} \quad \langle \mu, c \rangle$$

$$\text{subject to} \quad \Pi_1\mu - \Pi_2\mu = \nu_1 - \nu_2, \quad \mu \in M^+(X \times X), \tag{2.6}$$

where $c : X \times X \to \mathbb{R}$ is a given cost function and ν_1 and ν_2 are p.m.'s on X, a metric space. The so-called balance condition in (2.6) is equivalent to the following: there

exists $\Delta \in M(X)$ such that

$$\Pi_1\mu = \nu_1 + \Delta \quad \text{and} \quad \Pi_2\mu = \nu_2 - \Delta.$$

Hence, for any feasible measure μ for the mass transshipment problem we have

$$\mu(X \times X) = \Pi_1\mu(X) = \nu_1(X) + \Delta(X) = 1 + \Delta(X),$$

$$\mu(X \times X) = \Pi_2\mu(X) = \nu_2(X) - \Delta(X) = 1 - \Delta(X).$$

It follows that $\Delta(X) = 0$, and, therefore, μ is a p.m. Summarizing, if μ is feasible for the MT problem with $X = Y$ (see (2.2)), then μ is feasible for the mass transshipment problem. Moreover, the results for MT introduced in the following sections can be extended, with appropriate changes, to the mass transshipment problem. Research on this matter is in progress.

3. Convex sets of randomized strategies. In this section we introduce a result from [13] that characterizes the extreme points of feasible solutions to the problem $\mathbb{P}$ in (2.4)–(2.5). This result is applied to the MT problem in section 6. First, we need a definition. (Recall (2.3).)

DEFINITION 3.1. *A randomized strategy $\varphi \in \Phi_1$ is said to be a* randomization of at most $n+1$ *deterministic strategies, or simply an* $(n+1)$-randomization, *if there exist a positive integer $m \leq n+1$, functions $f_1, \ldots, f_m$ in $\mathbb{F}$, and nonnegative numbers $\alpha_1, \ldots, \alpha_m$ with $\alpha_1 + \cdots + \alpha_m = 1$ such that*

$$\varphi(B|x) = \sum_{j=1}^{m} \alpha_j \delta_{f_j(x)}(B) \quad \forall B \in \mathbb{B}(Y),\ x \in X. \tag{3.1}$$

In this case we write $\varphi \in \mathcal{R}_{n+1}(f_1, \ldots, f_m; \alpha_1, \ldots, \alpha_m)$.

The following result, which uses the notation in (2.5), is applied to the MT problem in section 6.

THEOREM 3.2. *Let X be a metric space, and Y a separable metric space. Fix an arbitrary p.m. ν on $\mathbb{B}(X)$, real-valued measurable functions $c_1, \ldots, c_n$ on $X \times Y$, and real numbers $b_1, \ldots, b_n$. Consider the set $\Delta \subset \Phi_1$ consisting of all the randomized strategies φ in Φ_1 for which, for all $i = 1, \ldots, n$,*

$$\int_X \int_Y |c_i(x,y)| \varphi(dy|x)\nu(dx) < \infty \tag{3.2}$$

and

$$\int_X \int_Y c_i(x,y) \varphi(dy|x)\nu(dx) \leq b_i. \tag{3.3}$$

Let $ex(\Delta)$ be the set of extreme points of Δ. Then

(a) *Δ is convex and*

$$ex(\Delta) \subset \mathcal{R}^0_{n+1}, \tag{3.4}$$

where $\mathcal{R}^0_{n+1}$ is the set of all the $(n+1)$-randomizations $\varphi \in \mathcal{R}_{n+1}(f_1, \ldots, f_m; \alpha_1, \ldots, \alpha_m)$ for which the vectors

$$\left(\int c_1(x, f_j(x))\nu(dx), \ldots, \int c_n(x, f_j(x))\nu(dx), 1 \right) \in \mathbb{R}^{n+1} \tag{3.5}$$

for $j = 1, \ldots, m$ are linearly independent.

(b) *If equality holds in* (3.3), *then the sets in* (3.4) *are equal.*

Proof. See [13]. □

4. An equivalent problem to MT. In this section we show that the MT problem is equivalent to the following optimization problem on the set Φ_1 of randomized strategies.

Let c, ν_1, and ν_2 be as in the MT problem (2.1)–(2.2). Consider

$$\textbf{MTK} \qquad \text{Minimize} \quad \int_X \int_Y c(x,y)\varphi(dy|x)\nu_1(dx), \tag{4.1}$$

$$\text{subject to} \quad \int_X \varphi(\cdot|x)\nu_1(dx) = \nu_2(\cdot),\ \varphi \in \Phi_1. \tag{4.2}$$

A randomized strategy $\varphi \in \Phi_1$ is said to be a *feasible solution* for MTK if it satisfies (4.2) and the integral in (4.1) is finite. The problem MTK is called *consistent* if the set of feasible solutions is nonempty, in which case its *value* is defined as the infimum, over the set of feasible solutions, of the integrals in (4.1). If the infimum is attained at some feasible solution, say φ^*, then MTK is said to be *solvable* and φ^* is called an *optimal solution* for MTK.

THEOREM 4.1. *MT and MTK are equivalent; that is, for each φ in Φ_1 that satisfies* (4.2) (*a feasible solution for MTK*), *there is a p.m. μ on $X \times Y$ that satisfies* (2.2) (*a feasible solution for MT*) *such that*

$$\langle \mu, c\rangle := \int_{X\times Y} c\ d\mu = \int_X \int_Y c(x,y)\varphi(dy|x)\nu_1(dx) =: \langle \varphi, c\rangle. \tag{4.3}$$

Conversely, for each feasible solution μ for MT, there exists a feasible solution φ for MTK such that φ and μ satisfy (4.3).

Proof. Let μ be a feasible solution for MT. Then, by a well-known result on the disintegration of product measures (see, e.g., [14, Proposition D.8, p. 184]), there is a stochastic kernel $\varphi \in \Phi_1$ such that

$$\mu(A \times B) = \int_A \varphi(B|x)\nu_1(dx) \quad \forall A \in \mathbb{B}(X), \quad B \in \mathbb{B}(Y). \tag{4.4}$$

Hence

$$\begin{aligned} \nu_2(B) = \Pi_2\mu(B) &= \mu(X \times B) \\ &= \int_X \varphi(B|x)\nu_1(dx), \end{aligned}$$

and thus φ is a feasible solution for MTK by (4.2). Moreover, by (4.4),

$$\int_{X\times Y} c\ d\mu = \int_X \int_Y c(x,y)\varphi(dy|x)\nu_1(dx),$$

and (4.3) follows.

Now, let φ be a feasible solution for MTK. Then there is unique p.m. μ on $X \times Y$ given by

$$\mu(A \times B) := \int_A \varphi(B|x)\nu_1(dx) \quad \forall A \in \mathbb{B}(X), \quad B \in \mathbb{B}(Y), \tag{4.5}$$

and for which

$$\int_{X\times Y} c\, d\mu = \int_X \int_Y c(x,y)\varphi(dy|x)\nu_1(dx).$$

Furthermore,

$$\Pi_1\mu(A) = \mu(A \times Y) = \int_A \varphi(Y|x)\nu_1(dx) = \nu_1(A),$$

and since φ is a feasible solution for MTK, (4.2) gives

$$\Pi_2\mu(B) = \mu(X \times B) = \int_X \varphi(B|x)\nu_1(dx) = \nu_2(B).$$

Therefore, μ satisfies (2.2) and thus it is a feasible solution for MT. □

A measure defined as in (4.5) will be denoted by $\mu := \varphi \cdot \nu_1$.

Part (c) in the following corollary uses (2.3).

COROLLARY 4.2.

(a) *MT is consistent if and only if MTK is consistent.*

(b) *MT is solvable if and only if MTK is solvable.*

(c) *If MTK has an optimal solution $\varphi \in \Phi_1$ which is a deterministic strategy, say $\varphi(\cdot|x) = \delta_{f(x)}(\cdot)$ for some $f \in \mathbb{F}$, then f is a so-called* optimal coupling *for MT (which means that the measure $\mu(A \times B) = \int_A \delta_{f(x)}(B)\nu_1(dx)$ is concentrated along $(x, f(x))$) and is an optimal solution for MT* [1, 18, 21, 22].

Let $\mathcal{F}_{MT}$ be the class of feasible solutions to MT and let $\mathcal{F}_{MTK}$ be the class of feasible solutions to MTK. In view of Corollary 4.2(b), in the following section we study the solvability of MTK. In fact, we show that MTK has an optimal solution which is an extreme point of $\mathcal{F}_{MTK}$.

5. Solvability of MTK. To show that MTK is solvable, first we introduce some concepts that will be used to define suitable topologies on sets of stochastic kernels. The following definitions can also be found in [5, 6], for instance. (Recall Definition 2.2 on Φ and Φ_1.)

DEFINITION 5.1. *A* uniformly finite kernel *from X to Y is a signed kernel $\varphi \in \Phi$ such that*

$$\sup_{x \in X} |\varphi(Y|x)| < \infty,$$

where, for each $x \in X$, $|\varphi(Y|x)|$ denotes the total variation of $\varphi(\cdot|x)$.

The set of all uniformly finite kernels is denoted by Φ_0. It is evident that Φ_0 is a linear space with the usual addition and scalar multiplication of signed kernels, and that, moreover, $\Phi_1 \subset \Phi_0 \subset \Phi$.

DEFINITION 5.2. *A* normal integrand bounded from below *on $X \times Y$ is a function $g : X \times Y \to (-\infty, \infty]$ such that*

(i) *$g(x, \cdot)$ is l.s.c. on Y for every $x \in X$,*

(ii) *g is $\mathbb{B}(X) \times \mathbb{B}(Y)$-measurable, and*

(iii) *there is a function $f \in L_1(\nu_1) := L_1(X, \mathbb{B}(X), \nu_1)$ such that $g(x, y) \geq f(x)$ for all $x \in X, y \in Y$.*

The set of all normal integrands bounded from below is denoted by $\mathcal{C}^{bb}$. We next introduce a suitable subset of $\mathcal{C}^{bb}$.

DEFINITION 5.3. *A* Carathéodory integrand *on $X \times Y$ is a function $g : X \times Y \to \mathbb{R}$ such that*

(i) *$g(x, \cdot)$ is continuous on Y for every $x \in X$,*

(ii) *g is $\mathbb{B}(X) \times \mathbb{B}(Y)$-measurable on $X \times Y$, and*

(iii) *$|g| \leq f$ on $X \times Y$ for some $f \in L_1(\nu_1)$.*

The set of all Carathéodory integrands on $X \times Y$ is denoted by $\mathcal{C}^c$. Observe that $\mathcal{C}^c$ contains the space $C_b(X \times Y)$ of continuous bounded functions on $X \times Y$. On the other hand, for each $g \in \mathcal{C}^c$ we define a functional $I_g : \Phi_0 \to \mathbb{R}$ by

$$I_g(\varphi) := \int_X \int_Y g(x,y)\varphi(dy|x)\nu_1(dx). \tag{5.1}$$

The *weak topology* on Φ_0 is defined as the coarsest topology for which all the functionals $I_g : \Phi_0 \to \mathbb{R}$, for $g \in \mathcal{C}^c$, are continuous. Similarly, the *weak topology* on Φ_1 is defined as the coarsest topology for which all the functionals $I_g : \Phi_1 \to \mathbb{R}$, for $g \in \mathcal{C}^c$, are continuous. It can be proved that Φ_0 is a Hausdorff topological vector space which is locally convex [5].

ASSUMPTION 5.4.

(a) *Y is a complete and separable metric space.*

(b) *The "cost" function $c(x,y)$ is nonnegative and $c(x,\cdot)$ is l.s.c. on Y for every $x \in X$. (Hence c is in $\mathcal{C}^{bb}$.)*

(c) *MT is consistent.*

Concerning the consistency in Assumption 5.4(c), see Remark 2.1.

We have the following result.

THEOREM 5.5. *Under Assumption* 5.4*, there exists a stochastic kernel $\varphi \in \Phi_1$ such that φ is an optimal solution to MTK and, moreover, φ is an extreme point of $\mathcal{F}_{MTK}$.*

Proof. First we will show that $\mathcal{F}_{MTK}$ is weakly closed. Let $(N, \leq)$ be a directed set and suppose that $\{\varphi_n, n \in N\}$ converges weakly to φ in Φ_1. Then, by Theorem 2.2(b) in [6], we have

$$\int_X \int_Y g(x,y)\varphi_n(dy|x)\nu_1(dx) \to \int_X \int_Y g(x,y)\varphi(dy|x)\nu_1(dx) \tag{5.2}$$

for all g in $\mathcal{C}^c$.

For φ and each φ_n, let μ and μ_n be the corresponding p.m.'s on $X \times Y$ defined by (4.5). By Theorem 4.1, $\{\mu_n, n \in N\}$ is a net in $\mathcal{F}_{MT}$. Moreover, by (4.5) again, for each f in $\mathcal{C}^c$ we have

$$\int_{X\times Y} f d\mu_n = \int_X \int_Y f(x,y)\varphi_n(dy|x)\nu_1(dx) \tag{5.3}$$

and

$$\int_{X\times Y} f d\mu = \int_X \int_Y f(x,y)\varphi(dy|x)\nu_1(dx). \tag{5.4}$$

In particular, if $f \in C_b(X \times Y)$, then by (5.2)–(5.4), we have

$$\begin{aligned} \int_{X\times Y} f d\mu_n &= \int_X \int_Y f(x,y)\varphi_n(dy|x)\nu_1(dx) \\ &\to \int_X \int_Y f(x,y)\varphi(dy|x)\nu_1(dx) \\ &= \int_{X\times Y} f d\mu. \end{aligned}$$

This implies that $\{\mu_n\}$ converges weakly to μ. In turn, the latter yields, by Lemma 2.7 in [15], that μ is a feasible solution to MT. Therefore, by Theorem 4.1, φ is in $\mathcal{F}_{MTK}$ and we conclude that $\mathcal{F}_{MTK}$ is weakly closed.

As Φ_1 is weakly compact (see Theorem 2.3(a) in [6]) and $\mathcal{F}_{MTK}$ is weakly closed, it follows that $\mathcal{F}_{MTK}$ is weakly compact. This fact, together with Assumption 5.6(b) and Theorem 2.3(b) in [6], gives that the functional

$$I_c(\varphi) := \int_X \int_Y c(x,y)\varphi(dy|x)\nu_1(dx) \tag{5.5}$$

is weakly inf-compact. Thus, I_c is a weakly l.s.c. functional and thus, by the Bauer extremum principle (see, for instance, Theorem 25.9 in [8]), it follows that there is an extreme point φ of $\mathcal{F}_{MTK}$ where I_c attains its minimum value. Finally comparing (5.5) and (4.1) we obtain the desired conclusion. □

6. An approximation approach to the MT problem. In this section we study the MT problem by means of an approximation procedure consisting of three steps. First, we assume that the marginal ν_2 has finite support, and we characterize an optimal solution for this problem (see Theorem 6.1). Second, we construct a sequence {MTi} of MT problems with the characteristic that the marginals ν_1^i and ν_2^i both have finite support. Then for each i, we find an optimal solution μ_i of MTi and show that $\{\mu_i\}$ is relatively compact. Therefore, there exist a measure μ^* and a subsequence $\{\mu_m\}$ such that $\{\mu_m\}$ converges weakly to μ^*. Finally, in the third step we prove that μ^* is an optimal solution to MT and, furthermore, $\lim_{m\to\infty}\langle \mu_m, c\rangle = \langle \mu^*, c\rangle$.

Step 1. Let us assume that ν_2 has finite support. Then there is a finite subset $\{y_1, y_2, \ldots, y_n\}$ of Y such that $\nu_2(\{y_i\}) := b_i \neq 0$ for all $i = 1, \ldots, n$, and $\nu_2(\{y_1, y_2, \ldots, y_n\}) = 1$.

We observe that the constraints (4.2) in this case are equivalent to

$$\int_X \varphi(\{y_i\}|x)\nu_1(dx) = \nu_2(\{y_i\}) = b_i \quad \forall i = 1, \ldots, n, \quad \varphi \in \Phi_1. \tag{6.1}$$

Consider $c_i(x,y) := I_{\{y_i\}}(y)$, where I_B is the indicator function of $B \in \mathbb{B}(X)$; that is, $I_B(x) := 1$ if $x \in B$ and $I_B(x) := 0$ if $x \notin B$. Then (6.1) can be expressed as

$$\int_X \int_Y I_{\{y_i\}}(y)\varphi(dy|x)\nu_1(dx) = b_i \quad \forall i = 1, \ldots, n, \quad \varphi \in \Phi_1$$

or, equivalently,

$$\int_X \int_Y c_i(x,y)\varphi(dy|x)\nu_1(dx) = b_i \quad \forall i = 1, \ldots, n, \quad \varphi \in \Phi_1.$$

Therefore, we have the following optimization problem, which is of the form (2.4)–(2.5):

$$\textbf{MTKn} \qquad \text{Minimize} \quad \int_X \int_Y c(x,y)\varphi(dy|x)\nu_1(dy) \tag{6.2}$$

$$\text{subject to} \quad \int_X \int_Y c_i(x,y)\varphi(dy|x)\nu_1(dx) = b_i \quad \forall i = 1, \ldots, n, \varphi \in \Phi_1. \tag{6.3}$$

Now, by Theorem 5.5, the MTKn problem achieves its minimum value at an extreme point φ_n^*. Moreover, by Theorem 3.2, φ_n^* is an $(n+1)$-randomization; that

is, there exist a positive integer $m \le n+1$, functions $f_1, \dots, f_m$ in $\mathbb{F}$, and nonnegative numbers $\alpha_1, \dots, \alpha_m$ such that $\alpha_1 + \cdots + \alpha_m = 1$ and

$$\varphi_n^*(B|x) = \sum_{j=1}^{m} \alpha_j \delta_{f_j(x)}(B) \quad \forall B \in \mathbb{B}(X), \quad x \in X. \tag{6.4}$$

In addition, by Corollary 4.2, the MT problem associated to MTKn is solvable.

Combining the latter facts we obtain the following theorem.

THEOREM 6.1. *If Assumption* 5.4 *holds and the marginal* ν_2 *has finite support, then there are a stochastic kernel* φ^* *of the form* (6.4) *and a probability measure* $\mu^* := \varphi^* \cdot \nu_1$ *on* $X \times Y$ *as in* (4.5)*; i.e.,*

$$\mu^*(A \times B) = \int_A \varphi^*(B|x)\nu_1(dx)$$

such that μ^* *is an optimal solution for the MT problem.*

Step 2. We need the following assumption.

ASSUMPTION 6.2.

(a) *The "cost" function* $c(x, y)$ *is nonnegative and continuous.*

(b) X *and* Y *are separable metric spaces.*

(c) *There exists a sequence* $\{H_n\}$ *of compact subsets of* $X \times Y$ *such that* $H_n \uparrow X \times Y$*; i.e., the* H_n *form a nondecreasing sequence that converges to* $\bigcup_n H_n = X \times Y$*.*

We observe that the Assumption 6.2(c) holds, for instance, if X and Y are both σ-compact metric spaces or if c is an inf-compact function.

The next proposition is a variant of a well-known result on the denseness of finitely supported measures in spaces of probability measures (see Theorem 4 in [20, p. 237]). Here for our approximation approach we need a more explicit result.

PROPOSITION 6.3. *If the Assumption* 6.2 *holds, then there exist two sequences of probability measures* $\{\nu_1^i\}$ *on* $\mathbb{B}(X)$ *and* $\{\nu_2^i\}$ *on* $\mathbb{B}(Y)$*, both with finite supports and such that* $\{\nu_1^i\}$ *converges weakly to* ν_1 *and* $\{\nu_2^i\}$ *converges weakly to* ν_2*.*

Proof. By Assumption 6.2(c), there are two sequences of compact sets F_t and G_t in X and Y, respectively, such that $F_t \times G_t \uparrow X \times Y$.

Let

$$E_i = \{(x, y) | c(x, y) \le i\}$$

for $i = 1, 2, \dots$. Let $\Pi_1(E_i)$ and $\Pi_2(E_i)$ be the projections of E_i on X and Y, respectively. Then there is a positive integer t_i such that

$$\nu_1(F_{t_i} \cap \Pi_1(E_i)) \ge \nu_1(\Pi_1(E_i)) - 1/4i^2 \tag{6.5}$$

and

$$\nu_2(G_{t_i} \cap \Pi_2(E_i)) \ge \nu_2(\Pi_2(E_i)) - 1/4i^2. \tag{6.6}$$

Since c is a continuous function and $F_{t_i} \times G_{t_i} \cap E_i$ is a compact set, there exists $\gamma_i > 0$ such that

$$\text{if } d_1(x, x') < \gamma_i \quad \text{and} \quad d_2(y, y') < \gamma_i, \quad \text{then } |c(x, y) - c(x', y')| < 1/i,$$

for $(x, y), (x', y') \in F_i \times G_i \cap E_i$. Let $\varepsilon_i = \min\{\gamma_i, 1/i\}$.

Let $B_r(z)$ be the open neighborhood with center in z and radius r. By the compactness of $F_{t_i} \times G_{t_i} \cap E_i$, there are positive integers M_i, N_i and finite sets $\{x_1^i, x_2^i, \ldots, x_{M_i}^i\}$ and $\{y_1^i, y_2^i, \ldots, y_{N_i}^i\}$ such that $F_{t_i} \cap \Pi_1(E_i) \subset \bigcup_{k=1}^{M_i} B_{\varepsilon_i}(x_k^i)$ and $G_i \cap \Pi_2(E_i) \subset \bigcup_{j=1}^{N_i} B_{\varepsilon_i}(y_j^i)$.

We define the disjoint sets

$$A_k^i = \left(B_{\varepsilon_i}(x_k^i) \Big\backslash \bigcup_{s=1}^{k-1} B_{\varepsilon_i}(x_s^i) \right) \cap F_{t_i} \cap \Pi_1(E_i) \quad \text{for } k = 1, 2, \ldots, M_i,$$

$$D_j^i = \left(B_{\varepsilon_i}(y_j^i) \Big\backslash \bigcup_{s=1}^{j-1} B_{\varepsilon_i}(y_s^i) \right) \cap G_{t_i} \cap \Pi_2(E_i) \quad \text{for } j = 1, 2, \ldots, N_i,$$

$$A_{M_i+1}^i = (F_{t_i} \cap \Pi_1(E_i))^c \quad \text{and} \quad D_{N_i+1}^i = (G_{t_i} \cap \Pi_2(E_i))^c.$$

We take $(x_{M_i+1}^i, y_{N_i+1}^i)$ in $(F_{t_i} \times G_{t_i}) \cap E_i$.

Now, we introduce measures ν_1^i and ν_2^i with finite supports

$$\{x_1^i, \ldots, x_{M_i+1}^i\} \subset F_{t_i} \cap \Pi_1(E_i)$$

and

$$\{y_1^i, \ldots, y_{N_i+1}^i\} \subset G_{t_i} \cap \Pi_2(E_i),$$

respectively, and defined as

$$\nu_1^i(A) := \sum_{k=1}^{M_i+1} \nu_1(A_k^i) \delta_{\{x_k^i\}}(A) \quad \text{for } A \in \mathbb{B}(X)$$

and

$$\nu_2^i(D) := \sum_{j=1}^{N_i+1} \nu_2(D_j^i) \delta_{\{y_j^i\}}(D) \quad \text{for } D \in \mathbb{B}(Y).$$

We claim that ν_1^i converges weakly to ν_1, and ν_2^i converges weakly to ν_2. The proofs of the last two facts are quite similar, and so we will prove only the latter, ν_2^i converges weakly to ν_2, i.e.,

$$\lim_{i \to \infty} \int f d\nu_2^i = \int f d\nu_2$$

for every bounded real-valued uniformly continuous function f (see, for instance, [20, Theorem 6.1(b), p. 40]).

Let $f : Y \to \mathbb{R}$ be a bounded uniformly continuous function. Let $M > 0$ be such that

$$|f(y)| \leq M \quad \forall y \in Y. \tag{6.7}$$

Pick an arbitrary $\varepsilon > 0$. As f is uniformly continuous, there exists $\gamma > 0$ such that

$$\text{if } d_2(y, y') < \gamma, \quad \text{then } |f(y) - f(y')| < \varepsilon/3. \tag{6.8}$$

Now let i_1 and i_2 be positive integers such that $1/i_1 < \gamma$ and $1/(i_2)^2 < \varepsilon/6M$. Moreover, as $\Pi_2(E_i) \uparrow Y$, there is an integer i_3 such that $\nu_2(E_{i_3}^c) < \varepsilon/6M$.

Let $i_0 = \max\{i_1, i_2, i_3\}$. Then, by (6.6), (6.7), (6.8), and the construction of ν_2^i, we have that for all $i \geq i_0$

$$\begin{aligned}
&\left| \int_Y f(y)\nu_2(dy) - \int_Y f(y)\nu_2^i(dy) \right| \\
&\quad \leq \sum_{j=1}^{N_i} \int_{D_j^i} |f(y) - f(y_j^i)|\nu_2(dy) + \int_{D_{N_i+1}} |f(y) - f(y_{N_i+1}^i)|\nu_2(dy) \\
&\quad \leq \frac{\varepsilon}{3} \sum_{j=1}^{N_i} \nu_2(D_j^i) + 2M\nu_2(\Pi_2(E_i) - G_{t_i}) + 2M\nu_2(\Pi_2(E_i)^c) \\
&\quad < \varepsilon. \qquad \square
\end{aligned}$$

In the following theorem $\{\nu_1^i\}$ and $\{\nu_2^i\}$ are the sequences given by Proposition 6.3, and

$$c_i := \min\{c, i\}.$$

THEOREM 6.4. *If the Assumption* 6.2 *holds and if μ is a feasible solution to the MT problem, then there exists a sequence of p.m.'s $\{\mu_i\}$ on $\mathbb{B}(X \times Y)$ with marginals $\Pi_1\mu_i = \nu_1^i$ and $\Pi_2\mu_i = \nu_2^i$, and such that μ_i converges weakly to μ. Moreover,*

$$\lim_{i\to\infty} \langle \mu_i, c_i \rangle = \langle \mu, c \rangle.$$

Proof. Let A_k^i, D_j^i, F_{t_i}, G_{t_i}, E_i, M_i, N_i, $\{x_1^i, \ldots, x_{M_i+1}^i\}$, and $\{y_1^i, \ldots, y_{N_i+1}^i\}$ be as in the proof of Proposition 6.3.

Now, let μ be a probability measure on $\mathbb{B}(X \times Y)$ such that $\Pi_1\mu = \nu_1$ and $\Pi_2\mu = \nu_2$. We define measures μ_i by

$$\mu_i(E) = \sum_{k=1}^{M_i+1} \sum_{j=1}^{N_i+1} \mu(A_k^i \times D_j^i)\delta_{(x_k^i, y_j^i)}(E)$$

for all E in $\mathbb{B}(X \times Y)$.

Observe that μ_i has a finite support contained in $(F_{t_i} \times G_{t_i}) \cap E_i$. We will now see that $\Pi_1\mu_i = \nu_1^i$.

Choose an arbitrary set A in $\mathbb{B}(X)$. Then the definition and properties of Dirac measures give

$$\begin{aligned}
\Pi_1\mu_i(A) &= \mu_i(A \times Y) \\
&= \sum_{k=1}^{M_i+1} \left(\sum_{j=1}^{N_i+1} \mu(D_k^i \times C_j^i) \right) \delta_{x_k^i}(A) \\
&= \sum_{k=1}^{M_i+1} \mu(D_k^i \times Y)\delta_{x_k^i}(A) \\
&= \sum_{k=1}^{M_i+1} \nu_1(D_k^i)\delta_{x_k^i}(A) \\
&= \nu_1^i(A);
\end{aligned}$$

that is, $\Pi_1\mu_i = \nu_1^i$. Similarly $\Pi_2\mu_i = \nu_2^i$.

We observe that, as in the proof of Proposition 6.3, $\{\mu_i\}$ converges weakly to μ. Finally we show that

$$\lim_{i\to\infty}\langle\mu_i, c_i\rangle = \langle\mu, c\rangle.$$

As $E_i \uparrow X\times Y$ and c is μ-integrable, given $\varepsilon>0$ there exists an integer i_1 such that, for all $i\geq i_1$, we have

$$\int_{E_i^c} c\,d\mu < \varepsilon/6. \tag{6.9}$$

Moreover, there exists i_2 such that, for all $i\geq i_2$, we have

$$1/i < \varepsilon/6. \tag{6.10}$$

Now, let $i_0=\min\{i_1,i_2\}$. By (6.5), (6.6), (6.9), and (6.10) we have that for all $i\geq i_0$

$$\begin{aligned}
\left|\int_{X\times Y} c\,d\mu - \int_{X\times Y} c_i d\mu_i\right| &= \left|\int_{E_i} c\,d\mu + \int_{E_i^c} c\,d\mu - \int_{E_i} c_i d\mu_i - \int_{E_i^c} c_i d\mu_i\right| \\
&= \left|\int_{E_i} c\,d\mu + \int_{E_i^c} c\,d\mu - \sum_{k=1}^{M_i+1}\sum_{j=1}^{N_i+1} c_i(x_k^i,y_j^i)\mu((A_k^i\times D_j^i)\cap E_i)\right. \\
&\quad \left. - \sum_{k=1}^{M_i+1}\sum_{j=1}^{N_i+1} c_i(x_k^i,y_j^i)\mu((A_k^i\times D_j^i)\cap E_i^c)\right| \\
&\leq 2\int_{E_i^c} c\,d\mu + \sum_{k=1}^{M_i}\sum_{j=1}^{N_i}\int_{(A_k^i\times D_j^i)\cap E_i} |c(x,y)-c_i(x_k^i,y_j^i)|d\mu \\
&\quad + \sum_{k=1}^{M_i}\int_{(A_k^i\times D_{N_i+1}^i)\cap E_i} c\,d\mu + \sum_{k=1}^{M_i+1} c_i(x_k^i,y_{N_i+1}^i)\mu((A_k^i\times D_{N_i+1}^i)\cap E_i) \\
&\quad + \sum_{j=1}^{N_i}\int_{(A_{M_i+1}^i\times D_j^i)\cap E_i} c\,d\mu + \sum_{j=1}^{N_i+1} c_i(x_{M_i+1}^i,y_j^i)\mu((A_{M_i+1}^i\times D_j^i)\cap E_i) \\
&\leq \varepsilon/3 + (1/i)\sum_{k=1}^{M_i}\sum_{j=1}^{N_i}\mu((A_k^i\times D_j^i)\cap E_i) \\
&\quad + 2i\{\mu((X\times D_{N_i+1}^i)\cap E_i) + \mu(((A_{M_i+1}^i)\times Y)\cap E_i)\} \\
&< \varepsilon/6 + 2i\{\mu(X\times(D_{N_i+1}^i\cap\Pi_2(E_i))) + \mu((A_{M_i+1}^i\cap\Pi_1(E_i))\times Y)\} \\
&= \varepsilon/6 + 2i\{\nu_2(D_{N_i+1}^i\cap\Pi_2(E_i)) + \nu_1(A_{M_i+1}^i\cap\Pi_1(E_i))\} \\
&\leq \varepsilon/6+\varepsilon/3<\varepsilon. \qquad \square
\end{aligned}$$

Step 3. We now complete the approximation procedure. We will need either one of the following assumptions.

ASSUMPTION 6.5.

(a) *X is a separable metric space.*

(b) *Y is a complete and separable metric space.*

(c) *The "cost" function $c(x,y)$ is nonnegative, inf-compact, and continuous.*

ASSUMPTION 6.6.
(a) *X is a σ-compact metric space.*
(b) *Y is a complete and σ-compact metric space.*
(c) *The "cost" function $c(x, y)$ is nonnegative and continuous.*

Finally, we state our main result, in which N_m is as in the proof of Proposition 6.3.

THEOREM 6.7. *If either Assumption* 6.5 *or Assumption* 6.6 *holds, then there exists a sequence $\{\mu_m^*\}$ of probability measures of the form*

$$\mu_m^* = \varphi_m^* \cdot \nu_1,$$

where φ_m^* is a randomization of at most N_m+1 deterministic strategies, and, moreover
(a) *μ_m^* converges weakly to a probability measure μ^*,*
(b) *μ^* is an optimal solution to MT, and*
(c) *$\lim_{m\to\infty}\langle c_m, \mu_m^*\rangle = \langle c, \mu^*\rangle = \min(MT)$, with $c_m = \min\{c, m\}$.*

Proof. Let $\{\nu_1^i\}$ and $\{\nu_2^i\}$ be sequences of p.m.'s as in Proposition 6.3. Consider the following MT problems:

$$\textbf{MTi} \qquad \text{Minimize} \quad \langle \mu, c_i\rangle$$
$$\text{subject to} \quad \Pi_1\mu = \nu_1^i, \quad \Pi_2\mu = \nu_2^i, \quad \mu \geq 0.$$

Now, by Theorem 6.1, for each i there exists a randomization of at most $N_i + 1$ deterministic strategies φ_{N_i+1} of the form (6.4) and such that $\mu_i = \varphi_{N_i+1} \cdot \nu_1^i$ is an optimal solution for MTi.

The hypothesis (Assumption 6.5 or 6.6) implies that the sequence $\{\mu_i\}$ is tight; see Lemma 2.4 and Remark 2.5 in [15]. Hence, by Prohorov's theorem there are a subsequence $\{\mu_m^*\}$ of $\{\mu_i\}$ and a p.m. μ^* on $\mathbb{B}(X \times Y)$ such that $\{\mu_m^*\}$ converges weakly to μ^*.

In addition, by Lemma 2.7 in [15], the marginals $\Pi_k\mu_m^*$ converge weakly to the marginal $\Pi_k\mu^*$ for $k = 1, 2$, which implies that $\Pi_1\mu^* = \nu_1$ and $\Pi_2\mu^* = \nu_2$; that is, μ^* is a feasible solution for MT.

Suppose now that μ is an optimal solution for MT. Then by Theorem 6.4 there is a sequence $\{\mu_m\}$ of p.m.'s on $\mathbb{B}(X \times Y)$ such that $\{\mu_m\}$ converges weakly to μ, and μ_m is a feasible solution for MTm. Since μ is an optimal solution for MT and μ^* is feasible for MT, we have

$$\langle \mu, c\rangle \leq \langle \mu^*, c\rangle. \tag{6.11}$$

On the other hand, for each m, we have that

$$\langle \mu_m, c_m\rangle \geq \langle \mu_m^*, c_m\rangle,$$

whereas by Theorem 6.4

$$\langle \mu, c\rangle = \lim_{m\to\infty}\langle \mu_m, c_m\rangle \geq \limsup_{m\to\infty}\langle \mu_m^*, c_m\rangle \geq 0. \tag{6.12}$$

Now, pick an arbitrary $\epsilon > 0$. Since $c_n \uparrow c$, there exists an integer n such that

$$\langle \mu^*, c\rangle \geq \langle \mu^*, c_n\rangle \geq \langle \mu^*, c\rangle - \varepsilon. \tag{6.13}$$

For each $m \geq n$ we have

$$0 \leq \langle \mu_m^*, c_n\rangle \leq \langle \mu_m^*, c_m\rangle,$$

and as $\mu_m^* \to \mu^*$ we obtain

$$\langle \mu^*, c_n \rangle = \lim_{m\to\infty} \langle \mu_m, c_n \rangle \le \limsup_{m\to\infty} \langle \mu_m^*, c_m \rangle.$$

Therefore, by (6.13), it follows that

$$\langle \mu^*, c \rangle \le \limsup_{m\to\infty} \langle \mu_m^*, c_m \rangle + \varepsilon.$$

Consequently, as ε was arbitrary,

$$\langle \mu^*, c \rangle \le \limsup_{m\to\infty} \langle \mu_m^*, c_m \rangle,$$

which together with (6.12) gives

$$\langle \mu^*, c \rangle \le \langle \mu, c \rangle. \tag{6.14}$$

Thus, from (6.11) and (6.14) we obtain $\langle \mu, c \rangle = \langle \mu^*, c \rangle$; that is, μ^* is an optimal solution for MT. □

Remark 6.8. Since the measures ν_1^i and ν_2^i have finite support, MTi is a classical transportation problem [23, 24].

7. Concluding remarks. In the previous sections we presented two new approaches to obtain the solvability of the MT problem in metric spaces. In the first one, we transform MT into an equivalent optimization problem, MTK, using randomized strategies. In the second approach we obtain the solution of MT, and the corresponding optimal value, as the limit of a sequence of MT problems with marginals supported on finite sets.

The second approach requires the cost function $c(x, y)$ to be continuous, but this condition can be weakened. For instance, if c is l.s.c. and bounded below, say, nonnegative, then it is the limit of a nondecreasing sequence of continuous bounded functions $c_n \ge 0$. In this case, we might solve the MT problem for c_n and then try to show that, as $n \to \infty$, in the limit we obtain the solution of the original MT problem. Another case is the following.

Suppose that the Assumption 5.4 holds and that Y is a σ-compact space. Let $\{X_l\}$ be a countable partition of X such that each X_l is a σ-compact subspace of X and the cost function c restricted to each $X_l \times Y$ is continuous. Without loss of generality we may take $\alpha_l := \nu_1(X_l) > 0$ for all $l = 1, 2, \ldots$. Let us define

$$\nu_1^l(B) := (1/\alpha_l)\nu_1(B \cap X_l) \quad \forall B \in \mathbb{B}(X), \quad l = 1, 2, \ldots.$$

Consider the following problem:

MTl Minimize $\langle \mu, c \rangle$

subject to $\Pi_1 \mu = \nu_1^l, \quad \Pi_2 \mu = \nu_2, \quad \mu \ge 0.$

Now, for each feasible measure μ to the MT problem, we obtain the decomposition $\mu = \sum_l \alpha_l \mu_l$, where $\mu_l(E) := \mu(E \cap (X_l \times Y))/\alpha_l$ for E in $\mathbb{B}(X \times X)$ and $l = 1, 2, \ldots$, and, moreover, μ_l is feasible for MTl.

In fact, μ is an optimal solution to MT if and only if μ_l is optimal to MTl for all $l \ge 1$. Hence we may apply the approximation scheme for each $X_l \times Y$.

REFERENCES

[1] T. Abdellaoui, *Optimal solution of a Monge–Kantorovich transportation problem*, J. Comput. Appl. Math., 96 (1998), pp. 149–161.

[2] E. J. Anderson and P. Nash, *Linear Programming in Infinite–Dimensional Spaces*, Wiley, Chichester, UK, 1987.

[3] J. Alvarez-Mena and O. Hernández-Lerma, *Convergence and approximation of optimization problems*, SIAM J. Optim., 15 (2005), pp. 527–539.

[4] R. J. Aumann and M. Perles, *A variational problem arising in economics*, J. Math. Anal. Appl., 11 (1965), pp. 488–503.

[5] E. J. Balder, *Lectures on Young Measures*, Preprint 9517, Cahiers de Mathématiques de la Décision, CEREMADE, Université Paris IX Dauphine, Paris, 1995.

[6] E. J. Balder, *Generalized equilibrium results for games with incomplete information*, Math. Oper. Res., 13 (1988), pp. 265–276.

[7] S. Cambanis, G. Simons, and W. Stout, *Inequalities for $Ek(X, Y)$ when the marginals are fixed*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 36 (1976), pp. 285–294.

[8] G. Choquet, *Lectures on Analysis*, Vol. 2, W. A. Benjamin, New York, 1969.

[9] J. A. Cuesta-Albertos, C. Matrán, S. T. Rachev, and L. R. Rüschendorf, *Mass transportation problems in probability theory*, Math. Sci., 21 (1996), pp. 37–72.

[10] J. A. Cuesta-Albertos and A. Tuero-Díaz, *A characterization for the solution of the Monge-Kantorovich mass transference problem*, Statist. Probab. Lett., 6 (1993), pp. 147–152.

[11] W. Gangbo, *The Monge mass transfer problem and its applications*, in Proceedings of the NSF-CBMS Conference, Contemp. Math. 226, AMS, Providence, RI, 1999, pp. 79–104.

[12] J. González-Hernández and J. R. Gabriel, *On the consistency of the mass transfer problem*, Oper. Res. Lett., 34 (2006), pp. 382–386.

[13] J. González-Hernández and O. Hernández-Lerma, *Extreme points of sets of randomized strategies in constrained optimization and control problems*, SIAM J. Optim., 15 (2005), pp. 1085–1104.

[14] O. Hernández-Lerma and J. B. Lasserre, *Discrete-Time Markov Control Processes: Basic Optimality Criteria*, Springer-Verlag, New York, 1996.

[15] O. Hernández-Lerma and J. R. Gabriel, *Strong duality of the Monge–Kantorovich mass transfer problem in metric spaces*, Math. Z., 239 (2002), pp. 579–591.

[16] P. Jiménez-Guerra and B. Rodríguez-Salinas, *A general solution of the Monge-Kantorovich mass-transfer problem*, J. Math. Anal. Appl., 202 (2002), pp. 492–510.

[17] L. V. Kantorovich, *On the translocation of masses*, C.R. (Doklady) Acad. Sci. URSS (N.S.), 37 (1942), pp. 199–201.

[18] R. J. McCann, *Exact solutions to the transportation problem on the line*, R. Soc. Lond., Proc. Ser. A Math. Phys. Eng. Sci., 455 (1999), pp. 1341–1380.

[19] G. Monge, *Mémoire sur la théorie des déblais et rémblais*, Histoire de l'Académie Royale des Sciences de Paris, avec les Mémoires de Mathématiques et de Physique pour la même année, 1781, pp. 666–704.

[20] K. R. Parthasarathy, *Probability Measures on Metrics Spaces*, Academic Press, New York, 1967.

[21] S. T. Rachev, *The Monge-Kantorovich mass transference problem and its stochastic application*, Theory Probab. Appl., 29 (1984), pp. 647–676.

[22] S. T. Rachev, *Probability Metrics and the Stability of Stochastic Models*, Wiley, New York, 1991.

[23] S. T. Rachev and L. Rüschendorf, *Mass Transportation Problems. Vol.* I. *Theory*, Springer-Verlag, New York, 1998.

[24] S. T. Rachev and L. Rüschendorf, *Mass Transportation Problems. Vol.* II. *Applications*, Springer-Verlag, New York, 1998.

[25] P. S. Ruzankin, *Construction of the optimal joint distribution of two random variables*, Theory Probab. Appl., 46 (2002), pp. 316–334.

[26] S. Y. Wu, *Extremal points and an algorithm for a class of continuous transportation problems*, J. Inf. and Optim. Sci., 13 (1992), pp. 97–106.

[27] M. E. Yaari, *On the existence of an optimal plan in continuous-time allocation processes*, Econometrica, 32 (1964), pp. 576–590.

SIAM J. OPTIM.
Vol. 17, No. 2, pp. 500–510
© 2006 Society for Industrial and Applied Mathematics

OPTIMALITY CONDITIONS FOR QUASICONVEX PROGRAMS*

NGUYEN THI HONG LINH† AND JEAN-PAUL PENOT‡

Abstract. We present necessary and sufficient optimality conditions for a problem with a convex set constraint and a quasiconvex objective function. We apply the obtained results to a mathematical programming problem involving quasiconvex functions.

Key words. convex programming, normal cone, optimality conditions, quasiconvex function, quasiconvex programming, subdifferential

AMS subject classifications. 90C26, 26B25, 52A41

DOI. 10.1137/040621843

1. Introduction. It is the purpose of this paper to present some optimality conditions for constrained optimization problems under generalized convexity assumptions. We essentially deal with quasiconvex functions, i.e., functions whose sublevel sets are convex. Such functions form the main class of generalized convex functions and are widely used in mathematical economics. We do not assume that the data of the problem are smooth. Thus, we replace the derivatives appearing in the classical results by subdifferentials. We only use the adapted subdifferentials of quasiconvex analysis, namely the Plastria subdifferential [35] and the infradifferential or Gutiérrez subdifferential [10]. Because these subdifferentials are useful for algorithmic purposes [20], [35] and have links with duality [17, Prop. 6.1], [29], [30], [31], their use in problems in which quasiconvexity properties occur seems to be sensible, although their calculus rules are not as rich as in the case of convex analysis [24], [33]. In [17, Prop. 6.1] Martínez-Legaz presented a result of the Kuhn–Tucker type using these subdifferentials; in [18, Thm. 4.1] and [19, Prop. 6.3] variants of these subdifferentials are used. In each of these results, a Slater condition and a semistrict quasiconvexity assumption (called strict quasiconvexity in [15]) are imposed. Here we essentially assume the functions are quasiconvex and we deduce the results from optimality conditions for problems with set constraints. We also point out the link with the differentiable case dealt with in the pioneer paper [1]. We do not make a comparison with results using the all-purpose subdifferentials of nonsmooth analysis (see [3], [22]). The reason is that these subdifferentials are local, whereas the ones we use are of global character; intermediate notions are presented in [4], [21], and [25]. On the other hand, our necessary conditions refine conditions using normal cones or subdifferentials related to normal cones to sublevels sets as the Greenberg–Pierskalla subdifferential [9] as in [26]. Numerous papers deal with optimality conditions for constrained problems under generalized convexity conditions (see [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [26], [27], [28], [29], [30], [31], [34], [36] for example). A link with the case we are dealing with here, which is essentially the quasiconvex case, could be found by assuming that the Gutiérrez or Plastria

*Received by the editors December 31, 2004; accepted for publication (in revised form) January 16, 2006; published electronically August 16, 2006.
http://www.siam.org/journals/siopt/17-2/62184.html
†Department of Mathematics, University of Natural Sciences, Ho Chi Minh City, Vietnam (honglinh98t1@yahoo.com).
‡Laboratoire de Mathématiques, CNRS UMR 5142, Faculté des Sciences, Av. de l'Université 64000 Pau, France (jean-paul.penot@univ-pau.fr).

subdifferentials are nonempty at each point. However, we do not wish to impose such a restrictive condition.

2. Preliminaries: Gutiérrez and Plastria functions. In what follows X is a normed vector space (n.v.s.) with closed unit ball B_X. We denote by $N(C,x)$ the *normal cone* at $x \in X$ to a convex subset C of X given by

$$N(C,x) := \{x^* \in X^* : \forall u \in C,\ \langle x^*, u-x\rangle \le 0\}.$$

When $x \in C$, it is the polar cone of the *tangent cone* $T(C,x)$ which is the closure of $\mathbb{R}_+(C-x)$.

A function $f : X \to \mathbb{R}_\infty := \mathbb{R} \cup \{+\infty\}$ is said to be *quasiconvex* if for each $\overline{x} \in X$ its sublevel set

$$S_f(\overline{x}) := \{x \in X : f(x) \le f(\overline{x})\}$$

is convex, or, equivalently, if for each $r \in \mathbb{R}$ the strict sublevel set $S_f^<(r) := \{x \in X : f(x) < r\}$ is convex.

Recall that the *lower subdifferential*, or *Plastria subdifferential* of a function $f : X \to \overline{\mathbb{R}}$ on a Banach space X at some point $\overline{x}$ of its domain $\operatorname{dom} f := \{x \in X : f(x) \in \mathbb{R}\}$ is the set

$$\partial^< f(\overline{x}) := \{\overline{x}^* \in X^* : \forall x \in S_f^<(\overline{x}),\ f(x) - f(\overline{x}) \ge \langle \overline{x}^*, x-\overline{x}\rangle\},$$

where $S_f^<(\overline{x}) := S_f^<(f(\overline{x}))$ is the strict sublevel set of f at $\overline{x}$. We will also use the following variant, called the *infradifferential* [10] or *Gutiérrez subdifferential*:

$$\partial^\le f(\overline{x}) := \{\overline{x}^* \in X^* : \forall x \in S_f(\overline{x}),\ f(x) - f(\overline{x}) \ge \langle \overline{x}^*, x-\overline{x}\rangle\}.$$

If no point of the level set $L_f(\overline{x}) := f^{-1}(f(\overline{x}))$ is a local minimizer of f, we have $\partial^< f(\overline{x}) = \partial^\le f(\overline{x})$. If $f(\overline{x}) > \inf f(X)$, this equality also holds when f is radially continuous (i.e., continuous along lines) and *semistrictly quasiconvex* in the sense that when $f(x) < f(y)$ one has $f((1-t)x+ty) < f(y)$ for any $t \in (0,1)$; in particular, this equality holds for convex continuous functions. In spite of the fact that the preceding constructions have a close similarity with the Fenchel subdifferential, they differ by significant features. In particular $\partial^< f(\overline{x})$ and $\partial^\le f(\overline{x})$ are unbounded or empty: it is easy to check that they are shady in the sense that they are closed, convex, and stable by homotheties with rate t greater than 1 since for $\overline{x}^* \in \partial^< f(\overline{x})$ and $x \in S_f^<(\overline{x})$ we have $\langle t\overline{x}^*, x-\overline{x}\rangle \le \langle \overline{x}^*, x-\overline{x}\rangle \le 0$ and a similar observation when $\overline{x}^* \in \partial^\le f(\overline{x})$ and $x \in S_f(\overline{x})$.

We say that f is a *Plastria function* at $\overline{x}$ if its strict sublevel set $S_f^<(\overline{x})$ is convex and such that

$$N(S_f^<(\overline{x}), \overline{x}) = \mathbb{R}_+ \partial^< f(\overline{x}). \tag{1}$$

We say that f is a *Gutiérrez function* at $\overline{x}$ if its sublevel set $S_f(\overline{x})$ is convex and such that

$$N(S_f(\overline{x}), \overline{x}) = \mathbb{R}_+ \partial^\le f(\overline{x}). \tag{2}$$

Since $\partial^< f(\overline{x})$ and $\partial^\le f(\overline{x})$ are shady in the sense that they are stable under multiplication by any $t \in [1,\infty)$, relations (1) and (2) are equivalent to $N(S_f^<(\overline{x}), \overline{x}) =$

$[0,1]\partial^< f(\overline{x})$ and $N(S_f(\overline{x}),\overline{x}) = [0,1]\partial^{\leq} f(\overline{x})$, respectively. These conditions being rather stringent, it may be useful to replace f by its extension by $+\infty$ outside some ball (see [20] for the case of nonconvex quadratic functions). However, we provide three criteria. The first one deals with convex transformable functions, an important class of quasiconvex functions.

PROPOSITION 1. *Let f be a proper convex function and let $\overline{x} \in \operatorname{dom} f := f^{-1}(\mathbb{R})$ be such that $f(\overline{x}) > \inf f(X)$ and $\mathbb{R}_+(\operatorname{dom} f - \overline{x}) = X$. Then f is a Gutiérrez function and a Plastria function at $\overline{x}$:*

$$N(S_f(\overline{x}),\overline{x}) = \mathbb{R}_+\partial^{\leq} f(\overline{x}) = \mathbb{R}_+\partial f(\overline{x}) = \mathbb{R}_+\partial^< f(\overline{x}) = N(S_f^<(\overline{x}),\overline{x}).$$

More generally, if $f := h \circ g$, where $g : \mathbb{R} \to \mathbb{R}_\infty$ is a convex function and $h : T \to \mathbb{R}_\infty$ is a strictly increasing function on some interval T of $\mathbb{R}_\infty$ containing $g(X)$, with $h(+\infty) = +\infty$, then f is a Gutiérrez function and a Plastria function at $\overline{x}$ provided $f(\overline{x}) > \inf f(X)$ and $\mathbb{R}_+(\operatorname{dom} f - \overline{x}) = X$. Moreover, $\mathbb{R}_+\partial^< f(\overline{x}) = \mathbb{R}_+\partial^< g(\overline{x}) = \mathbb{R}_+\partial^{\leq} g(\overline{x}) = \mathbb{R}_+\partial^{\leq} f(\overline{x})$.

Proof. By [32, Prop. 5.4] one has $N(S_f(\overline{x}),\overline{x}) = \mathbb{R}_+\partial f(\overline{x})$; here we use the fact that under the assumption $\mathbb{R}_+(\operatorname{dom} f - \overline{x}) = X$, we have $N(\operatorname{dom} f,\overline{x}) = \{0\}$. Since $\mathbb{R}_+\partial f(\overline{x}) \subset \mathbb{R}_+\partial^{\leq} f(\overline{x}) \subset N(S_f(\overline{x}),\overline{x})$, we get equality. Moreover, if $\overline{x}^* \in \partial^< f(\overline{x})$ and $x \in S_f(\overline{x})$, taking $z \in X$ such that $f(z) < f(\overline{x})$ and $t \in (0,1)$ we have $x_t := (1-t)x + tz \in S_f^<(\overline{x})$, hence

$$(1-t)f(x) + tf(z) \geq f(x_t) \geq f(\overline{x}) + \langle \overline{x}^*, x - \overline{x}\rangle,$$

and taking the limit as $t \to 0$, we get $f(x) \geq f(\overline{x}) + \langle \overline{x}^*, x - \overline{x}\rangle$, hence $\overline{x}^* \in \partial^{\leq} f(\overline{x})$. Moreover, the preceding argument shows that $S_f(\overline{x})$ is contained in the closure of $S_f^<(\overline{x})$, so that $N(S_f(\overline{x}),\overline{x}) = N(S_f^<(\overline{x}),\overline{x})$.

Now let $f := h \circ g$ be as in the second part of the statement. Since h is (strictly) increasing, we have $S_f(\overline{x}) = S_g(\overline{x})$, $S_f^<(\overline{x}) = S_g^<(\overline{x})$. Setting $\overline{r} := g(\overline{x})$, since $\partial^< h(\overline{r}) \subset (0,+\infty)$, using [33, Prop. 3.5], we have

$$N(S_f^<(\overline{x}),\overline{x}) = N(S_g^<(\overline{x}),\overline{x}) = \mathbb{R}_+\partial^< g(\overline{x}) = \mathbb{R}_+\partial^< h(\overline{r})\partial^< g(\overline{x}) \subset \mathbb{R}_+\partial^< f(\overline{x}),$$

hence equality, the reverse inclusion being obvious. The proof that f is a Gutiérrez function at $\overline{x}$ is similar. □

The second one is a slight improvement of previous results in [35], [17], [24]. It uses the now classical notion of calmness: $f : X \to \overline{\mathbb{R}}$ is said to be *calm* with rate c at $w \in X$ if $f(w)$ is finite and if

$$\forall x \in X \qquad\qquad f(x) - f(w) \geq -c\,\|x - w\|\,.$$

Such a condition is obviously satisfied if f is Lipschitzian with rate c or if f is Stepanovian (or stable) with rate c at w in the sense that $|f(x) - f(w)| \leq c\,\|x - w\|$ for any $x \in X$.

PROPOSITION 2. *Assume that f is radially continuous on X and calm with rate $c \in \mathbb{R}_+$ at each point of the level set $L_f(\overline{x}) := f^{-1}(f(\overline{x}))$ and that $S_f(\overline{x})$ and $S_f^<(\overline{x})$ are convex. Then*

$$N(S_f^<(\overline{x}),\overline{x})\backslash cB_X = \partial^< f(\overline{x})\backslash cB_X,$$
$$N(S_f(\overline{x}),\overline{x})\backslash cB_X = \partial^{\leq} f(\overline{x})\backslash cB_X.$$

If, moreover, $N(S_f^<(\overline{x}),\overline{x}) \neq \{0\}$*, then the function* f *is a Plastria function at* $\overline{x}$*, while if* $N(S_f(\overline{x}),\overline{x}) \neq \{0\}$*, then* f *is also a Gutiérrez function at* $\overline{x}$*.*

The condition $N(S_f(\overline{x}),\overline{x}) \neq \{0\}$ (or $N(S_f^<(\overline{x}),\overline{x}) \neq \{0\}$) is a rather mild condition when X is finite dimensional. However, when X is infinite dimensional, it may occur that a closed convex set $C \neq X$ is such that $N(C,\overline{x}) = \{0\}$ for some $\overline{x} \in C \backslash \mathrm{int} C$.

Proof. Let us first prove that whenever $\overline{x}^* \in N'_f(\overline{x}) := N(S_f^<(\overline{x}),\overline{x})$ satisfies $\|\overline{x}^*\| := \lambda c$ for some $\lambda > 1$, then $\overline{x}^* \in \partial^< f(\overline{x})$. Since $N'_f(\overline{x})$ is a cone and $\partial^< f(\overline{x})$ is w*-closed, this will show that $N'_f(\overline{x}) \backslash cB_X \subset \partial^< f(\overline{x}) \backslash cB_X$ and that equality holds. Let $\overline{y}^* := c^{-1}\lambda^{-1}\overline{x}^*$. Given $x \in [f < f(\overline{x})]$ we have by assumption $t := \langle \overline{y}^*, x - \overline{x}\rangle < 0$. Let us choose $v \in X$ such that $\|v\| \leq \lambda$ and $\langle \overline{y}^*, v\rangle = 1$ and set $w = x - \overline{x} - tv$. Then $\langle \overline{y}^*, w\rangle = 0$. Let us show that $f(\overline{x}+w) \geq f(\overline{x})$. In order to do so, we pick $z \in X$ such that $\langle \overline{y}^*, z\rangle > 0$; then for any $s > 0$, we have $\langle \overline{y}^*, w+sz\rangle > 0$, thus $f(\overline{x}+w+sz) \geq f(\overline{x})$. By radial continuity, $f(\overline{x}+w) \geq f(\overline{x})$.

Since $f\,|_{[x,\overline{x}+w]}$ is continuous and $f(x) < f(\overline{x}) \leq f(\overline{x}+w)$, there exists $x' \in [x, \overline{x}+w]$ such that $f(x') = f(\overline{x})$ and $f(x'') < f(\overline{x})$ for all $x'' \in [x, x')$. Then, since $\lambda > 1$, $\|\overline{y}^*\| = 1$ and $t = \langle \overline{y}^*, x - \overline{x}\rangle$, we have

$$\begin{aligned} f(x) - f(\overline{x}) = f(x) - f(x') &\geq -c\,\|x - x'\| \geq -c\,\|x - (\overline{x}+w)\| \\ &= -c\,\|tv\| = ct\,\|v\| \geq \lambda c\,\langle \overline{y}^*, x - \overline{x}\rangle = \langle \overline{x}^*, x - \overline{x}\rangle\,. \end{aligned}$$

Thus $f(x) - f(\overline{x}) \geq \langle \overline{x}^*, x - \overline{x}\rangle$ holds for all $x \in [f < f(\overline{x})]$ and $\overline{x}^* \in \partial^< f(\overline{x})$.

Assume now that $\overline{x}^* \in N(S_f(\overline{x}),\overline{x})$ is such that $\|\overline{x}^*\| \geq c$; then $\overline{x}^* \in N(S_f^<(\overline{x}),\overline{x})$, hence $\overline{x}^*$ is such that $f(x) - f(\overline{x}) \geq \langle \overline{x}^*, x - \overline{x}\rangle$ for each $x \in S_f^<(\overline{x})$, while if $f(x) = f(\overline{x})$, then

$$f(x) - f(\overline{x}) = 0 \geq \langle \overline{x}^*, x - \overline{x}\rangle\,.$$

This proves that $\overline{x}^* \in \partial^\leq f(\overline{x})$.

Now, given $\overline{x}^* \in N(S_f^<(\overline{x}),\overline{x})$, $\overline{x}^* \neq 0$, we can find $r > 0$ such that $\overline{x}^* = r\overline{w}^*$ for some $\overline{w}^* \in N(S_f^<(\overline{x}),\overline{x})$ satisfying $\|\overline{w}^*\| \geq c$. Thus $\overline{x}^* \in \mathbb{R}_+ \partial^< f(\overline{x})$. Since $\partial^< f(\overline{x})$ is nonempty by what precedes, we also have $0 \in \mathbb{R}_+ \partial^< f(\overline{x})$, hence $N(S_f^<(\overline{x}),\overline{x}) \subset \mathbb{R}_+ \partial^< f(\overline{x})$. The reverse inclusion being obvious, we get $N(S_f^<(\overline{x}),\overline{x}) = \mathbb{R}_+ \partial^< f(\overline{x})$. The equality $N(S_f(\overline{x}),\overline{x}) = \mathbb{R}_+ \partial^\leq f(\overline{x})$ is obtained similarly. □

The third criterion uses a differentiability assumption and brings some supplement to [23, Prop. 15].

Proposition 3. *Let* f *be quasiconvex, differentiable at* $\overline{x}$ *with a nonzero derivative. If* $\partial^< f(\overline{x})$ *is nonempty, then* f *is a Plastria function at* $\overline{x}$ *and there exists some* $\overline{r} \geq 1$ *such that* $\partial^< f(\overline{x}) = [\overline{r}, \infty) f'(\overline{x})$*. If* $\partial^\leq f(\overline{x})$ *is nonempty, then* f *is a Gutiérrez function at* $\overline{x}$ *and there exists some* $\overline{s} \geq 1$ *such that* $\partial^\leq f(\overline{x}) = [\overline{s}, \infty) f'(\overline{x})$*.*

Proof. Let us first prove that if f is quasiconvex, differentiable at $\overline{x}$ with a nonzero derivative one has $N(S_f^<(\overline{x}),\overline{x}) = \mathbb{R}_+ f'(\overline{x})$. We first observe that

$$f'(\overline{x})^{-1}((-\infty,0)) \subset T(S_f^<(\overline{x}),\overline{x}) \subset T(S_f(\overline{x}),\overline{x}) \subset f'(\overline{x})^{-1}((-\infty,0]).$$

Since $f'(\overline{x}) \neq 0$, we can find some $w \in X$ with $f'(\overline{x})w < 0$. Then, for any $v \in f'(\overline{x})^{-1}((-\infty,0])$ and any sequence $(r_n) \to 0_+$ we have

$$v_n := v + r_n w \in f'(\overline{x})^{-1}((-\infty,0)) \subset T(S_f^<(\overline{x}),\overline{x}),$$

and since $(v_n) \to v$ and $T(S_f^<(\overline{x}),\overline{x})$ is closed, we get $v \in T(S_f^<(\overline{x}),\overline{x})$. Thus $T(S_f^<(\overline{x}),\overline{x}) = T(S_f(\overline{x}),\overline{x}) = f'(\overline{x})^{-1}((-\infty,0])$. Then, the Farkas lemma ensures

that

$$N(S_f^<(\overline{x}),\overline{x}) = N(S_f(\overline{x}),\overline{x}) = \mathbb{R}_+ f'(\overline{x}).$$

Let us now assume that $0 \notin \partial^< f(\overline{x}) \neq \varnothing$ and let us pick some $\overline{x}^* \in \partial^< f(\overline{x})$. By what precedes, for any $v \in X$ such that $f'(\overline{x})v \leq 0$ we can find sequences $(t_n) \to 0_+$, $(v_n) \to v$ such that $f(\overline{x} + t_n v_n) < f(\overline{x})$ for each n. Then we get

$$\langle \overline{x}^*, v\rangle \leq \lim_n \frac{1}{t_n}\left(f(\overline{x} + t_n v_n) - f(\overline{x})\right) = f'(\overline{x})v \leq 0,$$

so that, by the Farkas lemma again, there exists some $r \in \mathbb{R}_+$ such that $\overline{x}^* = rf'(\overline{x})$. In fact, the preceding inequalities (taken with some v such that $f'(\overline{x})v < 0$) show that $r \geq 1$. Let $\overline{r} := \inf\{r \in \mathbb{R} : \exists \overline{x}^* \in \partial^< f(\overline{x}),\ \overline{x}^* = rf'(\overline{x})\}$. We have $\overline{r} \geq 1$ by what precedes, and, by closedness of $\partial^< f(\overline{x})$, $\overline{x}^* = \overline{r}f'(\overline{x})$ for some $\overline{x}^* \in \partial^< f(\overline{x})$. Thus, $\partial^< f(\overline{x}) \subset [\overline{r},\infty)f'(\overline{x})$ and since $\overline{r}f'(\overline{x}) = \overline{x}^* \in \partial^< f(\overline{x})$ and $[1,\infty)\partial^< f(\overline{x}) \subset \partial^< f(\overline{x})$, we get $\partial^< f(\overline{x}) = [\overline{r},\infty)f'(\overline{x})$. It follows that $\mathbb{R}_+\partial^< f(\overline{x}) = \mathbb{R}_+ f'(\overline{x})$.

The proof for $\partial^\leq f(\overline{x})$ is similar. □

The following example shows that one may have $\overline{r} > 1$.

Example 1. Let $X = \mathbb{R}$ and for $c < 0$ let f be given by $f(x) = c^3/3$ for $x \in (-\infty, c)$, $f(x) = x^3/3$ for $x \geq c$. Then, for $\overline{x} = 1$, we have $\partial^< f(\overline{x}) = \partial^\leq f(\overline{x}) = [\overline{r},\infty)$ with $\overline{r} = \max(1, (1 + c + c^2)/3)$.

A localization of the preceding concepts may enlarge the range of the optimality conditions which follow. Let us define the local normal cone to C at $\overline{x}$ as

$$N_{loc}(C,\overline{x}) := \bigcup_{r>0} N(C \cap B(\overline{x},r),\overline{x}),$$

where $B(\overline{x},r)$ denotes the open ball with center $\overline{x}$ and radius r. When C is convex, we have $N_{loc}(C,\overline{x}) = N(C,\overline{x})$. We also define the *local Gutiérrez subdifferential* and the *local Plastria subdifferential* of f at $\overline{x}$ by

$$\partial^\leq_{loc} f(\overline{x}) := \{\overline{x}^* \in X^* : \exists r > 0, \forall x \in S_f(\overline{x}) \cap B(\overline{x},r),\ f(x) - f(\overline{x}) \geq \langle \overline{x}^*, x - \overline{x}\rangle\},$$
$$\partial^<_{loc} f(\overline{x}) := \{\overline{x}^* \in X^* : \exists r > 0, \forall x \in S_f^<(\overline{x}) \cap B(\overline{x},r),\ f(x) - f(\overline{x}) \geq \langle \overline{x}^*, x - \overline{x}\rangle\},$$

respectively. We say that f is *locally a Plastria function at* $\overline{x}$ if there exists some $r > 0$ such that $S_f^<(\overline{x}) \cap B(\overline{x},r)$ is convex and if

$$N_{loc}(S_f^<(\overline{x}),\overline{x}) = \mathbb{R}_+ \partial^<_{loc} f(\overline{x}).$$

Locally Gutiérrez functions can be defined similarly.

3. Optimality conditions for constrained problems. In the present section we consider the minimization problem

$$(\mathcal{C}) \quad \text{minimize } f(x) \text{ subject to } x \in C,$$

where $f : X \to \overline{\mathbb{R}}$ is a function on the n.v.s. X and C is a convex subset of X.

Proposition 4. *Let f be an u.s.c. Plastria function at some solution $\overline{x}$ to $(\mathcal{C})$ which is not a local minimizer of f. Then one has*

$$0 \in \partial^< f(\overline{x}) + N(C,\overline{x}). \tag{3}$$

Proof. Since f is quasiconvex and u.s.c., the strict sublevel set $S_f^<(\overline{x})$ is open and convex; it is nonempty since $\overline{x}$ is not a minimizer of f. Since $\overline{x}$ is a solution to $(\mathcal{C})$, this sublevel set is disjoint from C. Thus, the Hahn–Banach separation theorem yields some $c \in \mathbb{R}$ and u^* in the unit sphere of X^* such that

$$\langle u^*, x - \overline{x}\rangle \geq c \geq \langle u^*, w - \overline{x}\rangle \qquad \forall w \in S_f^<(\overline{x}), \forall x \in C. \tag{4}$$

Taking $x = \overline{x}$, we see that $c \leq 0$. Moreover, since $\overline{x}$ is not a local minimizer of f, there exists a sequence $(w_n) \to \overline{x}$ such that $w_n \in S_f^<(\overline{x})$ for each n. Therefore $c = 0$. Then we have $u^* \in N(S_f^<(\overline{x}), \overline{x}) = \mathbb{R}_+ \partial^< f(x)$ and since $u^* \neq 0$, we can find $\overline{x}^* \in \partial^< f(\overline{x})$ and $r \in \mathbb{R}_+$ such that $\overline{x}^* = ru^*$. On the other hand, the first inequality of (4) means that $-u^* \in N(C, \overline{x})$. Thus, $\overline{x}^* + r(-u^*) = 0$ and (3) is satisfied. □

Example 2. The example (taken from [26]) of the function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x$ for $x \in (-\infty, 0)$, $f(x) = 0$ for $x \in [0, 1]$, $f(x) = x - 1$ for $x \in (1, +\infty)$ and $C := \mathbb{R}_+$, $\overline{x} = 1$ shows that the assumption that $\overline{x}$ is not a local minimizer cannot be dispensed within the preceding statement.

Now let us give a sufficient condition. Observe that no assumption is required on f besides finiteness at $\overline{x}$.

PROPOSITION 5. *Let $f : X \to \mathbb{R} \cup \{\infty\}$ be an arbitrary function finite at $\overline{x}$ satisfying relation* (3). *Then $\overline{x}$ is a solution to $(\mathcal{C})$.*

Proof. Let $\overline{x}^* \in \partial^< f(\overline{x})$ be such that $-\overline{x}^* \in N(C, \overline{x})$. Assume that $\overline{x}$ is not a solution to $(\mathcal{C})$: there exists some $x \in C$ such that $f(x) < f(\overline{x})$. Then one has, by the definitions of $\partial^< f(\overline{x})$ and $N(C, \overline{x})$,

$$0 > f(x) - f(\overline{x}) \geq \langle \overline{x}^*, x - \overline{x}\rangle,$$
$$\langle \overline{x}^*, x - \overline{x}\rangle \geq 0,$$

a contradiction. □

Let us observe that the preceding sufficient condition can also be derived from the one in [26, Prop. 2.1] which uses the Greenberg–Pierskalla subdifferential

$$\partial^* f(\overline{x}) := \{\overline{x}^* \in X^* : \forall x \in S_f^<(\overline{x})\ \langle \overline{x}^*, x - \overline{x}\rangle < 0\}$$

since $\partial^< f(\overline{x}) \subset \partial^* f(\overline{x})$. On the other hand, the necessary condition in Proposition 4 is more precise than the necessary condition in [26, Prop. 2.2].

A slight supplement to the preceding results can be given. It deals with *strict solutions* to $(\mathcal{C})$, i.e., points $\overline{x} \in C$ such that $f(\overline{x}) < f(x)$ for each $x \in C \backslash \{\overline{x}\}$. For the sufficient condition we assume that C is *strictly convex at* $\overline{x}$ in the sense that $\langle \overline{x}^*, x - \overline{x}\rangle < 0$ for every $x \in C \backslash \{\overline{x}\}$ and $\overline{x}^* \in N(C, \overline{x}) \backslash \{0\}$. Observe that if $N(C, \overline{x}) \backslash \{0\}$ is nonempty (in particular if C is a convex subset of a finite dimensional space) and if C is strictly convex at $\overline{x}$, then $\overline{x}$ is an extremal point of C (i.e., $C \backslash \{\overline{x}\}$ is convex).

PROPOSITION 6. *Given a function $f : X \to \mathbb{R} \cup \{\infty\}$ finite at $\overline{x}$ and a subset C of X which is strictly convex at $\overline{x}$, the following relation implies that $\overline{x}$ is a strict solution to $(\mathcal{C})$ or a global minimizer of f on X:*

$$0 \in \partial^{\leq} f(\overline{x}) + N(C, \overline{x}). \tag{5}$$

Conversely, when X is finite dimensional, C is a convex subset of X not reduced to $\{\overline{x}\}$, $\overline{x}$ is an extremal point of C, and f is a Gutiérrez function at $\overline{x}$; relation (5) *is necessary in order that $\overline{x}$ be a strict solution to $(\mathcal{C})$ or a global minimizer of f on X.*

Proof. Assume relation (5) holds and C is strictly convex at $\overline{x}$. If $\overline{x}$ is not a global minimizer of f on X there exists some $\overline{x}^* \in \partial^{\leq} f(\overline{x})$ such that $-\overline{x}^* \in N(C,\overline{x})$ and $\overline{x}^* \neq 0$. Then, if $x \in C\backslash\{\overline{x}\}$ is such that $f(x) \leq f(\overline{x})$ we have $\langle \overline{x}^*, x-\overline{x}\rangle \leq f(x) - f(\overline{x}) \leq 0$ since $\overline{x}^* \in \partial^{\leq} f(\overline{x})$ and $\langle -\overline{x}^*, x-\overline{x}\rangle < 0$ since $-\overline{x}^* \in N(C,\overline{x})\backslash\{0\}$, a contradiction. Thus $\overline{x}$ is a strict solution to $(\mathcal{C})$.

When $\overline{x}$ is a strict solution to $(\mathcal{C})$, the sets $C\backslash\{\overline{x}\}$ and $S_f(\overline{x})$ are disjoint. If, moreover, f is a Gutiérrez function at $\overline{x}$ and $\overline{x}$ is an extremal point of C but is not a global minimizer of f on X, and $C \neq \{\overline{x}\}$, then these sets are convex and nonempty. Thus, when X is finite dimensional, a separation theorem yields some $c \in \mathbb{R}$ and u^* in the unit sphere of X^* such that

$$\langle u^*, x-\overline{x}\rangle \geq c \geq \langle u^*, w-\overline{x}\rangle \qquad \forall w \in S_f(\overline{x}), \forall x \in C\backslash\{\overline{x}\}. \tag{6}$$

Since x can be arbitrarily close to $\overline{x}$, we have $c \leq 0$. On the other hand, since we can take $w = \overline{x}$, we have $c \geq 0$, hence $c = 0$. Thus $-u^* \in N(C,\overline{x})$ and $u^* \in N(S_f(\overline{x}),\overline{x}) = \mathbb{R}_+ \partial^{\leq} f(\overline{x})$ since f is a Gutiérrez function at $\overline{x}$. Since $u^* \neq 0$, one can find $r > 0$ and $\overline{x}^* \in \partial^{\leq} f(\overline{x})$ such that $u^* = r\overline{x}^*$ and $-\overline{x}^* \in N(C,\overline{x})$, so that relation (5) holds. When $\overline{x}$ is a global minimizer of f on X, we have $0 \in \partial^{\leq} f(\overline{x}) \cap (-N(C,\overline{x}))$. ∎

Now, let us give conditions for local minimization.

PROPOSITION 7. *Let f be an u.s.c. locally Plastria function at some local solution $\overline{x}$ to $(\mathcal{C})$ which is not a local minimizer of f. Then one has*

$$0 \in \partial^{<}_{loc} f(\overline{x}) + N(C,\overline{x}). \tag{7}$$

Conversely, for any function f finite at $\overline{x}$ which satisfies relation (7), $\overline{x}$ is a solution to $(\mathcal{C})$.

Proof. By assumption, we can find $r > 0$ such that $\overline{x}$ is a minimizer of f on $C \cap V$, where $V := B(\overline{x}, r)$. Taking a smaller r if necessary and setting $f_V(x) = f(x)$ if $x \in V$, $f_V(x) = +\infty$ if $x \in X\backslash V$, we may assume that f_V is a Plastria function at $\overline{x}$. Then relation (7) follows from Proposition 4.

The converse assertion follows from the sufficient condition and from the observation that if $\overline{x}^* \in \partial^{<}_{loc} f(\overline{x})$, then there is some neighborhood V of $\overline{x}$ such that $\overline{x}^* \in \partial^{<} f_V(\overline{x})$. ∎

4. Necessary condition for the mathematical programming problem. Let us consider now the case in which the constraint set C is defined by a finite family of inequalities, so that problem $(\mathcal{C})$ turns into the mathematical programming problem

$$(\mathcal{M}) \qquad \text{minimize } f(x) \text{ subject to } x \in C := \{x \in X : \ g_1(x) \leq 0, \ldots, g_n(x) \leq 0\}.$$

Let us first consider the case of a single constraint.

LEMMA 8. *Let $\overline{x}$ be a solution to $(\mathcal{M})$ in which $g_1 = \cdots = g_n = g$ and $\overline{x}$ is not a local minimizer of f. Assume that f is a Plastria function at $\overline{x}$ and that g is u.s.c. at $\overline{x}$ and a Gutiérrez function at $\overline{x}$. Then $g(\overline{x}) = 0$ and there exists some $y \in \mathbb{R}_+$ such that*

$$0 \in \partial^{<} f(\overline{x}) + y\partial^{\leq} g(\overline{x}).$$

Proof. By Proposition 4, there exists $\overline{x}^* \in \partial^{<} f(\overline{x})$ such that $-\overline{x}^* \in N(C,\overline{x})$. If $g(\overline{x}) < 0$, since g is u.s.c. at $\overline{x}$, $\overline{x}$ belongs to the interior of C, hence $\overline{x}$ is a local minimizer of f, and our assumption discards that case. Thus $g(\overline{x}) = 0$, and since g is

a Gutiérrez function at $\overline{x}$, we have $N(C,\overline{x}) = \mathbb{R}_+\partial^{\leq} g(\overline{x})$. Thus there exists $y \in \mathbb{R}_+$ such that $-\overline{x}^* \in y\partial^{\leq} g(\overline{x})$. □

Now let us turn to the general case. We will use the following lemma.

LEMMA 9. *Let $(g_i)_{i\in I}$ be a finite family of quasiconvex Gutiérrez functions at some $\overline{x} \in X$. For $i \in I$, let $C_i := g_i^{-1}((-\infty, 0])$. Assume g_i is u.s.c. at $\overline{x}$, $g_i(\overline{x}) = 0$ for each $i \in I$ and either*

(a) *there exist some $k \in I$ and some $z \in C_k$ such that $g_i(z) < 0$ for each $i \in I\backslash\{k\}$ (Slater condition), or*

(b) *C_i is closed for each $i \in I$ and $\mathbb{R}_+ \left(\Delta - \prod_{i\in I} C_i\right) = X^I$, where $\Delta := \{(x_i)_{i\in I} : \forall j, k \in I,\ x_j = x_k\}$ is the diagonal of X^I.*

Then, $h := \max_{i\in I} g_i$ is a Gutiérrez function at $\overline{x}$ and one has

$$\mathbb{R}_+\partial^{\leq} h(\overline{x}) = \sum_{i\in I} \mathbb{R}_+\partial^{\leq} g_i(\overline{x}). \tag{8}$$

Proof. In case (a) we have $C_k \cap \left(\bigcap_{i\in I\backslash\{k\}} \operatorname{int} C_i\right) \neq \varnothing$; hence, for $C := \bigcap_{i\in I} C_i$, we get

$$N(C,\overline{x}) = \overline{\operatorname{co}}\left(\bigcup_{i\in I} N(C_i,\overline{x})\right) = \sum_{i\in I} N(C_i,\overline{x}).$$

In case (b), this relation also holds by the Attouch–Brézis qualification condition [2]. Thus, since $\partial^{\leq} g_i(\overline{x}) \subset \partial^{\leq} h(\overline{x})$ and $\partial^{\leq} h(\overline{x})$ is convex,

$$\mathbb{R}_+\partial^{\leq} h(\overline{x}) \subset N(C,\overline{x}) = \sum_{i\in I} N(C_i,\overline{x}) = \sum_{i\in I} \mathbb{R}_+\partial^{\leq} g_i(\overline{x}) \subset \mathbb{R}_+\partial^{\leq} h(\overline{x}),$$

so that h is a Gutiérrez function at $\overline{x}$ and relation (8) holds. □

THEOREM 10. *Let $\overline{x}$ be a solution to $(\mathcal{M})$ which is not a local minimizer of f. Let $I := \{i \in \{1,\dots,n\} : g_i(\overline{x}) = 0\}$. Assume that one of the assumptions* (a) *or* (b) *of Lemma 9 is satisfied. Assume that f is a Plastria function at $\overline{x}$, $g_1,\dots,g_n$ are u.s.c. at $\overline{x}$ and that for every $i \in I$, g_i is a Gutiérrez functions at $\overline{x}$. Then, there exist some $y_1,\dots,y_n \in \mathbb{R}_+$ such that*

$$\begin{aligned} 0 &\in \partial^{<} f(\overline{x}) + y_1\partial^{\leq} g_1(\overline{x}) + \cdots + y_n\partial^{\leq} g_n(\overline{x}), \\ y_i g_i(\overline{x}) &= 0 \qquad\qquad \text{for } i = 1,\dots,n. \end{aligned}$$

Proof. Let $g := \max_{1\leq i\leq n} g_i$, $h := \max_{i\in I} g_i$, and let $D := h^{-1}((-\infty, 0])$. Then, for $i \in \{1,\dots,n\}\backslash I$, the point $\overline{x}$ belongs to the interior of $C_i := g_i^{-1}((-\infty, 0])$, so that for any $x \in C := g^{-1}((-\infty, 0])$ and any $t > 0$ small enough we have $\overline{x} + t(x - \overline{x}) \in D$. It follows that $N(D,\overline{x}) = N(C,\overline{x})$. By Proposition 4 there exists some $\overline{x}^* \in \partial^{<} f(\overline{x})$ such that $-\overline{x}^* \in N(D,\overline{x}) = N(C,\overline{x})$. Now h is u.s.c. at $\overline{x}$ and is a Gutiérrez function at $\overline{x}$ by Lemma 9. Then, by relation (8), there exist $y_i \in \mathbb{R}_+$, $\overline{y}_i^* \in \partial^{\leq} g_i(\overline{x})$ such that $-\overline{x}^* = y_1\overline{y}_1^* + \cdots + y_n\overline{y}_n^*$ and the result is proven. □

Proposition 1 shows that the preceding statement encompasses the classical result for convex mathematical programming. The next example illustrates the theorem; note that since the function f is not semistrictly quasiconvex [17, Prop. 6.1] and [19, Prop. 6.3] cannot be applied.

Example 3. Let f be as in Example 2 with $X := \mathbb{R}$ and let g_1 be given by $g_1(x) := -x$. Then f and g_1 are Gutiérrez and Plastria functions at $\overline{x} = 0$ and $\overline{x}$

is not a local minimizer of f. Moreover, the Slater condition is satisfied at $\overline{x}$. We can take $y_1 = 1$ as a multiplier since $\partial^< f(\overline{x}) = [1, +\infty)$ and $\partial^{\leq} g_1(\overline{x}) = (-\infty, -1]$, $g_1(\overline{x}) = 0$.

A link with the classical Karush, Kuhn, and Tucker theorem is delineated in the next statement.

COROLLARY 11. *Assume the hypothesis of the preceding proposition are satisfied and that $f, g_1, \ldots, g_n$ are differentiable at $\overline{x}$ with nonzero derivatives. Then there exist some $\lambda_1, \ldots, \lambda_n \in \mathbb{R}_+$ such that*

$$\begin{aligned} f'(\overline{x}) + \lambda_1 g_1'(\overline{x}) + \cdots + \lambda_n g_n'(\overline{x}) &= 0, \\ \lambda_i g_i(\overline{x}) &= 0 \qquad \text{for } i = 1, \ldots, n. \end{aligned}$$

Proof. By Proposition 3 and the preceding result, there exist some $r \geq 1, y_i \in \mathbb{R}_+$ and some $\overline{y}_i^* \in \partial^{\leq} g_i(\overline{x})$ for $i = 1, \ldots, n$ such that

$$r f'(\overline{x}) + y_1 \overline{y}_1^* + \cdots + y_n \overline{y}_n^* = 0;$$

also $\overline{y}_i^* = s_i g_i'(\overline{x})$ for some $s_i \geq 1$. Setting $\lambda_i = r^{-1} s_i y_i$, we get the result. □

Let us give a simple sufficient condition for the mathematical programming problem $(\mathcal{M})$.

THEOREM 12. *If $\overline{x} \in C$ is such that there exist $y_i \in \mathbb{R}_+$ for $i = 1, \ldots, n$ such that the following conditions are satisfied, then $\overline{x}$ is a solution to problem $(\mathcal{M})$:*

$$\begin{aligned} 0 &\in \partial^< f(\overline{x}) + y_1 \partial^{\leq} g_1(\overline{x}) + \cdots + y_n \partial^{\leq} g_n(\overline{x}), \\ g_1(\overline{x}) &\leq 0, \ldots, g_n(\overline{x}) \leq 0, \\ y_1 g_1(\overline{x}) &= 0, \ldots, y_n g_n(\overline{x}) = 0. \end{aligned}$$

Proof. Suppose to the contrary that there exists some $x \in C$ such that $f(x) < f(\overline{x})$. Let $I(\overline{x}) := \{i \in \{1, \ldots, n\} : g_i(\overline{x}) = 0\}$. Let $\overline{x}^* \in \partial^< f(\overline{x})$, $\overline{x}_i^* \in \partial^{\leq} g_i(\overline{x})$ for $i = 1, \ldots, n$ be such that

$$0 = \overline{x}^* + y_1 \overline{x}_1^* + \cdots + y_n \overline{x}_n^* = \overline{x}^* + \sum_{i \in I(\overline{x})} y_i \overline{x}_i^*.$$

Since $f(x) < f(\overline{x})$, $g_i(x) \leq 0 = g_i(\overline{x})$ for $i \in I(\overline{x})$, by the definitions of $\partial^< f(\overline{x})$, $\partial^{\leq} g_i(\overline{x})$ we have

$$\begin{aligned} \langle \overline{x}^*, x - \overline{x} \rangle &\leq f(x) - f(\overline{x}), \\ \langle \overline{x}_i^*, x - \overline{x} \rangle &\leq g_i(x) - g_i(\overline{x}), \qquad i \in I(\overline{x}). \end{aligned}$$

Multiplying each side of the last inequality by y_i and adding the obtained sides to the ones of the preceding relation, we get

$$\begin{aligned} 0 &= \langle \overline{x}^*, x - \overline{x} \rangle + \sum_{i \in I(\overline{x})} y_i \langle \overline{x}_i^*, x - \overline{x} \rangle \\ &\leq f(x) - f(\overline{x}) + \sum_{i \in I(\overline{x})} y_i \left(g_i(x) - g_i(\overline{x}) \right) \leq f(x) - f(\overline{x}), \end{aligned}$$

a contradiction. □

Example 4. Let X, $\overline{x}$, f, and g_1 be as in Example 3. Then, since $y_1 := 1$ is a multiplier we get that $\overline{x}$ is a solution to problem $(\mathcal{M})$.

Acknowledgments. The authors are grateful to the editor and two anonymous referees for detailed remarks which helped in improving the presentation of the paper. The stay of the first author at the University of Pau thanks to the support of the Région Aquitaine is gratefully acknowledged.

REFERENCES

[1] K. J. Arrow and A. C. Enthoven, *Quasi-concave programming*, Econometrica, 29 (1961), pp. 779–800.
[2] H. Attouch and H. Brézis, *Duality for the sum of convex functions in general Banach spaces*, in Aspects of Mathematics and Its Applications, J. A. Barroso, ed., North-Holland, Amsterdam, 1986, pp. 125–133.
[3] J. M. Borwein and Q. J. Zhu, *A survey of subdifferential calculus with applications*, Nonlinear Anal., 38 (1999), pp. 687–773.
[4] A. Daniilidis, N. Hadjisavvas, and J.-E. Martínez-Legaz, *An appropriate subdifferential for quasiconvex functions*, SIAM J. Optim., 12 (2001), pp. 407–420.
[5] J. Dutta, V. Vetrivel, and S. Nanda, *Semi-invex functions and their subdifferentials*, Bull. Austral. Math. Soc., 56 (1997), pp. 385–393.
[6] G. Giorgi and A. Guerraggio, *First order generalized optimality conditions for programming problems with a set constraint*, in Generalized Convexity. Proceedings of the 5th International Workshop on Generalized Convexity, Pécs, Hungary, 1992, S. Komlósi et al., eds., Lecture Notes Econ. Math. Syst. 405, Springer-Verlag, Berlin, 1994, pp. 171–185.
[7] G. Giorgi and A. Guerraggio, *Various types of nonsmooth invex functions*, J. Inform. Optim. Sci., 17 (1996), pp. 137–150.
[8] B. M. Glover and V. Jeyakumar, *Abstract nonsmooth nonconvex programming*, in Proceedings of the 5th International Workshop on Generalized Convexity Held at Janus Pannonius University, Pécs, Hungary, 1992, S. Komlósi et al., eds., Lecture Notes Econ. Math. Syst. 405, Springer-Verlag, Berlin, 1994, pp. 186–210.
[9] H. P. Greenberg and W. P. Pierskalla, *Quasiconjugate function and surrogate duality*, Cahiers du Centre d'Etude de Recherche Opérationnelle, 15 (1973), pp. 437–448.
[10] J. M. Gutiérrez Díez, *Infragradients and directions of decrease*, Rev. Real Acad. Cienc. Exact. Fís. Natur. Madrid, 78 (1984), pp. 523–532.
[11] J. Gwinner and V. Jeyakumar, *A solvability theorem and minimax fractional programming*, Z. Oper. Res., 37 (1993), pp. 1–12.
[12] V. Jeyakumar and B. M. Glover, *A new version of Farkas' lemma and global convex maximization*, Appl. Math. Lett., 6 (1993), pp. 39–43.
[13] V. Jeyakumar, W. Oettli, and M. Natividad, *A solvability theorem for a class of quasiconvex mappings with applications to optimization*, J. Math. Anal. Appl., 179 (1993), pp. 537–546.
[14] Z. A. Khan, *On nondifferentiable quasiconvex programming problem*, J. Inform. Optim. Sci., 12 (1991), pp. 57–64.
[15] O. L. Mangasarian, *Nonlinear Programming*, McGraw-Hill, New York, 1969.
[16] J. E. Martínez-Legaz, *Quasiconvex duality theory by generalized conjugation methods*, Optimization, 19 (1988), pp. 603–652.
[17] J.-E. Martínez-Legaz, *On lower subdifferentiable functions*, in Trends in Mathematical Optimization, K. H. Hoffmann et al., eds., Internat. Series Numer. Math. 84, Birkhauser, Basel, 1988, pp. 197–232.
[18] J.-E. Martínez-Legaz, *Weak lower subdifferentials and applications*, Optimization, 21 (1990), pp. 321–341.
[19] J.-E. Martínez-Legaz and S. Romano-Rodríguez, *α-lower subdifferentiable functions*, SIAM J. Optim., 3 (1993), pp. 800–825.
[20] J.-E. Martínez-Legaz and S. Romano-Rodríguez, *Lower subdifferentiability of quadratic functions*, Math. Programming, 60 (1993), pp. 93–113.
[21] J. E. Martínez-Legaz and P. H. Sach, *A new subdifferential in quasiconvex analysis*, J. Convex Anal., 6 (1999), pp. 1–11.
[22] H. Van Ngai and M. Théra, *Error bounds and implicit multifunction theorem in smooth Banach spaces and applications to optimization*, Set-Valued Anal., 12 (2004), pp. 195–223.
[23] J.-P. Penot, *Are generalized derivatives useful for generalized convex functions?*, in Generalized Convexity, Generalized Monotonicity: Recent Results, J.-P. Crouzeix, J. E. Martinez-Legaz, and M. Volle, eds., Kluwer, Dordrecht, The Netherlands, 1998, pp. 3–59.

[24] J.-P. Penot, *What is quasiconvex analysis?*, Optimization, 47 (2000), pp. 35–110.
[25] J.-P. Penot, *A variational subdifferential for quasiconvex functions*, J. Optim. Theory Appl., 111 (2001), pp. 165–171.
[26] J.-P. Penot, *Characterization of solution sets of quasiconvex programs*, J. Optim. Theory Appl., 117 (2003), pp. 627–636.
[27] J.-P. Penot, *A Lagrangian approach to quasiconvex analysis*, J. Optim. Theory Appl., 117 (2003), pp. 637–647.
[28] J.-P. Penot and P. H. Sach, *Generalized monotonicity of subdifferentials and generalized convexity*, J. Optim. Theory Appl., 94 (1997), pp. 251–262.
[29] J.-P. Penot and M. Volle, *Another duality scheme for quasiconvex problems*, in Trends in Mathematical Optimization, K. H. Hoffmann et al., eds., Internat. Series Numer. Math. 84, Birkhäuser, Basel, Switzerland, 1988, pp. 259–275.
[30] J.-P. Penot and M. Volle, *On quasi-convex duality*, Math. Oper. Res., 15 (1990), pp. 597–625.
[31] J.-P. Penot and M. Volle, *Surrogate programming and multipliers in quasiconvex programming*, SIAM J. Control Optim., 42 (2003), pp. 1994–2003.
[32] J.-P. Penot and C. Zalinescu, *Harmonic sums and duality*, J. Convex Anal., 7 (2000), pp. 95–113.
[33] J.-P. Penot and C. Zalinescu, *Elements of quasiconvex subdifferential calculus*, J. Convex Anal., 7 (2000), pp. 243–269.
[34] R. Pini, *Invexity and generalized convexity*, Optimization, 22 (1991), pp. 513–525.
[35] F. Plastria, *Lower subdifferentiable functions and their minimization by cutting planes*, J. Optim. Theory Appl., 46 (1985), pp. 37–53.
[36] P. H. Sach, D. S. Kim, and G. Lee, *Invexity as Necessary Optimality Condition in Nonsmooth Programs*, preprint 2000/30, Hanoi Institute of Mathematics, Hanoi, Vietnam.

SIAM J. OPTIM.
Vol. 17, No. 2, pp. 511–525
© 2006 Society for Industrial and Applied Mathematics

STABILITY OF MULTISTAGE STOCHASTIC PROGRAMS*

H. HEITSCH†, W. RÖMISCH†, AND C. STRUGAREK‡

Abstract. Quantitative stability of linear multistage stochastic programs is studied. It is shown that the infima of such programs behave (locally) Lipschitz continuous with respect to the sum of an L_r-distance and of a distance measure for the filtrations of the original and approximate stochastic (input) processes. Various issues of the result are discussed and an illustrative example is given. Consequences for the reduction of scenario trees are also discussed.

Key words. stochastic programming, multistage, nonanticipativity, stability, filtration, probability metrics

AMS subject classification. 90C15

DOI. 10.1137/050632865

1. Introduction. We consider a finite horizon sequential decision process under uncertainty, in which a decision made at t is based only on information available at t ($1 \le t \le T$). We assume that the information is given by a discrete time multivariate stochastic process $\{\xi_t\}_{t=1}^T$ defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and with ξ_t taking values in $\mathbb{R}^d$. The information available at t consists of the random vector $\xi^t := (\xi_1, \dots, \xi_t)$, and the stochastic decision x_t at t varying in $\mathbb{R}^{m_t}$ is assumed to depend only on ξ^t. The latter property is called *nonanticipativity* and is equivalent to the measurability of x_t with respect to the σ-field $\mathcal{F}_t \subseteq \mathcal{F}$, which is generated by ξ^t. Hence, we have $\mathcal{F}_t \subseteq \mathcal{F}_{t+1}$ for $t = 1, \dots, T-1$ and we assume that $\mathcal{F}_1 = \{\emptyset, \Omega\}$, i.e., ξ_1 and x_1 are deterministic and, with no loss of generality, that $\mathcal{F}_T = \mathcal{F}$.
More precisely, we consider the following *linear multistage stochastic program*:

$$\min \left\{ \mathbb{E}\left[\sum_{t=1}^{T} \langle b_t(\xi_t), x_t \rangle\right] \;\middle|\; \begin{array}{l} x_t \in X_t, \\ x_t \text{ is } \mathcal{F}_t\text{-measurable}, t = 1, \dots, T, \\ A_{t,0}x_t + A_{t,1}(\xi_t)x_{t-1} = h_t(\xi_t), t = 2, \dots, T \end{array} \right\}, \tag{1.1}$$

where the subsets X_t of $\mathbb{R}^{m_t}$ are nonempty, closed, and polyhedral; the cost coefficients $b_t(\xi_t)$ belong to $\mathbb{R}^{m_t}$; the right-hand sides $h_t(\xi_t)$ are in $\mathbb{R}^{n_t}$; $A_{t,0}$ are fixed (n_t, m_t)-matrices; and $A_{t,1}(\xi_t)$ are (n_t, m_{t-1})-matrices, respectively. We assume that $b_t(\cdot)$, $h_t(\cdot)$, and $A_{t,1}(\cdot)$ depend affinely linearly on ξ_t covering the situation that some of the components of b_t and h_t, and of the elements of $A_{t,1}$ are random.

The challenge of multistage models consists in the presence of two groups of entirely different constraints, namely of measurability and of pointwise constraints for the decisions x_t. This fact does not lead to consequences in the two-stage situation ($T = 2$). In general, however, it is the origin of both the theoretical and computa-

*Received by the editors June 1, 2005; accepted for publication (in revised form) February 9, 2006; published electronically August 16, 2006. This work was supported by the DFG Research Center MATHEON (Mathematics for key technologies) in Berlin and by a grant of EDF (Electricité de France).
http://www.siam.org/journals/siopt/17-2/63286.html
†Institute of Mathematics, Humboldt-University Berlin, D-10099 Berlin, Germany (heitsch@math.hu-berlin.de, romisch@math.hu-berlin.de).
‡EdF R&D, OSIRIS, 1 Avenue du Général de Gaulle F-92141 Clamart Cedex, France (cyrille.strugarek@edf.fr), Ecole Nationale des Ponts et Chaussées, Paris, France, and Ecole Nationale Supérieure de Techniques Avancées, Paris, France.

tional challenges of multistage models. In the present paper, it produces the essential difference of quantitative stability estimates compared to the two-stage case.

When solving multistage models computationally, the first step consists of approximating the stochastic process $\xi = \{\xi_t\}_{t=1}^T$ by a process having finitely many scenarios that exhibit tree structure and have its root at the fixed element ξ_1 of $\mathbb{R}^d$ (see the survey [4] for further information). In this way, both the random vectors ξ^t and the σ-fields $\mathcal{F}_t$ are approximated at each t. This process finally leads to linear programming models that are very large scale in most cases and may be solved by decomposition methods that exploit specific structures of the model (see [31] for additional background). In order to reduce the model dimension, it might be desirable to reduce the originally designed tree. The approaches to scenario reduction in [5, 11] and to scenario tree generation in [21, 14, 10] make use of probability metrics, i.e., of metric distances on spaces of probability measures, where the metrics are selected such that the optimal values of original and approximate stochastic program are close if the distance of the original probability distribution $P = \mathcal{L}(\xi)$ of ξ and its approximation Q is small.

Such quantitative stability results are well developed for two-stage models (cf. the survey [28]). It turned out that distances of probability measures are relevant which are given by certain Monge–Kantorovich mass transportation problems. Such problems are of the form

$$\inf\Big\{\int_{\Xi\times\Xi} c(\xi,\tilde{\xi})\eta(d\xi,d\tilde{\xi}) : \eta\in\mathcal{P}(\Xi\times\Xi),\ \pi_1\eta = P,\ \pi_2\eta = Q\Big\}, \tag{1.2}$$

where Ξ is a closed subset of some Euclidean space, π_1 and π_2 denote the projections onto the first and second components, respectively, c is a nonnegative, symmetric, and continuous cost function and P and Q belong to a set $\mathcal{P}_c(\Xi)$ of probability measures on Ξ, where all integrals are finite. Two types of cost functions have been used in stability analysis of stochastic programs [5, 29], namely,

$$c(\xi,\tilde{\xi}) := \|\xi-\tilde{\xi}\|^r \quad (\xi,\tilde{\xi}\in\Xi) \tag{1.3}$$

and

$$c(\xi,\tilde{\xi}) := \max\{1, \|\xi-\xi_0\|^{r-1}, \|\tilde{\xi}-\xi_0\|^{r-1}\}\|\xi-\tilde{\xi}\| \quad (\xi,\tilde{\xi}\in\Xi) \tag{1.4}$$

for some $r\geq 1$ and $\xi_0\in\Xi$. In both cases, the set $\mathcal{P}_c(\Xi)$ may be chosen as the set $\mathcal{P}_r(\Xi)$ of all probability measures on Ξ having absolute moments of order r. The cost (1.3) leads to *L_r-minimal metrics* ℓ_r [25], which are defined by

$$\ell_r(P,Q) := \inf\left\{\int_{\Xi\times\Xi}\|\xi-\tilde{\xi}\|^r\eta(d\xi,d\tilde{\xi})\,|\eta\in\mathcal{P}(\Xi\times\Xi),\ \pi_1\eta = P,\ \pi_2\eta = Q\right\}^{\frac{1}{r}} \tag{1.5}$$

and sometimes also called Wasserstein metrics of order r [9]. The mass transportation problem (1.2) with cost (1.4) defines the Monge–Kantorovich functionals $\hat{\mu}_r$ [22, 24]. A variant of the functional $\hat{\mu}_r$ appears if, in its definition by (1.2), the conditions $\eta\in\mathcal{P}(\Xi\times\Xi), \pi_1\eta = P,\ \pi_2\eta = Q$ are replaced by η being a finite measure on $\Xi\times\Xi$ such that $\pi_1\eta - \pi_2\eta = P - Q$. The corresponding functionals $\mathring{\mu}_r$ are smaller than $\hat{\mu}_r$ and turn out to be metrics on $\mathcal{P}_r(\Xi)$. They are called Fortet–Mourier metrics of order r [8, 22]. The convergence of sequences of probability measures, with respect to both metrics ℓ_r and $\mathring{\mu}_r$, is equivalent to their weak convergence and the convergence

of their rth order absolute moments. For $r = 1$ we have the identity $\mathring{\mu}_1 = \hat{\mu}_1 = \ell_1$ and the corresponding metric is also called Kantorovich distance. Two-stage models are known to behave stable with respect to Fortet–Mourier metrics [23].

Much less is known, however, of the multistage case. The present paper may be regarded as an extension of the quantitative analysis in [7], which considers a less general probabilistic setup and assumes implicitly that the filtrations of the original and approximate stochastic processes coincide. The paper [19] and the recent work [20] provide (qualitative) convergence results of approximations and [16, 32] deal with empirical estimates in multistage models. In the recent paper [34] the role of probability metrics for studying stability of multistage models is questioned critically. An example is given showing that closeness of original and approximate probability distributions in terms of some probability metric is not sufficient for the infima to be close in general. The recent thesis [1] focuses precisely on the question of information in stochastic programs. The conclusions of this work do not address stability, but only discretization of multistage stochastic programs. They illuminate the role which should be played by σ-field distances in order to obtain a consistent discretization of such programs.

The main result of the present paper (Theorem 2.1) provides stability of infima of the multistage model (1.1) with respect to a sum of the L_r-norm and of a distance of the information structures, i.e., the filtrations of σ-fields, of the original and approximate stochastic (input) processes. Hence, it enlightens the corresponding arguments in [34]. Several comments are given on the stability result, its assumptions, the filtration distance, and on the choice of the underlying probability space if the original and approximate (input) probability distributions are given in practical models. Furthermore, we provide an illustrative example which shows that the filtration distance is indispensable for stability (Example 2.6). Finally, some consequences for designing scenario reduction schemes in multistage models are discussed.

2. Stability of multistage models. Under weak hypotheses, the program (1.1) can be equivalently reformulated as a minimization problem for the deterministic first stage decision x_1 (see [31, Chapter 1] or [6, 26] for example). It is of the form

$$\min\Big\{\mathbb{E}[f(x_1,\xi)] = \int_{\Xi} f(x_1,\xi)P(d\xi) : x_1 \in X_1\Big\}, \tag{2.1}$$

where Ξ is a closed subset of $\mathbb{R}^{Td}$ containing the support of the probability distribution P of ξ, and f is an integrand on $\mathbb{R}^{m_1}\times\Xi$ given by the dynamic programming recursion

$$\begin{aligned} f(x_1,\xi) &:= \Phi_1(x_1,\xi^1) = \langle b_1(\xi_1), x_1\rangle + \Phi_2(x_1,\xi^2), \qquad (2.2)\\ \Phi_t(x_1,\dots,x_{t-1},\xi^t) &:= \inf\Big\{\langle b_t(\xi_t), x_t\rangle + \mathbb{E}\left[\Phi_{t+1}(x_1,\dots,x_t,\xi^{t+1})|\mathcal{F}_t\right] : x_t \in X_t,\\ &\qquad\qquad x_t \text{ is } \mathcal{F}_t\text{-measurable}, A_{t,0}x_t + A_{t,1}(\xi_t)x_{t-1} = h_t(\xi_t)\Big\}\\ &\qquad (t = 2,\dots,T),\\ \Phi_{T+1}(x_1,\dots,x_T,\xi^{T+1}) &:= 0. \end{aligned}$$

Using the representation (2.2) of the integrand f for $T = 2$ quantitative stability results are proved in [23, 28] with respect to Fortet–Mourier metrics of probability distributions and earlier in [29] with respect to L_r-minimal metrics. For $T > 2$, however, the integrand f depends on conditional expectations with respect to the σ-fields $\mathcal{F}_t$ and, hence, on the underlying probability measure $\mathbb{P}$ in a nonlinear way. Consequently, the methodology for studying quantitative stability properties of stochastic

programs of the form (2.1) developed in [23, 28] does not apply to multistage models in general.

An alternative for studying stability of multistage models consists in considering them as optimization problems in functional spaces (see also [18, 26]), where the Banach spaces $L_{r'}(\Omega, \mathcal{F}, \mathbb{P}; \mathbb{R}^m)$ with $m = \sum_{t=1}^T m_t$ and endowed with the norm

$$\|x\|_{r'} := \left(\sum_{t=1}^T \mathbb{E}[\|x_t\|^{r'}] \right)^{\frac{1}{r'}} \quad \text{for } r' \geq 1 \text{ and } \quad \|x\|_\infty := \max_{t=1,\dots,T} \operatorname{ess\,sup} \|x_t\|$$

are appropriate. Here, the stochastic input process ξ belongs to $L_r(\Omega, \mathcal{F}, \mathbb{P}; \mathbb{R}^s)$ for some $r \geq 1$ and $s := Td$, and r' is defined by

$$r' := \begin{cases} \frac{r}{r-1} & \text{if only costs are random,} \\ r & \text{if only right-hand sides are random,} \\ r = 2 & \text{if only costs and right-hand sides are random,} \\ \infty & \text{if all technology matrices are random and } r = T. \end{cases}$$

The number r corresponds to the order of (absolute) moments of ξ that are required to exist. The definition of the numbers r' implies that the objective function is well defined and finite. In the third case it may alternatively be required that the costs $b_t(\xi_t)$ have finite moments of order $\hat{r} \geq 1$. Then we choose $r' := \frac{\hat{r}}{\hat{r}-1}$ and require that $h_t(\xi_t)$ belongs to $L_{r'}$.

Let us introduce some notations. Let F denote the objective function defined on $L_r(\Omega, \mathcal{F}, \mathbb{P}; \mathbb{R}^s) \times L_{r'}(\Omega, \mathcal{F}, \mathbb{P}; \mathbb{R}^m) \to \mathbb{R}$ by $F(\xi, x) := \mathbb{E}[\sum_{t=1}^T \langle b_t(\xi_t), x_t \rangle]$, let

$$\mathcal{X}_t(x_{t-1}; \xi_t) := \{x_t \in X_t | A_{t,0} x_t + A_{t,1}(\xi_t) x_{t-1} = h_t(\xi_t)\}$$

denote the tth feasibility set for every $t = 2, \dots, T$, and

$$\mathcal{X}(\xi) := \{x \in \times_{t=1}^T L_{r'}(\Omega, \mathcal{F}_t, \mathbb{P}; \mathbb{R}^{m_t}) | x_1 \in X_1, x_t \in \mathcal{X}_t(x_{t-1}; \xi_t), t = 2, \dots, T\}$$

denote the set of feasible elements of the stochastic program (1.1) with input ξ. Then the stochastic program (1.1) may be rewritten in the form

$$\min\{F(\xi, x) : x \in \mathcal{X}(\xi)\}. \tag{2.3}$$

Let $v(\xi)$ denote the optimal value of (2.3) and let, for any $\alpha \geq 0$,

$$l_\alpha(F(\xi, \cdot)) := \{x \in \mathcal{X}(\xi) : F(\xi, x) \leq v(\xi) + \alpha\}$$

denote its *α-level set*. The following conditions are imposed on (2.3).

(A1) There exists a $\delta > 0$ such that for any $\tilde{\xi} \in L_r(\Omega, \mathcal{F}, \mathbb{P}; \mathbb{R}^s)$ with $\|\tilde{\xi} - \xi\|_r \leq \delta$, any $t = 2, \dots, T$ and any $x_1 \in X_1$, $x_\tau \in L_{r'}(\Omega, \mathcal{F}_t, \mathbb{P}; \mathbb{R}^{m_\tau})$ with $x_\tau \in \mathcal{X}_\tau(x_{\tau-1}; \tilde{\xi}_\tau)$, $\tau = 2, \dots, t-1$, the tth feasibility set $\mathcal{X}_t(x_{t-1}; \tilde{\xi}_t)$ is nonempty (*relatively complete recourse locally around ξ*).

(A2) The optimal values $v(\tilde{\xi})$ of (2.3) with input $\tilde{\xi}$ are finite for all $\tilde{\xi}$ in a neighborhood of ξ and the objective function F is *level-bounded locally uniformly at ξ*, i.e., for some $\alpha > 0$ there exists a $\delta > 0$ and a bounded subset B of $L_{r'}(\Omega, \mathcal{F}, \mathbb{P}; \mathbb{R}^m)$ such that $l_\alpha(F(\tilde{\xi}, \cdot))$ is nonempty and contained in B for all $\tilde{\xi} \in L_r(\Omega, \mathcal{F}, \mathbb{P}; \mathbb{R}^s)$ with $\|\tilde{\xi} - \xi\|_r \leq \delta$.

(A3) $\xi \in L_r(\Omega, \mathcal{F}, \mathbb{P}; \mathbb{R}^s)$ for some $r \geq 1$.

To state our main result we introduce the distance $D_{\mathrm{f}}(\xi, \tilde{\xi})$ of the filtrations of ξ and its approximation (or perturbation) $\tilde{\xi}$, respectively. It is defined by

$$D_{\mathrm{f}}(\xi, \tilde{\xi}) := \sup_{\varepsilon \in (0,\alpha]} D_{\mathrm{f},\varepsilon}(\xi, \tilde{\xi}) \tag{2.4}$$

and $D_{\mathrm{f},\varepsilon}(\xi, \tilde{\xi})$ denotes the ε-filtration distance given by

$$D_{\mathrm{f},\varepsilon}(\xi, \tilde{\xi}) := \inf \sum_{t=2}^{T-1} \max\{\|x_t - \mathbb{E}[x_t|\tilde{\mathcal{F}}_t]\|_{r'}, \|\tilde{x}_t - \mathbb{E}[\tilde{x}_t|\mathcal{F}_t]\|_{r'}\}, \tag{2.5}$$

where the infimum is taken with respect to all $x \in l_\varepsilon(F(\xi, \cdot))$ and $\tilde{x} \in l_\varepsilon(F(\tilde{\xi}, \cdot))$, respectively, i.e., with respect to all feasible decisions belonging to the ε-level sets of the original and perturbed programs. Furthermore, $\mathcal{F}_t$ and $\tilde{\mathcal{F}}_t$, $t = 1, \ldots, T$, denote the filtrations of ξ and $\tilde{\xi}$, respectively.

Now, we are ready to state our main stability result for multistage stochastic programs.

THEOREM 2.1. *Let* (A1), (A2), *and* (A3) *be satisfied and* X_1 *be bounded. Then there exists positive constants* L, α, *and* δ *such that the estimate*

$$|v(\xi) - v(\tilde{\xi})| \le L(\|\xi - \tilde{\xi}\|_r + D_{\mathrm{f}}(\xi, \tilde{\xi})) \tag{2.6}$$

holds for all random elements $\tilde{\xi} \in L_r(\Omega, \mathcal{F}, \mathbb{P}; \mathbb{R}^s)$ *with* $\|\tilde{\xi} - \xi\|_r \le \delta$.

Proof. Let M_t denote the set-valued mappings $u \mapsto \{x \in \mathbb{R}^{m_t} | A_{t,0}x = u,\ x \in X_t\}$ from $\mathbb{R}^{n_t}$ to $\mathbb{R}^{m_t}$ for $t = 2, \ldots, T$. The mappings have polyhedral graph and, hence, are Lipschitz continuous with respect to the Hausdorff distance on their domain $\operatorname{dom} M_t \subseteq \mathbb{R}^{n_t}$ [27, Example 9.35]. Hence, there exist positive constants l_t such that we have

$$\sup_{x \in M_t(\bar{u})} d(x, M_t(\tilde{u})) \le l_t \|\bar{u} - \tilde{u}\| \tag{2.7}$$

for all $\bar{u}, \tilde{u} \in \operatorname{dom} M_t$, where $d(x, C)$ denotes the distance of x to a nonempty set C in $\mathbb{R}^{m_t}$.

Now, let $\alpha > 0$ and $\delta > 0$ be selected as in (A1) and (A2). Let $\varepsilon \in (0, \alpha]$, $\tilde{\xi} \in L_r(\Omega, \mathcal{F}, \mathbb{P}; \mathbb{R}^s)$ be such that $\|\tilde{\xi} - \xi\|_r < \delta$ and $v(\tilde{\xi}) \in \mathbb{R}$, and let $\bar{x} \in l_\varepsilon(F(\xi, \cdot))$. By $\tilde{\mathcal{F}}_t$ we denote the σ-field generated by $\tilde{\xi}^t := (\tilde{\xi}_1, \ldots, \tilde{\xi}_t)$ for $t = 1, \ldots, T$. Now, we show recursively the existence of constants $\hat{L}_t > 0$ and of elements $\tilde{x}_t$ belonging to the appropriate spaces $L_{r'}(\Omega, \tilde{\mathcal{F}}_t, \mathbb{P}; \mathbb{R}^{m_t})$ for each $t = 1, \ldots, T$ such that $\tilde{x}_t \in X_t$, $t = 1, \ldots, T$, $A_{t,0}\tilde{x}_t + A_{t,1}(\tilde{\xi}_t)\tilde{x}_{t-1} = h_t(\tilde{\xi}_t)$, $t = 2, \ldots, T$, and that

$$\|\mathbb{E}[\bar{x}_t|\tilde{\mathcal{F}}_t] - \tilde{x}_t\|$$

can be estimated recursively with respect to t. Let $t = 1$, we then set $\tilde{x}_1 := \bar{x}_1$ and $\hat{L}_1 := 1$. For $t > 1$, we assume that $\hat{L}_{t-1}$ and $\tilde{x}_{t-1}$ have already been constructed, set $\bar{u}_t := h_t(\xi_t) - A_{t,1}(\xi_t)\bar{x}_{t-1}$, $\tilde{u}_t := h_t(\tilde{\xi}_t) - A_{t,1}(\tilde{\xi}_t)\tilde{x}_{t-1}$ and consider the following set-valued mappings from Ω to $\mathbb{R}^{m_t}$ given by

$$\omega \to M_t(\tilde{u}_t(\omega)) \qquad \text{and} \qquad \omega \to \arg\min_{x \in M_t(\tilde{u}_t(\omega))} \|\mathbb{E}[\bar{x}_t|\tilde{\mathcal{F}}_t](\omega) - x\|.$$

Both are measurable with respect to the σ-field $\tilde{\mathcal{F}}_t$ due to the measurability of $\tilde{x}_{t-1}$ with respect to $\tilde{\mathcal{F}}_{t-1}$ and well-known measurability results for set-valued mappings

(e.g., [27, Theorem 14.36]). In addition, the set-valued mapping $\omega \to M_t(\tilde{u}_t(\omega))$ is nonempty-valued due to (A1). Hence, by appealing to [27, Theorem 14.37] there exists a $\tilde{\mathcal{F}}_t$-measurable selection $\tilde{x}_t$ of the second mapping. Since $\mathbb{E}[\bar{x}_t|\tilde{\mathcal{F}}_t]$ belongs to $M_t(\mathbb{E}[\bar{u}_t|\tilde{\mathcal{F}}_t])$, (2.7) provides the estimate

$$\begin{aligned}
\|\mathbb{E}[\bar{x}_t|\tilde{\mathcal{F}}_t] - \tilde{x}_t\| &\le l_t\|\mathbb{E}[\bar{u}_t|\tilde{\mathcal{F}}_t] - \tilde{u}_t\| \\
&\le l_t(\|\mathbb{E}[h_t(\xi_t)|\tilde{\mathcal{F}}_t] - h_t(\tilde{\xi}_t)\| + \|\mathbb{E}[A_{t,1}(\xi_t)\bar{x}_{t-1}|\tilde{\mathcal{F}}_t] - A_{t,1}(\tilde{\xi}_t)\tilde{x}_{t-1}\|) \\
&\le l_t(K_t\|\mathbb{E}[\xi_t|\tilde{\mathcal{F}}_t] - \tilde{\xi}_t\| + \|\mathbb{E}[A_{t,1}(\xi_t)\bar{x}_{t-1} - A_{t,1}(\tilde{\xi}_t)\bar{x}_{t-1}|\tilde{\mathcal{F}}_t]\| \\
&\qquad + \|A_{t,1}(\tilde{\xi}_t)\|\|\mathbb{E}[\bar{x}_{t-1}|\tilde{\mathcal{F}}_t] - \tilde{x}_{t-1}\|) \\
&\le l_t\bar{K}_t(\|\mathbb{E}[\xi_t - \tilde{\xi}_t|\tilde{\mathcal{F}}_t]\| + \mathbb{E}[\|\xi_t - \tilde{\xi}_t\|\|\bar{x}_{t-1}\||\tilde{\mathcal{F}}_t] \\
&\qquad + \max\{1, \|\tilde{\xi}_t\|\}(\|\mathbb{E}[\bar{x}_{t-1} - \mathbb{E}[\bar{x}_{t-1}|\tilde{\mathcal{F}}_{t-1}]|\tilde{\mathcal{F}}_t]\| \\
&\qquad + \|\mathbb{E}[\bar{x}_{t-1}|\tilde{\mathcal{F}}_{t-1}] - \tilde{x}_{t-1}\|)),
\end{aligned}$$

where K_t and $\bar{K}_t$ are certain positive constants, the affine linearity of $h_t(\cdot)$ and $A_{t,1}(\cdot)$ and Jensen's inequality is used for the second summand. Clearly, we have $\|\tilde{\xi}_\tau\| \le C\|\tilde{\xi}^t\|$ with some constant C for all $\tau = 2, \dots, t$, $t = 2, \dots, T$, and the corresponding norms in $\mathbb{R}^d$ and $\mathbb{R}^{td}$. Using Jensen's inequality also in the first and third summand of the latter estimate we obtain recursively

$$\begin{aligned}
\|\mathbb{E}[\bar{x}_t|\tilde{\mathcal{F}}_t] - \tilde{x}_t\| \le \hat{L}_t\Big(&\sum_{\tau=2}^{t} \max\{1, \|\tilde{\xi}^t\|^{t-\tau}\}\mathbb{E}[(1 + \|\bar{x}_{\tau-1}\|)\|\xi_\tau - \tilde{\xi}_\tau\|\,|\tilde{\mathcal{F}}_\tau] \\
&+ \sum_{\tau=2}^{t-1} \max\{1, \|\tilde{\xi}^t\|^{t-\tau}\}\mathbb{E}[\|\bar{x}_\tau - \mathbb{E}[\bar{x}_\tau|\tilde{\mathcal{F}}_\tau]\|\,|\tilde{\mathcal{F}}_{\tau+1}]\Big)
\end{aligned} \tag{2.8}$$

with some positive constant $\hat{L}_t$ for $t = 2, \dots, T$. Note that the sum on the right-hand side of (2.8) disappears if only costs are random. The max-terms in (2.8) and the norms $\|x_{\tau-1}\|$ in (2.8) vanish if the technology matrices are not random. Inserting $\bar{x}$ and $\tilde{x}$ into the objective function we obtain

$$v(\tilde{\xi}) - v(\xi) \le F(\tilde{\xi}, \tilde{x}) - F(\xi, \bar{x}) + \varepsilon. \tag{2.9}$$

In case of only right-hand sides being random we continue (2.9) using (2.8) and obtain

$$\begin{aligned}
v(\tilde{\xi}) - v(\xi) &\le \sum_{t=2}^{T} \mathbb{E}[\langle b_t, \mathbb{E}[\tilde{x}_t - \bar{x}_t|\tilde{\mathcal{F}}_t]\rangle] + \varepsilon \le \sum_{t=2}^{T} \|b_t\|\mathbb{E}[\|\tilde{x}_t - \mathbb{E}[\bar{x}_t|\tilde{\mathcal{F}}_t]\|] + \varepsilon \\
&\le \hat{L}\sum_{t=2}^{T} \mathbb{E}\Big[\sum_{\tau=2}^{t} \mathbb{E}[\|\xi_\tau - \tilde{\xi}_\tau\||\tilde{\mathcal{F}}_\tau] + \sum_{\tau=2}^{t-1} \mathbb{E}[\|\bar{x}_\tau - \mathbb{E}[\bar{x}_\tau|\tilde{\mathcal{F}}_\tau]\|\,|\tilde{\mathcal{F}}_{\tau+1}]\Big] + \varepsilon \\
&\le \hat{L}T\mathbb{E}\Big[\sum_{t=2}^{T} \|\xi_t - \tilde{\xi}_t\| + \sum_{\tau=2}^{T-1} \|\bar{x}_\tau - \mathbb{E}[\bar{x}_\tau|\tilde{\mathcal{F}}_\tau]\|\Big] + \varepsilon \\
&\le \hat{L}T\Big(\|\xi - \tilde{\xi}\|_r + \sum_{\tau=2}^{T-1} \|\bar{x}_\tau - \mathbb{E}[\bar{x}_\tau|\tilde{\mathcal{F}}_\tau]\|_r\Big) + \varepsilon,
\end{aligned}$$

where $\hat{L} := \max_{t=1,\dots,T} \hat{L}_t\|b_t\|$. If costs are random, we obtain the estimate

$$v(\tilde{\xi}) - v(\xi) \le F(\tilde{\xi}, \tilde{x}) - F(\tilde{\xi}, \bar{x}) + F(\tilde{\xi}, \bar{x}) - F(\xi, \bar{x}) + \varepsilon$$

$$\leq \mathbb{E}\Big[\sum_{t=2}^{T}\langle b_t(\tilde{\xi}_t), \mathbb{E}[\tilde{x}_t - \bar{x}_t|\tilde{\mathcal{F}}_t]\rangle\Big] + \mathbb{E}\Big[\sum_{t=1}^{T}\langle b_t(\tilde{\xi}_t) - b_t(\xi_t), \bar{x}_t\rangle\Big] + \varepsilon$$

$$\leq \hat{K}\mathbb{E}\Big[\sum_{t=2}^{T}\max\{1, \|\tilde{\xi}_t\|\}\|\tilde{x}_t - \mathbb{E}[\bar{x}_t|\tilde{\mathcal{F}}_t]\| + \sum_{t=1}^{T}\|\tilde{\xi}_t - \xi_t\|\|\bar{x}_t\|\Big] + \varepsilon \tag{2.10}$$

with some positive constant $\hat{K}$. In case of only costs being random, i.e., $r' = \frac{r}{r-1}$, we continue with

$$v(\tilde{\xi}) - v(\xi) \leq \hat{K}\mathbb{E}\Big[\sum_{t=2}^{T}\max\{1, \|\tilde{\xi}_t\|\}\|\tilde{x}_t - \mathbb{E}[\bar{x}_t|\tilde{\mathcal{F}}_t]\|\Big] + \hat{K}\|\tilde{\xi} - \xi\|_r\|\bar{x}\|_{r'} + \varepsilon$$

$$\leq \hat{K}\mathbb{E}\Big[\sum_{t=2}^{T}\max\{1, \|\tilde{\xi}_t\|\}\|\tilde{x}_t - \mathbb{E}[\bar{x}_t|\tilde{\mathcal{F}}_t]\|\Big] + K\|\tilde{\xi} - \xi\|_r + \varepsilon,$$

where Hölder's inequality and the boundedness of $\|\bar{x}\|_{r'}$ according to (A2) were used leading to some constant $K > 0$. Using the estimate (2.8), we conclude that

$$v(\tilde{\xi}) - v(\xi) \leq L\Big(\|\tilde{\xi} - \xi\|_r + \sum_{t=2}^{T-1}\|\bar{x}_t - \mathbb{E}[\bar{x}_t|\tilde{\mathcal{F}}_t]\|_{r'}\Big) + \varepsilon,$$

where Hölder's inequality and the fact that $\tilde{\xi}$ varies in a bounded set in L_r were used leading to some constant $L > 0$ (depending on ξ).

Next, we consider the case $r = r' = 2$. Starting from (2.10) we use the Cauchy–Schwarz inequality and obtain

$$v(\tilde{\xi}) - v(\xi) \leq \hat{K}\Big[\Big(\sum_{t=2}^{T}\mathbb{E}[\max\{1, \|\tilde{\xi}_t\|^2\}]\Big)^{\frac{1}{2}}\Big(\sum_{t=2}^{T}\mathbb{E}[\|\tilde{x}_t - \mathbb{E}[\bar{x}_t|\tilde{\mathcal{F}}_t]\|^2]\Big)^{\frac{1}{2}} + \|\tilde{\xi} - \xi\|_2\|\bar{x}\|_2\Big] + \varepsilon$$

$$\leq \Big(\|\tilde{\xi} - \xi\|_2 + \sum_{t=2}^{T-1}\|\bar{x}_t - \mathbb{E}[\bar{x}_t|\tilde{\mathcal{F}}_t]\|_2\Big) + \varepsilon$$

with some constant $L > 0$ (depending on ξ) due to (2.8), (A2), and the fact that $\tilde{\xi}$ varies in some bounded set in L_2.

Finally, we consider the situation that costs, right-hand sides, and technology matrices are random, i.e., $r = T$ and $r' = \infty$. In this case, the estimate (2.8) attains the form

$$\|\mathbb{E}[\bar{x}_t|\tilde{\mathcal{F}}_t] - \tilde{x}_t\| \leq \hat{L}_t\Big(\sum_{\tau=2}^{t}\max\{1, \|\tilde{\xi}^t\|^{t-\tau}\}\mathbb{E}[\|\xi_\tau - \tilde{\xi}_\tau\|\,|\tilde{\mathcal{F}}_\tau] + \sum_{\tau=2}^{t-1}\max\{1, \|\tilde{\xi}^t\|^{t-\tau}\}\|\bar{x}_\tau - \mathbb{E}[\bar{x}_\tau|\tilde{\mathcal{F}}_\tau]\|_\infty\Big).$$

Now, we start again from (2.10) and use the latter estimate and obtain

$$v(\tilde{\xi}) - v(\xi) \leq \hat{L}\mathbb{E}\Big[\sum_{t=2}^{T}\Big(\sum_{\tau=2}^{t}\max\{1, \|\tilde{\xi}^t\|^{t+1-\tau}\}\mathbb{E}[\|\xi_\tau - \tilde{\xi}_\tau\|\,|\tilde{\mathcal{F}}_\tau]$$

$$
\begin{aligned}
&\quad + \sum_{\tau=2}^{t-1} \max\{1, \|\tilde{\xi}^t\|^{t+1-\tau}\} \|\bar{x}_\tau - \mathbb{E}[\bar{x}_\tau|\tilde{\mathcal{F}}_\tau]\|_\infty \Big) + \sum_{t=1}^{T} \|\tilde{\xi}_t - \xi_t\| \Big] + \varepsilon \\
&\leq \tilde{L}\mathbb{E}\Big[\sum_{t=2}^{T} \max\{1, \|\tilde{\xi}^t\|^{t-1}\} \mathbb{E}[\|\xi_t - \tilde{\xi}_t\| \, |\tilde{\mathcal{F}}_t] \Big] \\
&\quad + \sum_{t=2}^{T-1} \mathbb{E}[\max\{1, \|\tilde{\xi}^t\|^{t-1}\}] \|\bar{x}_t - \mathbb{E}[\bar{x}_t|\tilde{\mathcal{F}}_t]\|_\infty + \|\tilde{\xi} - \xi\|_1 + \varepsilon \\
&\leq \bar{L}\mathbb{E}[\max\{1, \|\tilde{\xi}\|^T\}] \Big(\|\xi - \tilde{\xi}\|_T + \sum_{t=2}^{T-1} \|\bar{x}_t - \mathbb{E}[\bar{x}_t|\tilde{\mathcal{F}}_t]\|_\infty \Big) + \varepsilon,
\end{aligned} \tag{2.11}
$$

where $\hat{L}, \tilde{L}, \bar{L}$ are certain positive constants and Hölder's inequality was used. Since $\tilde{\xi}$ varies in a bounded subset of L_T, there exists a constant $L > 0$ (depending on ξ) such that

$$
v(\tilde{\xi}) - v(\xi) \leq L\Big(\|\xi - \tilde{\xi}\|_r + \sum_{t=2}^{T-1} \|\bar{x}_t - \mathbb{E}[\bar{x}_t|\tilde{\mathcal{F}}_t]\|_{r'} \Big) + \varepsilon, \tag{2.12}
$$

where $r = T$ and $r' = \infty$. Hence, an estimate of the form (2.12) is obtained in all cases. Changing the role of ξ and $\tilde{\xi}$ leads to an estimate of the form

$$
v(\xi) - v(\tilde{\xi}) \leq L\Big(\|\xi - \tilde{\xi}\|_r + \sum_{t=2}^{T-1} \|\tilde{x}_t - \mathbb{E}[\tilde{x}_t|\mathcal{F}_t]\|_{r'} \Big) + \varepsilon. \tag{2.13}
$$

We note that the second summands in the estimates (2.12) and (2.13) are bounded by

$$
\sum_{t=2}^{T-1} \max\{\|\bar{x}_t - \mathbb{E}[\bar{x}_t|\tilde{\mathcal{F}}_t]\|_{r'}, \|\tilde{x}_t - \mathbb{E}[\tilde{x}_t|\mathcal{F}_t]\|_{r'}\}. \tag{2.14}
$$

Since the estimates (2.12) and (2.13) are valid for all $\bar{x} \in l_\varepsilon(F(\xi, \cdot))$ and $\tilde{x} \in l_\varepsilon(F(\tilde{\xi}, \cdot))$, we arrive at the estimate

$$
|v(\xi) - v(\tilde{\xi})| \leq L\Big(\|\xi - \tilde{\xi}\|_r + D_{\mathrm{f},\varepsilon}(\xi, \tilde{\xi}) \Big) + \varepsilon \leq L\Big(\|\xi - \tilde{\xi}\|_r + \sup_{\varepsilon \in (0,\alpha]} D_{\mathrm{f},\varepsilon}(\xi, \tilde{\xi}) \Big) + \varepsilon.
$$

Finally, it remains to take the infimum of the right-hand side with respect to $\varepsilon > 0$ and the proof is complete. □

Remark 2.2. A sufficient condition for (A1) to hold is the *complete fixed recourse* condition on all matrices $A_{t,0}$, i.e., the sets X_t are polyhedral cones and $A_{t,0}X_t = \mathbb{R}^{n_t}$ holds for $t = 2, \ldots, T$. Assumption (A2) on the locally uniform level-boundedness of the objective function F is quite standard in perturbation results for optimization problems (see, e.g., [27, Theorem 1.17]). The finiteness condition for the optimal values is needed because it is not implied by the level-boundedness of F for all relevant pairs (r, r'). In the case that Ω is finite or $1 < r' < \infty$, the existence of solutions of (2.3) (and, thus, the finiteness of $v(\xi)$) is a simple consequence of the compactness or the weak sequential compactness of $l_\alpha(F(\xi, \cdot))$ in the reflexive Banach space $L_{r'}(\Omega, \mathcal{F}, \mathbb{P}; \mathbb{R}^m)$ and of the linearity of the objective. Then the filtration distance is of the form

$$
D_{\mathrm{f}}(\xi, \tilde{\xi}) = \inf\Big\{ \sum_{t=2}^{T-1} D_t(\xi, \tilde{\xi}) : x \in l_0(F(\xi, \cdot)), \tilde{x} \in l_0(F(\tilde{\xi}, \cdot)) \Big\}, \tag{2.15}
$$

where $D_t(\xi, \tilde{\xi})$ is defined by

$$D_t(\xi,\tilde{\xi}) := \max\{\|x_t - \mathbb{E}[x_t|\tilde{\mathcal{F}}_t]\|_{r'}, \|\tilde{x}_t - \mathbb{E}[\tilde{x}_t|\mathcal{F}_t]\|_{r'}\} = \max\{\|x_t - \mathbb{E}[x_t|\tilde{\xi}_1,\dots,\tilde{\xi}_t]\|_{r'}, \|\tilde{x}_t - \mathbb{E}[\tilde{x}_t|\xi_1,\dots,\xi_t]\|_{r'}\}. \tag{2.16}$$

Remark 2.3. In practical situations, the available knowledge on the stochastic input consists in (partial or complete) information on its probability distribution. Which probability space should be selected? A natural answer certainly is: Take a probability space where the L_r-distance $\|\xi-\tilde{\xi}\|_r$ and the $L_{r'}$-distances $\|x_t-\mathbb{E}[x_t|\tilde{\xi}_1,\dots,\tilde{\xi}_t]\|_{r'}$ and $\|\tilde{x}_t - \mathbb{E}[\tilde{x}_t|\xi_1,\dots,\xi_t]\|_{r'}$, $t = 2,\dots,T-1$, are minimal. Let us explain this minimality condition in case of the L_r-distance $\|\xi-\tilde{\xi}\|_r$. Let P and Q in $\mathcal{P}_r(\Xi)$ be the probability distributions of ξ and $\tilde{\xi}$. Then there exists an optimal solution $\eta^* \in \mathcal{P}(\Xi\times\Xi)$ of the mass transportation problem (1.5) [22, Theorem 8.1.1], i.e.,

$$\ell_r^r(P,Q) = \int_{\Xi\times\Xi} \|\xi-\tilde{\xi}\|^r \eta^*(d\xi, d\tilde{\xi}),$$

where $\pi_1\eta^* = P$ and $\pi_2\eta^* = Q$. Furthermore, there exists a probability space $(\Omega',\mathcal{F}',\mathbb{P}')$ and an optimal coupling, i.e., a pair $(\xi'(\cdot),\tilde{\xi}'(\cdot))$ of Ξ-valued random elements defined on it, such that the probability distribution of $(\xi'(\cdot),\tilde{\xi}'(\cdot))$ is just η^* [22, Theorem 2.5.1]. In particular, we have that the distance in $L_r(\Omega',\mathcal{F}',\mathbb{P}';\mathbb{R}^s)$ is just the L_r-minimal distance of the probability distributions, i.e.,

$$\ell_r(P,Q) = \|\xi'(\cdot) - \tilde{\xi}'(\cdot)\|_r.$$

In the same way, the relevant minimal $L_{r'}$-distances $\|x_t - \mathbb{E}[x_t|\tilde{\xi}_1,\dots,\tilde{\xi}_t]\|_{r'}$ and $\|\tilde{x}_t - \mathbb{E}[\tilde{x}_t|\xi_1,\dots,\xi_t]\|_{r'}$ correspond to the $\ell_{r'}$-distance of the probability distributions of $x(t)$ and $\mathbb{E}[x_t|\tilde{\xi}_1,\dots,\tilde{\xi}_t]$, and of $\tilde{x}(t)$ and $\mathbb{E}[\tilde{x}_t|\xi_1,\dots,\xi_t]$, respectively.

Remark 2.4 (stability of first-stage solutions). Using the same technique as for proving [28, Theorem 9], the continuity property of infima in Theorem 2.1 can be supplemented by a quantitative stability property of the solution set $S(\xi)$ of (2.1), i.e., of the set of first stage solutions. Namely, there exists a constant $\hat{L} > 0$ such that

$$\sup_{x\in S(\tilde{\xi})} d(x, S(\xi)) \le \Psi_\xi^{-1}(\hat{L}(\|\xi-\tilde{\xi}\|_r + D_{\mathrm{f}}(\xi,\tilde{\xi}))), \tag{2.17}$$

where $\Psi_\xi(\tau) := \inf\left\{\mathbb{E}[f(x_1,\xi)] - v(\xi) : d(x_1, S(\xi)) \ge \tau,\, x_1 \in X_1\right\}$ with $\Psi_\xi^{-1}(\alpha) := \sup\{\tau \in \mathbb{R}_+ : \Psi_\xi(\tau) \le \alpha\}$ $(\alpha \in \mathbb{R}_+)$ is the growth function of the original problem (2.1) near its solution set $S(\xi)$. The boundedness condition for X_1 in Theorem 2.1 can be relaxed to the assumption that the set $S(\xi)$ is bounded. In the latter case a version of (2.6) is derived that contains localized optimal values. Then the estimate (2.6) is valid whenever its right-hand side is sufficiently small.

Remark 2.5 (convergence of filtrations). This remark aims at precising the link between the filtration distance (2.4) and previous work on *convergence of information*. A distance between σ-fields was introduced in [2]. It metrizes a topology called uniform topology on the set of σ-fields. Due to the work of [30] and [17], this distance reads, for all $\mathcal{B},\mathcal{B}'$ sub-σ-fields of $\mathcal{F}$

$$d_B(\mathcal{B},\mathcal{B}') := \sup_{f\in\Phi} \|\mathbb{E}[f|\mathcal{B}] - \mathbb{E}[f|\mathcal{B}']\|_1, \tag{2.18}$$

with Φ the set of all $\mathcal{F}$-measurable functions f such that for all $\omega \in \Omega$, $\|f(\omega)\| \le 1$. Thanks to [15], a filtration can be said to converge to another one if and only if

each σ-field at each time step converges according to the distance d_B. Hence, a distance between filtrations can be introduced, based on the sum of the distances between σ-fields. The second summand in our stability result can be seen as such a distance between the filtrations generated by the two stochastic processes ξ and $\tilde{\xi}$. This summand is not exactly the same as the sum of distances d_B, but it has the same sense: If the feasible set of the stochastic program is bounded, the filtration distance (2.4) is bounded by a sum of distances d_B. Other distances between filtrations and σ-fields have been introduced (see, e.g., [3]) to fit with stochastic optimization problems. The thesis [1] provides a good survey and a few new results on the application of such information distances.

The following example shows that filtration distances are indispensable for the stability of multistage models.

Example 2.6. We consider a multistage stochastic program that models the optimal purchase over time under cost uncertainty. Its decisions x_t correspond to the amounts to be purchased at each time period. The uncertain prices are ξ_t, $t = 1, \dots, T$, and the objective consists in minimizing the expected costs such that a prescribed amount $a > 0$ is achieved at the end of a given time horizon. The problem is of the form

$$\min \left\{ \mathbb{E}\left[\sum_{t=1}^{T} \xi_t x_t\right] \;\middle|\; \begin{array}{l} (x_t, s_t) \in X_t = \mathbb{R}^2_+, \\ (x_t, s_t) \text{ is } \mathcal{F}_t\text{-measurable}, \\ s_t - s_{t-1} = x_t,\ t = 2, \dots, T, \\ s_1 = 0, \quad s_T = a \end{array} \right\},$$

where the state variable s_t corresponds to the amount at time t and $\mathcal{F}_t := \sigma\{\xi_1, \dots, \xi_t\}$. Let $T := 3$ and P_ε denote the probability distribution of the stochastic price process. P_ε is given by the two scenarios $\xi^1_\varepsilon = (3, 2+\varepsilon, 3)$ ($\varepsilon \in [0, 1)$) and $\xi^2_\varepsilon = (3, 2, 1)$ each endowed with probability $\frac{1}{2}$. Let $Q := P_0$ denote the approximation of P_ε given by the two scenarios $\tilde{\xi}^1 = (3, 2, 3)$ and $\tilde{\xi}^2 = (3, 2, 1)$ with the same probabilities $\frac{1}{2}$. We assume that the scenario trees of the processes ξ_ε and $\tilde{\xi}$ are of the form displayed in Figure 2.1, i.e., the filtrations of σ-fields generated by ξ_ε and $\tilde{\xi}$ do not coincide.

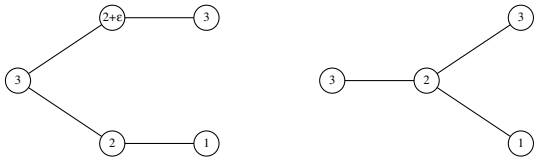


FIG. 2.1. *Scenario trees for P_ε (left) and Q.*

We obtain

$$v(\xi_\varepsilon) = \frac{3+\varepsilon}{2} a \quad \text{and} \quad v(\tilde{\xi}) = 2a\,, \text{ but } \quad \ell_1(P_\varepsilon, Q) = \|\xi_\varepsilon - \tilde{\xi}\|_1 = \frac{\varepsilon}{2}.$$

Hence, the multistage stochastic purchasing model is *not stable* with respect to the L_1-distance $\|\cdot\|_1$. However, the estimate for $|v(\xi) - v(\tilde{\xi})|$ in Theorem 2.1 is valid with $L = 1$ since $D_{\mathrm{f}}(\xi, \tilde{\xi}) = \frac{a}{2}$ holds for the filtration distance (with $r' = \infty$).

Finally, let us consider the case of discrete probability measures P and Q. Let P have scenarios ξ^i with probabilities $p_i > 0$, $i = 1, \dots, N$, and Q scenarios $\tilde{\xi}^j$ and probabilities $q_j > 0$, $j = 1, \dots, M$. Clearly, $\sum_{i=1}^N p_i = 1$ and $\sum_{j=1}^M q_j = 1$. Then $\ell_r^r(P, Q)$ is the optimal value of a finite-dimensional linear transportation problem

(e.g., [24]) and there exist optimal weights $\eta_{ij} \geq 0$ of the scenario pair $(\xi^i, \tilde{\xi}^j)$, $i = 1, \ldots, N$, $j = 1, \ldots, M$. Hence, there exists a pair $(\xi, \tilde{\xi})$ of random vectors on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where $\Omega = \{\omega_{ij} : i = 1, \ldots, N, j = 1, \ldots, M\}$ and $\mathbb{P}(\omega_{ij}) = \eta_{ij}$, $i = 1, \ldots, N$, $j = 1, \ldots, M$. We define $\xi(\omega_{ij}) = \xi^i$ for every $j = 1, \ldots, M$ and $\tilde{\xi}(\omega_{ij}) = \tilde{\xi}^j$ for every $i = 1, \ldots, N$.

Now, our aim is to study the second term in the stability estimate in Theorem 2.1, namely, the distance of filtrations. Let $\mathcal{F}_t$ and $\tilde{\mathcal{F}}_t$ denote the σ-fields generated by $(\xi_1, \ldots, \xi_t)$ and $(\tilde{\xi}_1, \ldots, \tilde{\xi}_t)$, respectively. Let I_t and $\tilde{I}_t$ denote the index set of realizations of ξ_t and $\tilde{\xi}_t$, respectively. Furthermore, let $\mathcal{E}_t$ and $\tilde{\mathcal{E}}_t$ denote families of nonempty elements of $\mathcal{F}_t$ and $\tilde{\mathcal{F}}_t$, respectively, that form partitions of Ω and generate the corresponding σ-fields. We set $E_{ts} := \{\omega \in \Omega : (\xi_1(\omega), \ldots, \xi_t(\omega)) = (\xi_1^s, \ldots, \xi_t^s)\}$, $s \in I_t$, and $\tilde{E}_{ts} := \{\omega \in \Omega : (\tilde{\xi}_1(\omega), \ldots, \tilde{\xi}_t(\omega)) = (\tilde{\xi}_1^s, \ldots, \tilde{\xi}_t^s)\}$, $s \in \tilde{I}_t$.

We set $r = r' = 1$ and require conditions (A1) and (A2) to hold. Since (2.3) is finite-dimensional in this case, optimal solutions x and $\tilde{x}$ exist and we obtain according to Remark 2.2 that

$$
\begin{aligned}
D_t(\xi, \tilde{\xi}) &= \max\Big\{ \sum_{i,j} \eta_{ij} \|x_t(\omega_{ij}) - \mathbb{E}[x_t|\tilde{\mathcal{F}}_t](\omega_{ij})\|, \\
&\qquad \sum_{i,j} \eta_{ij} \|\tilde{x}_t(\omega_{ij}) - \mathbb{E}[\tilde{x}_t|\mathcal{F}_t](\omega_{ij})\| \Big\} \\
&= \max\Big\{ \sum_{s \in \tilde{I}_t} \sum_{\omega_{ij} \in \tilde{E}_{ts}} \eta_{ij} \Big\| x_t(\omega_{ij}) - \frac{\sum_{\omega_{kl} \in \tilde{E}_{ts}} \eta_{kl} x_t(\omega_{kl})}{\sum_{\omega_{kl} \in \tilde{E}_{ts}} \eta_{kl}} \Big\|, \\
&\qquad \sum_{s \in I_t} \sum_{\omega_{ij} \in E_{ts}} \eta_{ij} \Big\| \tilde{x}_t(\omega_{ij}) - \frac{\sum_{\omega_{kl} \in E_{ts}} \eta_{kl} \tilde{x}_t(\omega_{kl})}{\sum_{\omega_{kl} \in E_{ts}} \eta_{kl}} \Big\| \Big\}.
\end{aligned} \tag{2.19}
$$

The latter representation of D_t has potential to be further estimated in specific cases. In particular, it simplifies considerably for the situation of scenario reduction.

Example 2.7 (scenario reduction). Let us consider the case of deleting scenario $l \in \{1, \ldots, N\}$ of ξ according to the methodology in [5, 11] for the distance ℓ_1 and $r = r' = 1$. Then $\tilde{\xi}$ has the scenarios $\xi^1, \ldots, \xi^{l-1}, \xi^{l+1}, \ldots, \xi^N$ and the probabilities of ξ^j are $q_j = p_j$ for every $j \notin \{j(l), l\}$ and $q_{j(l)} = p_{j(l)} + p_l$, where $j(l) \in \arg\min_{j \neq l} \|\xi^j - \xi^l\|$ (see [5, Theorem 2]). This corresponds to $\tilde{\xi}(\omega_{ij}) = \xi^j$ for every $i = 1, \ldots, N$, $j = 1, \ldots, N$, $j \neq l$. We also infer from [5, Theorem 2] that the optimal weights of the transportation problem defining $\ell_1(P, Q)$ are

$$
\eta_{ij} = \begin{cases} p_l, & i = l,\ j = j(l), \\ p_j, & i = j \neq l, \\ 0 & \text{otherwise.} \end{cases}
$$

We set $\hat{\omega}_j := \omega_{jj}$ for every $j = 1, \ldots, N$, $j \neq l$, $\hat{\omega}_l = \omega_{lj(l)}$ and introduce the notation E_{ts_j} and $\tilde{E}_{ts_j}$ for the sets in $\mathcal{E}_t$ and $\tilde{\mathcal{E}}_t$, respectively, that contain $\hat{\omega}_j$. From (2.19) we conclude the following representations of D_t:

$$
D_t(\xi, \tilde{\xi}) = \max\left\{ \sum_{s \in \tilde{I}_t} \sum_{\hat{\omega}_j \in \tilde{E}_{ts}} p_j \Big\| x_t(\hat{\omega}_j) - \frac{\sum_{\hat{\omega}_k \in \tilde{E}_{ts}} p_k x_t(\hat{\omega}_k)}{\sum_{\hat{\omega}_k \in \tilde{E}_{ts}} p_k} \Big\|, \right.
$$

$$\left. \sum_{s\in I_t} \sum_{\hat\omega_j\in E_{ts}} p_j \Big\| \tilde x_t(\hat\omega_j) - \frac{\sum_{\omega_k\in E_{ts}} p_k \tilde x_t(\hat\omega_k)}{\sum_{\hat\omega_k\in E_{ts}} p_k} \Big\| \right\}$$

$$= \max \left\{ \sum_{s\in \tilde I_t} \frac{1}{\sum_{\hat\omega_k\in \tilde E_{ts}} p_k} \sum_{\hat\omega_j\in \tilde E_{ts}} \Big\| \sum_{\hat\omega_k\in \tilde E_{ts}} p_k p_j [x_t(\hat\omega_j) - x_t(\hat\omega_k)] \Big\|, \right.$$

$$\left. \sum_{s\in I_t} \frac{1}{\sum_{\hat\omega_k\in E_{ts}} p_k} \sum_{\hat\omega_j\in E_{ts}} \Big\| \sum_{\hat\omega_k\in E_{ts}} p_k p_j [\tilde x_t(\hat\omega_j) - \tilde x_t(\hat\omega_k)] \Big\| \right\}$$

$$= \max \left\{ \sum_{s\in \tilde I_t} \frac{1}{\sum_{\hat\omega_k\in \tilde E_{ts}} p_k} \sum_{\hat\omega_j\in \tilde E_{ts}} \Big\| \sum_{\hat\omega_k\in \tilde E_{ts}\setminus E_{ts_j}} p_k p_j [x_t(\hat\omega_j) - x_t(\hat\omega_k)] \Big\|, \right.$$

$$\left. \sum_{s\in I_t} \frac{1}{\sum_{\hat\omega_k\in E_{ts}} p_k} \sum_{\hat\omega_j\in E_{ts}} \Big\| \sum_{\hat\omega_k\in E_{ts}\setminus \tilde E_{ts_j}} p_k p_j [\tilde x_t(\hat\omega_j) - \tilde x_t(\hat\omega_k)] \Big\| \right\},$$

where the final equality is a consequence of the corresponding measurability properties of x_t, which imply $x_t(\hat\omega_j) = x_t(\hat\omega_k)$ if $\hat\omega_k \in E_{ts}\cap \tilde E_{ts_j}$ and $\hat\omega_k \in \tilde E_{ts}\cap E_{ts_j}$, respectively. Since $E_{ts_j} = \tilde E_{ts_j}$ for $j \notin \{l, j(l)\}$ and $\tilde E_{ts_l} = E_{tj(l)} \cup \{\hat\omega_l\}$, we may continue with

$$D_t(\xi,\tilde\xi) = \max \left\{ \frac{1}{\sum_{\hat\omega_k\in \tilde E_{ts_l}} p_k} \sum_{\hat\omega_j\in \tilde E_{ts_l}} \Big\| \sum_{\hat\omega_k\in \tilde E_{ts_l}\setminus E_{ts_j}} p_k p_j [x_t(\hat\omega_j) - x_t(\hat\omega_k)] \Big\|, \right.$$

$$\left. \frac{1}{\sum_{\hat\omega_k\in E_{ts_l}} p_k} \sum_{\hat\omega_j\in E_{ts_l}} \Big\| \sum_{\hat\omega_k\in E_{ts_l}\setminus \tilde E_{ts_j}} p_k p_j [\tilde x_t(\hat\omega_j) - \tilde x_t(\hat\omega_k)] \Big\| \right\}$$

$$= \max \left\{ \frac{1}{\sum_{\hat\omega_k\in \tilde E_{ts_l}} p_k} \left\{ \sum_{\hat\omega_k\in E_{ts_{j(l)}}} \Big\| p_l p_k [x_t(\hat\omega_k) - x_t(\hat\omega_l)] \Big\| \right.\right.$$

$$\left. + \Big\| \sum_{\hat\omega_k\in E_{ts_{j(l)}}} p_k p_l [\tilde x_t(\hat\omega_l) - \tilde x_t(\hat\omega_k)] \Big\| \right\},$$

$$\frac{1}{\sum_{\hat\omega_k\in E_{ts_l}} p_k} \left\{ \sum_{\hat\omega_k\in E_{ts_l}\setminus\{\hat\omega_l\}} \Big\| p_l p_k [x_t(\hat\omega_k) - x_t(\hat\omega_l)] \Big\| \right.$$

$$\left.\left. + \Big\| \sum_{\hat\omega_k\in E_{ts_l}\setminus\{\hat\omega_l\}} p_k p_l [\tilde x_t(\hat\omega_l) - \tilde x_t(\hat\omega_k)] \Big\| \right\} \right\}$$

$$\le \max \left\{ \frac{\sum_{\hat\omega_k\in E_{ts_{j(l)}}} 2p_l p_k \|x_t(\hat\omega_k) - x_t(\hat\omega_l)\|}{p_l + \sum_{\hat\omega_k\in E_{ts_{j(l)}}} p_k}, \frac{\sum_{\hat\omega_k\in E_{ts_l}\setminus\{\hat\omega_l\}} 2p_l p_k \|\tilde x_t(\hat\omega_k) - \tilde x_t(\hat\omega_l)\|}{p_l + \sum_{\hat\omega_k\in E_{ts_l}\setminus\{\hat\omega_l\}} p_k} \right\}$$

$$\le 2p_l \max \Big\{ \|x_t(\hat\omega_{j(l)}) - x_t(\hat\omega_l)\|, \min_{\hat\omega_k\in E_{ts_l}\setminus\{\hat\omega_l\}} \|\tilde x_t(\hat\omega_k) - \tilde x_t(\hat\omega_l)\| \Big\}, \tag{2.20}$$

where the convention is used that $\min_{\hat\omega_k\in E_{ts_l}\setminus\{\hat\omega_l\}} = 0$ if $E_{ts_l} \setminus \{\hat\omega_l\} = \emptyset$. The final

estimate makes use of the fact that all $x_t(\hat{\omega}_k)$ with $\hat{\omega}_k \in E_{ts_{j(l)}}$ and $\hat{\omega}_k \in E_{ts_l} \setminus \{\hat{\omega}_l\}$, respectively, coincide.

In the following two cases, the above estimate simplifies to

$$D_t(\xi, \tilde{\xi}) \leq \begin{cases} 0 & \text{if } \hat{\omega}_l \in E_{ts_{j(l)}}, \\ 2p_l \|x_t(\hat{\omega}_{j(l)}) - x_t(\hat{\omega}_l)\| & \text{if } E_{ts_l} = \{\hat{\omega}_l\}. \end{cases}$$

As the sets $l_0(F(\xi, \cdot))$ and $l_0(F(\tilde{\xi}, \cdot))$ of solutions of the original and perturbed multistage models are bounded in $L_{r'}$ due to (A2), there exists a constant $K > 0$ such that

$$D_{\mathrm{f}}(\xi, \tilde{\xi}) \leq K p_l.$$

Hence, if the probability p_l of the deleted scenario is small, the filtration distance is also small. Then there is no need to modify the deletion procedure based on best approximations with respect to the metric ℓ_1. This is mostly the case if the tree is bushy, i.e., contains many scenarios.

A more reliable estimate for the filtration distance may be obtained by solving the stochastic program for an approximation $\hat{\xi}$ of ξ (on $\{\hat{\omega}_1, \ldots, \hat{\omega}_N\}$), which contains much less scenarios than ξ. Then an estimate for the filtration distance may be obtained by computing

$$2p_l \sum_{t=2}^{T-1} \max \Big\{ \|\hat{x}_t(\hat{\omega}_{j(l)}) - \hat{x}_t(\hat{\omega}_l)\|, \min_{\hat{\omega}_k \in E_{ts_l} \setminus \{\hat{\omega}_l\}} \|\hat{x}_t(\hat{\omega}_k) - \hat{x}_t(\hat{\omega}_l)\| \Big\},$$

where $\hat{x} \in l_0(F(\hat{\xi}, \cdot))$ is the corresponding solution. Altogether, some scenario deletion suggested by the strategy in [5, 11] can either be carried out if the bound (2.20) on the filtration distance remains small or is rejected.

3. Conclusions. While quantitative stability results for two-stage stochastic programs have to take into account only a suitable distance of probability distributions, this is no longer the case for multistage models, where the filtration distance enters stability estimates. This fact demonstrates the importance of the conditional structure of multistage stochastic programs. This is in line with the observations and results of [32]. In a sense, it also seems to illustrate the complexity results obtained in the recent paper [33]. It is shown there that multistage stochastic programs have higher complexity than two-stage models. Techniques for generating and reducing *scenario trees* in multistage stochastic programs, which are based on stability arguments, have to respect *both* probability *and* filtration distances as both contribute to changes of optimal values. Example 2.7 provides upper bounds for the filtration distance if some scenario is deleted. Bounding the filtration distance is also possible for the forward and backward scenario tree generation algorithms developed in [10] and [12]. Such bounds are derived and discussed in the companion paper [13].

Acknowledgments. The first two authors wish to thank the members of the OSIRIS Division at R&D of EDF for several stimulating discussions on scenario trees and stability. We extend our gratitude to René Henrion (WIAS Berlin) for his comments on an earlier version of this paper and to two anonymous referees for their insightful comments.

REFERENCES

[1] K. Barty, *Contributions à la discrétisation des contraintes de mesurabilité pour les problèmes d'optimisation stochastique*, Thèse de Doctorat, École Nationale des Ponts et Chaussées, Paris, 2004.

[2] E. S. Boylan, *Equiconvergence of martingales*, Ann. Math. Statist., 42 (1971), pp. 552–559.

[3] K. D. Cotter, *Convergence of information, random variables and noise*, J. Math. Econom., 16 (1987), pp. 39–51.

[4] J. Dupačová, G. Consigli, and S. W. Wallace, *Scenarios for multistage stochastic programs*, Ann. Oper. Res., 100 (2000), pp. 25–53.

[5] J. Dupačová, N. Gröwe-Kuska, and W. Römisch, *Scenario reduction in stochastic programming: An approach using probability metrics*, Math. Program., 95 (2003), pp. 493–511.

[6] I. Evstigneev, *Measurable selection and dynamic programming*, Math. Oper. Res., 1 (1976), pp. 267–272.

[7] O. Fiedler and W. Römisch, *Stability in multistage stochastic programming*, Ann. Oper. Res., 56 (1995), pp. 79–93.

[8] R. Fortet and E. Mourier, *Convergence de la répartition empirique vers la répartition théorique*, Ann. Sci. Ecole Norm. Sup. (3), 70 (1953), pp. 267–285.

[9] C. R. Givens and R. M. Shortt, *A class of Wasserstein metrics for probability distributions*, Michigan Math. J., 31 (1984), pp. 231–240.

[10] N. Gröwe-Kuska, H. Heitsch, and W. Römisch, *Scenario reduction and scenario tree construction for power management problems*, A. Borghetti, C.A. Nucci, and M. Paolone eds., IEEE Bologna Power Tech Proceedings, Bologna, Italy, 2003.

[11] H. Heitsch and W. Römisch, *Scenario reduction algorithms in stochastic programming*, Comput. Optim. Appl., 24 (2003), pp. 187–206.

[12] H. Heitsch and W. Römisch, *Generation of multivariate scenario trees to model stochasticity in power management*, IEEE St. Petersburg Power Tech Proceedings, St. Petersburg, Russia, 2005.

[13] H. Heitsch and W. Römisch, *Scenario tree modelling for multistage stochastic programs*, preprint 296, DFG Research Center Matheon (Mathematics for key technologies), 2005, Berlin, Germany, (www.matheon.de).

[14] R. Hochreiter and G. Ch. Pflug, *Financial scenario generation for stochastic multi-stage decision processes as facility location problem*, Ann. Oper. Res., to appear.

[15] D. N. Hoover, *Convergence in distribution and Skorokhod convergence for the general theory of processes*, Probab. Theory Related Fields, 89 (1991), pp. 239–259.

[16] V. Kaňková, *Empirical estimates in multistage stochastic programs*, Report No. 1930, Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Prague, 1998.

[17] J. Neveu, *Note on the tightness of the metric on the set of complete sub σ-algebras of a probability space*, Ann. Math. Statist., 43 (1972), pp. 1369–1371.

[18] P. Olsen, *Multistage stochastic programming with recourse as mathematical programming in an L_p-space*, SIAM J. Control Optim., 14 (1976), pp. 528–537.

[19] P. Olsen, *Discretizations of multistage stochastic programming problems*, Math. Programming Stud., 6 (1976), pp. 111–124.

[20] T. Pennanen, *Epi-convergent discretizations of multistage stochastic programs via integration quadratures*, Stochastic Programming E-Print Series 19–2004 (www.speps.org) and Math. Program., Ser. B, to appear.

[21] G. Ch. Pflug, *Scenario tree generation for multiperiod financial optimization by optimal discretization*, Math. Program., 89 (2001), pp. 251–271.

[22] S. T. Rachev, *Probability Metrics and the Stability of Stochastic Models*, Wiley and Sons, Chichester, UK, 1991.

[23] S. T. Rachev and W. Römisch, *Quantitative stability in stochastic programming: The method of probability metrics*, Math. Oper. Res., 27 (2002), pp. 792–818.

[24] S. T. Rachev and L. Rüschendorf, *Mass Transportation Problems*, Vol. I and II, Springer, Berlin, 1998.

[25] S. T. Rachev and A. Schief, *On L_p-minimal metrics*, Probab. Math. Statist., 13 (1992), pp. 311–320.

[26] R. T. Rockafellar and R. J-B Wets, *Nonanticipativity and $\mathcal{L}^1$-martingales in stochastic optimization problems*, Math. Programming Stud., 6 (1976), pp. 170–187.

[27] R. T. Rockafellar and R. J-B Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.

[28] W. Römisch, *Stability of stochastic programming problems*, in Stochastic Programming (A. Ruszczyński and A. Shapiro Eds.), Handbooks in Operations Research and Management

Science, 10, Elsevier, Amsterdam, 2003, pp. 483–554.
[29] W. Römisch and R. Schultz, *Stability analysis for stochastic programs*, Ann. Oper. Res., 30 (1991), pp. 241–266.
[30] L. Rogge, *Uniform inequalities for conditional expectations*, Ann. Probab., 2 (1974), pp. 486–489.
[31] A. Ruszczyński and A. Shapiro, eds., *Stochastic Programming*, Handbooks in Operations Research and Management Science, 10, Elsevier, Amsterdam, 2003.
[32] A. Shapiro, *Inference of statistical bounds for multistage stochastic programming problems*, Math. Methods Oper. Res., 58 (2003), pp. 57–68.
[33] A. Shapiro and A. Nemirovski, *On complexity of stochastic programming problems*, in Continuous Optimization: Current Trends and Applications, V. Jeyakumar and A. M. Rubinov eds., Springer, New York, 2005, pp. 111–144.
[34] C. Strugarek, *On the Fortet-Mourier metric for the stability of stochastic programming problems*, Stochastic Programming E-Print Series 25-2004 (www.speps.org).

SIAM J. OPTIM.
Vol. 17, No. 2, pp. 526–557
© 2006 Society for Industrial and Applied Mathematics

A NEW ACTIVE SET ALGORITHM FOR BOX CONSTRAINED OPTIMIZATION*

WILLIAM W. HAGER† AND HONGCHAO ZHANG†

Abstract. An active set algorithm (ASA) for box constrained optimization is developed. The algorithm consists of a nonmonotone gradient projection step, an unconstrained optimization step, and a set of rules for branching between the two steps. Global convergence to a stationary point is established. For a nondegenerate stationary point, the algorithm eventually reduces to unconstrained optimization without restarts. Similarly, for a degenerate stationary point, where the strong second-order sufficient optimality condition holds, the algorithm eventually reduces to unconstrained optimization without restarts. A specific implementation of the ASA is given which exploits the recently developed cyclic Barzilai–Borwein (CBB) algorithm for the gradient projection step and the recently developed conjugate gradient algorithm CG_DESCENT for unconstrained optimization. Numerical experiments are presented using box constrained problems in the CUTEr and MINPACK-2 test problem libraries.

Key words. nonmonotone gradient projection, box constrained optimization, active set algorithm, ASA, cyclic BB method, CBB, conjugate gradient method, CG_DESCENT, degenerate optimization

AMS subject classifications. 90C06, 90C26, 65Y20

DOI. 10.1137/050635225

1. Introduction.

We develop an active set method for the box constrained optimization problem

$$\min \{f(\mathbf{x}) : \mathbf{x} \in \mathcal{B}\}, \tag{1.1}$$

where f is a real-valued, continuously differentiable function defined on the set

$$\mathcal{B} = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{l} \le \mathbf{x} \le \mathbf{u}\}. \tag{1.2}$$

Here $\mathbf{l} < \mathbf{u}$, and possibly $l_i = -\infty$ or $u_i = \infty$.

The box constrained optimization problem appears in a wide range of applications, including the obstacle problem [67], the elastic-plastic torsion problem [47], optimal design problems [7], journal bearing lubrication [20], inversion problems in elastic wave propagation [6], and molecular conformation analysis [48]. Problem (1.1) is often a subproblem of augmented Lagrangian or penalty schemes for general constrained optimization (see [24, 25, 37, 38, 43, 46, 52, 53, 65]). Thus the development of numerical algorithms to efficiently solve (1.1), especially when the dimension is large, is important in both theory and applications.

We begin with an overview of the development of active set methods. A seminal paper is Polyak's 1969 paper [68] which considers a convex, quadratic cost function. The conjugate gradient method is used to explore a face of the feasible set, and the negative gradient is used to leave a face. Since Polyak's algorithm added or dropped

*Received by the editors July 5, 2005; accepted for publication (in revised form) February 9, 2006; published electronically August 16, 2006. This material is based upon work supported by the National Science Foundation under grant 0203270.

http://www.siam.org/journals/siopt/17-2/63522.html

†Department of Mathematics, University of Florida, P.O. Box 118105, Gainesville, FL 32611-8105 (hager@math.ufl.edu, http://www.math.ufl.edu/∼hager; hzhang@math.ufl.edu, http://www.math.ufl.edu/∼hzhang).

only one constraint in each iteration, Dembo and Tulowitzki proposed [32] the conjugate gradient projection (CGP) algorithm which could add and drop many constraints in an iteration. Later, Yang and Tolle [79] further developed this algorithm to obtain finite termination, even when the problem was degenerate at a local minimizer $\mathbf{x}^*$. That is, for some i, $x_i^* = l_i$ or $x_i^* = u_i$ and $\nabla f(\mathbf{x}^*)_i = 0$. Another variation of the CGP algorithm, for which there is a rigorous convergence theory, is developed by Wright [77]. Moré and Toraldo [67] point out that when the CGP scheme starts far from the solution, many iterations may be required to identify a suitable working face. Hence, they propose using the gradient projection method to identify a working face, followed by the conjugate gradient method to explore the face. Their algorithm, called GPCG, has finite termination for nondegenerate quadratic problems. Recently, adaptive conjugate gradient algorithms have been developed by Dostál [35, 36] and Dostál, Friedlander, and Santos [38] which have finite termination for a strictly convex quadratic cost function, even when the problem is degenerate.

For general nonlinear functions, some of the earlier research [3, 19, 49, 61, 66, 71] focused on gradient projection methods. To accelerate the convergence, more recent research has developed Newton and trust region methods (see [26] for an in-depth analysis). In [4, 17, 24, 42] superlinear and quadratic convergence is established for nondegenerate problems, while [44, 46, 60, 63] establish analogous convergence results, even for degenerate problems. Although computing a Newton step can be computationally expensive, approximation techniques, such as a sparse, incomplete Cholesky factorization [62], could be used to reduce the computational expense. Nonetheless, for large-dimensional problems or for problems in which the initial guess is far from the solution, the Newton/trust region approach can be inefficient. In cases when the Newton step is unacceptable, a gradient projection step is preferred.

The affine-scaling interior-point method of Coleman and Li [21, 22, 23] (also see Branch, Coleman, and Li [14]) is a different approach to (1.1), related to the trust region algorithm. More recent research on this strategy includes [33, 58, 59, 76, 83]. These methods are based on a reformulation of the necessary optimality conditions obtained by multiplication with a scaling matrix. The resulting system is often solved by Newton-type methods. Without assuming strict complementarity (i.e., for degenerate problems), the affine-scaling interior-point method converges superlinearly or quadratically, for a suitable choice of the scaling matrix, when the strong second-order sufficient optimality condition [70] holds. When the dimension is large, forming and solving the system of equations at each iteration can be time consuming, unless the problem has special structure. Recently, Zhang [83] proposed an interior-point gradient approach for solving the system at each iteration. Convergence results for other interior-point methods applied to more general constrained optimization appear in [39, 40, 78].

The method developed in this paper is an active set algorithm (ASA) which consists of a nonmonotone gradient projection step, an unconstrained optimization step, and a set of rules for branching between the steps. Global convergence to a stationary point is established. For a nondegenerate stationary point, the ASA eventually reduces to unconstrained optimization without restarts. Similarly, for a degenerate stationary point, where the strong second-order sufficient optimality condition holds, the ASA eventually reduces to unconstrained optimization without restarts. If strict complementarity holds and all the constraints are active at a stationary point, then convergence occurs in a finite number of iterations. In general, our analysis does not show that the strictly active constraints are identified in a finite number of iterations;

instead, when the strong second-order sufficient optimality condition holds, we show that the ASA eventually branches to the unconstrained optimization step, and henceforth, the active set does not change. Thus in the limit, the ASA reduces to unconstrained optimization without restarts. Furthermore, if the ith constraint in (1.1) is strictly active at a stationary point $\mathbf{x}^*$ (i.e., $\nabla f(\mathbf{x}^*)_i \neq 0$) and the iterates $\mathbf{x}_k$ converge to $\mathbf{x}^*$, then the distance between the ith component of $\mathbf{x}_k$ and the associated limit, either l_i or u_i, is on the order of the square of the distance between $\mathbf{x}_k$ and $\mathbf{x}^*$.

A specific implementation of the ASA is given, which utilizes our recently developed cyclic Barzilai–Borwein (CBB) algorithm [30] for the gradient projection step and our recently developed conjugate gradient algorithm CG_DESCENT [54, 55, 56, 57] for the unconstrained optimization step. Recent numerical results [27, 45, 50, 51, 74, 81] indicate that in some cases, a nonmonotone line search is superior to a monotone line search. Moreover, gradient methods based on a Barzilai–Borwein (BB) step [2] have exhibited impressive performance in a variety of applications [7, 10, 28, 29, 48, 64, 72]. The BB methods developed in [8, 9, 10, 11, 12, 69] are all based on a Grippo–Lampariello–Lucidi (GLL) type of line search [50]. We have obtained better performance using an adaptive, nonmonotone line search which originates from [31, 75]. Using the adaptive nonmonotone line search, more constraints can be added or dropped in a single iteration. In addition, the cyclic implementation of the BB step [30], in which the same BB stepsize is reused for several iterations, performs better than the original BB step. Hence, in the gradient projection phase of the ASA, we use the CBB scheme of [30] and an adaptive nonmonotone line search.

After detecting a suitable working face, the ASA branches to the unconstrained optimization algorithm, which operates in a lower-dimensional space since some components of $\mathbf{x}$ are fixed. For the numerical experiments, we implement this step using our conjugate gradient algorithm CG_DESCENT. An attractive feature of this algorithm is that the search directions are always sufficient descent directions; furthermore, when the cost function is a strongly convex quadratic, the ASA converges in a finite number of iterations, even when strict complementary slackness does not hold.

Our paper is organized as follows. In section 2 we present the nonmonotone gradient projection algorithm (NGPA) and analyze its global convergence properties. Section 3 presents the ASA and specifies the requirements of the unconstrained optimization algorithm. Section 4 establishes global convergence results for the ASA, while section 5 analyzes local convergence. Section 6 presents numerical comparisons using box constrained problems in the CUTEr [13] and MINPACK-2 [1] test problem libraries. Finally, the appendix gives a specific implementation of the nonmonotone gradient projection method based on our CBB method.

Throughout this paper, we use the following notation. For any set $\mathcal{S}$, $|\mathcal{S}|$ stands for the number of elements (cardinality) of $\mathcal{S}$, while $\mathcal{S}^c$ is the complement of $\mathcal{S}$. $\|\cdot\|$ is the Euclidean norm of a vector. The subscript k is often used to denote the iteration number in an algorithm, while x_{ki} stands for the ith component of the iterate $\mathbf{x}_k$. The gradient $\nabla f(\mathbf{x})$ is a row vector, while $\mathbf{g}(\mathbf{x}) = \nabla f(\mathbf{x})^\mathsf{T}$ is a column vector; here $^\mathsf{T}$ denotes transpose. The gradient at the iterate $\mathbf{x}_k$ is $\mathbf{g}_k = \mathbf{g}(\mathbf{x}_k)$. We let $\nabla^2 f(\mathbf{x})$ denote the Hessian of f at $\mathbf{x}$. The ball with center $\mathbf{x}$ and radius ρ is denoted $B_\rho(\mathbf{x})$.

2. Nonmonotone gradient projection algorithm. In this section, we consider a generalization of (1.1) in which the box $\mathcal{B}$ is replaced with a nonempty, closed convex set Ω:

$$\min \{f(\mathbf{x}) : \mathbf{x} \in \Omega\}. \tag{2.1}$$

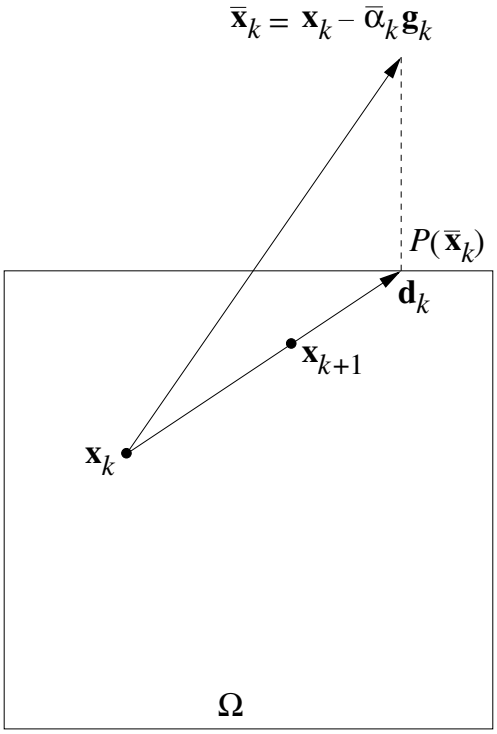


FIG. 2.1. *The gradient projection step.*

We begin with an overview of our gradient projection algorithm. Step k in our algorithm is depicted in Figure 2.1. Here P denotes the projection onto Ω:

$$P(\mathbf{x}) = \arg \min_{\mathbf{y} \in \Omega} \|\mathbf{x} - \mathbf{y}\|. \tag{2.2}$$

Starting at the current iterate $\mathbf{x}_k$, we compute an initial iterate $\bar{\mathbf{x}}_k = \mathbf{x}_k - \bar{\alpha}_k \mathbf{g}_k$. The only constraint on the initial steplength $\bar{\alpha}_k$ is that $\bar{\alpha}_k \in [\alpha_{\min}, \alpha_{\max}]$, where $\alpha_{\min}$ and $\alpha_{\max}$ are fixed, positive constants, independent of k. Since the nominal iterate may lie outside Ω, we compute its projection $P(\bar{\mathbf{x}}_k)$ onto Ω. The search direction is $\mathbf{d}_k = P(\bar{\mathbf{x}}_k) - \mathbf{x}_k$, similar to the choice made in SPG2 [11]. Using a nonmonotone line search along the line segment connecting $\mathbf{x}_k$ and $P(\bar{\mathbf{x}}_k)$, we arrive at the new iterate $\mathbf{x}_{k+1}$.

In the statement of the NGPA given below, f_k^r denotes the "reference" function value. A monotone line search corresponds to the choice $f_k^r = f(\mathbf{x}_k)$. The nonmonotone GLL scheme takes $f_k^r = f_k^{\max}$, where

$$f_k^{\max} = \max\{f(\mathbf{x}_{k-i}) : 0 \le i \le \min(k, M-1)\}. \tag{2.3}$$

Here $M > 0$ is a fixed integer, the memory. In the appendix, we give a procedure for choosing the reference function value based on our CBB scheme.

NGPA PARAMETERS.

- $\epsilon \in [0, \infty)$, error tolerance
- $\delta \in (0, 1)$, descent parameter used in Armijo line search

- $\eta \in (0,1)$, decay factor for stepsize in Armijo line search
- $[\alpha_{\min}, \alpha_{\max}] \subset (0,\infty)$, interval containing initial stepsize

NONMONOTONE GRADIENT PROJECTION ALGORITHM (NGPA).

Initialize $k = 0$, $\mathbf{x}_0$ = starting guess, and $f_{-1}^r = f(\mathbf{x}_0)$.

While $\|P(\mathbf{x}_k - \mathbf{g}_k) - \mathbf{x}_k\| > \epsilon$

1. Choose $\overline{\alpha}_k \in [\alpha_{\min}, \alpha_{\max}]$ and set $\mathbf{d}_k = P(\mathbf{x}_k - \overline{\alpha}_k \mathbf{g}_k) - \mathbf{x}_k$.
2. Choose f_k^r so that $f(\mathbf{x}_k) \le f_k^r \le \max\{f_{k-1}^r, f_k^{\max}\}$ and $f_k^r \le f_k^{\max}$ infinitely often.
3. Let f_R be either f_k^r or $\min\{f_k^{\max}, f_k^r\}$. If $f(\mathbf{x}_k + \mathbf{d}_k) \le f_R + \delta \mathbf{g}_k^{\mathsf{T}} \mathbf{d}_k$, then $\alpha_k = 1$.
4. If $f(\mathbf{x}_k + \mathbf{d}_k) > f_R + \delta \mathbf{g}_k^{\mathsf{T}} \mathbf{d}_k$, then $\alpha_k = \eta^j$, where $j > 0$ is the smallest integer such that
$$f(\mathbf{x}_k + \eta^j \mathbf{d}_k) \le f_R + \eta^j \delta \mathbf{g}_k^{\mathsf{T}} \mathbf{d}_k. \tag{2.4}$$
5. Set $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k$ and $k = k + 1$.

End

The condition $f(\mathbf{x}_k) \le f_k^r$ guarantees that the Armijo line search in step 4 can be satisfied. The requirement that "$f_k^r \le f_k^{\max}$ infinitely often" in step 2 is needed for the global convergence result, Theorem 2.2. This is a rather weak requirement which can be satisfied by many strategies. For example, at every L iteration, we could simply set $f_k^r = f_k^{\max}$. Another strategy, closer in spirit to the one used in the numerical experiments, is to choose a decrease parameter $\Delta > 0$ and an integer $L > 0$ and set $f_k^r = f_k^{\max}$ if $f(\mathbf{x}_{k-L}) - f(\mathbf{x}_k) \le \Delta$.

To begin the convergence analysis, recall that $\mathbf{x}^*$ is a stationary point for (2.1) if the first-order optimality condition holds:

$$\nabla f(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}^*) \ge 0 \quad \text{for all } \mathbf{x} \in \Omega. \tag{2.5}$$

Let $\mathbf{d}^\alpha(\mathbf{x})$, $\alpha \in \mathbb{R}$, be defined in terms of the gradient $\mathbf{g}(\mathbf{x}) = \nabla f(\mathbf{x})^{\mathsf{T}}$ as follows:

$$\mathbf{d}^\alpha(\mathbf{x}) = P(\mathbf{x} - \alpha \mathbf{g}(\mathbf{x})) - \mathbf{x}.$$

In the NGPA, the search direction is $\mathbf{d}_k = \mathbf{d}^{\overline{\alpha}_k}(\mathbf{x}_k)$. For unconstrained optimization, $\mathbf{d}^\alpha(\mathbf{x})$ points along the negative gradient at $\mathbf{x}$ when $\alpha > 0$. Some properties of P and $\mathbf{d}^\alpha$ are summarized below.

PROPOSITION 2.1 (Properties of P and $\mathbf{d}^\alpha$).

P1. $(P(\mathbf{x}) - \mathbf{x})^{\mathsf{T}}(\mathbf{y} - P(\mathbf{x})) \ge 0$ *for all* $\mathbf{x} \in \mathbb{R}^n$ *and* $\mathbf{y} \in \Omega$.

P2. $(P(\mathbf{x}) - P(\mathbf{y}))^{\mathsf{T}}(\mathbf{x} - \mathbf{y}) \ge \|P(\mathbf{x}) - P(\mathbf{y})\|^2$ *for all* $\mathbf{x}$ *and* $\mathbf{y} \in \mathbb{R}^n$.

P3. $\|P(\mathbf{x}) - P(\mathbf{y})\| \le \|\mathbf{x} - \mathbf{y}\|$ *for all* $\mathbf{x}$ *and* $\mathbf{y} \in \mathbb{R}^n$.

P4. $\|\mathbf{d}^\alpha(\mathbf{x})\|$ *is nondecreasing in* $\alpha > 0$ *for any* $\mathbf{x} \in \Omega$.

P5. $\|\mathbf{d}^\alpha(\mathbf{x})\|/\alpha$ *is nonincreasing in* $\alpha > 0$ *for any* $\mathbf{x} \in \Omega$.

P6. $\mathbf{g}(\mathbf{x})^{\mathsf{T}} \mathbf{d}^\alpha(\mathbf{x}) \le -\|\mathbf{d}^\alpha(\mathbf{x})\|^2/\alpha$ *for any* $\mathbf{x} \in \Omega$ *and* $\alpha > 0$.

P7. *For any* $\mathbf{x} \in \Omega$ *and* $\alpha > 0$, $\mathbf{d}^\alpha(\mathbf{x}) = \mathbf{0}$ *if and only if* $\mathbf{x}$ *is a stationary point for* (2.1).

P8. *Suppose* $\mathbf{x}^*$ *is a stationary point for* (2.1). *If for some* $\mathbf{x} \in \mathbb{R}^n$ *there exist positive scalars* λ *and* γ *such that*
$$(\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{x}^*))^{\mathsf{T}}(\mathbf{x} - \mathbf{x}^*) \ge \gamma \|\mathbf{x} - \mathbf{x}^*\|^2 \tag{2.6}$$
and
$$\|\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{x}^*)\| \le \lambda \|\mathbf{x} - \mathbf{x}^*\|, \tag{2.7}$$

then we have

$$\|\mathbf{x}-\mathbf{x}^*\| \le \left(\frac{1+\lambda}{\gamma}\right)\|\mathbf{d}^1(\mathbf{x})\|.$$

Proof. P1 is the first-order optimality condition associated with the solution of (2.2). Replacing $\mathbf{y}$ with $P(\mathbf{y})$ in P1 gives

$$(P(\mathbf{x})-\mathbf{x})^{\mathsf{T}}(P(\mathbf{y})-P(\mathbf{x})) \ge 0.$$

Adding this to the corresponding inequality obtained by interchanging $\mathbf{x}$ and $\mathbf{y}$ yields P2 (see [80]). P3 is the nonexpansive property of a projection (for example, see [5, Prop. 2.1.3]). P4 is given in [73]. For P5, see [5, Lem. 2.3.1]. P6 is obtained from P1 by replacing $\mathbf{x}$ with $\mathbf{x}-\alpha\mathbf{g}(\mathbf{x})$ and replacing $\mathbf{y}$ with $\mathbf{x}$. If $\mathbf{x}^*$ is a stationary point satisfying (2.5), then P6 with $\mathbf{x}$ replaced by $\mathbf{x}^*$ yields $\mathbf{d}^\alpha(\mathbf{x}^*)=\mathbf{0}$. Conversely, if $\mathbf{d}^\alpha(\mathbf{x}^*)=\mathbf{0}$, then by P1 with $\mathbf{x}$ replaced by $\mathbf{x}^*-\alpha\mathbf{g}(\mathbf{x}^*)$, we obtain

$$0 \le \alpha\mathbf{g}(\mathbf{x}^*)^{\mathsf{T}}(\mathbf{y}-P(\mathbf{x}^*-\alpha\mathbf{g}(\mathbf{x}^*)) = \alpha\mathbf{g}(\mathbf{x}^*)^{\mathsf{T}}(\mathbf{y}-\mathbf{x}^*),$$

which implies that $\mathbf{x}^*$ is a stationary point (see [5, Fig. 2.3.2]).

Finally, let us consider P8. Replacing $\mathbf{x}$ with $\mathbf{x}-\mathbf{g}(\mathbf{x})$ and replacing $\mathbf{y}$ with $\mathbf{x}^*$ in P1 gives

$$[P(\mathbf{x}-\mathbf{g}(\mathbf{x}))-\mathbf{x}+\mathbf{g}(\mathbf{x})]^{\mathsf{T}}[\mathbf{x}^*-P(\mathbf{x}-\mathbf{g}(\mathbf{x}))] \ge 0. \tag{2.8}$$

By the definition of $\mathbf{d}^\alpha(\mathbf{x})$, (2.8) is equivalent to

$$[\mathbf{d}^1(\mathbf{x})+\mathbf{g}(\mathbf{x})]^{\mathsf{T}}[\mathbf{x}^*-\mathbf{x}-\mathbf{d}^1(\mathbf{x})] \ge 0.$$

Rearranging this and utilizing (2.6) gives

$$\begin{aligned}\mathbf{d}^1(\mathbf{x})^{\mathsf{T}}(\mathbf{x}^*-\mathbf{x})-\mathbf{g}(\mathbf{x})^{\mathsf{T}}\mathbf{d}^1(\mathbf{x})-\|\mathbf{d}^1(\mathbf{x})\|^2 &\ge \mathbf{g}(\mathbf{x})^{\mathsf{T}}(\mathbf{x}-\mathbf{x}^*)\\ &\ge \gamma\|\mathbf{x}-\mathbf{x}^*\|^2+\mathbf{g}(\mathbf{x}^*)^{\mathsf{T}}(\mathbf{x}-\mathbf{x}^*).\end{aligned} \tag{2.9}$$

Focusing on the terms involving $\mathbf{g}$ and utilizing (2.7), we have

$$\begin{aligned}\mathbf{g}(\mathbf{x}^*)^{\mathsf{T}}(\mathbf{x}^*-\mathbf{x})-\mathbf{g}(\mathbf{x})^{\mathsf{T}}\mathbf{d}^1(\mathbf{x}) &\le \lambda\|\mathbf{x}-\mathbf{x}^*\|\,\|\mathbf{d}^1(\mathbf{x})\|+\mathbf{g}(\mathbf{x}^*)^{\mathsf{T}}(\mathbf{x}^*-\mathbf{x}-\mathbf{d}^1(\mathbf{x}))\\ &= \lambda\|\mathbf{x}-\mathbf{x}^*\|\,\|\mathbf{d}^1(\mathbf{x})\|+\mathbf{g}(\mathbf{x}^*)^{\mathsf{T}}[\mathbf{x}^*-P(\mathbf{x}-\mathbf{g}(\mathbf{x}))]\\ &\le \lambda\|\mathbf{x}-\mathbf{x}^*\|\,\|\mathbf{d}^1(\mathbf{x})\|\end{aligned} \tag{2.10}$$

by (2.5), since $P(\mathbf{x}-\mathbf{g}(\mathbf{x})) \in \Omega$. Combining (2.9) and (2.10), the proof is complete. ∎

Next, we establish a convergence result for the NGPA.

THEOREM 2.2. *Let $\mathcal{L}$ be the level set defined by*

$$\mathcal{L}=\{\mathbf{x}\in\Omega : f(\mathbf{x}) \le f(\mathbf{x}_0)\}. \tag{2.11}$$

We assume the following conditions hold:

G1. *f is bounded from below on $\mathcal{L}$ and $d_{\max}=\sup_k\|\mathbf{d}_k\|<\infty$.*

G2. *If $\bar{\mathcal{L}}$ is the collection of $\mathbf{x}\in\Omega$ whose distance to $\mathcal{L}$ is at most $d_{\max}$, then ∇f is Lipschitz continuous on $\bar{\mathcal{L}}$.*

Then either the NGPA with $\epsilon = 0$ terminates in a finite number of iterations at a stationary point, or we have

$$\liminf_{k\to\infty} \|\mathbf{d}^1(\mathbf{x}_k)\| = 0.$$

Proof. By P6, the search direction $\mathbf{d}_k$ generated in step 1 of the NGPA is a descent direction. Since $f_k^r \geq f(\mathbf{x}_k)$ and $\delta < 1$, the Armijo line search condition (2.4) is satisfied for j sufficiently large. We now show that $\mathbf{x}_k \in \mathcal{L}$ for each k. Since $f_0^{\max} = f_{-1}^r = f(\mathbf{x}_0)$, step 2 of the NGPA implies that $f_0^r \leq f(\mathbf{x}_0)$. Proceeding by induction, suppose that for some $k \geq 0$, we have

$$f_j^r \leq f(\mathbf{x}_0) \quad \text{and} \quad f_j^{\max} \leq f(\mathbf{x}_0) \tag{2.12}$$

for all $j \in [0, k]$. Again, since the search direction $\mathbf{d}_k$ generated in step 1 of the NGPA is a descent direction, it follows from steps 3 and 4 of the NGPA and the induction hypothesis that

$$f(\mathbf{x}_{k+1}) \leq f_k^r \leq f(\mathbf{x}_0). \tag{2.13}$$

Hence, $f_{k+1}^{\max} \leq f(\mathbf{x}_0)$ and $f_{k+1}^r \leq \max\{f_k^r, f_{k+1}^{\max}\} \leq f(\mathbf{x}_0)$. This completes the induction. Thus (2.12) holds for all j. Consequently, we have $f_R \leq f(\mathbf{x}_0)$ in steps 3 and 4 of the NGPA. Again, since the search direction $\mathbf{d}_k$ generated in step 1 of the NGPA is a descent direction, it follows from steps 3 and 4 that $f(\mathbf{x}_k) \leq f(\mathbf{x}_0)$, which implies that $\mathbf{x}_k \in \mathcal{L}$ for each k.

Let λ be the Lipschitz constant for ∇f on $\bar{\mathcal{L}}$. As in [81, Lem. 2.1], we have

$$\alpha_k \geq \min\left\{1, \left(\frac{2\eta(1-\delta)}{\lambda}\right)\frac{|\mathbf{g}_k^{\mathsf{T}}\mathbf{d}_k|}{\|\mathbf{d}_k\|^2}\right\} \tag{2.14}$$

for all k. By P6,

$$|\mathbf{g}_k^{\mathsf{T}}\mathbf{d}_k| \geq \frac{\|\mathbf{d}_k\|^2}{\overline{\alpha}_k} \geq \frac{\|\mathbf{d}_k\|^2}{\alpha_{\max}}.$$

It follows from (2.14) that

$$\alpha_k \geq \min\left\{1, \left(\frac{2\eta(1-\delta)}{\lambda\alpha_{\max}}\right)\right\} := c. \tag{2.15}$$

By steps 3 and 4 of the NGPA and P6, we conclude that

$$f(\mathbf{x}_{k+1}) \leq f_k^r + \delta c\mathbf{g}_k^{\mathsf{T}}\mathbf{d}_k \leq f_k^r - \delta c\|\mathbf{d}_k\|^2/\overline{\alpha}_k \leq f_k^r - \delta c\|\mathbf{d}_k\|^2/\alpha_{\max}. \tag{2.16}$$

We now prove that $\liminf_{k\to\infty} \|\mathbf{d}_k\| = 0$. Suppose, to the contrary, that there exists a constant $\gamma > 0$ such that $\|\mathbf{d}_k\| \geq \gamma$ for all k. By (2.16), we have

$$f(\mathbf{x}_{k+1}) \leq f_k^r - \tau, \quad \text{where } \tau = \delta c\gamma^2/\alpha_{\max}. \tag{2.17}$$

Let k_i, $i = 0, 1, \ldots$, denote an increasing sequence of integers with the property that $f_j^r \leq f_j^{\max}$ for $j = k_i$ and $f_j^r \leq f_{j-1}^r$ when $k_i < j < k_{i+1}$. Such a sequence exists by the requirement on f_k^r given in step 2 of the NGPA. Hence, we have

$$f_j^r \leq f_{k_i}^r \leq f_{k_i}^{\max} \quad \text{when } k_i \leq j < k_{i+1}. \tag{2.18}$$

By (2.17) it follows that

$$f(\mathbf{x}_j) \le f_{j-1}^r - \tau \le f_{k_i}^{\max} - \tau \quad \text{when } k_i < j \le k_{i+1}. \tag{2.19}$$

It follows that

$$f_{k_{i+1}}^r \le f_{k_{i+1}}^{\max} \le f_{k_i}^{\max}. \tag{2.20}$$

Hence, if $a = k_{i_1}$ and $b = k_{i_2}$, where $i_1 > i_2$ and $a - b > M$, then by (2.18)–(2.20) we have

$$f_a^{\max} = \max_{0 \le j < M} f(\mathbf{x}_{a-j}) \le \max_{1 \le j \le M} f_{a-j}^r - \tau \le f_b^{\max} - \tau.$$

Since the sequence k_i, $i = 0, 1, \ldots$, is infinite, this contradicts the fact that f is bounded from below. Consequently, $\liminf_{k\to\infty} \|\mathbf{d}_k\| = 0$. By P4 and P5, it follows that

$$\|\mathbf{d}_k\| \ge \min\{\alpha_{\min}, 1\} \|\mathbf{d}^1(\mathbf{x}_k)\|.$$

Thus $\liminf_{k\to\infty} \|\mathbf{d}^1(\mathbf{x}_k)\| = 0$. □

Recall that f is strongly convex on Ω if there exists a scalar $\gamma > 0$ such that

$$f(\mathbf{x}) \ge f(\mathbf{y}) + \nabla f(\mathbf{y})(\mathbf{x} - \mathbf{y}) + \frac{\gamma}{2}\|\mathbf{x} - \mathbf{y}\|^2 \tag{2.21}$$

for all $\mathbf{x}$ and $\mathbf{y} \in \Omega$. Interchanging $\mathbf{x}$ and $\mathbf{y}$ in (2.21) and adding, we obtain the (usual) monotonicity condition

$$(\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}))(\mathbf{y} - \mathbf{x}) \ge \gamma \|\mathbf{y} - \mathbf{x}\|^2. \tag{2.22}$$

For a strongly convex function, (2.1) has a unique minimizer $\mathbf{x}^*$, and the conclusion of Theorem 2.2 can be strengthened as follows.

COROLLARY 2.3. *Suppose f is strongly convex and twice continuously differentiable on Ω, and there is a positive integer L with the property that for each k, there exists $j \in [k, k + L)$ such that $f_j^r \le f_j^{\max}$. Then the iterates $\mathbf{x}_k$ of the NGPA with $\epsilon = 0$ converge to the global minimizer $\mathbf{x}^*$.*

Proof. As shown at the start of the proof of Theorem 2.2, $f(\mathbf{x}_k) \le f(\mathbf{x}_0)$ for each k. Hence, $\mathbf{x}_k$ lies in the level set $\mathcal{L}$ defined in (2.11). Since f is strongly convex, $\mathcal{L}$ is a bounded set; since f is twice continuously differentiable, $\|\nabla f(\mathbf{x}_k)\|$ is bounded uniformly in k. For any $\mathbf{x} \in \Omega$, we have $P(\mathbf{x}) = \mathbf{x}$. By P3, it follows that

$$\|\mathbf{d}^\alpha\| = \|P(\mathbf{x} - \alpha \mathbf{g}(\mathbf{x})) - \mathbf{x}\| = \|P(\mathbf{x} - \alpha \mathbf{g}(\mathbf{x})) - P(\mathbf{x})\| \le \alpha \|\mathbf{g}(\mathbf{x})\|.$$

Since $\bar{\alpha}_k \in [\alpha_{\min}, \alpha_{\max}]$, $d_{\max} = \sup_k \|\mathbf{d}_k\| < \infty$. Consequently, the set $\bar{\mathcal{L}}$ defined in G2 is bounded. Again, since f is twice continuously differentiable, ∇f is Lipschitz continuous on $\bar{\mathcal{L}}$. By assumption, $f_k^r \le f_k^{\max}$ infinitely often. Consequently, the hypotheses of Theorem 2.2 are satisfied, and either the NGPA with $\epsilon = 0$ terminates in a finite number of iterations at a stationary point, or we have

$$\liminf_{k\to\infty} \|\mathbf{d}^1(\mathbf{x}_k)\| = 0. \tag{2.23}$$

Since f is strongly convex on Ω, $\mathbf{x}^*$ is the unique stationary point for (2.1). Hence, when the iterates converge in a finite number of steps, they converge to $\mathbf{x}^*$. Otherwise,

(2.23) holds, in which case there exists an infinite sequence $l_1 < l_2 < \cdots$ such that $\|\mathbf{d}^1(\mathbf{x}_{l_j})\|$ approaches zero as j tends to ∞. Since (2.22) holds, it follows from P8 that $\mathbf{x}_{l_j}$ approaches $\mathbf{x}^*$ as j tends to ∞. By P4 and P5, we have

$$\|\mathbf{d}^\alpha(\mathbf{x})\| \le \max\{1, \alpha\}\|\mathbf{d}^1(\mathbf{x})\|.$$

Since $\bar{\alpha}_k \in [\alpha_{\min}, \alpha_{\max}]$, it follows that

$$\|\mathbf{d}_k\| \le \max\{1, \alpha_{\max}\}\|\mathbf{d}^1(\mathbf{x}_k)\|.$$

Since the stepsize $\alpha_k \in (0, 1]$, we deduce that

$$\|\mathbf{x}_{k+1} - \mathbf{x}_k\| = \alpha_k\|\mathbf{d}_k\| \le \|\mathbf{d}_k\| \le \max\{1, \alpha_{\max}\}\|\mathbf{d}^1(\mathbf{x}_k)\|. \tag{2.24}$$

By P3, P is continuous; consequently, $\mathbf{d}^\alpha(\mathbf{x})$ is a continuous function of $\mathbf{x}$. The continuity of $\mathbf{d}^\alpha(\cdot)$ and $f(\cdot)$ combined with (2.24) and the fact that $\mathbf{x}_{l_j}$ converges to $\mathbf{x}^*$ implies that for any $\delta > 0$ and for j sufficiently large, we have

$$f(\mathbf{x}_k) \le f(\mathbf{x}^*) + \delta \quad \text{for all } k \in [l_j, l_j + M + L].$$

By the definition of $f_k^{\max}$,

$$f_k^{\max} \le f(\mathbf{x}^*) + \delta \quad \text{for all } k \in [l_j + M, l_j + M + L]. \tag{2.25}$$

As in the proof of Theorem 2.2, let k_i, $i = 0, 1, \ldots$, denote an increasing sequence of integers with the property that $f_j^r \le f_j^{\max}$ for $j = k_i$ and $f_j^r \le f_{j-1}^r$ when $k_i < j < k_{i+1}$. As shown in (2.20),

$$f_{k_{i+1}}^{\max} \le f_{k_i}^{\max} \tag{2.26}$$

for each i. The assumption that for each k, there exists $j \in [k, k + L)$ such that $f_j^r \le f_j^{\max}$, implies that

$$k_{i+1} - k_i \le L. \tag{2.27}$$

Combining (2.25) and (2.27), for each l_j, there exists some $k_i \in [l_j + M, l_j + M + L]$ and

$$f_{k_i}^{\max} \le f(\mathbf{x}^*) + \delta. \tag{2.28}$$

Since δ was arbitrary, it follows from (2.26) and (2.28) that

$$\lim_{i\to\infty} f_{k_i}^{\max} = f(\mathbf{x}^*); \tag{2.29}$$

the convergence is monotone by (2.26). By the choice of k_i and by the inequality $f(\mathbf{x}_k) \le f_k^r$ in step 2, we have

$$f(\mathbf{x}_k) \le f_k^r \le f_{k_i}^{\max} \quad \text{for all } k \ge k_i. \tag{2.30}$$

Combining (2.29) and (2.30),

$$\lim_{k\to\infty} f(\mathbf{x}_k) = f(\mathbf{x}^*). \tag{2.31}$$

Together, (2.5) and (2.21) yield

$$f(\mathbf{x}_k) \ge f(\mathbf{x}^*) + \frac{\gamma}{2}\|\mathbf{x}_k - \mathbf{x}^*\|^2. \tag{2.32}$$

Combining this with (2.31), the proof is complete. ∎

3. The active set algorithm. Starting with this section, we focus on the box constrained problem (1.1). To simplify the exposition, we consider the special case when $\mathbf{l} = \mathbf{0}$ and $\mathbf{u} = \boldsymbol{\infty}$:

$$\min \{f(\mathbf{x}) : \mathbf{x} \geq \mathbf{0}\}.$$

We emphasize that the analysis and algorithm apply to the general box constrained problem (1.1) with both upper and lower bounds.

Although the gradient projection scheme of the NGPA has an attractive global convergence theory, the convergence rate can be slow in a neighborhood of a local minimizer. In contrast, for unconstrained optimization, the conjugate gradient algorithm often exhibits superlinear convergence in a neighborhood of a local minimizer. We develop an ASA which uses the NGPA to identify active constraints, and which uses an unconstrained optimization algorithm, such as the CG_DESCENT scheme in [54, 55, 57, 56], to optimize f over a face identified by the NGPA.

We begin with some notation. For any $\mathbf{x} \in \Omega$, let $\mathcal{A}(\mathbf{x})$ and $\mathcal{I}(\mathbf{x})$ denote the active and inactive indices, respectively:

$$\mathcal{A}(\mathbf{x}) = \{i \in [1, n] : x_i = 0\},$$
$$\mathcal{I}(\mathbf{x}) = \{i \in [1, n] : x_i > 0\}.$$

The active indices are further subdivided into those indices satisfying strict complementarity and the degenerate indices:

$$\mathcal{A}_+(\mathbf{x}) = \{i \in \mathcal{A}(\mathbf{x}) : g_i(\mathbf{x}) > 0\},$$
$$\mathcal{A}_0(\mathbf{x}) = \{i \in \mathcal{A}(\mathbf{x}) : g_i(\mathbf{x}) = 0\}.$$

We let $\mathbf{g}_I(\mathbf{x})$ denote the vector whose components associated with the set $\mathcal{I}(\mathbf{x})$ are identical to those of $\mathbf{g}(\mathbf{x})$, while the components associated with $\mathcal{A}(\mathbf{x})$ are zero:

$$g_{Ii}(\mathbf{x}) = \begin{cases} 0 & \text{if } x_i = 0, \\ g_i(\mathbf{x}) & \text{if } x_i > 0. \end{cases}$$

An important feature of our algorithm is that we try to distinguish between active constraints satisfying strict complementarity and active constraints that are degenerate using an identification strategy, which is related to the idea of an identification function introduced in [41]. Given fixed parameters $\alpha \in (0, 1)$ and $\beta \in (1, 2)$, we define the (undecided index) set $\mathcal{U}$ at $\mathbf{x} \in \mathcal{B}$ as follows:

$$\mathcal{U}(\mathbf{x}) = \{i \in [1, n] : |g_i(\mathbf{x})| \geq \|\mathbf{d}^1(\mathbf{x})\|^\alpha \text{ and } x_i \geq \|\mathbf{d}^1(\mathbf{x})\|^\beta\}.$$

In the numerical experiments, we take $\alpha = 1/2$ and $\beta = 3/2$. In practice, $\mathcal{U}$ is almost always empty when we reach a neighborhood of a minimizer, and the specific choice of α and β does not have a significant effect on convergence. The introduction of the $\mathcal{U}$ set leads to a strong local convergence theory developed in section 5.

The indices in $\mathcal{U}$ correspond to components of $\mathbf{x}$ for which the associated gradient component $g_i(\mathbf{x})$ is relatively large, while x_i is not close to 0 (in the sense that $x_i \geq \|\mathbf{d}^1(\mathbf{x})\|^\beta$). When the set $\mathcal{U}$ of uncertain indices is empty, we feel that the indices with large associated gradient components are almost identified. In this case we prefer the unconstrained optimization algorithm.

Although our numerical experiments are based on the conjugate gradient code CG_DESCENT, a broad class of unconstrained optimization algorithms (UAs) can

be applied. The following requirements for the UA are sufficient for establishing the convergence results that follow. Conditions U1–U3 are sufficient for global convergence, while U1–U4 are sufficient for the local convergence analysis. Condition U4 could be replaced with another descent condition for the initial line search; however, the analysis of section 5 has been carried out under U4.

Unconstrained algorithm (UA) requirements.

U1. $\mathbf{x}_k \geq \mathbf{0}$ and $f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k)$ for each k.

U2. $\mathcal{A}(\mathbf{x}_k) \subset \mathcal{A}(\mathbf{x}_{k+1})$ for each k.

U3. If $\mathcal{A}(\mathbf{x}_{j+1}) = \mathcal{A}(\mathbf{x}_j)$ for $j \geq k$, then $\liminf_{j\to\infty} \|\mathbf{g}_I(\mathbf{x}_j)\| = 0$.

U4. Whenever the UA is started, $\mathbf{x}_{k+1} = P(\mathbf{x}_k - \alpha_k \mathbf{g}_I(\mathbf{x}_k))$, where α_k is obtained from a Wolfe line search. That is, α_k is chosen to satisfy

$$\phi(\alpha_k) \leq \phi(0) + \delta\alpha_k\phi'(0) \quad \text{and} \quad \phi'(\alpha_k) \geq \sigma\phi'(0), \tag{3.1}$$

where

$$\phi(\alpha) = f(P(\mathbf{x}_k - \alpha\mathbf{g}_I(\mathbf{x}_k))), \quad 0 < \delta < \sigma < 1. \tag{3.2}$$

Condition U1 implies that the UA is a monotone algorithm, so that the cost function can only decrease in each iteration. Condition U2 concerns how the algorithm behaves when an infeasible iterate is generated. Condition U3 describes the global convergence of the UA when the active set does not change. In U4, $\phi'(\alpha)$ is the derivative from the right side of α; α_k exists since ϕ is piecewise smooth with a finite number of discontinuities in its derivative, and $\phi'(\alpha)$ is continuous at $\alpha = 0$.

Our ASA is presented in Figure 3.1. In the first step of the algorithm, we execute the NGPA until we feel that the active constraints satisfying strict complementarity have been identified. In step 2, we execute the UA until a subproblem has been solved (step 2a). When new constraints become active in step 2b, we may decide to restart either the NGPA or the UA. By restarting the NGPA, we mean that $\mathbf{x}_0$ in the NGPA is identified with the current iterate $\mathbf{x}_k$. By restarting the UA, we mean that iterates are generated by the UA using the current iterate as the starting point.

4. Global convergence. We begin with a global convergence result for the ASA.

Theorem 4.1. *Let $\mathcal{L}$ be the level set defined by*

$$\mathcal{L} = \{\mathbf{x} \in \mathcal{B} : f(\mathbf{x}) \leq f(\mathbf{x}_0)\}.$$

Assume the following conditions hold:

A1. *f is bounded from below on $\mathcal{L}$ and $d_{\max} = \sup_k \|\mathbf{d}_k\| < \infty$.*

A2. *If $\bar{\mathcal{L}}$ is the collection of $\mathbf{x} \in \mathcal{B}$ whose distance to $\mathcal{L}$ is at most $d_{\max}$, then ∇f is Lipschitz continuous on $\bar{\mathcal{L}}$.*

A3. *The UA satisfies* U1–U3.

Then either the ASA with $\epsilon = 0$ terminates in a finite number of iterations at a stationary point, or we have

$$\liminf_{k\to\infty} \|\mathbf{d}^1(\mathbf{x}_k)\| = 0. \tag{4.1}$$

Proof. If only the NGPA is performed for large k, then (4.1) follows from Theorem 2.2. If only the UA is performed for large k, then by U2, the active sets $\mathcal{A}(\mathbf{x}_k)$ must approach a limit. Since μ does not change in the UA, it follows from U3 and the condition $\|\mathbf{g}_I(\mathbf{x}_k)\| \geq \mu\|\mathbf{d}^1(\mathbf{x}_k)\|$ that (4.1) holds. Finally, suppose that the NGPA

ASA PARAMETERS.

- $\epsilon \in [0, \infty)$, error tolerance, stop when $\|\mathbf{d}^1(\mathbf{x}_k)\| \leq \epsilon$
- $\mu \in (0, 1)$, $\|\mathbf{g}_I(\mathbf{x}_k)\| < \mu \|\mathbf{d}^1(\mathbf{x}_k)\|$ implies subproblem solved
- $\rho \in (0, 1)$, decay factor for μ tolerance
- $n_1 \in [1, n)$, number of repeated $\mathcal{A}(\mathbf{x}_k)$ before switch from the NGPA to the UA
- $n_2 \in [1, n)$, used in switch from the UA to the NGPA

ACTIVE SET ALGORITHM (ASA).

1. While $\|\mathbf{d}^1(\mathbf{x}_k)\| > \epsilon$ execute the NGPA and check the following:
   a. If $\mathcal{U}(\mathbf{x}_k) = \emptyset$, then
      If $\|\mathbf{g}_I(\mathbf{x}_k)\| < \mu \|\mathbf{d}^1(\mathbf{x}_k)\|$, then $\mu = \rho\mu$.
      Otherwise, goto step 2.
   b. Else if $\mathcal{A}(\mathbf{x}_k) = \mathcal{A}(\mathbf{x}_{k-1}) = \cdots = \mathcal{A}(\mathbf{x}_{k-n_1})$, then
      If $\|\mathbf{g}_I(\mathbf{x}_k)\| \geq \mu \|\mathbf{d}^1(\mathbf{x}_k)\|$, then goto step 2.

   End

2. While $\|\mathbf{d}^1(\mathbf{x}_k)\| > \epsilon$ execute the UA and check the following:
   a. If $\|\mathbf{g}_I(\mathbf{x}_k)\| < \mu \|\mathbf{d}^1(\mathbf{x}_k)\|$, then restart the NGPA (step 1).
   b. If $|\mathcal{A}(\mathbf{x}_{k-1})| < |\mathcal{A}(\mathbf{x}_k)|$, then
      If $\mathcal{U}(\mathbf{x}_k) = \emptyset$ or $|\mathcal{A}(\mathbf{x}_k)| > |\mathcal{A}(\mathbf{x}_{k-1})| + n_2$, restart the UA at $\mathbf{x}_k$.
      Else restart the NGPA.
      End

   End

FIG. 3.1. *Statement of the ASA.*

is restarted an infinite number of times at $k_1 < k_2 < \cdots$ and that it terminates at $k_1 + l_1 < k_2 + l_2 < \cdots$, respectively. Thus $k_i < k_i + l_i \leq k_{i+1}$ for each i. If (4.1) does not hold, then by (2.19) and (2.20), we have

$$f(\mathbf{x}_{k_i + l_i}) \leq f(\mathbf{x}_{k_i}) - \tau. \tag{4.2}$$

By U1,

$$f(\mathbf{x}_{k_{i+1}}) \leq f(\mathbf{x}_{k_i + l_i}). \tag{4.3}$$

Combining (4.2) and (4.3), we have $f(\mathbf{x}_{k_{i+1}}) \leq f(\mathbf{x}_{k_i}) - \tau$, which contradicts the assumption that f is bounded from below. □

When f is strongly convex, the entire sequence of iterates converges to the global minimizer $\mathbf{x}^*$, as stated in the following corollary. Since the proof of this result relies on the local convergence analysis, the proof is delayed until the end of section 5.

COROLLARY 4.2. *If f is strongly convex and twice continuously differentiable on $\mathcal{B}$, and assumption* A3 *of Theorem* 4.1 *is satisfied, then the iterates $\mathbf{x}_k$ of the ASA with $\epsilon = 0$ converge to the global minimizer $\mathbf{x}^*$.*

5. Local convergence. In the next series of lemmas, we analyze local convergence properties of the ASA. We begin by focusing on nondegenerate stationary points; that is, stationary points $\mathbf{x}^*$ with the property that $g_i(\mathbf{x}^*) > 0$ whenever $x_i^* = 0$.

5.1. Nondegenerate problems. In this case, it is relatively easy to show that the ASA eventually performs only the UA without restarts. The analogous result for degenerate problems is established in section 5.2.

THEOREM 5.1. *If f is continuously differentiable, $0 < \mu \le 1$, and the iterates $\mathbf{x}_k$ generated by the ASA with $\epsilon = 0$ converge to a nondegenerate stationary point $\mathbf{x}^*$, then after a finite number of iterations, the ASA performs only the UA without restarts.*

Proof. Since $\mathbf{x}^*$ is a nondegenerate stationary point and f is continuously differentiable, there exists $\rho > 0$ with the property that for all $\mathbf{x} \in \mathcal{B}_\rho(\mathbf{x}^*)$, we have

$$g_i(\mathbf{x}) > 0 \text{ if } i \in \mathcal{A}(\mathbf{x}^*) \quad \text{and} \quad x_i > 0 \text{ if } i \in \mathcal{A}(\mathbf{x}^*)^c. \tag{5.1}$$

Let k_+ be chosen large enough that $\mathbf{x}_k \in \mathcal{B}_\rho(\mathbf{x}^*)$ for all $k \ge k_+$. If $k \ge k_+$ and $x_{ki} = 0$, then $d_{ki} = 0$ in step 1 of the NGPA. Hence, $x_{k+1,i} = 0$ if $\mathbf{x}_{k+1}$ is generated by the NGPA. By U2, the UA cannot free a bound constraint. It follows that if $k \ge k_+$ and $x_{ki} = 0$, then $x_{ji} = 0$ for all $j \ge k$. Consequently, there exists an index $K \ge k_+$ with the property that $\mathcal{A}(\mathbf{x}_k) = \mathcal{A}(\mathbf{x}_j)$ for all $j \ge k \ge K$.

For any index i, $|d_i^1(\mathbf{x})| \le |g_i(\mathbf{x})|$. Suppose $\mathbf{x} \in \mathcal{B}_\rho(\mathbf{x}^*)$; by (5.1), $d_i^1(\mathbf{x}) = 0$ if $x_i = 0$. Hence,

$$\|\mathbf{d}^1(\mathbf{x})\| \le \|\mathbf{g}_I(\mathbf{x})\| \tag{5.2}$$

for all $\mathbf{x} \in \mathcal{B}_\rho(\mathbf{x}^*)$. If $k > K + n_1$, then in step 1b of the ASA, it follows from (5.2) and the assumption $\mu \in (0, 1]$ that the NGPA will branch to step 2 (UA). In step 2, the condition "$\|\mathbf{g}_I(\mathbf{x}_k)\| < \mu\|\mathbf{d}^1(\mathbf{x}_k)\|$" of step 2a is never satisfied by (5.2). Moreover, the condition "$|\mathcal{A}(\mathbf{x}_{k-1})| < |\mathcal{A}(\mathbf{x}_k)|$" of step 2b is never satisfied since $k > K$. Hence, the iterates never branch from the UA to the NPGA and the UA is never restarted. ∎

5.2. Degenerate problems. We now focus on degenerate problems and show that a result analogous to Theorem 5.1 holds under the strong second-order sufficient optimality condition. We begin with a series of preliminary results.

LEMMA 5.2. *If f is twice-continuously differentiable and there exists an infinite sequence of iterates $\mathbf{x}_k$ generated by the ASA with $\epsilon = 0$ converging to a stationary point $\mathbf{x}^*$, $\mathbf{x}_k \ne \mathbf{x}^*$ for each k, then for each $i \in \mathcal{A}_+(\mathbf{x}^*)$ we have*

$$\limsup_{k\to\infty} \frac{x_{ki}}{\|\mathbf{x}_k - \mathbf{x}^*\|^2} < \infty. \tag{5.3}$$

Proof. Assume that $\mathcal{A}_+(\mathbf{x}^*)$ is nonempty; otherwise there is nothing to prove. Let k_+ be chosen large enough that $g_i(\mathbf{x}_k) > 0$ for all $i \in \mathcal{A}_+(\mathbf{x}^*)$ and $k \ge k_+$. Since f is twice-continuously differentiable, ∇f is Lipschitz continuous in a neighborhood of $\mathbf{x}^*$. Choose $\rho > 0$ and let λ be the Lipschitz constant for ∇f in the ball $B_\rho(\mathbf{x}^*)$ with center $\mathbf{x}^*$ and radius ρ. Since $\mathbf{d}^1(\mathbf{x}^*) = \mathbf{0}$, it follows from the continuity of $\mathbf{d}^1(\cdot)$ that $\mathbf{d}_k$ tends to $\mathbf{0}$ (see (2.24)). Choose k_+ large enough that the ball with center $\mathbf{x}_k$ and radius $\|\mathbf{d}_k\|$ is contained in $B_\rho(\mathbf{x}^*)$ for all $k \ge k_+$. If $x_{li} = 0$ for some $i \in \mathcal{A}_+(\mathbf{x}^*)$ and $l \ge k_+$, then by the definition of $\mathbf{d}_k$ in the NGPA, we have $d_{ki} = 0$ for all $k \ge l$. Hence, $x_{ki} = 0$ for each $k \ge l$ in the NGPA. Likewise, in the UA it follows from U2 that $x_{ji} = 0$ for $j \ge k$ when $x_{ki} = 0$; that is, the UA does not free an active constraint. In other words, when an index $i \in \mathcal{A}_+(\mathbf{x}^*)$ becomes active at iterate $\mathbf{x}_k$, $k \ge k_+$, it remains active for all the subsequent iterations. Thus (5.3) holds trivially for any $i \in \mathcal{A}_+(\mathbf{x}^*)$ with the property that $x_{ki} = 0$ for some $k \ge k_+$.

Now, let us focus on the nontrivial indices in $\mathcal{A}_+(\mathbf{x}^*)$. That is, suppose that there exists $l \in \mathcal{A}_+(\mathbf{x}^*)$ and $x_{kl} > 0$ for all $k \geq k_+$. By the analysis given in the previous paragraph, when k_+ is sufficiently large,

$$\text{either } x_{ki} > 0 \quad \text{or} \quad x_{ki} = 0 \tag{5.4}$$

for all $k \geq k_+$ and $i \in \mathcal{A}_+(\mathbf{x}^*)$ (since an index $i \in \mathcal{A}_+(\mathbf{x}^*)$, which becomes active at iterate $\mathbf{x}_k$, remains active for all the subsequent iterations). We consider the following possible cases.

Case 1. For an infinite number of iterations k, $\mathbf{x}_k$ is generated by the UA, and the UA is restarted a finite number of times.

In this case, the ASA eventually performs only the UA, without restarts. By U2 and U3, we have $\liminf_{k\to\infty} \|\mathbf{g}_I(\mathbf{x}_k)\| = 0$. On the other hand, by assumption, $l \in \mathcal{I}(\mathbf{x}_k)$ for $k \geq k_+$ and $g_l(\mathbf{x}^*) > 0$, which is a contradiction since $g_l(\mathbf{x}_k)$ converges to $g_l(\mathbf{x}^*)$.

Case 2. For an infinite number of iterations k, $\mathbf{x}_k$ is generated by the UA, and the UA is restarted an infinite number of times.

In this case, we will show that after a finite number of iterations, $x_{ki} = 0$ for all $i \in \mathcal{A}_+(\mathbf{x}^*)$. Suppose, to the contrary, that there exists an $l \in \mathcal{A}_+(\mathbf{x}^*)$ such that $x_{kl} > 0$ for all $k \geq k_+$. By U4, each time the UA is restarted, we perform a Wolfe line search. By the second half of (3.1), we have

$$\phi'(\alpha_k) - \phi'(0) \geq (\sigma - 1)\phi'(0). \tag{5.5}$$

It follows from the definition (3.2) of $\phi(\alpha)$ that

$$\phi'(0) = -\sum_{i\in\mathcal{I}(x_k)} g_{ki}^2 = -\|\mathbf{g}_I(\mathbf{x}_k)\|^2 \quad \text{and} \tag{5.6}$$

$$\begin{aligned}\phi'(\alpha_k) &= -\sum_{i\in\mathcal{I}(x_{k+1})} g_{ki}g_{k+1,i} \\ &= -\sum_{i\in\mathcal{I}(x_k)} g_{ki}g_{k+1,i} + \sum_{i\in\mathcal{A}(x_{k+1})\setminus\mathcal{A}(x_k)} g_{ki}g_{k+1,i}. \end{aligned} \tag{5.7}$$

By the Lipschitz continuity of ∇f and P3, we have

$$\begin{aligned}\|\mathbf{g}(\mathbf{x}_k) - \mathbf{g}(\mathbf{x}_{k+1})\| &= \|\mathbf{g}(P(\mathbf{x}_k)) - \mathbf{g}(P(\mathbf{x}_k - \alpha_k\mathbf{g}_I(\mathbf{x}_k)))\| \\ &\leq \lambda\alpha_k\|\mathbf{g}_I(\mathbf{x}_k)\|.\end{aligned}$$

Hence, by the Schwarz inequality,

$$\left|\sum_{i\in\mathcal{I}(\mathbf{x}_k)} g_{ki}(g_{ki} - g_{k+1,i})\right| \leq \lambda\alpha_k\|\mathbf{g}_I(\mathbf{x}_k)\|^2. \tag{5.8}$$

Since $\mathcal{A}(\mathbf{x}_{k+1}) \setminus \mathcal{A}(\mathbf{x}_k) \subset \mathcal{I}(\mathbf{x}_k)$, the Schwarz inequality also gives

$$\sum_{i\in\mathcal{A}(x_{k+1})\setminus\mathcal{A}(x_k)} g_{ki}g_{k+1,i} \leq \|\mathbf{g}_I(\mathbf{x}_k)\|\|\mathbf{g}_{k+1}\|_{\mathcal{N}}, \tag{5.9}$$

where

$$\|\mathbf{g}_{k+1}\|_{\mathcal{N}}^2 = \sum_{i\in\mathcal{A}(x_{k+1})\setminus\mathcal{A}(x_k)} g_{k+1,i}^2.$$

Here $\mathcal{N} = \mathcal{A}(\mathbf{x}_{k+1}) \setminus \mathcal{A}(\mathbf{x}_k)$ corresponds to the set of constraints that are newly activated as we move from $\mathbf{x}_k$ to $\mathbf{x}_{k+1}$. Combining (5.5)–(5.9),

$$\alpha_k \geq \frac{1-\sigma}{\lambda} - \frac{\|\mathbf{g}_{k+1}\|_{\mathcal{N}}}{\lambda \|\mathbf{g}_I(\mathbf{x}_k)\|}, \quad \text{where } \|\mathbf{g}_{k+1}\|_{\mathcal{N}}^2 = \sum_{i \in \mathcal{A}(x_{k+1}) \setminus \mathcal{A}(x_k)} g_{k+1,i}^2. \tag{5.10}$$

For k sufficiently large, (5.4) implies that the newly activated constraints $\mathcal{A}(\mathbf{x}_{k+1}) \setminus \mathcal{A}(\mathbf{x}_k)$ exclude all members of $\mathcal{A}_+(\mathbf{x}^*)$. Since the $\mathbf{x}_k$ converge to $\mathbf{x}^*$, $\|\mathbf{g}_{k+1}\|_{\mathcal{N}}$ tends to zero. On the other hand, $\|\mathbf{g}_I(\mathbf{x}_k)\|$ is bounded away from zero since the index l is contained in $\mathcal{I}(\mathbf{x}_k)$. Hence, the last term in (5.10) tends to 0 as k increases, and the lower bound for α_k approaches $(1-\sigma)/\lambda$. Since $x_l^* = 0$, it follows that x_{kl} approaches 0. Since the lower bound for α_k approaches $(1-\sigma)/\lambda$, $g_l(\mathbf{x}^*) > 0$, and $\mathbf{x}_k$ converges to $\mathbf{x}^*$, we conclude that

$$x_{k+1,l} = x_{kl} - \alpha_k g_{kl} < 0$$

for k sufficiently large. This contradicts the initial assumption that constraint l is inactive for k sufficiently large. Hence, in a finite number of iterations, $x_{ki} = 0$ for all $i \in \mathcal{A}_+(\mathbf{x}^*)$.

Case 3. The UA is executed a finite number of iterations.

In this case, the iterates are generated by the NGPA for k sufficiently large. Suppose that (5.3) is violated for some $l \in \mathcal{A}_+(\mathbf{x}^*)$. We show that this leads to a contradiction. By (5.4), $x_{kl} > 0$ for all $k \geq k_+$. Since $\mathbf{x}_k$ converges to $\mathbf{x}^*$, $\mathbf{x}_l^* = 0$, and $g_l(\mathbf{x}^*) > 0$, it is possible to choose k larger, if necessary, so that

$$x_{kl} - g_{kl}\alpha_{\min} < 0. \tag{5.11}$$

Since (5.3) is violated and $\mathbf{x}_k$ converges to $\mathbf{x}^*$, we can choose k larger, if necessary, so that

$$\frac{x_{kl}}{\|\mathbf{x}_k - \mathbf{x}^*\|^2} \geq \frac{\lambda(2+\lambda)^2 \max\{1, \alpha_{\max}\}^2}{2(1-\delta)g_{kl}}, \tag{5.12}$$

where $0 < \delta < 1$ is the parameter appearing in step 3 of the NGPA, and λ is the Lipschitz constant for ∇f. We will show that for this k, we have

$$f(\mathbf{x}_k + \mathbf{d}_k) \leq f_R + \delta \mathbf{g}_k^{\mathsf{T}} \mathbf{d}_k, \tag{5.13}$$

where f_R is specified in step 3 of the NGPA. According to step 3 of the NGPA, when (5.13) holds, $\alpha_k = 1$, which implies that

$$x_{k+1,l} = x_{kl} + d_{kl}. \tag{5.14}$$

Since (5.11) holds and $\bar{\alpha}_k \geq \alpha_{\min}$, we have

$$d_{kl} = \max\{x_{kl} - \bar{\alpha}_k g_{kl}, 0\} - x_{kl} = -x_{kl}. \tag{5.15}$$

This substitution in (5.14) gives $x_{k+1,l} = 0$, which contradicts the fact that $x_{kl} > 0$ for all $k \geq k_+$.

To complete the proof, we need to show that when (5.12) holds, (5.13) is satisfied. Expanding in a Taylor series around $\mathbf{x}_k$ and utilizing (5.15) gives

$$\begin{aligned} f(\mathbf{x}_k + \mathbf{d}_k) &= f(\mathbf{x}_k) + \int_0^1 f'(\mathbf{x}_k + t\mathbf{d}_k)dt \\ &= f(\mathbf{x}_k) + \mathbf{g}_k^{\mathsf{T}}\mathbf{d}_k + \int_0^1 (\nabla f(\mathbf{x}_k + t\mathbf{d}_k) - \mathbf{g}_k^{\mathsf{T}})\mathbf{d}_k dt \\ &\le f(\mathbf{x}_k) + \mathbf{g}_k^{\mathsf{T}}\mathbf{d}_k + \frac{\lambda}{2}\|\mathbf{d}_k\|^2 \\ &= f(\mathbf{x}_k) + \delta\mathbf{g}_k^{\mathsf{T}}\mathbf{d}_k + (1-\delta)\mathbf{g}_k^{\mathsf{T}}\mathbf{d}_k + \frac{\lambda}{2}\|\mathbf{d}_k\|^2 \\ &\le f(\mathbf{x}_k) + \delta\mathbf{g}_k^{\mathsf{T}}\mathbf{d}_k + (1-\delta)g_{kl}d_{kl} + \frac{\lambda}{2}\|\mathbf{d}_k\|^2 \qquad \text{(5.16a)} \\ &= f(\mathbf{x}_k) + \delta\mathbf{g}_k^{\mathsf{T}}\mathbf{d}_k - (1-\delta)g_{kl}x_{kl} + \frac{\lambda}{2}\|\mathbf{d}_k\|^2. \qquad \text{(5.16b)} \end{aligned}$$

The inequality (5.16a) is due to the fact that $g_{ki}d_{ki} \le 0$ for each i. By P3, P4, P5, and P7, and by the Lipschitz continuity of ∇f, we have

$$\begin{aligned} \|\mathbf{d}_k\| &\le \max\{1, \alpha_{\max}\}\|\mathbf{d}^1(\mathbf{x}_k)\| \\ &= \max\{1, \alpha_{\max}\}\|\mathbf{d}^1(\mathbf{x}_k) - \mathbf{d}^1(\mathbf{x}^*)\| \\ &= \max\{1, \alpha_{\max}\}\|P(\mathbf{x}_k - \mathbf{g}_k) - \mathbf{x}_k - P(\mathbf{x}^* - \mathbf{g}(\mathbf{x}^*)) + \mathbf{x}^*\| \\ &\le \max\{1, \alpha_{\max}\}(\|\mathbf{x}_k - \mathbf{x}^*\| + \|P(\mathbf{x}_k - \mathbf{g}_k) - P(\mathbf{x}^* - \mathbf{g}(\mathbf{x}^*))\|) \\ &\le \max\{1, \alpha_{\max}\}(\|\mathbf{x}_k - \mathbf{x}^*\| + \|\mathbf{x}_k - \mathbf{g}_k - (\mathbf{x}^* - \mathbf{g}(\mathbf{x}^*))\|) \\ &\le \max\{1, \alpha_{\max}\}(2\|\mathbf{x}_k - \mathbf{x}^*\| + \|\mathbf{g}_k - \mathbf{g}(\mathbf{x}^*)\|) \\ &\le \max\{1, \alpha_{\max}\}(2 + \lambda)\|\mathbf{x}_k - \mathbf{x}^*\|. \end{aligned}$$

Combining this upper bound for $\|\mathbf{d}_k\|$ with the lower bound (5.12) for x_{kl}, we conclude that

$$\begin{aligned} \frac{\lambda}{2}\|\mathbf{d}_k\|^2 &\le \frac{\lambda}{2}\max\{1, \alpha_{\max}\}^2(2+\lambda)^2\|\mathbf{x}_k - \mathbf{x}^*\|^2 \\ &\le \frac{1}{2}\left(\frac{2(1-\delta)x_{kl}g_{kl}}{\|\mathbf{x}_k - \mathbf{x}^*\|^2}\right)\|\mathbf{x}_k - \mathbf{x}^*\|^2 \\ &= (1-\delta)x_{kl}g_{kl}. \end{aligned}$$

Hence, by (5.16b) and by the choice for f_R specified in step 3 of the NGPA, we have

$$f(\mathbf{x}_k + \mathbf{d}_k) \le f(\mathbf{x}_k) + \delta\mathbf{g}_k^{\mathsf{T}}\mathbf{d}_k \le f_R + \delta\mathbf{g}_k^{\mathsf{T}}\mathbf{d}_k. \qquad \text{(5.17)}$$

This completes the proof of (5.13). □

There is a fundamental difference between the gradient projection algorithm presented in this paper and algorithms based on a "piecewise projected gradient" [15, 16, 17]. For our gradient projection algorithm, we perform a single projection, and then we backtrack towards the starting point. Thus we are unable to show that the active constraints are identified in a finite number of iterations; in contrast, with the piecewise project gradient approach, where a series of projections may be performed, the active constraints can be identified in a finite number of iterations. In Lemma 5.2 we show that even though we do not identify the active constraints, the

components of $\mathbf{x}_k$ corresponding to the strictly active constraints are on the order of the error in $\mathbf{x}_k$ squared.

If all the constraints are active at a stationary point $\mathbf{x}^*$ and strict complementarity holds, then convergence is achieved in a finite number of iterations.

COROLLARY 5.3. *If f is twice-continuously differentiable, the iterates $\mathbf{x}_k$ generated by the ASA with $\epsilon = 0$ converge to a stationary point $\mathbf{x}^*$, and $|\mathcal{A}_+(\mathbf{x}^*)| = n$, then $\mathbf{x}_k = \mathbf{x}^*$ after a finite number of iterations.*

Proof. Let $\mathbf{x}_{k,\max}$ denote the largest component of $\mathbf{x}_k$. Since $\|\mathbf{x}_k\|^2 \le n\mathbf{x}_{k,\max}^2$, we have

$$\frac{\mathbf{x}_{k,\max}}{\|\mathbf{x}_k\|^2} \ge \frac{1}{n\mathbf{x}_{k,\max}}. \tag{5.18}$$

Since all the constraints are active at $\mathbf{x}^*$, $\mathbf{x}_{k,\max}$ tends to zero. By (5.18) the conclusion (5.3) of Lemma 5.2 does not hold. Hence, after a finite number of iterations, $\mathbf{x}_k = \mathbf{x}^*$. ∎

Recall [70] that for any stationary point $\mathbf{x}^*$ of (1.1), the strong second-order sufficient optimality condition holds if there exists $\gamma > 0$ such that

$$\mathbf{d}^{\mathsf{T}}\nabla^2 f(\mathbf{x}^*)\mathbf{d} \ge \gamma\|\mathbf{d}\|^2 \quad \text{whenever} \quad d_i = 0 \text{ for all } i \in \mathcal{A}_+(\mathbf{x}^*). \tag{5.19}$$

Using P8, we establish the following.

LEMMA 5.4. *If f is twice-continuously differentiable near a stationary point $\mathbf{x}^*$ of (1.1) satisfying the strong second-order sufficient optimality condition, then there exists $\rho > 0$ with the following property:*

$$\|\mathbf{x} - \mathbf{x}^*\| \le \sqrt{1 + \left(\frac{(1+\lambda)^2}{.5\gamma}\right)^2}\,\|\mathbf{d}^1(\mathbf{x})\| \tag{5.20}$$

for all $\mathbf{x} \in B_\rho(\mathbf{x}^*)$, where λ is any Lipschitz constant for ∇f over $B_\rho(\mathbf{x}^*)$.

Proof. By the continuity of the second derivative of f, it follows from (5.19) that for $\rho > 0$ sufficiently small,

$$(\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{x}^*))^{\mathsf{T}}(\mathbf{x} - \mathbf{x}^*) \ge .5\gamma\|\mathbf{x} - \mathbf{x}^*\|^2 \tag{5.21}$$

for all $\mathbf{x} \in B_\rho(\mathbf{x}^*)$ with $x_i = 0$ for all $i \in \mathcal{A}_+(\mathbf{x}^*)$. Choose ρ smaller if necessary so that

$$x_i - g_i(\mathbf{x}) \le 0 \text{ for all } i \in \mathcal{A}_+(\mathbf{x}^*) \text{ and } \mathbf{x} \in B_\rho(\mathbf{x}^*). \tag{5.22}$$

Let $\bar{\mathbf{x}}$ be defined as follows:

$$\bar{x}_i = \begin{cases} 0 \text{ if } i \in \mathcal{A}_+(\mathbf{x}^*), \\ x_i \text{ otherwise.} \end{cases} \tag{5.23}$$

Since (5.22) holds, it follows that

$$\|\mathbf{x} - \bar{\mathbf{x}}\| \le \|\mathbf{d}^1(\mathbf{x})\| \tag{5.24}$$

for all $\mathbf{x} \in B_\rho(\mathbf{x}^*)$. Also, by (5.22), we have

$$[P(\bar{\mathbf{x}} - \mathbf{g}(\mathbf{x})) - \bar{\mathbf{x}}]_i = 0 \quad \text{and} \quad \mathbf{d}^1(\mathbf{x})_i = [P(\mathbf{x} - \mathbf{g}(\mathbf{x})) - \mathbf{x}]_i = -x_i$$

for all $i \in \mathcal{A}_+(\mathbf{x}^*)$, while

$$[P(\bar{\mathbf{x}} - \mathbf{g}(\mathbf{x})) - \bar{\mathbf{x}}]_i = \mathbf{d}^1(\mathbf{x})_i = [P(\mathbf{x} - \mathbf{g}(\mathbf{x})) - \mathbf{x}]_i$$

for $i \notin \mathcal{A}_+(\mathbf{x}^*)$. Hence, we have

$$\|P(\bar{\mathbf{x}} - \mathbf{g}(\mathbf{x})) - \bar{\mathbf{x}}\| \leq \|\mathbf{d}^1(\mathbf{x})\| \tag{5.25}$$

for all $\mathbf{x} \in B_\rho(\mathbf{x}^*)$. By the Lipschitz continuity of $\mathbf{g}$, (5.24), (5.25), and P3, it follows that

$$\begin{aligned} \|\mathbf{d}^1(\bar{\mathbf{x}})\| &= \|P(\bar{\mathbf{x}} - \mathbf{g}(\bar{\mathbf{x}})) - P(\bar{\mathbf{x}} - \mathbf{g}(\mathbf{x})) + P(\bar{\mathbf{x}} - \mathbf{g}(\mathbf{x})) - \bar{\mathbf{x}}\| \\ &\leq \lambda\|\bar{\mathbf{x}} - \mathbf{x}\| + \|\mathbf{d}^1(\mathbf{x})\| \\ &\leq (1+\lambda)\|\mathbf{d}^1(\mathbf{x})\| \end{aligned} \tag{5.26}$$

for all $\mathbf{x} \in B_\rho(\mathbf{x}^*)$. By P8, (5.21), and (5.26), we have

$$\|\bar{\mathbf{x}} - \mathbf{x}^*\| \leq \left(\frac{1+\lambda}{.5\gamma}\right)\|\mathbf{d}^1(\bar{\mathbf{x}})\| \leq \left(\frac{(1+\lambda)^2}{.5\gamma}\right)\|\mathbf{d}^1(\mathbf{x})\|. \tag{5.27}$$

Since $\|\mathbf{x} - \bar{\mathbf{x}}\|^2 + \|\bar{\mathbf{x}} - \mathbf{x}^*\|^2 = \|\mathbf{x} - \mathbf{x}^*\|^2$, the proof is completed by squaring and adding (5.27) and (5.24). □

We now show that the undecided index set $\mathcal{U}$ becomes empty as the iterates approach a stationary point, where the strong second-order sufficient optimality condition holds.

LEMMA 5.5. *Suppose f is twice-continuously differentiable, $\mathbf{x}^*$ is a stationary point of* (1.1) *satisfying the strong second-order sufficient optimality condition, and $\mathbf{x}_k$, $k = 0, 1, \ldots$, is an infinite sequence of feasible iterates for* (1.1) *converging to $\mathbf{x}^*$, $\mathbf{x}_k \neq \mathbf{x}^*$ for each k. If there exists a constant ξ such that*

$$\limsup_{k\to\infty} \frac{x_{ki}}{\|\mathbf{x}_k - \mathbf{x}^*\|^2} \leq \xi < \infty \tag{5.28}$$

for all $i \in \mathcal{A}_+(\mathbf{x}^*)$, then $\mathcal{U}(\mathbf{x}_k)$ is empty for k sufficiently large.

Proof. To prove that $\mathcal{U}(\mathbf{x})$ is empty, we must show that for each $i \in [1, n]$, one of the following inequalities is violated:

$$|g_i(\mathbf{x})| \geq \|\mathbf{d}^1(\mathbf{x})\|^\alpha \text{ or} \tag{5.29}$$

$$x_i \geq \|\mathbf{d}^1(\mathbf{x})\|^\beta. \tag{5.30}$$

By Lemma 5.4, there exists a constant c such that $\|\mathbf{x} - \mathbf{x}^*\| \leq c\|\mathbf{d}^1(\mathbf{x})\|$ for all $\mathbf{x}$ near $\mathbf{x}^*$. If $i \in \mathcal{A}_+(\mathbf{x}^*)$, then by (5.28), we have

$$\limsup_{k\to\infty} \frac{x_{ki}}{\|\mathbf{d}^1(\mathbf{x}_k)\|^\beta} \leq \limsup_{k\to\infty} \frac{\xi\|\mathbf{x}_k - \mathbf{x}^*\|^2}{\|\mathbf{d}^1(\mathbf{x}_k)\|^\beta} \leq \limsup_{k\to\infty} \xi c^2 \|\mathbf{d}^1(\mathbf{x}_k)\|^{2-\beta} = 0$$

since $\beta \in (1, 2)$. Hence, for each $i \in \mathcal{A}_+(\mathbf{x}^*)$, (5.30) is violated for k sufficiently large.

If $i \notin \mathcal{A}_+(\mathbf{x}^*)$, then $g_i(\mathbf{x}^*) = 0$. By Lemma 5.4, we have

$$\begin{aligned} \limsup_{k\to\infty} \frac{|g_i(\mathbf{x}_k)|}{\|\mathbf{d}^1(\mathbf{x}_k)\|^\alpha} &= \limsup_{k\to\infty} \frac{|g_i(\mathbf{x}_k) - g_i(\mathbf{x}^*)|}{\|\mathbf{d}^1(\mathbf{x}_k)\|^\alpha} \\ &\leq \limsup_{k\to\infty} \frac{\lambda\|\mathbf{x}_k - \mathbf{x}^*\|}{\|\mathbf{d}^1(\mathbf{x}_k)\|^\alpha} \\ &\leq \limsup_{k\to\infty} \lambda c \|\mathbf{d}^1(\mathbf{x}_k)\|^{1-\alpha} = 0, \end{aligned}$$

since $\alpha \in (0,1)$. Here, λ is a Lipschitz constant for $\mathbf{g}$ in a neighborhood of $\mathbf{x}^*$. Hence, (5.29) is violated if $i \notin \mathcal{A}_+(\mathbf{x}^*)$. ∎

Remark. If $i \in \mathcal{A}_+(\mathbf{x}^*)$ and the iterates $\mathbf{x}_k$ converge to a stationary point $\mathbf{x}^*$, then $g_i(\mathbf{x}_k)$ is bounded away from 0 for k sufficiently large. Since $\mathbf{d}^1(\mathbf{x}_k)$ tends to zero, the inequality $|g_i(\mathbf{x}_k)| \geq \|\mathbf{d}^1(\mathbf{x}_k)\|^\alpha$ is satisfied for k sufficiently large. Hence, if $\mathcal{U}(\mathbf{x}_k)$ is empty and $i \in \mathcal{A}_+(\mathbf{x}^*)$, then $x_{ki} < \|\mathbf{d}^1(\mathbf{x}_k)\|^\beta$ where $\beta \in (1,2)$. In other words, when $\mathcal{U}(\mathbf{x}_k)$ is empty, the components of $\mathbf{x}_k$ associated with strictly active indices $\mathcal{A}_+(\mathbf{x}^*)$ are going to zero faster than the error $\|\mathbf{d}^1(\mathbf{x}_k)\|$.

LEMMA 5.6. *Suppose f is twice-continuously differentiable, $\mathbf{x}^*$ is a stationary point of* (1.1) *satisfying the strong second-order sufficient optimality condition, and $\mathbf{x}_k$, $k = 0, 1, \ldots$, is an infinite sequence of feasible iterates for* (1.1) *converging to $\mathbf{x}^*$, $\mathbf{x}_k \neq \mathbf{x}^*$ for each k. If there exists a constant ξ such that*

$$\limsup_{k\to\infty} \frac{x_{ki}}{\|\mathbf{x}_k - \mathbf{x}^*\|^2} \leq \xi < \infty \tag{5.31}$$

for all $i \in \mathcal{A}_+(\mathbf{x}^*)$, then there exist $\mu^* > 0$ such that

$$\|\mathbf{g}_I(\mathbf{x}_k)\| \geq \mu^* \|\mathbf{d}^1(\mathbf{x}_k)\| \tag{5.32}$$

for k sufficiently large.

Proof. Choose $\rho > 0$, and let λ be the Lipschitz constant for ∇f in $B_\rho(\mathbf{x}^*)$. As in (5.23), let $\bar{\mathbf{x}}$ be defined by $\bar{x}_i = 0$ if $i \in \mathcal{A}_+(\mathbf{x}^*)$ and $\bar{x}_i = x_i$ otherwise. If $\mathbf{x}_k \in B_\rho(\mathbf{x}^*)$, we have

$$\begin{aligned} \|\mathbf{d}^1(\mathbf{x}_k)\| &\leq \|\mathbf{d}^1(\mathbf{x}_k) - \mathbf{d}^1(\mathbf{x}^*)\| \\ &\leq \|\mathbf{d}^1(\mathbf{x}_k) - \mathbf{d}^1(\bar{\mathbf{x}}_k)\| + \|\mathbf{d}^1(\bar{\mathbf{x}}_k) - \mathbf{d}^1(\mathbf{x}^*)\| \\ &\leq (2+\lambda)(\|\mathbf{x}_k - \bar{\mathbf{x}}_k\| + \|\bar{\mathbf{x}}_k - \mathbf{x}^*\|). \end{aligned} \tag{5.33}$$

Utilizing (5.31) gives

$$\begin{aligned} \|\bar{\mathbf{x}}_k - \mathbf{x}_k\| &\leq \sum_{i=1}^n |\bar{x}_{ki} - x_{ki}| \\ &= \sum_{i \in \mathcal{A}_+(\mathbf{x}^*)} x_{ki} \leq n\xi \|\mathbf{x}_k - \mathbf{x}^*\|^2 \\ &\leq n\xi \|\mathbf{x}_k - \mathbf{x}^*\|(\|\mathbf{x}_k - \bar{\mathbf{x}}_k\| + \|\bar{\mathbf{x}}_k - \mathbf{x}^*\|). \end{aligned}$$

Since $\mathbf{x}_k$ converges to $\mathbf{x}^*$, it follows that for any $\epsilon > 0$,

$$\|\bar{\mathbf{x}}_k - \mathbf{x}_k\| \leq \epsilon \|\bar{\mathbf{x}}_k - \mathbf{x}^*\| \tag{5.34}$$

when k is sufficiently large. Combining (5.33) and (5.34), there exists a constant $c > 0$ such that

$$\|\mathbf{d}^1(\mathbf{x}_k)\| \leq c \|\bar{\mathbf{x}}_k - \mathbf{x}^*\| \tag{5.35}$$

for k sufficiently large.

Let k be chosen large enough that

$$\|\mathbf{x}_k - \mathbf{x}^*\| < \min\{x_i^* : i \in \mathcal{I}(\mathbf{x}^*)\}. \tag{5.36}$$

Suppose, in this case, that $i \in \mathcal{A}(\mathbf{x}_k)$. If $x_i^* > 0$, then $\|\mathbf{x}_k - \mathbf{x}^*\| \geq x_i^*$, which contradicts (5.36). Hence, $\bar{x}_{ki} = x_i^* = 0$. Moreover, if $i \in \mathcal{A}_+(\mathbf{x}^*)$, then by the definition (5.23), $\bar{x}_{ki} = x_i^* = 0$. In summary,

$$\begin{cases} \bar{x}_{ki} = x_i^* = 0 & \text{for each } i \in \mathcal{A}(\mathbf{x}_k) \cup \mathcal{A}_+(\mathbf{x}^*), \\ g_i(\mathbf{x}^*) = 0 & \text{for each } i \in \mathcal{A}_+(\mathbf{x}^*)^c, \end{cases} \tag{5.37}$$

where $\mathcal{A}_+(\mathbf{x}^*)^c$ is the complement of $\mathcal{A}_+(\mathbf{x}^*)$. Define $\mathcal{Z} = \mathcal{A}(\mathbf{x}_k)^c \cap \mathbf{A}_+(\mathbf{x}^*)^c$.

By the strong second-order sufficient optimality condition and for $\mathbf{x}$ near $\mathbf{x}^*$, we have

$$\begin{aligned} \frac{\gamma}{2}\|\bar{\mathbf{x}} - \mathbf{x}^*\|^2 &\leq [\bar{\mathbf{x}} - \mathbf{x}^*]^\mathsf{T} \int_0^1 \nabla^2 f(\mathbf{x}^* + t(\bar{\mathbf{x}} - \mathbf{x}^*))dt \;\; [\bar{\mathbf{x}} - \mathbf{x}^*] \\ &= (\bar{\mathbf{x}} - \mathbf{x}^*)^\mathsf{T}(\mathbf{g}(\bar{\mathbf{x}}) - \mathbf{g}(\mathbf{x}^*)). \end{aligned} \tag{5.38}$$

We substitute $\mathbf{x} = \mathbf{x}_k$ in (5.38) and utilize (5.37) to obtain

$$\begin{aligned} (\bar{\mathbf{x}}_k - \mathbf{x}^*)^\mathsf{T}(\mathbf{g}(\bar{\mathbf{x}}_k) - \mathbf{g}(\mathbf{x}^*)) &= \sum_{i=1}^n (\bar{x}_{ki} - x_i^*)(g_i(\bar{\mathbf{x}}_k) - g_i(\mathbf{x}^*)) \\ &= \sum_{i \in \mathcal{Z}} (\bar{x}_{ki} - x_i^*) g_i(\bar{\mathbf{x}}_k) \\ &\leq \|\bar{\mathbf{x}}_k - \mathbf{x}^*\| \left(\sum_{i \in \mathcal{I}(\mathbf{x}_k)} g_i(\bar{\mathbf{x}}_k)^2 \right)^{1/2}, \end{aligned} \tag{5.39}$$

since $\mathcal{Z} \subset \mathcal{A}(\mathbf{x}_k)^c = \mathcal{I}(\mathbf{x}_k)$. Exploiting the Lipschitz continuity of ∇f, (5.39) gives

$$(\bar{\mathbf{x}}_k - \mathbf{x}^*)^\mathsf{T}(\mathbf{g}(\bar{\mathbf{x}}_k) - \mathbf{g}(\mathbf{x}^*)) \leq \|\bar{\mathbf{x}}_k - \mathbf{x}^*\|(\|\mathbf{g}_I(\mathbf{x}_k)\| + \lambda\|\bar{\mathbf{x}}_k - \mathbf{x}_k\|). \tag{5.40}$$

Combining (5.34), (5.38), and (5.40), we conclude that for k sufficiently large,

$$\frac{\gamma}{4}\|\bar{\mathbf{x}}_k - \mathbf{x}^*\| \leq \|\mathbf{g}_I(\mathbf{x}_k)\|. \tag{5.41}$$

Combining (5.35) and (5.41), the proof is complete. ∎

Remark. If $\mathbf{x}_k$ is a sequence converging to a nondegenerate stationary point $\mathbf{x}^*$, then (5.32) holds with $\mu^* = 1$, without assuming either the strong second-order sufficient optimality condition or (5.31)—see Theorem 5.1. In Lemma 5.6, the optimization problem could be degenerate.

We now show that after a finite number of iterations, the ASA will perform only the UA with a fixed active constraint set.

THEOREM 5.7. *If f is twice-continuously differentiable and the iterates $\mathbf{x}_k$ generated by the ASA with $\epsilon = 0$ converge to a stationary point $\mathbf{x}^*$ satisfying the strong second-order sufficient optimality condition, then after a finite number of iterations, the ASA performs only the UA without restarts.*

Proof. By Lemma 5.2, the hypotheses (5.28) and (5.31) of Lemmas 5.5 and 5.6 are satisfied. Hence, for k sufficiently large, the undecided set $\mathcal{U}(\mathbf{x}_k)$ is empty and the lower bound (5.32) holds. In step 1a, if $\|\mathbf{g}_I(\mathbf{x}_k)\| < \mu\|\mathbf{d}^1(\mathbf{x}_k)\|$, then μ is multiplied by the factor $\rho < 1$. When $\mu < \mu^*$, Lemma 5.6 implies that $\|\mathbf{g}_I(\mathbf{x}_k)\| \geq \mu\|\mathbf{d}^1(\mathbf{x}_k)\|$. Hence, step 1a of the ASA branches to step 2, while step 2 cannot branch to step 1 since the condition $\|\mathbf{g}_I(\mathbf{x}_k)\| < \mu\|\mathbf{d}^1(\mathbf{x}_k)\|$ is never satisfied in step 2a and $\mathcal{U}(\mathbf{x}_k)$

is empty in step 2b for k sufficiently large. Since the UA only adds constraints, we conclude that after a finite number of iterations, the active set does not change. ∎

Remark. If f is a strongly convex quadratic function, then by Corollary 4.2, the iterates $\mathbf{x}_k$ converge to the global minimizer $\mathbf{x}^*$. If the UA is based on the conjugate gradient method for which there is finite convergence when applied to a convex quadratic, it follows from Theorem 5.7 that the ASA converges in a finite number of iterations.

We now give the proof of Corollary 4.2; that is, when f is strongly convex and twice-continuously differentiable on $\mathcal{B}$, and assumption A3 of Theorem 4.1 is satisfied, then the entire sequence of iterates generated by the ASA converges to the global minimizer $\mathbf{x}^*$. Note that the assumptions of Corollary 4.2 are weaker than those of Corollary 2.3 (global convergence of the NGPA) since Corollary 4.2 requires only that $f_k^r \le f_k^{\max}$ infinitely often in the NGPA.

Proof. For a strongly convex function, A1 and A2 always hold. Since all the assumptions of Theorem 4.1 are satisfied, there exists a subsequence $\mathbf{x}_{k_j}$, $j = 1, 2, \ldots$, of the iterates such that

$$\lim_{j\to\infty} \|\mathbf{d}^1(\mathbf{x}_{k_j})\| = 0.$$

Since the UA is monotone and since the NGPA satisfies (2.12) and (2.13), it follows from the strong convexity of f that the $\mathbf{x}_{k_j}$ are contained in a bounded set. Since $\mathbf{d}^1(\cdot)$ is continuous, there exists a subsequence, also denoted $\mathbf{x}_{k_j}$, converging to a limit $\mathbf{x}^*$ with $\mathbf{d}^1(\mathbf{x}^*) = \mathbf{0}$. Since the unique stationary point of a strongly convex function is its global minimizer, $\mathbf{x}^*$ is the global solution of (1.1).

Case A. There exists an infinite subsequence, also denoted $\{\mathbf{x}_{k_j}\}$, with the property that $\mathbf{x}_{k_j+1}$ is generated by the UA.

In this case, we are done since the UA is monotone and the inequality

$$f(\mathbf{x}_k) \le f(\mathbf{x}_{k_j}) \tag{5.42}$$

holds for all $k \ge k_j$ (see (2.12) and (2.13)). Since $\mathbf{x}_{k_j}$ converges to $\mathbf{x}^*$, it follows that $f(\mathbf{x}_{k_j})$ converges to $f(\mathbf{x}^*)$, and hence, by (5.42) and (2.32), the entire sequence converges to $\mathbf{x}^*$.

Case B. There exists an infinite subsequence, also denoted $\{\mathbf{x}_{k_j}\}$, with the property that $\mathbf{x}_{k_j+1}$ is generated by the NGPA.

Either

$$\limsup_{j\to\infty} \frac{(\mathbf{x}_{k_j})_i}{\|\mathbf{x}_{k_j} - \mathbf{x}^*\|^2} < \infty \quad \text{for all } i \in \mathcal{A}_+(\mathbf{x}^*) \tag{5.43}$$

holds or (5.43) is violated. By the analysis given in Case 3 of the proof of Lemma 5.2, when (5.43) is violated, (5.13) holds, from which it follows that for j sufficiently large,

$$\mathbf{x}_{k_j+1,i} = 0 \quad \text{for all } i \in \mathcal{A}_+(\mathbf{x}^*). \tag{5.44}$$

Hence, either the sequence $\mathbf{x}_{k_j}$ satisfies (5.43) or the sequence $\mathbf{x}_{k_j+1}$ satisfies (5.44). In this latter case, it follows from (5.17) that

$$f(\mathbf{x}_{k_j+1}) \le f(\mathbf{x}_{k_j}).$$

Since $f(\mathbf{x}_{k_j})$ converges to $f(\mathbf{x}^*)$, we conclude that $f(\mathbf{x}_{k_j+1})$ converges to $f(\mathbf{x}^*)$, and $\mathbf{x}_{k_j+1}$ converges to $\mathbf{x}^*$.

In either case (5.43) or (5.44), there exists a sequence K_j (either $K_j = k_j$ or $K_j = k_j + 1$) with the property that $\mathbf{x}_{K_j}$ converges to $\mathbf{x}^*$ and

$$\limsup_{j\to\infty} \frac{(\mathbf{x}_{K_j})_i}{\|\mathbf{x}_{K_j} - \mathbf{x}^*\|^2} < \infty \quad \text{for all } i \in \mathcal{A}_+(\mathbf{x}^*).$$

By Lemma 5.5, $\mathcal{U}(\mathbf{x}_{K_j})$ is empty for j sufficiently large. By Lemma 5.6, there exists $\mu^* > 0$ such that

$$\|\mathbf{g}_I(\mathbf{x}_{K_j})\| \geq \mu^* \|\mathbf{d}^1(\mathbf{x}_{K_j})\|$$

for j sufficiently large. As in the proof of Theorem 5.7, at iteration K_j for j sufficiently large, the ASA jumps from step 1 to the UA in step 2. Hence, for j sufficiently large, $\mathbf{x}_{K_j+1}$ is generated by the UA, which implies that Case A holds. □

6. Numerical experiments. This section compares the CPU time performance of the ASA, implemented using the nonlinear conjugate gradient code CG_DESCENT for the UA and the CBB method (see the appendix) for the NGPA, to the performance of the following codes:

- L-BFGS-B [18, 84]: The limited memory quasi-Newton method of Zhu, Byrd, and Nocedal (ACM algorithm 778).
- SPG2 version 2.1 [10, 11]: The nonmonotone spectral projected gradient method of Birgin, Martínez, and Raydan (ACM algorithm 813).
- GENCAN [9]: The monotone active set method with spectral projected gradients developed by Birgin and Martínez.
- TRON version 1.2 [63]: A Newton trust region method with incomplete Cholesky preconditioning developed by Lin and Moré.

A detailed description of our implementation of the ASA is given in the appendix.

L-BFGS-B was downloaded from Jorge Nocedal's Web page (http://www.ece.northwestern.edu/∼nocedal/lbfgsb.html); TRON was downloaded from Jorge Moré's Web page (http://www-unix.mcs.anl.gov/∼more/tron/); and SPG2 and GENCAN were downloaded on June 28, 2005, from the TANGO Web page maintained by Ernesto Birgin (http://www.ime.usp.br/∼egbirgin/tango/downloads.php). All codes are written in Fortran and compiled with f77 (default compiler settings) on a Sun workstation. The stopping condition was

$$\|P(\mathbf{x} - \mathbf{g}(\mathbf{x})) - \mathbf{x}\|_\infty \leq 10^{-6},$$

where $\|\cdot\|_\infty$ denotes the sup-norm of a vector. In running any of these codes, default values were used for all parameters. In the NGPA, we chose the following parameter values:

$$\alpha_{\min} = 10^{-20}, \quad \alpha_{\max} = 10^{+20}, \quad \eta = .5, \quad \delta = 10^{-4}, \quad M = 8.$$

Here M is the memory used to evaluate $f_k^{\max}$ (see (2.3)). In the ASA the parameter values were as follows:

$$\mu = .1, \quad \rho = .5, \quad n_1 = 2, \quad n_2 = 1.$$

In the CBB method (see the appendix), the parameter values were the following:

$$\theta = .975, \quad L = 3, \quad A = 40, \quad m = 4, \quad \gamma_1 = M/L, \quad \gamma_2 = A/M.$$

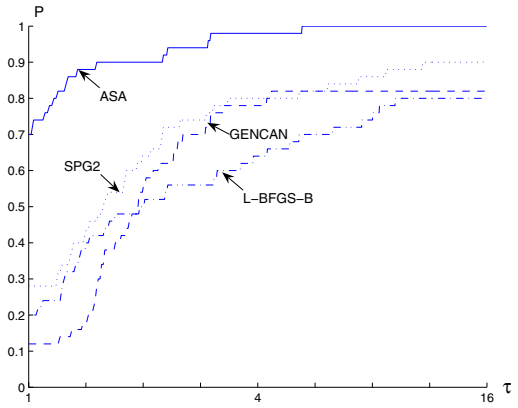


FIG. 6.1. *Performance profiles, CPU time metric, 50 CUTEr test problems, gradient-based methods.*

The separation parameter Δ in condition R4 of the appendix was the natural separation between floating point numbers. That is, R4 was satisfied when the floating point version of f_{k+1} was strictly less than the floating point version of $f_k^{\min}$.

The test set consisted of all 50 box constrained problems in the CUTEr library [13] with dimensions between 50 and 15,625, and all 23 box constrained problems in the MINPACK-2 library [1] with dimension 2500. TRON is somewhat different from the other codes since it employs Hessian information and an incomplete Cholesky preconditioner, while the codes ASA, L-BFGS-B, SPG2, and GENCAN utilize only gradient information. When we compare our code to TRON, we use the same Lin–Moré preconditioner [62] used by TRON for our unconstrained algorithm. The preconditioned ASA code is called P-ASA. Since TRON is targeted to large-sparse problems, we compare our code to TRON using the 23 MINPACK-2 problems and the 42 sparsest CUTEr problems (the number of nonzeros in the Hessian was at most 1/5 the total number of entries). The codes L-BFGS-B, SPG2, and GENCAN were implemented for the CUTEr test problems, while ASA and TRON were implemented for both test sets CUTEr and MINPACK-2.

The CPU time in seconds and the number of iterations, function evaluations, gradient evaluations, and Hessian evaluations for each of the methods are posted at the following Web site: http://www.math.ufl.edu/∼hager/papers/CG. In running the numerical experiments, we checked whether different codes converged to different local minimizers; when comparing the codes, we restricted ourselves to test problems in which all codes converged to the same local minimizer, and where the running time of the fastest code exceeded .01 seconds. The numerical results are now analyzed.

The performance of the algorithms, relative to CPU time, was evaluated using the performance profiles of Dolan and Moré [34]. That is, for each method, we plot the fraction P of problems for which the method is within a factor τ of the best time. In Figure 6.1, we compare the performance of the four codes ASA, L-BFGS-B, SPG2, and GENCAN using the 50 CUTEr test problems. The left side of the figure gives the percentage of the test problems for which a method is the fastest; the right side gives the percentage of the test problems that were successfully solved by each of the methods. The top curve is the method that solved the most problems in a time that was within a factor τ of the best time. Since the top curve in Figure 6.1 corresponds

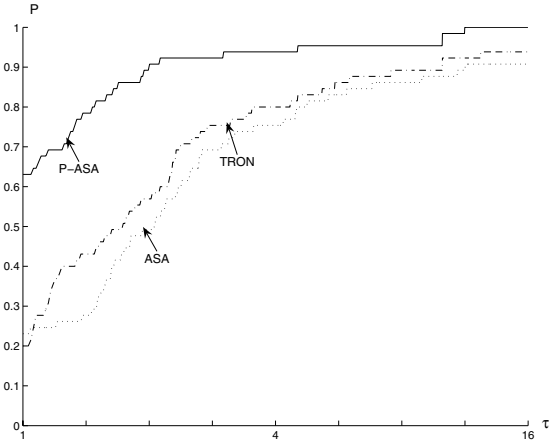


FIG. 6.2. *Performance profiles, CPU time metric,* 42 *sparsest CUTEr problems,* 23 *MINPACK-2 problems,* $\epsilon = 10^{-6}$.

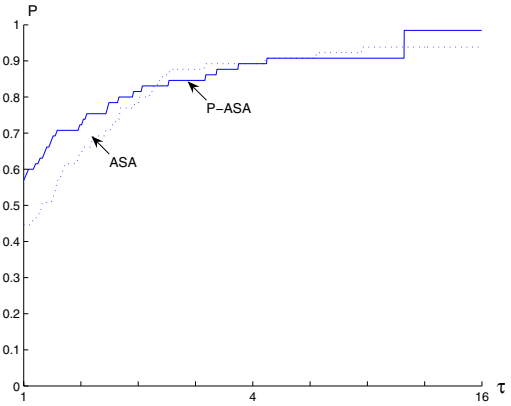


FIG. 6.3. *Performance profiles, CPU time metric,* $\epsilon = 10^{-2}\|\mathbf{d}^1(\mathbf{x}_0)\|_\infty$.

to the ASA, this algorithm is clearly fastest for this set of 50 test problems with dimensions ranging from 50 to 15,625. The relative difference in performance between the ASA and the competing methods seen in Figure 6.1 is greater than the relative difference in performance between CG_DESCENT and the competing methods, as seen in the figures given in [55, 57]. Hence, both the gradient projection algorithm and the conjugate gradient algorithm are contributing to the better performance of the ASA.

In Figure 6.2 we compare the performance of TRON to P-ASA and ASA for the 42 sparsest CUTEr test problems and the 23 MINPACK-2 problems. Observe that P-ASA has the top performance, and that ASA, which utilizes only the gradient, performs almost as well as the Hessian-based code TRON. The number of conjugate gradient iterations performed by the P-ASA code is much less than the number of conjugate gradient iterations performed by the ASA code. Finally, in Figure 6.3 we

compare the performance of P-ASA to ASA for the relaxed convergence tolerance $\epsilon = 10^{-2}\|\mathbf{d}^1(\mathbf{x}_0)\|_\infty$. Based on Figures 6.2 and 6.3, the preconditioned ASA scheme is more efficient than unconditioned ASA for the more stringent stopping criterion, while the unconditioned and preconditioned schemes are equally effective for a more relaxed stopping criterion. Although the performance profile for ASA is beneath 1 in Figure 6.2, it reaches 1 as τ increases—there are some problems in which P-ASA is more than 16 times faster than ASA. Due to these difficult problems, the ASA profile is still beneath 1 for $\tau = 16$.

When we solve an optimization problem, the solution time consists of two parts, as follows:

T1. The time associated with the evaluation of the function or its gradient or its Hessian.

T2. The remaining time, which is often dominated by the time used in the linear algebra.

The CPU time performance profile measures a mixture of T1 and T2 for a set of test problems. In some applications, T1 (the evaluation time) may dominate. In order to assess how the algorithms may perform in the limit, when T2 is negligible compared to T1, we could ignore T2 and compare the algorithms based on T1. In the next set of experiments, we explore how the algorithms perform in the limit, as T1 becomes infinitely large relative to T2.

Typically, the time to evaluate the gradient of a function is greater than the time to evaluate the function itself. Also, the time to evaluate the Hessian is greater than the time to evaluate the gradient. If the time to evaluate the function is 1, then the average time to evaluate the gradient and Hessian for the CUTEr bound constrained test set is as follows:

$$\text{function } = 1, \quad \text{gradient } = 2.6, \quad \text{Hessian } = 21.0.$$

Similarly, for the MINPACK-2 test set, the relative evaluation times are

$$\text{function } = 1, \quad \text{gradient } = 2.0, \quad \text{Hessian } = 40.5$$

on average.

For each method and for each test problem, we compute an "evaluation time" where the time for a function evaluation is 1, the time for a gradient evaluation is either 2.6 (CUTEr) or 2.0 (MINPACK-2), and the time for a Hessian evaluation is either 21.0 (CUTEr) or 40.5 (MINPACK-2). In Figure 6.4 we compare the performance of gradient-based methods, and in Figure 6.5 we compare the performance of the gradient-based ASA and the method which exploits the Hessian (P-ASA or TRON).

In Figure 6.4 we see that for the evaluation metric and τ near 1, L-BFGS-B performs better than ASA, but as τ increases, ASA dominates L-BFGS-B. In other words, in the evaluation metric, there are more problems in which L-BFGS-B is faster than the other methods; however, ASA is not much slower than L-BFGS-B. When τ reaches 1.5, ASA starts to dominate L-BFGS-B.

In Figure 6.5 we see that P-ASA dominates TRON in the evaluation metric. Hence, even though TRON uses far fewer function evaluations, it uses many more Hessian evaluations. Since the time to evaluate the Hessian is much greater than the time to evaluate the function, P-ASA has better performance. In summary, by neglecting the time associated with the linear algebra, the relative gap between P-ASA and TRON decreases, while the relative gap between TRON and ASA increases, as seen in Figure 6.5. Nonetheless, in the evaluation metric, the performance profile for P-ASA is still above the profile for TRON.

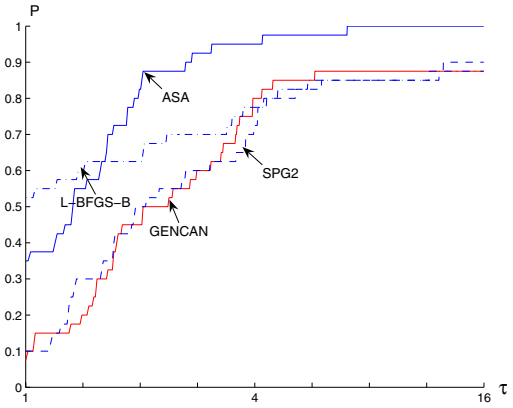


FIG. 6.4. *Performance profiles, evaluation metric,* 50 *CUTEr test problems, gradient-based methods.*

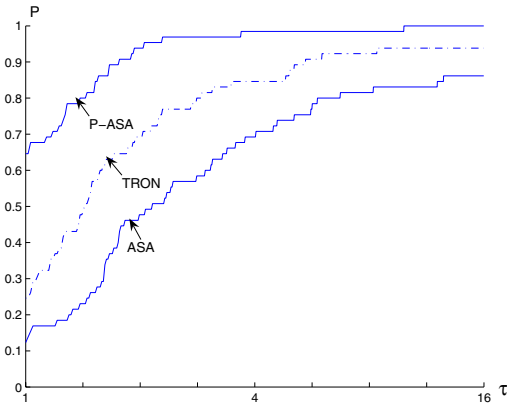


FIG. 6.5. *Performance profiles, evaluation metric,* 42 *sparsest CUTEr problems,* 23 *MINPACK-2 problems.*

7. Conclusions. We have presented a new ASA for solving box constrained optimization problems. The algorithm consists of a nonmonotone gradient projection phase and an unconstrained optimization phase. Rules are given for deciding when to branch from one phase to the other. The branching criteria are based on whether the set of undecided indices is empty or the active set subproblem is solved with sufficient accuracy. We show that for a nondegenerate stationary point, the algorithm eventually reduces to unconstrained optimization without restarts. The analogous result for a degenerate stationary point is established under the strong second-order sufficient optimality condition.

For an implementation of the ASA which uses the CBB method [30] for the nonmonotone gradient projection and which uses CG_DESCENT [54, 55, 56, 57] for unconstrained optimization, we obtained higher CPU time performance profiles than

those of L-BFGS-B, SPG2, GENCAN, and TRON for a test set consisting of all 50 CUTEr [13] box constrained problems with dimension greater than 50, and all 23 MINPACK-2 [1] box constrained problems.

Appendix. An implementation of the ASA. For the numerical results in section 6, our choice for the UA is the conjugate gradient algorithm CG_DESCENT [54, 55, 57, 56]. When an iterate lands outside the feasible set, we may increase the size of the active set using an approach similar to that in [9]. Roughly, we perform an approximate line search for the function

$$\phi(\alpha) = f(P(\mathbf{x}_k + \alpha \mathbf{d}_k))$$

along the current search direction $\mathbf{d}_k$, and any components of $\mathbf{x}_{k+1} = P(\mathbf{x}_k + \alpha_k \mathbf{d}_k)$ which reach the boundary are added to the current active set.

The initial stepsize $\bar{\alpha}_k$ in the NGPA is generated using the CBB method [30]. In the remainder of this section, we explain in detail the initial stepsize computation and choice for the reference function value f_k^r in the NGPA (see [82] for preliminary numerical results based on a closely related initial stepsize and reference function value). We show that these choices satisfy the hypotheses of Theorem 2.2.

The BB stepsize [2] is given by

$$\overline{\alpha}_{k+1}^{BB} = \frac{\mathbf{s}_k^{\mathsf{T}} \mathbf{s}_k}{\mathbf{s}_k^{\mathsf{T}} \mathbf{y}_k}, \tag{A.1}$$

where $\mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$ and $\mathbf{y}_k = \mathbf{g}_{k+1} - \mathbf{g}_k$. An attractive feature of the BB stepsize is that for unconstrained optimization and without a line search, linear convergence is achieved [30] for a starting guess in a neighborhood of the local minimizer with a positive definite Hessian. Moreover, if the same BB stepsize is repeated for several iterations, then even faster convergence is often achieved (see [30]). We refer to schemes that employ the same BB stepsize for several iterations as cyclic BB (CBB) schemes. From an asymptotic perspective, either BB or CBB schemes are inferior to conjugate gradient schemes, for which the convergence rate can be superlinear. On the other hand, for a bound constrained optimization problem, where the active constraints at an optimal solution are unknown, the asymptotic convergence rate is irrelevant until the active constraints are identified. A nonmonotone BB or CBB iteration yields an efficient strategy for identifying active constraints.

When possible, the initial stepsize $\overline{\alpha}_k$ is given by the CBB formula

$$\overline{\alpha}_{k+j} = \overline{\alpha}_k^{BB} \quad \text{for } j = 0, \ldots, m-1,$$

where the BB step appears in (A.1) and m is the number of times the BB step is reused. When $\overline{\alpha}_k^{BB} \notin [\alpha_{\min}, \alpha_{\max}]$, we project it on the interval $[\alpha_{\min}, \alpha_{\max}]$.

We now provide a more detailed statement of our algorithm for computing the initial stepsize. The integer j counts the number of times the current BB step has been reused, while the parameter m is the CBB memory (the maximum number of times the BB step will be reused).

Initial stepsize.

I0. If $k = 0$, choose $\overline{\alpha}_0 \in [\alpha_{\min}, \alpha_{\max}]$ and a parameter $\theta < 1$ near 1; set $j = 0$ and $flag = 1$. If $k > 0$, set $flag = 0$.
I1. If $0 < |d_{ki}| < \bar{\alpha}_k |g_{ki}|$ for some i, then set $flag = 1$.
I2. If $\alpha_k = 1$ in the NGPA, then set $j = j + 1$.
I3. If $\alpha_k < 1$ in the NGPA, then set $flag = 1$.

I4. If $j \geq m$ or $flag = 1$ or $\mathbf{s}_k^T\mathbf{y}_k/(\|\mathbf{s}_k\|\|\mathbf{y}_k\|) \geq \theta$, then
  a. If $\mathbf{s}_k^T\mathbf{y}_k \leq 0$, then
    1. If $j \geq 1.5m$, then set $t = \min\{\|\mathbf{x}_k\|_\infty, 1\}/\|\mathbf{d}^1(\mathbf{x}_k)\|_\infty$, $\overline{\alpha}_{k+1} = \min\{\alpha_{\max}, \max[t, \alpha_k]\}$, and $j = 0$.
    2. Else set $\overline{\alpha}_{k+1} = \overline{\alpha}_k$.
  b. Else set $\overline{\alpha}_{k+1} = \min\{\alpha_{\max}, \max[\alpha_{\min}, \mathbf{s}_k^T\mathbf{s}_k/\mathbf{s}_k^T\mathbf{y}_k]\}$ and $j = 0$.

Since this procedure always generates an initial stepsize $\overline{\alpha}_k \in [\alpha_{\min}, \alpha_{\max}]$, it complies with the requirement in step 1 of the NGPA. If the original BB step is truncated (see I1), or an Armijo line search is performed (see I3), or the cycle number j reaches m (see I4), or $\mathbf{s}_k^T\mathbf{y}_k/(\|\mathbf{s}_k\|\|\mathbf{y}_k\|)$ is close to 1 (see I4), then we try to compute a new BB step. The BB stepsize computation appears in step I4b. One motivation for computing a new BB step when $\mathbf{s}_k^T\mathbf{y}_k/(\|\mathbf{s}_k\|\|\mathbf{y}_k\|)$ is close to 1 is given in [30]; when f is a quadratic, this condition is satisfied when the step $\mathbf{s}_k$ is close to an eigenvector of the Hessian. When $\mathbf{s}_k^T\mathbf{y}_k \leq 0$ (see I4a), the function is not convex on the line segment connecting $\mathbf{x}_k$ and $\mathbf{x}_{k+1}$, and a relatively large stepsize is used in the next iteration. A rationale for the step taken in this case appears in [57].

Now consider the reference function value f_k^r. Let f_k denote $f(\mathbf{x}_k)$. In the algorithm which follows, the integer a counts the number of consecutive iterations that $\alpha_k = 1$ in the NGPA (and the Armijo line search in step 4 is skipped). The integer l counts the number of iterations since the function value is strictly decreased by an amount $\Delta > 0$.

Reference function value.

R0. If $k = 0$, choose parameters $A > L > 0$, $\gamma_1 > 1$, $\gamma_2 > 1$, and $\Delta > 0$; initialize $a = l = 0$ and $f_0^{\min} = f_0^{\max\min} = f_0^r = f_{-1}^r = f_0$.

R1. Update f_k^r as follows:
  a. If $l = L$, then set $l = 0$ and
$$f_k^r = \begin{cases} f_k^{\max\min} & \text{if } \frac{f_k^{\max} - f_k^{\min}}{f_k^{\max\min} - f_k^{\min}} \geq \gamma_1, \\ f_k^{\max} & \text{otherwise.} \end{cases}$$
  b. Else if $a > A$, then set
$$f_k^r = \begin{cases} f_k^{\max} & \text{if } f_k^{\max} > f_k \text{ and } \frac{f_{k-1}^r - f_k}{f_k^{\max} - f_k} \geq \gamma_2, \\ f_{k-1}^r & \text{otherwise.} \end{cases}$$
  c. Otherwise, $f_k^r = f_{k-1}^r$.

R2. Set f_R as follows in step 3 of the NGPA:
  a. If $j = 0$ (first iterate in a CBB cycle), then $f_R = f_k^r$.
  b. If $j > 0$, then $f_R = \min\{f_k^{\max}, f_k^r\}$.

  If $\alpha_k < 1$ in the NGPA, then set $a = 0$.

R3. If $\alpha_k = 1$ in the NGPA, then set $a = a + 1$.

R4. If $f_{k+1} \leq f_k^{\min} - \Delta$, then set $f_{k+1}^{\max\min} = f_{k+1}^{\min} = f_{k+1}$ and $l = 0$; otherwise, put $l = l + 1$, $f_{k+1}^{\min} = f_k^{\min}$, and $f_{k+1}^{\max\min} = \max\{f_k^{\max\min}, f_{k+1}\}$.

The variable $f_k^{\max}$, defined in (2.3), stores the maximum of recent function values. The variable $f_k^{\min}$ stores the minimum function value to within the tolerance Δ. The variable $f_k^{\max\min}$ stores the maximum function value since the last new minimum was recorded in $f_k^{\min}$. More explanations concerning the choice of the reference function value are given in [30, 31]. Now, let us check that the choice for f_k^r given above satisfies the requirements in step 2 of the NGPA.

Proof that $f_k \leq f_k^r$. In R1, we set

(i) $f_k^r = f_k^{\max\min}$ or
(ii) $f_k^r = f_k^{\max}$ or
(iii) $f_k^r = f_{k-1}^r$.

By R4, $f_k^{\max\min} \geq f_k$. In case (ii), $f_k^{\max} \geq f_k$ by the definition of $f_k^{\max}$. In steps 3 and 4 of the NGPA, we have $f_k \leq f_R \leq f_{k-1}^r$. Hence, in each of the cases (i)–(iii), we have $f_k \leq f_k^r$. □

Proof that $f_k^r \leq \max\{f_{k-1}^r, f_k^{\max}\}$. In R1a, f_k^r is equal to either $f_k^{\max}$ or $f_k^{\max\min}$. Since $\gamma_1 > 1$, we set only $f_k^r = f_k^{\max\min}$ when $f_k^{\max\min} \leq f_k^{\max}$. Hence, in R1a, $f_k^r \leq f_k^{\max}$. In R1b, f_k^r is equal to either $f_k^{\max}$ or f_{k-1}^r. Since $\gamma_2 > 1$, we set only $f_k^r = f_{k-1}^r$ when $f_{k-1}^r \geq f_k^{\max}$. Hence, in R1b, $f_k^r \leq f_k^{\max}$. In R1c, we set $f_k^r = f_{k-1}^r$. Combining these observations, $f_k^r \leq \max\{f_{k-1}^r, f_k^{\max}\}$ in R1a–R1c. □

Proof that $f_k^r \leq f_k^{\max}$ *infinitely often.* The condition $f_{k+1} \leq f_k^{\min} - \Delta$ in R4 is satisfied only a finite number of times when f is bounded from below. Thus for k sufficiently large, f_k^r is updated in R1a every L iterations. In this case, since $\gamma_1 > 1$, $f_k^r = f_k^{\max\min}$ only when $f_k^{\max\min} \leq f_k^{\max}$, which implies that $f_k^r \leq f_k^{\max}$. Hence, for large k, $f_k^r \leq f_k^{\max}$ every L iterations. □

Acknowledgments. Constructive comments by the referees are gratefully acknowledged. In particular, the idea of splitting the local convergence analysis into the nondegenerate case (where there is no need to assume the strong second-order sufficient optimality condition), followed by the degenerate case, was suggested by one of the referees.

REFERENCES

[1] B. M. Averick, R. G. Carter, J. J. Moré, and G. L. Xue, *The MINPACK-2 Test Problem Collection*, Tech. report, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, 1992.
[2] J. Barzilai and J. M. Borwein, *Two point step size gradient methods*, IMA J. Numer. Anal., 8 (1988), pp. 141–148.
[3] D. P. Bertsekas, *On the Goldstein-Levitin-Polyak gradient projection method*, IEEE Trans. Automat. Control, 21 (1976), pp. 174–184.
[4] D. P. Bertsekas, *Projected Newton methods for optimization problems with simple constraints*, SIAM J. Control Optim., 20 (1982), pp. 221–246.
[5] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, Belmont, MA, 1999.
[6] E. G. Birgin, R. Biloti, M. Tygel, and L. T. Santos, *Restricted optimization: A clue to a fast and accurate implementation of the common reflection surface stack method*, J. Appl. Geophys., 42 (1999), pp. 143–155.
[7] E. G. Birgin, I. Chambouleyron, and J. M. Martínez, *Estimation of the optical constants and the thickness of thin films using unconstrained optimization*, J. Comput. Phys., 151 (1999), pp. 862–880.
[8] E. G. Birgin and J. M. Martínez, *A box-constrained optimization algorithm with negative curvature directions and spectral projected gradients*, in Topics in Numerical Analysis, Comput. Suppl. 15, Springer, Vienna, 2001, pp. 49–60.
[9] E. G. Birgin and J. M. Martínez, *Large-scale active-set box-constrained optimization method with spectral projected gradients*, Comput. Optim. Appl., 23 (2002), pp. 101–125.
[10] E. G. Birgin, J. M. Martínez, and M. Raydan, *Nonmonotone spectral projected gradient methods on convex sets*, SIAM J. Optim., 10 (2000), pp. 1196–1211.
[11] E. G. Birgin, J. M. Martínez, and M. Raydan, *Algorithm* 813: *SPG—software for convex-constrained optimization*, ACM Trans. Math. Software, 27 (2001), pp. 340–349.
[12] E. G. Birgin, J. M. Martínez, and M. Raydan, *Inexact spectral projected gradient methods on convex sets*, IMA J. Numer. Anal., 23 (2003), pp. 539–559.
[13] I. Bongartz, A. R. Conn, N. I. M. Gould, and P. L. Toint, *CUTE: Constrained and unconstrained testing environments*, ACM Trans. Math. Software, 21 (1995), pp. 123–160.
[14] M. A. Branch, T. F. Coleman, and Y. Li, *A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems*, SIAM J. Sci. Comput., 21 (1999), pp. 1–23.

[15] J. V. Burke and J. J. Moré, *On the identification of active constraints*, SIAM J. Numer. Anal., 25 (1988), pp. 1197–1211.
[16] J. V. Burke and J. J. Moré, *Exposing constraints*, SIAM J. Optim., 4 (1994), pp. 573–595.
[17] J. V. Burke, J. J. Moré, and G. Toraldo, *Convergence properties of trust region methods for linear and convex constraints*, Math. Program., 47 (1990), pp. 305–336.
[18] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, *A limited memory algorithm for bound constrained optimization*, SIAM J. Sci. Comput., 16 (1995), pp. 1190–1208.
[19] P. Calamai and J. Moré, *Projected gradient for linearly constrained problems*, Math. Program., 39 (1987), pp. 93–116.
[20] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.
[21] T. F. Coleman and Y. Li, *On the convergence of interior-reflective Newton methods for nonlinear minimization subject to bounds*, Math. Program., 67 (1994), pp. 189–224.
[22] 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.
[23] T. F. Coleman and Y. Li, *A Trust Region and Affine Scaling Interior Point Method for Nonconvex Minimization with Linear Inequality Constraints*, Tech. report, Cornell University, Ithaca, NY, 1997.
[24] A. R. Conn, N. I. M. Gould, and Ph. L. Toint, *Global convergence of a class of trust region algorithms for optimization with simple bounds*, SIAM J. Numer. Anal., 25 (1988), pp. 433–460.
[25] 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.
[26] A. R. Conn, N. I. M. Gould, and Ph. L. Toint, *Trust Region Methods*, SIAM, Philadelphia, 2000.
[27] Y. H. Dai, *On the nonmonotone line search*, J. Optim. Theory Appl., 112 (2002), pp. 315–330.
[28] Y. H. Dai and R. Fletcher, *Projected Barzilai-Borwein methods for large-scale box-constrained quadratic programming*, Numer. Math., 100 (2005), pp. 21–47.
[29] Y. H. Dai and R. Fletcher, *New algorithms for singly linearly constrained quadratic programs subject to lower and upper bounds*, Math. Program., 106 (2006), pp. 403–421.
[30] Y. H. Dai, W. W. Hager, K. Schittkowski, and H. Zhang, *The cyclic Barzilai-Borwein method for unconstrained optimization*, IMA J. Numer. Anal., to appear.
[31] Y. H. Dai and H. Zhang, *An adaptive two-point stepsize gradient algorithm*, Numer. Algorithms, 27 (2001), pp. 377–385.
[32] R. S. Dembo and U. Tulowitzki, *On the Minimization of Quadratic Functions Subject to Box Constraints*, Tech. report, School of Organization and Management, Yale University, New Haven, CT, 1983.
[33] J. E. Dennis, M. Heinkenschloss, and L. N. Vicente, *Trust-region interior-point algorithms for a class of nonlinear programming problems*, SIAM J. Control Optim., 36 (1998), pp. 1750–1794.
[34] E. D. Dolan and J. J. Moré, *Benchmarking optimization software with performance profiles*, Math. Program., 91 (2002), pp. 201–213.
[35] Z. Dostál, *Box constrained quadratic programming with proportioning and projections*, SIAM J. Optim., 7 (1997), pp. 871–887.
[36] Z. Dostál, *A proportioning based algorithm for bound constrained quadratic programming with the rate of convergence*, Numer. Algorithms, 34 (2003), pp. 293–302.
[37] Z. Dostál, A. Friedlander, and S. A. Santos, *Solution of coercive and semicoercive contact problems by FETI domain decomposition*, Contemp. Math., 218 (1998), pp. 82–93.
[38] Z. Dostál, A. Friedlander, and S. A. Santos, *Augmented Lagrangians with adaptive precision control for quadratic programming with simple bounds and equality constraints*, SIAM J. Optim., 13 (2003), pp. 1120–1140.
[39] A. S. El-Bakry, R. A. Tapia, T. Tsuchiya, and Y. Zhang, *On the formulation and theory of the primal-dual Newton interior-point method for nonlinear programming*, J. Optim. Theory Appl., 89 (1996), pp. 507–541.
[40] A. S. El-Bakry, R. A. Tapia, and Y. Zhang, *On the convergence rate of Newton interior-point methods in the absence of strict complementarity*, Comput. Optim. Appl., 6 (1996), pp. 157–167.
[41] F. Facchinei, A. Fischer, and C. Kanzow, *On the accurate identification of active constraints*, SIAM J. Optim., 9 (1998), pp. 14–32.
[42] 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.

[43] F. Facchinei and S. Lucidi, *A class of penalty functions for optimization problems with bound constraints*, Optimization, 26 (1992), pp. 239–259.
[44] 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.
[45] R. Fletcher, *On the Barzilai-Borwein Method*, Tech. report, Department of Mathematics, University of Dundee, Dundee, Scotland, 2001.
[46] 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.
[47] R. Glowinski, *Numerical Methods for Nonlinear Variational Problems*, Springer-Verlag, Berlin, New York, 1984.
[48] W. Glunt, T. L. Hayden, and M. Raydan, *Molecular conformations from distance matrices*, J. Comput. Chem., 14 (1993), pp. 114–120.
[49] A. A. Goldstein, *Convex programming in Hilbert space*, Bull. Amer. Math. Soc., 70 (1964), pp. 709–710.
[50] L. Grippo, F. Lampariello, and S. Lucidi, *A nonmonotone line search technique for Newton's method*, SIAM J. Numer. Anal., 23 (1986), pp. 707–716.
[51] L. Grippo and M. Sciandrone, *Nonmonotone globalization techniques for the Barzilai-Borwein gradient method*, Comput. Optim. Appl., 23 (2002), pp. 143–169.
[52] W. W. Hager, *Dual techniques for constrained optimization*, J. Optim. Theory Appl., 55 (1987), pp. 37–71.
[53] W. W. Hager, *Analysis and implementation of a dual algorithm for constrained optimization*, J. Optim. Theory Appl., 79 (1993), pp. 427–462.
[54] W. W. Hager and H. Zhang, *CG_DESCENT User's Guide*, Tech. report, Department of Mathematics, University of Florida, Gainesville, FL, 2004.
[55] W. W. Hager and H. Zhang, *A new conjugate gradient method with guaranteed descent and an efficient line search*, SIAM J. Optim., 16 (2005), pp. 170–192.
[56] W. W. Hager and H. Zhang, *A survey of nonlinear conjugate gradient methods*, Pacific J. Optim., 2 (2006), pp. 35–58.
[57] W. W. Hager and H. Zhang, *Algorithm* 851: *CG_DESCENT, a conjugate gradient method with guaranteed descent*, ACM Trans. Math. Software, 32 (2006), pp. 113–137.
[58] M. Heinkenschloss, M. Ulbrich, and S. Ulbrich, *Superlinear and quadratic convergence of affine-scaling interior-point Newton methods for problems with simple bounds without strict complementarity assumption*, Math. Program., 86 (1999), pp. 615–635.
[59] C. Kanzow and A. Klug, *On affine-scaling interior-point Newton methods for nonlinear minimization with bound constraints*, Comput. Optim. Appl., to appear.
[60] M. Lescrenier, *Convergence of trust region algorithms for optimization with bounds when strict complementarity does not hold*, SIAM J. Numer. Anal., 28 (1991), pp. 476–495.
[61] E. S. Levitin and B. T. Polyak, *Constrained minimization problems*, USSR Comput. Math. Math. Physics, 6 (1966), pp. 1–50.
[62] C.-J. Lin and J. J. Moré, *Incomplete Cholesky factorizations with limited memory*, SIAM J. Sci. Comput., 21 (1999), pp. 24–45.
[63] C.-J. Lin and J. J. Moré, *Newton's method for large bound-constrained optimization problems*, SIAM J. Optim., 9 (1999), pp. 1100–1127.
[64] W. B. Liu and Y. H. Dai, *Minimization algorithms based on supervisor and searcher cooperation*, J. Optim. Theory Appl., 111 (2001), pp. 359–379.
[65] J. M. Martínez, *BOX-QUACAN and the implementation of augmented Lagrangian algorithms for minimization with inequality constraints*, J. Comput. Appl. Math., 19 (2000), pp. 31–56.
[66] G. P. McCormick and R. A. Tapia, *The gradient projection method under mild differentiability conditions*, SIAM J. Control, 10 (1972), pp. 93–98.
[67] 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.
[68] B. T. Polyak, *The conjugate gradient method in extremal problems*, USSR Comput. Math. Math. Phys., 9 (1969), pp. 94–112.
[69] M. Raydan, *The Barzilai and Borwein gradient method for the large scale unconstrained minimization problem*, SIAM J. Optim., 7 (1997), pp. 26–33.
[70] S. M. Robinson, *Strongly regular generalized equations*, Math. Oper. Res., 5 (1980), pp. 43–62.
[71] A. Schwartz and E. Polak, *Family of projected descent methods for optimization problems with simple bounds*, J. Optim. Theory Appl., 92 (1997), pp. 1–31.
[72] T. Serafini, G. Zanghirati, and L. Zanni, *Gradient projection methods for quadratic programs and applications in training support vector machines*, Optim. Methods Softw., 20 (2005), pp. 353–378.

[73] 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.

[74] P. L. Toint, *An assessment of nonmonotone line search techniques for unconstrained optimization*, SIAM J. Sci. Comput., 17 (1996), pp. 725–739.

[75] P. L. Toint, *A non-monotone trust region algorithm for nonlinear optimization subject to convex constraints*, Math. Program., 77 (1997), pp. 69–94.

[76] 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.

[77] S. J. Wright, *Implementing proximal point methods for linear programming*, J. Optim. Theory Appl., 65 (1990), pp. 531–554.

[78] H. Yamashita and H. Yabe, *Superlinear and quadratic convergence of some primal-dual interior-point methods for constrained optimization*, Math. Program., 75 (1996), pp. 377–397.

[79] E. K. Yang and J. W. Tolle, *A class of methods for solving large convex quadratic programs subject to box constraints*, Math. Program., 51 (1991), pp. 223–228.

[80] E. H. Zarantonello, *Projections on convex sets in Hilbert space and spectral theory*, in Contributions to Nonlinear Functional Analysis, E. H. Zarantonello, ed., Academic Press, New York, 1971, pp. 237–424.

[81] H. Zhang and W. W. Hager, *A nonmonotone line search technique and its application to unconstrained optimization*, SIAM J. Optim., 14 (2004), pp. 1043–1056.

[82] H. Zhang and W. W. Hager, *PACBB: A projected adaptive cyclic Barzilai-Borwein method for box constrained optimization*, in Multiscale Optimization Methods and Applications, William W. Hager, Shu-Jen Huang, Panos M. Pardalos, and Oleg A. Prokopyev, eds., Springer, New York, 2005, pp. 387–392.

[83] Y. Zhang, *Interior-point Gradient Methods with Diagonal-scalings for Simple-bound Constrained Optimization*, Tech. report TR04-06, Department of Computational and Applied Mathematics, Rice University, Houston, TX, 2004.

[84] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, *Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization*, ACM Trans. Math. Software, 23 (1997), pp. 550–560.

SIAM J. OPTIM.
Vol. 17, No. 2, pp. 558–576
© 2006 Society for Industrial and Applied Mathematics

JORDAN-ALGEBRAIC APPROACH TO CONVEXITY THEOREMS FOR QUADRATIC MAPPINGS*

L. FAYBUSOVICH†

Abstract. We describe a Jordan-algebraic version of results related to convexity of images of quadratic mappings as well as related results on exactness of symmetric relaxations of certain classes of nonconvex optimization problems. The exactness of relaxations is proved based on rank estimates. Our approach provides a unifying viewpoint on a large number of classical results related to cones of Hermitian matrices over real and complex numbers. We describe (apparently new) results related to cones of Hermitian matrices with quaternion entries and to the exceptional 27-dimensional Euclidean Jordan algebra.

Key words. Jordan-algebraic technique, symmetric relaxations, convexity theorems

AMS subject classifications. 90C26, 52A41

DOI. 10.1137/050635560

1. Introduction. Starting with [F1, F2], the Jordan-algebraic technique proved to be useful as a unifying tool for the description and analysis of interior-point algorithms. In the present paper we use this technique for similar goals, but with the difference being that we use it to study the convexity images of quadratic mappings between finite-dimensional vector spaces. This circle of problems has numerous connections with optimization theory (see, e.g., [Pol] and the references therein for a discussion of various connections of this type). In particular, questions such as under what assumptions are semidefinite relaxations of quadratically constrained quadratic programming problems exact (see, e.g., [YZ] and the references therein) or when can one omit rank constraints in semidefinite programming problems (see [BM] and the references therein) are important for modern optimization theory. Another very interesting connection is with the famous S-lemma (see [BN, pp. 300–314]). Our approach is modeled on the work of Barvinok [B2, B3, B1] but is developed within a more general framework of Jordan algebras. The paper is organized as follows. In section 2 we briefly describe Jordan-algebraic concepts related to our discussion. In section 3 we give a complete description of the facial structure of a symmetric cone in the form somewhat different from one in [FK]. Section 4 is of central importance and provides estimates on the rank of a feasible point in the intersection of an affine subspace and a symmetric cone. Our results are a direct (but not an immediate!) generalization of the results of Barvinok, who considered the cones of Hermitian matrices over $\mathbf{R}$ and $\mathbf{C}$. In section 5 we derive from rank estimates some convexity results and results about exact convex relaxations of generally nonconvex optimization problems. It is done within the general Jordan-algebraic context. The central object here is the manifold of primitive idempotents in a simple Euclidean Jordan algebra (or its conic hull). A very general Jordan-algebraic version of the well-known S-lemma is given. In section 6 we interpret the results of section 5 for concrete symmetric cones. The

*Received by the editors July 8, 2005; accepted for publication (in revised form) February 13, 2006; published electronically August 16, 2006. This research was supported in part by NSF grant DMS-042740.

http://www.siam.org/journals/siopt/17-2/63556.html

†Department of Mathematics, University of Notre Dame, 255 Hurley Hall, Notre Dame, IN 46545 (leonid.faybusovich.1@ud.edu).

cases of symmetric cones corresponding to algebras of Hermitian matrices over **H** and exceptional 27-dimensional algebra seem to lead to new results.

Another type of convexity result (a Jordan-algebraic version of the Horn–Schur theorem) was obtained in [LKF].

2. Jordan-algebraic concepts. We stick to the notation of an excellent book [FK]. We do not attempt to describe the Jordan-algebraic language here but instead provide detailed references to [FK]. Throughout this paper, we use the following notation:

- V is a simple Euclidean Jordan algebra;
- $\mathrm{rank}(V)$ stands for the rank of V;
- $x \circ y$ is the Jordan algebraic multiplication for $x, y \in V$;
- $\langle x, y\rangle = \mathrm{tr}(x \circ y)$ is the canonical scalar product in V; here tr is the trace operator on V;
- Ω is the cone of invertible squares in V;
- $\bar{\Omega}$ is the closure of Ω in V;
- an element $f \in V$ such that $f^2 = f$ and $\mathrm{tr}(f) = 1$ is called a primitive idempotent in V;
- the set $\mathcal{T}(V)$ of primitive idempotents is a smooth compact connected submanifold in V;
- given $x \in V$, we denote by $L(x)$ the corresponding multiplication operator on V, i.e.,

$$L(x)y = x \circ y, \quad y \in V;$$

- given $x \in V$, we denote by $P(x)$ the so-called quadratic representation of x, i.e.,

$$P(x) = 2L(x)^2 - L(x^2).$$

Given $x \in V$, there exist idempotents $f_1, \dots, f_k$ in V such that $f_i \circ f_j = 0$ for $i \neq j$ and such that $f_1 + f_2 + \cdots + f_k = e$, and distinct real numbers $\lambda_1, \dots, \lambda_k$ with the following property:

$$x = \sum_{i=1}^{k} \lambda_i f_i. \tag{1}$$

The numbers λ_i and idempotents f_i are uniquely defined by x (see Theorem III.1.1 in [FK]).

The representation (1) is called the spectral decomposition of x. Within the context of this paper the notion of the rank of x is very important. By definition,

$$\mathrm{rank}(x) = \sum_{i:\lambda_i \neq 0} \mathrm{tr}(f_i). \tag{2}$$

Given $x \in V$, the operator $L(x)$ is symmetric with respect to the canonical scalar product. If f is an idempotent in V, it turns out that the spectrum of $L(f)$ belongs to $\{0, \frac{1}{2}, 1\}$. Following [FK], we denote by $V(1, f), V(\frac{1}{2}, f), V(0, f)$ corresponding eigenspaces.

It is clear that

$$V = V(0, f) \oplus V(1, f) \oplus V\left(\frac{1}{2}, f\right) \tag{3}$$

and the eigenspaces are pairwise orthogonal with respect to the scalar product $\langle , \rangle$. This is the so-called Peirce decomposition of V with respect to an idempotent f. However, eigenspaces have more structure (see [FK, Proposition IV.1.1]). In particular, $V(0,f), V(1,f)$ are subalgebras in V. Let f_1, f_2 be two primitive orthogonal idempotents. It turns out that

$$\dim V\left(\frac{1}{2}, f_1\right) \cap V\left(\frac{1}{2}, f_2\right)$$

does not depend on the choice of the pair f_1, f_2 (see Corollary IV.2.6, p. 71 in [FK]). It is called the degree of V (notation $\mathrm{d}(V)$).

If V is a simple Euclidean Jordan algebra, then

$$\dim V = \operatorname{rank}(V) + \frac{\mathrm{d}(V)}{2} \operatorname{rank}(V)(\operatorname{rank}(V) - 1).$$

Note that two simple Euclidean Jordan algebras are isomorphic if and only if their ranks and degrees coincide.

The next proposition will be frequently used in what follows.

PROPOSITION 1. *Let $x, y \in \bar{\Omega}$. Then $\langle x, y\rangle \geq 0$; $\langle x, y\rangle = 0$ if and only if $x \circ y = 0$.*

For a proof see, e.g., [F2].

We summarize some of the properties of algebras $V(1,f)$.

PROPOSITION 2. *Let f be an idempotent in a simple Euclidean Jordan algebra V. Then $V(1,f)$ is a simple Euclidean Jordan algebra with identity element f. Moreover,*

$$\operatorname{rank}(V(1,f)) = \operatorname{rank}(f),$$
$$\mathrm{d}(V(1,f)) = \mathrm{d}(V).$$

The trace operator on $V(1,f)$ coincides with the restriction of the trace operator on V. If $\tilde{\Omega}$ is the cone of invertible squares in $V(1,f)$, then $\bar{\tilde{\Omega}} = \bar{\Omega} \cap V(1,f)$.

Proposition 2 easily follows from the properties of Peirce decomposition on V (see section IV.2 in [FK]). Notice that if c is a primitive idempotent in $V(1,f)$, then c is primitive idempotent in V, i.e., $\mathcal{T}(V(1,f)) = \mathcal{T}(V) \cap V(1,f)$.

Indeed, let $c = c_1 + c_2$ where $c_1, c_2 \in V$ and $c_1^2 = c_1, c_2^2 = c_2$. Since $c \in V(1,f), c \circ (e-f) = 0$, i.e., $(e-f) \circ c_1 + (e-f) \circ c_2 = 0$.

Hence, $\langle e-f, c_1\rangle + \langle e-f, c_2\rangle = 0$. But $e-f, c_1, c_2 \in \bar{\Omega}$. Hence, $\langle e-f, c_1\rangle \geq 0, \langle e-f, c_2\rangle \geq 0$. We conclude that $\langle e-f, c_1\rangle = \langle e-f, c_2\rangle = 0$. By Proposition 1 $(e-f) \circ c_1 = (e-f) \circ c_2 = 0$, i.e., $c_1, c_2 \in V(1,f)$. But c is primitive in $V(1,f)$. Hence $c_1 = 0$ or $c_2 = 0$, which proves that c is primitive in V.

Let $f_1, \ldots, f_r$, where $r = \operatorname{rank}(V)$, be a system of primitive idempotents such that $f_i \circ f_j = 0$ for $i \neq j$ and $f_1 + \cdots + f_r = e$. Such a system is called a Jordan frame. Given $x \in V$, there exists a Jordan frame $f_1, \ldots, f_r$ and real numbers $\lambda_1, \ldots, \lambda_r$ such that

$$x = \sum_{i=1}^{r} \lambda_i f_i.$$

The numbers λ_i (with their multiplicities) are uniquely determined by x (see Theorem III.1.2 in [FK]).

It is clear that

$$\operatorname{tr}(x) = \sum_{i=1}^{r} \lambda_i, \quad \operatorname{rank}(x) = \operatorname{card}\{i \in [1,r] : \lambda_i \neq 0\}.$$

Since primitive idempotents in $V(1,f)$ remain primitive in V, it easily follows that the rank of $x \in V(1,f)$ is the same as its rank in V.

3. Facial structure of the cone of squares. Throughout this paper we will use notation B_ϵ for an open ball in V with the center at 0 and of radius ϵ (with respect to the norm induced by the canonical Euclidean product). Given a subset $S \subset V$, we denote by $\mathrm{Aff}(S)$ the smallest affine subspace in V containing S (affine hull of S). The notation $\mathrm{ri}(S)$ is used for the relative interior of S:

$$\mathrm{ri}(S) = \{x \in S : \exists \epsilon > 0, (x + B_\epsilon) \cap \mathrm{Aff}(S) \subset S\}.$$

Let S be a convex subset of V. A face of S is a convex subset T of S such that whenever $\lambda x + \mu y \in T$, where $x, y \in S, \lambda, \mu > 0, \lambda + \mu = 1$, then $x, y \in T$. Recall the following theorem (Theorem 2.6.10 in [W]).

THEOREM 1. *Let a be a point of convex set S in V. Let $\mathcal{F}_a$ be the intersection of all faces of S containing a. Then $\mathcal{F}_a$ is a face of S. Moreover, $a \in \mathrm{ri}(\mathcal{F}_a)$ and the relative interiors of the faces of S form a partition of S.*

In this section we describe the facial structure of the cone of squares $\bar{\Omega}$ in V in a form somewhat different from the one given in Proposition IV.3.1 of [FK].

THEOREM 2. *Let $x \in \partial\Omega = \bar{\Omega} \backslash \Omega$ and*

$$x = \sum_{i=1}^{k+1} \lambda_i(x) f_i(x)$$

be the spectral decomposition of x, where $\lambda_i(x) > 0$, $i = 1, 2, \ldots, k$, $\lambda_{k+1}(x) = 0$ are pairwise distinct eigenvalues of x. Let

$$\mathcal{F}_x = \{z \in \bar{\Omega} : \langle f_{k+1}(x), z\rangle = 0\}.$$

Then $\mathcal{F}_x$ is a face of $\bar{\Omega}$. Moreover, $x \in \mathrm{ri}(\mathcal{F}_x), \mathrm{Aff}(\mathcal{F}_x) = V(0, f_{k+1}(x)) = V(1, f_1(x) + \cdots + f_k(x)); \mathrm{ri}(\mathcal{F}_x) = \tilde{\Omega}$, where $\tilde{\Omega}$ is the cone of invertible squares in $V(0, f_{k+1}(x))$, $\bar{\tilde{\Omega}} = \bar{\Omega} \cap V(0, f_{k+1}(x))$.

COROLLARY 1. *In particular,*

$$\dim \mathcal{F}_x = \dim V(1, f_1(x) + \cdots + f_k(x)) = \varphi_d(\mathrm{rank}(x)),$$
$$\varphi_d(x) = x + \frac{dx(x-1)}{2}.$$

Here $d = d(V)$ is the degree of V.

Proof of Theorem 2. Since $x \in \partial\Omega, f_{k+1}(x) \neq 0$. Let $H_x = \{z \in V; \langle z, f_{k+1}(x)\rangle = 0\}$. If $z \in \bar{\Omega}$, then $\langle z, f_{k+1}(x)\rangle \geq 0$ ($f_{k+1}^2(x) = f_{k+1}(x)$; hence, $f_{k+1}(x) \in \bar{\Omega}$). This implies that H_x is a supporting hyperplane to $\bar{\Omega}$ ($x \in H_x$). Thus, $\mathcal{F} = H_x \cap \bar{\Omega}$ is a face of $\bar{\Omega}$ and, moreover, $x \in \mathcal{F}$. Furthermore, $y \in \mathcal{F}$ is equivalent to $y \in \bar{\Omega}$ and $\langle y, f_{k+1}(x)\rangle = 0$. But then, by Proposition 1, $y \circ f_{k+1}(x) = 0$ and consequently $y \in V(0, f_{k+1}(x)) = V(1, e - f_{k+1}(x))$. Thus, $\mathcal{F} \subset V(0, f_{k+1}(x))$, which implies $\mathrm{Aff}(\mathcal{F}) \subset V(0, f_{k+1}(x))$. It is clear that $x \in \tilde{\Omega}$ (with inverse $\sum_{i=1}^k (1/\lambda_i) f_i(x)$ in $V(0, f_{k+1}(x))$).

Thus $\mathrm{rank} V(0, f_{k+1}(x)) = \mathrm{rank}(x)$. Since $\bar{\tilde{\Omega}} = \bar{\Omega} \cap V(0, f_{k+1}(x))$ by Proposition 2 we conclude that $\tilde{\Omega} \subset \mathcal{F}$, which implies that $V(0, f_{k+1}(x)) = \mathrm{Aff}(\tilde{\Omega}) \subset \mathrm{Aff}(\mathcal{F})$. Thus

$$V(0, f_{k+1}(x)) = \mathrm{Aff}(\mathcal{F}).$$

Since $\tilde{\Omega}$ is relatively open in $V(0, f_{k+1}(x))$, we see that $\tilde{\Omega} \subset \mathrm{ri}(\mathcal{F})$. On the other hand, $\mathcal{F} = \mathcal{F} \cap \mathrm{Aff}(\mathcal{F}) = \bar{\Omega} \cap V(0, f_{k+1}(x)) = \bar{\tilde{\Omega}}$. Hence, $\mathrm{ri}(\mathcal{F}) = \mathrm{ri}(\bar{\tilde{\Omega}}) = \tilde{\Omega}$.

4. Rank estimates. We are now in a position to generalize the main results of [B2, B3] to arbitrary symmetric cones.

THEOREM 3. *Let $\mathcal{A}$ be an affine subspace in V such that*

$$S = \bar{\Omega} \cap \mathcal{A} \neq \emptyset.$$

Then there exists $x \in S$ such that

$$\varphi_d(\text{rank}(x)) \leq \text{codim}_V \mathcal{A}.$$

Here d is the degree of V.

Proof. Since S is closed, nonempty, and does not contain straight lines, it contains an extreme point x (see, e.g., [B1, p. 53], [B2, B3]). Let $\text{rank}(x) = m$. There exists a unique face $\mathcal{F}_x$ of $\bar{\Omega}$ such that $x \in \text{ri}(\mathcal{F}_x)$. By Theorem 2, $\dim\mathcal{F}_x = \varphi_d(m)$. It is clear that $\mathcal{F}_x \cap \mathcal{A}$ is a face of S and, moreover, $x \in \text{ri}(\mathcal{F}_x \cap \mathcal{A})$. Indeed, $x \in \text{ri}(\mathcal{F}_x)$ implies that $\exists \epsilon > 0$ such that $(x + B_\epsilon) \cap \text{Aff}(\mathcal{F}_x) \subset \mathcal{F}_x$. But then $(x + B_\epsilon) \cap \text{Aff}(\mathcal{F}_x) \cap \mathcal{A} \subset \mathcal{F}_x \cap \mathcal{A}$. Since $\text{Aff}(\mathcal{F}_x \cap \mathcal{A}) \subset \text{Aff}(\mathcal{F}_x) \cap \mathcal{A}$ we have $x \in \text{ri}(\mathcal{F}_x \cap A)$. Since x is an extreme point of S, $x \in \text{ri}(\mathcal{F}_x \cap \mathcal{A})$, and $\mathcal{F}_x \cap \mathcal{A}$ is a face of S, we conclude that $\mathcal{F}_x \cap \mathcal{A} = \{x\}$ (by Theorem 1, there exists a unique face $\mathcal{F}$ of S such that $x \in \text{ri}(\mathcal{F})$).

Let $\text{Aff}(\mathcal{F}_x) = x + X, \mathcal{A} = x + Y$, where X, Y is a vector subspace of V. We are going to show that $X \cap Y = 0$. We know that there exists $\epsilon > 0$ such that $(x + B_\epsilon) \cap (x + X) \subset \mathcal{F}_x$. Hence, $(x + B_\epsilon) \cap (x + X) \cap (x + Y) = x + (B_\epsilon \cap X \cap Y) \subset \mathcal{F}_x \cap \mathcal{A} = \{x\}$. If $X \cap Y \neq 0$, we would arrive at a contradiction. Now, $X \cap Y = 0$ implies that $\dim(X + Y) = \dim X + \dim Y$. On the other hand, $\dim(X + Y) \leq \dim V$. Hence, $\dim\mathcal{F}_x = \dim X \leq \dim V - \dim Y = \text{codim}_V \mathcal{A}$. We noticed before that $\dim\mathcal{F}_x = \varphi_d(\text{rank}(x))$. The result follows. □

Remark. Let $a_1, \ldots, a_k \in V, b_1, \ldots, b_k \in \mathbf{R}$, and

$$\mathcal{A} = \{z \in V : \langle a_i, z\rangle = b_i, i = 1, \ldots, k\}.$$

It is clear that $\text{codim}_V \mathcal{A} \leq k$, provided $\mathcal{A} \neq \emptyset$. In this case Theorem 3 implies that $\varphi_d(\text{rank}(x)) \leq k$.

Remark. In the case where V is the Jordan algebra of real symmetric matrices, Theorem 3 coincides with the result on p. 194 of [B2]. See also [Pat]. In this case $\text{d}(V) = 1$.

THEOREM 4. *Let $\mathcal{A}$ be an affine subspace in V such that*

$$S = \bar{\Omega} \cap \mathcal{A}$$

is nonempty and bounded. Suppose that there exists an integer $r \geq 1$ such that $\text{codim}_V(\mathcal{A}) \leq \varphi_d(r + 1)$, $\text{rank}(V) \geq r + 2$. Then there exists $x \in S$ such that $\text{rank}(x) \leq r$. Here $d = d(V)$ is the degree of V.

Proof. We need to consider several cases.

(i) Let $\mathcal{A} \cap \Omega = \emptyset$. But $\mathcal{A} \cap \partial\Omega \neq \emptyset$. Let $y \in \mathcal{A} \cap \partial\Omega$. Since $\mathcal{A} \cap \Omega = \emptyset$, there exists a hyperplane H in V separating $\mathcal{A}$ and Ω. Since $\mathcal{A} \cap \partial\Omega \neq \emptyset$ we should have that H is a supporting hyperplane to $\bar{\Omega}$ and A (as an affine subspace) is a subset in H. Then $\mathcal{F} = H \cap \bar{\Omega}$ is a proper face of $\bar{\Omega}$. By Theorem 2 $H \cap \bar{\Omega}$ is the face of the form $\bar{\Omega}_0$ where $\bar{\Omega}_0$ is the cone of squares in the algebra $V(0, f)$ and f is a nonzero idempotent in V. Since $\mathcal{A} \subset H$, we have $S = \mathcal{A} \cap \bar{\Omega} \subset H \cap \bar{\Omega} = \bar{\Omega}_0$. Thus $S \subset \bar{\Omega}_0 \cap (V(0, f) \cap \mathcal{A})$. Since $\bar{\Omega}_0 = \bar{\Omega} \cap V(0, f) \subset \bar{\Omega}$, we have $\bar{\Omega}_0 \cap (V(0, f) \cap \mathcal{A}) \subset \bar{\Omega} \cap \mathcal{A}$. Thus

$$S = \bar{\Omega}_0 \cap (V(0, f) \cap \mathcal{A}).$$

Let us estimate $\mathrm{codim}_{V(0,f)}\mathcal{A}\cap V(0,f)$.

Let $\mathcal{A}=y+Y$, where $y\in S, Y$ is a vector subspace in V. Denote $V(0,f)$ by W. We have $\dim(Y\cap W)=\dim(W)+\dim Y-\dim(W+Y)$. Now $W+Y=(y+Y)+W\subset H$. Hence, $\dim(W+Y)\le \dim H=\dim V-1$. Consequently, $\dim W-\dim(Y\cap W)=\dim(W+Y)-\dim Y\le \dim V-\dim Y-1=\mathrm{codim}_V\mathcal{A}-1$. Thus $\mathrm{codim}_W(Y\cap W)\le \mathrm{codim}_V\mathcal{A}-1\le \varphi_d(r+1)-1$.

The last inequality is due to the assumptions of the theorem. We can apply Theorem 3 to $S=(\mathcal{A}\cap W)\cap\bar\Omega_0\subset W$ to conclude that there is $x\in S$ such that $\varphi_d(\mathrm{rank}(x))\le\varphi_d(r+1)-1$. Since φ_d is an increasing function on $[0,+\infty)$ (notice that $d\ge 1$), we conclude that $\mathrm{rank}(x)\le r$.

Suppose now that $\mathcal{A}\cap\Omega\neq\emptyset$.

(ii) Consider first the case $\mathrm{rank}(V)=r+2, r\ge 1, \mathcal{A}\subset V$, and $\mathrm{codim}_V(\mathcal{A})=\varphi_d(r+1)$. Since $\mathcal{A}\cap\Omega\neq\emptyset$, we have $\mathrm{ri}(S)=\mathcal{A}\cap\Omega$, $\mathrm{Aff}(S)=\mathcal{A}$. Hence, $\dim S=\dim\mathcal{A}=\dim V-\mathrm{codim}_V\mathcal{A}=\varphi_d(r+2)-\varphi_d(r+1)=1+d(r+1)$. It is also clear that $\partial_1 S=\mathrm{rebd}S=\mathcal{A}\cap\partial\Omega$. If there exists $x\in\partial_1 S$ such that $\mathrm{rank}(x)\le r$, there is nothing to prove. For $x\in\mathrm{ri}(S)$, $\mathrm{rank}(x)=\mathrm{rank}(V)=r+2$. Otherwise, notice that for $x\in\partial_1 S$ $\mathrm{rank}(x)<\mathrm{rank}(V)=r+2$. Thus we should have $\mathrm{rank}(x)=r+1$ for all $x\in\partial_1 S$. Take $y\in\mathcal{A}\cap\Omega$. There exists $\epsilon>0$ such that $(y+B_\epsilon)\cap\mathcal{A}\subset S$. Let $\mathcal{A}=y+Y, Y$ be a vector subspace in V. Consider $\mathbf{S}_\epsilon=\partial_1(B_\epsilon\cap Y)=\partial B_\epsilon\cap Y$. It is clear that $\mathbf{S}_\epsilon$ is homeomorphic to the $d(r+1)$-dimensional sphere. Given $z\in\mathbf{S}_\epsilon$, there exists a unique positive $t(z)$ such that $y+t(z)z\in\partial_1 S=\mathcal{A}\cap\partial\Omega$ (recall that S is a convex compact set). The map $\psi:\mathbf{S}_\epsilon\to V, \psi(z)=y+t(z)z$ is clearly continuous. Since $\psi(\mathbf{S}_\epsilon)\subset\partial_1 S$, we have $\mathrm{rank}\psi(z)=r+1$ for any $z\in\mathbf{S}_\epsilon$.

We need the following lemma.

LEMMA 1. *Let V be a simple Euclidean Jordan algebra,* $\mathrm{rank}(V)=l$. *Suppose that* $0<s<l$ *and*

$$\bar\Omega_s=\{x\in\bar\Omega:\mathrm{rank}(x)=s\}.$$

For $x\in\bar\Omega_s$ consider the spectral decomposition

$$x=\sum_{j=1}^{k+1}\lambda_j(x)f_j(x),$$

where $\lambda_{k+1}=0$. The map $\gamma_s=\bar\Omega_s\to V, \gamma_s(x)=f_{k+1}(x)$ is continuous.

We postpone the proof of the lemma and continue with the proof of the theorem.

Consider $\tilde\psi:\mathbf{S}_\epsilon\to V, \tilde\psi(z)=\gamma_{r+1}(\psi(z))$. Since $\psi(\mathbf{S}_\epsilon)\subset\bar\Omega_{r+1}$, the map $\tilde\psi$ is continuous. Notice that $\tilde\psi(\mathbf{S}_\epsilon)\subset\mathcal{T}(V)$ (the manifold of primitive idempotents in V). Indeed, let $\psi(z)=\sum_{j=1}^{k+1}\lambda_j f_j$ be spectral decomposition and $\lambda_{k+1}=0$. Since $\mathrm{rank}(\psi(z))=r+1$, we have $\sum_{j=1}^{k}\mathrm{tr}(f_j)=r+1$.

But $\sum_{j=1}^{k+1}f_j=e$. Hence, $\sum_{j=1}^{k+1}\mathrm{tr}(f_j)=\mathrm{tr}(e)=\mathrm{rank}(V)=r+2$.

Thus, $\mathrm{tr}f_{k+1}=1$, i.e., $f_{k+1}\in\mathcal{T}(V)$. Notice that $\dim\mathcal{T}(V)=\mathrm{d}(r+1)$ (see exercise 4a, p. 78 in [FK]). Hence $\tilde\psi:\mathbf{S}_\epsilon\to\mathcal{T}(V)$ is a continuous map between two compact connected manifolds of the same dimension. But then $\tilde\psi$ cannot be injective. Indeed, if $\tilde\psi$ is injective, then $\tilde\psi$ should be a homeomorphism of $\mathbf{S}_\epsilon$ onto $\mathcal{T}(V)$ (see, e.g., Corollary 28.4, p. 172 in [Ha]). However, under our assumptions $\mathcal{T}(V)$ is not homeomorphic to a sphere. Indeed, we assume that $r\ge 1$ and $\mathrm{rank}(V)=r+2$, i.e., $\mathrm{rank}(V)\ge 3$. (Notice that if $\mathrm{rank}(V)=2$, then $\mathcal{T}(V)$ is homeomorphic to a sphere.) We need to consider two separate cases.

If $d = 1$, then $\mathcal{T}(V)$ is homeomorphic to $\mathbf{P}_{r+1}(\mathbf{R})$ (see exercise 5, p. 99 in [FK]), which is not homeomorphic to sphere for $r \geq 1$.

If $d > 1, r \geq 3$ (the only possible choices are $d = 2, d = 4, d = 8$), then according to [H, p. 351], the following holds. Denote by $b_i(\mathcal{T}(V))$, $i = 0, 1, \ldots, d(r+1)$, the Betti numbers of $\mathcal{T}(V)$. Then

$$b_i(\mathcal{T}(V)) = \begin{cases} 1 & \text{when } i = 0(\text{mod } d), \\ 0 & \text{otherwise,} \end{cases}$$

whereas $b_i(\mathbf{S}_\epsilon) = 1, i = 0, i = d(r+1)$, and $b_i(\mathbf{S}_\epsilon) = 0$ otherwise. It is then clear that, say, $\beta_d(\mathcal{T}(V)) = 1$ but $\beta_d(\mathbf{S}_\epsilon) = 0$ (recall that $r \geq 1$!). Thus under our assumption, there exist $z_1 \neq z_2$ in $\mathbf{S}_\epsilon$ such that $\tilde{\psi}(z_1) = \tilde{\psi}(z_2)$. Since $z_1 \neq z_2$, it is clear that $\psi(z_1) \neq \psi(z_2)$. Let $c = \tilde{\psi}(z_1) = \tilde{\psi}(z_2)$. According to our construction of c, we have that $\psi(z_1), \psi(z_2) \in V(0, c)$ and, moreover, if Ω_0 is the cone of invertible squares in $V(0, c)$, then both $\psi(z_1)$ and $\psi(z_2) \in \Omega_0$ (see Theorem 2). Let L be a line passing through $\psi(z_1)$ and $\psi(z_2)$. It is clear that $L \subset \mathcal{A}$, and since Ω_0 does not contain lines, L hits its relative boundary at some point z_0. Then $\text{rank}(z_0) < \text{rank}(\psi(z_1)) = \text{rank}(\psi(z_2)) = r + 1$. Clearly, $z_0 \in \bar{\Omega}_0 \cap \mathcal{A} \subset S$. Thus, $\text{rank}(z_0) \leq r$.

(iii) Consider now a general case, where $\text{codim}_V \mathcal{A} \leq \varphi_d(r+1)$, $\dim V \geq r+2, r \geq 1$. (We still assume that $\mathcal{A} \cap \Omega \neq \emptyset$.) If $\text{codim}_V \mathcal{A} < \varphi_d(r+1)$, then the result follows from Theorem 3. It suffices to consider the case $\text{codim}_V \mathcal{A} = \varphi_d(r+1)$. By Theorem 3 there exists $y \in S$ such that $\text{rank}(y) \leq r + 1$. If $\text{rank}(y) < r + 1$, we are done. Consider the case $\text{rank}(y) = r + 1$. Let

$$y = \sum_{j+1}^{k+1} \lambda_j f_j(y)$$

be spectral decomposition of y and $\lambda_{k+1} = 0$. Let $f_{k+1}(y) = c_1 + \cdots + c_s$, where $c_1, \ldots, c_s$ are primitive pairwise orthogonal idempotents.

$$\text{rank}(f_{k+1}(y)) = \text{rank}(V) - \text{rank}(y) = \text{rank}(V) - (r+1) \geq 1.$$

Thus $s = \text{rank}(f_{k+1}(y)) \geq 1$. Let $W = V(0, c_1 + c_2 + \cdots + c_{s-1})$. Notice that $\langle y, f_{k+1}(y) \rangle = 0$ implies $\langle y, c_i \rangle = 0$, $i = 1, 2, \ldots, s$. Hence $y \in W$. Further, $\text{rank}(W) = \text{rank}(y) + 1 = r + 2$. Notice that $\text{codim}_W(\mathcal{A} \cap W) \leq \text{codim}_V(\mathcal{A}) = \varphi_d(r+1)$ (as we saw in case (i)).

Let Ω_W be the cone of invertible squares in W. Since $y \in \bar{\Omega}_W \cap (\mathcal{A} \cap W) = S \cap W$, the result follows from Theorem 3 if $\text{codim}_W(\mathcal{A} \cap W) < \text{codim}_V(\mathcal{A})$, or we are in case (ii) if $\text{codim}_W(\mathcal{A} \cap W) = \text{codim}_V \mathcal{A} = \varphi_d(r+2)$. This completes the proof of the theorem. □

Proof of Lemma 1. Since $\bar{\Omega}_s$ is an orbit of connected component of the group of automorphisms of Ω (see Proposition IV.3.1 (iii) in [FK]), it is a smooth connected submanifold in V.

Let $x \in \bar{\Omega}_s$ have a spectral decomposition

$$x = \sum_{j=1}^{k+1} \lambda_j f_j(x),$$

$\lambda_{k+1} = 0$. Using the Peirce decomposition associated with (complete) system of orthogonal idempotents $f_1(x), \ldots, f_{k+1}(x)$ (see section IV.2 in [FK]), one can easily see

that

$$\mathrm{Im}P(x) = V(1, f_1(x) + f_2(x) + \cdots + f_k(x)) = V(0, f_{k+1}(x)).$$

Recall that $P(x)$ is the quadratic representation of x. But $\dim \mathrm{Im}P(x) = \mathrm{rank}(P(x))$ (rank of the $\mathbf{R}$-linear map $P(x) : V \to V$) and $\mathrm{rank}(P(x))$ is constant when x varies over $\bar{\Omega}_s$ (see Proposition IV.3.1 (iv) in [FK]). Since $\bar{\Omega}_s$ is connected and the map $x \to P(x)$ is continuous, we conclude that the map $x \to \mathrm{Im}P(x)$ is continuous (see Proposition 13.6.1, p. 408 in [GLR]). Let $\pi(x) : V \to \mathrm{Im}P(x)$ be an orthogonal projection (with respect to the canonical scalar product $\langle , \rangle$). The continuity of the map $x \to \mathrm{Im}P(x)$ is equivalent to the continuity of the map $x \to \pi(x)$ (see [GLR, Chapter 13]). But $\pi(x) = P(f_1(x) + f_2(x) + \cdots + f_k(x))$ (see [FK, p. 65]). On the other hand, $P(f_1(x) + f_2(x) + \cdots + f_k(x))e = f_1(x) + \cdots + f_k(x) = e - f_{k+1}(x)$, which implies that the map $x \to f_{k+1}(x) = \gamma_s(x)$ is continuous.

Remark. If V is the algebra of symmetric matrices with real entries, Theorem 4 coincides with Theorem 1.2 in [B3].

5. Some applications. The natural question within the Jordan-algebraic approach developed here concerns the convexity of the image of the manifold $\mathcal{T}(V)$ of primitive idempotents (or its conic hull) under linear maps. We will show how to transform it to the setting of quadratic maps in the section 6.

PROPOSITION 3. *Let V be a simple Euclidean Jordan algebra of degree d. Given $a_1, \ldots, a_k \in V$, consider the linear map*

$$N : V \to \mathbf{R}^k,$$

$$N(x) = \begin{bmatrix} \langle a_1, x \rangle \\ \vdots \\ \langle a_k, x \rangle \end{bmatrix}. \tag{4}$$

If $\varphi_d^{-1}(k) < 2$, then

$$N\left(\bigcup_{\lambda \geq 0} \lambda \mathcal{T}(V)\right) = N(\bar{\Omega}).$$

In particular, $N(\bigcup_{\lambda \geq 0} \lambda \mathcal{T}(V))$ is a convex cone.

Proof. Denote $\bigcup_{\lambda \geq 0} \lambda \mathcal{T}(V)$ by K. Since $K \subset \bar{\Omega}$, it is clear that $N(K) \subset N(\bar{\Omega})$. Let $b = (b_1, \ldots, b_k)^T \in N(\bar{\Omega})$. Then there exists $x \in \bar{\Omega}$ such that $N(x) = b$. Consider $S = \{y \in \bar{\Omega} : N(y) = b\}$. It is clear that $x \in S$, i.e., $S \neq \emptyset$. According to Theorem 3, there exists $z \in S$ such that $\varphi_d(\mathrm{rank}(z)) \leq k$ or (using monotonicity of φ_d) $\mathrm{rank}(z) \leq \varphi_d^{-1}(k) < 2$. Hence, $\mathrm{rank}(z) = 1$ or $\mathrm{rank}(z) = 0$. In both cases, it is clear that $z \in K$. Thus $b \in N(K)$. □

PROPOSITION 4. *Let V be a simple Euclidean Jordan algebra, $d(V) = d$, $\mathrm{rank}(V) \geq 3$. Suppose that $k = \varphi_d(2), a_1, \ldots, a_k \in V$ are such that there exist $\tau_1, \ldots, \tau_k \in \mathbf{R}$ with the property*

$$\sum_{i=1}^{k} \tau_i a_i \in \Omega. \tag{5}$$

Then

$$N\left(\bigcup_{\lambda\geq 0}\lambda\mathcal{T}(V))=N(\bar{\Omega}\right) \text{ is a closed convex cone.}$$

Proof. In the notation of the proof of Proposition 3, it is clear that $N(K)\subset N(\bar{\Omega})$. Let $b\in N(\bar{\Omega})$, i.e., there exists $x\in\bar{\Omega}$ such that $N(x)=b$. The set $S=\{y\in\bar{\Omega}: N(y)=b\}$ is nonempty. Moreover, it is bounded. Indeed,

$$S\subset\left\{y\in\bar{\Omega}:\left\langle\sum_{i=1}^{k}\tau_i a_i,y\right\rangle=\sum_{i=1}^{k}\tau_i b_i\right\}=T.$$

Since $\sum_{i=1}^{k}\tau_i a_i\in\Omega$, the set T is bounded (see Corollary I.1.6 in [FK]). Hence, S is bounded. By Theorem 4 (with $r=1$) there exists $z\in\bar{\Omega}$ such that $\operatorname{rank}(z)\leq 1$. Hence, $z\in K$, i.e., $N(K)=N(\bar{\Omega})$. The closeness of $N(\bar{\Omega})$ immediately follows from the fact that $\operatorname{Ker}N\cap\bar{\Omega}=0$, which in turn easily follows from (5). □

PROPOSITION 5. *Let $d(V)=d$, let $\operatorname{rank}(V)\geq 3$, let N be defined as in* (4)*, and let $k=\varphi_d(2)-1$. Then $N(\mathcal{T}(V))$ is convex.*

Proof. Consider $\tilde{N}:V\to\mathbf{R}^{k+1}$,

$$\tilde{N}(x)=\begin{bmatrix}\langle a_1,x\rangle\\ \vdots\\ \langle a_k,x\rangle\\ \operatorname{tr}(x)=\langle e,x\rangle\end{bmatrix}.$$

By Proposition 4 $\tilde{N}(\bigcup_{\lambda\geq 0}\lambda\mathcal{T}(V))=\tilde{N}(\bar{\Omega})$. Indeed, (5) is clearly satisfied if we take $\tau_1=\tau_2=\cdots=\tau_k=0,\tau_{k+1}=1$. Consider $H=\{(b_1,\dots,b_{k+1})^T\in\mathbf{R}^{k+1}:b_{k+1}=1\}$. H is a hyperplane in $\mathbf{R}^{k+1}$ and $H\cap\tilde{N}(\bar{\Omega})$ is convex. Denote by $\tilde{N}_1$ the restriction of $\tilde{N}$ on $(\bigcup_{\lambda\geq 0}\lambda\mathcal{T}(V))$. It is clear that

$$\tilde{N}_1^{-1}(H\cap\tilde{N}(\bar{\Omega}))=\left\{z\in\bigcup_{\lambda\geq 0}\lambda\mathcal{T}(V):\operatorname{tr}(z)=1\right\}=\mathcal{T}(V).$$

This means that

$$N(\mathcal{T}(V))=H\cap\tilde{N}(\bar{\Omega}).\qquad\square$$

PROPOSITION 6. *Let $d(V)=d$, $\operatorname{rank}(V)\geq 3$, $k<\varphi_d(2)$. Let $c,a_i,i=1,2,\dots,k$, in V be such that there exist $\tau,\tau_i,i=1,2,\dots,k$, in $\mathbf{R}$ with the property*

$$\tau c+\sum_{i=1}^{k}\tau_i a_i\in\Omega.$$

Further, let $b_i,i=1,2,\dots,k$, be in $\mathbf{R}$.

Then

$$\inf\left\{\langle c,x\rangle:\langle a_i,x\rangle=b_i,i=1,2,\dots,k,\quad x\in\bigcup_{\lambda\geq 0}\lambda\mathcal{T}(V)\right\}$$
$$=\inf\{\langle c,x\rangle:\langle a_i,x\rangle=b_i,i=1,2,\dots,k,\quad x\in\bar{\Omega}\}.$$

Proof. We assume that the set $S = \{x \in \bar{\Omega} : \langle a_i, x\rangle = b_i, i = 1, 2, \ldots, k\}$ is not empty. Otherwise, there is nothing to prove.

Let $y \in S, \langle c, y\rangle = t$. Consider the map $N : \bar{\Omega} \to \mathbf{R}^{k+1}$,

$$N(x) = \begin{bmatrix} \langle a_1, x\rangle \\ \vdots \\ \langle a_k, x\rangle \\ \langle c, x\rangle \end{bmatrix}.$$

By Proposition 4 $N(\bar{\Omega}) = N(\bigcup_{\lambda\geq 0} \lambda\mathcal{T}(V))$.

Since $(b_1, \ldots, b_k, t)^T \in N(\bar{\Omega})$, we have $(b_1, \ldots, b_k, t)^T \in N(\bigcup_{\lambda\geq 0} \lambda\mathcal{T}(V))$. The result follows. □

PROPOSITION 7. *Let $d(V) = d$, $\operatorname{rank}(V) \geq 3$. Further, let $c, a_i, i = 1, 2, \ldots, k, k < \varphi_d(2) - 1$. Suppose that the set*

$$S = \{x \in \bar{\Omega} : \langle a_i, x\rangle = b_i, i = 1, 2, \ldots, k, \operatorname{tr}(x) = 1\} \text{ is not empty.}$$

Then

$$\begin{aligned} &\min\{\langle c, x\rangle : x \in \mathcal{T}(V), \langle a_i, x\rangle = b_i, i = 1, 2, \ldots, k\} \\ &= \min\{\langle c, x\rangle : \langle a_i, x\rangle = b_i, i = 1, 2, \ldots, k, \operatorname{tr}(x) = 1, x \in \bar{\Omega}\}. \end{aligned}$$

Proof. Let $y \in S, \langle c, y\rangle = t$. Consider the map $N : V \to \mathbf{R}^{k+2}$,

$$N(z) = \begin{bmatrix} \langle a_1, z\rangle \\ \vdots \\ \langle a_k, z\rangle \\ \langle e, z\rangle = \operatorname{tr}(z) \\ \langle c, z\rangle \end{bmatrix}.$$

We have $N(\bar{\Omega}) \cap \{(d_1, \ldots, d_{k+2})^T \in \mathbf{R}^{k+2} : d_{k+1} = 1\} = N(\mathcal{T}(V))$ by Proposition 5 (or more precisely its proof). It is clear that $(b_1, \ldots, b_k, 1, t) \in N(\bar{\Omega}) \cap \{(d_1, \ldots, d_{k+2})^T \in \mathbf{R}^{k+2} : d_{k+1} = 1\}$. The result follows. □

PROPOSITION 8. *Let $d(V) = d$, $r \geq 1$, and $1 \leq k < \varphi_d(r + 1)$ be such that $\operatorname{rank}(V) \geq r + 2$. Let $a_1, \ldots, a_k \in V$. Consider the map N described as in (4). Then every point of convex hull $\operatorname{conv}(N(\mathcal{T}(V)))$ can be represented as a convex combination of r (not necessarily distinct) points of $N(\mathcal{T}(V))$.*

Proof. Let $b = (b_1, \ldots, b_k)^T \in \operatorname{conv}(N(\mathcal{T}(V)))$. Thus

$$b = \sum_{i=1}^{m} \lambda_i N(x_i)$$

for some $m \geq 1$, $x_i \in \mathcal{T}(V)$, $\lambda_i \geq 0$, $\sum_{i=1}^m \lambda_i = 1$. Let $x = \sum_{i=1}^m \lambda_i x_i$. It is clear that $x \in \bar{\Omega}$, $N(x) = b$, $\operatorname{tr}(x) = \sum_{i=1}^m \lambda_i \operatorname{tr}(x_i) = 1$.

It is clear that $S = \{z \in \bar{\Omega} : N(x) = b, \operatorname{tr}(x) = 1\}$ is nonempty and bounded. By Theorem 4 there exist $z \in S$ such that $\operatorname{rank}(z) \leq r$. Let $\mu_1, \ldots, \mu_t, f_1, \ldots, f_t, t = \operatorname{rank}(V)$ be such that $f_1, \ldots, f_t$ is a Jordan frame and

$$z = \sum_{s=1}^{t} \mu_s f_s.$$

We notice earlier that $\operatorname{rank}(z) \leq r$ is equivalent to $\operatorname{card}\{s \in [1,t] : \mu_s > 0\} \leq r$.

Let $J = \{s \in [1,t] : \mu_s > 0\}$. We have

$$z = \sum_{s \in J}^{t} \mu_s f_s.$$

Since $\operatorname{tr}(z) = \sum_{s \in J} \mu_s = 1$, and $f_s \in \mathcal{T}(V)$ for all s, we conclude that

$$N(z) = b, \quad N(z) = \sum_{s \in J} \mu_s N(f_s).$$

The result follows. □

The next proposition can be interpreted as an abstract version of the well-known S-lemma (see, e.g., [BN]).

PROPOSITION 9. *Let $c, a_i, i = 1, 2, \ldots, k$, in V be such that $N(\mathcal{T}(V))$ is a convex set. Here $N : V \to \mathbf{R}^{k+1}$,*

$$N(x) = \begin{bmatrix} \langle a_1, x \rangle \\ \vdots \\ \langle a_k, x \rangle \\ \langle c, x \rangle \end{bmatrix}.$$

Suppose that there exists $x_0 \in \mathcal{T}(V)$ such that $\langle a_i, x_0 \rangle > 0$ for $i = 1, \ldots, k$. Further, let

$$\Gamma = \{x \in \mathcal{T}(V) : \langle a_i, x \rangle \geq 0, i = 1, \ldots, k\}.$$

Then $\langle c, x \rangle \geq 0$, for all $x \in \Gamma$ if and only if there exist nonnegative $\lambda_1, \ldots, \lambda_k$ such that

$$c - \sum_{i=1}^{k} \lambda_i a_i \in \bar{\Omega}.$$

Proof. We prove the (nontrivial) "only if" part. Let

$$Y = N\left(\bigcup_{\lambda \geq 0} \lambda \mathcal{T}(V)\right)$$

and

$$Z = \{z \in \mathbf{R}^{k+1} : z_i \geq 0, i = 1, \ldots, k, z_{k+1} < 0\}.$$

Then by our assumptions $Y \bigcap Z = \emptyset$. Both Y and Z are convex. Hence, by the separation theorem there exist real $\mu_1, \ldots, \mu_k, \lambda$ not all equal to zero, and real a such that

$$\sum_{i=1}^{k} \mu_i y_i + \lambda y_{k+1} \geq a \quad \text{for all } y \in Y,$$

$$\sum_{i=1}^{k} \mu_i z_i + \lambda z_{k+1} \leq a \quad \text{for all } z \in Z.$$

The standard reasoning then shows that $\mu_i \le 0$ for all i, $\lambda \ge 0$, and $a = 0$. Let us show that $\lambda > 0$. If $\lambda = 0$, then

$$\sum_{i=1}^{k} \mu_i y_i \ge 0$$

for all $y \in Y$, i.e.,

$$\left\langle \sum_{i=1}^{k} \mu_i a_i, x \right\rangle \ge 0$$

for any $x \in \mathcal{T}(V)$. This implies

$$\sum_{i=1}^{k} \mu_i a_i \in \bar{\Omega}.$$

By our assumptions there exists $x_0 \in \mathcal{T}(V)$ such that $\langle a_i, x_0 \rangle > 0$ for all i. We arrive at the contradiction, since all μ_i are nonpositive and not equal to zero simultaneously. Hence, $\lambda > 0$. But then

$$c - \sum_{i=1}^{k} \lambda_i a_i \in \bar{\Omega}$$

for $\lambda_i = -\mu_i/\lambda$. □

6. Interpretation in terms of quadratic mappings. To interpret the results of section 5 in terms of quadratic mappings, we need to understand the structure of manifolds $\mathcal{T}(V)$ for various simple Euclidean Jordan algebras. If $\operatorname{rank}(V) \ge 3$, every such algebra is of the type $\operatorname{Herm}(m, A)$, where $A = \mathbf{R}, \mathbf{C}, \mathbf{H}$, or $\mathbf{O}$. Here $\mathbf{R}, \mathbf{C}, \mathbf{H}, \mathbf{O}$ are algebras of real, complex, quaternion, and octonion numbers, respectively, and $\operatorname{Herm}(m, A)$ stands for the Jordan algebra of Hermitian matrices of size $m \times m$ with entries in A. Notice that if $A = \mathbf{O}, m \le 3$. The Jordan-algebraic multiplication in all these cases is the same:

Given $C, D \in \operatorname{Herm}(m, A)$,

$$C \circ D = \frac{CD + DC}{2},$$

where CD is the usual matrix multiplication. The list of corresponding manifolds $\mathcal{T}(V)$ is given on p. 99 of [FK]. We now consider the situation for concrete series $\operatorname{Herm}(m, A)$.

Case 1. Let $A = \mathbf{R}$. In this case the Jordan-algebraic operator tr coincides with the usual operator Tr of the matrix.

Thus

$$\langle C, D \rangle = \operatorname{Tr}(CD), \quad C, D \in \operatorname{Herm}(m, \mathbf{R}),$$

and $\operatorname{Herm}(m, \mathbf{R})$ is the algebra of $m \times m$ symmetric matrices with real entries.

$$\mathcal{T}(V) = \{C \in \operatorname{Herm}(m, \mathbf{R}) : C^2 = C, \operatorname{Tr}(C) = 1\},$$

i.e., $\mathcal{T}(V)$ is a manifold of one-dimensional orthogonal projections. Consider the map $\mu : \mathbf{R}^m \to \mathcal{T}(V)$, $\mu(x) = xx^T$. It is very well known that $\mu(\mathbf{S}^{m-1}) = \mathcal{T}(V)$.

PROPOSITION 10. *Let $q_i(x) = x^T C_i x$, $i = 1, 2$, be two quadratic forms on $\mathbf{R}^m$. Here $C_1, C_2 \in \mathrm{Herm}(m, \mathbf{R})$. Consider the map $\nu : \mathbf{R}^m \to \mathbf{R}^2$, $\nu(x) = (q_1(x), q_2(x))^T$. Then $\nu(\mathbf{R}^m)$ is a convex cone in $\mathbf{R}^2$*

Proof. It is clear that $\mu(\mathbf{R}^n) = \bigcup_{\lambda \geq 0} \lambda \mathcal{T}(V)$. We are going to use Proposition 3. Notice that $d(\mathrm{Herm}(m, \mathbf{R})) = 1$ for any $m \geq 2$. Hence, $\varphi_d(2) = 3$. Thus for $k = 2$ the image $N(\bigcup_{\lambda \geq 0} \lambda \mathcal{T}(V))$ is convex. In our case

$$N \circ \mu(x) = (x^T C_1 x, x^T C_2 x).$$

The result follows. □

Remark. Proposition 10 is a classical theorem of Dines [D]. Similarly to Proposition 4 we obtain the following result.

PROPOSITION 11. *Let $q_i(x) = x^T C_i x$, $i = 1, 2, 3$, be three quadratic forms on $\mathbf{R}^m$. Here $C_1, C_2, C_3 \in \mathrm{Herm}(m, \mathbf{R})$ are such that there exist real τ_1, τ_2, τ_3 with the property that*

$$\tau_1 C_1 + \tau_2 C_2 + \tau_3 C_3 > 0$$

(i.e., the corresponding matrix is positive definite). Then for $m \geq 3$ the image $\nu(\mathbf{R}^m)$ is a convex closed cone.

Remark. Notice that $\mathrm{rank}(\mathrm{Herm}(m, \mathbf{R})) = m$. The result of Proposition 11 is central in [Pol].

Similarly, Proposition 5 yields Brickman's classical theorem [Br].

Case 2. Let $A = \mathbf{C}$. Notice that $\mathrm{rank}(\mathrm{Herm}(m, \mathbf{C})) = m$, $d(\mathrm{Herm}(m, \mathbf{C})) = 2$. In this case $\mathrm{tr}(C) = \mathrm{Tr}(C)$, where Tr is the usual matrix trace. Hence, for the canonical scalar product we obtain

$$\langle C, D \rangle = \mathrm{Tr}(CD), \quad C, D \in \mathrm{Herm}(m, \mathbf{C}).$$

We have

$$\mathcal{T}(\mathrm{Herm}(m, \mathbf{C})) = \{C \in \mathrm{Herm}(m, \mathbf{C}) : C^2 = C, \mathrm{Tr}(C) = 1\}.$$

Once again $\mathcal{T}(V)$ in this case is the manifold of orthogonal projections on (complex) one-dimensional subspaces in $\mathbf{C}^m$. The map $\mu : \mathbf{C}^m \to \mathrm{Herm}(m, \mathbf{C})$, $\mu(x) = xx^*$, where $x^* = \bar{x}^T$ maps the unit sphere $\mathbf{S}^{2m-1}$ onto $\mathcal{T}(\mathrm{Herm}(m, \mathbf{C}))$. Notice that in this case $\varphi_d(2) = 4$ and all propositions from section 5 admit natural interpretation. For example, Proposition 3 leads to the following result.

PROPOSITION 12. *Let $q_i(x) = x^* C_i x$, $x \in \mathbf{C}^m$, $i = 1, 2, 3$, be three Hermitian forms. Consider the map $\nu : \mathbf{C}^m \to \mathbf{R}^3$,*

$$\nu(x) = (q_1(x), q_2(x), q_3(x)).$$

Then $\nu(\mathbf{C}^m)$ is a convex cone.

Remark. This result is also known (see, e.g., [Pol]).

Proposition 5 takes the following form.

PROPOSITION 13. *Under the assumption of Proposition* 12 *let $m \geq 3$. Then*

$$\nu(\mathbf{S}^{2m-1}) \text{ is convex.}$$

Here $\mathbf{S}^{2m-1} = \{x \in \mathbf{C}^m : x^*x = 1\}$.

Remark. This result is in [AP].

Here is an interpretation of Proposition 8.

PROPOSITION 14. *Let* r, m, k *be such that* $r \geq 1$, $m \geq r+2$, *and* $1 \leq k < \varphi_2(r+1) = (r+1)^2$. *Let* $C_1, \ldots, C_k \in \mathrm{Herm}(m, \mathbf{C})$. *Consider the map* $\nu : \mathbf{C}^m \to \mathbf{R}^k$, $\nu(x) = (x^*C_1x, x^*C_2x, \ldots, x^*C_kx)$. *Then every element of* $\mathrm{conv}(\nu(\mathbf{S}^{2m-1}))$ *can be represented as a convex combination of* r *(not necessarily distinct) points of the form* $\nu(x), x \in \mathbf{S}^{2m-1}$.

Remark. This result is essentially in [Poon].

The next proposition immediately follows from Proposition 6.

PROPOSITION 15. *Let* $C_0, \ldots, C_3 \in \mathrm{Herm}(m, \mathbf{C})$ *be such that*

$$\sum_{i=0}^{3} \tau_i C_i > 0$$

for some real τ_i. *Further, let* $m \geq 3$. *Consider the following quadratic optimization problem:*

$$q_0(x) \to \min, \quad q_i(x) = b_i, \quad i = 1, 2, 3, \ x \in \mathbf{C^m}.$$

Here b_i *are some real numbers. Further, consider its semidefinite relaxation:*

$$\mathrm{Tr}(C_0Y) \to \min, \quad \mathrm{Tr}(C_iY) = b_i, \quad i = 1, 2, 3, \quad Y \geq 0, \quad Y \in \mathrm{Herm}(m, \mathbf{C}).$$

Then this semidefinite relaxation is exact.

Case 3. Consider the case $A = \mathbf{H}$.

In principle, the same approach as that in Cases 1 and 2 works here. However, we prefer to work with complex Hermitian matrices. Notice that $d(\mathrm{Herm}(m, \mathbf{H})) = 4$, $\mathrm{rank}(\mathrm{Herm}(m, \mathbf{H})) = m$.

Let $J = \left[\begin{smallmatrix} 0 & I_m \\ -I_m & 0 \end{smallmatrix}\right]$. Consider a subalgebra

$$V = \{C \in \mathrm{Herm}(2m, \mathbf{C}) : JC = \bar{C}J\}.$$

It is shown in [FK] (see, in particular, p. 88 and exercise 1 of Chapter 3) that $\mathrm{Herm}(m, \mathbf{H})$ is isomorphic (as a Jordan algebra) to subalgebra V of $\mathrm{Herm}(2m, \mathbf{C})$.

Let $C = \left[\begin{smallmatrix} C_1 & C_2 \\ C_3 & C_4 \end{smallmatrix}\right]$ be a partition of a $2m \times 2m$ matrix with complex entries into four $m \times m$ blocks. Then $C \in V$ if and only if $C_1^* = C_1, C_4 = \bar{C}_1, C_2^T = -C_2$, and $C_3 = -\bar{C}_2$ (a direct computation). In other words, a typical element from V looks like this:

$$\left[\begin{array}{cc} C_1 & C_2 \\ -\bar{C}_2 & \bar{C}_1 \end{array}\right], \tag{6}$$

where $C_1^* = C_1, C_2^T = -C_2$.

We need to describe $\mathcal{T}(V)$.

LEMMA 2. *Let* $\xi \in \mathbf{C}^{2m}$ *be such that* $\xi^*\xi = 1$. *Consider* $f(\xi) = \xi\xi^* + (J\bar{\xi})(J\bar{\xi})^*$. *Then* $f(\xi) \in V$, $f(\xi)^2 = f(\xi)$.

Proof. The proof is a direct computation. □

LEMMA 3. *Let* $\xi_1, \ldots, \xi_m \in \mathbf{C}^{2m}$ *be such that* $\xi_i^*\xi_i = 1$, $i = 1, \ldots, m$, $\xi_i^*\xi_j = 0$, $\xi_i^*J\bar{\xi}_j = 0$ *for* $i \neq j$. *Then*

$$f(\xi_i) \circ f(\xi_j) = \delta_{ij} f(\xi_i), \qquad i, j = 1, 2, \ldots, m,$$

and

$$f(\xi_1) + \cdots + f(\xi_m) = I_{2m}.$$

Proof. Notice that under our assumptions $\xi_1, \ldots, \xi_m, J\bar{\xi}_1, \ldots, J\bar{xi}_m$ form an orthonormal basis in $\mathbf{C}^{2m}$. The result follows by a direct computation.

Since $\operatorname{rank}(V) = m$, we see that $f(\xi_1), \ldots, f(\xi_m)$ form a Jordan frame in V. In particular, $\operatorname{tr}(f(\xi)) = 1$ if $\xi^*\xi = 1$. Since $\operatorname{Tr}(f(\xi)) = 2$, we conclude that

$$\operatorname{tr}(X) = \frac{1}{2}\operatorname{Tr}(X), \ X \in V,$$

where Tr is the usual matrix trace. □

LEMMA 4.

$$\mathcal{T}(V) = \{f(\xi) : \xi \in \mathbf{S}^{4m-1}\}.$$

Proof. By Lemma 3 we have $f(\xi) \in \mathcal{T}(V)$ if $\xi^*\xi = 1$. The connected component of the identity of the group $O(V)$ of (Jordan-algebra) automorphisms of V acts transitively on $\mathcal{T}(V)$. In our case

$$O(V) = \{C \in \operatorname{Mat}(2m, \mathbf{C}) : C^* = C^{-1}, JC = \bar{C}J\}.$$

See p. 98 in [FK]. Let $C \in O(V)$. Then

$$C \cdot f(\xi) = C\xi\xi^*C^* + C(J\bar{\xi})(J\bar{\xi})^*C^* = (C\xi)(C\xi)^* + (J\overline{C\xi})(J\overline{C\xi})^* = f(C\xi)).$$

Notice that $C\xi \in \mathbf{S}^{4m-1}$. We see that $O(V)$ maps $\{f(\xi) : \xi \in \mathbf{S}^{4m-1}\}$ onto itself. Hence we have the result. □

Let us compute

$$\Delta = \operatorname{tr}(C \circ f(\xi)) = \langle C, f(\xi)\rangle \ \text{ for } C \in V.$$

We have

$$\Delta = \frac{1}{2}\operatorname{Tr}(Cf(\xi)) = \frac{1}{2}\operatorname{Tr}(\xi^*C\xi + (J\bar{\xi})^*C(J\bar{\xi})).$$

Now, $(J\bar{\xi})^*C(J\bar{\xi}) = -\bar{\xi}^*JCJ\bar{\xi} = -\bar{\xi}^*\bar{C}J^2\bar{\xi} = \bar{\xi}^*\bar{C}\bar{\xi} = \overline{\xi^*C\xi}$. But $\xi^*C\xi$ is real, since C is Hermitian. Hence, $\langle C, f(\xi)\rangle = \xi^*C\xi$.

We summarize our results in the following proposition.

PROPOSITION 16. *Consider a realization of* $\operatorname{Herm}(m, \mathbf{H})$ *in the form*

$$V = \{C \in \operatorname{Herm}(2m, \mathbf{C}) : JC = \bar{C}J\}.$$

Then

$$\mathcal{T}(V) = \{f(\xi) = \xi^*\xi + (J\bar{\xi})(J\bar{\xi})^* : \xi \in \mathbf{S}^{2m-1}\}.$$

Given $C \in V, \langle C, f(\xi)\rangle = \xi^*C\xi$.

In particular, we see that $\mu : \mathbf{C}^{2m} \to V$, $\mu(\xi) = \xi\xi^* + (J\bar{\xi})(J\bar{\xi})^*$ is such that $\mu(\mathbf{S}^{2m-1}) = \mathcal{T}(V)$.

We see now that Propositions 3–9 admit a natural interpretation in terms of convexity of images of families of quadratic forms. Notice that $\varphi_4(2) = 6$.

As an example, consider the reformulation of Proposition 5.

PROPOSITION 17. *Let $D_1, \dots, D_5$ be matrices of the form* (6). *Let $q_i(x) = x^* D_i x, x \in \mathbf{C^{2m}}$, $m \geq 3$, with $\mu : \mathbf{C}^{2m} \to \mathbf{R}^5$ defined as*

$$\mu(x) = (q_1(x), \dots, q_5(x))^T.$$

Then $\mu(\mathbf{S}^{4m-1})$ is convex.

Here $\mathbf{S}^{4m-1} = \{x \in \mathbf{C}^{2m} : x^* x = 1\}$.

Let now $D_0, \dots, D_5$ be matrices of the form (6) and such that

$$\sum_{i=0}^{5} \tau_i D_i > 0$$

for some real τ_i. The next two propositions immediately follow from Propositions 6 and 9.

PROPOSITION 18. *Let $m \geq 3$. Consider the following quadratic optimization problem:*

$$q_0(x) \to \min, \quad q_i(x) = b_i, \quad i = 1, \dots, 5, \ x \in \mathbf{C}^{2m}.$$

Here b_i are some real numbers and $q_i(x) = x^* D_i x$. Consider, further, its semidefinite relaxation:

$$\operatorname{Tr}(D_0 Y) \to \min, \quad \operatorname{Tr}(D_i Y) = b_i, \quad i = 1, \dots, 5,$$

$$Y \geq 0, \quad Y \in \operatorname{Herm}(m, \mathbf{H})$$

(*i.e., Y is of the form* (6)). *Then this semidefinite relaxation is exact.*

PROPOSITION 19. *Let $m \geq 3$,*

$$\Gamma = \{x \in \mathbf{C^{2m}} : q_i(x) \geq 0, i = 1, \dots, 5\}.$$

Suppose that there exists $x_0 \in \mathbf{C^{2m}}$ such that $q_i(x) > 0, i = 1, \dots, 5$. Further, let $q_0(x) \geq 0$ for all $x \in \Gamma$. Then there exist real nonnegative $\lambda_1, \dots, \lambda_5$ such that

$$D_0 - \sum_{i=1}^{5} \lambda_i D_i \geq 0.$$

Remark. Notice that all results obtained from Propositions 3–9 for $A = \mathbf{H}$ seem to be new.

Case 4. Consider the last case $A = \mathbf{O}$, $V = \operatorname{Herm}(3, \mathbf{O})$. In this case $\operatorname{rank}(V) = 3, d(V) = 8, \dim V = 27$. Notice that $\varphi_8(2) = 10$.

Let

$$C = \begin{bmatrix} \xi_1 & x_3 & \bar{x_2} \\ \bar{x_3} & \xi_2 & x_1 \\ x_2 & \bar{x_1} & \xi_3 \end{bmatrix}, \quad D = \begin{bmatrix} \eta_1 & y_3 & \bar{y_2} \\ \bar{y_3} & \eta_2 & y_1 \\ y_2 & \bar{y_1} & \eta_3 \end{bmatrix} \in \operatorname{Herm}(3, \mathbf{O}). \tag{7}$$

Here $\xi_i, \eta_i \in \mathbf{R}$, $x_i, y_i \in \mathbf{O}, i = 1, 2, 3$.

Recall (see, e.g., [E]) that octonions can be identified with the pair of quaternions: $z \in \mathbf{O} \Leftrightarrow z = (z_1, z_2), z_1, z_2 \in \mathbf{H}$.

Moreover, if $t = (t_1, t_2) \in \mathbf{O}$, then

$$zt = (z_1 t_1 - \bar{z}_2 t_2, z_2 \bar{t}_1 + t_2 z_1),$$

$\bar{z} = (\bar{z}_1, -z_2)$. The trace operator tr coincides with the matrix Tr (see [FK, p. 88–90]). Hence,

$$\langle C, D \rangle = \mathrm{Tr}\ (C \circ D) = \mathrm{Tr}\ \frac{(CD + DC)}{2}.$$

A short computation with C, D in the form (7) yields

$$\langle C, D \rangle = \sum_{i=1}^{3} \xi_i \eta_i + \mathrm{Re} \left(\sum_{i=1}^{3} \bar{x}_i y_i + x_i \bar{y}_i \right) = \sum_{i=1}^{3} \xi_i \eta_i + 2 \sum_{i=1}^{3} \langle x_i, y_i \rangle_{\mathbf{O}}.$$

Here $\langle x, y \rangle_{\mathbf{O}} = \mathrm{Re}(x\bar{y})$, $x, y \in \mathbf{O}$. Notice that the last equality follows from $\mathrm{Re}(xy) = \mathrm{Re}(yx), x, y \in \mathbf{O}$. See Proposition V.1.2 in [FK].

As usual,

$$\mathcal{T}(V) = \{C \in V : C^2 = C, \mathrm{Tr}(C) = 1\}.$$

A direct computation (see, e.g., [BP]) yields the following.

PROPOSITION 20. *Let C be parameterized as in* (7). *Then* $\mathcal{T}(Herm(3, \mathbf{O})) = \{(x_1, x_2, x_3, \xi_1, \xi_2, \xi_3) \in \mathbf{O}^3 \times \mathbf{R}^3 : \xi_1 + \xi_2 + \xi_3 = 1,\ \xi_1 = \xi_1^2 + \|x_2\|^2 + \|x_3\|^2,\ \xi_2 = \xi_2^2 + \|x_1\|^2 + \|x_3\|^2,\ \xi_3 = \xi_3^2 + \|x_1\|^2 + \|x_2\|^2,\ \xi_1 \bar{x}_1 = x_2 x_3,\ \xi_2 \bar{x}_2 = x_3 x_1,\ \xi_3 \bar{x}_3 = x_1 x_2\}$.

Here $\|x\|^2 = x\bar{x} = \bar{x}x$.

Propositions 3–8 are the statements about the convexity of linear images of the manifold $\mathcal{T}(V)$ or its conic hull. Notice that $\dim \mathcal{T}(V) = d(V)(\mathrm{rank}(V) - 1) = 16$.

It is not so easy, however, to translate these results into ones concerning images of quadratic forms. Acting in analogy with the cases of $A = \mathbf{R}, \mathbf{C}, \mathbf{H}$, we need to consider the map $\mu : \mathbf{O}^3 \to \mathrm{Herm}(3, \mathbf{O})$,

$$\mu(d_1, d_2, d_3) = (d_i \bar{d}_j), \quad i, j = 1, 2, 3.$$

Let $\mathbf{S}^{23} = \{(d_1, d_2, d_3) \in \mathbf{O}^3 : \|d_1\|^2 + \|d_2\|^2 + \|d_3\|^2 = 1\}$.

Unfortunately, $\mu(\mathbf{S}^{23}) \neq \mathcal{T}(\mathrm{Herm}(3, \mathbf{O}))$. More precisely $\mathcal{T}(\mathrm{Herm}(3, \mathbf{O})) \subset \mu(\mathbf{S}^{23})$, but the inclusion is strict. Similar construction for $A = \mathbf{R}, \mathbf{C}, \mathbf{H}$ yields a coincidence of corresponding sets.

The problem is due to the fact that multiplication in $\mathbf{O}$ is not associative (see [BP] for details). Nevertheless, we have the following proposition.

PROPOSITION 21. *Let $\tilde{\mu}$ be the restriction of μ to $\mathbf{O} \times \mathbf{O} \times \mathbf{R}$. Then* $\tilde{\mu}(\mathbf{S}^{16}) = \mathcal{T}(V)$.

Here $\mathbf{S}^{16} = \{(d_1, d_2, \zeta) \in \mathbf{O} \times \mathbf{O} \times \mathbf{R} : \|d_1\|^2 + \|d_2\|^2 + \zeta^2 = 1\},\ \|d\| = \sqrt{d\bar{d}}, d \in \mathbf{O}$.

Proof. Let us show first that $\tilde{\mu}(\mathbf{S}^{16}) \subset \mathcal{T}(V)$. We simply need to check all conditions of Proposition 20.

We have that $\mu(d_1, d_2, \zeta) = (x_1, x_2, x_3, \xi_1, \xi_2, \xi_3)$ is equivalent to $\xi_i = \|d_i\|^2, i = 1, 2, 3,\ x_3 = d_1 \bar{d}_2,\ \bar{x}_2 = \zeta d_1, x_1 = \zeta d_2$. Let us check, for example, that $\xi_1 = \xi_1^2 + \|x_2\|^2 + \|x_3\|^2$. We have $\Delta = \xi_1^2 + \|x_2\|^2 + \|x_3\|^2 = \|d_1\|^4 + \zeta^2 \|d_1\|^2 + \|d_1\|^2 \|d_2\|^2$. Here we used $\|d_1 d_2\| = \|d_1\| \|d_2\|$. Hence, $\Delta = \|d_1\|^2 (\|d_1\|^2 + \zeta^2 + \|d_2\|^2) = \|d_1\|^2 = \xi_1$.

The other conditions are verified similarly. Let us show that $\tilde{\mu}(\mathbf{S}^{16}) \supset \mathcal{T}(V)$. Let $(\xi_1, \xi_2, \xi_3, x_1, x_2, x_3) \in \mathcal{T}(V)$. Consider, first, the case where $\xi_3 > 0$.

Take $d_1 = \frac{\bar{x_2}}{\sqrt{\xi_3}}$, $d_2 = \frac{x_1}{\sqrt{\xi_3}}$, $\zeta = \sqrt{\xi_3}$.

Notice that $\|d_1\|^2+\|d_2\|^2+\zeta^2 = \frac{\|x_2\|^2+\|x_1\|^2+\xi_3^2}{\xi_3} = 1$ because of the one of the defining relations for $\mathcal{T}(V)$. We can easily check that $\mu(d_1, d_2, \zeta) = (\xi_1, \xi_2, \xi_3, x_1, x_2, x_3)$.

Consider now the case $\xi_3 = 0$. Due to the condition $\xi_3 = \xi_3^2 + \|x_1\|^2 + \|x_2\|^2$, we obtain $x_1 = x_2 = 0$. Hence, we have

$$\xi_1 = \xi_1^2 + \|x_3\|^2, \quad \xi_2 = \xi_2^2 + \|x_3\|^2, \quad \xi_1 + \xi_2 = 1.$$

This system has two solutions:

$$\xi_1 = \frac{1}{2} \pm \frac{\sqrt{1 - 4\|x_3\|^2}}{2},$$

$$\xi_2 = \frac{1}{2} \mp \frac{\sqrt{1 - 4\|x_3\|^2}}{2},$$

provided $\|x_3\| \le \frac{1}{2}$. Take $d_1 = \frac{\sqrt{\xi_1}x_3}{\|x_3\|}, d_2 = \sqrt{\xi_2}, \zeta = 0$ if $x_3 \neq 0$. If $x_3 = 0$, then $\xi_1 = 0, \xi_2 = 1$ or $\xi_1 = 1, \xi_2 = 0$. In both cases take $d_1 = \sqrt{\xi_1}, d_2 = \sqrt{\xi_2}, \zeta = 0$. We easily check that $\mu(d_1, d_2, \eta) = (\xi_1, \xi_2, \xi_3, x_1, x_2, x_3)$. □

We identify $\mathbf{O}\times\mathbf{O}\times\mathbf{R}$ with $\mathbf{R}^{17}$. Consider on $\mathbf{R}^{17}$ quadratic forms of the following type:

$$\begin{aligned} f_{y,\eta}(d_1, d_2, \zeta) = {} & \eta_1\|d_1\|^2 + \eta_2\|d_2\|^2 + \eta_3\zeta^2 \\ & + 2\langle y_1, d_2\rangle_{\mathbf{O}}\zeta + 2\langle y_2, \bar{d_1}\rangle_{\mathbf{O}}\zeta + 2\langle y_3, d_1\bar{d_2}\rangle_{\mathbf{O}}\zeta. \end{aligned} \tag{8}$$

Here $y_1, y_2, y_3 \in \mathbf{O}, \eta_1, \eta_2, \eta_3 \in \mathbf{R}$.

Notice that if D is constructed from $(y_1, y_2, y_3, \eta_1, \eta_2, \eta_3)$ as in (7), then

$$f_{y,\eta}(d_1, d_2, \zeta) = \mathrm{Tr}(D \circ \tilde{\mu}(d_1, d_2, \zeta)).$$

We can now easily reformulate Propositions 3–8 in terms of quadratic forms $f_{y,\eta}$ on $\mathbf{R}^{17}$. For example, Proposition 3 leads to the following result. Notice that $\varphi_8(2) = 10$.

PROPOSITION 22. *Let $q_i, i = 1, 2, \ldots, 9$, be nine quadratic forms of type* (8) *on $\mathbf{R}^{17}$ identified with $\mathbf{O} \times \mathbf{O} \times \mathbf{R}$. Consider the map*

$$\nu(d_1, d_2, \zeta) = (q_1(d_1, d_2, \zeta), \ldots, q_9(d_1, d_2, \zeta)).$$

Then $\nu(\mathbf{R}^{17})$ is a convex cone.

7. Concluding remarks. In the present paper we have considered a large number of classical results related to the convexity of image of quadratic mappings in a general context of Euclidean Jordan algebras. The technique used is a generalization of semidefinite relaxation technique which has been used by Barvinok for similar purposes. Our context is more general and allows one to obtain convexity result corresponding to the series $\mathrm{Herm}(m, \mathbf{H})$ of Euclidean Jordan algebra and exceptional 27-dimensional algebra, which seem to be new. The present paper does not exhaust all possibilities offered by the Jordan-algebraic technique for the analysis of this circle of questions. We plan to address the remaining issues elsewhere.

REFERENCES

[AP] Y. H. Au-Yeung and Y. T. Poon, *A remark on convexity and positive definiteness concerning Hermitian matrices*, Southeast Asian Bull. Math., 3 (1979), pp. 85–92.
[B1] A. I. Barvinok, *A Course in Convexity*, AMS, Providence, RI, 2002.
[B2] A. I. Barvinok, *Problem of distance geometry and convex properties of quadratic maps*, Discrete Comput. Geom., 13 (1995), pp. 189–202.
[B3] A. I. Barvinok, *A remark on the rank of positive semidefinite matrices subject to affine constraints*, Discrete Comput. Geom., 25 (2001), pp. 23–31.
[BN] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization*, SIAM, Philadelphia, 2001.
[BP] C. Brada and F. Pecant-Tison, *Geometrie du plan projectif des octaves de Cayley*, Geom. Dedicata, 23 (1987), pp. 131–154.
[Br] L. Brickman, *On the field of values of a matrix*, Proc. Amer. Math. Soc., 12 (1961), pp. 61–66.
[BM] S. Burer and R. Monteiro, *Local minima and convergence in low-rank semidefinite programming*, Math. Program., 103 (2005), pp. 427–444.
[D] L. L. Dines, *On the mapping of quadratic forms*, Bull. Amer. Math. Soc., 47 (1941), pp. 494–498.
[E] H.-D. Ebbinghaus, H. Hermes, F. Hirzebruch, M. Koecher, K. Mainzer, J. Neukirch, A. Prestel, and R. Remmert, *Numbers*, Springer-Verlag, New York, 1991.
[FK] J. Faraut and A. Koranyi, *Analysis on Symmetric Cones*, Clarendon Press, New York, 1994.
[F1] L. Faybusovich, *Euclidean Jordan algebras and interior-point algorithms*, J. Positivity, 1 (1997), pp. 331–357.
[F2] L. Faybusovich, *Linear system in Jordan algebras and primal-dual interior-point algorithms*, J. Comput. Appl. Math., 86 (1997), pp. 149–175.
[GLR] I. Gohberg, P. Lancaster, and L. Rodman, *Invariant Subspace of Matrices with Applications*, John Wiley and Sons, New York, 1986.
[Ha] A. Hatcher, *Algebraic Topology*, Cambridge University Press, Cambridge, UK, 2002.
[H] U. Hirzebruch, *Über Jordan-Algebren and kompakte Riemannsche symmetrische Raiime vom Rang* 1, Math. Zeitschr., 90 (1965), pp. 339–354.
[Pat] G. Pataki, *On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues*, Math. Oper. Res., 23 (1998), pp. 339–358.
[Pol] B. T. Polyak, *Convexity of quadratic transformations and its use in control and optimization*, J. Optim. Theory Appl., 99 (1998), pp. 553–583.
[Poon] Y. T. Poon, *On the convex hull of the multiform numerical range*, Linear Multilinear Algebra, 37 (1994), pp. 221–223.
[LKF] Y. Lim, J. Kim, and L. Faybusovich, *Simultaneous diagonalization on simple Euclidean Jordan algebras and its applications*, Forum Math., 15 (2003), pp. 639–644.
[W] R. Webster, *Convexity*, Oxford University Press, New York, 1994.
[YZ] Y. Ye and S. Zhang, *New results on quadratic minimization*, SIAM J. Optim., 14 (2003), pp. 245–267.

SIAM J. OPTIM.
Vol. 17, No. 2, pp. 577–605
© 2006 Society for Industrial and Applied Mathematics

ACTIVE SET IDENTIFICATION IN NONLINEAR PROGRAMMING*

CHRISTINA OBERLIN† AND STEPHEN J. WRIGHT†

Abstract. Techniques that identify the active constraints at a solution of a nonlinear programming problem from a point near the solution can be a useful adjunct to nonlinear programming algorithms. They have the potential to improve the local convergence behavior of these algorithms and in the best case can reduce an inequality constrained problem to an equality constrained problem with the same solution. This paper describes several techniques that do not require good Lagrange multiplier estimates for the constraints to be available a priori, but depend only on function and first derivative information. Computational tests comparing the effectiveness of these techniques on a variety of test problems are described. Many tests involve degenerate cases, in which the constraint gradients are not linearly independent and/or strict complementarity does not hold.

Key words. nonlinear programming, active constraint identification, degeneracy

AMS subject classifications. 90C30, 90C46

DOI. 10.1137/050626776

1. Introduction.

Consider the following nonlinear programming problem:

$$\min f(x) \text{ subject to } h(x) = 0, \quad c(x) \leq 0, \tag{1.1}$$

where $x \in \mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$, $c : \mathbb{R}^n \to \mathbb{R}^m$, and $h : \mathbb{R}^n \to \mathbb{R}^p$ are twice continuously differentiable functions.

In this paper, we examine identification of active inequality constraints—the components of c for which equality holds at a local solution x^*—using information available at a point x near x^*. We focus on identification schemes that do not require good estimates of the Lagrange multipliers to be available a priori. Rather, in some cases, such estimates are computed as an adjunct to the identification technique. In most of our results, we relax the "standard" nondegeneracy assumptions at x^* to allow linearly dependent active constraint gradients and weakly active constraints. We consider three schemes that require solution of linear programs and one that requires solution of a mixed integer program. We analyze the effectiveness of these schemes and discuss computational issues of solving the linear and mixed-integer programs. Finally, we present results obtained on randomly generated problems and on degenerate problems from the CUTEr test set [13].

One area in which identification schemes are useful is in "EQP" approaches to sequential quadratic programming (SQP) algorithms, in which each iteration consists of an estimation of the active set followed by solution of an equality constrained quadratic program that enforces the apparently active constraints and ignores the apparently inactive ones. The "IQP" variant of SQP, in which an inequality constrained subproblem is solved (thereby estimating the active set implicitly), has been more widely studied in the past two decades, but the EQP variant has been revived recently by Byrd et al. [5, 6].

*Received by the editors March 13, 2005; accepted for publication (in revised form) February 26, 2006; published electronically August 16, 2006. Research supported by NSF grants ATM-0296033, CNS-0127857, CCF-0113051, SCI-0330538, DMS-0427689, CCF-0430504, CTS-0456694, and CNS-0540147, DOE grant DE-FG02-04ER25627, and an NSF Graduate Research Fellowship.
http://www.siam.org/journals/siopt/17-2/62677.html
†Computer Sciences Department, University of Wisconsin, 1210 W. Dayton Street, Madison, WI 53706 (coberlin@cs.wisc.edu, swright@cs.wisc.edu).

1.1. Assumptions and background. We describe here the notation and assumptions used in the remainder of the paper.

The Lagrangian for (1.1) is

$$\mathcal{L}(x,\mu,\lambda) = f(x) + \mu^T h(x) + \lambda^T c(x), \tag{1.2}$$

where $\mu \in \mathbb{R}^p$ and $\lambda \in \mathbb{R}^m$ are Lagrange multipliers. First-order necessary conditions for x^* to be a solution of (1.1), assuming a constraint qualification, are that there exist multipliers (μ^*, λ^*) such that

$$\nabla_x \mathcal{L}(x^*, \mu^*, \lambda^*) = 0, \tag{1.3a}$$

$$h(x^*) = 0, \tag{1.3b}$$

$$0 \geq c(x^*) \perp \lambda^* \geq 0, \tag{1.3c}$$

where the symbol $\perp$ denotes vector complementarity; that is, $a \perp b$ means $a^T b = 0$. We define the "dual" solution set as follows:

$$\mathcal{S}_D \stackrel{\text{def}}{=} \{(\mu^*, \lambda^*) \text{ satisfying (1.3)}\}, \tag{1.4}$$

while the primal-dual solution set $\mathcal{S}$ is

$$\mathcal{S} \stackrel{\text{def}}{=} \{x^*\} \times \mathcal{S}_D.$$

The set of *active inequality constraints* at x^* is defined as follows:

$$\mathcal{A}^* = \{i = 1, 2, \ldots, m \,|\, c_i(x^*) = 0\}.$$

The *weakly active inequality constraints* $\mathcal{A}_0^*$ are those active constraints i for which $\lambda_i^* = 0$ for all optimal multipliers (μ^*, λ^*); that is,

$$\mathcal{A}_0^* = \{i \in \mathcal{A}^* \,|\, \lambda_i^* = 0 \text{ for all } (\mu^*, \lambda^*) \in \mathcal{S}_D\}. \tag{1.5}$$

The constraints $\mathcal{A}^* \setminus \mathcal{A}_0^*$ are said to be the *strongly active inequalities*.

In this paper, we make use of the following two constraint qualifications at x^*. The linear independence constraint qualification (LICQ) is that

$$\{\nabla h_i(x^*),\ i = 1, 2, \ldots, p\} \cup \{\nabla c_i(x^*),\ i \in \mathcal{A}^*\} \text{ is linearly independent.} \tag{1.6}$$

The Mangasarian–Fromovitz constraint qualification (MFCQ) is that there is a vector $v \in \mathbb{R}^n$ such that

$$\nabla c_i(x^*)^T v < 0,\ \ i \in \mathcal{A}^*; \quad \nabla h_i(x^*)^T v = 0,\ \ i = 1, 2, \ldots, p, \tag{1.7a}$$

$$\{\nabla h_i(x^*),\ i = 1, 2, \ldots, p\} \text{ is linearly independent.} \tag{1.7b}$$

In some places, we use the following second-order sufficient condition: Defining

$$\begin{aligned} \mathcal{C} \stackrel{\text{def}}{=} \{v \,|\, & \nabla c_i(x^*)^T v = 0,\ i \in \mathcal{A}^* \setminus \mathcal{A}_0^*,\ \ \nabla c_i(x^*)^T v \leq 0,\ i \in \mathcal{A}_0^*, \\ & \nabla h_i(x^*)^T v = 0,\ i = 1, 2, \ldots, p\}, \end{aligned} \tag{1.8}$$

we require that

$$v^T \nabla^2_{xx} \mathcal{L}(x^*, \mu^*, \lambda^*) v > 0 \text{ for all } v \in \mathcal{C} \setminus \{0\} \text{ and all } (\mu^*, \lambda^*) \in \mathcal{S}_D. \tag{1.9}$$

The following notation is used for first derivatives of the objective and constraint functions at x:

$$g(x) = \nabla f(x), \quad J(x) = [\nabla h_i(x)^T]_{i=1,2,\ldots,p}, \quad A(x) = [\nabla c_i(x)^T]_{i=1,2,\ldots,m}.$$

We use $A_i(x) = \nabla c_i(x)^T$ to denote the ith row of $A(x)$, while for any index set $\mathcal{T} \subset \{1, 2, \ldots, m\}$, we use $A_{\mathcal{T}}(x)$ to denote the $|\mathcal{T}| \times n$ submatrix corresponding to $\mathcal{T}$. In some subsections, the argument x is omitted from the quantities $c(x)$, $A(x)$, $A_i(x)$, and $A_{\mathcal{T}}(x)$ if the dependence on x is clear from the context. In some instances, we also use ∇c_i^*, g^*, etc., to denote $\nabla c_i(x^*)$, $g(x^*)$, etc., respectively.

Given a matrix $B \in \mathbb{R}^{n\times q}$ we denote

$$\text{range}[B] = \{Bz \,|\, z \in \mathbb{R}^q\}, \quad \text{pos}[B] = \{Bz \,|\, z \in \mathbb{R}^q,\ z \geq 0\}.$$

The norms $\|\cdot\|_1$, $\|\cdot\|_2$, and $\|\cdot\|_\infty$ all appear in the paper. When the subscript is omitted, the Euclidean norm $\|\cdot\|_2$ is intended.

We use the usual definition of the distance function $\text{dist}\,(\cdot,\cdot)$ between sets, that is,

$$\text{dist}\,(\mathcal{S}_1, \mathcal{S}_2) = \inf_{s_1\in\mathcal{S}_1, s_2\in\mathcal{S}_2} \|s_1 - s_2\|. \tag{1.10}$$

(Distance between a point and a set is defined similarly.)

For a vector z, function $\max(z, 0)$ (defined componentwise) is denoted by z_+, while z_- denotes $\max(-z, 0)$. We use the notation e throughout the paper to denote the vector $(1, 1, \ldots, 1)^T$. (The dimension of e is not specified but is clear from the context.)

In assessing the accuracy of an active set estimate, a *false positive* is an index i that is identified as active by our scheme but which actually does not belong to $\mathcal{A}^*$, while a *false negative* is an index $i \in \mathcal{A}^*$ which is wrongly identified as inactive.

1.2. Related work. Some previous works have studied the behavior of nonlinear programming algorithms in identifying active constraint sets, more or less as a by-product of their progress toward a solution. Other papers have described the use of these active set estimates to speed the convergence of the algorithm in its final stages. We mention several works of both types here, in nonlinear programming and in the context of other optimization and complementarity problems.

Bertsekas [1] proposed a two-metric algorithm for minimizing a nonlinear function subject to bound constraints on the components of x. A key aspect of this method is estimation of the active bounds at the solution. (Different second-order scalings are applied to the apparently active components and the free components.) Strongly active constraints are identified for all feasible x in a neighborhood of x^*. The latter result is also proved by Lescrenier [16] for a trust-region algorithm.

Burke and Moré [3, 4] take a geometric approach, assuming the constraints to be expressed in the form $x \in \Omega$ for a convex set Ω. This set can be partitioned into faces, where a face F is defined to be a subset of Ω such that every line segment in Ω whose relative interior meets F is contained in F. In this context, active set identification corresponds to the identification of the face that contains the solution x^*. In [3], it is shown that "quasi-polyhedral" faces are identified for all x close to x^* provided that a geometric nondegeneracy condition akin to strict complementarity is satisfied. (Quasi-polyhedrality is defined in [3, Definition 2.5]; curved faces are not quasi-polyhedral.) Burke [2] takes a partly algebraic viewpoint and shows that the set

of active indices of a linear approximation to the problem at x near x^* are sufficient for the objective gradient to be contained in the cone of active constraint gradients—a result not unlike Theorem 3.2 below.

Wright [21] also uses a hybrid geometric-algebraic viewpoint and considers convex constraint sets Ω with (possibly curved) boundaries defined by (possibly nonlinear) inequalities. The concept of a "class-$\mathcal{C}^p$ identifiable surface" is defined, and it is shown that this surface is identified at all x close to x^* provided that a nondegeneracy condition is satisfied. Hare and Lewis [15] extend these concepts to nonconvex sets, using concepts of prox-regularity and partly smooth functions developed elsewhere by Lewis [17] and others.

Facchinei, Fischer, and Kanzow [10] describe a technique based on the algebraic representation of the constraint set that uses estimates of the Lagrange multipliers (μ, λ) along with the current x to obtain a two-sided estimate of the distance of (x, μ, λ) to the primal-dual solution set. This estimate is used in a threshold test to obtain an estimate of $\mathcal{A}^*$. We discuss this technique further in section 2.

Conn, Gould, and Toint [8, Chapter 12] discuss the case of convex constraints, solved with a trust-region algorithm in which a "generalized Cauchy point" is obtained via gradient projection. They prove that when assumptions akin to strict complementarity and LICQ hold at the solution x^*, their approach identifies the active set once the iterates enter a neighborhood of x^*; see, for example, [8, Theorem 12.3.8].

Active constraint identification has played an important role in finite termination strategies for linear programming (LP). Ye [25] proposed such a strategy, which determined the active set estimate by a simple comparison of the primal variables with the dual slacks. (An equality constrained quadratic program, whose formulation depends crucially on the active set estimate, is solved in an attempt to "jump to" an optimal point.) El-Bakry, Tapia, and Zhang [9] discuss methods based on "indicators" for identifying the active constraints for LP.

Similar active identification and finite termination strategies are available for monotone linear complementarity problems; see, for example, the paper of Monteiro and Wright [18]. For monotone nonlinear complementarity problems, Yamashita, Dan, and Fukushima [24] describe a technique for classifying indices (including degenerate indices) at the limit point of a proximal point algorithm. This threshold is defined similarly to the one in [10], while the classification test is similar to that of [18].

1.3. Organization of the paper. In section 2, we review a technique for identifying the active set using an estimate (x, μ, λ) of the primal-dual optimum. This technique provides the basis for the identification techniques of subsections 3.2 and 3.3. Section 3 describes the main techniques for identifying the active set without assuming that reliable estimates of the Lagrange multipliers (μ, λ) are available. Subsection 3.1 describes a technique used by Byrd et al. [5, 6] along with a dual variant; subsection 3.2 describes a technique based on minimizing the primal-dual measure of section 2, which can be formulated as a mixed integer program; and subsection 3.3 derives an LP approximation to the latter technique. In all cases, we prove results about the effectiveness of these schemes and discuss their relationship to each other. In section 4, we describe our implementation of the identification schemes and present results obtained on randomly generated problems (with controlled degeneracy) and on degenerate problems from the CUTEr test set. Some conclusions appear in section 5.

2. Identification from a primal-dual point. In this section, we suppose that along with an estimate x of the solution x^*, we have estimates of the Lagrange multi-

pliers (μ, λ). We describe a threshold test based on the function ψ defined as follows:

$$\psi(x,\mu,\lambda) = \left\| \begin{bmatrix} \nabla_x \mathcal{L}(x,\mu,\lambda) \\ h(x) \\ \min(\lambda, -c(x)) \end{bmatrix} \right\|_1, \tag{2.1}$$

where the $\min(\cdot,\cdot)$ is taken componentwise. (Other norms could be used in this definition, including weighted norms, but the ℓ_1 norm is convenient for computation in later contexts.) The test based on ψ provides the starting point for the LPEC (linear program with equilibrium constraints) scheme of subsection 3.2, where we fix x and choose (μ, λ) to minimize ψ, rather than assuming that (μ, λ) are given.

The following result shows that for (x, μ, λ) close to $\mathcal{S}$, this function provides a two-sided estimate of the distance to the solution. (See Facchinei, Fischer, and Kanzow [10, Theorem 3.6], Hager and Gowda [14], and Wright [22, Theorem A.1] for proofs of results similar or identical to this one.)

THEOREM 2.1. *Suppose the KKT conditions* (1.3)*, the MFCQ* (1.7)*, and the second-order condition* (1.9) *are satisfied at* x^**. There are constants* $\epsilon \in (0, 1]$ *and* $C > 0$ *such that, for all* (x, μ, λ) *with* $\lambda \geq 0$ *and* $\operatorname{dist}((x,\mu,\lambda), \mathcal{S}) \leq \epsilon$*, we have*

$$C^{-1}\psi(x,\mu,\lambda) \leq \operatorname{dist}((x,\mu,\lambda), \mathcal{S}) \leq C\psi(x,\mu,\lambda). \tag{2.2}$$

(The upper bound of 1 *in the definition of* ϵ *is needed to simplify later arguments.)*

For future reference, we define L to be a Lipschitz constant for the functions g, c, h, A, and J in the neighborhood $\|x - x^*\| \leq \epsilon$, for the ϵ given in Theorem 2.1. In particular, we have

$$\begin{aligned} &\|g(x) - g(x^*)\| \leq L\|x - x^*\|, \quad \|c(x) - c(x^*)\| \leq L\|x - x^*\|, \\ &\|A(x) - A(x^*)\| \leq L\|x - x^*\|, \quad \|h(x) - h(x^*)\| \leq L\|x - x^*\|, \\ &\|J(x) - J(x^*)\| \leq L\|x - x^*\|, \quad \text{for all } x \text{ with } \|x - x^*\| \leq \epsilon. \end{aligned} \tag{2.3}$$

We define a constant K_1 such that the following condition is satisfied:

$$K_1 = \max\left(\|c(x^*)\|_\infty, \max_{(\mu^*,\lambda^*)\in\mathcal{S}_D} \|(\mu^*, \lambda^*)\|_\infty \right) + 1. \tag{2.4}$$

(Note that finiteness of K_1 is assured under MFCQ.)

The active set estimate is a threshold test, defined as follows for a given parameter $\sigma \in (0, 1)$:

$$\mathcal{A}(x,\mu,\lambda) = \{i \mid c_i(x) \geq -\psi(x,\mu,\lambda)^\sigma\}. \tag{2.5}$$

The following result is an immediate consequence of Theorem 2.1. It has been proved in earlier works (see, for example, [10]), but since the proof is short and illustrative, we repeat it here.

THEOREM 2.2. *Suppose that the KKT conditions* (1.3)*, the MFCQ* (1.7)*, and the second-order condition* (1.9) *are satisfied at* x^**. Then there is* $\bar{\epsilon}_1 > 0$ *such that for all* (x, μ, λ) *with* $\lambda \geq 0$ *and* $\operatorname{dist}((x,\mu,\lambda), \mathcal{S}) \leq \bar{\epsilon}_1$*, we have that* $\mathcal{A}(x,\mu,\lambda) = \mathcal{A}^*$*.*

Proof. First set $\bar{\epsilon}_1 = \epsilon$, where ϵ is small enough to satisfy the conditions in Theorem 2.1. Taking any $i \notin \mathcal{A}^*$, we can decrease $\bar{\epsilon}_1$ if necessary to ensure that the following inequalities hold for all (x, μ, λ) with $\operatorname{dist}((x,\mu,\lambda), \mathcal{S}) \leq \bar{\epsilon}_1$:

$$c_i(x) < (1/2)c_i(x^*) \leq -\psi(x,\mu,\lambda)^\sigma,$$

thus ensuring that $i \notin \mathcal{A}(x, \mu, \lambda)$.

We can reduce $\bar{\epsilon}_1$ again if necessary to ensure that the following relation holds for all $i \in \mathcal{A}^*$ and all (x, μ, λ) with $\operatorname{dist}((x, \mu, \lambda), \mathcal{S}) \leq \bar{\epsilon}_1$:

$$|c_i(x)| \leq L\|x - x^*\| \leq L \operatorname{dist}((x, \mu, \lambda), \mathcal{S}) \leq LC\psi(x, \mu, \lambda) \leq \psi(x, \mu, \lambda)^\sigma,$$

where L is the Lipschitz constant defined in (2.3). We conclude that $i \in \mathcal{A}(x, \mu, \lambda)$. □

High-quality estimates of the optimal Lagrange multipliers may be available in primal-dual interior-point algorithms and augmented Lagrangian algorithms. In SQP algorithms, an estimate (μ, λ) may be available from the QP subproblem solved at the previous iteration, or from an approximation procedure based on the current estimate of the active set (which usually also derives from the QP subproblem). However, the use of devices such as trust regions or ℓ_1 penalty terms in the subproblem may interfere with the accuracy of the Lagrange multiplier estimates. Moreover, in many algorithms, there is not a particularly strong motivation for obtaining accurate estimates of (μ, λ). For instance, in SQP algorithms that use exact second derivatives, rapid convergence of the primal iterates to x^* can be obtained even when (μ, λ) do not converge to $\mathcal{S}_D$; see Theorem 12.4.1 of Fletcher [11] and the comments that follow this result. The QP subproblem of the primal-dual algorithms in the `Knitro` software package may return only the primal variables, in which case the multipliers must be approximated using primal information [7].

Even in cases in which an estimate of (μ, λ) is available from the algorithm, it may be desirable to seek alternative values of (μ, λ) that decrease the value of $\psi(x, \mu, \lambda)$, thereby tightening the tolerance in the threshold test (2.5). This approach forms the basis of the techniques described in subsections 3.2 and 3.3, which provide asymptotically accurate estimates of the Lagrange multipliers as well as of the active set $\mathcal{A}^*$.

3. Identification from a primal point. We describe a number of techniques for estimating $\mathcal{A}^*$ for a given x near the solution x^*. We discuss the relationships between these techniques and conditions under which they provide asymptotically accurate estimates of $\mathcal{A}^*$.

3.1. Linear programming techniques. We describe here techniques based on a linearization of the ℓ_1 penalty formulation of (1.1). A linearized trust-region subproblem is solved and an estimate of $\mathcal{A}^*$ is extracted from the solution. One of these techniques is used by Byrd et al. [5, 6] as part of their SQP-EQP approach. (The idea of a linearized trust-region subproblem was proposed initially by Fletcher and Sainz de la Maza [12].)

The following subproblem forms the basis of the techniques in this section:

$$\min_d \; g^T d + \nu\|Jd + h\|_1 + \nu\|(c + Ad)_+\|_1 \quad \text{subject to } \|d\|_\infty \leq \Delta, \tag{3.1}$$

where ν is a penalty parameter, Δ is the trust-region radius, and all functions are assumed to be evaluated at x. This problem can be formulated explicitly as a linear program by introducing auxiliary variables r, s, and t, and writing

$$\min_{(d,r,s,t)} \; g^T d + \nu e^T r + \nu e^T s + \nu e^T t \quad \text{subject to} \tag{3.2a}$$

$$Ad + c \leq r, \;\; Jd + h = t - s, \;\; -\Delta e \leq d \leq \Delta e, \;\; (r, s, t) \geq 0, \tag{3.2b}$$

where, as mentioned in the introduction, we have $e = (1, 1, \ldots, 1)^T$. The dual of this problem is as follows:

$$\min_{(\lambda,\mu,u,v)} \; -c^T\lambda - h^T\mu + \Delta e^T u + \Delta e^T v \quad \text{subject to} \tag{3.3a}$$

$$A^T\lambda + J^T\mu + g = u - v, \;\; 0 \le \lambda \le \nu e, \;\; -\nu e \le \mu \le \nu e, \;\; (u, v) \ge 0. \tag{3.3b}$$

This formulation can be written more compactly as follows:

$$\min_{(\lambda,\mu)} \; -c^T\lambda - h^T\mu + \Delta\|A^T\lambda + J^T\mu + g\|_1 \quad \text{subject to} \tag{3.4a}$$

$$0 \le \lambda \le \nu e, \;\; -\nu e \le \mu \le \nu e. \tag{3.4b}$$

The formulations above are feasible and bounded. Moreover, they admit some invariance to scaling the constraints. Suppose, for some constraint c_i, we have that the λ_i component of the dual solution is strictly less than its upper bound of ν. By duality, we then have $r_i = 0$ at the solution of (3.2). If we scale constraint c_i by some $\sigma_i > 0$ (that is, we set $c_i \leftarrow \sigma_i c_i$ and $A_i \leftarrow \sigma_i A_i$), constraints (3.2b) and (3.3b) continue to be satisfied, while the objectives (3.2a) and (3.3a) remain unchanged (and therefore optimal) if we set $\lambda_i \leftarrow \lambda_i/\sigma_i$, provided that $\lambda_i/\sigma_i \le \nu$. Similar comments apply regarding the components of h.

The active set estimate can be derived from the solution of these linear programs in different ways. We mention the following three possibilities:

$$\mathcal{A}_c(x) = \{i \,|\, A_i d + c_i \ge 0\}, \tag{3.5a}$$

$$\mathcal{A}_\lambda(x) = \{i \,|\, \lambda_i > 0\}, \tag{3.5b}$$

$$\mathcal{A}_B(x) = \{i \,|\, \lambda_i \text{ is in the optimal basis for (3.3)}\}. \tag{3.5c}$$

The first of these activity tests (3.5a) cannot be expected to identify weakly active constraints except when $x = x^*$. The second test (3.5b) will generally not identify weakly active constraints, and will also fail to identify a strongly active constraint i if the particular multiplier estimate used in the test happens to have $\lambda_i = 0$. The third test (3.5c) does not attempt to estimate the full active set, but rather a "sufficient" subset of it that can be used in subsequent calculations requiring a nonsingular basis matrix for the active constraint gradients.

For the remainder of this section, we focus on $\mathcal{A}_c(x)$. The following simple lemma shows that, for x sufficiently near x^* and Δ sufficiently small, this activity test does not contain false positives.

LEMMA 3.1. *There are positive constants $\bar{\epsilon}_2$ and $\bar{\Delta}$ such that when $\|x - x^*\| \le \bar{\epsilon}_2$ and $\Delta \le \bar{\Delta}$, we have $\mathcal{A}_c(x) \subset \mathcal{A}^*$.*

Proof. We first choose $\bar{\epsilon}_2$ small enough such that for any x with $\|x - x^*\| \le \bar{\epsilon}_2$ and any $i \notin \mathcal{A}^*$ we have $c_i(x) \le \frac{1}{2}c_i(x^*) < 0$. By decreasing $\bar{\Delta}$ if necessary, we also have, for any $\|d\|_\infty \le \Delta \le \bar{\Delta}$ with $\|x - x^*\| \le \bar{\epsilon}_2$, that $i \notin \mathcal{A}^* \Rightarrow A_i(x)d + c_i(x) < 0$. The result follows from the definition (3.5a) of $\mathcal{A}_c(x)$. ∎

When the trust-region radius Δ is bounded in terms of $\|x - x^*\|$ and a constraint qualification holds, we can show that the set identified by (3.5a) is at least extensive enough to "cover" the objective gradient g^*.

THEOREM 3.2. *If MFCQ holds at x^*, for any $\zeta \in (0, 1)$, there are positive constants $\bar{\nu}$, $\bar{\epsilon}_2$, and $\bar{\Delta}$ such that whenever the conditions $\nu \ge \bar{\nu}$, $\|x - x^*\| \le \bar{\epsilon}_2$, and $\Delta \in [\|x - x^*\|^\zeta, \bar{\Delta}]$ are satisfied, we have*

$$-g^* \in \text{range}[\nabla h^*] + \text{pos}[(\nabla c_i^*)_{i \in \mathcal{A}_c(x)}]. \tag{3.6}$$

Proof. We start by defining $\bar{\epsilon}_2$ and $\bar{\Delta}$ as in Lemma 3.1. For these values (and any smaller values) we have immediately that $\mathcal{A}_c(x) \subset \mathcal{A}^*$.

We require $\nu \geq \bar{\nu}$, where

$$\bar{\nu} \stackrel{\text{def}}{=} \max\ (\{\|(\mu^*, \lambda^*)\|_\infty \mid (\mu^*, \lambda^*) \in \mathcal{S}_D\}) + 1. \tag{3.7}$$

Note that $\bar{\nu}$ is well defined because the KKT and MFCQ conditions guarantee the nonemptiness and boundedness of $\mathcal{S}_D$.

For any $(\mu^*, \lambda^*) \in \mathcal{S}_D$, the dual problem (3.3) at x^* with $(\mu, \lambda, u, v) = (\mu^*, \lambda^*, 0, 0)$ has objective value 0 because of the complementarity condition (1.3c). For the problem with $x \neq x^*$, we obtain a feasible point for (3.3) by setting

$$(\mu, \lambda, u, v) = \left(\mu^*, \lambda^*, (A^T\lambda^* + J^T\mu^* + g)_+, (A^T\lambda^* + J^T\mu^* + g)_-\right).$$

The objective at this point is

$$\begin{aligned}
&-c^T\lambda^* - h^T\mu^* + \Delta\|A^T\lambda^* + J^T\mu^* + g\|_1 \\
&= (c(x^*) - c(x))^T\lambda^* + (h(x^*) - h(x))^T\mu^* \\
&\qquad +\Delta\|(A(x^*)^T - A(x)^T)\lambda^* + (J(x^*)^T - J(x)^T)\mu^* + (g(x^*) - g(x))\|_1 \\
&= O(\|x - x^*\|).
\end{aligned} \tag{3.8}$$

The first equality is due to (1.3), while the second is due to the continuous differentiability of f, c, and h and the boundedness of $\mathcal{S}_D$. The optimal point for (3.3) must therefore have an objective value that is bounded above by a positive number of size $O(\|x - x^*\|)$.

Suppose for contradiction that regardless of how small we choose $\bar{\epsilon}_2$, there is an x with $\|x - x^*\| \leq \bar{\epsilon}_2$ such that the active set $\mathcal{A}_c(x)$ has the property that $-g^* \notin \text{range}[\nabla h^*] + \text{pos}[(\nabla c_i^*)_{i\in\mathcal{A}_c(x)}]$. Since there are only a finite number of possible sets $\mathcal{A}_c(x)$, we pick one of them for which this property holds for x arbitrarily close to x^* and call it $\mathcal{A}_1$. The set $\text{range}[\nabla h^*] + \text{pos}[(\nabla c_i^*)_{i\in\mathcal{A}_1}]$ is finitely generated and is therefore closed; see Rockafellar [19, Theorem 19.1].

Using the definition for dist $(\cdot, \cdot)$ (1.10), we have that τ defined by

$$\tau \stackrel{\text{def}}{=} (0.5)\text{dist}\ (-g^*, \text{range}[\nabla h^*] + \text{pos}[(\nabla c_i^*)_{i\in\mathcal{A}_1}]) \tag{3.9}$$

is strictly positive. After a possible reduction of $\bar{\epsilon}_2$, we have that

$$\text{dist}\ (-g(x), \text{range}[\nabla h(x)] + \text{pos}[(\nabla c_i(x))_{i\in\mathcal{A}_1}]) \geq \tau \tag{3.10}$$

for the given $\mathcal{A}_1$ and *all* x with $\|x - x^*\| \leq \bar{\epsilon}_2$. (The proof of the latter claim makes use of standard arguments and appears in Appendix A.)

Given x with $\mathcal{A}_c(x) = \mathcal{A}_1$, let the solutions to the problems (3.2) and (3.3) at x be denoted by (d_x, r_x, s_x, t_x) and $(\mu_x, \lambda_x, u_x, v_x)$, respectively. For all $i \notin \mathcal{A}_1$, we have by (3.5a) and complementarity that $A_i d_x + c_i < 0$, $(r_x)_i = 0$, and

$$(\lambda_x)_i = 0 \ \text{ for all } i \notin \mathcal{A}_1. \tag{3.11}$$

We now consider the objective of the dual problem (3.3) in two parts. We have by using the property (3.11) that

$$\begin{aligned}
\Delta e^T u_x + \Delta e^T v_x &\geq \Delta \min_{\lambda\geq 0, \mu} \left\| g + J^T\mu + \sum_{i\in\mathcal{A}_1} \lambda_i \nabla c_i \right\|_1 \\
&= \Delta\text{dist}\ (-g, \text{range}[\nabla h] + \text{pos}[(\nabla c_i)_{i\in\mathcal{A}_1}]) \\
&\geq \Delta\tau.
\end{aligned}$$

From $\nu \geq \bar{\nu}$ and (3.7), we also have

$$-c^T\lambda_x - h^T\mu_x \geq -\nu\|c_+\|_1 - \nu\|h\|_1.$$

By substituting these relations into the dual objective (3.3), we have

$$-c^T\lambda_x - h^T\mu_x + \Delta e^T u_x + \Delta e^T v_x \geq \Delta\tau - \nu\|c_+\|_1 - \nu\|h\|_1.$$

Finally, we decrease $\bar{\epsilon}_2$ further if necessary so that

$$\Delta\tau - \nu\|c(x)_+\|_1 - \nu\|h(x)\|_1 \geq (\tau/2)\|x - x^*\|^\zeta$$

for $\|x - x^*\| \leq \bar{\epsilon}_2$. We note that such a choice is possible since $\Delta \geq \|x - x^*\|^\zeta$ and $h(x)$ and $(c(x))_+$ are both $O(\|x - x^*\|)$. Hence the optimal objective in (3.3) is bounded below by $(\tau/2)\|x - x^*\|^\zeta$. This bound contradicts our earlier observation in (3.8) that the optimal objective is bounded above by a multiple of $\|x - x^*\|$. We conclude that $\tau = 0$ in (3.9), and thus $-g^* \in \text{range}[\nabla h^*] + \text{pos}[(\nabla c_i^*)_{i\in\mathcal{A}_c(x)}]$, as claimed. ∎

When the assumptions are made stronger, we obtain the following result.

COROLLARY 3.3. *If LICQ holds at x^*, for any ζ, $\bar{\epsilon}_2$, Δ, ν, and x satisfying the conditions of Theorem* 3.2, $\mathcal{A}^* \setminus \mathcal{A}_0^* \subset \mathcal{A}_c(x) \subset \mathcal{A}^*$. *If strict complementarity also holds at x^*, then $\mathcal{A}_c(x) = \mathcal{A}^*$.*

Proof. When LICQ holds at x^*, the multiplier (μ^*, λ^*) which satisfies equations (1.3) is unique, and $\lambda_i^* > 0$ for all $i \in \mathcal{A}^* \setminus \mathcal{A}_0^*$. For ζ, $\bar{\epsilon}_2$, Δ, and ν defined in Theorem 3.2, we must have $i \in \mathcal{A}_c(x)$ whenever $\lambda_i^* > 0$, since otherwise (3.6) would not hold. Thus, $\mathcal{A}^* \setminus \mathcal{A}_0^* \subset \mathcal{A}_c(x)$. Lemma 3.1 supplies $\mathcal{A}_c(x) \subset \mathcal{A}^*$. The final statement follows trivially from the equivalence of strict complementarity with $\mathcal{A}_0^* = \emptyset$. ∎

The implementation of SQP-EQP known as `Active` [5, 6], which is contained in the `Knitro` package, solves the formulation (3.2) using variants of the simplex method. It is observed (Waltz [20]) that many simplex iterations are spent in resolving the trust-region bounds $-\Delta e \leq d \leq \Delta e$. This effort would seem to be wasted; we are much more interested in the question of which linearized inequality constraints from (1.1) are active at the solution of (3.2) (and, ultimately, of (1.1)) than in the trust-region bounds. The authors of `Active` have tried various techniques to terminate the solution of (3.2) prematurely at an inexact solution, but these appear to increase the number of "outer" iterations of the SQP algorithm.

Because there is no curvature, trust-region bounds in (3.1) may be active, regardless of the size of Δ, even when x is arbitrarily close to x^*. The theorems above highlight the importance of choosing Δ large enough to allow constraints in $\mathcal{A}^*$ to become active in (3.1) but small enough to prevent inactive constraints (those not in $\mathcal{A}^*$) from becoming active in (3.1). Byrd et al. [5, section 3] describe a heuristic for `Active` in which Δ is adjusted from its value at the previous iteration of the outer algorithm according to the success of the QP step, the norms of the QP step, the solution d of (3.1), and whether or not the minimizer of the quadratic model in this direction d lies at the boundary of the trust region.

The performance of these schemes also depends strongly on the value of ν in (3.1) and (3.4). The bound

$$\nu \geq \max\left(\max_j \lambda_j^*, \max_k |\mu_k^*|\right) \tag{3.12}$$

ensures global convergence. However, excessively large estimates of ν can slow convergence. The heuristic used in `Active` [5, section 9] re-solves the LP for increasing values of ν whenever a substantial decrease in infeasibility is possible. In addition, ν is decreased whenever the bound (3.12) (using the current multiplier estimates) is inactive for several consecutive successful, feasible LP iterations.

Theorem 3.2 and Corollary 3.3 suggest that the approaches of this section may give false negatives for constraints that are weakly active, or which may have an optimal Lagrange multiplier of zero. However, it is not obvious that failure to identify such constraints would adversely affect the performance of nonlinear programming algorithms. To first order, they are not critical to satisfying the KKT conditions.

3.2. A technique based on the primal-dual estimate. Here we describe a scheme based on explicit minimization of $\psi(x, \mu, \lambda)$ in (2.1) with respect to (μ, λ) for $\lambda \geq 0$. We show that this minimization problem can be formulated as a linear program with equilibrium constraints (LPEC), one that is related to the linear programs discussed in subsection 3.1. However, in contrast to this earlier approach, we use a threshold test like (2.5) to estimate the active set, rather than active set or Lagrange multiplier information from the subproblem.

The LPEC subproblem is as follows:

$$\omega(x) \stackrel{\text{def}}{=} \min_{\lambda \geq 0, \mu} \sum_{i=1}^{m} |\min(\lambda_i, -c_i)| + \|h\|_1 + \|A^T\lambda + J^T\mu + g\|_1. \tag{3.13}$$

The activity test $\mathcal{A}_{\text{lpec}}$ is defined as

$$\mathcal{A}_{\text{lpec}}(x) = \{i \,|\, c_i(x) \geq -(\beta\omega(x))^\sigma\}, \tag{3.14}$$

where $\beta > 0$ and $\sigma \in (0, 1)$ are constants.

The problem (3.13) can be formulated as the following LPEC:

$$\omega(x) \stackrel{\text{def}}{=} \min_{(\lambda,\mu,s,u,v)} e^T s + \sum_{c_i \geq 0} c_i + \|h\|_1 + e^T u + e^T v \quad \text{subject to} \tag{3.15a}$$

$$0 \leq (-c)_+ - s \perp \lambda - s \geq 0, \tag{3.15b}$$

$$A^T\lambda + J^T\mu + g = u - v, \qquad (\lambda, s, u, v) \geq 0. \tag{3.15c}$$

By introducing a large constant M and binary variables y_i, $i = 1, 2, \ldots, m$ (which take on the value 0 if the minimum in $\min(-c_i, \lambda_i)$ is achieved by $-c_i$ and 1 if it is achieved by λ_i), we can write (3.15) as the following mixed-integer (binary) program:

$$\omega(x) \stackrel{\text{def}}{=} \min_{(\lambda,\mu,s,y,u,v)} e^T s + \|h\|_1 + e^T u + e^T v \quad \text{subject to} \tag{3.16a}$$

$$-c_i - s_i \leq -c_i y_i, \quad i = 1, 2, \ldots, m, \tag{3.16b}$$

$$\lambda_i - s_i \leq M(1 - y_i), \quad i = 1, 2, \ldots, m, \tag{3.16c}$$

$$A^T\lambda + J^T\mu + g = u - v, \tag{3.16d}$$

$$(\lambda, u, v) \geq 0, \qquad s \geq (c)_+, \qquad y_i \in \{0, 1\}, \quad i = 1, 2, \ldots m. \tag{3.16e}$$

The validity of this formulation for (3.13) is based on the nonnegativity of λ and the minimization of the $e^T s$ term. The large parameter M is necessary for (3.16c) but not for (3.16b), because $-c$ is a parameter while λ is a variable in the program.

There are notable similarities between the formulation (3.13) of the LPEC subproblem and the dual formulation (3.4) of the previous subsection. First, the term

$\|A^T\lambda + J^T\mu + g\|_1$ appears in both objectives, though in (3.4) it is weighted by the trust-region radius Δ. Second, the term $\|h\|_1$ in (3.13) (which is constant in (3.13) and (3.16a)) corresponds to the rather different term $-\mu^T h$ in (3.4). Third, the parameter ν which penalizes constraint violation does not appear in (3.13). Fourth, and perhaps most interestingly, the minimum $|\min(-c_i, \lambda_i)|$ in (3.13) is replaced by the product $(-c_i)\lambda_i$ in (3.4). While the use of the minimum may lead to stronger identification properties (see below), it is responsible for the presence of equilibrium constraints in (3.13) and therefore makes the subproblem much harder to solve. In addition, the attractive scale invariance property possessed by the $-c_i\lambda_i$ term is lost. If we multiply c_i and A_i by some $\sigma_i > 0$ and replace $\lambda_i \leftarrow \lambda_i/\sigma_i$ to maintain constancy of the product $A_i\lambda_i$, the minimum $|\min(-c_i, \lambda_i)|$ will be replaced by $|\min(-\sigma_i c_i, \lambda_i/\sigma_i)|$, which has a different value in general.

We now show that $\omega(x)$ defined in (3.13) provides a two-sided estimate of the distance to the solution and that the identification scheme (3.14) eventually is successful, under appropriate assumptions.

THEOREM 3.4. *Suppose that the KKT conditions* (1.3)*, the MFCQ* (1.7)*, and the second-order condition* (1.9) *are satisfied at* x^**, and let* ϵ *be as defined in Theorem* 2.1*. Then there are positive constants* $\bar{\epsilon} \in (0, \epsilon/2]$ *and* $\bar{C}$ *such that for all* x *with* $\|x - x^*\| \le \bar{\epsilon}$*, we have that*

(i) *the minimum in* (3.13) *is achieved at some* (μ, λ) *with* $\operatorname{dist}((\mu, \lambda), \mathcal{S}_D) \le \epsilon/2$*;*
(ii) $\bar{C}^{-1}\omega(x) \le \|x - x^*\| \le \bar{C}\omega(x)$*; and*
(iii) $\mathcal{A}_{\mathrm{lpec}}(x) = \mathcal{A}^*$.

Proof.

(i) Note first that for any $(\mu^*, \lambda^*) \in \mathcal{S}_D$ and any x with $\|x - x^*\| \le \epsilon$, we have that

$$\begin{aligned}
\omega(x) &\le \psi(x, \mu^*, \lambda^*) \\
&= \sum_{i=1}^{m} |\min(\lambda_i^*, -c_i(x))| + \|h(x)\|_1 + \|A(x)^T\lambda^* + J(x)^T\mu^* + g(x)\|_1 \\
&\le \sum_{i=1}^{m} |c_i(x) - c_i(x^*)| + \|h(x) - h(x^*)\|_1 \\
&\qquad + \|(A(x) - A(x^*))^T\lambda^* + (J(x) - J(x^*))^T\mu^* + (g(x) - g(x^*))\|_1 \\
&\le C_1\|x - x^*\| \qquad (3.17)
\end{aligned}$$

for some constant C_1. (In the second-to-last inequality, we used $\min(\lambda_i^*, -c_i(x^*)) = 0$, which follows from (1.3c).) Hence, if the minimum in (3.13) occurs outside the set $\{(\mu, \lambda) | \lambda \ge 0, \operatorname{dist}((\mu, \lambda), \mathcal{S}_D) \le \epsilon/2\}$ for x arbitrarily close to x^*, we must be able to choose a sequence (x^k, μ^k, λ^k) with $x^k \to x^*$, $\lambda^k \ge 0$, and $\operatorname{dist}((\mu^k, \lambda^k), \mathcal{S}_D) > \epsilon/2$ such that

$$\psi(x^k, \mu^k, \lambda^k) \le \psi(x^k, \mu^*, \lambda^*) \le C_1\|x^k - x^*\| \text{ for all } k.$$

In particular we have $\psi(x^k, \mu^k, \lambda^k) \to 0$. Consider first the case in which (μ^k, λ^k) is unbounded. By taking a subsequence if necessary, we can assume that

$$\|(\mu^k, \lambda^k)\| \to \infty, \quad \frac{(\mu^k, \lambda^k)}{\|(\mu^k, \lambda^k)\|} \to (\mu^*, \lambda^*), \quad \|(\mu^*, \lambda^*)\| = 1, \quad \lambda^* \ge 0.$$

For any $i \notin \mathcal{A}^*$, we have, by taking a further subsequence if necessary, that $c_i(x^k) < (1/2)c_i(x^*) < 0$ for all k. Since $|\min(\lambda_i^k, -c_i(x^k))| \le \psi(x^k, \mu^k, \lambda^k) \to 0$, we have that

$\lambda_i^k \to 0$ and thus $\lambda_i^* = 0$ for all $i \notin \mathcal{A}^*$. We also have that $A(x^k)^T\lambda^k + J(x^k)^T\mu^k + g(x^k) \to 0$; thus when we divide this expression by $\|(\mu^k, \lambda^k)\|$ and take limits, we obtain

$$A(x^*)^T\lambda^* + J(x^*)^T\mu^* = A_{\mathcal{A}^*}(x^*)^T\lambda^*_{\mathcal{A}^*} + J(x^*)^T\mu^* = 0.$$

We can now use the usual argument based on the MFCQ property (1.7) (see Appendix A) to deduce that $\lambda^*_{\mathcal{A}^*} = 0$ and then $\mu^* = 0$, contradicting $\|(\mu^*, \lambda^*)\| = 1$. Hence, the sequence (μ^k, λ^k) must be bounded.

By taking a subsequence if necessary, we can define a vector $(\hat{\mu}, \hat{\lambda})$ such that

$$(\mu^k, \lambda^k) \to (\hat{\mu}, \hat{\lambda}), \quad \hat{\lambda} \geq 0.$$

The limit $\psi(x^k, \mu^k, \lambda^k) \to 0$ thus implies that $\psi(x^*, \hat{\mu}, \hat{\lambda}) = 0$, which in turn implies that $(\hat{\mu}, \hat{\lambda}) \in \mathcal{S}_D$, contradicting $\text{dist}\,((\mu^k, \lambda^k), \mathcal{S}_D) > \epsilon/2$. Thus, there is some $\bar{\epsilon}$ such that for all x with $\|x - x^*\| \leq \bar{\epsilon}$ the minimum occurs in the set $\{(\mu, \lambda) | \lambda \geq 0, \text{dist}\,((\mu, \lambda), \mathcal{S}_D) \leq \epsilon/2\}$. Since this set is compact (boundedness of $\mathcal{S}_D$ follows from the MFCQ (1.7)), we conclude that the minimum in (3.13) is attained by some (μ, λ) in this set.

(ii) The left-hand inequality is already proved by (3.17). We now show that, for the $\bar{\epsilon} \in (0, \epsilon/2]$ determined in part (i), we have

$$\|x - x^*\| \leq C\omega(x) \text{ for all } x \text{ with } \|x - x^*\| \leq \bar{\epsilon} \tag{3.18}$$

for C defined in Theorem 2.1. First note that for any (μ, λ) with $\text{dist}\,((\mu, \lambda), \mathcal{S}_D) \leq \epsilon/2$, we have

$$\text{dist}\,((x, \mu, \lambda), \mathcal{S}) \leq \|x - x^*\| + \text{dist}\,((\mu, \lambda), \mathcal{S}_D) \leq \bar{\epsilon} + \epsilon/2 \leq \epsilon,$$

so that from Theorem 2.1 we have

$$\|x - x^*\| \leq \text{dist}\,((x, \mu, \lambda), \mathcal{S}) \leq C\psi(x, \mu, \lambda) \tag{3.19}$$

for all (μ, λ) with $\text{dist}\,((\mu, \lambda), \mathcal{S}_D) \leq \epsilon/2$ and $\lambda \geq 0$. We showed in part (i) that the minimum of $\psi(x, \mu, \lambda)$ is attained in this set for sufficiently small choice of $\bar{\epsilon}$. Hence, we have

$$\|x - x^*\| \leq C \min_{\lambda \geq 0, \text{dist}\,((\mu,\lambda),\mathcal{S}_D) \leq \epsilon/2} \psi(x, \mu, \lambda) = C\omega(x),$$

as required. The result follows by taking $\bar{C} = \max(C, C_1)$, where C_1 is from (3.17).

(iii) The proof of this final claim follows from an argument like that of Theorem 2.2. ∎

We note that an exact solution of (3.13) (or (3.16)) is not needed to estimate the active set. In fact, any approximate solution whose objective value is within a chosen fixed factor of the optimal objective value will suffice to produce an asymptotically accurate estimate. Computationally speaking, we can terminate the branch-and-bound procedure at the current incumbent once the lower bound is within a fixed factor of the incumbent objective value. Moreover, we can derive an excellent starting point for (3.16) from the solution of the dual subproblem (3.3) of the previous subsection or from the LP subproblem of the next section. (As our experiments of section 4 show, the branch-and-bound procedure often terminates at the root node, without

doing any expansion of the branch-and-bound tree at all. When this occurs, the main computational cost is the cost of solving a single LP relaxation of (3.16).)

The main differences between the schemes of this subsection and the previous one can be summarized as follows:

- When a second-order sufficient condition holds, the scheme of this subsection accurately estimates $\mathcal{A}^*$ (including the weakly active constraints), whereas the schemes of the previous subsection may identify only those active constraints that are instrumental in satisfying the first KKT condition (1.3a).
- Effectiveness of the techniques of the previous subsection depends critically on the choice of trust-region radius Δ, whereas no such parameter is present in this subsection. However, the practical performance of the latter approach may depend on the scaling of the constraints c_i and their multipliers λ_i. Performance may be improved for some problems by changing the relative weightings of the terms $\|h\|_1$ and $\|A^T\lambda + J^T\mu + g\|_1$ in $\omega(x)$. However, it is difficult to determine a choice of weights that works reliably for a range of problems.

3.3. A linear programming approximation to the LPEC. In this section, we describe a technique that has the same identification properties as the scheme of the previous subsection, as described in Theorem 3.4, but requires only the solution of a linear program, rather than an LPEC. The key to the scheme is to obtain a two-sided bound on $\omega(x)$, defined in (3.13), that can be obtained by solving a single linear program.

We start by defining the following functions:

$$\rho(x,\mu,\lambda) \stackrel{\text{def}}{=} \sum_{c_i<0} -c_i\lambda_i + \sum_{c_i\geq 0} c_i + \|h\|_1 + \|A^T\lambda + J^T\mu + g\|_1, \tag{3.20}$$

$$\bar{\rho}(x,\mu,\lambda) \stackrel{\text{def}}{=} \sum_{c_i<0} (-c_i\lambda_i)^{1/2} + \sum_{c_i\geq 0} c_i + \|h\|_1 + \|A^T\lambda + J^T\mu + g\|_1. \tag{3.21}$$

These functions are related in the following elementary fashion.

LEMMA 3.5. *For any (μ,λ) with $\lambda \geq 0$, we have*

$$\bar{\rho}(x,\mu,\lambda) \leq \rho(x,\mu,\lambda) + \sqrt{m}\rho(x,\mu,\lambda)^{1/2}.$$

Proof.

$$\begin{aligned}
\bar{\rho}(x,\mu,\lambda) &= \left\| \left[(-c_i\lambda_i)^{1/2} \right]_{c_i<0} \right\|_1 + \sum_{c_i\geq 0} c_i + \|h\|_1 + \|A^T\lambda + J^T\mu + g\|_1 \\
&\leq \sqrt{m} \left\| \left[(-c_i\lambda_i)^{1/2} \right]_{c_i<0} \right\|_2 + \sum_{c_i\geq 0} c_i + \|h\|_1 + \|A^T\lambda + J^T\mu + g\|_1 \\
&= \sqrt{m} \left[\sum_{c_i<0} (-c_i\lambda_i) \right]^{1/2} + \sum_{c_i\geq 0} c_i + \|h\|_1 + \|A^T\lambda + J^T\mu + g\|_1 \\
&\leq \sqrt{m}\rho(x,\mu,\lambda)^{1/2} + \rho(x,\mu,\lambda). \qquad \square
\end{aligned}$$

The next result defines the relationship between ρ, $\bar{\rho}$, and the proximality measure ψ defined in (2.1).

LEMMA 3.6. *Let $K_2 \geq 1$ be given. Then for all (x, μ, λ) with $\lambda \geq 0$ and*

$$\|c\|_\infty \leq K_2, \quad \|\lambda\|_\infty \leq K_2, \tag{3.22}$$

we have that

$$K_2^{-1}\rho(x,\mu,\lambda) \leq \psi(x,\mu,\lambda) \leq \bar{\rho}(x,\mu,\lambda). \tag{3.23}$$

Proof. For $c_i < 0$ and $\lambda_i \geq 0$, we have

$$-c_i\lambda_i = \min(-c_i, \lambda_i)\max(-c_i, \lambda_i) \geq \min(-c_i, \lambda_i)^2 \tag{3.24}$$

and also

$$-c_i\lambda_i \leq K_2 \min(-c_i, \lambda_i). \tag{3.25}$$

From (3.24) we have

$$\begin{aligned}\psi(x,\mu,\lambda) &= \sum_{c_i<0} |\min(-c_i,\lambda_i)| + \sum_{c_i\geq 0} c_i + \|h\|_1 + \|g + A^T\lambda + J^T\mu\|_1 \\ &\leq \sum_{c_i<0} (-c_i\lambda_i)^{1/2} + \sum_{c_i\geq 0} c_i + \|h\|_1 + \|g + A^T\lambda + J^T\mu\|_1 \\ &= \bar{\rho}(x,\mu,\lambda),\end{aligned}$$

thereby proving the right-hand inequality in (3.23).

For the left-hand inequality, we have from (3.25) and $K_2 \geq 1$ that

$$\psi(x,\mu,\lambda) \geq K_2^{-1}\sum_{c_i<0}(-c_i\lambda_i) + \sum_{c_i\geq 0} c_i + \|h\|_1 + \|A^T\lambda + J^T\mu + g\|_1 \geq K_2^{-1}\rho(x,\mu,\lambda),$$

as required. ∎

We are particularly interested in the solution (μ_x, λ_x) to the program

$$\min_{\mu, 0\leq\lambda\leq K_1 e} \rho(x,\mu,\lambda), \tag{3.26}$$

where K_1 is the constant defined in (2.4). The problem of determining (μ_x, λ_x) can also be expressed as the following linear program:

$$\min_{(\lambda,\mu,u,v)} \sum_{c_i<0}(-c_i\lambda_i) + \sum_{c_i\geq 0} c_i + \|h\|_1 + e^Tu + e^Tv \quad \text{subject to} \tag{3.27a}$$

$$A^T\lambda + J^T\mu + g = u - v, \;\; 0 \leq \lambda \leq K_1 e, \;\; (u,v) \geq 0. \tag{3.27b}$$

We define the activity test associated with $\bar{\rho}$ as follows:

$$\mathcal{A}_{\bar{\rho}}(x) = \{i = 1, 2, \ldots, m \,|\, c_i(x) \geq -(\beta\bar{\rho}(x,\mu_x,\lambda_x))^{\bar{\sigma}}\} \tag{3.28}$$

for given constants $\beta > 0$ and $\bar{\sigma} \in (0, 1)$.

We now prove a result similar to Theorem 3.4, showing in particular that under the same assumptions as the earlier result, the identification scheme above is asymptotically successful.

THEOREM 3.7. *Suppose that the KKT conditions* (1.3)*, the MFCQ* (1.7)*, and the second-order condition* (1.9) *are satisfied at x^*, and let ϵ be as defined in Theorem* 2.1*. Then there exists a positive constant $\hat{\epsilon} \in (0, \epsilon/2]$ such that for all x with $\|x - x^*\| \leq \hat{\epsilon}$, we have*

(i) *the minimum in* (3.27) *is attained at some* (μ, λ) *with* $\operatorname{dist}((\mu, \lambda), \mathcal{S}_D) \le \epsilon/2$*;*
(ii) $K_1^{-1}\rho(x, \mu_x, \lambda_x) \le \omega(x) \le \bar{\rho}(x, \mu_x, \lambda_x)$*, where* K_1 *is the constant defined in* (2.4)*; and*
(iii) $\mathcal{A}_{\bar{\rho}}(x) = \mathcal{A}^*$.

Proof.

(i) Note that for any $(\mu^*, \lambda^*) \in \mathcal{S}_D$ and any x with $\|x - x^*\| < \epsilon$, we have

$$\begin{aligned}
&\rho(x, \mu^*, \lambda^*) \\
&= \sum_{c_i<0} -c_i(x)\lambda_i^* + \sum_{c_i\ge 0} c_i(x) + \|h(x)\|_1 + \|A(x)^T\lambda^* + J(x)^T\mu^* + g(x)\|_1 \\
&\le \sum_{c_i<0} (c_i(x^*) - c_i(x))\lambda_i^* + \sum_{c_i\ge 0} (c_i(x) - c_i(x^*)) + \|h(x) - h(x^*)\|_1 \\
&\qquad + \|(A(x) - A(x^*))^T\lambda^* + (J(x) - J(x^*))^T\mu^* + (g(x) - g(x^*))\|_1 \\
&\le C_2\|x - x^*\|
\end{aligned}$$

for some constant C_2. (In the first inequality above, we used the fact $\lambda_i^* c_i^* = 0$ for all i to bound the first summation, and the fact that $c_i^* \le 0$ for all i to bound the second summation.) Note that since $\|(\mu^*, \lambda^*)\|_\infty \le K_1$, we have $0 \le \lambda^* \le K_1 e$, so that (μ^*, λ^*), together with an obvious choice of (u, v), is feasible for (3.27). We note also that any $(\hat{\mu}, \hat{\lambda})$ for which $\rho(x^*, \hat{\mu}, \hat{\lambda}) = 0$ and $\hat{\lambda} \ge 0$ satisfies $(\hat{\mu}, \hat{\lambda}) \in \mathcal{S}_D$. Using these observations, the remainder of the proof closely parallels that of Theorem 3.4(i), so we omit the details.

(ii) Reduce $\hat{\epsilon}$ if necessary to ensure that $\hat{\epsilon} \le \bar{\epsilon} \le \epsilon/2$, where $\bar{\epsilon}$ is defined in Theorem 3.4. Reduce $\hat{\epsilon}$ further if necessary to ensure that $\|c(x)\|_\infty \le K_1$ for all x with $\|x - x^*\| \le \hat{\epsilon}$. Note that by Theorem 3.4(i), the minimizer of (3.13) has $\operatorname{dist}((\mu, \lambda), \mathcal{S}_D) \le \epsilon/2$, and therefore $\|\lambda\|_\infty \le \|\lambda^*\|_\infty + 1 \le K_1$ for any $(\mu^*, \lambda^*) \in \mathcal{S}_D$.

Using the result of Lemma 3.6 (with K_1 replacing K_2), we have that

$$\begin{aligned}
&K_1^{-1}\rho(x, \mu_x, \lambda_x) \\
&= \min_{\mu, 0\le\lambda\le K_1 e} K_1^{-1}\rho(x, \mu, \lambda) \le \min_{\mu, 0\le\lambda\le K_1 e} \psi(x, \mu, \lambda) \le \psi(x, \mu_x, \lambda_x) \le \bar{\rho}(x, \mu_x, \lambda_x).
\end{aligned}$$

However, as we showed in Theorem 3.4(i), the minimizer of $\psi(x, \mu, \lambda)$ over the set of (μ, λ) with $\lambda \ge 0$ is attained at values of (μ, λ) that satisfy the restriction $\|\lambda\|_\infty \le K_1$; thus we can write

$$K_1^{-1}\rho(x, \mu_x, \lambda_x) \le \min_{\mu, 0\le\lambda} \psi(x, \mu, \lambda) \le \bar{\rho}(x, \mu_x, \lambda_x),$$

which yields the result, by (3.13).

(iii) We have from Lemma 3.5, Theorem 3.4(ii), and part (ii) of this theorem that $\bar{\rho}(x, \mu_x, \lambda_x) \to 0$ as $x \to x^*$. Therefore, using continuity of c_i, $i = 1, 2, \ldots, m$, we can decrease $\hat{\epsilon}$ if necessary to ensure that for $\|x - x^*\| \le \hat{\epsilon}$, we have

$$c_i(x) < (1/2)c_i(x^*) \le -(\beta\bar{\rho}(x, \mu_x, \lambda_x))^{\bar{\sigma}} \quad \text{for all } i \notin \mathcal{A}^*.$$

Hence, $i \notin \mathcal{A}_{\bar{\rho}}(x)$ for all such x.

For $i \in \mathcal{A}^*$, we have for the Lipschitz constant L defined in (2.3), and using

Theorem 3.4(ii) and part (ii) of this theorem, that

$$\begin{aligned}|c_i(x)| &\le L\|x-x^*\| \\ &= L\|x-x^*\|^{1-\bar\sigma}\|x-x^*\|^{\bar\sigma} \\ &\le L\|x-x^*\|^{1-\bar\sigma}\bar C^{\bar\sigma}\omega(x)^{\bar\sigma} \\ &\le \left[L\|x-x^*\|^{1-\bar\sigma}\bar C^{\bar\sigma}/\beta^{\bar\sigma}\right](\beta\bar\rho(x,\mu_x,\lambda_x))^{\bar\sigma} \\ &\le (\beta\bar\rho(x,\mu_x,\lambda_x))^{\bar\sigma}\end{aligned}$$

for $\hat\epsilon$ sufficiently small. Hence, we have $i \in \mathcal{A}_{\bar\rho}(x)$ for all x with $\|x-x^*\| \le \hat\epsilon$. □

Near the solution, $\omega(x)$ may be (and often is) much smaller than $\bar\rho(x,\mu_x,\lambda_x)$, because of the looseness of the estimate (3.24). To compensate for this difference, we set $\bar\sigma$ in the definition of $\mathcal{A}_{\bar\rho}$ (3.28) to be larger than σ in the definition of $\mathcal{A}_{\rm lpec}$ (3.14) in the tests described in the next section.

A referee has pointed out that some interesting insights are available from examination of the dual of the subproblem (3.27). Ignoring the upper bound $\lambda \le K_1 e$, we can write the dual as

$$\begin{aligned}\min g^T d \quad & \text{subject to} \\ A_i d + c_i \le 0 \quad & \text{for } i \text{ with } c_i < 0, \\ A_i d \le 0 \quad & \text{for } i \text{ with } c_i \ge 0, \\ Jd = 0, \quad & -e \le d \le e.\end{aligned}$$

It is not difficult to construct examples for which $A_i d + c_i = 0$ for an inactive constraint i, even when x is arbitrarily close to x^*. (The referee gave the example of minimizing a scalar x^2 subject to $-x - 0.5 \le 0$ for x slightly greater than the optimum $x^* = 0$.) Thus, if the active set estimate were obtained from formulae such as (3.5), it may not be asymptotically accurate. Hence, the use of the threshold test (3.28) in place of activity tests (3.5) is key to the effectiveness of the approach of this section. In this vein, it can be shown that if the (μ,λ) components of the solution of the earlier LP problem (3.3) are inserted into the threshold test (3.28) in place of (μ_x,λ_x), an asymptotically accurate estimate is obtained, under certain reasonable assumptions. We omit a formal statement and proof of this claim, as we believe (μ_x,λ_x) to be a better choice of the Lagrange multipliers, because their calculation does not depend on the parameters ν and Δ that appear in (3.3).

4. Computational results. In this section, we apply the techniques of section 3 to a variety of problems in which x is slightly perturbed from its (approximately) known solution x^*. The resulting active set estimate is compared with our best guess of the active set at the solution. We report the false positives and false negatives associated with each technique, along with the runtimes required to execute the tests.

The LP techniques of subsection 3.1 are referred to as LP-P for the primal formulation (3.2) and LP-D for the dual formulation (3.3). For these formulations, we use both activity tests $\mathcal{A}_c$ and $\mathcal{A}_\lambda$ of (3.5), modified slightly with activity thresholds. We also implement the LPEC scheme of subsection 3.2 and the LP approximation scheme of subsection 3.3, which we refer to below as LPEC-A. We implemented all tests in C, using the CPLEX callable library (version 9.0) to solve the linear and mixed-integer programs.

The times required to implement the tests are related to the size and density of the constraint matrix for each formulation. The matrix dimensions for each formulation

TABLE 4.1
Problem dimensions as a function of the number of inequalities (m), variables (n), and equalities (p).

	LP-D	LP-P	LPEC-A	LPEC
Rows	n	$m+p$	n	$5m+n$
Columns	$m+2n+p$	$m+n+2p$	$m+2n+p$	$4m+2n+p$

are given in Table 4.1. Except for problems with many equality constraints, the LPEC formulation has the largest constraint matrix. Further, it is the only formulation with binary variables.

Subsection 4.1 discusses some specifics of the formulations, such as the choice of parameters and tolerances in the identification procedures. In subsection 4.2, we apply the identification techniques to a set of random problems, for which we have control over the dimensions and amount of degeneracy. In subsection 4.3, we consider a subset of constrained problems from the CUTEr test set, a conglomeration of problems arising from theory, modeling, and real applications [13]. While the random problems are well scaled with dense constraint Jacobians, the CUTEr problems may be poorly scaled and typically have sparse constraint Jacobians. Subsection 4.4 contains some remarks about additional testing.

4.1. Implementation specifics.

4.1.1. Choosing parameters and tolerances. The following implementation details are common to both random and CUTEr test sets. We bound the ℓ_∞ norm of the perturbation x from the (approximately) optimal point x^* by a noise parameter `noise`. Denoting by ϕ a random variable drawn from the uniform distribution on $[-1, 1]$, we define the perturbed point x as follows:

$$x_i = x_i^* + \frac{\texttt{noise}}{n}\phi, \quad i = 1, 2, \ldots, n. \tag{4.1}$$

A second parameter `DeltaFac` controls the bound on the trust-region radius for the LP-P and LP-D programs. We set

$$\Delta = \texttt{DeltaFac}\frac{\texttt{noise}}{n},$$

so that when $\texttt{DeltaFac} \geq 1$, the trust region is large enough to contain the true solution x^*. For the results tabulated below, we use $\texttt{DeltaFac} = 4$. This value is particularly felicitous for the LP-P and LP-D schemes, as it yields a Δ large enough to encompass the solution yet small enough to exclude many inactive constraints. The number of false positives therefore tends to be small for LP-P and LP-D in our tables. The relatively small trust region also allows the CPLEX presolver to streamline the LP formulations before calling the simplex code, thus reducing the solve times for the linear programs in LP-P and LP-D. (Specifically, for each inequality constraint that is inactive over the entire trust region, the LP-P subproblem is reduced by one row and column, while the LP-D subproblem is reduced by one column.) It is unlikely that a nonlinear programming algorithm that uses LP-P or LP-D as its identification technique could in practice choose a value of Δ as nice as the one used in these tests.

The activity tests $\mathcal{A}_c$ (3.5a) and $\mathcal{A}_\lambda$ (3.5b) were modified to include a tolerance, as follows:

$$\mathcal{A}_c(x) = \{i \mid A_i d + c_i \geq -\epsilon_0\} \tag{4.2}$$

and

$$\mathcal{A}_\lambda(x) = \{i \,|\, \lambda_i \geq \epsilon_0\}, \tag{4.3}$$

with $\epsilon_0 = 10^{-4}$.

In the tests $\mathcal{A}_{\text{lpec}}(x)$ (3.14) for LPEC and $\mathcal{A}_{\bar{\rho}}$ (3.28) for LPEC-A, we set $\beta = 1/(m+n+p)$. The value 0.75 is used for σ in $\mathcal{A}_{\text{lpec}}$ (3.14), while the larger value 0.90 is used for $\bar{\sigma}$ in $\mathcal{A}_{\bar{\rho}}$ (3.28).

By default, the mixed-integer solver in CPLEX makes use of various cut generation schemes, including flow covers, MIR cuts, implied bound cuts, and Gomory fractional cuts. We disabled these schemes because, given our usually excellent starting point for the LPEC test, the cost of cut generation is excessive compared to the cost of solving the root relaxation. However, for all tests, we allowed both linear and mixed-integer solvers to perform their standard presolving procedures, as they generally improved the performance. For the mixed-integer solver in LPEC, we accept the solution if it is within a factor of 2 of the lower bound.

For both linear and integer programming solvers, we tightened the general feasibility tolerance `eprhs` from 10^{-6} to 10^{-9} because problems in the CUTEr test set (notably BRAINPC0 and BRAINPC3) report infeasibilities after scaling for LPEC-A under the default tolerance. In addition, we observed LP-P objective values that were too negative when using default `eprhs` values. Specifically, for some of the constraints $A_i d + c_i - r_i \leq 0$ in (3.2), the solver would find a d with $A_i d + c_i$ slightly positive, while setting r_i to zero. Thus, the constraint would be satisfied to the specified tolerance, while avoiding the larger value of $g^T d$ that would be incurred if it were satisfied exactly.

4.1.2. Formulation details. For all test problems, the parameter ν of the LP-P and LP-D programs is assigned a value large enough to ensure that the known optimal multipliers (μ^*, λ^*) are feasible for (3.4). The results of this paper, theoretical and computational, are otherwise insensitive to the choice of ν. (However, the choice of ν appears to be important for global convergence of the nonlinear programming algorithm, as discussed in Byrd et al. [6].)

The computational efficiency of the LPEC mixed-integer program (3.16) is sensitive to the magnitude of M. Recall that the formulation (3.16) is identical to the LPEC (3.15) provided that M is sufficiently large, in particular, larger than $\|c+\lambda^*\|_\infty$, where λ^* is an optimal multiplier. However, excessively large M values may result in long runtimes. We observed runtime reductions of as much as 50% when we replaced heuristically chosen values of M with near-minimal values.

We describe some heuristics for setting M and ν in the following subsections.

In solving LPEC, we use a starting point based on the solution for LPEC-A. Specifically, we set λ, μ, u, and v to their optimal values from (3.27); set $y_i = 0$ if $-c_i < \lambda_i$ and $y_i = 1$ otherwise; and set $s_i = |\min(-c_i, \lambda_i)|$. In most cases, this starting point is close to an acceptable solution for LPEC and little extra work is needed beyond solving an LP relaxation of the LPEC at the root node and verifying that the starting point is not far from the lower bound obtained from this relaxation. The solution to LP-D also provides a useful starting point for LPEC in most cases.

For LPEC and LPEC-A, no attempt is made to scale the constraints c_i or the components of the threshold functions $\omega(x)$ and $\bar{\rho}(x, \mu_x, \lambda_x)$. Heuristics to adjust such weightings may improve the performance of LPEC and LPEC-A techniques.

4.2. Random problems. We generate random problems involving dense Jacobians J and A to mimic the behavior of a nonlinear program near a local solution x^*. Besides choosing the dimensions n, m, and p, we influence the amount of degeneracy in the problem by specifying the row rank of J and A and the proportion of weakly active constraints.

4.2.1. Problem setup. Parameters specific to the random problem setup are `fStrong` and `fWeak` (approximate proportion of strongly and weakly active inequality constraints, respectively) and `degenA` and `degenJ` (proportional to the ranks of the null spaces of A and J, respectively). We first fill out the first $(1 - \texttt{degenA})m$ rows of the optimal inequality constraint Jacobian A^* with components 5ϕ (where, as above, ϕ represents a random variable uniformly distributed in $[-1, 1]$). We then set the last $(\texttt{degenA})m$ rows of A^* to be random linear combinations of the first $(1-\texttt{degenA})m$ rows, where the coefficients of the linear combinations are chosen from ϕ. A similar process is used to choose the optimal equality constraint Jacobian J^* using the parameter `degenJ`.

We set the solution to be $x^* = 0$. Recall that x is a perturbation of x^* (4.1). First, we set each component of μ^* to $\frac{1}{2}\phi(\phi + 1)$. Next, we randomly classify each index $i \in \{1, 2, \ldots, m\}$ as "strongly active," "weakly active," or "inactive," such that the proportion in each category is approximately `fStrong`, `fWeak`, and $(1-\texttt{fStrong}-\texttt{fWeak})$, respectively. For the inactive components, we set $c_i^* = -\frac{5}{2}(\phi + 1)^2$, while for the strongly active components, we set $\lambda_i^* = \frac{5}{2}(\phi + 1)^2$. Other components of c^* and λ^* are set to zero. To make the optimality condition (1.3a) consistent, we now set $g^* = -(A^*)^T\lambda^* - (J^*)^T\mu^*$. Naturally, $h^* = 0$.

In accordance with the assumed Lipschitz properties, we set

$$\begin{aligned} g_i &= g_i^* + (\texttt{noise}/n)\phi, \quad i = 1, 2, \ldots, m, \\ A_{ij} &= A_{ij}^* + (\texttt{noise}/n)\phi, \quad i = 1, 2, \ldots, m, \quad j = 1, 2, \ldots, n, \\ J_{ij} &= J_{ij}^* + (\texttt{noise}/n)\phi, \quad i = 1, 2, \ldots, p, \quad j = 1, 2, \ldots, n. \end{aligned}$$

Since $c(x) = c^* + A^*(x - x^*) + O(\|x - x^*\|^2) = c^* + A^*x + O(\|x\|^2)$, we set

$$c_i = c_i^* + A_i^*x + (\texttt{noise}/n)^2\phi, \quad i = 1, 2, \ldots, m.$$

A similar scheme is used to set h.

The data thus generated is consistent to first order, but there is no explicit assurance that the second-order condition holds. (This condition is required for Theorems 3.4 and 3.7 concerning the exact identification properties of the LPEC and LPEC-A schemes.)

By setting ν to the large value 100, we ensure that solutions of LP-P and LP-D have $r = 0$ and $s = t = 0$. For the LPEC problem (3.16), we define $M = 5 \max_j(|c_j|)$. This value is large enough to secure local optimality of the LPEC programs of our test problems.

4.2.2. Nondegenerate problems. Results for a set of random nondegenerate problems are shown in Table 4.2, with runtimes in Table 4.3. Nondegeneracy is assured by setting $\texttt{fWeak} = 0$, $\texttt{degenA} = 0$, and $\texttt{degenJ} = 0$. The number of equality constraints p is $n/5$, and we set $\texttt{noise} = 10^{-3}$. Each entry in Table 4.2 shows the numbers of false positives and false negatives for the problem and test in question. For LP-P and LP-D, we report both active sets $\mathcal{A}_c$ (4.2) and $\mathcal{A}_\lambda$ (4.3). The case $m = 400$, $n = 200$, `fStrong=0.50` does not appear because the expected number of degrees of freedom $(n - p - (\texttt{fStrong} + \texttt{fWeak})m)$ is nonpositive.

TABLE 4.2
Nondegenerate random problems: False positives/false negatives. $p = n/5$, `noise` $= 10^{-3}$, `fWeak` $= 0.00$, `degenA` $= 0.00$, `DeltaFac` $= 4.00$.

m	n	fStrong	LP-D		LP-P		LPEC-A	LPEC
			$\mathcal{A}_c$	$\mathcal{A}_\lambda$	$\mathcal{A}_c$	$\mathcal{A}_\lambda$		
50	200	0.10	0/0	0/0	0/0	0/0	1/0	1/0
50	200	0.50	0/0	0/0	0/0	0/0	0/0	0/0
50	1000	0.10	0/0	0/0	0/0	0/0	0/0	0/0
50	1000	0.50	0/0	0/0	0/0	0/0	0/0	0/0
100	200	0.10	0/0	0/0	0/0	0/0	0/0	0/0
100	200	0.50	0/0	0/0	0/0	0/0	1/0	0/0
100	1000	0.10	0/0	0/0	0/0	0/0	0/0	0/0
100	1000	0.50	0/0	0/1	0/0	0/1	0/0	0/0
400	200	0.10	0/0	0/0	0/0	0/0	2/0	1/1
400	1000	0.10	1/0	0/0	1/0	0/0	3/0	1/0
400	1000	0.50	1/0	0/0	1/0	0/0	4/0	3/0

TABLE 4.3
Nondegenerate random problems: Time (secs). $p = n/5$, `noise` $= 10^{-3}$, `fWeak` $= 0.00$, `degenA` $= 0.00$, `DeltaFac` $= 4.00$.

m	n	fStrong	LP-D	LP-P	LPEC-A	LPEC
50	200	0.10	0.07	0.03	0.11	0.16
50	200	0.50	0.09	0.05	0.11	0.13
50	1000	0.10	7.08	4.61	7.34	7.94
50	1000	0.50	7.73	4.79	6.98	7.85
100	200	0.10	0.08	0.03	0.19	0.26
100	200	0.50	0.13	0.10	0.18	0.26
100	1000	0.10	7.05	4.46	9.35	9.99
100	1000	0.50	8.84	6.77	9.41	10.40
400	200	0.10	0.15	0.11	0.47	8.05
400	1000	0.10	9.80	5.59	20.70	171.24
400	1000	0.50	17.87	18.17	21.57	26.35

The identification techniques are accurate on these problems. Because the LICQ conditions hold (to high probability), even the LP-P and LP-D procedures are guaranteed to be asymptotically correct. Indeed, the LP-P and LP-D schemes generally perform best; the LPEC-A and LPEC schemes show a few false positives on the larger examples. For these problems, it is not necessary for the LPEC to search beyond the root node in the branch-and-bound tree, except in the case $m = 400$, $n = 200$, `fStrong` $= 0.10$, for which one additional node is considered.

In agreement with the theory of section 3, the false positives reported in LPEC-A and LPEC results disappear for smaller noise values. In particular, for `noise` $= 10^{-7}$ the identification results are perfect for the LPEC-A and LPEC methods, while the LP-D and LP-P methods still give some errors.

Runtimes are shown in Table 4.3. The differences between the approaches are not significant, except for two of the $n = 400$ cases, for which LPEC is substantially slower than LPEC-A.

4.2.3. Degenerate problems. Results for a set of random degenerate problems are shown in Table 4.4, with runtimes in Table 4.5. In these tables, we fixed `fStrong` $= 0.2$, `noise` $= 10^{-3}$, `degenJ` $= 0$, and $p = n/5$. The values of `fWeak` and `degenA` were varied, along with the dimensions m and n.

TABLE 4.4
Degenerate random problems: False positives/false negatives: $p = n/5$, `noise` $= 10^{-3}$, `fStrong` $= 0.20$, `DeltaFac` $= 4.00$.

m	n	fWeak	degenA	LP-D		LP-P		LPEC-A	LPEC
				$\mathcal{A}_c$	$\mathcal{A}_\lambda$	$\mathcal{A}_c$	$\mathcal{A}_\lambda$		
50	200	0.05	0.0	1/1	0/2	1/1	0/2	1/0	1/0
50	200	0.05	0.1	0/2	0/2	0/2	0/2	0/0	0/1
50	200	0.05	0.3	0/0	0/2	0/0	0/2	0/0	0/0
50	200	0.20	0.0	1/3	0/6	1/3	0/6	1/0	1/0
50	200	0.20	0.1	0/2	0/6	0/2	0/6	0/0	0/0
50	200	0.20	0.3	0/4	0/6	0/4	0/6	0/0	0/0
50	1000	0.05	0.0	0/0	0/2	0/0	0/2	0/0	0/0
50	1000	0.05	0.1	0/2	0/2	0/2	0/2	0/0	0/0
50	1000	0.05	0.3	0/2	0/2	0/2	0/2	0/1	0/1
50	1000	0.20	0.0	0/3	0/6	0/3	0/6	0/0	0/0
50	1000	0.20	0.1	0/3	0/6	0/3	0/6	0/0	0/0
50	1000	0.20	0.3	0/4	0/6	0/4	0/6	0/2	0/1
400	200	0.05	0.0	0/10	0/19	0/10	0/19	2/0	1/0
400	200	0.05	0.1	1/5	0/19	1/5	0/19	7/0	3/5
400	200	0.05	0.3	1/11	0/19	1/11	0/19	6/1	2/6
400	1000	0.05	0.0	2/8	0/19	2/9	0/19	3/0	1/0
400	1000	0.05	0.1	1/8	0/19	1/8	0/19	1/0	0/1
400	1000	0.05	0.3	4/7	0/19	4/7	0/19	6/0	5/7
400	1000	0.20	0.0	1/25	0/77	1/25	0/77	4/0	1/0
400	1000	0.20	0.1	0/22	0/77	0/22	0/77	1/0	0/6
400	1000	0.20	0.3	1/28	0/77	1/28	0/77	4/2	2/10

TABLE 4.5
Degenerate random problems: Time (secs). $p = n/5$, `noise` $= 10^{-3}$, `fStrong` $= 0.20$, `DeltaFac` $= 4.00$.

m	n	fWeak	degenA	LP-D	LP-P	LPEC-A	LPEC
50	200	0.05	0.0	0.08	0.04	0.11	0.15
50	200	0.05	0.1	0.08	0.04	0.10	0.15
50	200	0.05	0.3	0.08	0.04	0.12	0.16
50	200	0.20	0.0	0.06	0.04	0.10	0.13
50	200	0.20	0.1	0.09	0.05	0.11	0.16
50	200	0.20	0.3	0.09	0.04	0.11	0.38
50	1000	0.05	0.0	7.51	4.05	6.99	7.81
50	1000	0.05	0.1	7.26	4.70	6.71	7.70
50	1000	0.05	0.3	7.08	4.47	7.22	17.15
50	1000	0.20	0.0	7.54	4.43	7.47	8.04
50	1000	0.20	0.1	8.11	4.28	6.80	7.75
50	1000	0.20	0.3	7.52	4.80	7.28	17.04
400	200	0.05	0.0	0.23	0.28	0.41	2.71
400	200	0.05	0.1	0.27	0.28	0.41	9.69
400	200	0.05	0.3	0.22	0.30	0.39	3.81
400	1000	0.05	0.0	12.06	7.99	21.35	71.89
400	1000	0.05	0.1	10.99	9.73	21.99	29.76
400	1000	0.05	0.3	11.36	9.71	22.78	138.38
400	1000	0.20	0.0	15.59	12.33	20.98	141.82
400	1000	0.20	0.1	13.97	10.94	22.39	29.67
400	1000	0.20	0.3	13.47	11.24	22.56	131.26

All methods perform well when $m = 50$. The LPEC and LPEC-A approaches rarely make an identification error on these problems, whereas LP-P and LP-D record a few false negatives. For the problems with 400 inequality constraints, the numbers of errors made by LPEC-A and LPEC are lower than those made by LP-P and LP-D. The misidentifications for LP-D and LP-P tend to be false negatives, and their numbers increase with the number of weakly active constraints. This experience is in accordance with the theory of subsection 3.1, which gives no guarantee that the weakly active constraints will be identified. The numbers of false negatives are larger for test $\mathcal{A}_\lambda$ than for $\mathcal{A}_c$—nearly as large as the number of degenerate constraints. (For $m = 400$, `fWeak` $= 0.05$ there are 20 such constraints, while for $m = 400$ and `fWeak` $= 0.20$ there are 80.) This observation indicates that the multiplier (μ, λ) determined by the LP-P and LP-D solution is similar to the optimal multiplier (μ^*, λ^*) for the original problem, for which $\lambda_i^* = 0$ when constraint i is weakly active. The errors for LPEC-A and LPEC contain both false positives and false negatives, indicating that the values of σ and $\bar{\sigma}$ and the factor β that we use in the activity test are appropriate. (For larger values of σ and $\bar{\sigma}$, the numbers of false negatives increase dramatically.)

The methods can usually be ranked in order of speed as LP-P, LP-D, LPEC-A, and LPEC. The differences between LP-P and LP-D are likely due to problem size reductions by the presolver, which are greater for LP-P, and which are significant because the matrix is dense. As expected (see our discussion in subsection 4.1.1), we observed size reductions corresponding to the number of inactive constraints. In contrast, no presolver reductions were observed for LPEC-A.

For the mixed-integer program arising in the LPEC test, an additional node beyond the root node of the branch-and-bound tree is considered only for the case $m = 400$, $n = 200$, `fWeak` $= 0.05$, and `degenA` $= 0.1$. We observed large initial scaled dual infeasibilities and runtimes that are sensitive to the LPEC parameter M. For the case $m = 400$, the relative slowness of the LPEC method may be due to the relatively large size of the matrix generated by the LPEC (see Table 4.1).

4.3. CUTEr problems. We now consider a subset of constrained minimization problems from the CUTEr test set [13]. The subset contains degenerate problems of small or medium size for which the `Interior/Direct` algorithm of `Knitro 4.x` terminates successfully within 3000 iterations (with default parameter values). From the output of this code, we obtain approximations x^* to a solution and (μ^*, λ^*) to the optimal Lagrange multipliers.

The format of the CUTEr test problems differs from that of (1.1) in that bound constraints are treated separately from general constraints and all constraints are two-sided; that is, they have both lower and upper bounds. We implemented alternative formulations for our four tests which treated the bounds explicitly and combined them with the trust-region constraints, thereby reducing the total number of constraints and/or variables. We found that these formulations gave little or no improvement in performance, so we do not report on them further. For the results below, we rewrite the CUTEr test problems in the format (1.1), treating bound constraints in the same way as general inequality constraints.

4.3.1. Determining the "true" active set. In contrast to the random problems of section 4.2, the true active set $\mathcal{A}^*$ is not known but must be estimated from the solution determined by `Interior/Direct`. Inevitably, this solution is approximate; the code terminates when the constraint-multiplier product for each inequality constraint falls below a given tolerance, set by default to 10^{-6} (see Byrd, Hribar, and Nocedal [7]). If one of λ_i or $-c_i$ is much smaller than the other, classification

of the constraint is easy, but in many cases these two quantities are of comparable magnitude. For example, the `Interior/Direct` solutions of problems such as CAR2, BRAINPC*, and READING1 (when formulated as (1.1)) display patterns in which $-c_i$ increases steadily with i while λ_i decreases steadily, or vice versa. It is difficult to tell at which index i the line should be drawn between activity and inactivity.

In our tables below, we define $\mathcal{A}^*$ by applying the LPEC-A test (3.28) with $\bar{\sigma} = .75$ and $\beta = 1/(m+n+p)$ to the solution returned by `Knitro`. (LPEC could be used in place of LPEC-A to estimate $\mathcal{A}^*$ because both schemes are theoretically guaranteed to return the true active set for x close enough to x^*.) We also wish to determine the weakly active inequality set $\mathcal{A}_0^*$, defined by (1.5). Procedure ID0 from [23, section 3], which involves repeated solution of linear programs, could be used to determine this set. However, for purposes of Table 4.6, we populated $\mathcal{A}_0^*$ with those indices in the estimated $\mathcal{A}^*$ that fail the test (4.3) when applied to the multipliers returned by LPEC-A at x^*. Note that this technique produces a superset of $\mathcal{A}_0^*$ in general.

4.3.2. Implementation details. The penalty parameter in the LP-P and LP-D formulations is defined by

$$\nu = 1.5 \max(\max_j(\lambda_j^*), \max_k(|\mu_k^*|), 1),$$

where (μ^*, λ^*) are the approximately optimal multipliers that were reported by the `Interior/Direct` algorithm. This heuristic guarantees that these particular multipliers (μ^*, λ^*) are feasible for the LP-D formulation (3.3) at the `Interior/Direct` approximate solution x^*. For the parameter M in (3.16) we use

$$M = 3 \max(\max_j(\lambda_j^*), \max_j(|c_j(x)|)).$$

Function and gradient evaluations are obtained through the Fortran and C tools contained in the CUTEr distribution and through a driver modeled after the routine `loqoma.c` (an interface for the code LOQO), which is also contained in CUTEr.

4.3.3. Test results and runtimes. Results for `noise` $= 10^{-3}$ are shown in Table 4.6. The numbers of elements in our estimate of the optimal active set $\mathcal{A}^*$ and weakly active set $\mathcal{A}_0^*$ are listed as $|\mathcal{A}^*|$ and $|\mathcal{A}_0^*|$. Each entry in the main part of the table contains the false positive/false negative count for each combination of test problem and identification technique. Table 4.7 shows the dimensions of each problem in the format $m/n/p$, with the main part of the table displaying runtimes in seconds. The LPEC column additionally reports the number of nodes beyond the root needed to solve the LPEC to the required (loose) tolerance. For many of the problems, the root node is within a factor of 2 of the optimal solution, and the reported number is therefore zero. If the LPEC test exceeds our time limit of 180 seconds, we qualify the approximate solution with the symbol "†".

Trends. In Table 4.6, the LP-D and LP-P results are nearly identical, indicating that the two methods usually find the same solution. The errors for both tests are mostly false negatives, which is expected, because the theory of section 3.1 gives no guarantee that weakly active constraints will be identified. Further, false positives are unlikely because the nice value of Δ (set up by the choice of parameter `DeltaFac` $= 4$) excludes most inactive constraints from the trust region. The number of false negatives for the $\mathcal{A}_\lambda$ test is usually higher than for $\mathcal{A}_c$, because weakly active constraints will generally fail the $\mathcal{A}_\lambda$ test (4.3), while they may pass the $\mathcal{A}_c$ test (4.2). This behavior

TABLE 4.6
CUTEr problems: False positives/false negatives. `noise` $= 10^{-3}$, $\sigma = 0.75$, $\bar{\sigma} = 0.90$, `DeltaFac` $= 4.00$.

Problem	$\lvert\mathcal{A}^*\rvert/\lvert\mathcal{A}_0{}^*\rvert$	LP-D		LP-P		LPEC-A	LPEC
		$\mathcal{A}_c$	$\mathcal{A}_\lambda$	$\mathcal{A}_c$	$\mathcal{A}_\lambda$		
A4X12	191/88	0/21	0/122	0/21	0/122	6/0	0/32
AVION2	21/5	0/6	0/9	0/6	0/10	7/0	4/0
BIGBANK	0/0	0/0	0/0	0/0	0/0	0/0	0/0
BRAINPC0	3/3	0/0	0/1	0/0	0/3	65/0	67/0
BRAINPC1	3/3	4/0	0/2	4/0	0/3	33/0	38/0
BRAINPC3	3/3	0/0	0/1	0/0	0/3	67/0	69/0
BRAINPC4	9/9	6/0	0/9	6/0	0/9	22/0	56/0
CAR2	883/1	0/312	0/883	0/321	0/883	0/112	53/0†
CORE1	21/3	0/0	0/7	0/0	0/7	0/0	0/0
CORKSCRW	505/6	0/3	0/190	0/3	0/189	0/3	0/6
C-RELOAD	136/7	0/38	0/124	0/38	0/124	0/19	0/18†
DALLASM	3/1	0/0	0/1	0/0	0/1	0/0	0/0
DALLASS	1/0	0/0	0/1	0/0	0/1	0/0	0/0
DEMBO7	21/8	0/1	0/7	0/1	0/11	0/0	0/1
FEEDLOC	20/19	0/0	0/19	0/0	0/19	0/7	0/0
GMNCASE4	350/175	0/0	0/175	0/0	0/175	0/0	0/0
GROUPING	100/100	0/0	0/44	0/0	0/44	0/8	0/0
HANGING	2310/40	0/48	0/72	0/48	0/72	0/12	0/68
HELSBY	8/2	0/0	0/5	0/0	0/1	0/0	0/0
HIMMELBK	20/10	0/0	0/10	0/0	0/9	0/0	1/0
HUES-MOD	277/0	0/1	0/78	0/1	0/78	0/1	0/277
KISSING2	181/87	0/0	0/88	0/0	0/88	0/0	0/2
LISWET10	1999/0	0/2	0/237	0/2	0/254	1/0	0/6
LSNNODOC	3/1	0/0	0/1	0/0	0/1	0/0	0/0
MAKELA3	20/19	0/0	0/19	0/0	0/19	0/0	0/20
MINPERM	0/0	0/0	0/0	0/0	0/0	0/0	0/0
NET1	7/2	0/0	0/2	0/0	0/2	0/0	0/0
NGONE	102/0	0/0	0/86	0/0	0/86	0/0	0/5
OET7	110/105	0/15	0/105	0/15	0/105	38/21	86/20
POLYGON	105/4	0/0	0/4	0/0	0/4	0/0	0/17
PRIMALC8	505/2	0/4	0/505	0/4	0/505	0/0	0/0
PRODPL0	39/0	0/0	0/0	0/0	0/0	0/0	0/0
QPCBLEND	80/42	0/24	0/45	0/24	0/45	0/12	0/24
QPCBOEI1	309/49	0/0	0/47	0/0	0/49	4/0	0/18
QPCSTAIR	163/20	0/50	0/59	0/50	0/56	20/0	0/51
READING1	174/147	0/173	0/174	0/173	0/174	0/141	0/86†
SARO	675/2	0/43	0/675	0/43	0/675	0/44	0/58†
SAROMM	343/0	0/0	0/343	0/0	0/343	0/0	0/10†
SMBANK	0/0	0/0	0/0	0/0	0/0	0/0	0/0
SMMPSF	481/1	0/5	0/66	0/5	0/66	0/1	0/10
SOSQP1	2500/2500	0/2500	0/2500	0/2500	0/2500	0/2500	0/0
SREADIN3	180/154	0/180	0/180	0/180	0/180	0/146	0/104†
SSEBNLN	133/25	0/2	0/35	0/2	0/25	0/0	0/2
STEENBRA	381/95	0/0	0/55	0/0	0/51	0/0	0/0
TRIMLOSS	94/51	1/69	0/93	0/67	0/93	0/7	0/33
TRUSPYR2	8/1	0/0	0/1	0/0	0/0	0/0	4/0
TWIRIMD1	660/80	0/257	0/659	0/258	0/659	0/56	0/56†
TWIRISM1	140/29	0/15	0/83	0/15	0/84	0/15	0/18
ZAMB2	1259/0	0/673	0/1259	0/673	0/1259	0/102	0/102†

TABLE 4.7
CUTEr problems: Time (secs). `noise` $= 10^{-3}$, $\sigma = 0.75$, $\bar{\sigma} = 0.90$, `DeltaFac` $= 4.00$.

Problem	$m/n/p$	$\lvert\mathcal{A}^*\rvert/\lvert\mathcal{A}_0{}^*\rvert$	LP-D	LP-P	LPEC-A	LPEC/nodes
A4X12	384/ 130/ 16	191/ 88	0.02	0.03	0.01	5.62/66
AVION2	98/ 49/ 15	21/ 5	0.00	0.00	0.00	0.03/5
BIGBANK	3844/2230/1420	0/ 0	0.42	0.17	0.12	1.16/0
BRAINPC0	6905/6907/6902	3/ 3	3.98	6.82	22.52	31.47/0
BRAINPC1	6905/6907/6902	3/ 3	4.84	18.81	0.44	1.44/0
BRAINPC3	6905/6907/6902	3/ 3	2.87	6.64	25.42	58.82/0
BRAINPC4	6905/6907/6902	9/ 9	4.06	5.98	8.75	2.95/0
CAR2	4997/5999/4004	883/ 1	5.13	0.54	0.65	$189^{\dagger}$/7065
CORE1	139/ 65/ 41	21/ 3	0.00	0.00	0.00	0.01/0
CORKSCRW	4500/4506/3009	505/ 6	0.92	0.80	0.19	93.25/1
C-RELOAD	684/ 342/ 200	136/ 7	0.10	0.07	0.04	$181^{\dagger}$/40420
DALLASM	392/ 196/ 151	3/ 1	0.01	0.01	0.00	0.13/0
DALLASS	92/ 46/ 31	1/ 0	0.00	0.00	0.00	0.03/0
DEMBO7	53/ 16/ 0	21/ 8	0.00	0.00	0.00	0.03/1
FEEDLOC	462/ 90/ 22	20/ 19	0.00	0.00	0.00	0.18/0
GMNCASE4	350/ 175/ 0	350/ 175	0.05	0.08	0.04	0.12/0
GROUPING	200/ 100/ 125	100/ 100	0.00	0.00	0.00	0.01/0
HANGING	2330/3600/ 12	2310/ 40	27.40	44.70	6.46	10.04/0
HELSBY	685/1408/1399	8/ 2	0.22	0.20	0.03	0.50/0
HIMMELBK	24/ 24/ 14	20/ 10	0.00	0.00	0.00	0.00/0
HUES-MOD	5000/5000/ 2	277/ 0	1.34	0.15	0.24	1.08/0
KISSING2	625/ 100/ 6	181/ 87	0.01	0.02	0.01	14.89/492
LISWET10	2000/2002/ 0	1999/ 0	0.31	0.12	0.31	20.04/0
LSNNODOC	6/ 5/ 4	3/ 1	0.00	0.00	0.00	0.01/0
MAKELA3	20/ 21/ 0	20/ 19	0.00	0.00	0.00	0.00/0
MINPERM	1213/1113/1033	0/ 0	0.20	0.06	0.11	15.55/0
NET1	65/ 48/ 43	7/ 2	0.00	0.00	0.00	0.01/0
NGONE	5246/ 200/ 3	102/ 0	0.02	0.03	0.02	21.18/1
OET7	1002/ 7/ 0	110/ 105	0.01	0.02	0.00	0.29/0
POLYGON	5445/ 200/ 2	105/ 4	0.02	0.03	0.02	28.98/1
PRIMALC8	511/ 520/ 0	505/ 2	0.04	0.03	0.01	0.11/0
PRODPL0	69/ 60/ 20	39/ 0	0.00	0.00	0.00	0.02/0
QPCBLEND	114/ 83/ 43	80/ 42	0.01	0.01	0.00	0.50/141
QPCBOEI1	971/ 384/ 9	309/ 49	0.03	0.02	0.02	17.07/26
QPCSTAIR	532/ 467/ 291	163/ 20	0.05	0.03	0.02	2.65/38
READING1	8002/4002/2001	174/ 147	0.61	0.36	0.22	$192^{\dagger}$/2600
SARO	2920/4754/4025	675/ 2	3.17	4.57	2.67	$182^{\dagger}$/347
SAROMM	2920/5120/4390	343/ 0	4.68	7.13	1.71	$182^{\dagger}$/19
SMBANK	234/ 117/ 64	0/ 0	0.01	0.00	0.00	0.02/0
SMMPSF	743/ 720/ 240	481/ 1	0.11	0.04	0.03	4.82/255
SOSQP1	10000/5000/2501	2500/2500	0.08	0.08	0.30	7.94/0
SREADIN3	8004/4002/2001	180/ 154	0.80	0.32	0.23	$187^{\dagger}$/2385
SSEBNLN	384/ 194/ 74	133/ 25	0.01	0.01	0.01	0.03/0
STEENBRA	432/ 432/ 108	381/ 95	0.02	0.01	0.01	0.44/1
TRIMLOSS	319/ 142/ 20	94/ 51	0.00	0.00	0.01	0.45/31
TRUSPYR2	16/ 11/ 3	8/ 1	0.00	0.00	0.00	0.01/0
TWIRIMD1	2685/1247/ 521	660/ 80	3.86	0.97	1.02	$182^{\dagger}$/110
TWIRISM1	775/ 343/ 224	140/ 29	0.09	0.06	0.04	18.87/501
ZAMB2	7920/3966/1446	1259/ 0	0.50	0.14	0.21	$190^{\dagger}$/2025

is highlighted in the results for the problems GMNCASE4 and OET7. Their numbers of false negatives for the $\mathcal{A}_\lambda$ test correspond exactly to $|\mathcal{A}_0^*|$, while the corresponding numbers of false negatives for the $\mathcal{A}_c$ test are much lower.

In contrast to the results for LP-D and LP-P, the results for LPEC-A and LPEC show a mixture of false positives and false negatives. Further, the results for LPEC-A and LPEC are similar for most problems. For several problems, for example, TWIRIMD1 and ZAMB2, the results for LPEC-A agree with those for LPEC but not with those for LP-D and LP-P.

The runtimes given in Table 4.7 are typically much shorter than for the random problems in Tables 4.3 and 4.5 because the constraint Jacobians in the CUTEr problems are usually sparse (OET7 is an exception). The LP-D times are similar to the LP-P times, and LPEC-A times are generally comparable. With few exceptions, LPEC requires more execution time than LPEC-A. In cases for which LPEC requires significantly more time than LPEC-A, the LPEC identification performance is not better in general.

The LPEC method is usually the slowest, despite initialization from a good starting point. (The use of this starting point reduces significantly the solve time and the number of searched nodes for several problems, including HANGING, NGONE, SMMPSF, TRIMLOSS, and TWIRISM1.)

Anomalies. For several problems, the numbers of false negatives for LP-D and LP-P with test $\mathcal{A}_\lambda$ are larger than $|\mathcal{A}_0^*|$; see, for example, SREADIN3. This may happen because the LP-P and LP-D programs find a sparse λ, one that has many more zeros than the λ^* that we used to form our estimate of $\mathcal{A}_0^*$ as described above.

For certain problems, *all* methods return large numbers of false negatives. These problems often contain many bound constraints, for example, C-RELOAD, READING1, SREADIN3, TRIMLOSS, TWIRIMD1, and ZAMB2. We note that these errors still occurred when we reformulated the tests to treat the bound constraints explicitly.

For LPEC-A and LPEC, the BRAINPC* and OET7 problems have many false positives, as a result of many inactive constraints having values of $c_i(x)$ close to zero, below the threshold for determining activity.

For the problem SOSQP1, a quadratic program, only the LPEC method detects any active constraints; in fact, it makes no identification errors. A smaller choice for the parameter $\bar{\sigma}$ in the LPEC-A identification test would produce perfect identification for the LPEC-A technique also.

We remark on a few more of the anomalies in Table 4.7. Runtimes for HANGING are especially large, given its size. The LP solvers performed many perturbations, and the MIP solver for the LPEC test reports trouble identifying an integer solution. For LPEC, an extremely large number of iterations and nodes is reported for C-RELOAD, again due to difficulty in finding feasible integer solutions. Allowing the use of heuristics by the MIP solver yielded a large reduction in the number of considered nodes for this problem, but the runtime did not change significantly.

4.4. Additional remarks. We conclude this section with some general comments on the numerical results and on additional testing not reported above.

In general, the LP-P and LP-D tests give similar identification results, with a tendency to underestimate the active set (that is, false negatives). The primal activity test $\mathcal{A}_c$ is superior to the dual activity test $\mathcal{A}_\lambda$ for these methods. LP-P tended to take less time to solve, probably because of the greater reductions due to presolving.

We tested the effect of using a much larger Δ in the LP-P and LP-D formulations for the random test problems. Runtimes were slightly longer on the largest problems,

and the time advantage that LP-P has for smaller Δ disappears. The $\mathcal{A}_\lambda$ activity test returned the same poor underestimate of the active set as for the smaller Δ, while the $\mathcal{A}_c$ activity test made many more identification errors.

The LPEC-A test obviously is preferable to LPEC, as the results are similar (with the anomalies easily explained) and the runtimes are sometimes much shorter. We note that it might be possible to improve the performance of these methods by better scaling of the constraints.

We used a `noise` value of 10^{-3} in all reported results, but performed additional experiments with other values of this parameter. For smaller values of `noise`, LP-P and LP-D tend to have similar false positive counts, but show higher false negative counts in some cases. LPEC and LPEC-A show an overall improvement; for example, at `noise` $= 10^{-7}$ the BRAINPC* problems' results for LPEC and LPEC-A are nearly perfect. However, more false negatives are reported on some CUTEr problems. These difficult problems are the ones for which our estimate of the true active set $\mathcal{A}^*$ is sensitive to the parameters β, σ, and $\bar{\sigma}$ used in the threshold test (see subsection 4.3.1), and for which the estimate of the true active set changes significantly if we use LPEC in place of LPEC-A. Specifically, on problems LISWET10, OET7, and READING1, the additional false negatives that were reported when `noise` was decreased from 10^{-3} to 10^{-7} disappeared when $\bar{\sigma}$ was changed or when LPEC was used in place of LPEC-A in the determination of $\mathcal{A}^*$.

As expected, the results of the random problems in Tables 4.2 and 4.3 for the LPEC and LPEC-A techniques are nearly perfect for `noise` $= 10^{-7}$. (`noise` must be decreased to an even smaller value to remove a single false positive in some cases; this identification error is caused by a constraint that is only very slightly inactive.)

For values of `noise` larger than 10^{-3}, LP-P and LP-D report more false positives on the random problems and fewer false negatives on the CUTEr problems. The LPEC and LPEC-A tests tend to give more false positives, while the false negative count decreases on the CUTEr problems and increases on the random problems.

Following a suggestion of a referee, and in line with the discussion at the end of section 3, we obtained a new identification technique by inserting the solution of (3.3) in place of (μ_x, λ_x) in the threshold test (3.28). We found that, indeed, this "threshold LP-D" estimate of the active set was more accurate than those obtained from (3.5a) and (3.5b), as is done in the standard LP-D technique. On the random problem set, the results for threshold LP-D for `noise` $= 10^{-7}$ are identical to those for LPEC-A, in accordance with our claim that both techniques are asymptotically exact.

5. Conclusions. We have described several schemes for predicting the active set for a nonlinear program with inequality constraints, given an estimate x of a solution x^*. The effectiveness of some of these schemes in identifying the correct active set for x sufficiently close to x^* is proved, under certain assumptions. In particular, the scheme of subsection 3.3 has reasonable computational requirements and strong identification properties and appears to be novel. Computational tests are reported which show the properties of the various schemes on random problems and on degenerate problems from the CUTEr test set.

Knowledge of the correct active set considerably simplifies algorithms for inequality constrained nonlinear programming, as it removes the "combinatorial" aspect from the problem. However, it remains to determine how the schemes above can be used effectively as an element of a practical algorithm for solving nonlinear programs. It may be that reliable convergence can be obtained in general without complete knowl-

edge of the active set; some "sufficient subset" may suffice. What are the required properties of such a subset, and can we devise inexpensive identification schemes, based on the ones described in this paper, that identify it? We leave these and other issues to future research.

Appendix A. Proof of (3.10).

We prove this statement by contradiction. Suppose that there is a sequence $\{x^k\}$ with $x^k \to x^*$ such that

$$\text{dist } \left(-g(x^k), \text{range}[\nabla h(x^k)] + \text{pos}[(\nabla c_i(x^k))_{i\in\mathcal{A}_1}]\right) < \tau \tag{A.1}$$

for all k. By closedness, there must be vectors z^k and $y^k \geq 0$ such that the

$$\begin{aligned}&\text{dist } \left(-g(x^k), \text{range}[\nabla h(x^k)] + \text{pos}[(\nabla c_i(x^k))_{i\in\mathcal{A}_1}]\right)\\ &= \left\|\nabla h(x^k)z^k + \sum_{i\in\mathcal{A}_1} \nabla c_i(x^k)y_i^k + g(x^k)\right\| \leq \tau\end{aligned}$$

for all k. If $\{(z^k, y^k)\}$ is unbounded, we have by compactness of the unit ball, and by taking a subsequence if necessary, that $\|(z^k, y^k)\| \uparrow \infty$ and $(z^k, y^k)/\|(z^k, y^k)\| \to (z^*, y^*)$ with $\|(z^*, y^*)\| = 1$ and $y^* \geq 0$. Hence, by dividing both sides in the expression above by $\|(z^k, y^k)\|$ and taking limits, we have

$$(\nabla h^*)z^* + \sum_{i\in\mathcal{A}_1} (\nabla c_i^*)y_i^* = 0. \tag{A.2}$$

From Lemma 3.1, we have $\mathcal{A}_1 \subset \mathcal{A}^*$, so that the MFCQ condition (1.7) holds at x^* for $\mathcal{A}_1$ replacing $\mathcal{A}^*$. Hence, for the vector v in this condition, we have that $\nabla h(x^*)$ has full column rank and that $(\nabla h^*)^T v = 0$ and $\nabla(c_i^*)^T v < 0$ for all $i \in \mathcal{A}_1$. By taking inner products of (A.2) with v, we can deduce first that $y^* = 0$ and subsequently that $z^* = 0$, by a standard argument, contradicting $\|(z^*, y^*)\| = 1$. Therefore, the sequence $\{(z^k, y^k)\}$ must be bounded. Since the sequence remains in a ball about the origin (that is, a compact set), it has an accumulation point.

By taking a subsequence again if necessary, suppose that $(z^k, y^k) \to (\hat{z}, \hat{y})$. We then have that

$$\begin{aligned}&\left\|\nabla h(x^k)\hat{z} + \sum_{i\in\mathcal{A}_1} \nabla c_i(x^k)\hat{y}_i + g(x^k)\right\|\\ &\leq \left\|\nabla h(x^k)z^k + \sum_{i\in\mathcal{A}_1} \nabla c_i(x^k)y_i^k + g(x^k)\right\| + \|\nabla h(x^k)\|\|z^k - \hat{z}\|\\ &\qquad\qquad + \sum_{i\in\mathcal{A}_1} \|\nabla c_i(x^k)\|\,|y_i^k - \hat{y}_i|\\ &\leq \tau + o(1)\end{aligned}$$

for all k sufficiently large. By taking limits in this expression, we deduce that

$$\text{dist } (-g^*, \text{range}[\nabla h^*] + \text{pos}[(\nabla c_i^*)_{i\in\mathcal{A}_1}]) \leq \tau,$$

which contradicts the definition of τ, for $\tau > 0$. Hence, a sequence $\{x^k\}$ with the property (A.1) cannot exist, so (3.10) holds for all $\bar{\epsilon}_2$ sufficiently small. ∎

Acknowledgments. We thank Richard Waltz for many discussions during the early part of this project, for supplying us with the `Knitro` results, and for advice about the implementations. We also thank Dominique Orban for providing helpful advice about using CUTEr. Finally, we are most grateful to two anonymous referees for extremely thorough and helpful comments on the first version of this paper.

REFERENCES

[1] D. P. Bertsekas, *Projected Newton methods for optimization problems with simple constraints*, SIAM J. Control Optim., 20 (1982), pp. 221–246.
[2] J. Burke, *On the identification of active constraints* II: *The nonconvex case*, SIAM J. Numer. Anal., 27 (1990), pp. 1081–1102.
[3] J. V. Burke and J. J. Moré, *On the identification of active constraints*, SIAM J. Numer. Anal., 25 (1988), pp. 1197–1211.
[4] J. V. Burke and J. J. Moré, *Exposing constraints*, SIAM J. Optim., 4 (1994), pp. 573–595.
[5] R. Byrd, N. I. M. Gould, J. Nocedal, and R. A. Waltz, *An algorithm for nonlinear optimization using linear programming and equality constrained subproblems*, Math. Program., 100 (2004), pp. 27–48.
[6] R. H. Byrd, N. I. M. Gould, J. Nocedal, and R. A. Waltz, *On the convergence of successive linear-quadratic programming algorithms*, SIAM J. Optim., 16 (2005), pp. 471–489.
[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] A. R. Conn, N. I. M. Gould, and P. H. L. Toint, *Trust-Region Methods*, MPS/SIAM Ser. Optim. 1, SIAM, Philadelphia, 2000.
[9] A. S. El-Bakry, R. A. Tapia, and Y. Zhang, *A study of indicators for identifying zero variables in interior-point methods*, SIAM Rev., 36 (1994), pp. 45–72.
[10] F. Facchinei, A. Fischer, and C. Kanzow, *On the accurate identification of active constraints*, SIAM J. Optim., 9 (1998), pp. 14–32.
[11] R. Fletcher, *Practical Methods of Optimization*, 2nd ed., John Wiley and Sons, New York, 1987.
[12] R. Fletcher and E. Sainz de la Maza, *Nonlinear programming and nonsmooth optimization by successive linear programming*, Math. Programming, 43 (1989), pp. 235–256.
[13] N. I. M. Gould, D. Orban, and P. L. Toint, *CUTEr (and SifDec), a Constrained and Unconstrained Testing Environment, Revisited*, Technical report TR/PA/01/04, CERFACS, Toulouse, France, 2001.
[14] W. W. Hager and M. S. Gowda, *Stability in the presence of degeneracy and error estimation*, Math. Program., 85 (1999), pp. 181–192.
[15] W. Hare and A. Lewis, *Identifying active constraints via partial smoothness and prox-regularity*, J. Convex Anal., 11 (2004), pp. 251–266.
[16] M. Lescrenier, *Convergence of trust region algorithms for optimization with bounds when strict complementarity does not hold*, SIAM J. Numer. Anal., 28 (1991), pp. 476–495.
[17] A. S. Lewis, *Active sets, nonsmoothness, and sensitivity*, SIAM J. Optim., 13 (2002), pp. 702–725.
[18] R. D. C. Monteiro and S. J. Wright, *Local convergence of interior-point algorithms for degenerate monotone LCP*, Comput. Optim. Appl., 3 (1994), pp. 131–155.
[19] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[20] R. Waltz, *An Active-Set Trust-Region Algorithm for Nonlinear Optimization*, Presentation at the 18th International Symposium on Mathematical Programming (ISMP), Copenhagen, Denmark, 2003.
[21] S. J. Wright, *Identifiable surfaces in constrained optimization*, SIAM J. Control Optim., 31 (1993), pp. 1063–1079.
[22] S. J. Wright, *Modifying SQP for degenerate problems*, SIAM J. Optim., 13 (2002), pp. 470–497.
[23] S. J. Wright, *Constraint identification and algorithm stabilization for degenerate nonlinear programs*, Math. Program., 95 (2003), pp. 137–160.
[24] N. Yamashita, H. Dan, and M. Fukushima, *On the identification of degenerate indices in the nonlinear complementarity problem with the proximal point algorithm*, Math. Program., 99 (2004), pp. 377–397.
[25] Y. Ye, *On the finite convergence of interior-point algorithms for linear programming*, Math. Programming, 57 (1992), pp. 325–336.

SIAM J. OPTIM.
Vol. 17, No. 2, pp. 606–619
© 2006 Society for Industrial and Applied Mathematics

CONVERGENCE OF MESH ADAPTIVE DIRECT SEARCH TO SECOND-ORDER STATIONARY POINTS*

MARK A. ABRAMSON† AND CHARLES AUDET‡

Abstract. A previous analysis of second-order behavior of generalized pattern search algorithms for unconstrained and linearly constrained minimization is extended to the more general class of mesh adaptive direct search (MADS) algorithms for general constrained optimization. Because of the ability of MADS to generate an asymptotically dense set of search directions, we are able to establish reasonable conditions under which a subsequence of MADS iterates converges to a limit point satisfying second-order necessary or sufficient optimality conditions for general set-constrained optimization problems.

Key words. nonlinear programming, mesh adaptive direct search, derivative-free optimization, convergence analysis, second-order optimality conditions

AMS subject classifications. 90C30, 90C56, 65K05

DOI. 10.1137/050638382

1. Introduction. In this paper, we consider the class of derivative-free mesh adaptive direct search (MADS) algorithms applied to general constrained optimization problems of the form

$$\min_{x\in\Omega} f(x) \tag{1.1}$$

with $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and $\Omega \subseteq \mathbb{R}^n$.

We treat the constraints by the "barrier" approach of applying the algorithm, not to f, but to the barrier objective function $f_\Omega = f + \psi_\Omega$, where ψ_Ω is the indicator function for Ω; i.e., it is zero on Ω, and infinity elsewhere. If a point x is not in Ω, then we set $f_\Omega(x) = \infty$, and f is not evaluated. This is important in many practical engineering problems in which f is expensive to evaluate.

The class of MADS algorithms was introduced and analyzed in [4], as an extension of generalized pattern search (GPS) algorithms [3, 21] for solving nonlinearly constrained problems. Rather than applying a penalty function [18] or filter [5] approach to handle the nonlinear constraints, MADS defines an additional parameter that enables the algorithm to perform an exploration of the space of variables in an asymptotically dense set of directions. Under mild assumptions, the Clarke [9] calculus together with three types of tangent cones (hypertangent, Clarke tangent, and

*Received by the editors August 18, 2005; accepted for publication (in revised form) February 17, 2006; published electronically August 16, 2006. This work was performed by an employee of the U.S. Government or under U.S. Government contract. The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. Copyright is owned by SIAM to the extent not limited by these rights.

http://www.siam.org/journals/siopt/17-2/63838.html

†Department of Mathematics and Statistics, Air Force Institute of Technology, 2950 Hobson Way, Wright Patterson AFB, OH 45433 (Mark.Abramson@afit.edu, http://www.afit.edu/en/ENC/Faculty/MAbramson/abramson.html).

‡Département de Mathématiques et de Génie Industriel, École Polytechnique de Montréal and GERAD, C.P. 6079, Succ. Centre-ville, Montréal QB H3C 3A7, Canada (Charles.Audet@gerad.ca, http://www.gerad.ca/Charles.Audet). The work of this author was supported by FCAR grant NC72792 and NSERC grant 239436-05, AFOSR F49620-01-1-0013, and ExxonMobil Upstream Research Company.

contingent cones) is used to prove convergence of a subsequence of iterates to a point satisfying certain first-order conditions for optimality. An implementable instance of MADS is introduced in [4], in which positive spanning directions are chosen in a random fashion and almost sure convergence to a first-order stationary point is obtained. A similar first-order analysis is done in [15] for the DIRECT algorithm.

This paper extends the MADS analysis to show convergence to points satisfying certain second-order stationarity properties, in a manner similar to that of [1] for GPS. An important result of [1] is that the iterates produced by a GPS algorithm on a sufficiently smooth problem cannot converge in an infinite number of steps to a local maximizer. We show here that it may, unfortunately, converge in an infinite number of steps to a saddle point. The analysis in the present paper gives sufficient conditions under which a subsequence of the iterates produced by a MADS algorithm converges to a strict local minimizer. The necessary optimality condition is not based on any of the three tangent cones used in [4] but rather on the cone of feasible directions.

The paper is outlined as follows. The MADS algorithm is briefly described in section 2, with first-order properties restated in section 3. Section 4 introduces the generalized Hessian [16] with some associated properties, followed by second-order necessary and sufficient optimality conditions and convergence results. Section 5 provides some examples to illustrate the strength of these results, and section 6 offers some concluding remarks.

Notation. $\mathbb{R}$, $\mathbb{Z}$, and $\mathbb{N}$ denote the set of real numbers, integers, and nonnegative integers, respectively. For any set S, $\text{int}(S)$ denotes its interior, and $\text{cl}(S)$ its closure. For any matrix A, the notation $a \in A$ means that a is a column of A. For $x \in \mathbb{R}^n$ and $\varepsilon > 0$, we denote by $B_\varepsilon(x)$ the open ball $\{y \in \mathbb{R}^n : \|y - x\| < \epsilon\}$. We say that f is $C^{1,1}$ near x if there exists an open set S containing x such that f is continuously differentiable with Lipschitz derivatives for every point in S. The reader is invited to consult [16] for a discussion and examples of $C^{1,1}$ functions.

2. Mesh adaptive direct search. Like GPS methods, each iteration k of a MADS algorithm is characterized by two steps—an optional SEARCH step and a local POLL step, in which f_Ω is evaluated at specified points that lie on a mesh. The mesh is constructed from a finite fixed set of n_D directions $D \subset \mathbb{R}^n$ scaled by a mesh size parameter $\Delta_k^m > 0$. The directions form a positive spanning set [14] (i.e., nonnegative linear combinations of D must span $\mathbb{R}^n$), and each direction $d \in D$ must be constructed as the product Gz, where $G \in \mathbb{R}^{n\times n}$ is a nonsingular generating matrix and $z \in \mathbb{Z}^n$ is a vector of integers.

The following definition, taken from [4] and [5], precisely defines the current mesh so that all previously visited points lie on the current mesh.

DEFINITION 2.1. *At iteration k, the* current mesh *is defined to be the following union:*

$$M_k = \bigcup_{x \in S_k} \{x + \Delta_k^m D z : z \in \mathbb{N}^{n_D}\},$$

where S_k is the finite set of points where the objective function f has been evaluated by the start of iteration k and S_0 is a finite set of initial feasible points.

In both the SEARCH and POLL steps, the algorithm seeks to find an *improved mesh point*; i.e., a point $y \in M_k$ for which $f_\Omega(y) < f_\Omega(x_k)$, where x_k is the current iterate or incumbent best iterate found thus far.

The SEARCH step allows evaluation of f_Ω at any finite set of mesh points. Any strategy may be used, including none. This is more restrictive than the frame meth-

ods of Coope and Price [12], but it helps to ensure convergence without a sufficient decrease condition or any other assumptions on mesh directions. The SEARCH step adds nothing to the convergence theory, but well-chosen SEARCH strategies can greatly improve algorithm performance (see [2, 6, 7, 19]).

In the POLL step, f_Ω is evaluated at points adjacent to the current iterate in a subset of the mesh directions. Unlike GPS, the class of MADS algorithms has a second mesh parameter Δ_k^p, called the *poll size parameter*, which satisfies $\Delta_k^m \leq \Delta_k^p$ for all k, and also

$$\lim_{k\in K} \Delta_k^m = 0 \Leftrightarrow \lim_{k\in K} \Delta_k^p = 0 \text{ for any infinite subset of indices } K. \tag{2.1}$$

Under this construction, GPS methods now become the specific MADS instance in which $\Delta_k = \Delta_k^p = \Delta_k^m$.

The set of points generated in the POLL step is called a *frame*, with x_k referred to as the *frame center*. These terms are now formally defined as follows.

DEFINITION 2.2. *At iteration k, the* MADS frame *is defined to be the set*

$$P_k = \{x_k + \Delta_k^m d : d \in D_k\} \subset M_k,$$

where D_k is a positive spanning set such that for each $d \in D_k$, the following hold:

- *$d \neq 0$ can be written as a nonnegative integer combination of the directions in D: $d = Du$ for some vector $u \in \mathbb{N}^{n_D}$ that may depend on the iteration number k.*
- *The distance from the frame center x_k to a poll point $x_k + \Delta_k^m d$ is bounded by a constant times the poll size parameter: $\Delta_k^m \|d\| \leq \Delta_k^p \max\{\|d'\| : d' \in D\}$.*
- *Limits (as defined in Coope and Price* [11]*) of the normalized sets D_k are positive spanning sets.*

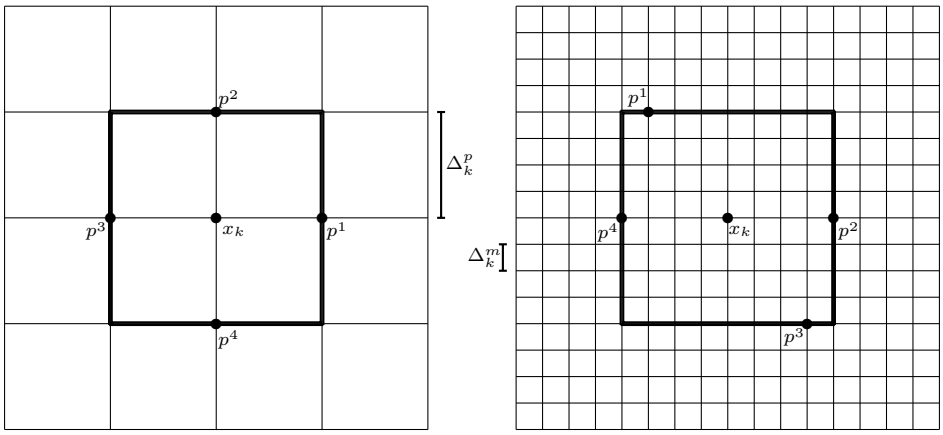
In GPS, the set of directions D_k used to construct the frame is a subset of the finite set D. There is more flexibility in MADS. In [4], an instance of MADS is presented in which the closure of the cone generated by the set $\bigcup_{k=1}^{\infty}\{\frac{d}{\|d\|} : d \in D_k\}$ equals $\mathbb{R}^n$. In this case, we say that the set of poll directions is *asymptotically dense* in $\mathbb{R}^n$.

Figure 2.1 illustrates typical GPS and MADS frames in $\mathbb{R}^2$ using the standard $2n$ coordinate directions. In each case, the mesh M_k is the set of points at the intersections of the horizontal and vertical lines. The thick lines delimit the points that are at a relative distance equal to the poll size parameter Δ_k^p from the frame center x_k. In MADS, the mesh size parameter Δ_k^m is much smaller than the poll size parameter; this allows many more possibilities in the frame construction.

If the POLL step fails to produce an improved mesh point, P_k is said to be a *minimal frame* with *minimal frame center* x_k. If either the SEARCH or POLL step is successful in finding an improved mesh point, the improved mesh point becomes the new current iterate $x_{k+1} \in \Omega$ and the mesh is either retained or coarsened. If neither step is successful, then the minimal frame center is retained as the current iterate (i.e., $x_{k+1} = x_k$) and the mesh is refined.

Rules for refining and coarsening the mesh are as follows. Given a fixed rational number $\tau > 1$ and two integers $w^- \leq -1$ and $w^+ \geq 0$, the mesh size parameter Δ_k^m is updated according to the rule

$$\Delta_{k+1}^m = \tau^{w_k}\Delta_k^m$$
$$\text{for some } w_k \in \begin{cases} \{0, 1, \ldots, w^+\} & \text{if an improved mesh point is found,} \\ \{w^-, w^- + 1, \ldots, -1\} & \text{otherwise.} \end{cases} \tag{2.2}$$



A GPS frame A MADS frame

FIG. 2.1. *GPS and MADS frames* $P_k = \{p^1, p^2, p^3, p^4\}$ *around the frame center* x_k*, with the same poll size parameter* Δ_k^p*.*

The class of MADS algorithms is stated formally as follows:

A GENERAL MADS ALGORITHM

0. INITIALIZATION: Let $x_0 \in \Omega$, set $\Delta_0^p \geq \Delta_0^m > 0$. Set iteration counter to $k = 0$.
1. SEARCH AND POLL STEP: Perform the SEARCH and POLL steps to completion, or until an improved mesh point x_{k+1} is found on the mesh M_k (see Definition 2.1).
   - OPTIONAL SEARCH: Evaluate f_Ω on a finite subset of trial points on the mesh M_k.
   - LOCAL POLL: Evaluate f_Ω on the frame P_k (see Definition 2.2).
2. PARAMETER UPDATE: Update Δ_{k+1}^m according to (2.2), and Δ_{k+1}^p according to (2.1). Increment $k \leftarrow k + 1$ and go to step 1.

3. Existing first-order stationarity results. Before presenting new results, we reproduce known convergence properties of MADS, originally published in [4]. All results are based on the following assumptions:

A1. A feasible initial point x_0 is provided.

A2. The initial objective function value $f(x_0)$ is finite.

A3. All iterates $\{x_k\}$ generated by MADS lie in a compact set.

Under these assumptions, Audet and Dennis [4] proved that

$$\liminf_{k \to +\infty} \Delta_k^p \;=\; \liminf_{k \to +\infty} \Delta_k^m \;=\; 0.$$

This ensures the existence of infinitely many minimal frame centers, since Δ_k^m shrinks only when a minimal frame is found. The following definition, taken from [4], is needed for later results.

DEFINITION 3.1. *A subsequence of the MADS iterates consisting of minimal frame centers,* $\{x_k\}_{k \in K}$ *for some subset of indices* K*, is said to be a* refining subsequence *if* $\{\Delta_k^p\}_{k \in K}$ *converges to zero.*

Let $\hat{x}$ *be the limit of a convergent refining subsequence. If* $\lim_{k \in L} \frac{d_k}{\|d_k\|}$ *exists for some subset* $L \subseteq K$ *with poll direction* $d_k \in D_k$*, and if* $x_k + \Delta_k^m d_k \in \Omega$ *for infinitely many* $k \in L$*, then this limit is said to be a* refining direction *for* $\hat{x}$*.*

Existence of refining subsequences for MADS was proved in [4]. The following four definitions [9, 17, 20] are needed in the main theorems.

DEFINITION 3.2. *A vector $v \in \mathbb{R}^n$ is said to be a* hypertangent vector *to the set $\Omega \subset \mathbb{R}^n$ at the point $x \in \Omega$ if there exists a scalar $\epsilon > 0$ such that*

$$y + tw \in \Omega \qquad \text{for all } y \in \Omega \cap B_\epsilon(x), \quad w \in B_\epsilon(v), \quad \text{and} \quad 0 < t < \epsilon. \tag{3.1}$$

The set of hypertangent vectors to Ω at x is called the hypertangent cone to Ω at x *and is denoted by $T_\Omega^H(x)$.*

DEFINITION 3.3. *A vector $v \in \mathbb{R}^n$ is said to be a* Clarke tangent vector *to the set $\Omega \subset \mathbb{R}^n$ at the point $x \in \mathrm{cl}(\Omega)$ if for every sequence $\{y_k\}$ of elements of Ω that converges to x and for every sequence of positive real numbers $\{t_k\}$ converging to zero, there exists a sequence of vectors $\{w_k\}$ converging to v such that $y_k + t_k w_k \in \Omega$. The set $T_\Omega^{Cl}(x)$ of all Clarke tangent vectors to Ω at x is called the* Clarke tangent cone *to Ω at x.*

DEFINITION 3.4. *A vector $v \in \mathbb{R}^n$ is said to be a* tangent vector *to the set $\Omega \subset \mathbb{R}^n$ at the point $x \in \mathrm{cl}(\Omega)$ if there exists a sequence $\{y_k\}$ of elements of Ω that converges to x and a sequence of positive real numbers $\{\lambda_k\}$ for which $v = \lim_k \lambda_k(y_k - x)$. The set $T_\Omega^{Co}(x)$ of all tangent vectors to Ω at x is called the* contingent cone *(or* sequential Bouligand tangent cone*) to Ω at x.*

DEFINITION 3.5. *The set Ω is said to be* regular *at x if $T_\Omega^{Cl}(x) = T_\Omega^{Co}(x)$.*

In addition to these definitions, we add the following clarifying remarks, due to Clarke [9] unless otherwise noted:

- Any convex set is regular at each of its points.
- Both $T_\Omega^{Co}(x)$ and $T_\Omega^{Cl}(x)$ are closed, and both $T_\Omega^{Cl}(x)$ and $T_\Omega^H(x)$ are convex.
- $T_\Omega^H(x) \subseteq T_\Omega^{Cl}(x) \subseteq T_\Omega^{Co}(x)$.
- Rockafellar [20] showed that if $T_\Omega^H(x)$ is nonempty, $T_\Omega^H(x) = \mathrm{int}(T_\Omega^{Cl}(x))$, and therefore, $T_\Omega^{Cl}(x) = \mathrm{cl}(T_\Omega^H(x))$.

In order to establish the results of this section, we apply a generalization of the Clarke [9] directional derivative, as presented in [17], in which function evaluations are restricted to points in the domain. Specifically, the Clarke generalized directional derivative of the locally Lipschitz function f at $x \in \Omega$ in the direction $v \in \mathbb{R}^n$ is defined by

$$f^\circ(x; v) := \limsup_{\substack{y \to x,\ y \in \Omega \\ t \downarrow 0,\ y + tv \in \Omega}} \frac{f(y + tv) - f(y)}{t}. \tag{3.2}$$

The fundamental result upon which the entire first order convergence analysis [4] of MADS relies is that if f is Lipschitz near the limit point $\hat{x}$ of a refining subsequence, then $f^\circ(\hat{x}; v) \geq 0$ for any refining direction v in the hypertangent cone $T_\Omega^H(\hat{x})$. The next definition, also from [4], provides some nonsmooth terminology for stationarity.

DEFINITION 3.6. *Let f be Lipschitz near $x^* \in \Omega$. Then x^* is said to be a Clarke (resp., contingent) stationary point of f over Ω if $f^\circ(x^*; v) \geq 0$ for every direction v in the Clarke tangent cone (resp., contingent cone) to Ω at x^*.*

In addition, x^* is said to be a Clarke (resp., contingent) KKT stationary point of f over Ω if $-\nabla f(x^*)$ exists and belongs to the polar of the Clarke tangent cone (resp., contingent cone) to Ω at x^*.

If $\Omega = \mathbb{R}^n$ or x^* lies in the interior of Ω, then a stationary point as described by Definition 3.6 meets the condition that $f^\circ(x^*; v) \geq 0$ for all $v \in \mathbb{R}^n$. This is equivalent

to $0 \in \partial f(x^*)$, the generalized gradient of f at x^* [9], which is defined by

$$\partial f(x) := \left\{ s \in \mathbb{R}^n : f^\circ(x; v) \geq v^T s \text{ for all } v \in \mathbb{R}^n \right\}.$$

The function f is said to be *strictly differentiable* at x if the generalized gradient of f at x is a singleton; i.e., $\partial f(x) = \{\nabla f(x)\}$.

The main results of [4] can now be summarized in the next theorem.

THEOREM 3.7. *Let $\hat{x} \in \Omega$ be the limit of a refining subsequence, and assume that $T_\Omega^H(\hat{x}) \neq \emptyset$ and the set of refining directions is dense in $T_\Omega^H(\hat{x})$.*

1. *If f is Lipschitz near $\hat{x}$, then $\hat{x}$ is a Clarke stationary point of f on Ω.*
2. *If f is strictly differentiable at $\hat{x}$, then $\hat{x}$ is a Clarke KKT stationary point of f on Ω.*

Furthermore, if Ω is regular at $\hat{x}$, then the following hold:

1. *If f is Lipschitz near $\hat{x}$, then $\hat{x}$ is a contingent stationary point of f on Ω.*
2. *If f is strictly differentiable at $\hat{x}$, then $\hat{x}$ is a contingent KKT stationary point of f on Ω.*

4. New second-order stationarity results. This section contains second-order convergence theory for MADS. In section 4.1 we recall the definition of the generalized Hessian and identify some useful properties. In section 4.2 we present second-order necessary and sufficient conditions for optimality for set-constrained optimization problems. Finally, in section 4.3, we establish conditions under which convergence of MADS iterates to a point satisfying second-order necessary and sufficient conditions is achieved.

4.1. Generalized second-order derivatives. Before proving convergence to second-order points, we present nonsmooth notions of second derivatives and introduce second-order optimality conditions. Generalized second-order directional derivatives are developed in [10] and [16], consistent with the Clarke [9] calculus for first-order derivatives. In this paper, we follow the Hiriart-Urruty, Strodiot, and Nguyen [16] definition of a generalized Hessian, given as follows.

DEFINITION 4.1. *Let $g : \mathbb{R}^n \to \mathbb{R}$ be $C^{1,1}$ near $x \in \Omega \subseteq \mathbb{R}^n$. The* generalized Hessian *of g at x, denoted by $\partial^2 g(x)$, is the set of matrices defined as the convex hull of the set*

$$\{A \in \mathbb{R}^{n \times n} : \exists\, x_k \to x \text{ with } g \text{ twice differentiable at } x_k \text{ and } \nabla^2 g(x_k) \to A\}.$$

By construction, $\partial^2 g(x)$ is a nonempty, compact, and convex set of symmetric matrices [16]. The function g is said to be *twice strictly differentiable* at x if the generalized Hessian is a singleton; i.e., $\partial^2 g(x) = \{\nabla^2 g(x)\}$. Furthermore, as a set-valued mapping, $x \rightrightarrows \partial^2 g(x)$ has two key properties, also identified in [16], which are necessary to establish optimality conditions in the next section.

- $\partial^2 g(x)$ is a *locally bounded* set-valued mapping:
  Given a matrix norm $\|\cdot\|$, there exist an $\varepsilon > 0$ and $\eta \in \mathbb{R}$ such that
  $$\sup\{\|A\| : A \in \partial^2 g(y), y \in B_\varepsilon(x)\} \ \leq \ \eta;$$
- $\partial^2 g(x)$ is a *closed* set-valued mapping:
  If $x_k \to x$ and $A_k \to A$ with $A_k \in \partial^2 g(x_k)$ for all k, then $A \in \partial^2 g(x)$.

The following second-order Taylor series result also comes from [16].

THEOREM 4.2. *Let $g : \mathbb{R}^n \to \mathbb{R}$ be $C^{1,1}$ in a open set $U \subset \mathbb{R}^n$, and let $[a, b] \subset U$ be a line segment. Then there exist an $x \in]a, b[$ and a matrix $A_x \in \partial^2 f(x)$ such that*

$$g(b) \ = \ g(a) + (b - a)^T \nabla f(a) + \frac{1}{2}(b - a)^T A_x (b - a).$$

In the next section, we apply this result to feasible points that may lie on the boundary of Ω. We are able to do this because our assumptions on the local smoothness of f are independent of Ω.

4.2. Second-order optimality conditions. Second-order necessary and sufficient optimality conditions for constrained problems are traditionally expressed in terms of the Lagrangian function. However, our use of the barrier approach in handling constraints provides no useful information about the constraint gradients, and thus prevents us from proving anything with respect to traditional optimality conditions. Therefore, instead of dealing with the Lagrangian function, we extend optimality conditions for set-constrained problems (see [8] for further discussions).

We now establish Clarke-based second-order necessary and sufficient conditions for set-constrained optimality. The proof for the former is very similar to one found in [16] for unconstrained problems, the only difference being the first-order condition satisfied by the local minimizer. It is expressed in terms of feasible directions, formally given in Definition 4.3.

DEFINITION 4.3. *The direction $v \in \mathbb{R}^n$ is said to be feasible for $\Omega \subset \mathbb{R}^n$ at $x \in \Omega$ if there exists an $\epsilon > 0$ for which $x + tv \in \Omega$ for all $0 \leq t < \epsilon$. The set of feasible directions for Ω at $x \in \Omega$ is a cone and is denoted by $T_\Omega^F(x)$.*

It follows immediately that $T_\Omega^H(x) \subseteq T_\Omega^F(x) \subseteq T_\Omega^{Co}(x)$ for any $x \in \Omega$. Moreover, if $T_\Omega^H(x) \neq \emptyset$ for some $x \in \Omega$, and if Ω is regular at x, then $\mathrm{cl}(T_\Omega^H(x)) = \mathrm{cl}(T_\Omega^F(x)) = T_\Omega^{Cl}(x) = T_\Omega^{Co}(x)$. However, without regularity it is possible that either of the following holds:

- $T_\Omega^{Cl}(x) \subset \mathrm{int}(T_\Omega^F(x))$: e.g., if $\Omega = \{(a,b) \in \mathbb{R}^2 : a \geq 0 \text{ or } b \geq 0\}$, then $T_\Omega^{Cl}(0,0) = \mathbb{R}^2_+$ and $T_\Omega^F(0,0) = \Omega$,
- $\mathrm{cl}(T_\Omega^F(x)) \subset \mathrm{int}(T_\Omega^{Cl}(x))$: e.g., if $\Omega = \mathbb{R}^2 \backslash \{(-\frac{1}{k}, b) \in \mathbb{R}^2 : b \in \mathbb{R}, k = 1, 2, \dots\}$, then $T_\Omega^F(0,0) = \{(a,b) \in \mathbb{R}^2 : a \geq 0\}$ and $T_\Omega^{Cl}(0,0) = \mathbb{R}^2$.

THEOREM 4.4 (second-order necessary condition for set-constrained optimality). *Let $x^* \in \Omega$ be a local solution of (1.1). If f is $C^{1,1}$ near x^*, then any feasible direction $v \in T_\Omega^F(x^*)$ for which $v^T \nabla f(x^*) = 0$ satisfies $v^T A v \geq 0$ for some $A \in \partial^2 f(x^*)$.*

Proof. Let $v \in \mathbb{R}^n$ be a feasible direction that satisfies $v^T \nabla f(x^*) = 0$, and consider the sequence $\{x_k\}$, where $x_k = x^* + \frac{1}{k} v$. It follows that $x_k \in \Omega$ when k is sufficiently large. Then by second-order Taylor series in a neighborhood of the local minimizer x^*, we have for each k sufficiently large

$$0 \;\leq\; f(x_k) - f(x^*) \;=\; \frac{1}{k} \nabla f(x^*)^T v + \frac{1}{2k^2} v^T A_k v \;=\; \frac{1}{2k^2} v^T A_k v, \tag{4.1}$$

where $A_k \in \partial^2 f(\bar{x}_k)$ for some $\bar{x}_k \in]x^*, x_k[$.

Since $\partial^2 f$ is locally bounded and $\bar{x}_k \to x^*$, the sequence $\{A_k\}$ is locally bounded and thus possesses an accumulation point A. Furthermore, since $\partial^2 f$ is a closed set-valued mapping, we have $A \in \partial^2 f(x^*)$. Taking limits in (4.1) leads to $v^T A v \geq 0$. ∎

Theorem 4.4 applies to the set of hypertangent vectors as well as feasible directions since the set of feasible directions contains the hypertangent cone. However, this necessary condition does not necessarily hold for directions in the Clarke tangent or contingent cone, as the following example shows.

Example 4.5. Consider the quadratic optimization problem, in which $f : \mathbb{R}^2 \to \mathbb{R}$ is defined by $f(a,b) = -(a^2 + b^2)$, and $\Omega = \{(a,b) \in \mathbb{R}^2 : a^2 + (b-1)^2 \leq 1\}$. The optimal solution is at $(0,2)$, where

$$T_\Omega^H(0,2) = T_\Omega^F(0,2) = \{(v_1, v_2) : v_2 < 0\},$$

$$T_\Omega^{Cl}(0,2) = T_\Omega^{Co}(0,2) = \{(v_1, v_2) : v_2 \leq 0\}.$$

The direction $v = (1,0)^T \in T_\Omega^{Cl}(0,2) = T_\Omega^{Co}(0,2)$ is not a feasible direction and makes a zero inner product with $\nabla f(0,2) = (0,-4)^T$, but the Hessian matrix is given by

$$\nabla^2 f(0,2) \;=\; \begin{bmatrix} -2 & 0 \\ 0 & -2 \end{bmatrix},$$

which yields $v^T \nabla^2 f(0,2) v = -2 < 0$.

THEOREM 4.6 (second-order sufficient condition for set-constrained optimality). *Let $x^* \in \Omega$ be a contingent stationary point for the optimization problem defined in* (1.1)*, and suppose that $T_\Omega^H(x^*) \neq \emptyset$ and that Ω is convex near x^*. If f is $C^{1,1}$ near x^*, and if $v^T A v > 0$ for all matrices $A \in \partial^2 f(x^*)$ and all nonzero tangent directions $v \in T_\Omega^{Co}(x)$ that satisfy $v^T \nabla f(x^*) = 0$, then x^* is a strict local solution of* (1.1).

Proof. The proof is by contraposition. Suppose that x^* is not a strict local minimizer. Then there exists a sequence $\{y_k\} \subset \Omega$ (with $y_k \neq x^*$) converging to x^* satisfying $f(y_k) \leq f(x^*)$ for all k. By taking subsequences if necessary, we can assume that the sequence $\{w_k\}$ with $w_k = \frac{y_k - x^*}{\|y_k - x^*\|}$ converges to some vector $v \in \mathbb{R}^n$.

Local convexity of Ω near x^* implies that v and w_k are contingent directions for all $k \geq \ell$, for some integer $\ell \geq 0$. Moreover, since x^* is assumed to be a contingent stationary point, and since f is continuously differentiable, then $v^T \nabla f(x^*) \geq 0$ and $w_k^T \nabla f(x^*) \geq 0$ for all $k \geq \ell$. However, since $f(y_k) \leq f(x^*)$ for all k, then $v^T \nabla f(x^*) = 0$.

Theorem 4.2 on Taylor series ensures that for each $k \geq \ell$, there exists some matrix $A_k \in \partial^2 f(\bar{x}_k)$ with $\bar{x}_k \in]x^*, y_k[$ such that

$$\begin{aligned} 0 \;\geq\; f(y_k) - f(x^*) &= (y_k - x^*)^T \nabla f(x^*) + \frac{1}{2}(y_k - x^*)^T A_k (y_k - x^*) \\ &\geq \frac{1}{2}(y_k - x^*)^T A_k (y_k - x^*). \end{aligned} \tag{4.2}$$

Now, since $\bar{x}_k \to x^*$, and since $\partial^2 f(x^*)$ is a closed locally bounded set-valued mapping, there exists an accumulation point $A \in \partial^2 f(x^*)$ of the sequence $\{A_k\}$. Dividing (4.2) by $\|y_k - x^*\|^2$ and taking limits leads to $0 \geq \frac{1}{2} v^T A v$, where $v \neq 0$ belongs to $T_\Omega^{Co}(x^*)$ and satisfies $v^T \nabla f(x^*) = 0$. ∎

The previous theorem requires as an assumption that the set Ω is locally convex. The following example shows that regularity of Ω is not sufficient to guarantee a local minimizer.

Example 4.7. Consider the quadratic optimization problem, in which $f : \mathbb{R}^2 \to \mathbb{R}$ is defined by $f(a,b) = a^2 + 2b$, and $\Omega = \{(a,b) \in \mathbb{R}^2 : 2a^2 + b \geq 0\}$. The solution $x^T = (0,0)$ is a contingent stationary point, and Ω is regular at x since

$$T_\Omega^{Cl}(x) = T_\Omega^{Co}(x) = \{(v_1, v_2) : v_2 \geq 0\}.$$

The vector $v = (1,0)^T \in T_\Omega^{Co}(x)$ satisfies $v^T \nabla f(x) = 0$ and $v^T \nabla^2 f(x) v = 2 > 0$. However, $(\epsilon, -\epsilon^2)$ belongs to the strict interior of Ω for all $\epsilon \neq 0$, and $f(\epsilon, -\epsilon^2) = -\epsilon^2 < 0 = f(x)$.

4.3. Second-order stationarity results for MADS. The next two results are the main contributions of this paper. The first theorem establishes convergence of

a subsequence of MADS iterates to a point satisfying the second-order necessary condition identified in Theorem 4.4, and the second establishes the sufficiency conditions of Theorem 4.6.

THEOREM 4.8. *Let f be $C^{1,1}$ near a limit $\hat{x}$ of a refining subsequence, and assume that $T_\Omega^H(\hat{x}) \neq \emptyset$ and that Ω is regular near $\hat{x}$. If the set of refining directions is dense in $T_\Omega^H(\hat{x})$, then $\hat{x}$ satisfies the second-order necessary condition for set-constrained optimality.*

Proof. Let $v \in \mathbb{R}^n$ be any nonzero feasible direction that satisfies $v^T \nabla f(\hat{x}) = 0$, and suppose, by way of contradiction, that $v^T \hat{A} v < 0$ for all matrices $\hat{A} \in \partial^2 f(\hat{x})$. Since $\partial^2 f(\hat{x})$ is nonempty and compact, and $\partial^2 f$ is a closed set-valued mapping, there exists some $\varepsilon > 0$ such that $v^T A v < 0$ for all $A \in \partial^2 f(x)$ and for all $x \in B_\varepsilon(\hat{x})$.

Let K denote the set of indices of unsuccessful iterations. Regularity of Ω, together with the assumption that $T_\Omega^H(\hat{x}) \neq \emptyset$, guarantees that $\mathrm{cl}(T_\Omega^H(\hat{x})) = T_\Omega^{Co}(\hat{x}) = \mathrm{cl}(T_\Omega^F(\hat{x}))$. Therefore, the denseness of the set of refining directions in $T_\Omega^H(\hat{x})$ ensures the existence of $\{w_k\}_{k \in K}$ converging to v with $w_k = \frac{d_k}{\|d_k\|}$, $d_k \in D_k$, for each $k \in K$. Applying Taylor series yields

$$f(x_k + \Delta_k^p d_k) - f(x_k) = \Delta_k^p d_k^T \nabla f(x_k) + \frac{1}{2}(\Delta_k^p)^2 d_k^T A_k^+ d_k, \tag{4.3}$$

$$f(x_k - \Delta_k^p d_k) - f(x_k) = -\Delta_k^p d_k^T \nabla f(x_k) + \frac{1}{2}(\Delta_k^p)^2 d_k^T A_k^- d_k, \tag{4.4}$$

where $A_k^+ \in \partial^2 f(x^+)$ for some $x^+ \in]x_k, x_k + \Delta_k^p d_k[$ and $A_k^- \in \partial^2 f(x^-)$ for some $x^- \in]x_k, x_k - \Delta_k^p d_k[$. Since $\Delta_k \to 0^+$ and $x_k \to \hat{x}$, there is a subsequence for which A_k^+ converges to some $A^+ \in \partial^2 f(\hat{x})$, and A_k^- converges to some $A^- \in \partial^2 f(\hat{x})$. Moreover, since $\partial^2 f(\hat{x})$ is a convex set, $A = \frac{1}{2}(A^+ + A^-) \in \partial^2 f(\hat{x})$.

Adding (4.3) and (4.4) and substituting $d_k = \|d_k\| w_k$ yields

$$\frac{1}{\Delta_k^p \|d_k\|} \left[\frac{f(x_k + \Delta_k^p \|d_k\| w_k) - f(x_k)}{\Delta_k^p \|d_k\|} + \frac{f(x_k - \Delta_k^p \|d_k\| w_k) - f(x_k)}{\Delta_k^p \|d_k\|} \right] = w_k^T A_k w_k, \tag{4.5}$$

where $A_k = \frac{1}{2}(A_k^+ + A_k^-)$. Furthermore, since $w_k \to v$ and $v^T A v < 0$, there exists $\gamma < 0$ such that $w_k^T A_k w_k \leq \gamma < 0$ for all sufficiently large $k \in K$, which forces the left-hand side of (4.5) to also be negative and bounded away from zero. But since $d_k \in D_k$ for all sufficiently large $k \in K$, we have that $f(x_k) \leq f(x_k + \Delta_k^p d_k)$, which makes nonnegative the first term of the left-hand side of (4.5) (for all sufficiently large $k \in K$). Thus it must be the case that

$$\frac{f(x_k - \Delta_k^p \|d_k\| w_k) - f(x_k)}{\Delta_k^p \|d_k\|} \leq \gamma < 0 \tag{4.6}$$

for all sufficiently large $k \in K$. Taking the limit of (4.6) as $k \to \infty$ in K yields $\nabla f(\hat{x})^T(-v) < 0$, or $\nabla f(\hat{x})^T v > 0$, which contradicts the assumption that $\nabla f(\hat{x})^T v = 0$. ∎

The following result shows that the sufficient conditions of Theorem 4.6 can be satisfied by a subsequence of MADS iterates, given stronger hypotheses than those of Theorem 4.8.

THEOREM 4.9. *Let f be twice strictly differentiable at a limit $\hat{x}$ of a refining subsequence, and assume that $T_\Omega^H(\hat{x}) \neq \emptyset$, Ω is convex near $\hat{x}$, and $\nabla^2 f(\hat{x})$ is nonsingular. If the set of refining directions is dense in $T_\Omega^H(\hat{x})$, then $\hat{x}$ is a strict local minimizer of f on Ω.*

Proof. Since f is twice strictly differentiable at $\hat{x}$, $\partial^2 f(\hat{x}) = \{\nabla^2 f(\hat{x})\}$. Thus, it follows from Theorem 4.8 that $v^T \nabla^2 f(\hat{x}) v \geq 0$ for all feasible directions $v \in T_\Omega^F(\hat{x})$ satisfying $\nabla f(\hat{x})^T v = 0$. But since $\nabla^2 f(\hat{x})$ is assumed to be nonsingular, this inequality is strict. Furthermore, by Theorem 3.7 and the smoothness of f near $\hat{x}$, $\hat{x}$ is a first-order contingent stationary point. Thus the hypotheses of Theorem 4.6 are satisfied, and the result is proved. ∎

Clearly, these are strong results for a direct search method. However, in practice, achieving denseness of the refining directions in the hypertangent cone (a key assumption) requires increasingly more poll directions per iteration. To overcome this problem, an implementable instance of MADS is introduced in [4], called LTMADS, in which the positive spanning directions used at each iteration are limited in number but are chosen randomly from among the increasing number of possible poll directions. While this is not difficult to implement, the drawback is that denseness of the refining directions is only achieved *almost surely* (i.e., with probability one). Thus, in practice, the convergence results proved both here and in [4] are only attained *almost surely*. This is a weaker measure of convergence, but it works well in practice [4]. We apply LTMADS to one of the numerical examples in the next section.

5. Examples. Second order results for GPS are presented in [1]. They are not as strong as those presented here for MADS. In this section, we illustrate this difference through two quadratic examples in $\mathbb{R}^2$. The first shows how GPS, but not MADS, can converge in an infinite number of iterations to a saddle point with wide cones of descent. This result is actually proved, but doing so requires an uncommon set of parameter choices. The second example [1] uses more realistic parameter choices, and numerical tests show that GPS stalls at a saddle point with narrow cones of descent, but MADS successfully avoids it.

5.1. An example where GPS converges in an infinite number of iterations to a saddle point. Consider the unconstrained quadratic optimization problem in which the polynomial objective function in $\mathbb{R}^2$ is $f(a, b) = a^2 + 3ab + b^2$. The point (0,0) is a saddle point, at which the descent directions lie in the cone generated by $a = \frac{1}{2}b(-3 \pm \sqrt{5})$.

We apply an instance of GPS where $D_k = D = \{e_1, e_2, -e_1, -e_2\}$ is constant throughout all iterations. On iterations that fail to improve the incumbent, the mesh size parameter is divided by 16. On successful iterations that follow an unsuccessful one, the mesh size is kept constant, and on other successful iterations, the mesh size parameter is multiplied by 8. Thus, the GPS parameters are $G = I$ (the identity matrix), $Z = D = [I;\ -I]$, $\tau = 2$, $w^- = -4$, and $w^+ = 3$.

Furthermore, we use an empty SEARCH and an opportunistic POLL, i.e., an iteration terminates as soon as an improved mesh point is generated. Moreover, when the iteration number k modulo 3 is 1, the POLL step first evaluates $x_k - \Delta_k e_2$, and otherwise, the POLL step first evaluates $x_k - \Delta_k e_1$. The order in which the other poll points are explored is irrelevant to this example.

The initial parameters are $x_0^T = (1, 1)$ with $f(x_0) = 5$ and $\Delta_0 = 8$. Figure 5.1 displays the first iterates generated by the algorithm. The figure also displays some level sets of f.

We next show that the entire sequence of iterates converges to the origin. This happens because this instance of GPS never generates any trial points in the cone where f is negative. It either jumps over the cone, which results in an unsuccessful iteration, or takes a small step which falls short of reaching the cone. For example,

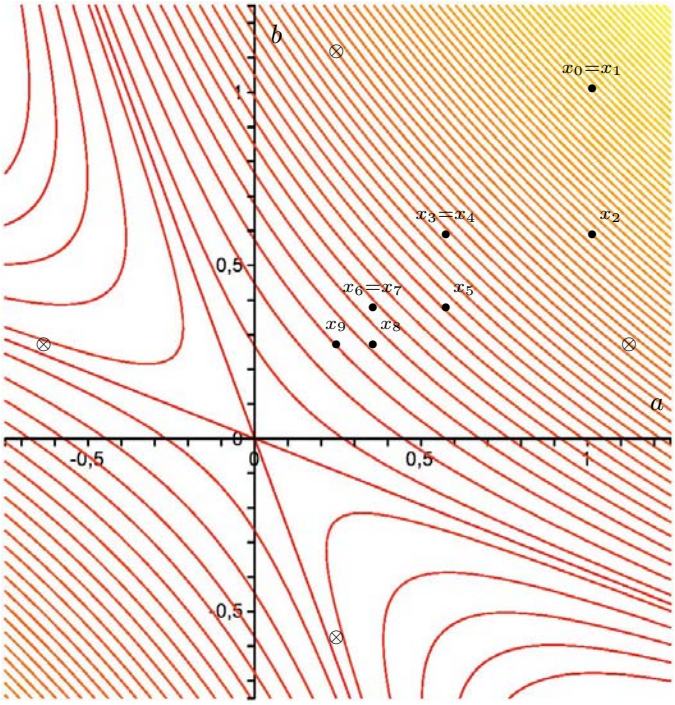


FIG. 5.1. *Initial GPS iterates.*

TABLE 5.1
Iterations $k = 3\ell$ to $k = 3(\ell+1)$.

k	x_k	Δ_k	Trial poll points at $x_k = (a_k, b_k)$
3ℓ	$(2^{-\ell}, 2^{-\ell})$	$2^{3-\ell}$	$f(a_k+\Delta_k, b_k) = f(9\times 2^{-\ell}, 2^{-\ell}) = 109\times 4^{-\ell}$
			$= f(2^{-\ell}, 9\times 2^{-\ell}) = f(a_k, b_k+\Delta_k)$
			$f(a_k-\Delta_k, b_k) = f(-7\times 2^{-\ell}, 2^{-\ell}) = 29\times 4^{-\ell}$
			$= f(2^{-\ell}, -7\times 2^{-\ell}) = f(a_k, b_k+\Delta_k)$
$3\ell+1$	$(2^{-\ell}, 2^{-\ell})$	$2^{-\ell-1}$	$f(a_k, b_k-\Delta_k) = f(2^{-\ell}, 2^{-\ell-1}) = 11\times 4^{-\ell-1}$
$3\ell+2$	$(2^{-\ell}, 2^{-\ell-1})$	$2^{-\ell-1}$	$f(a_k-\Delta_k, b_k) = f(2^{-\ell-1}, 2^{-\ell-1}) = 5\times 4^{-\ell-1}$
$3(\ell+1)$	$(2^{-\ell-1}, 2^{-\ell-1})$	$2^{2-\ell}$	...

at iteration $k = 9$, the trial poll points are $(\frac{9}{8}, 1), (1, \frac{9}{8}), (-\frac{7}{8}, 1)$, and $(1, -\frac{7}{8})$. These four trial points are represented by the symbol $\otimes$ in the figure.

PROPOSITION 5.1. *For any integer $\ell \geq 0$, the GPS iterates are such that $x_{3\ell} = x_{3\ell+1} = \left(2^{-\ell}, 2^{-\ell}\right)$, $x_{3\ell+2} = \left(2^{-\ell}, 2^{-\ell-1}\right)$, and $\Delta_{3\ell} = 2^{3-\ell}$, $\Delta_{3\ell+1} = \Delta_{3\ell+2} = 2^{-\ell-1}$.*

Proof. The proof is done by induction. The result is true for the initial iteration $k = 0$. Suppose that iteration $k = 3\ell$ is initiated with $\Delta_k = 2^{3-\ell}$ and $x_k = \left(2^{-\ell}, 2^{-\ell}\right)$. The current objective function value is $f(x_k) = 5\times 4^{-\ell}$. Table 5.1 details the objective function values at the poll points for iterations $k = 3\ell, 3\ell+1$, and $3\ell+2$. Trial points that improve the incumbent appear in shaded boxes. This table shows that the iterate for $k = 3(\ell+1)$ is $(2^{-\ell-1}, 2^{-\ell-1})$ and that the corresponding mesh size parameter is $2^{2-\ell}$. This concludes the proof. ∎

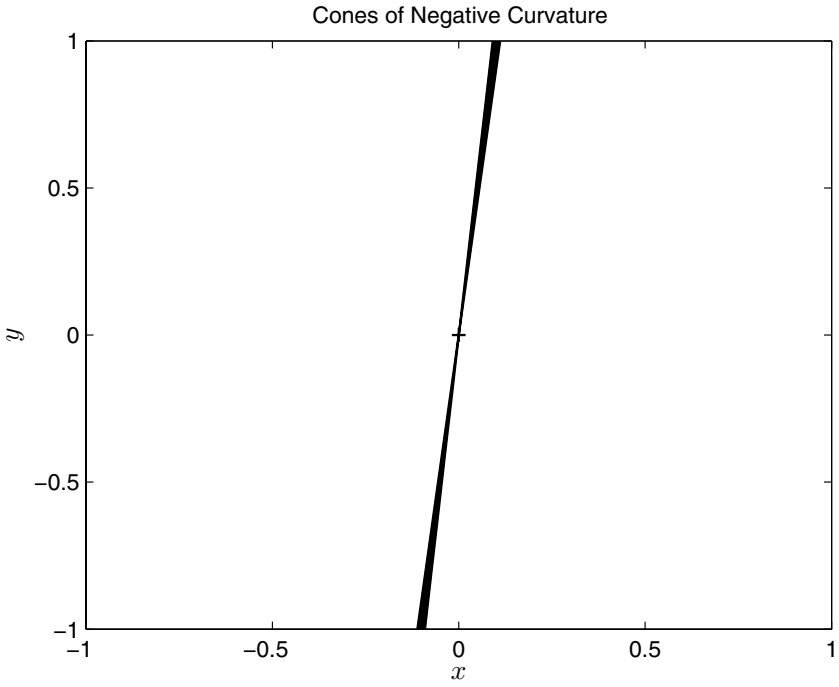


FIG. 5.2. *For $f(a,b) = (9a - b)(11a - b)$, the cones of descent at the saddle point (0,0) are shown in the shaded area between the lines $b = 9a$ and $b = 11a$.*

The previous proposition shows that the entire sequence of iterates generated by GPS converges to the saddle point $(0, 0)$, which is not a local minimizer. Theorem 4.8 ensures that any MADS instance with an asymptotically dense set of refining directions will not converge to that saddle point, since the necessary optimality condition is not satisfied: $v^T = (-1, 1)$ is a feasible direction for which $v^T \nabla f(0, 0) = 0$, but

$$v^T \nabla^2 f(0,0) v \;=\; v^T \begin{bmatrix} 2 & 3 \\ 3 & 2 \end{bmatrix} v^T \;=\; -2$$

is negative.

5.2. An example where GPS reaches and stalls at a saddle point. Consider the bound constrained problem

$$\min_{-2 \le a,b \le 2} \; f(a,b) \;=\; 99a^2 - 20ab + b^2 \;=\; (9a - b)(11a - b).$$

At the saddle point $(0, 0)$, directions of descent lie only in the narrow cone formed by the lines $b = 9a$ and $b = 11a$. Thus to avoid stalling at the saddle point, GPS or MADS would have to generate a feasible iterate that lies inside this cone (see Figure 5.2). In this example, the SEARCH step is empty and the initial point is chosen to be $(1.01,\ 0.93)$. This starting point is chosen to be nonintegral to make it more difficult for GPS to reach the integral point $(0, 0)$. Both GPS and MADS were run using the NOMAD software package [13] with primarily default settings: $G = I$, $Z = D = [I;\ -I]$, $\tau = 2$, $w^- = -1$, $w^+ = 0$, and standard $2n$ coordinate directions as poll directions.

GPS reaches the saddle point at the 358th function evaluation with a poll size parameter of 10^{-17}. This implies that, regardless of the termination tolerance chosen, it stalls there because none of the poll directions are directions of descent. On the other hand, NOMAD's implementation of LTMADS successfully moved off of the saddle point to reach a local minimizer in 100 of 100 runs. This is again consistent with Theorem 4.8, since $v^T = (1, 10)$ is a feasible direction for which $v^T \nabla f(0,0) = 0$ but

$$v^T \nabla^2 f(0,0) v \;=\; v^T \begin{bmatrix} 198 & -20 \\ -20 & 2 \end{bmatrix} v^T \;=\; -2$$

is negative.

6. Concluding remarks. The theoretical results presented here establish strong convergence results for MADS. In spite of MADS being a derivative-free method, we have shown convergence of a subsequence of MADS iterates to a second-order stationary point under conditions weaker than standard Newton assumptions, namely, that f is continuously differentiable with Lipschitz derivatives near the limit point. Moreover, if Ω is locally convex and f is twice strictly differentiable near the limit point, then the limit point is a local minimizer for (1.1).

In section 5, we provided examples to illustrate the superior convergence properties of MADS over GPS. However, since our implementation involves random selection of positive spanning directions, the convergence properties established in section 4.3 are achieved, in practice, with probability one. We envision a future area of research being the clever enumeration of these directions so that the stronger type of convergence is retained by an implementable instance of the algorithm. Specifically, we would like to deterministically generate an asymptotically dense set of directions in such a way that, after any finite number of iterations, the directions used by the algorithm are uniformly spaced (or as close to it as possible).

Acknowledgments. The authors wish to thank John Dennis for his support, useful discussions, and suggestions for improving the paper, along with two anonymous referees for their timely and helpful suggestions that have improved the presentation.

The views expressed in this paper are those of the authors and do not reflect the official policy or position of the U.S. Air Force, Department of Defense, or U.S. Government.

REFERENCES

[1] M. A. Abramson, *Second-order behavior of pattern search*, SIAM J. Optim., 16 (2005), pp. 515–530.

[2] M. A. Abramson, C. Audet, and J. E. Dennis, Jr., *Generalized pattern searches with derivative information*, Math. Program., 100 (2004), pp. 3–25.

[3] C. Audet and J. E. Dennis, Jr., *Analysis of generalized pattern searches*, SIAM J. Optim., 13 (2003), pp. 889–903.

[4] C. Audet and J. E. Dennis, Jr., *Mesh adaptive direct search algorithms for constrained optimization*, SIAM J. Optim., 17 (2006), pp. 188–217.

[5] C. Audet and J. E. Dennis, Jr., *A pattern search filter method for nonlinear programming without derivatives*, SIAM J. Optim., 14 (2004), pp. 980–1010.

[6] C. Audet and D. Orban, *Finding optimal algorithmic parameters using derivative-free optimization*, SIAM J. Optim, to appear.

[7] 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.

[8] K. P. E. Chong and S. H. Zak, *Linear and Nonlinear Programming*, John Wiley and Sons, New York, 2001.
[9] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Classics Appl. Math. 5, SIAM, Philadelphia, 1990.
[10] R. Cominetti and R. Correa, *A generalized second-order derivative in nonsmooth optimization*, SIAM J. Control Optim., 28 (1990), pp. 789–809.
[11] I. D. Coope and C. J. Price, *Frame-based methods for unconstrained optimization*, J. Optim. Theory Appl., 107 (2000), pp. 261–274.
[12] 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.
[13] G. Couture, C. Audet, J. E. Dennis, Jr., and M. A. Abramson, *The NOMAD project*, http://www.gerad.ca/NOMAD/.
[14] C. Davis, *Theory of positive linear dependence*, Amer. J. Math., 76 (1954), pp. 733–746.
[15] D. E. Finkel and C. T. Kelley, *Convergence Analysis of the DIRECT Algorithm*, Technical report CRSC-TR04-28, Department of Mathematics, North Carolina State University, Raleigh, NC, 2004.
[16] J.-B. Hiriart-Urruty, J.-J. Strodiot, and V. H. Nguyen, *Generalized Hessian matrix and second-order optimality conditions for problems with $C^{1,1}$ data*, Appl. Math. Optim., 11 (1984), pp. 43–56.
[17] J. Jahn, *Introduction to the Theory of Nonlinear Optimization*, Springer, Berlin, 1994.
[18] 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.
[19] M. D. McKay, W. J. Conover, and R. J. Beckman, *A comparison of three methods for selecting values of input variables in the analysis of output from a computer code*, Technometrics, 21 (1979), pp. 239–245.
[20] R. T. Rockafellar, *Generalized directional derivatives and subgradients of nonconvex functions*, Canad. J. Math., 32 (1980), pp. 257–280.
[21] V. Torczon, *On the convergence of pattern search algorithms*, SIAM J. Optim., 7 (1997), pp. 1–25.

SIAM J. OPTIM.
Vol. 17, No. 3, pp. 621–641
© 2006 Society for Industrial and Applied Mathematics

ON THE MINIMUM VOLUME COVERING ELLIPSOID OF ELLIPSOIDS*

E. ALPER YILDIRIM†

Abstract. Let $\mathcal{S}$ denote the convex hull of m full-dimensional ellipsoids in $\mathbb{R}^n$. Given $\epsilon > 0$ and $\delta > 0$, we study the problems of computing a $(1+\epsilon)$-approximation to the minimum volume covering ellipsoid of $\mathcal{S}$ and a $(1+\delta)n$-rounding of $\mathcal{S}$. We extend the first-order algorithm of Kumar and Yıldırım [*J. Optim. Theory Appl.*, 126 (2005), pp. 1–21] that computes an approximation to the minimum volume covering ellipsoid of a finite set of points in $\mathbb{R}^n$, which, in turn, is a modification of Khachiyan's algorithm [L. G. Khachiyan, *Math. Oper. Res.*, 21 (1996), pp. 307–320]. Our algorithm can also compute a $(1+\delta)n$-rounding of $\mathcal{S}$. For fixed $\epsilon > 0$ and $\delta > 0$, we establish polynomial-time complexity results for the respective problems, each of which is linear in the number of ellipsoids m. In particular, our algorithm can approximate the minimum volume covering ellipsoid of $\mathcal{S}$ in asymptotically the same number of iterations as that required by the algorithm of Kumar and Yıldırım to approximate the minimum volume covering ellipsoid of a set of m points. The main ingredient in our analysis is the extension of polynomial-time complexity of certain subroutines in the algorithm from a set of points to a set of ellipsoids. As a byproduct, our algorithm returns a finite "core" set $\mathcal{X} \subseteq \mathcal{S}$ with the property that the minimum volume covering ellipsoid of $\mathcal{X}$ provides a good approximation to the minimum volume covering ellipsoid of $\mathcal{S}$. Furthermore, the size of the core set depends only on the dimension n and the approximation parameter ϵ, but not on the number of ellipsoids m. We also discuss the extent to which our algorithm can be used to compute an approximate minimum volume covering ellipsoid and an approximate n-rounding of the convex hull of other sets in $\mathbb{R}^n$. We adopt the real number model of computation in our analysis.

Key words. minimum volume covering ellipsoids, Löwner ellipsoids, core sets, rounding of convex sets, approximation algorithms

AMS subject classifications. 90C25, 65K05, 90C22

DOI. 10.1137/050622560

1. Introduction.

Given m full-dimensional ellipsoids $\mathcal{E}_1, \ldots, \mathcal{E}_m$ in $\mathbb{R}^n$, let $\mathcal{S}$ denote their convex hull. In this paper, we are concerned with the problems of approximating the minimum volume covering ellipsoid (MVCE) of $\mathcal{S}$, denoted by MVCE($\mathcal{S}$), also known as the Löwner ellipsoid of $\mathcal{S}$, and computing an approximate n-rounding of $\mathcal{S}$.

Ellipsoidal approximations of a compact convex set $\mathcal{S} \subset \mathbb{R}^n$ with a nonempty interior play an important role in several applications. By the Löwner–John theorem (see Theorem 2.1), MVCE($\mathcal{S}$) provides a good rounding of the set $\mathcal{S}$, which implies that certain characteristics of $\mathcal{S}$ can be approximated using an ellipsoidal rounding as long as MVCE($\mathcal{S}$) can be computed efficiently. For instance, an ellipsoidal rounding of $\mathcal{S}$ can be used to efficiently compute lower and upper bounds for a quadratic optimization problem over $\mathcal{S}$ (see Proposition 2.6).

The idea of approximating complicated objects using simpler ones is widely used in computational geometry and computer graphics. A common approach is to replace a complicated but more realistic model of a complex object with a simpler model of a less complex object covering the original object such as a minimum volume box

*Received by the editors January 12, 2005; accepted for publication (in revised form) March 17, 2006; published electronically September 15, 2006.
http://www.siam.org/journals/siopt/17-3/62256.html
†Department of Industrial Engineering, Bilkent University, 06800 Bilkent, Ankara, Turkey (yildirim@bilkent.edu.tr). This research was supported in part by NSF through CAREER grant DMI-0237415.

or a sphere. More recently, ellipsoidal models have been proposed in the literature as they usually provide better approximations than bounding boxes or spheres (see, e.g., [25, 26, 14, 10]). The key idea is to construct a so-called bounding volume hierarchy [11], which is simply a tree of bounding volumes. The bounding volume at a given node encloses the bounding volumes of its children. The bounding volume of a leaf encloses a primitive. Such a data structure can be used for detection collision or ray tracing. For instance, if a ray misses the bounding volume of a particular node, then the ray will miss all of its children, and the children can be skipped. The ray tracing algorithm traverses this hierarchy, usually in depth-first order, and determines if the ray intersects an object. Therefore, if an ellipsoidal approximation is used, the construction of a bounding volume hierarchy requires the computation of the MVCE of a union of ellipsoids at every node.

There is an extensive body of research on MVCEs of a finite set of points. We refer the reader to [15, 29, 18] and the references therein for a detailed account of numerous applications and several algorithms. In contrast, we are not aware of any specialized algorithms for the MVCE of ellipsoids in the literature. It is known that the problem can be formulated as an instance of a convex determinant optimization problem with linear matrix inequalities [5, 2, 6], which is amenable to theoretically efficient algorithms proposed in [32, 31]. Our main objective in this paper is to establish that the problem of MVCE of ellipsoids admits a sufficiently rich structure that enables us to extend the first-order algorithm of Kumar and Yıldırım [18], which, in turn, is a modification of Khachiyan's algorithm [15], that computes an approximate MVCE of a finite set of points in an almost verbatim fashion to a set of ellipsoids. The main ingredient in our analysis is the extension of polynomial-time complexity of certain subroutines in the algorithm of [18] from a set of points to a set of ellipsoids. We mainly rely on the complexity results of Porkolab and Khachiyan [21] on semidefinite optimization with a fixed number of constraints, which leads to the polynomial-time complexity of quadratic optimization over an ellipsoid—one of the subroutines in our algorithm (see Proposition 2.6). Throughout this paper, we adopt the real number model of computation [4]; i.e., arithmetic operations with real numbers and comparisons can be done with unit cost.

Given $\epsilon > 0$ and a compact convex set $\mathcal{S} \subset \mathbb{R}^n$, an ellipsoid $\mathcal{E}$ is said to be a $(1+\epsilon)$-approximation to MVCE($\mathcal{S}$) if

$$\mathcal{E} \supseteq \mathcal{S}, \quad \text{vol } \mathcal{E} \leq (1+\epsilon) \text{ vol MVCE}(\mathcal{S}), \tag{1}$$

where vol $\mathcal{E}$ denotes the volume of $\mathcal{E}$. Given $\delta > 0$ and a compact convex set $\mathcal{S} \subset \mathbb{R}^n$, an ellipsoid $\tilde{\mathcal{E}}$ is said to be a $(1+\delta)n$-rounding of $\mathcal{S}$ if

$$\frac{1}{(1+\delta)n}\,\tilde{\mathcal{E}} \subseteq \mathcal{S} \subseteq \tilde{\mathcal{E}}, \tag{2}$$

where the ellipsoid on the left-hand side of (2) is obtained by scaling $\tilde{\mathcal{E}}$ around its center by a factor of $1/((1+\delta)n)$. If $\mathcal{S}$ is centrally symmetric (i.e., $\mathcal{S} = -\mathcal{S}$), then we replace the factor on the left-hand side by $1/\sqrt{(1+\delta)n}$. In this paper, we extend the first-order algorithm of [18] to compute a $(1+\epsilon)$-approximation to the MVCE of ellipsoids for $\epsilon > 0$. In particular, we establish that our extension has precisely the same iteration complexity as that of the algorithm of [18] (see Theorem 4.7). Furthermore, the overall complexity result is given by $O(mn^{O(1)}(\log n + [(1+\epsilon)^{2/n} - 1]^{-1}))$, which depends only linearly on the number of ellipsoids m (see Theorem 4.8). In addition, our algorithm can also compute a $(1+\delta)n$-rounding of the convex hull of a finite

number of ellipsoids for $\delta > 0$ in $O(mn^{O(1)}(\log n + \delta^{-1}))$ arithmetic operations (see Corollary 5.1). In both complexity results, $O(1)$ denotes a universal constant greater than four that does not depend on the particular instance. Therefore, our algorithm has polynomial-time complexity for fixed $\epsilon > 0$ and for fixed $\delta > 0$ and is especially well-suited for instances with $m \gg n$ and moderately small values of ϵ or δ.

As a byproduct, our algorithm computes a finite set $\mathcal{X} \subset \cup_{i=1,\ldots,m}\mathcal{E}_i$ with the property that the convex hull of $\mathcal{X}$, denoted by $\mathrm{conv}(\mathcal{X})$, provides a good approximation of $\mathcal{S} = \mathrm{conv}\,(\cup_{i=1,\ldots,m}\mathcal{E}_i)$. Moreover, the size of $\mathcal{X}$ depends only on the dimension n and the parameter ϵ but is independent of the number of ellipsoids m. In particular, $\mathcal{X}$ satisfies

$$\mathrm{vol\ MVCE}(\mathcal{X}) \leq \mathrm{vol\ MVCE}(\mathcal{S}) \leq \mathrm{vol}\ \mathcal{E} \leq (1+\epsilon)\mathrm{vol\ MVCE}(\mathcal{X}) \leq (1+\epsilon)\mathrm{vol\ MVCE}(\mathcal{S}),$$

where $\mathcal{E}$ denotes the $(1+\epsilon)$-approximation to the MVCE of $\mathcal{S}$ computed by our algorithm, which implies that $\mathcal{E}$ is simultaneously a $(1+\epsilon)$-approximation to $\mathrm{MVCE}(\mathcal{X})$ and to $\mathrm{MVCE}(\mathcal{S})$ (see Proposition 4.9).

Following the literature, we refer to $\mathcal{X}$ as an "ϵ-core set" (or a "core set") [8, 7, 17, 18] since $\mathrm{conv}(\mathcal{X})$ provides a compact approximation to the input set $\mathcal{S}$. Recently, core sets have received significant attention, and small core set results have been established for several geometric optimization problems such as the minimum enclosing ball problem and related clustering problems [17, 8, 7, 9, 1, 18]. Small core set results form a basis for developing practical algorithms for large-scale problems since many geometric optimization problems can be solved efficiently for small input sets.

The paper is organized as follows. We define our notation in the remainder of this section. In section 2, we present some preliminary results and discuss the complexity of semidefinite feasibility and optimization. We then establish that the ellipsoid containment problem can be cast as a linear matrix inequality and can therefore be checked in polynomial time. Section 3 is devoted to a deterministic volume approximation algorithm that will serve as an initialization stage for our algorithm. In section 4, we present and analyze a first-order algorithm for the MVCE problem. Section 5 establishes that our algorithm can also be used to compute an approximate n-rounding. We discuss how to extend our algorithm to other input sets in section 6. Section 7 concludes the paper with future research directions.

1.1. Notation. Vectors will be denoted by lowercase roman letters. For a vector u, u_i denotes its ith component. Inequalities on vectors will apply to each component. e will be reserved for the vector of ones in the appropriate dimension, which will be clear from the context. e^j is the jth unit vector. Uppercase roman letters will be reserved for matrices. $\mathcal{S}^n$ denotes the space of $n \times n$ real symmetric matrices. The inner product in $\mathcal{S}^n$ is given by $U \bullet V := \mathrm{trace}(UV) = \sum_{i,j} U_{ij}V_{ij}$ for any $U, V \in \mathcal{S}^n$. Note that $u^T Au = A \bullet uu^T$ for any $A \in \mathcal{S}^n$ and $u \in \mathbb{R}^n$. For $A \in \mathcal{S}^n$, $A \succ 0$ ($A \succeq 0$) indicates that A is positive definite (semidefinite) (i.e., the eigenvalues of A are strictly positive (nonnegative)). $\det(A)$ and $\mathrm{rank}(A)$ denote the determinant and the rank of a square matrix A, respectively. The identity matrix will be denoted by I. For a finite set of vectors $\mathcal{V}$, $\mathrm{span}(\mathcal{V})$ denotes the linear subspace spanned by the vectors in $\mathcal{V}$. The convex hull of a set $\mathcal{T} \in \mathbb{R}^n$ is referred to as $\mathrm{conv}(\mathcal{T})$. For a function $f : \mathbb{R}^n \to \mathbb{R}$, we use $x^* = \arg\max\ f(x)$ and $x_* = \arg\min\ f(x)$ to denote a global maximizer and a global minimizer of f, respectively. Superscripts will be used to refer to members of a sequence of vectors or matrices. Lowercase Greek letters will represent scalars. i, j, and k will be reserved for indexing purposes, and m and n will refer to the problem

data. Uppercase calligraphic letters will be used for all other objects such as sets, operators, and ellipsoids.

2. Preliminaries. A full-dimensional ellipsoid $\mathcal{E}$ in $\mathbb{R}^n$ admits a representation that is specified by an $n \times n$ symmetric positive definite matrix Q and a center $c \in \mathbb{R}^n$ and is defined as

$$\mathcal{E} = \{x \in \mathbb{R}^n : (x-c)^T Q(x-c) \leq 1\}. \tag{3}$$

The matrix Q determines the shape and the orientation of $\mathcal{E}$. In particular, the axes of $\mathcal{E}$ are the eigenvectors $d^1, \ldots, d^n \in \mathbb{R}^n$ of Q, and the length of each axis is given by $1/\sqrt{\lambda_1}, \ldots, 1/\sqrt{\lambda_n}$, where $\lambda_1, \ldots, \lambda_n$ are the corresponding eigenvalues of Q. Therefore, the volume of $\mathcal{E}$, denoted by vol $\mathcal{E}$, is given by

$$\text{vol } \mathcal{E} = \eta \det Q^{-\frac{1}{2}} = \eta \left(1/\sqrt{\Pi_{i=1}^n \lambda_i}\right), \tag{4}$$

where η is the volume of the unit ball in $\mathbb{R}^n$ [12]. Note that an ellipsoid $\mathcal{E}$ induces a norm on $\mathbb{R}^n$ via $\|x\|_{\mathcal{E}} := (x^T Q x)^{1/2}$. Therefore, every ellipsoid can be viewed as a translation of the unit ball in terms of the ellipsoidal norm induced by it.

Throughout this paper, we will assume that each of the input ellipsoids $\mathcal{E}_1, \ldots, \mathcal{E}_m \subset \mathbb{R}^n$ is full-dimensional. Note that this assumption is without loss of generality since any lower-dimensional ellipsoid can easily be approximated by a "thin" full-dimensional one. We remark that this assumption is merely for technical convenience, which allows us to have a uniform representation of each of the ellipsoids in the form given by (3). In addition, this assumption guarantees that $\text{conv}(\cup_{i=1}^m \mathcal{E}_i)$ is full-dimensional and leads to a simpler characterization of the ellipsoid containment problem (see Proposition 2.7). In particular, the full-dimensionality assumption on each of the ellipsoids can be relaxed by the weaker assumption that $\text{conv}(\cup_{i=1}^m \mathcal{E}_i)$ is full-dimensional and our analysis would still carry over to this slightly more general setting (see the discussion after Proposition 2.6). We refer the reader to [2] for further discussions on extremal ellipsoids.

We start with a classical result on the quality of the approximation of MVCE($\mathcal{S}$) of a convex set $\mathcal{S} \subset \mathbb{R}^n$.

THEOREM 2.1 (Löwner–John [13]). *Let $\mathcal{S} \subset \mathbb{R}^n$ be a compact, convex set with a nonempty interior. Then, MVCE($\mathcal{S}$) exists and is unique and satisfies*

$$\frac{1}{n} MVCE(\mathcal{S}) \subseteq \mathcal{S} \subseteq MVCE(\mathcal{S}), \tag{5}$$

where the ellipsoid on the left-hand side is obtained by scaling MVCE($\mathcal{S}$) around its center by a factor of $1/n$. Furthermore, if $\mathcal{S}$ is symmetric around the origin, then the factor on the left-hand side of (5) can be improved to $1/\sqrt{n}$.

We next state a well-known lemma that will be useful for our analysis.

LEMMA 2.2 (Schur complement). *Let*

$$A = \begin{bmatrix} B & C \\ C^T & D \end{bmatrix}$$

be a symmetric matrix with $B \in \mathcal{S}^\alpha$ and $D \in \mathcal{S}^\beta$. Assume that $D \succ 0$. Then, $A \succeq 0$ if and only if $B - CD^{-1}C^T \succeq 0$.

2.1. Complexity of semidefinite feasibility and optimization. Consider the following feasibility problems:

1. **(PF)** Given $A_1, A_2, \ldots, A_\kappa \in \mathcal{S}^n$ and $\beta_1, \ldots, \beta_\kappa \in \mathbb{R}$, determine whether there exists a matrix $X \in \mathcal{S}^n$ such that
$$A_i \bullet X \leq \beta_i, \quad i = 1, \ldots, \kappa, \quad X \succeq 0.$$
2. **(DF)** Given $B_0, B_1, \ldots, B_\kappa \in \mathcal{S}^n$, determine whether there exist real numbers $y_1, \ldots, y_\kappa$ such that
$$B_0 + y_1 B_1 + y_2 B_2 + \cdots + y_\kappa B_\kappa \succeq 0.$$

The complexity of the problems **(PF)** and **(DF)** is still a fundamental open problem. In the real number model of computation, both problems are in NP since one can check in polynomial time whether a given symmetric matrix is positive semidefinite using Cholesky factorization. Ramana [22] proved that both problems belong to NP $\cap$ co-NP. Porkolab and Khachiyan [21] established the following complexity results, which, in turn, are mainly based on the first-order theory of the reals developed by Renegar [24].

THEOREM 2.3. *Problems* **(PF)** *and* **(DF)** *can be solved in* $\kappa n^{O(\min\{\kappa, n^2\})}$ *and* $O(\kappa n^4) + n^{O(\min\{\kappa, n^2\})}$ *operations over the reals, respectively.*

In addition, let us consider the following optimization versions:

1. **(PO)** Given $D, A_1, A_2, \ldots, A_\kappa \in \mathcal{S}^n$ and $\beta_1, \ldots, \beta_\kappa \in \mathbb{R}$, solve
$$\alpha^* := \inf_{X \in \mathcal{S}^n} \{D \bullet X : A_i \bullet X \leq \beta_i, \quad i = 1, \ldots, \kappa, \quad X \succeq 0\}.$$
2. **(DO)** Given $B_0, B_1, \ldots, B_\kappa \in \mathcal{S}^n$ and $d \in \mathbb{R}^\kappa$, solve
$$\beta^* := \sup_{y_1, \ldots, y_\kappa \in \mathbb{R}} \left\{ \sum_{i=1}^{\kappa} d_i y_i : B_0 + y_1 B_1 + y_2 B_2 + \cdots + y_\kappa B_\kappa \succeq 0 \right\}.$$

The complexity results of Theorem 2.3 also extend to the optimization versions **(PO)** and **(DO)** [21].

THEOREM 2.4. *For problems* **(PO)** *and* **(DO)**, *each of the following can be solved in* $\kappa n^{O(\min\{\kappa, n^2\})}$ *and* $O(\kappa n^4) + n^{O(\min\{\kappa, n^2\})}$ *operations over the reals, respectively:* (i) *feasibility,* (ii) *boundedness,* (iii) *attainment of the optimal value, and* (iv) *computation of a least norm optimal solution.*

One important consequence of Theorems 2.3 and 2.4 is that semidefinite feasibility and semidefinite optimization can be solved in polynomial time if κ is fixed. We state this as a separate corollary.

COROLLARY 2.5. *Each of the four problems* **(PF)**, **(DF)**, **(PO)**, *and* **(DO)** *can be solved in polynomial time for fixed* κ.

This result will play a key role in our algorithm as the semidefinite feasibility and semidefinite optimization problems we will encounter will always satisfy the condition of the corollary.

2.2. Ellipsoid containment. In this section, we study the problem of deciding whether a given full-dimensional ellipsoid $\mathcal{E}$ is contained in another full-dimensional ellipsoid $\mathcal{E}^*$. Furthermore, we establish how to efficiently compute a point in $\mathcal{E}$ that is furthest from the center of $\mathcal{E}^*$ in terms of the ellipsoidal norm induced by $\mathcal{E}^*$.

We start with the following well-known result about polynomiality of quadratic optimization over an ellipsoid (see, e.g., [34]). We remark that this result can be found elsewhere in the literature (see, e.g., [27, 23, 28, 36, 6]). We mainly include it here for the sake of completeness. Our treatment can be considered as a special case of the more general proof of [28] and relies on the fact that the possibly nonconvex optimization problem admits a tight semidefinite programming (SDP) relaxation, whose optimal solution can be used to recover an optimal solution for the original problem.

PROPOSITION 2.6. *Any quadratic function $f : \mathbb{R}^n \to \mathbb{R}$ can be maximized over a full-dimensional ellipsoid in $O(n^{O(1)})$ operations, where $O(1)$ is a universal constant greater than three.*

Proof. Let $f(x) := x^T Ax + 2b^T x + \gamma$, where $A \in \mathcal{S}^n$, $b \in \mathbb{R}^n$, and $\gamma \in \mathbb{R}$, and let $\mathcal{E} \subset \mathbb{R}^n$ denote a full-dimensional ellipsoid, which admits a representation given by $\mathcal{E} := \{x \in \mathbb{R}^n : (x-c)^T Q(x-c) \leq 1\}$, where $Q \in \mathcal{S}^n$ is positive definite and $c \in \mathbb{R}^n$. We wish to solve

$$\text{(P)} \quad \max_{x \in \mathbb{R}^n} \{f(x) : x \in \mathcal{E}\}.$$

We consider the following SDP relaxation:

$$\text{(SP)} \quad \max_{X \in \mathcal{S}^{n+1}} \{F \bullet X : G \bullet X \leq 0,\, E_{n+1} \bullet X = 1,\, X \succeq 0\},$$

where

$$F := \begin{bmatrix} A & b \\ b^T & \gamma \end{bmatrix}, \quad G := \begin{bmatrix} Q & -Qc \\ -c^T Q & c^T Qc - 1 \end{bmatrix}, \quad E_{n+1} = e^{n+1}(e^{n+1})^T.$$

Note that (SP) is a relaxation of (P) since for any feasible solution $x \in \mathbb{R}^n$ of (P),

$$\begin{bmatrix} x \\ 1 \end{bmatrix} [x^T \quad 1] = \begin{bmatrix} xx^T & x \\ x^T & 1 \end{bmatrix} \succeq 0$$

is a feasible solution of (SP) with the same objective function value. We claim that the relaxation is exact in the sense that the optimal values of (P) and (SP) coincide and an optimal solution of (SP) can be converted into an optimal solution of (P).

Consider the following Lagrangian dual of (SP):

$$\text{(SD)} \quad \min_{\lambda, \beta} \{\beta :\ \lambda G + \beta E_{n+1} \succeq F,\, \lambda \geq 0\}.$$

We now make several observations about (SP) and (SD). Note that (SP) satisfies the Slater condition since the solution given by

$$\tilde{X} := \begin{bmatrix} cc^T + \alpha I & c \\ c^T & 1 \end{bmatrix}$$

satisfies $E_{n+1} \bullet \tilde{X} = 1$, $G \bullet \tilde{X} = -1 + \alpha\, Q \bullet I < 0$, for sufficiently small $\alpha > 0$, and $\tilde{X} \succ 0$, which implies that $\tilde{X}$ is a strictly feasible solution of (SP). Therefore, strong duality holds between (SP) and (SD), and the optimal value is attained in (SD). Furthermore, the feasible set of (SP) is bounded because the only solution to the system

$$G \bullet Y \leq 0, \quad E_{n+1} \bullet Y = 0, \quad Y \succeq 0, \quad Y \in \mathcal{S}^{n+1}$$

is $Y = 0$ since $Q \succ 0$. Therefore, the optimal value is also attained in (SP).

By Corollary 2.5, we can solve (SP) in $O(n^{O(1)})$ time (one can replace the equality constraint with two inequality constraints). Let X^* and (λ^*, β^*) denote optimal solutions of (SP) and (SD), respectively. It follows from optimality conditions that

$$X^* \bullet (\lambda^* G + \beta^* E_{n+1} - F) = 0, \quad \lambda^*(G \bullet X^*) = 0. \tag{6}$$

Since $G \bullet X^* \leq 0$, we can compute a rank-one decomposition of $X^* := \sum_{i=1}^{\rho} p^i (p^i)^T$, where $\rho := \text{rank}(X^*) \geq 1$ and $p^i \in \mathbb{R}^{n+1}$, $p^i \neq 0$, $i = 1, \ldots, \rho$, in $O(n^3)$ operations such that $(p^i)^T G p^i \leq 0$, $i = 1, \ldots, \rho$ [28, Proposition 3]. We now construct a rank-one optimal solution of (SP) using this decomposition.

By (6), $\sum_{i=1}^{\rho} (p^i)^T (\lambda^* G + \beta^* E_{n+1} - F) p^i = 0$, which implies that

$$(p^i)^T (\lambda^* G + \beta^* E_{n+1} - F) p^i = 0, \quad i = 1, \ldots, \rho, \tag{7}$$

by dual feasibility. Similarly, $\lambda^*(G \bullet X^*) = \lambda^* \sum_{i=1}^{\rho} (p^i)^T G p^i = 0$, which implies that

$$\lambda^* (p^i)^T G p^i = 0, \quad i = 1, \ldots, \rho, \tag{8}$$

since $(p^i)^T G p^i \leq 0$, $i = 1, \ldots, \rho$, and $\lambda^* \geq 0$.

Let j be any index in $\{1, 2, \ldots, \rho\}$ and let us define

$$p^j = \begin{bmatrix} x^j \\ \alpha^j \end{bmatrix},$$

where $x^j \in \mathbb{R}^n$ and $\alpha^j \in \mathbb{R}$. We claim that $\alpha^j \neq 0$. Otherwise, $0 \geq (p^j)^T G p^j = (x^j)^T Q x^j$, which implies that $x^j = 0$ since $Q \succ 0$, contradicting the fact that $p^j \neq 0$. We now let $x_*^j := (1/\alpha^j) p^j$. Since $G \bullet x_*^j (x_*^j)^T \leq 0$ and $E_{n+1} \bullet x_*^j (x_*^j)^T = 1$, it follows from (7) and (8) that $x_*^j (x_*^j)^T$ is a rank-one optimal solution of (SP), which implies that $(1/\alpha^j) x^j$ is an optimal solution of (P). (We remark that each of the indices in $\{1, 2, \ldots, \rho\}$ can be used to compute a different optimal solution of (P).) ∎

In fact, Proposition 2.6 holds true even if the ellipsoid defining the feasible region of the optimization problem is lower-dimensional. In this case, one can restrict the quadratic function f to the smallest affine subspace containing the ellipsoid and invoke the same analysis in the proof. We now use Proposition 2.6 to give a simple proof of the well-known characterization of the ellipsoid containment problem.

PROPOSITION 2.7. *Let $\mathcal{E} \subset \mathbb{R}^n$ and $\mathcal{E}^* \subset \mathbb{R}^n$ denote two full-dimensional ellipsoids with representations given by $\mathcal{E} := \{x \in \mathbb{R}^n : (x - c)^T Q (x - c) \leq 1\}$ and $\mathcal{E}^* := \{x \in \mathbb{R}^n : (x - c^*)^T Q^* (x - c^*) \leq 1\}$, where $Q \in \mathcal{S}^n$ and $Q^* \in \mathcal{S}^n$ are positive definite and $c \in \mathbb{R}^n$ and $c^* \in \mathbb{R}^n$. Then, $\mathcal{E} \subseteq \mathcal{E}^*$ if and only if there exists $\tau > 0$ such that*

$$\tau \begin{bmatrix} Q & -Qc \\ -c^T Q & c^T Q c - 1 \end{bmatrix} \succeq \begin{bmatrix} Q^* & -Q^* c^* \\ -c^{*T} Q^* & c^{*T} Q^* c^* - 1 \end{bmatrix}. \tag{9}$$

Proof. The statement follows directly from the $\mathcal{S}$-lemma [33] (see also [20] for a comprehensive treatment). However, we give a simple proof using standard duality arguments.

If (9) is satisfied, then we must have $\tau > 0$ since $Q \succ 0$ and $Q^* \succ 0$. Consider

$$\text{(P)} \quad \max_{x \in \mathbb{R}^n} \{(x - c^*)^T Q^* (x - c^*) - 1 : \ (x - c)^T Q (x - c) - 1 \leq 0\}.$$

By an argument similar to that in the proof of Proposition 2.6, it follows that

$$\text{(SP)} \quad \max_{X \in \mathcal{S}^{n+1}} \{F \bullet X : \ G \bullet X \leq 0, \, E_{n+1} \bullet X = 1, \, X \succeq 0\}$$

is a tight SDP relaxation of (P), where $F \in \mathcal{S}^{n+1}$ and $G \in \mathcal{S}^{n+1}$ are respectively given by

$$F = \begin{bmatrix} Q^* & -Q^*c^* \\ -c^{*T}Q^* & c^{*T}Q^*c^* - 1 \end{bmatrix}, \quad G = \begin{bmatrix} Q & -Qc \\ -c^TQ & c^TQc - 1 \end{bmatrix}.$$

The dual of (SP) is

$$\text{(SD)} \quad \min_{\lambda, \beta} \{\beta : \ \lambda G + \beta E_{n+1} \succeq F, \, \lambda \geq 0\}.$$

Let $v(P), v(SP)$, and $v(SD)$ denote the optimal values of (P), (SP), and (SD), respectively. It follows from the proof of Proposition 2.6 that

$$v(P) = v(SP) = v(SD). \tag{10}$$

Obviously, $\mathcal{E} \subseteq \mathcal{E}^*$ if and only if $v(P) \leq 0$. If (9) is feasible, then $(\lambda, \beta) = (\tau, 0)$ is a feasible solution of (SD), which implies that $v(P) = v(SD) \leq 0$. Conversely, if $v(P) \leq 0$, then let (λ^*, β^*) be an optimal solution of (SD) with optimal value $v(SD) = v(P) = \beta^* \leq 0$. Then

$$\lambda^* G \succeq \lambda^* G + \beta^* E_{n+1} \succeq F,$$

since $E_{n+1} \succeq 0$ and $\beta^* \leq 0$, which implies that λ^* is a feasible solution of (9). This completes the proof. □

We close this subsection by giving an equivalent characterization of (9).

LEMMA 2.8. *Condition* (9) *is equivalent to*

$$\tau \begin{bmatrix} Q & -Qc & 0 \\ -c^TQ & c^TQc - 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \succeq \begin{bmatrix} Q^* & -Q^*c^* & 0 \\ -c^{*T}Q^* & -1 & c^{*T}Q^* \\ 0 & Q^*c^* & -Q^* \end{bmatrix}. \tag{11}$$

Proof. We use the notation of Lemma 2.2. After rewriting (11) as a constraint of the form $A \succeq 0$, we let B denote the top left 2×2 block and define C and D accordingly. The equivalence now simply follows from the Schur complement lemma since $D := Q^* \succ 0$. □

We remark that condition (9) (or, equivalently, condition (11)) is a semidefinite constraint in a single variable. Therefore, it follows from Corollary 2.5 that ellipsoid containment can be checked in polynomial time.

It follows from (11) that the problem of computing the MVCE of a set of m full-dimensional ellipsoids can be formulated as a convex determinant maximization problem (see, e.g., [5, 6, 2]) with m linear matrix inequalities of size $(2n+1) \times (2n+1)$, m nonnegative variables $\tau_1, \ldots, \tau_m$, an $n \times n$ positive definite matrix variable Q^* that determines the shape and the orientation of the optimal ellipsoid, and an n-dimensional vector variable $z^* := Q^*c^*$, from which the center of the optimal ellipsoid can be recovered. As the dimension of the problem grows, the computational cost of interior-point algorithms [32, 31] quickly becomes prohibitive. This is one of our motivations to develop a specialized algorithm for the MVCE problem.

3. Initial volume approximation. Let $\mathcal{E}_1, \ldots, \mathcal{E}_m$ denote m full-dimensional ellipsoids, which admit representations given by

$$\mathcal{E}_i := \{x \in \mathbb{R}^n : (x - c^i)^T Q^i (x - c^i) \leq 1\},\ i = 1, \ldots, m, \tag{12}$$

where $Q^i \in \mathcal{S}^n$ is positive definite and $c^i \in \mathbb{R}^n,\ i = 1, \ldots, m$. We define $\mathcal{S} := \operatorname{conv}(\cup_{i=1}^m \mathcal{E}_i)$. In this section, we present a simple deterministic algorithm that identifies a finite subset $\mathcal{X}_0 \subset \cup_{i=1}^m \mathcal{E}_i$ of size $2n$ such that vol MVCE($\mathcal{X}_0$) is a provable approximation to vol MVCE($\mathcal{S}$).

ALGORITHM 3.1 (volume approximation algorithm).

Require: Input set $\mathcal{E}_1, \ldots, \mathcal{E}_m \subset \mathbb{R}^n$

1: $\Psi \leftarrow \{0\}$, $\mathcal{X}_0 \leftarrow \emptyset$, $k \leftarrow 0$.
2: While $\mathbb{R}^n \setminus \Psi \neq \emptyset$ do
3: **loop**
4: $k \leftarrow k+1$. Pick an arbitrary unit vector $b^k \in \mathbb{R}^n$ in the orthogonal complement of Ψ.
5: $x^{2k-1} \leftarrow \arg\max_{i=1,\ldots,m}\{(b^k)^T x : x \in \mathcal{E}_i\}$, $\mathcal{X}_0 \leftarrow \mathcal{X}_0 \cup \{x^{2k-1}\}$.
6: $x^{2k} \leftarrow \arg\min_{i=1,\ldots,m}\{(b^k)^T x : x \in \mathcal{E}_i\}$, $\mathcal{X}_0 \leftarrow \mathcal{X}_0 \cup \{x^{2k}\}$.
7: $\Psi \leftarrow \operatorname{span}(\Psi, \{x^{2k-1} - x^{2k}\})$.
8: **end loop**
9: **Output:** $\mathcal{X}_0$

LEMMA 3.1. *Algorithm* 3.1 *terminates after* $O(mn^3)$ *arithmetic operations and returns a subset* $\mathcal{X}_0 \subset \cup_{i=1}^m \mathcal{E}_i$ *with* $|\mathcal{X}_0| = 2n$ *such that*

$$\text{vol MVCE}(\mathcal{S}) \leq n^{2n} \text{vol MVCE}(\mathcal{X}_0). \tag{13}$$

Proof. We first establish the running time of Algorithm 3.1. At step k, Ψ is given by the span of k linearly independent vectors since $\mathcal{S}$ is full-dimensional. Hence, upon termination, $\Psi = \mathbb{R}^n$. It follows that $|\mathcal{X}_0| = 2n$. At each step, we optimize a linear function over each of the m ellipsoids $\mathcal{E}_i$. Let $Q^i = (U^i)^T U^i,\ i = 1, \ldots, m$, denote the Cholesky factorization of $Q^i,\ i = 1, \ldots, m$, which can be computed in $O(mn^3)$ operations. Note that $\mathcal{E}_i = \{x \in \mathbb{R}^n : x = (U^i)^{-1}u + c^i,\ \|u\| \leq 1\},\ i = 1, \ldots, m$. Therefore, at step k, each optimization problem has a closed form solution given by $\tilde{x}^{i,k}_{\max,\min} := c^i \pm (1/\|(U^i)^{-T} b^k\|)(U^i)^{-1}(U^i)^{-T} b^k$ with an optimal value of $(b^k)^T c^i \pm (1/\|(U^i)^{-T} b^k\|)(b^k)^T (U^i)^{-1}(U^i)^{-T} b^k$. For each ellipsoid $\mathcal{E}_i$, $\tilde{x}^{i,k}_{\max,\min}$ can be computed in $O(n^2)$ operations since U^i is upper triangular, which yields an overall computational cost of $O(mn^3)$ operations after n steps. Therefore, Algorithm 3.1 terminates after $O(mn^3)$ arithmetic operations.

We now prove (13). It follows from the results of Betke and Henk [3] that vol $\mathcal{S} \leq n!$ vol conv($\mathcal{X}_0$). Combining this inequality with Theorem 2.1, we obtain

$$\frac{1}{n^n} \text{ vol MVCE}(\mathcal{S}) \leq \text{vol } \mathcal{S} \leq n! \text{ vol conv}(\mathcal{X}_0) \leq n! \text{ vol MVCE}(\mathcal{X}_0),$$

which implies that vol MVCE($\mathcal{S}$) $\leq n! n^n$ vol MVCE($\mathcal{X}_0$) $\leq n^{2n}$ vol MVCE($\mathcal{X}_0$). ∎

4. A first-order algorithm. In this section, we present a first-order algorithm to compute a $(1 + \epsilon)$-approximation to the MVCE of the union of a set of full-dimensional ellipsoids $\mathcal{E}_1, \ldots, \mathcal{E}_m \subset \mathbb{R}^n$ for $\epsilon > 0$. Our algorithm is a generalization of the first-order algorithm presented in [18] to compute the MVCE of a finite set of m points, which, in turn, is obtained from a modification of Khachiyan's algorithm [15].

Our treatment closely follows the interpretation of Khachiyan's algorithm presented in [18].

As a by-product, we establish the existence of a finite core set $\mathcal{X} \subset \cup_{i=1,\ldots,m} \mathcal{E}_i$ whose size depends on only the dimension n and the parameter ϵ, but is independent of the number of ellipsoids m.

ALGORITHM 4.1 (a first-order algorithm that computes a $(1+\epsilon)$-approximation to MVCE($\mathcal{S}$)).

Require: Input set of ellipsoids $\mathcal{E}_1, \ldots, \mathcal{E}_m \subset \mathbb{R}^n$ given by (12) and $\epsilon > 0$.

1: Run Algorithm 3.1 on $\mathcal{E}_1, \ldots, \mathcal{E}_m$ to obtain output $\mathcal{X}_0 := \{x^1, \ldots, x^{2n}\}$.
2: $u^0 \leftarrow (1/2n)e \in \mathbb{R}^{2n}$.
3: $w^0 \leftarrow \sum_{j=1}^{2n} x^j u_j^0$.
4: $(M^0)^{-1} \leftarrow n \sum_{j=1}^{2n} u_j^0 (x^j - w^0)(x^j - w^0)^T$.
5: $\mathcal{F}_0 \leftarrow \{x \in \mathbb{R}^n : (x - w^0)^T M^0 (x - w^0) \leq 1\}$.
6: $x^{2n+1} \leftarrow \arg\max_{i=1,\ldots,m} \{(x - w^0)^T M^0 (x - w^0) : x \in \mathcal{E}_i\}$.
7: $\epsilon_0 \leftarrow (x^{2n+1} - w^0)^T M^0 (x^{2n+1} - w^0) - 1$.
8: $k \leftarrow 0$.
9: While $\epsilon_k > (1+\epsilon)^{2/n} - 1$ do
10: **loop**
11: $\beta_k \leftarrow \frac{\epsilon_k}{(n+1)(1+\epsilon_k)}$.
12: $k \leftarrow k + 1$.
13: $u^k \leftarrow \begin{bmatrix} (1-\beta_{k-1})u^{k-1} \\ \beta_{k-1} \end{bmatrix}$.
14: $w^k \leftarrow \sum_{j=1}^{2n+k} x^j u_j^k$.
15: $(M^k)^{-1} \leftarrow n \sum_{j=1}^{2n+k} u_j^k (x^j - w^k)(x^j - w^k)^T$.
16: $\mathcal{F}_k \leftarrow \{x \in \mathbb{R}^n : (x - w^k)^T M^k (x - w^k) \leq 1\}$.
17: $\mathcal{X}_k \leftarrow \mathcal{X}_{k-1} \cup \{x^{2n+k}\}$.
18: $x^{2n+k+1} \leftarrow \arg\max_{i=1,\ldots,m} \{(x - w^k)^T M^k (x - w^k) : x \in \mathcal{E}_i\}$.
19: $\epsilon_k \leftarrow (x^{2n+k+1} - w^k)^T M^k (x^{2n+k+1} - w^k) - 1$.
20: **end loop**
21: **Output:** $\sqrt{1+\epsilon_k}\,\mathcal{F}_k$, $\mathcal{X}_k$

We now describe Algorithm 4.1. Given m full-dimensional ellipsoids $\mathcal{E}_1, \ldots, \mathcal{E}_m \subset \mathbb{R}^n$ with representations given by (12), the algorithm calls Algorithm 3.1 and computes a finite set $\mathcal{X}_0 \subset \cup_{i=1}^m \mathcal{E}_i$ with $|\mathcal{X}_0| = 2n$. Next, a "trial ellipsoid" $\mathcal{F}_0$ is defined. Note that the center w^0 of $\mathcal{F}_0$ is simply the sample mean of $\mathcal{X}_0$ and M^0 is the inverse of the (scaled) sample covariance matrix of $\mathcal{X}_0$. ϵ_k measures the extent to which $\mathcal{F}_k$ should be enlarged around its center in order to cover $\mathcal{S} := \text{conv}(\cup_{i=1}^m \mathcal{E}_i)$. u^k can be viewed as a nonnegative weight vector whose components sum up to one. Note that the dimension of u^k increases by one at each iteration and is equal to $|\mathcal{X}_k|$. Unless the termination criterion is satisfied, the algorithm proceeds in an iterative manner as follows: At Step 13, u^k gets updated and is used to define w^k and M^k for the next trial ellipsoid $\mathcal{F}_k$. Observe that x^{2n+k} is precisely the farthest point in $\mathcal{S}$ from the center of the trial ellipsoid $\mathcal{F}_{k-1}$ in terms of its ellipsoidal norm. It is straightforward to verify that

$$w^k = (1-\beta_{k-1})w^{k-1} + \beta_{k-1}x^{2n+k}, \quad k = 1, 2, \ldots, \tag{14}$$

and

$$(M^k)^{-1} = (1-\beta_{k-1})(M^{k-1})^{-1} + n(1-\beta_{k-1})\beta_{k-1} d^k (d^k)^T \tag{15}$$

for $k = 1, 2, \ldots$, where $d^k := x^{2n+k} - w^{k-1}$. It follows that the next trial ellipsoid $\mathcal{F}_k$ is obtained by shifting the center of $\mathcal{F}_{k-1}$ towards x^{2n+k}, and its shape is determined by a nonnegative combination of $(M^{k-1})^{-1}$ and a rank-one update. This update can be viewed as "enriching the eigenspace of $(M^{k-1})^{-1}$ in the direction $d^k := x^{2n+k} - w^{k-1}$." We refer the reader to [18, section 4.3] for a related discussion. The parameter $\beta_{k-1} \in [0, 1)$ solves the following line search problem as observed by Khachiyan [15]:

$$(\mathrm{LS}(k)) \qquad \max_{\beta \in [0,1]} \log\det\left[(1-\beta)(M^{k-1})^{-1} + n(1-\beta)\beta d^k (d^k)^T\right]$$

for $k = 1, 2, \ldots$, where $d^k := x^{2n+k} - w^{k-1}$. Algorithm 4.1 terminates when the desired accuracy is achieved.

Algorithm 4.1 is an extension of the one proposed in [18] that computes a $(1+\epsilon)$-approximation to the MVCE of a finite set of *m points* in $\mathbb{R}^n$, which, in turn, is a modification of Khachiyan's algorithm [15]. The algorithm in [15] can be viewed as a sequential linear programming algorithm (or, equivalently, as a Frank–Wolfe algorithm [29]) for the nonlinear optimization problem arising from the dual formulation of the MVCE problem (cf. $(\mathbf{D}(\mathcal{X}_\mathbf{0}))$ in the proof of Theorem 4.1) for a finite set of points (see, e.g., the discussion in [18, section 4.1]). Algorithm 4.1 is motivated by the simple observation that the union of a set of ellipsoids in $\mathbb{R}^n$ can be viewed as an infinite set of points. Despite the fact that the finite-dimensional optimization formulation on which Khachiyan's algorithm is based no longer carries over to this more general setting, our main goal in this paper is to establish that essentially the same framework can be used with proper modifications to approximate the MVCE of the union of a finite number of ellipsoids. Since the algorithm is driven by linearizing the nonlinear objective function of the dual optimization formulation, we continue to refer to Algorithm 4.1 as a first-order algorithm. We remark that the interior-point algorithms of [32, 31] also rely on the second-order information arising from the Hessian of the objective function.

We next analyze the complexity of Algorithm 4.1. Our analysis resembles those of Khachiyan [15] and Kumar and Yıldırım [18]. The key ingredient in the complexity analysis is to demonstrate that Algorithm 4.1 produces a sequence $\{\mathcal{F}_k\}$ of trial ellipsoids with strictly increasing volumes. We utilize Lemma 3.1 to show that vol $\mathcal{F}_0$ is already a provable approximation to vol MVCE($\mathcal{S}$). The analysis will then be complete by establishing that each step of Algorithm 4.1 can be executed efficiently.

We start by proving that vol $\mathcal{F}_0$ is a provable approximation to vol MVCE($\mathcal{S}$).

THEOREM 4.1. *The ellipsoid $\mathcal{F}_0 \subset \mathbb{R}^n$ defined in Algorithm 4.1 satisfies*

$$\log \mathrm{vol}\ \mathcal{F}_0 \le \log \mathrm{vol}\ \mathrm{MVCE}(\mathcal{S}) \le \log \mathrm{vol}\ \mathcal{F}_0 + 2n \log n + \frac{n}{2} \log 2. \tag{16}$$

Proof. We first establish that

$$\log \mathrm{vol}\ \mathcal{F}_0 \le \log \mathrm{vol}\ \mathrm{MVCE}(\mathcal{X}_0), \tag{17}$$

where $\mathcal{X}_0 = \{x^1, \ldots, x^{2n}\}$ denotes the set of $2n$ points returned by Algorithm 3.1. Consider the following dual formulation to compute MVCE($\mathcal{X}_0$) (see, e.g., [15] or [29]):

$$(\mathbf{D}(\mathcal{X}_\mathbf{0})) \quad \begin{array}{ll} \max_u & \log\det \Pi_0(u) \\ \text{s.t.} & e^T u = 1, \\ & u \ge 0, \end{array}$$

where $u \in \mathbb{R}^{2n}$ is the decision variable and $\Pi_0 : \mathbb{R}^{2n} \to \mathcal{S}^{n+1}$ is a linear operator given by

$$\Pi_0(u) := \sum_{j=1}^{2n} u_j \begin{bmatrix} x^j(x^j)^T & x^j \\ (x^j)^T & 1 \end{bmatrix}.$$

MVCE($\mathcal{X}_0$) can be recovered from an optimal solution u^* of ($\mathbf{D}(\mathcal{X}_\mathbf{0})$) [18, Lemma 2.1]. Furthermore, the optimal value of ($\mathbf{D}(\mathcal{X}_\mathbf{0})$) satisfies

$$\log \text{vol MVCE}(\mathcal{X}_0) = \log \eta + \frac{n}{2} \log n + \frac{1}{2} \log \det \Pi_0(u^*), \tag{19}$$

where η is the volume of the unit ball in $\mathbb{R}^n$.

Let us consider the feasible solution $u^0 := (1/2n)e \in \mathbb{R}^{2n}$ of ($\mathbf{D}(\mathcal{X}_\mathbf{0})$). We have

$$\begin{aligned} \Pi_0(u^0) &= \begin{bmatrix} (1/2n)\sum_{j=1}^{2n} x^j(x^j)^T & w^0 \\ (w^0)^T & 1 \end{bmatrix}, \\ &= \begin{bmatrix} I & w^0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} (1/n)(M^0)^{-1} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} I & 0 \\ (w^0)^T & 1 \end{bmatrix}, \end{aligned} \tag{20}$$

which implies that

$$\log \det \Pi_0(u^0) = -n \log n + \log \det (M^0)^{-1} = -n \log n + 2 \log \det (M^0)^{-1/2}. \tag{21}$$

However, $\log \text{vol } \mathcal{F}_0 = \log \eta + \log \det (M^0)^{-1/2}$. Combining this equality with (21), we obtain

$$\log \text{vol } \mathcal{F}_0 = \log \eta + \frac{n}{2} \log n + \frac{1}{2} \log \det \Pi_0(u^0).$$

Since u^0 is a feasible solution for the maximization problem ($\mathbf{D}(\mathcal{X}_\mathbf{0})$), it follows from (19) that $\log \text{vol } \mathcal{F}_0 \leq \log \text{vol MVCE}(\mathcal{X}_0)$.

Since $\mathcal{X}_0 \subset \mathcal{S}$, we clearly have $\log \text{vol MVCE}(\mathcal{X}_0) \leq \log \text{vol MVCE}(\mathcal{S})$, which proves the first inequality in (16). To prove the second inequality, let

$$B := [x^1, \ldots, x^{2n}] \in \mathbb{R}^{n \times 2n}.$$

Then, $w^0 = (1/2n)Be$ and it is easy to verify that $(M^0)^{-1} = (1/2)BPB^T$, where $P := I - (1/2n)ee^T$ is an orthogonal projection matrix onto the orthogonal complement of the vector e. Note that $Pe^j = e^j - (1/2n)e$, $j = 1, \ldots, 2n$. Therefore, for any $j = 1, \ldots, 2n$, we have

$$\begin{aligned} (x^j - w^0)^T M^0 (x^j - w^0) &= 2(e^j - (1/2n)e)^T B^T (BPB^T)^{-1} B(e^j - (1/2n)e), \\ &= 2(Pe^j)^T PB^T (BP^2B^T)^{-1} BP(Pe^j), \\ &\leq 2\|Pe^j\|^2, \\ &= \frac{2n-1}{n}, \\ &< 2, \end{aligned}$$

where we used $P = P^2$ on the second line and the fact that $PB^T(BP^2B^T)^{-1}BP$ is an orthogonal projection matrix to derive the first inequality. Consequently, the

ellipsoid $\mathcal{G} := \{x \in \mathbb{R}^n : (x - w^0)^T (1/2) M^0 (x - w^0) \le 1\}$ covers $\mathcal{X}_0$. Therefore,

$$\begin{aligned}\log \text{vol MVCE}(\mathcal{X}_0) &\le \log \text{vol } \mathcal{G},\\ &= \log \eta + \frac{n}{2}\log 2 + \log \det (M^0)^{-1/2},\\ &= \frac{n}{2}\log 2 + \log \text{vol } \mathcal{F}_0,\end{aligned}$$

which implies that $\log \text{vol } \mathcal{F}_0 \ge \log \text{vol MVCE}(\mathcal{X}_0) - (n/2)\log 2$. By Lemma 3.1, we have $\log \text{vol MVCE}(\mathcal{X}_0) \ge \log \text{vol MVCE}(\mathcal{S}) - 2n \log n$. Combining these two inequalities, we obtain $\log \text{vol } \mathcal{F}_0 + 2n \log n + (n/2)\log 2 \ge \log \text{vol MVCE}(\mathcal{S})$ as desired. ∎

The next lemma relates $\log \text{vol } \mathcal{F}_k$ to $\log \text{vol MVCE}(\mathcal{S})$.

LEMMA 4.2. *For any $k = 0, 1, 2, \ldots$, we have*

$$\log \text{vol } \mathcal{F}_k \le \log \text{vol MVCE}(\mathcal{S}) \le \log \text{vol } \mathcal{F}_k + \frac{n}{2}\log(1 + \epsilon_k). \tag{22}$$

Proof. By definition of ϵ_k, $\sqrt{1+\epsilon_k}\,\mathcal{F}_k \supseteq \mathcal{S}$, where $\sqrt{1+\epsilon_k}\,\mathcal{F}_k$ is given by expanding $\mathcal{F}_k$ around its center w^k by a factor of $\sqrt{1+\epsilon_k}$. Therefore, $\log \text{vol MVCE}(\mathcal{S}) \le \log \text{vol } \mathcal{F}_k + (n/2)\log(1+\epsilon_k)$, which proves the second inequality in (22).

We follow an argument similar to that in the proof of Theorem 4.1 to establish the first inequality (cf. (20), (21), and (19)). At step k of Algorithm 4.1, $u^k \in \mathbb{R}^{2n+k}$ is a feasible solution of the optimization problem $(\mathbf{D}(\mathcal{X}_\mathbf{k}))$. Therefore,

$$\begin{aligned}\log \text{vol } \mathcal{F}_k &= \log \eta + \frac{n}{2}\log n + \frac{1}{2}\log \det \Pi_k(u^k),\\ &\le \log \eta + \frac{n}{2}\log n + \frac{1}{2}\log \det \Pi_k(u^k_*),\\ &= \log \text{vol MVCE}(\mathcal{X}_k),\end{aligned}$$

where u^k_* denotes the optimal solution of $(\mathbf{D}(\mathcal{X}_\mathbf{k}))$ and $\Pi_k : \mathbb{R}^{2n+k} \to \mathcal{S}^{n+1}$ is a linear operator given by

$$\Pi_k(u) := \sum_{j=1}^{2n+k} u_j \begin{bmatrix} x^j (x^j)^T & x^j \\ (x^j)^T & 1 \end{bmatrix}. \tag{23}$$

Since $\mathcal{X}_k \subset \mathcal{S}$, the first inequality follows. ∎

The following corollary immediately follows from Lemma 4.2.

COROLLARY 4.3. *For any $k = 0, 1, 2, \ldots$, $\epsilon_k \ge 0$. Furthermore, if Algorithm 4.1 does not terminate at step k, then $\epsilon_k > (1+\epsilon)^{2/n} - 1$.*

So far, we have established the following results: (i) vol $\mathcal{F}_0$ is a provable approximation to vol MVCE($\mathcal{S}$) and (ii) the sequence of ellipsoids $\mathcal{F}_k$ generated by Algorithm 4.1 yields a sequence of lower bounds on vol MVCE($\mathcal{S}$). Our next goal is to demonstrate that $\{\text{vol } \mathcal{F}_k\}, k = 0, 1, \ldots$, is a strictly increasing sequence, which implies that Algorithm 4.1 produces increasingly sharper lower bounds to vol MVCE($\mathcal{S}$). At this stage, it is worth noticing that the line search problem LS(k) precisely computes the next trial ellipsoid which yields the largest increase in the volume for the particular updating scheme of Algorithm 4.1.

PROPOSITION 4.4. *For any $k = 0, 1, 2, \ldots$,*

$$\log \text{vol } \mathcal{F}_{k+1} \ge \log \text{vol } \mathcal{F}_k + \begin{cases} \frac{1}{2}\log 2 - \frac{1}{4} > 0 & \text{if } \epsilon_k \ge \frac{n+1}{n}, \\ \frac{1}{16}\epsilon_k^2 & \text{if } \epsilon_k < \frac{n+1}{n}. \end{cases} \tag{24}$$

Proof. Our proof mimics Khachiyan's argument [15]. By the definition of ϵ_k, we have $1+\epsilon_k = (x^{2n+k+1}-w^k)^T M^k (x^{2n+k+1}-w^k)$. Let $z^k := x^{2n+k+1}-w^k$. It follows from (15) that

$$\begin{aligned}
\log\det(M^{k+1})^{-1} &= \log\det\left\{(1-\beta_k)\left[(M^k)^{-1}+n\beta_k z^k(z^k)^T\right]\right\},\\
&= n\log(1-\beta_k)+\log\det\left[(M^k)^{-1}\left(I+n\beta_k M^k z^k(z^k)^T\right)\right],\\
&= \log\det(M^k)^{-1}+n\log(1-\beta_k)+\log\left[1+n\beta_k(1+\epsilon_k)\right],\\
&= \log\det(M^k)^{-1}-n\log\left(1+\frac{\beta_k}{1-\beta_k}\right)+\log\left(1+\left(\frac{n}{n+1}\right)\epsilon_k\right),\\
&= \log\det(M^k)^{-1}-n\log\left(1+\frac{\epsilon_k}{(n+1)(1+\epsilon_k)-\epsilon_k}\right)\\
&\qquad +\log\left(1+\left(\frac{n}{n+1}\right)\epsilon_k\right),\\
&\geq \log\det(M^k)^{-1}-\frac{\left(\frac{n}{n+1}\right)\epsilon_k}{1+\left(\frac{n}{n+1}\right)\epsilon_k}+\log\left(1+\left(\frac{n}{n+1}\right)\epsilon_k\right),
\end{aligned}$$

where we used the definition of β_k in the last two equalities and the inequality $\log(1+\zeta)\leq\zeta$ for $\zeta>-1$. Since $\log\text{vol}\,\mathcal{F}_k = \log\eta+\log\det(M^k)^{-1/2} = \log\eta+(1/2)\log\det(M^k)^{-1}$, it follows that

$$\log\text{vol}\,\mathcal{F}_{k+1}\geq\log\text{vol}\,\mathcal{F}_k+\frac{1}{2}\log\left(1+\left(\frac{n}{n+1}\right)\epsilon_k\right)-\frac{\left(\frac{n}{n+1}\right)\epsilon_k}{2\left(1+\left(\frac{n}{n+1}\right)\epsilon_k\right)}.$$

The assertion follows from the observation that $f(\nu) := (1/2)\log(1+\nu)-\nu/[2(1+\nu)]$ is a strictly increasing function for $\nu\geq 0$ and $f(\nu)\geq\nu^2/16$ for $\nu\in[0,1]$. ∎

We are now ready to analyze the iteration complexity of Algorithm 4.1. To this end, we define the following parameters:

$$\tau_\rho := \min\left\{k:\left(\frac{n}{n+1}\right)\epsilon_k\leq 1/2^\rho\right\},\quad \rho=0,1,2,\ldots. \tag{25}$$

The next lemma establishes certain properties of τ_ρ.

LEMMA 4.5. *τ_ρ satisfies the following relationships:*

$$\tau_0 = O(n\log n), \tag{26}$$

$$\tau_\rho-\tau_{\rho-1}\leq n2^{\rho+5},\quad \rho=1,2,\ldots. \tag{27}$$

Proof. By Theorem 4.1, $\log\text{vol}\,\mathcal{F}_0\leq\log\text{vol MVCE}(\mathcal{S})\leq\log\text{vol}\,\mathcal{F}_0+2n\log n+(n/2)\log 2$. At every iteration k with $\epsilon_k>(n+1)/n$, we have $\log\text{vol}\,\mathcal{F}_{k+1}-\log\text{vol}\,\mathcal{F}_k\geq(1/2)\log 2-1/4>0$ by Proposition 4.4. Therefore, $\tau_0=O(n\log n)$.

Let us now consider $\tau_\rho-\tau_{\rho-1}$, $\rho\geq 1$. For simplicity, let $\gamma := \tau_{\rho-1}$. By definition of $\tau_{\rho-1}$, it follows from Lemma 4.2 that $\log\text{vol}\,\mathcal{F}_\gamma\leq\log\text{vol MVCE}\,(\mathcal{S})\leq\log\text{vol}\,\mathcal{F}_\gamma+(n/2)\log(1+[(n+1)/n]2^{-(\rho-1)})\leq\log\text{vol}\,\mathcal{F}_\gamma+(n+1)2^{-\rho}$. By Proposition 4.4, $\log\text{vol}\,\mathcal{F}_{k+1}-\log\text{vol}\,\mathcal{F}_k\geq[(n+1)/n]^2 2^{-(2\rho+4)}\geq 2^{-(2\rho+4)}$ at every iteration k with $\epsilon_k>[(n+1)/n]2^{-\rho}$. Therefore, $\tau_\rho-\tau_{\rho-1}\leq[(n+1)2^{-\rho}]/2^{-(2\rho+4)}=(n+1)2^{\rho+4}\leq n2^{\rho+5}$, which completes the proof. ∎

Lemma 4.5 enables us to establish the following result.

LEMMA 4.6. *Let $\mu \in (0,1)$. Algorithm* 4.1 *computes an iterate with $\epsilon_k \leq \mu$ in $O(n(\log n + \mu^{-1}))$ iterations.*

Proof. Let σ be a positive integer such that $[(n+1)/n]2^{-\sigma} \leq \mu \leq [(n+1)/n]2^{1-\sigma}$. Therefore, after $k = \tau_\sigma$ iterations, we already have $\epsilon_k \leq [(n+1)/n]2^{-\sigma} \leq \mu$. However,

$$\tau_\sigma = \tau_0 + \sum_{\rho=1}^{\sigma}(\tau_\rho - \tau_{\rho-1}) \leq \tau_0 + 64n\sum_{\rho=1}^{\sigma} 2^{\rho-1} \leq \tau_0 + 64n2^\sigma \leq O\left(n\log n + \frac{n}{\mu}\right),$$

where we used Lemma 4.5 and the inequality $2^\sigma \leq 4/\mu$. ∎

We are now in a position to establish the iteration complexity of Algorithm 4.1.

THEOREM 4.7. *Let $\epsilon > 0$. Algorithm* 4.1 *computes a $(1+\epsilon)$-approximation to MVCE($\mathcal{S}$) after at most $O(n(\log n + [(1+\epsilon)^{2/n} - 1]^{-1}))$ iterations.*

Proof. We first establish that Algorithm 4.1 returns a $(1+\epsilon)$-approximation to MVCE($\mathcal{S}$) upon termination. Let κ denote the index of the final iterate. We have $\epsilon_\kappa \leq (1+\epsilon)^{2/n} - 1$. The trial ellipsoid $\mathcal{F}_\kappa$ satisfies $\mathcal{S} \subseteq \sqrt{1+\epsilon_\kappa}\,\mathcal{F}_\kappa$, which together with Lemma 4.2 implies that

$$\text{vol } \mathcal{F}_\kappa \leq \text{vol MVCE}(\mathcal{S}) \leq \text{vol}\sqrt{1+\epsilon_\kappa}\mathcal{F}_\kappa = (1+\epsilon_\kappa)^{n/2}\text{vol } \mathcal{F}_\kappa \leq (1+\epsilon)\text{vol } \mathcal{F}_\kappa.$$

Therefore, $\sqrt{1+\epsilon_\kappa}\,\mathcal{F}_\kappa$ is indeed a $(1+\epsilon)$-approximation to MVCE($\mathcal{S}$).

We now prove the iteration complexity. If $\epsilon \geq [2+(1/n)]^{n/2} - 1$, then $(1+\epsilon)^{2/n} - 1 \geq (n+1)/n$, which implies that at most $\tau_0 = O(n\log n)$ iterations already suffice. Otherwise, the result follows from Lemma 4.6. ∎

Remark 1. The iteration complexity of Algorithm 4.1 is asymptotically identical to that of the algorithm of Kumar and Yıldırım [18] that computes a $(1+\epsilon)$-approximation to the MVCE of a finite set of m points.

We now establish the overall complexity of Algorithm 4.1.

THEOREM 4.8. *Algorithm* 4.1 *computes a $(1+\epsilon)$-approximation to MVCE($\mathcal{S}$) in*

$$O\left(mn^{O(1)}(\log n + [(1+\epsilon)^{2/n} - 1]^{-1})\right)$$

operations, where $O(1)$ denotes a universal constant greater than four.

Proof. We already have the iteration complexity from Theorem 4.7. We need only analyze the computational cost of each iteration.

Let us start with the initialization stage. By Lemma 3.1, Algorithm 3.1 runs in $O(mn^3)$ operations. w^0 and $(M^0)^{-1}$ can be computed in $O(n^2)$ and $O(n^3)$ operations, respectively. The furthest point x^{2n+1} from the center of $\mathcal{F}_0$ can be determined by solving m separate quadratic optimization problems with a single ellipsoidal constraint. By Proposition 2.6, each optimization problem can be solved in $O(n^{O(1)}+n^3)$ operations. Finally, it takes $O(n^2)$ operations to compute ϵ_0. Therefore, the overall complexity of the initialization step is $O(m(n^{O(1)} + n^3))$ operations. Similarly, at iteration k, the major work is the computation of the furthest point x^{2n+k+1}, which can be performed in $O(m(n^{O(1)}+n^3))$ operations. Therefore, the overall running time of Algorithm 4.1 is given by $O(mn^{O(1)}(\log n + [(1+\epsilon)^{2/n} - 1]^{-1}))$ operations. ∎

Remark 2. The overall complexity of Algorithm 4.1 is linear in m, the number of ellipsoids. This suggests that, in theory, Algorithm 4.1 is especially well-suited for instances of the MVCE problem that satisfy $m \gg n$ and for moderate values of ϵ. In addition, if $\epsilon \in (0,1)$, we have $(1+\epsilon)^{2/n} - 1 = \Theta(\epsilon/n)$, in which case the running time of Algorithm 4.1 can be simplified to $O((1/\epsilon)mn^{O(1)})$ operations, where $O(1)$ is now

a universal constant greater than five. Note that the running time of Algorithm 4.1 is polynomial for fixed ϵ.

We close this section by establishing that the convex hull of the finite set of points collected by Algorithm 4.1 serves as a reasonably good approximation to $\mathcal{S}$ in the sense that their respective MVCEs are closely related.

PROPOSITION 4.9. *Let κ denote the index of the final iterate of Algorithm* 4.1. *Then, $\mathcal{X}_\kappa$ satisfies*

$$\text{vol } MVCE(\mathcal{X}_\kappa) \leq \text{vol } MVCE(\mathcal{S}) \leq (1+\epsilon)\text{vol } MVCE(\mathcal{X}_\kappa). \tag{28}$$

In addition,

$$|\mathcal{X}_\kappa| = O\left(n(\log n + [(1+\epsilon)^{2/n} - 1]^{-1})\right). \tag{29}$$

Proof. We first prove (28). Note that the first inequality is obvious since $\mathcal{X}_\kappa \subset \mathcal{S}$. The second inequality follows from the relationships vol $\mathcal{F}_\kappa \leq$ vol MVCE$(\mathcal{X}_\kappa) \leq$ vol MVCE$(\mathcal{S})$ (see the proof of Lemma 4.2) and vol MVCE$(\mathcal{S}) \leq (1+\epsilon)$vol $\mathcal{F}_\kappa$ (see the proof of Theorem 4.7).

Since $|\mathcal{X}_\kappa| = 2n + \kappa$, (29) simply follows from Theorem 4.7. ∎

Remark 3. Proposition 4.9 establishes that Algorithm 4.1 computes a finite set of points $\mathcal{X}_\kappa \subset \mathcal{S}$ whose MVCE is related to MVCE$(\mathcal{S})$ via (28). In addition, $|\mathcal{X}_\kappa|$ depends only on the dimension n and the approximation factor ϵ but is independent of the number of ellipsoids m. Furthermore, for $\epsilon \in (0,1)$, we have $|\mathcal{X}_\kappa| = O(n^2/\epsilon)$. Therefore, $\mathcal{X}_\kappa$ serves as a finite core set for $\mathcal{S}$. Viewed from this perspective, Proposition 4.9 is an addition to the previous core set results for other geometric optimization problems [17, 8, 7, 9, 1, 18].

Remark 4. In [18], a similar core set result has been established for the MVCE problem for a finite set of m points in $\mathbb{R}^n$. It is remarkable that asymptotically the same result holds regardless of the difference in the underlying geometric structures of the two input sets. In particular, the main ingredient in [18] that leads to the improved complexity result over Khachiyan's algorithm [15] as well as the core set result is the initial volume approximation. In a similar manner, the counterpart of this initialization stage (cf. Algorithm 3.1) enables us to extend the algorithm of Kumar and Yıldırım to a set of ellipsoids. Khachiyan's algorithm cannot be extended to a set of ellipsoids as it relies on the finiteness property of the input set at the initialization stage.

5. Rounding. In this section, we establish that Algorithm 4.1 can also be used to compute a $(1+\delta)n$-rounding of $\mathcal{S} := \text{conv}(\cup_{i=1}^m \mathcal{E}_i)$, where $\mathcal{E}_1, \ldots, \mathcal{E}_m \subset \mathbb{R}^n$ are full-dimensional ellipsoids and $\delta > 0$. We assume that $\mathcal{S}$ is not symmetric around the origin.

Our analysis closely follows Khachiyan's treatment for an input set of a finite number of points in $\mathbb{R}^n$ [15]. At iteration k of Algorithm 4.1, let $q^j := [(x^j)^T, 1]^T \in \mathbb{R}^{n+1}$, $j = 1, \ldots, 2n+k$, and let $\mathcal{Q}_k := \text{conv}(\{\pm q^1, \ldots, \pm q^{2n+k}\})$, which is a centrally symmetric polytope in $\mathbb{R}^{n+1}$ (i.e., $\mathcal{Q}_k = -\mathcal{Q}_k$).

Let $u^k \in \mathbb{R}^{2n+k}$ denote the iterate at iteration k of Algorithm 4.1. Let us define a full-dimensional ellipsoid $\mathcal{G}_k \subset \mathbb{R}^{n+1}$ given by

$$\mathcal{G}_k := \{y \in \mathbb{R}^{n+1} : y^T \Pi_k (u^k)^{-1} y \leq 1\}, \tag{30}$$

where $\Pi_k : \mathbb{R}^{2n+k} \to \mathcal{S}^{n+1}$ is a linear operator defined by (23). Since u^k is a feasible solution of $(\mathbf{D}(\mathcal{X}_\mathbf{k}))$, it follows from [15, Lemma 2] that $\mathcal{G}_k \subseteq \mathcal{Q}_k$. Furthermore, for any $q^j,\ j = 1, \ldots, 2n+k$, we have

$$\begin{aligned}(q^j)^T \Pi_k (u^k)^{-1} q^j &= [(x^j)^T \quad 1] \begin{bmatrix} I & 0 \\ -(w^k)^T & 1 \end{bmatrix} \begin{bmatrix} nM_k & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} I & -w^k \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x^j \\ 1 \end{bmatrix}, \\ &= n(x^j - w^k)^T M_k (x^j - w^k) + 1, \\ &\leq n(1+\epsilon_k) + 1,\end{aligned}$$

which, together with the previous inclusion, implies that

$$\mathcal{G}_k \subseteq \mathcal{Q}_k \subseteq \sqrt{1 + n(1+\epsilon_k)}\, \mathcal{G}_k; \tag{31}$$

i.e., $\sqrt{1+n(1+\epsilon_k)}\,\mathcal{G}_k$ is a $\sqrt{(1+\tilde{\delta})(n+1)}$-rounding of $\mathcal{Q}_k$, where $\tilde{\delta} := (n\epsilon_k)/(n+1)$.

Let $\mathcal{H}_k := \{x \in \mathbb{R}^n : [x^T \quad 1]^T \in \sqrt{1+n(1+\epsilon_k)}\,\mathcal{G}_k \cap \Lambda\}$, where

$$\Lambda := \{y \in \mathbb{R}^{n+1} : y_{n+1} = 1\}. \tag{32}$$

Note that $\mathcal{H}_k \subset \mathbb{R}^n$ is a full-dimensional ellipsoid. By [15, Lemma 5],

$$\frac{1}{(1+\epsilon_k)n}\, \mathcal{H}_k \subseteq \operatorname{conv}(\mathcal{X}_k) \subseteq \mathcal{H}_k; \tag{33}$$

i.e., $\mathcal{H}_k \subset \mathbb{R}^n$ is a $(1+\epsilon_k)n$-rounding of $\operatorname{conv}(\mathcal{X}_k)$. However, it is straightforward to verify that $x \in \mathcal{H}_k$ if and only if $(x - w^k)^T M^k (x - w^k) \leq 1 + \epsilon_k$, which implies that $\mathcal{H}_k = \sqrt{1+\epsilon_k}\,\mathcal{F}_k$. Since $\operatorname{conv}(\mathcal{X}_k) \subseteq \mathcal{S} \subseteq \sqrt{1+\epsilon_k}\,\mathcal{F}_k$, it follows from (33) that

$$\frac{1}{(1+\epsilon_k)n}\, \mathcal{H}_k \subseteq \operatorname{conv}(\mathcal{X}_k) \subseteq \mathcal{S} \subseteq \mathcal{H}_k; \tag{34}$$

i.e., $\mathcal{H}_k$ is simultaneously a $(1+\epsilon_k)n$-rounding of $\operatorname{conv}(\mathcal{X}_k)$ and of $\mathcal{S}$. Therefore, in order to obtain a $(1+\delta)n$-rounding of $\mathcal{S}$, it suffices to run Algorithm 4.1 until $\epsilon_k \leq \delta$. We summarize this result in the following corollary, whose proof follows directly from Lemma 4.6.

COROLLARY 5.1. *Given $\delta > 0$, Algorithm* 4.1 *computes a $(1+\delta)n$-rounding of $\mathcal{S}$ in*

$$O(mn^{O(1)}(\log n + \delta^{-1}))$$

arithmetic operations, where $O(1)$ is a universal constant greater than four. In addition, upon termination of Algorithm 4.1*, the ellipsoid computed by Algorithm* 4.1 *is also a $(1+\delta)n$-rounding of the convex hull of a finite subset $\mathcal{X}_k \subset \mathcal{S}$ with the property that*

$$|\mathcal{X}_k| = O\left(n(\log n + \delta^{-1})\right). \tag{35}$$

Remark 5. Upon termination of Algorithm 4.1 with a $(1+\delta)n$-rounding of $\mathcal{S}$, Corollary 5.1 establishes that $\mathcal{X}_k$ is an $\tilde{\epsilon}$-core set for $\mathcal{S}$, where $\tilde{\epsilon} := (1+\delta)^{n/2} - 1$. In fact, Khachiyan's algorithm [15] is motivated by first computing a $(1+\delta)n$-rounding of $\mathcal{S}$ and then choosing δ in such a way that the ellipsoid computed by the algorithm is a $(1+\epsilon)$-approximation to the MVCE.

We close this section by noting that Corollary 5.1 can be improved if $\mathcal{S}$ is centrally symmetric. In this case, we no longer need to "lift" the vectors in $\mathcal{X}_k$ to $\mathbb{R}^{n+1}$. A similar argument can be invoked to establish that $\mathcal{H}_k := \sqrt{1+\epsilon_k}\,\mathcal{F}_k$ satisfies

$$\frac{1}{\sqrt{(1+\epsilon_k)n}}\, \mathcal{H}_k \subseteq \operatorname{conv}(\{\pm x^1, \ldots, \pm x^{2n+k}\}) \subseteq \mathcal{S} \subseteq \mathcal{H}_k.$$

6. Extensions to other sets. In this section, we discuss the extent to which Algorithm 4.1 can be used to compute an approximate MVCE and an approximate n-rounding of other input sets. Let $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_m$ denote m objects in $\mathbb{R}^n$ and let $\mathcal{T} := \cup_{i=1}^m \mathcal{T}_i$. In order to extend Algorithm 4.1, we identify the following two subroutines that need to be implemented efficiently:

1. *Subroutine* 1: Optimizing a linear function over $\mathcal{T}$.
2. *Subroutine* 2: Maximizing a quadratic function over $\mathcal{T}$.

Note that Subroutine 1 is required by Algorithm 3.1. Similarly, the computation of the furthest point from the center of a trial ellipsoid (in its ellipsoidal norm) is equivalent to Subroutine 2. All of the other operations of Algorithm 4.1 can be performed efficiently for any input set $\mathcal{T}$.

Let us now consider specific examples of input sets. Clearly, a finite set of points and a finite set of balls would be special cases of a finite set of ellipsoids. Therefore, Algorithm 4.1 would be trivially applicable in these cases. We just remark that certain subroutines can be implemented more efficiently for these input sets. For a finite set of points, Subroutines 1 and 2 can be performed in $O(mn)$ and $O(mn^2)$ operations, respectively. We then recover the algorithm of Kumar and Yıldırım [18]. For a finite set of balls, while Subroutine 1 can still be implemented in $O(mn)$ operations, Subroutine 2 would require the same computational cost as that required by a set of ellipsoids. Therefore, the running time of Algorithm 4.1 would asymptotically remain the same for a finite set of balls. A similar argument also holds for an input set of ellipsoids, each of which is defined by the same matrix $Q = Q^i,\ i = 1, \ldots, m$, since such an input set can be a priori transformed into a set of balls.

6.1. Set of half ellipsoids. Let $\mathcal{T}_i := \{x \in \mathbb{R}^n : (x - c^i)^T Q^i (x - c^i) \leq 1, (f^i)^T x \leq \alpha^i\}$, where $c^i \in \mathbb{R}^n$ and $Q^i \in \mathcal{S}^n$ are positive definite, $f^i \in \mathbb{R}^n$ with $f^i \neq 0$, and $\alpha^i \in \mathbb{R},\ i = 1, \ldots, m$. $\mathcal{T}_i$ is simply given by the intersection of a full-dimensional ellipsoid and a half-space. We will refer to each $\mathcal{T}_i$ as a half-ellipsoid. To avoid trivialities, we assume that each $\mathcal{T}_i$ has a nonempty interior. It follows from the results of Sturm and Zhang [28] that the problem of optimizing any quadratic (hence linear) objective function over $\mathcal{T}_i$ can be cast as an equivalent SDP problem with a fixed number of constraints using a technique similar to that used in the proof of Proposition 2.6. Since both Subroutines 1 and 2 naturally decompose into a linear and quadratic optimization problem over each $\mathcal{T}_i$, respectively, it follows from Corollary 2.5 that both of them can be implemented in polynomial time. Therefore, Algorithm 4.1 can compute an approximate MVCE and an approximate n-rounding of a set of half-ellipsoids in polynomial time.

6.2. Set of intersections of a pair of ellipsoids. Let $\mathcal{T}_i := \{x \in \mathbb{R}^n : (x - c^i)^T Q^i (x - c^i) \leq 1, (x - h^i)^T Q^i (x - h^i) \leq 1\}$, where $c^i \in \mathbb{R}^n$, $h^i \in \mathbb{R}^n$, and $Q^i \in \mathcal{S}^n$ are positive definite, $i = 1, \ldots, m$. Note that each $\mathcal{T}_i$ is given by the intersection of two ellipsoids defined by the same matrix Q^i with different centers. Similarly to the previous case, Sturm and Zhang [28] establish that the problem of optimizing any quadratic (hence linear) objective function over $\mathcal{T}_i$ can be decomposed into two quadratic (linear) optimization problems over a half-ellipsoid, each of which can be solved in polynomial time. Therefore, Algorithm 4.1 can compute an approximate MVCE and an approximate n-rounding of a set of intersections of a pair of ellipsoids in polynomial time. We remark that the complexity of solving a general quadratic optimization problem over the intersection of two arbitrary ellipsoids is still an open problem.

6.3. Other sets and limitations. Based on the previous examples, it is clear that Algorithm 4.1 can be applied to any input set as long as Subroutines 1 and 2 admit efficient implementations. While Subroutine 1 can be performed efficiently for a rather large class of input sets (e.g., classes of convex sets that admit efficiently computable barrier functions [19]), Subroutine 2 can be efficiently implemented only in very special cases, some of which have been outlined in this section.

For instance, if $\mathcal{T}$ is given by the union of polytopes $\mathcal{T}_i := \{x \in \mathbb{R}^n : A^i x \leq b^i\}$, where $A^i \in \mathbb{R}^{m \times n}$ and $b^i \in \mathbb{R}^m$, $i = 1, \ldots, m$, then Subroutine 1 reduces to linear programming, which can be solved efficiently using interior-point methods combined with a finite termination strategy [35]. However, maximizing a convex quadratic function over a polytope is in general an NP-hard problem. Therefore, even in the case of a single polytope defined by linear inequalities, the problem of computing an approximate MVCE is computationally intractable. We remark that the maximum volume inscribed ellipsoid in a polytope defined by linear inequalities can be efficiently approximated (see, e.g., [16]).

In summary, the extent to which Algorithm 4.1 can be applied to other input sets is largely determined by whether Subroutine 2 can be implemented efficiently. Since quadratic optimization over various feasible regions is an active area of research [28, 36], further progress in establishing polynomial complexity may widen the domain of input sets to which Algorithm 4.1 can be applied.

7. Concluding remarks. In this paper, we established that the first-order algorithm of Kumar and Yıldırım [18] that computes an approximate MVCE of a finite set of points can be extended to compute the MVCE of the union of finitely many full-dimensional ellipsoids without compromising the polynomial-time complexity for a fixed approximation parameter $\epsilon > 0$. Moreover, the iteration complexity of our extension and the core set size remain asymptotically identical. In addition, we establish that our algorithm can also compute an appproximate n-rounding of the convex hull of a finite number of ellipsoids. We discuss how the framework of our algorithm can be extended to compute an approximate MVCE and an approximate n-rounding of other input sets in polynomial time and present certain limitations. Our core set result is an addition to the recent sequence of works on core sets for several geometric optimization problems [17, 8, 7, 9, 1, 18].

While our algorithm has a polynomial-time complexity in theory, it would be especially well suited for instances of the MVCE problem with $m \gg n$ and moderately small values of ϵ. In particular, our algorithm would be applicable to the problem of constructing a bounding volume hierarchy as the objects lie in three-dimensional space (i.e., $n = 3$) and a fixed parameter ϵ usually suffices for practical applications. To the best of our knowledge, this is the first result in the literature towards approximating the convex hull of a union of ellipsoids by that of a finite subset whose size depends on only the dimension n and the parameter ϵ.

On the other hand, our algorithm would probably not be practical if a higher accuracy (i.e., a smaller ϵ) were required or if the dimension n were large. In addition, it is well known that first-order algorithms in general suffer from slow convergence in practice, especially for smaller values of ϵ. Our preliminary computational results indicate that both of the first-order algorithms of [15, 18] for an input set of points tend to take an excessive number of iterations as ϵ is decreased, which suggests that the practical performance of these algorithms is indeed closely related to the worst-case theoretical complexity bounds. Motivated by the core set result established in this paper and the encouraging computational results based on a similar core set

result for the minimum enclosing ball problem [17], we intend to work on a column generation algorithm for the MVCE problem with an emphasis on establishing an upper bound on the number of subproblems solved to obtain a desired accuracy.

Very recently, Todd and Yıldırım [30] proposed a modification of the algorithm of Kumar and Yıldırım [18] that computes an approximate MVCE and an approximate n-rounding of a finite set of points. Their modification allows "dropping" points from a working core set throughout the algorithm and maintains the same complexity bound as that of the algorithm of [18]. As such, it has the potential of computing smaller core sets in practice. We remark that the same idea can easily be incorporated into our algorithm for a set of ellipsoids without any increase in the asymptotic complexity bound.

Acknowledgments. I am grateful to Piyush Kumar for several inspiring discussions that led to the fruition of this work and for bringing to my attention the initial volume approximation result, which provides the backbone of our algorithm. I would like to thank Mike Todd and two anonymous referees for their careful, helpful, and perceptive comments and suggestions, which led to a significant improvement in the exposition. In particular, I would like to acknowledge an anonymous referee who pointed out a flaw in an earlier draft, suggested several improvements in section 4, and made comments that prompted the addition of section 5.

REFERENCES

[1] P. K. Agarwal, R. Poreddy, K. R. Varadarajan, and H. Yu, *Practical methods for shape fitting and kinetic data structures using core sets*, in Proceedings of the 20th Annual ACM Symposium on Computational Geometry, 2004.

[2] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*, MPS/SIAM Ser. Optim. 2, SIAM, Philadelphia, 2001.

[3] U. Betke and M. Henk, *Approximating the volume of convex bodies*, Discrete Comput. Geom., 10 (1993), pp. 15–21.

[4] L. Blum, M. Shub, and S. Smale, *On a theory of computation and complexity over the real numbers: NP-completeness, recursive functions, and universal machines*, Bull. Amer. Math. Soc. (N.S.), 21 (1989), pp. 1–46.

[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] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004.

[7] M. Bădoiu and K. L. Clarkson, *Smaller core-sets for balls*, in Proceedings of the 14th ACM-SIAM Symposium on Discrete Algorithms, 2003, Baltimore, MD, pp. 801–802.

[8] M. Bădoiu, S. Har-Peled, and P. Indyk, *Approximate clustering via core-sets*, in Proceedings of the 34th Annual ACM Symposium on Theory of Computing, 2002, pp. 250–257.

[9] T. M. Chan, *Faster core-set constructions and data-stream algorithms in fixed dimensions*, Comput. Geom., 35 (2006), pp. 20–35.

[10] Y. Chien and J. Liu, *Improvements to ellipsoidal fit based collision detection*, Tech. report TR-IIS-03-001, Academia Sinica, Institute of Information Science, Taiwan, Taipei, Republic of China, 2003.

[11] J. H. Clark, *Hierarchical geometric models for visible surface algorithms*, Commun. ACM, 19 (1976), pp. 547–554.

[12] M. Grötschel, L. Lovász, and A. Schrijver, *Geometric Algorithms and Combinatorial Optimization*, Springer, New York, 1988.

[13] F. John, *Extremum problems with inequalities as subsidiary conditions*, in Studies and Essays Presented to R. Courant on his 60th Birthday, January 8, 1948, Interscience, New York, 1948, pp. 187–204; reprinted in *Fritz John, Collected Papers*, Vol. 2, J. Moser, ed., Birkhäuser Boston, Boston, 1985, pp. 543–560.

[14] M. Ju, J. Liu, S. Shiang, Y. Chien, K. Hwang, and W. Lee, *Fast and accurate collision detection based on enclosed ellipsoids*, Robotica, 19 (2001), pp. 381–394.

[15] L. G. Khachiyan, *Rounding of polytopes in the real number model of computation*, Math. Oper. Res., 21 (1996), pp. 307–320.
[16] L. G. Khachiyan and M. J. Todd, *On the complexity of approximating the maximal inscribed ellipsoid for a polytope*, Math. Programming, 61 (1993), pp. 137–159.
[17] P. Kumar, J. S. B. Mitchell, and E. A. Yildirim, *Approximate minimum enclosing balls in high dimensions using core-sets*, ACM J. Exp. Algorithmics, 8 (2003), 29 pp.
[18] P. Kumar and E. A. Yildirim, *Minimum volume enclosing ellipsoids and core sets*, J. Optim. Theory Appl., 126 (2005), pp. 1–21.
[19] Yu. Nesterov and A. Nemirovskii, *Interior Point Polynomial Algorithms in Convex Programming*, SIAM Stud. Appl. Math. 13, SIAM, Philadelphia, 1994.
[20] I. Pólik and T. Terlaky, *A comprehensive study of the $\mathcal{S}$-lemma*, Tech. report, McMaster University, Advanced Optimization Laboratory, Hamilton, ON, Canada, 2004.
[21] L. Porkolab and L. Khachiyan, *On the complexity of semidefinite programs*, J. Global Optim., 10 (1997), pp. 351–365.
[22] M. V. Ramana, *An exact duality theory for semidefinite programming and its complexity implications*, Math. Programming, 77 (1997), pp. 129–162.
[23] F. Rendl and H. Wolkowicz, *A semidefinite framework for trust region subproblems with applications to large scale minimization*, Math. Programming, 77 (1997), pp. 273–299.
[24] J. Renegar, *On the computational complexity and geometry of the first-order theory of the reals, Part* I: *Introduction; preliminaries; the geometry of the semi-algebraic sets; the decision problem for the existential theory of the reals*, J. Symbolic Comput., 13 (1992), pp. 255–299.
[25] E. Rimon and S. Boyd, *Obstacle collision detection using best ellipsoid fit*, J. Intell. Robot. Syst., 18 (1997), pp. 105–126.
[26] S. Shiang, J. Liu, and Y. Chien, *Estimate of minimum distance between convex polyhedra based on enclosed ellipsoids*, in Proceedings of the 2000 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2000, pp. 739–744.
[27] R. J. Stern and H. Wolkowicz, *Indefinite trust region subproblems and nonsymmetric eigenvalue perturbations*, SIAM J. Optim., 5 (1995), pp. 286–313.
[28] J. F. Sturm and S. Z. Zhang, *On cones of nonnegative quadratic functions*, Math. Oper. Res., 28 (2003), pp. 246–267.
[29] P. Sun and R. M. Freund, *Computation of minimum volume covering ellipsoids*, Oper. Res., 52 (2004), pp. 690–706.
[30] M. J. Todd and E. A. Yildirim, *On Khachiyan's algorithm for the computation of minimum volume enclosing ellipsoids*, Tech. report TR 1435, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY, 2005.
[31] K. C. Toh, *Primal-dual path-following algorithms for determinant maximization problems with linear matrix inequalities*, Comput. Optim. Appl., 14 (1999), pp. 309–330.
[32] 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.
[33] V. A. Yakubovich, *$\mathcal{S}$-procedure in nonlinear control theory*, Vestn. Leningr. Univ., 4 (1977), pp. 73–93 (in English).
[34] Y. Ye, *On affine scaling algorithms for nonconvex quadratic programming*, Math. Programming, 56 (1992), pp. 285–300.
[35] Y. Ye, *On the finite convergence of interior-point algorithms for linear programming*, Math. Programming, 57 (1992), pp. 325–335.
[36] Y. Ye and S. Zhang, *New results on quadratic minimization*, SIAM J. Optim., 14 (2003), pp. 245–267.

SIAM J. OPTIM.
Vol. 17, No. 3, pp. 642–664
© 2006 Society for Industrial and Applied Mathematics

FINDING OPTIMAL ALGORITHMIC PARAMETERS USING DERIVATIVE-FREE OPTIMIZATION*

CHARLES AUDET† AND DOMINIQUE ORBAN†

Abstract. The objectives of this paper are twofold. We devise a general framework for identifying locally optimal algorithmic parameters. Algorithmic parameters are treated as decision variables in a problem for which no derivative knowledge or existence is assumed. A derivative-free method for optimization seeks to minimize some measure of performance of the algorithm being fine-tuned. This measure is treated as a black-box and may be chosen by the user. Examples are given in the text. The second objective is to illustrate this framework by specializing it to the identification of locally optimal trust-region parameters in unconstrained optimization. The derivative-free method chosen to guide the process is the mesh adaptive direct search, a generalization of pattern search methods. We illustrate the flexibility of the latter and in particular make provision for surrogate objectives. Locally, optimal parameters with respect to overall computational time on a set of test problems are identified. Each function call may take several hours and may not always return a predictable result. A tailored surrogate function is used to guide the search towards a local solution. The parameters thus identified differ from traditionally used values, and allow one to solve a problem that remained otherwise unsolved in a reasonable time using traditional values.

Key words. derivative-free optimization, black-box optimization, parameter estimation, surrogate functions, mesh adaptive direct search, trust-region methods

AMS subject classifications. 90C56, 90C90, 90C31

DOI. 10.1137/040620886

1. Introduction.

Most algorithms, be it in optimization or any other field, depend more or less critically on a number of parameters. Some parameters may be real numbers, such as an initial trust-region radius, or a scalar that dictates the precision at which a subproblem needs to be solved. These parameters may be required to remain between two, possibly infinite bounds, or may be constrained in a more complex way. They may also be discrete, such as the maximal number of iterations or the number of banned directions in a taboo search heuristic, or even categorical, such as a Boolean indicating whether exact or inexact Hessian is used, or which preconditioner to use. The overall behavior of the algorithm is influenced by the values of these parameters. Unfortunately, for most practical cases, it remains unclear how a user should proceed to determine *good*, let alone optimal, values for those parameters. We devise a framework for fine-tuning such parameters which is general enough to encompass most numerical algorithms from engineering, numerical analysis and optimization. The design of the framework relies on the observation that measures of *performance* can be derived from the dependency of the algorithm on its parameters. These measures are context- and problem-dependent and for this reason, we wish to

*Received by the editors December 15, 2004; accepted for publication (in revised form) March 20, 2006; published electronically September 15, 2006.

http://www.siam.org/journals/siopt/17-3/62088.html

†GERAD and Département de Mathématiques et de Génie Industriel, École Polytechnique de Montréal, C.P. 6079, Succ. Centre-ville, Montréal, Québec, H3C 3A7 Canada (Charles.Audet@gerad.ca, http://www.gerad.ca/Charles.Audet; Dominique.Orban@gerad.ca, http://www.mgi.polymtl.ca/Dominique.Orban). The work of the first author was supported by NSERC grant 239436-01, AFOSR F49620-01-1-0013, and ExxonMobil R62291. The work of the second author was supported by NSERC grant 299010-04.

treat them as black-boxes in the remainder of this paper. We shall, however, give examples in the context of a particular application.

An optimization problem is formulated where such a measure of performance is minimized as a function of the parameters, over a domain of acceptable values. We use a class of nonsmooth optimization algorithms to solve this problem which makes provision for an optional *surrogate* function to guide the search strategy. A surrogate function is a simplification of the real objective function that possesses similar behavior, but is believed to be less costly to evaluate, or easier to manipulate, in some sense. Surrogate functions range from simplified physical models to approximation surfaces, such as Kriging models. The reader is invited to consult [6] for a general framework for the use of surrogates in an optimization context.

As an illustration of how to use this framework, we address the study of standard parameters present in trust-region algorithms for unconstrained optimization and try to identify some locally optimal values. The quality of a set of parameters is measured by the overall computational time required by a trust-region algorithm to solve a significant number of test problems to a prescribed precision. Other possibilities of performance measures include the overall number of function calls, the number of failures, the total number of iterations or the agreement of the algorithm with some prescribed behavior. In the numerical tests presented, the test problems originate from the CUTEr [19] collection. We formulate an optimization problem, where each evaluation of the objective function requires solving a collection of test problems. This objective function is time-consuming to evaluate, is highly nonlinear and no derivative is available or even proved to exist. Moreover, evaluating the objective twice with the same arguments may lead to slightly different function values since the computational time is influenced by the current machine load and fluctuates with network activity. We opted for the mesh adaptive direct search (MADS) [4] as derivative-free method for reasons which we explain below. In our context, a surrogate function is obtained by applying the same trust-region algorithm to a set of easier problems. The trust-region parameters obtained by MADS allow the solution of a problem which remained otherwise unsolved in reasonable time by the trust-region method.

Related work on locally optimal parameter identification in a similar trust-region framework is presented in [17], where the parameter space is discretized and a thorough search is carried out. However, even for a modest number of discretized values, devising a mechanism able to compare so many variants of an algorithm becomes an issue. In recent years, performance profiles [13] have been extensively used to compare algorithms, but it remains unclear how to use them to efficiently compare more than five or six. We circumvent this issue in the present paper by delegating the task to another optimization algorithm.

In recent years, pattern-search methods have proved to perform decently on molecular geometry and conformation problems [1, 28]. The paper [23] provides a review of direct search methods.

The paper is structured as follows. In section 2, we describe a derivative-free framework, then outline a specific implementation of the MADS class of algorithms, and highlight the main convergence results. We also describe how a surrogate function can be used within this algorithm. Section 3 describes a standard trust-region algorithm, and discusses the four algorithmic parameters to be fine-tuned. In section 4, we present a methodology specializing our framework to identify locally optimal trust-region parameters using a MADS algorithm. Results are presented in secton 5, and we give concluding remarks in section 7.

2. A framework for nondifferentiable problems. A general optimization problem may be stated as

$$\min_{p\in\Omega} \quad \psi(p), \tag{2.1}$$

with $\psi : \Omega \subseteq \mathbb{R}^\ell \to \mathbb{R} \cup \{+\infty\}$. The nature and structure of the function ψ and the domain Ω limit the type of algorithms that may be used to attempt to solve this problem. Global optimization is often only possible when the problem structure is sufficiently rich and exploitable, and when the problem size is reasonable, but frequently out of reach in acceptable time. Under appropriate smoothness assumptions, we are thus often content with first-order critical solutions. For example, when ψ is continuously differentiable over Ω, an appropriate variant of Newton's method combined with a globalizing scheme yields a critical point under reasonable assumptions. When ψ is nondifferentiable, discontinuous or fails to evaluate for some values of its argument $p \in \Omega$, problem (2.1) cannot be satisfactorily approached by such a method. This is often the case when evaluating the objective entices running a computer code. In order to evaluate, the code may, for instance, have to solve a coupled system of differential equations, and may, for some internal reasons, fail to return a meaningful value. In this case, the function value is simply considered to be infinite. In a helicopter rotor blade design application [5], the objective function failed to return a value two times out of three. Randomness may also be present in the evaluation of a function, as in [29], where two evaluations of ψ at the same point p return slightly different values. In this less optimistic case, the best optimality condition which can be hoped for is to find a *refined point*. We shed light on this concept in section 2.3.

2.1. A general overview of pattern search type methods with surrogate. In the special case where Ω is defined by finitely many linear inequalities, algorithms from the broad class of generalized pattern search (GPS) methods [3] are natural candidates to perform the minimization. They have the advantage of being relatively simple and easy to implement.

Algorithms of the pattern search type attempt to locate a minimizer of the function ψ over Ω by means of the *barrier* function

$$\psi_\Omega(p) = \begin{cases} +\infty & \text{if } p \notin \Omega \\ \psi(p) & \text{otherwise.} \end{cases} \tag{2.2}$$

We will refer to ψ as being the *truth function* or, sometimes, simply the truth.

As is frequent in nonlinear programming, a second function, $\sigma : \Omega \to \mathbb{R} \cup \{+\infty\}$, playing the role of a *model* for ψ, may be used to steer the iterates towards promising regions. In the context of nondifferentiable optimization and pattern search methods, a model is often referred to as a *surrogate*. The surrogate may be an approximation to the truth, or it may be a simplified function whose behavior is similar to that of the truth. An important feature of the surrogate is that it should be *cheaper* than the truth, in some sense—less costly in terms of time or other. The previous sentences are left intentionally vague, since the formal convergence analysis is independent of the quality of the approximation of ψ by σ. However, in practice, appropriate surrogates may improve the convergence speed. A *barrier* surrogate σ_Ω is defined similarly to (2.2).

Pattern search type methods are iterative procedures where each iteration essentially consists of two steps. First, a global exploration of the space of variables

is conducted in hopes of improving the best feasible solution so far, or *incumbent*, $p_k \in \Omega$. This flexible exploration stage, called the SEARCH step, returns a set of candidates, but the truth function need not be evaluated at all of them. To decide at which of these the truth ψ_Ω will be evaluated, they are ordered in increasing surrogate function values. After ordering, the set of candidates $\mathcal{L} = \{q^1, q^2, \ldots, q^m\}$ satisfies $\sigma_\Omega(q^1) \leq \sigma_\Omega(q^2) \leq \cdots \leq \sigma_\Omega(q^m)$. A candidate $q^j \in \mathcal{L}$ will be considered promising if $\sigma_\Omega(q^j) \leq \sigma_\Omega(p_k) + v|\sigma_\Omega(p_k)|$, where $v \in \mathbb{R}_+ \cup \{+\infty\}$ is a threshold supplied by the user. Candidates which are not promising are eliminated from the SEARCH list.

The truth function is then evaluated at the promising candidates in $\mathcal{L}$ with increasing values of i, and terminating the process as soon as $\psi_\Omega(q^i) < \psi_\Omega(p_k)$. In this case, an improved incumbent is found and we set $p_{k+1} = q^i$.

In the event where the SEARCH fails to identify an improved iterate, a local exploration about p_k is performed. This is called the POLL step. Again, the surrogate function is used to order the trial points. The convergence analysis relies only on this step and it must obey stricter rules. In particular, it prohibits the pruning of candidates.

A inconvenience about GPS algorithm is that they require some prior knowledge on the structure of Ω to decide on a set of appropriate search directions. Lack of such knowledge may prevent progress towards a minimizer [26]. So as to make a provision for more general Ω and bypass such a prior knowledge requirement, we opted for an extension of GPS named the mesh adaptive direct search (MADS) [4], which is also based on the above principles. MADS allows a more general exploration of the space of variables ensuring stronger convergence results than GPS. Indeed, GPS confines the POLL at each iteration to a fixed finite set of directions, while the set of directions for the MADS POLL may vary at each iteration, and in the limit the union of these poll directions over all iterations is dense in the whole space.

2.2. An iteration of a MADS algorithm. We now present a lower-level description of MADS. The reader is invited to refer to [4] for a complete algorithmic description, a detailed convergence analysis and numerical comparisons with GPS. The version presented here is specialized to our purposes and some algorithmic choices were made.

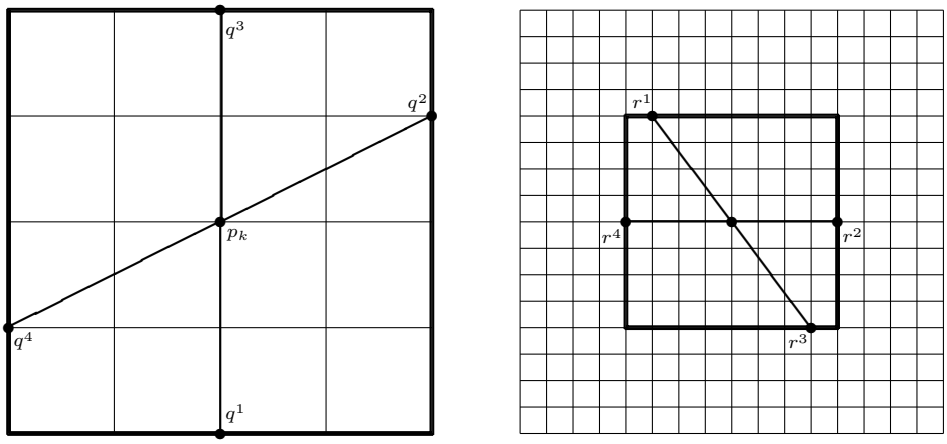
Let $S_0 \subset \Omega$ denote a finite set of initial guesses, provided by the user (a strategy exploiting the surrogate to determine S_0 is presented in section 4.4). Set p_0 to be the best initial guess in S_0. A MADS algorithm is constructed in such a way that any trial point generated by the SEARCH or POLL step is required to lie on the *current mesh*, the coarseness of which is governed by a *mesh size parameter* $\Delta_k \in \mathbb{R}_+$. The mesh is formally defined in Definition 2.1.

DEFINITION 2.1. *At iteration k, the current mesh is defined to be the union*

$$M_k = \bigcup_{p \in S_k} \left\{ p + \Delta_k z : z \in \mathbb{Z}^\ell \right\},$$

where S_k is the set of points where the objective function ψ has been evaluated by the start of iteration k and $\mathbb{Z}$ denotes the set of integers.

The goal of the iteration is to find a trial mesh point with a lower objective function value than the current incumbent value $\psi_\Omega(p_k)$. Such a trial point is called an *improved mesh point*, and the iteration is called a *successful* iteration. There are no sufficient decrease requirements: any improvement in ψ_Ω leads to a successful iteration. The iteration is said to be *unsuccessful* if no improved point is found.



Iteration k (unsuccessful) Iteration $k+1$ with $p_{k+1} = p_k$

FIG. 1. *Two consecutive frames F_k, F_{k+1} in $\mathbb{R}^2$ with $\Delta_k = \frac{1}{2}$, $\Delta_{k+1} = \frac{1}{8}$.*

The POLL step evaluates ψ_Ω at 2ℓ trial points surrounding the current incumbent. These neighboring points are called the *frame*, and are denoted by

$$F_k = \{p_k \pm \Delta_k d : d \in D_k\}, \tag{2.3}$$

where $D_k = \{d^1, d^2, \ldots, d^\ell\}$ is a basis in $\mathbb{Z}^\ell$. To ensure convergence, the radii $(\max\{\|\Delta_k d\|_\infty : d \in D_k\})$ of successive frames must converge to zero at a slower rate than the mesh size parameter. The construction of the basis D_k proposed in [4] ensures that

$$\|\Delta_k d\|_\infty = O(\sqrt{\Delta_k}) \quad \text{for all} \quad d \in D_k. \tag{2.4}$$

Figure 1 shows an example of two consecutive frames in $\mathbb{R}^2$. The figure on the left represents iteration k. The mesh M_k is represented by the intersection of all lines. Suppose that $\Delta_k = \frac{1}{2}$. The thick lines delimit the frame, i.e., the region in which all four POLL points must lie. In this example, the frame points q_1 and q_3 are obtained by the randomly generated direction $d_1 = (0, -2)$, and q_2 and q_4 are obtained by $d_2 = (2, 1)$. The figure on the right displays a possible frame if iteration k is unsuccessful. The mesh is finer at iteration $k+1$ than it was at iteration k, and there are more possibilities in choosing a frame. More precisely, $\Delta_{k+1} = \frac{\Delta_k}{4} = \frac{1}{8}$ and, as in (2.4), the distance from the boundary of the frame to the incumbent is reduced by a factor of 2; $\|r^i - p_{k+1}\|_\infty = \sqrt{\frac{1}{4}}\|q^j - p_k\|_\infty$ for all i, j. The direction used to construct the frame points r_1 and r_3 are $d_1 = (-3, 4)$, and the direction for r_2 and r_4 are $d_2 = (4, 0)$. In the event that iteration $k+1$ is successful at the mesh POLL point r_3, iteration $k+2$ would be initiated at the new incumbent $p_{k+2} = r_3$ with a larger mesh size parameter $\Delta_{k+2} = 4\Delta_{k+1} = \frac{1}{2}$.

When the POLL step fails to generate an improved mesh point then the frame is called a *minimal frame*, and the frame center p_k is said to be a *minimal frame center*. At iteration k, the rule for updating the mesh size parameter is

$$\Delta_{k+1} = \begin{cases} \Delta_k/4 & \text{if } p_k \text{ is a minimal frame center,} \\ 4\Delta_k & \text{if an improved mesh point is found, and } \Delta_k \le \frac{1}{4}, \\ \Delta_k & \text{otherwise.} \end{cases} \tag{2.5}$$

ALGORITHM 2.1. [A MADS ALGORITHM]

Step 0 [Initialization] Let S_0 be given, $p_0 \in \arg\min\{\psi(p) : p \in S_0\}$, $\Delta_0 > 0$, and $v \in \mathbb{R}_+ \cup \{+\infty\}$. Set the iteration counter $k = 0$, and go to Step 1.

Step 1 [SEARCH step] Let $\mathcal{L} = \{q^1, q^2, \ldots, q^m\} \subset M_k$ be a finite (possibly empty) set of mesh points such that $\sigma_\Omega(q^i) \le \sigma_\Omega(q^j) \le \sigma_\Omega(p_k) + v|\sigma_\Omega(p_k)|$ when $1 \le i < j \le m$.
Let i_0 be the smallest $i \in \{1, \ldots, m\}$ such that $\psi_\Omega(q^i) < \psi_\Omega(p_k)$.
If no such index i_0 exists, go to Step 2.
Otherwise, declare k successful, set $p_{k+1} = q^{i_0}$ and go to Step 3.

Step 2 [POLL step] Construct the frame $F_k = \{q^1, q^2, \ldots, q^{2\ell}\}$ as in (2.3) and order the points so that $\sigma_\Omega(q^i) \le \sigma_\Omega(q^j)$ when $1 \le i < j \le 2\ell$.
Let i_0 be the smallest $i \in \{1, \ldots, 2\ell\}$ such that $\psi_\Omega(q^i) < \psi_\Omega(p_k)$.
If no such index i_0 exists, declare k unsuccessful and go to Step 3.
Otherwise, declare k successful, set $p_{k+1} = q^{i_0}$ and go to Step 3.

Step 3 [Parameter update] If iteration k was declared unsuccessful, then p_k is a minimal frame center and p_{k+1} is set to p_k. Otherwise p_{k+1} is an improved mesh point.
Update Δ_{k+1} according to (2.5). Increase $k \leftarrow k + 1$ and go back to Step 1.

The previous description is summarized in Algorithm 2.1.

Given the functions ψ and σ, the only steps that are not completely defined in Algorithm 2.1 are the selection of the set of initial guesses S_0 and the SEARCH strategy. Particular choices in the framework of an application are discussed in section 4.4.

2.3. Convergence properties of MADS. We restrict ourselves to the case where Ω is convex and full dimensional, i.e., with nonempty interior. This encompasses the case where Ω is defined by (possibly strict) linear inequalities. It seems reasonable to assume that in practice, most algorithms depend on parameters satisfying this assumption. Moreover, the MADS convergence analysis greatly simplifies. This section presents the specialized results. The proofs of these results are special cases of the proofs in [4].

Let $\mathrm{cl}(A)$ denote the closure of a set A. Under our conditions, the tangent cone to Ω at some $\hat{p} \in \mathrm{cl}(\Omega)$, denoted by $T_\Omega(\hat{p})$, becomes the closure of the set $\{\mu(v - \hat{p}) : \mu \in \mathbb{R}_+,\ v \in \Omega\}$. The normal cone to Ω at $\hat{p} \in \mathrm{cl}(\Omega)$ is the polar of the tangent cone $N_\Omega(\hat{p}) = T_\Omega(\hat{p})^\circ = \{w \in \mathbb{R}^\ell \mid w^T v \le 0\ \forall v \in T_\Omega(\hat{p})\}$. If the gradient $\nabla\psi(\hat{p})$ exists at $\hat{p} \in \mathrm{cl}(\Omega)$, we say that $\hat{p}$ is *first-order critical* for (2.1) if $-\nabla\psi(\hat{p}) \in N_\Omega(\hat{p})$. For nonconvex domains, more general definitions of tangency from nonsmooth analysis are used in [4].

The analysis relies on Assumption 2.1.

ASSUMPTION 2.1. *At least one initial guess $p_0 \in S_0 \subseteq \Omega$ has finite $\psi(p_0)$ value and all iterates $\{p_k\}$ produced by Algorithm* 2.1 *lie in a compact set.*

The mechanism of Algorithm 2.1 ensures the following property.

LEMMA 2.2. *The sequence of mesh size parameters satisfies* $\liminf_{k\to\infty} \Delta_k = 0$. *Moreover, since Δ_k shrinks only at minimal frames, it follows that there are infinitely many minimal frame centers.*

Definition 2.3 specifies important subsequences of iterates and limiting directions.

DEFINITION 2.3. *A subsequence of the* MADS *iterates consisting of minimal frame centers, $\{p_k\}_{k\in K}$ for some subset of indices K, is said to be a* refining subsequence

if $\{\Delta_k\}_{k\in K}$ *converges to zero. Any accumulation point of* $\{p_k\}_{k\in K}$ *will be called a refined point.*

Let $\{p_k\}_{k\in K}$ *be a convergent refining subsequence, with refined point* $\hat{p}$*, and let* v *be any accumulation point of the set* $\{\frac{d_k}{\|d_k\|} : p_k + \Delta_k d_k \in \Omega,\ k \in K\} \subset \mathbb{R}^\ell$*. Then* v *is said to be a refining direction for* $\hat{p}$*.*

Along a refining subsequence, we thus always have $\psi(p_{k+1}) \leq \psi(p_k)$. Note that under Assumption 2.1, there always exists at least one convergent refining subsequence, one refining point and a positive spanning set of refining directions.

Recall that the Clarke generalized directional derivative [7, 22] of ψ at $\hat{p}$ in the direction $v \neq 0$ is defined as

$$\psi^\circ(\hat{p}; v) \equiv \limsup_{\substack{y \to \hat{p},\ y \in \Omega, \\ t \downarrow 0,\ y + tv \in \Omega}} \frac{\psi(y + tv) - \psi(y)}{t}.$$

The following result presents a hierarchy of optimality conditions satisfied at a refining point. The weakest conclusion is a direct consequence of Definition 2.3. The next conclusions result from increasingly strong assumptions on ψ. The main result is that the Clarke derivatives of ψ at a refined point $\hat{p}$ are nonnegative for all directions in the tangent cone.

THEOREM 2.4. *Let* $\hat{p} \in cl(\Omega)$ *be a refined point of a refining subsequence* $\{p_k\}_{k\in K}$*, and assume that the set of refining directions for* $\hat{p}$ *is dense in* $T_\Omega(\hat{p})$*.*

- $\hat{p}$ *is the limit of minimal frame center on frames that become infinitely fine,*
- *if* ψ *is lower semicontinuous at* $\hat{p}$*, then* $\psi(\hat{p}) \leq \lim_{k\in K} \psi(p_k)$*,*
- *if* ψ *is Lipschitz near* $\hat{p}$*, then* $\psi^\circ(\hat{p}, v) \geq 0$ *for every* $v \in T_\Omega(\hat{p})$*,*
- *if* ψ *is Lipschitz near* $\hat{p} \in int(\Omega)$*, then* $0 \in \partial\psi(\hat{p}) \equiv \{s \in \mathbb{R}^\ell : \psi^\circ(\hat{x}; v) \geq v^T s,\ \forall v \in \mathbb{R}^\ell\}$*,*
- *if* ψ *is strictly differentiable* [24] *at* $\hat{p} \in \Omega$*, then* $\hat{p}$ *is first-order critical for* (2.1) *and is a KKT point of* (2.1)*.*

Note that the above convergence results rely only on the POLL step, and are independent of the surrogate function and of the SEARCH step. Furthermore, even though Algorithm 2.1 is applied to ψ_Ω instead of ψ, the convergence results are linked to the local smoothness of ψ and not ψ_Ω, which is obviously discontinuous on the boundary of Ω.

Practical implementations of MADS ensure that the density assumption of Theorem 2.4 is satisfied [4].

3. Trust-region methods. In this section, we briefly cover a globally convergent framework for unconstrained programming. This framework depends on a number of parameters which influence the impact of the main convergence result.

To successfully tackle a smooth nonlinear nonconvex programming problem from a remote starting guess, the iteration must often be embedded into a *globalization* scheme, the most popular of which are the linesearch, the trust region [10], and the more recent filters [14, 20]. The first two are probably the most well known and their philosophies may be seen as dual; a linesearch strategy computes a step length along a predetermined direction, while a trust-region strategy considers all acceptable directions but limits the maximal step length.

3.1. A basic trust-region algorithm. Trust-region methods appear to date back to a 1944 paper in which they were used to solve nonlinear least-squares problems [25]. After updating rules were introduced in 1966 [15] for the size of the region,

global convergence of a particular algorithm was proved in 1970 [30]. Trust-region methods now form one of the most popular globalization schemes and are often praised for their robustness and flexibility. They are used throughout optimization, from regularization problems to derivative-free and interior-point methods. For lists of references, historical notes, and thorough theoretical developments across the whole optimization spectrum we refer the interested reader to [10].

For simplicity, assume we wish to solve the unconstrained programming problem

$$\min_{x \in \mathbb{R}^n} \quad f(x), \tag{3.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a twice-continuously differentiable function which might be highly nonlinear and/or costly to evaluate. For problem (3.1) to be well defined, we will assume that f is bounded below. At iteration k, instead of manipulating f directly, f is replaced by a suitable local model m_k which is easier and cheaper to evaluate. A region $\mathcal{B}_k \subset \mathbb{R}^n$, referred to as the *trust region*, is defined around x_k to represent the extent to which m_k is believed to reasonably model f. The trust region is defined as the ball

$$\mathcal{B}_k \equiv \{x_k + s \in \mathbb{R}^n \ : \ \|s\| \leq \delta_k\},$$

where $\delta_k > 0$ is the current trust-region radius and $\|\cdot\|$ represents any norm on $\mathbb{R}^n$. To simplify the exposition, we choose the Euclidean norm but other choices are acceptable [10]. The model m_k is approximately minimized within $\mathcal{B}_k$. If the decrease thus achieved is sufficient, and if the agreement between f and m_k at the trial point is satisfactory, the step is accepted and the radius δ_k is possibly increased. Otherwise, the step is rejected and the radius is decreased. This last option indicates that m_k might have been trusted in too large a neighborhood of x_k.

Global convergence of trust-region schemes is ensured by mild assumptions on m_k and on the decrease that should be achieved at each iteration. In practice, one of the most popular models is the quadratic model

$$m_k(x_k + s) = f(x_k) + \nabla f(x_k)^T s + \frac{1}{2} s^T \nabla^2 f(x_k) s. \tag{3.2}$$

Sufficient decrease in the model at a trial point $x_k + s$ is achieved if the decrease in m_k is at least a fraction of that obtained at the *Cauchy point* x_k^{C}—the minimizer of m_k along the steepest descent direction $d = -\nabla m_k(x_k)$ within $\mathcal{B}_k$—i.e.,

$$m_k(x_k) - m_k(x_k + s) \geq \theta \left[m_k(x_k) - m_k(x_k^{\mathrm{C}}) \right], \tag{3.3}$$

where $0 < \theta < 1$ is independent of k.

A typical trust-region framework for problem (3.1) may be stated as Algorithm 3.1.

The updating rule (3.4) at Step 3 of Algorithm 3.1 is not the only one used in practice, but most likely the most common one. Other rules involve polynomial interpolation of $\rho_k = \rho(s_k)$ as a function of the step s_k [12], while others devise more sophisticated functions to obtain the new radius [21, 38].

The lengthy subject of how to solve the subproblems at Step 1 of Algorithm 3.1 while ensuring (3.3) is out of the scope of this paper. We shall simply argue in section 4 that the method used in our implementation ensures this.

3.2. Convergence properties of the basic algorithm. We recall in this section the important global convergence properties of Algorithm 3.1 without proof. The proofs may be found in [10].

ALGORITHM 3.1. [BASIC TRUST-REGION ALGORITHM]

Step 0 [Initialization] An initial point $x_0 \in \mathbb{R}^n$ and an initial trust-region radius $\delta_0 > 0$ are given, as well as parameters

$$0 \leq \eta_1 < \eta_2 < 1 \quad \text{and} \quad 0 < \alpha_1 < 1 < \alpha_2.$$

Compute $f(x_0)$ and set $k = 0$.

Step 1 [Step calculation] Define the model (3.2) of $f(x_k + s)$ in $\mathcal{B}_k$ and compute a step $s_k \in \mathcal{B}_k$ which satisfies (3.3).

Step 2 [Acceptance of the trial point] Compute $f(x_k + s_k)$ and

$$\rho_k = \frac{f(x_k) - f(x_k + s_k)}{m_k(x_k) - m_k(x_k + s_k)}.$$

If $\eta_1 \leq \rho_k$, then set $x_{k+1} = x_k + s_k$; otherwise, set $x_{k+1} = x_k$.

Step 3 [Trust-region radius update] Set

$$\delta_{k+1} = \begin{cases} \alpha_1 \|s_k\| & \text{if} \quad \rho_k < \eta_1 \\ \delta_k & \text{if} \quad \eta_1 \leq \rho_k < \eta_2 \\ \max[\alpha_2 \|s_k\|, \delta_k] & \text{if} \quad \eta_2 \leq \rho_k. \end{cases} \tag{3.4}$$

Increment k by one, and go to Step 1.

Step 2 of Algorithm 3.1 is often referred to as the computation of achieved versus predicted reduction. Achieved reduction is the actual reduction in the objective f, defined by $\text{ared}_k = f(x_k) - f(x_k + s_k)$. Predicted reduction is the reduction suggested by the model, $\text{pred}_k = m_k(x_k) - m_k(x_k + s_k)$. A step s_k is thus accepted whenever $\text{ared}_k \geq \eta_1 \text{pred}_k$, an iteration we refer to as *successful*.

Requirements on the function f and each model m_k are gathered in Assumption 3.1.

ASSUMPTION 3.1. *The function f is bounded below and its Hessian matrix remains bounded over a set containing all iterates x_k.*

The first stage in the global convergence analysis of Algorithm 3.1 is usually summarized by Theorem 3.1.

THEOREM 3.1. *Suppose that Assumption* 3.1 *is satisfied. Then*

$$\liminf_{k \to \infty} \|\nabla f(x_k)\| = 0.$$

Theorem 3.1 was first proved in [30] in a framework where $\eta_1 = 0$, i.e., where all trial points are accepted as soon as they produce a decrease in the objective. This result proves that if $\{x_k\}$ has limit points, at least one of them is critical. In fact, this is as good a convergence result as we can obtain when $\eta_1 = 0$ [39]. Algorithm 3.1 is more demanding on the trial point—*sufficient* reduction must be achieved. This sheds some light on the importance of the value of η_1 in the framework, for as the next result shows, a much stronger conclusion holds in this case.

THEOREM 3.2. *Suppose Assumption* 3.1 *is satisfied, and that $\eta_1 > 0$. Then*

$$\lim_{k \to \infty} \|\nabla f(x_k)\| = 0.$$

In other words, Theorem 3.2 shows that *all* limit points are first-order critical for (3.1). The distinction between Theorem 3.1 and Theorem 3.2 was reinforced by the

careful example of [39], where it is shown that an algorithm with $\eta_1 = 0$ may very well produce limit points which are not critical.

The importance of the parameters η_1 and η_2, but also α_1 and α_2 of Algorithm 3.1 will be of interest to us in the remainder of this paper. In particular, we shall come back to the issue of reduction versus sufficient reduction.

4. Methodology. An objective of the paper is to address a long-standing question of identifying four optimal parameters found in a trust-region update (3.4), namely η_1, η_2, α_1, and α_2. In this section, we present a general methodology to address this issue.

4.1. A black-box approach to parameter estimation. Suppose that Algorithm $\mathcal{A}$ depends on a set of continuous parameters p restricted to lie in $\Omega \subset \mathbb{R}^\ell$, where ℓ is typically small. Let $\mathcal{P}_\mathcal{O} = \{\mathcal{P}_i \mid i \in \mathcal{O}\}$ be a set of $n_\mathcal{O} \geq 1$ problem instances believed to be representative of the class of problems for which Algorithm $\mathcal{A}$ was designed, or to be of particular interest in the context of Algorithm $\mathcal{A}$. Define a function $\psi : \Omega \to \mathbb{R}$ so that for any $p \in \Omega$, $\psi(p)$ is some measure of the performance of Algorithm $\mathcal{A}$ in solving the set of problems $\mathcal{P}_i \in \mathcal{P}_\mathcal{O}$ and such that, the smaller the value of $\psi(p)$, the better the performance of the algorithm, in a context-dependent sense.

In an optimization context, examples of a function ψ would include the total CPU time required to solve the complete test set, the number of problems unsuccessfully solved or the cumulative number of iterations or of function evaluations. In other contexts, any appropriate measure may be used.

The above description qualifies as a black-box optimization problem in the sense that a computer program must in general be run in order to evaluate $\psi(\cdot)$ at a given parameter value $p \in \Omega$. For all allowed values of the parameters p, we seek to minimize a global measure of the performance of Algorithm $\mathcal{A}$. In other words, we wish to solve problem (2.1).

Problem (2.1) is usually a small-dimensional nondifferentiable optimization problem with expensive black-box function evaluation. As an additional difficulty, evaluating the objective function of (2.1) twice at the same value of p might produce two sligthly different results.[1] It therefore seems natural to use the MADS algorithm to approach it.

In the present context, there is a natural way to define a less expensive surrogate function that would have an overall behavior similar to that of ψ. Let $\mathcal{P}_\mathcal{S} = \{\mathcal{P}_j \mid j \in \mathcal{S}\}$ be a set of $n_\mathcal{S} \geq 1$ *easy* problems and for any $p \in \Omega$, define $\sigma(p)$ to be the same measure (as with ψ) of performance of Algorithm $\mathcal{A}$ in solving the set $\mathcal{P}_\mathcal{S}$ of problems. The quality of an approximation of the behavior of ψ by the surrogate function σ depends on $n_\mathcal{S}$ and on the features of the problems in $\mathcal{P}_\mathcal{S}$. The more problems, the better the approximation, but the cost of surrogate evaluations will increase. There is therefore a trade-off between the quality and the cost of a surrogate function. It would thus make sense to include in $\mathcal{P}_\mathcal{S}$ problems that are less expensive to solve and possibly, but not necessarily, we may choose $\mathcal{S} \subseteq \mathcal{O}$.

Another interesting possibility for a surrogate would be to solve the problems in $\mathcal{P}_\mathcal{S}$ to a looser accuracy.

Note that this framework is sufficiently general to encompass algorithmic parameter estimation in almost any branch of applied mathematics or engineering.

[1] And thus technically, ψ is not a function in the mathematical sense.

4.2. An implementation of the basic trust-region Algorithm 3.1. Our implementation of the trust-region method is relatively conventional and relies on the building blocks of the GALAHAD library for optimization [18]. Trust-region subproblems are solved by means of the Generalized Lanczos method for Trust Regions GLTR [16, 18]. This method is attractive for its ability to naturally handle negative curvature and to stop at the Steihaug–Toint point, i.e., the intersection of the trust-region boundary with a direction of sufficiently negative curvature [35, 37]. It ensures satisfaction of (3.3).

We now discuss the termination criteria of applying the trust-region Algorithm 3.1 to the unconstrained problem

$$(\mathcal{P}_i) \;\equiv\; \min_{x\in\mathbb{R}^{n_i}} \quad f_i(x),$$

where $f_i : \mathbb{R}^{n_i} \to \mathbb{R}$, for $i \in \mathcal{O} \cup \mathcal{S}$. The trust-region algorithm stops as soon as an iterate x_k satisfies

$$\|\nabla f_i(x_k)\|_2 \leq 10^{-5}.$$

The algorithm is also terminated when this criterion was not met in the first 1000 iterations. We elected against scaling the problem or the stopping test, as this introduced a new parameter in the procedure. Following the Steihaug–Toint procedure, the subproblem in $s \in \mathbb{R}^{n_i}$

$$\begin{array}{ll} \min & \nabla f_i(x_k)^T s + \frac{1}{2} s^T \nabla^2 f_i(x_k) s \\ \text{s.t.} & \|s\|_2 \leq \delta_k \end{array}$$

is successfully solved as soon as

$$\|\nabla f_i(x_k + s)\|_2 \;\leq\; \min\left[\frac{1}{10}, \|\nabla f_i(x_k)\|_2^{1/2}\right] \quad \text{or} \quad \|s\|_2 = \delta_k.$$

The initial trust-region radius was chosen according to

$$\delta_0 \;=\; \max\left(\frac{1}{10}\|\nabla f_i(x_0)\|_2,\; 1\right). \tag{4.1}$$

4.3. Measures of performance. In our experiments below, a single function $\psi(\cdot)$ is considered, the evaluation of which involves running a computer program which has a given trust-region algorithm for unconstrained optimization solve a series of problems with given values of the parameters. The parameters are $p = (\eta_1, \eta_2, \alpha_1, \alpha_2)$ from Step 3 of Algorithm 3.1, $\ell = 4$, and

$$\Omega = \left\{p \in \mathbb{R}^4 \mid 0 \leq \eta_1 < \eta_2 < 1 \text{ and } 0 < \alpha_1 < 1 < \alpha_2 \leq 10\right\}. \tag{4.2}$$

A strategy to fine-tune the parameter δ_0 appears in [32]. We have chosen to exclude this parameter from Ω, as the problem-dependent value (4.1) seems more sensible both conceptually and in our tests. This also makes comparisons with the results of [17] more manageable.

Note that Assumption 2.1 is satisfied since the domain Ω is a full-dimensional bounded convex set. The upper bound on α_2 was introduced on the one hand to

satisfy Assumption 2.1, and on the other hand because it does not appear intuitive, or useful, to enlarge the trust-region by an arbitrarily large factor on very successful iterations. The domain Ω is handled by MADS through the barrier approach, i.e., if a parameter value violates any linear constraint, the resulting value of the objective is set to infinity.

To ensure that a change in Δ in Algorithm 2.1 is comparable for all four parameters, the latter are scaled using

$$\begin{bmatrix} \tilde{\eta}_1 \\ \tilde{\eta}_2 \\ \tilde{\alpha}_1 \\ \tilde{\alpha}_2 \end{bmatrix} = \begin{bmatrix} 1000 & & & \\ & 100 & & \\ & & 100 & \\ & & & 10 \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \\ \alpha_1 \\ \alpha_2 \end{bmatrix}. \tag{4.3}$$

MADS then works with the variables $(\tilde{\eta}_1, \tilde{\eta}_2, \tilde{\alpha}_1, \tilde{\alpha}_2)$.

Suppose a set of $n_{\mathcal{O}}$ unconstrained problems from the CUTEr [19] collection is chosen. An evaluation of the black-box function $\psi(\cdot)$ is defined by the solution of these problems using Algorithm 3.1 and the current parameter values $p \in \Omega$. The outcome of this evaluation is either a real number—our measure of performance—or an infinite value, resulting from a computer or algorithmic failure in one or more problems. Failures may occur because the maximum number of iterations has been exceeded, because of memory or system-dependent errors or perhaps because of a floating-point exception.

Noticeably, in the context of Algorithm 3.1, the same parameter values

$$p^{\mathrm{C}} = (\eta_1, \eta_2, \alpha_1, \alpha_2) = \left(\frac{1}{4}, \frac{3}{4}, \frac{1}{2}, 2\right) \tag{4.4}$$

are often recommended in the literature [8, 9, 11, 31, 33, 34]. We shall refer to those as the *classical* parameter values. Our contention is to show that the values (4.4) are arbitrary and that much better options are available. We use Algorithm MADS to identify them. In our tests, the initial trial point considered by MADS is p^{C}.

If evaluating the objective function is relatively cheap compared to the algorithm's internal work, which might be the case when the dimension of the problem is large and the linear algebra therefore more expensive, then the overall CPU time is an obvious choice for measuring performance. Since this turned out to be the case for several problems in our objective list, it is the measure of performance which we chose. If we denote by $\tau_i(p)$ the CPU time which was necessary to solve problem $\mathcal{P}_i$ with the parameters p, we may define

$$\psi^{\mathrm{cpu}}(p) = \sum_{i \in \mathcal{O}} \tau_i(p). \tag{4.5}$$

Obviously, similar performance functions may be defined to minimize the overall number of function evaluations—a measure justified in the frequent case where objective function evaluations are computationally costly and dominate the other internal work of the algorithm.

From the MADS point of view, each objective function evaluation requires the computation of $n_{\mathcal{O}}$ values: $\tau_i(\cdot)$ for $i \in \mathcal{O}$.

4.4. The role and choice of surrogate function σ. A surrogate function plays three important roles in the present context. First, it is used as if it were the

real objective function, for the sole purpose of obtaining a better starting point than p^{C} before restarting the procedure with the objective function defined by the objective list. Its purpose is thus to postpone the long computations until a basin containing a local minimizer is reached. No surrogate approximation of the surrogate σ was used. Starting from p^{C} given by (4.4), MADS terminates with some solution p^{S}. After having obtained these parameters, we can now apply MADS on the truth function ψ and use the surrogate function σ to guide the algorithm. The set of initial guesses was chosen as $S_0 = \{p^{\mathrm{C}}, p^{\mathrm{S}}\}$.

Second, the surrogate is used to order trial points generated by the MADS POLL or SEARCH steps, as described in section 2.1. If the surrogate is appropriate, the ordering should produce a successful iterate for the real objective function before all directions have been explored.

Finally, the third role of the surrogate is to eliminate from consideration the trial SEARCH points at which the surrogate function value exceeds the user threshold value v (introduced in section 2.1), which is in our case set to 0.1.

In the present application, the SEARCH strategy differs from one iteration to another, and goes as follows. When $k = 0$, the SEARCH consists of a 64 point Latin hypercube sampling of Ω in hopes of identifying promising basins [27, 36]. At iteration $k \geq 1$, the SEARCH consists of evaluating the surrogate barrier function at 8 randomly generated mesh points and at the point produced by the *dynamic search* described in [4] (when applicable). This dynamic search is only called after a successful iteration, and essentially consists of evaluating ψ at the next mesh point in a previously successful direction.

In the present context, a function evaluation consists in solving a list of problems with the trust-region Algorithm 3.1 and combining the results on each problem into a unique real number. The objective function was defined by a list of relatively hard problems of significant dimension. A surrogate function σ, as described in section 4.1, was defined by a set of smaller problems which could be expected to be solved in overall reasonable time. The surrogate function is simply

$$\sigma^{\mathrm{cpu}}(p) = \sum_{j \in \mathcal{S}} \tau_j(p). \tag{4.6}$$

The problems were chosen as follows.

The trust-region algorithm described in section 4.2 with the classical parameter values p^{C} was run on the 163 unconstrained regular problems from the CUTEr collection, using the default dimensions. From those, some problems failed to be solved for memory reasons and some reached the maximum number of iterations of 1000. Two test lists were extracted from the results. The *surrogate* list $\mathcal{S}$ consists of those problems for which $n_i < 1000$ and $0.01 \leq \tau_i(p^{\mathrm{C}}) \leq 30$ (measured in seconds). The surrogate list contains 54 problems of small to moderate dimension, $2 \leq n_i \leq 500$ and such that $\sum_{i \in \mathcal{S}} \tau_i(p^{\mathrm{C}}) = 68.20$ seconds. We may thus expect that running through the surrogate list, i.e., evaluating the surrogate, should not take much longer than two minutes. Problems in this list and their dimension are summarized in Table 5. The *objective* list $\mathcal{O}$ consists in those problems for which $n_i \geq 1000$ and $\tau_i(p^{\mathrm{C}}) \leq 3600$. This yields a list of 55 problems with $1000 \leq n_i \leq 20000$ and such that $\sum_{i \in \mathcal{O}} \tau_i(p^{\mathrm{C}}) = 13461$ seconds, which amounts to approximately 3 hours and 45 minutes. The latter is the time that *one* objective function evaluation may be expected to take. Problems in this list and their dimension are summarized in Table 4.

TABLE 1
Minimization of the surrogate function to improve the initial solution. The measure $\sigma^{cpu}(p)$ is in seconds.

#f evals	$\sigma^{\text{cpu}}(p)$	$\tilde{\eta}_1$	$\tilde{\eta}_2$	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$
1	68.20	250	75	50	20
3	66.11	538.86719	77.480469	43.765625	17.914062
57	56.94	221.65625	89.175781	39.511719	23.042969
138	56.53	221.65625	89.300781	39.402344	23.136719
139	54.42	221.65625	89.675781	39.074219	23.417969
141	53.88	221.65625	89.675781	39.074219	22.917969
154	53.67	221.65625	90.175781	39.074219	22.667969
194	52.57	221.6875	90.175781	38.996094	22.792969
224	52.43	221.625	90.203125	38.996094	22.792969
289	52.35	221.625	90.207031	38.996094	22.792969

5. Numerical results. All tests were run on a 500 MHz Sun Blade 100 running SunOS 5.8. The implementation of MADS is a C++ package called NOMAD.[2]

5.1. Improving the initial solution with a surrogate function. The first run consisted of applying MADS to the surrogate function (4.6) from the classical parameters p^{C}. Results are reported in Table 1. The first column contains the number of MADS function evaluations required to improve the measure to the value in the second column. The other columns contain the corresponding parameter values.

MADS stopped after 310 function evaluations as the mesh size parameter Δ_k dropped below the stopping tolerance of 10^{-6}. The best set of parameters

$$p^{\text{S}} = (0.221625, 0.90207031, 0.38996094, 2.2792969)$$

was identified at the 289th evaluation. This strategy allowed the improvement of the truth initial value from $\psi^{\text{cpu}}(p^{\text{C}}) = 13461$ to $\psi^{\text{cpu}}(p^{\text{S}}) = 11498$ seconds, i.e., an improvement of approximately 33 minutes or 14.58%.

Note that the value of the surrogate at p^{C} had to be evaluated again during the first iteration and that its differred by approximately 1% from the one we had first obtained section 4.4 before building the surrogate. This is an illustration of the nondeterministic aspect of such objective functions.

5.2. Performance profiles. Comparison between the initial and final values of the objective function, i.e., the benchmarking of the trust-region method on a set of nonlinear unconstrained programs for the initial and final values of the parameters, will be presented using performance profiles. Originally introduced in [13], we briefly recall here how to read them.

Suppose that a given algorithm $\mathcal{A}_i$ from a competing set $\mathcal{A}$ reports a statistic $u_{ij} \geq 0$ when run on problem j from a test set $\mathcal{S}$, and that the smaller this statistic the better the algorithm is considered. Let the function

$$\omega(u, u^*, \alpha) = \begin{cases} 1 & \text{if } u \leq \alpha u^* \\ 0 & \text{otherwise} \end{cases}$$

be defined for all u, u^* and all $\alpha \geq 1$. The *performance profile* of algorithm $\mathcal{A}_i$ is the function

$$\pi_i(\alpha) = \frac{\sum_{j \in \mathcal{S}} \omega(u_{ij}, u_j^*, \alpha)}{|\mathcal{S}|} \quad \text{with} \quad \alpha \geq 1,$$

[2]May be downloaded from www.gerad.ca/NOMAD.

TABLE 2

Minimization of the objective function from improved starting point. The measure $\psi^{cpu}(p)$ is in seconds.

$\#f$ evals	$\psi^{\text{cpu}}(p)$	$\tilde{\eta}_1$	$\tilde{\eta}_2$	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$
1	11498.02	221.625	90.207031	38.996094	22.792969
5	11241.23	221.375	90.207031	38.871094	22.917969
7	10757.54	221.375	90.207031	38.871094	23.417969
13	10693.04	219.375	90.207031	38.871094	24.417969
40	10691.43	219.25	90.207031	38.871094	24.417969
42	10617.41	219.25	90.207031	38.621094	24.417969
73	10617.41	219.25	90.207031	38.621094	24.417969
74	10279.85	221.25	94.457031	37.996094	23.042969
77	10183.95	221.25	94.457031	37.933594	23.042969
97	10183.95	221.25	94.457031	37.933594	23.042969
98	10195.39	221.25	94.457031	37.933594	23.042969
115	10195.39	221.25	94.457031	37.933594	23.042969
116	10193.14	221.25	94.457031	37.933594	23.042969
142	10193.14	221.25	94.457031	37.933594	23.042969

where $u_j^* = \min_{i \in \mathcal{A}} u_{ij}$. Thus, $\pi_i(1)$ gives the fraction of the number of problems for which algorithm $\mathcal{A}_i$ was the most effective, according to the statistics u_{ij}, $\pi_i(2)$ gives the fraction for which algorithm $\mathcal{A}_i$ is within a factor of 2 of the best, and $\lim_{\alpha \to \infty} \pi_i(\alpha)$ gives the fraction of examples for which the algorithm succeeded.

5.3. Minimizing the total computing time. Algorithm 2.1 was restarted using the objective function $\psi^{\text{cpu}}(\cdot)$. The starting point was the parameter values suggested by the surrogate function in Table 1. The stopping condition this time was to perform a maximum of 150 truth evaluations. Based on the estimate of roughly 3 hours and 45 minutes per function evaluation, this amounts to an expected total running time of just about three weeks. Fragments of the evolution of $\psi^{\text{cpu}}(\cdot)$ are given in Table 2.

The final iterate

$$p^* = (0.22125, 0.94457031, 0.37933594, 2.3042969) \tag{5.1}$$

produced by MADS gives a value $\psi^{\text{cpu}}(p^*) = 10193.14$, and reduces the total computing time to just under 2 hours and 50 minutes; a reduction of more than 11% of the computing time when compared to p^{S} and of almost 25% when compared to p^{C}. This p^* is clearly in favor of a sufficient decrease condition rather than of $\eta_1 \approx 0$.

The values $\psi^{\text{cpu}}(p^{\text{C}})$ and $\psi^{\text{cpu}}(p^*)$ can be visualized in the profile of Figure 2 which compares the cpu-time profiles of Algorithm 3.1 applied to the objective list for the initial and final parameter values. The profile of Figure 3 presents a similar comparison, using the number of function evaluations.

5.4. Interpretation of the results. The above results must be interpreted in light of the objective function used in the minimization procedure. Figures 2 and 3 result from a minimization of $\psi^{\text{cpu}}(\cdot)$. A minimization of the total number of trust-region function evaluations $\psi^{\text{fval}}(\cdot)$ would likely have produced a different set of parameters. Moreover, the simplicity of $\psi^{\text{fval}}(\cdot)$ and $\psi^{\text{cpu}}(\cdot)$ is counterbalanced by their disadvantage of computing *global* measures. More sophisticated objectives in the present application could penalize the fact that a particular problem took a long time to fail for some parameter values while for others, failure was quickly detected.

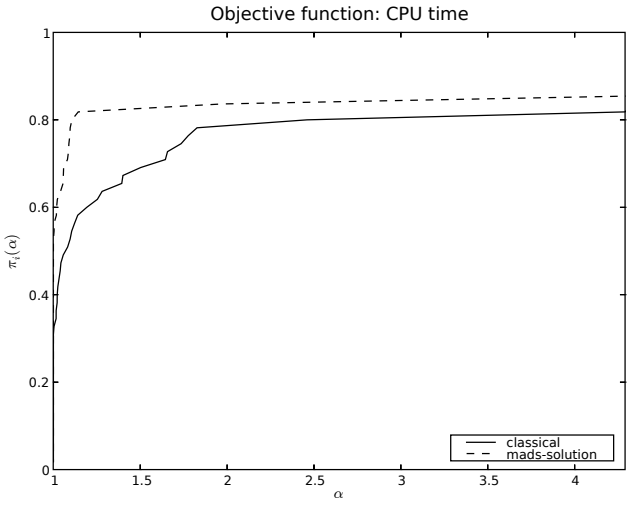


FIG. 2. *Profile comparing the CPU time required for one evaluation of the* MADS *objective for the initial and final parameter values.*

Similarly, they do not treat differently problems which are uniformly solved in a fraction of a second for nearly all parameter values and problems whose running time varies with great amplitude. Such effects might, and do, cause MADS to elect against exploring certain regions.

To illustrate this point, we note that the parameter values recommended by MADS differ significantly from those recommended in [17]. In a separate preliminary series

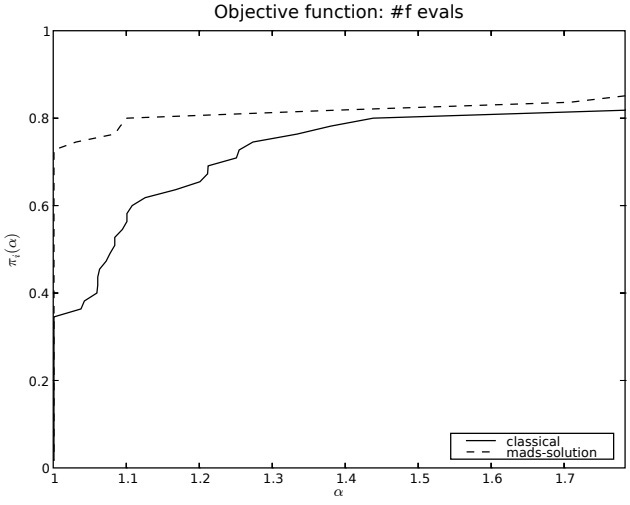


FIG. 3. *Profile comparing the number of function evaluations required for one evaluation of the* MADS *objective for the initial and final parameter values. The scale of the* α *axis has been extended.*

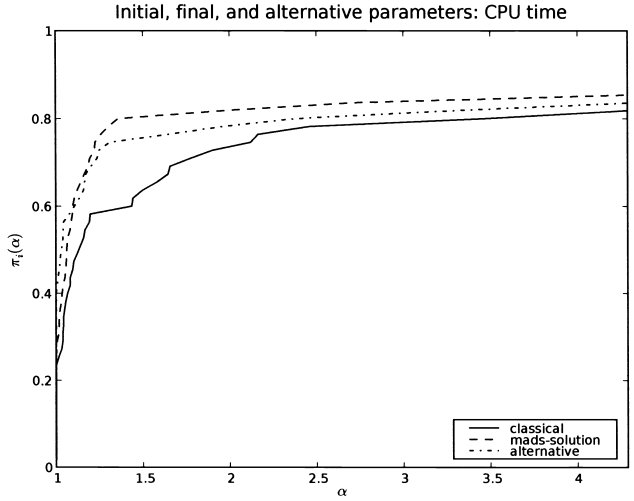


FIG. 4. *Profile comparing the CPU time required for one evaluation of the* MADS *objective for the initial, final, and alternative parameter values.*

of tests, the surrogate function $\sigma^{\text{cpu}}(\cdot)$ was minimized over Ω (4.2) without using the scaling (4.3). The final, *alternative*, set of parameters thus produced

$$p^{\text{A}} = (0.000008, 0.9906, 0.3032, 3.4021), \tag{5.2}$$

is rather close to the recommendations of [17] and seems to indicate that enforcing sufficient descent is not particularly beneficial in practice. The value $\psi^{\text{cpu}}(p^{\text{A}}) = 13707$ is surprisingly higher than $\psi^{\text{cpu}}(p^{\text{C}}) = 13461$. However, the corresponding profile appears significantly better than the reference algorithm using p^{C} as illustrated by Figures 4 and 5. Problems with long solution times are gathered in Table 3. The first three of those do not appear to influence the behavior of MADS by much, as their solution time varies little. Some problems failed to be solved for any values of the parameter. Among those, `GENHUMPS` is particularly detrimental to the value $\psi^{\text{cpu}}(p^{\text{A}})$ as the failure takes 10 times longer to be detected than at p^*. Likely, the value p^{A} would produce much better results if `GENHUMPS` were not present. We see nonetheless in the present case that MADS performed its task as it should have and that, perhaps, it is the objective function $\psi^{\text{cpu}}(\cdot)$ which should take such outliers into account, since the presence of problems like `GENHUMPS` cannot be anticipated.

TABLE 3

Problems with relatively high and varying solution times. A "F" indicates a failure. CPU times are in seconds.

Problem $\mathcal{P}_i$	$\tau_i(p^{\text{C}})$	$\tau_i(p^*)$	$\tau_i(p^{\text{A}})$
`DIXON3DQ`	2728.24	1572.75	1286.14
`EIGENALS`	1768.25	1177.19	1119.76
`NCB20B`	1444.47	1152.80	964.25
`CHAINWOO`	1224.25 F	1224.10 F	1392.78 F
`GENHUMPS`	615.29 F	444.96 F	4028.99 F

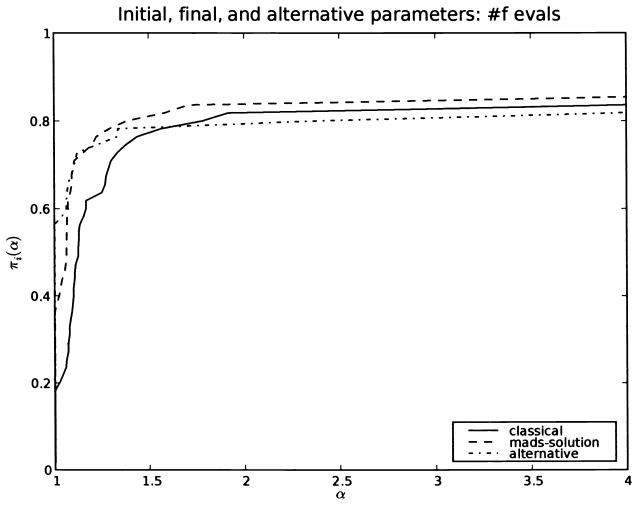


FIG. 5. *Profile comparing the number of function evaluations required for one evaluation of the* MADS *objective for the initial, final, and alternative parameter values.*

A phenomenon of a much more optimistic kind is revealed by problem `CRAGGLVY` which could not be solved in less than 1000 iterations using p^{C}, which took 83.3 seconds, but was solved in 20 iterations and 6.59 seconds using p^*.

6. Discussion. As mentioned earlier, the determination of useful parameter values for the management of the trust region was examined in the past in [17], where, on a much smaller list of 24 problems and after discretizing Ω into 3960 values for the parameters, the final value p^{A} (5.2) was identified. In order to supplement Figures 4 and 5 and to validate the final parameter value p^* (5.1) produced by MADS, we performed an additional series of tests.

We restarted MADS with the initial value p^{A} to determine whether or not it would identify p^{A} as a local minimum, or escape and move to a different region. To make the task more difficult for MADS, it was initialized with a small initial mesh size $\Delta_0 = 4^{-4} = 0.00390625$. The initial search was disabled to force the algorithm to explore only nearby regions. As before, the surrogate was used and all other algorithmic parameters were left unaltered. To limit the effort, we imposed a maximum of 100 truth function evaluations.

After almost 16 days of computation and 100 evaluations of the truth function, MADS moved away from p^* and exited with final parameter values

$$\overline{p} = (0.0023205, 0.999233, 0.356989, 3.39077)$$

and objective value $\psi^{\mathrm{cpu}}(\overline{p}) = 11144$. Most noticeably, the values of the first and third parameters changed the most, causing an improvement of 42 minutes. From the output of the algorithm, it is likely that it would have kept moving away towards larger values of η_1. We believe this is due to the diverse nature of the problems in $\mathcal{O}$ and in particular, `GENHUMPS`, cf. Table 3. This indicates that, in the present case, p^{A} is not a local minimizer.

To simply compare the computational burden involved in the tests with MADS and those of [17], notice that one function evaluation in [17] takes between 8.76 and 63.9 seconds, which is comparable to the cost of our surrogate. According to Table 1, it only took 289 evaluations of the surrogate to identify p^{S} which is already rather close to p^*. Even if $\psi^{\mathrm{cpu}}(p^{\mathrm{S}})$ is not as good as $\psi^{\mathrm{cpu}}(p^*)$, it is much lower than $\psi^{\mathrm{cpu}}(p^{\mathrm{A}})$. The overall cost of the tests of section 5.1 on the surrogate function is about 4 hours—only marginally more than an expected evaluation of ψ^{cpu}. We could therefore estimate that the total cost of identifying p^* is the 142 evaluations reported in Table 2 plus the evaluation of section 5.1, or 143 evaluations. This is 27.7 times less than the 3960 evaluations required by the exhaustive exploration.

With an expected 3 hours and 45 minutes per function evluation, discretizing as in [17] and evaluating ψ^{cpu} at the 3960 different parameter values might have taken us as much as 3960×3.75 hours ≈ 619 days. Pessimistically speaking, this almost represents 2 years of computation. While this might be perceived as the price to pay for global information, there still is no guarantee that a global minimizer has been found. We prefer to think in terms of regions instead of precise parameter values.

7. Conclusion. We presented a framework for the optimization of algorithmic parameters, which is general enough to be applied to many branches of engineering and computational science. Using the algorithm presented in [2], this framework may also be extended to the case where some parameters are categorical. The framework is illustrated on an example which at the same time addresses the long-standing question of determining locally optimal trust-region parameters in unconstrained minimization. The MADS algorithm for nonsmooth optimization of expensive functions [4] is at the core of the framework.

The very notion of optimality for such problems is not well defined. Hence, our aim in designing this framework was to suggest values for the parameters which seem to perform *better*, in a sense specified by the user, on a set of problems which are context-dependent and can also be specified by the user. In real applications, we believe this black-box approach is beneficial since it allows users to take full advantage of their knowledge of the context to design appropriate test sets and performance measures. As our numerical experience indicates, the choice of objective to be optimized will likely influence the results.

We reserve the exploration of more elaborate objective functions, making provision for outliers, and the study of scaling strategies for MADS for future work. We also wish to explore modifications of the algorithm to accept more general constraints.

Appendix. Tables. Tables 4 and 5 report numerical results on the objective and surrogate lists. In these tables, n is the number of variables, $\#f$ eval is the number of function evaluations, $\tau_i(\cdot)$ is the CPU time in seconds, and $\varphi_i(\cdot)$ is the number of trust-region function evaluations. These statistics are reported at the three sets of parameters p^{C}, p^*, and p^{A}.

Acknowledgments. We wish to thank Gilles Couture for developing NOMAD, the C++ implementation of MADS. We also wish to acknowledge the constructive comments of two anonymous referees which helped improve an earlier version of this paper.

TABLE 4

Results on the objective list $\mathcal{O}$. Timings $\tau_i(\cdot)$ are in seconds. The measure $\varphi_i(\cdot)$ is the number of function evaluations in each case. Failures occur when $\varphi_i(\cdot) = 1001$.

Name	n	$\tau_i(p^{\mathrm{C}})$	$\tau_i(p^*)$	$\tau_i(p^{\mathrm{A}})$	$\varphi_i(p^{\mathrm{C}})$	$\varphi_i(p^*)$	$\varphi_i(p^{\mathrm{A}})$
ARWHEAD	5000	0.95	0.92	0.93	7	7	7
BDQRTIC	5000	5.24	5.26	5.24	18	18	17
BROYDN7D	5000	443.70	242.95	205.03	331	264	261
BRYBND	5000	5.83	6.35	4.87	15	15	13
CHAINWOO	4000	1224.25	1224.10	1392.78	1001	1001	1001
COSINE	10000	1.83	1.63	1.70	14	12	13
CRAGGLVY	5000	83.30	6.59	6.39	1001	20	19
DIXMAANA	3000	0.53	0.56	0.48	12	12	11
DIXMAANB	3000	0.46	0.51	0.51	12	13	12
DIXMAANC	3000	0.55	0.52	0.53	14	13	13
DIXMAAND	3000	0.64	0.63	0.56	15	15	14
DIXMAANE	3000	7.19	7.87	8.90	15	15	16
DIXMAANF	3000	37.29	21.00	24.47	26	24	20
DIXMAANG	3000	30.91	7.20	23.65	24	18	24
DIXMAANH	3000	12.84	25.33	30.62	22	24	27
DIXMAANI	3000	178.43	193.92	185.57	17	17	16
DIXMAANJ	3000	455.60	399.68	317.44	35	32	30
DIXMAANL	3000	461.46	278.59	364.61	40	29	32
DIXON3DQ	10000	2728.24	1572.75	1286.14	10	8	8
DQDRTIC	5000	1.25	1.13	1.07	13	12	12
DQRTIC	5000	2.33	2.37	2.22	53	50	47
EDENSCH	2000	0.98	1.06	1.02	21	21	19
EG2	1000	0.06	0.07	0.06	4	4	4
EIGENALS	2550	1768.25	1177.19	1119.76	98	77	80
ENGVAL1	5000	2.47	1.77	1.71	18	15	14
EXTROSNB	1000	78.29	96.82	69.68	1001	1001	1001
FLETCBV3	5000	314.52	420.17	454.31	1001	1001	1001
FLETCHBV	5000	314.96	408.22	449.52	1001	1001	1001
FLETCHCR	1000	77.14	85.28	76.15	1001	1001	1001
FMINSRF2	5625	151.81	166.47	177.89	140	239	333
FMINSURF	5625	134.09	129.05	212.60	122	169	488
FREUROTH	5000	2.37	2.19	67.56	18	16	1001
GENHUMPS	5000	615.29	444.96	4028.99	1001	1001	1001
INDEF	5000	260.57	200.16	273.50	1001	1001	1001
LIARWHD	5000	1.74	1.84	1.50	20	22	18
MODBEALE	20000	38.46	23.39	28.13	23	16	17
NCB20	5010	639.83	457.02	336.32	86	71	55
NCB20B	5000	1444.47	1152.80	964.25	23	23	21
NONCVXU2	5000	647.52	539.34	538.52	1001	1001	1001
NONCVXUN	5000	708.77	555.75	551.77	1001	1001	1001
NONDIA	5000	0.61	0.51	0.58	8	8	8
NONDQUAR	5000	415.12	168.80	331.19	164	92	123
PENALTY1	1000	0.43	0.44	0.50	62	56	62
POWELLSG	5000	1.85	1.81	1.78	25	24	24
POWER	10000	54.74	54.96	54.65	44	44	43
QUARTC	5000	2.36	2.31	2.18	53	50	47
SCHMVETT	5000	3.85	3.79	3.82	11	10	10
SINQUAD	5000	2.72	2.77	2.78	16	16	17
SPARSQUR	10000	17.10	17.04	16.03	28	27	26
SROSENBR	5000	0.56	0.51	0.55	11	10	10
TESTQUAD	5000	37.97	36.40	37.64	18	17	16
TOINTGSS	5000	3.60	2.81	1.03	23	19	12
TQUARTIC	5000	1.20	1.17	1.15	15	14	15
TRIDIA	5000	28.40	30.07	29.15	17	16	15
WOODS	4000	6.09	6.34	7.61	66	68	74
Total	ψ	13641.01	10193.14	13707	11837	10771	12173

TABLE 5

Results on the surrogate list $\mathcal{S}$. Timings $\tau_i(\cdot)$ are in seconds. The measure $\varphi_i(\cdot)$ is the number of function evaluations in each case. Failures occur when $\varphi_i(\cdot) = 1001$.

Name	n	$\tau_i(p^{\mathrm{C}})$	$\tau_i(p^*)$	$\tau_i(p^{\mathrm{A}})$	$\varphi_i(p^{\mathrm{C}})$	$\varphi_i(p^*)$	$\varphi_i(p^{\mathrm{A}})$
3PK	30	0.14	0.10	0.12	74	58	61
ARGLINA	200	0.45	0.46	0.42	6	6	5
BIGGS6	6	0.02	0.02	0.00	27	26	25
BOX3	3	0.03	0.01	0.02	9	9	9
BROWNAL	200	0.30	0.30	0.30	7	7	7
BROWNBS	2	0.02	0.01	0.01	26	24	18
BROWNDEN	4	0.02	0.01	0.01	13	12	12
CHNROSNB	50	0.32	0.29	0.54	69	70	109
CLIFF	2	0.02	0.00	0.02	30	30	30
CUBE	2	0.02	0.01	0.02	48	43	42
DECONVU	61	2.91	0.75	0.57	93	35	27
DENSCHND	3	0.03	0.00	0.02	34	34	35
DIXMAANK	15	0.02	0.01	0.02	12	12	12
DJTL	2	0.08	0.07	0.08	171	150	156
ERRINROS	50	0.25	0.22	0.22	74	73	72
GENROSE	500	28.54	25.04	21.78	465	413	377
GROWTHLS	3	0.12	0.12	0.08	186	189	168
GULF	3	0.18	0.16	0.14	58	49	44
HAIRY	2	0.04	0.03	0.02	82	47	41
HEART6LS	6	0.59	0.59	0.57	1001	1001	1001
HEART8LS	8	0.25	0.12	0.14	263	148	186
HIELOW	3	2.73	1.81	1.95	15	10	11
HIMMELBF	4	0.11	0.07	0.08	205	113	113
HUMPS	2	0.44	0.43	0.27	1001	1001	588
LOGHAIRY	2	0.44	0.48	0.44	1001	1001	1001
MANCINO	100	11.61	6.13	5.54	24	20	18
MARATOSB	2	0.34	0.32	0.34	1001	1001	1001
MEYER3	3	0.54	0.56	0.48	1001	1001	1001
OSBORNEA	5	0.09	0.08	0.08	74	80	75
OSBORNEB	11	0.19	0.16	0.15	25	22	24
PALMER1C	8	0.68	0.70	0.73	1001	1001	1001
PALMER1D	7	0.06	0.02	0.02	54	19	19
PALMER2C	8	0.59	0.61	0.63	1001	1001	1001
PALMER3C	8	0.31	0.50	0.17	651	1001	301
PALMER4C	8	0.08	0.39	0.09	85	1001	132
PALMER6C	8	0.13	0.07	0.05	234	112	100
PALMER7C	8	0.41	0.47	0.06	1001	1001	193
PALMER8C	8	0.14	0.11	0.08	273	233	167
PENALTY2	200	0.39	0.38	0.37	18	18	18
PFIT1LS	3	0.28	0.16	0.23	630	363	477
PFIT2LS	3	0.10	0.09	0.11	247	184	222
PFIT3LS	3	0.11	0.09	0.11	280	204	215
PFIT4LS	3	0.21	0.16	0.23	515	317	412
SENSORS	100	11.13	9.47	11.21	22	20	33
SISSER	2	0.02	0.01	0.00	15	15	15
SNAIL	2	0.04	0.03	0.07	80	76	154
TOINTGOR	50	0.04	0.04	0.04	12	12	11
TOINTPSP	50	0.02	0.04	0.03	21	25	19
TOINTQOR	50	0.02	0.01	0.02	11	10	9
VARDIM	200	0.02	0.05	0.03	31	30	30
VAREIGVL	50	0.09	0.07	0.07	15	15	14
VIBRBEAM	8	2.39	2.39	2.33	1001	1001	1001
WATSON	12	0.04	0.03	0.04	11	11	11
YFITU	3	0.06	0.05	0.07	77	76	140
Total	σ	68.20	54.30	51.22	14381	14431	11964

REFERENCES

[1] P. Alberto, F. Nogueira, H. Rocha, and L. N. Vicente, *Pattern-search methods for user-provided points: Application to molecular geometry problems*, SIAM J. Optim., 14 (2004), pp. 1216–1236.

[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., *Analysis of generalized pattern searches*, SIAM J. Optim., 13 (2003), pp. 889–903.

[4] C. Audet and J. E. Dennis, Jr., *Mesh adaptive direct search algorithms for constrained optimization*, SIAM J. Optim., 17 (2006), pp. 188–217.

[5] A. J. Booker, J. E. Dennis, Jr., P. D. Frank, D. B. Serafini, and V. Torczon, *Optimization using surrogate objectives on a helicopter test example*, in Optimal Design and Control, Progress in Systems and Control Theory, J. Borggaard, J. Burns, E. Cliff, and S. Schreck, eds., Birkhäuser, Cambridge, MA, 1998, pp. 49–58.

[6] 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 Optimization, 17 (1999), pp. 1–13.

[7] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983, reissued as Classics in Applied Mathematics 5, SIAM, Philadelphia, 1990.

[8] 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.

[9] A. R. Conn, N. I. M. Gould, and Ph. L. Toint, *Testing a class of methods for solving minimization problems with simple bounds on the variables*, Math. Comp., 50 (1988), pp. 399–430.

[10] A. R. Conn, N. I. M. Gould, and Ph. L. Toint, *Trust-Region Methods*, SIAM, Philadelphia, 2000.

[11] 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.

[12] J. E. Dennis and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Classics in Applied Mathematics, SIAM, Philadelphia, 1996.

[13] E. Dolan and J. Moré, *Benchmarking optimization software with performance profiles*, Math. Program., 91 (2002), pp. 201–213.

[14] R. Fletcher and S. Leyffer, *Nonlinear programming without a penalty function*, Math. Program., 91 (2002), pp. 239–270.

[15] S. M. Goldfeldt, R. E. Quandt, and H. F. Trotter, *Maximization by quadratic hill-climbing*, Econometrica, 34 (1966), pp. 541–551.

[16] N. I. M. Gould, S. Lucidi, M. Roma, and Ph. L. Toint, *Solving the trust-region subproblem using the Lanczos method*, SIAM J. Optim., 9 (1999), pp. 504–525.

[17] N. I. M. Gould, D. Orban, A. Sartenaer, and Ph. L. Toint, *Sensitivity of trust-region algorithms*, 4OR, 3 (2005), pp. 227–241.

[18] N. I. M. Gould, D. Orban, and Ph. L. Toint, GALAHAD—*a library of thread-safe Fortran* 90 *packages for large-scale nonlinear optimization*, Trans. ACM Math. Softw., 29 (2003), pp. 353–372.

[19] N. I. M. Gould, D. Orban, and Ph. L. Toint, CUTEr *and* SifDec*, a constrained and unconstrained testing environment, revisited*, Trans. ACM Math. Softw., 29 (2003), pp. 373–394.

[20] N. I. M. Gould, C. Sainvitu, and Ph. L. Toint, *A filter-trust-region method for unconstrained optimization*, SIAM J. Optim., 16 (2005), pp. 341–357.

[21] L. Hei, *A self-adaptive trust region algorithm*, J. Comput. Math., 21 (2003), pp. 229–236.

[22] J. Jahn, *Introduction to the Theory of Nonlinear Optimization*, Springer, Berlin, 1994.

[23] T. G. Kolda, R. M. Lewis, and V. Torczon, *Optimization by direct search: New perspectives on some classical and modern methods*, SIAM Rev., 45 (2003), pp. 385–482.

[24] E. B. Leach, *A note on inverse function theorem*, Proc. Amer. Math. Soc., 12 (1961), pp. 694–697.

[25] K. Levenberg, *A method for the solution of certain problems in least squares*, Quart. J. Appl. Math., 2 (1944), pp. 164–168.

[26] R. M. Lewis and V. Torczon, *Pattern search algorithms for bound constrained minimization*, SIAM J. Optim., 9 (1999), pp. 1082–1099.

[27] M. D. McKay, W. J. Conover, and R. J. Beckman, *A comparison of three methods for selecting values of input variables in the analysis of output from a computer code*, Technometrics, 21 (1979), pp. 239–245.

[28] J. C. Meza and M. L. Martinez, *Direct search methods for the molecular conformation problem*, Journal of Computational Chemistry, 15 (1994), pp. 627–632.

[29] M. S. Ouali, H. Aoudjit, and C. Audet, *Optimisation des stratégies de maintenance*, Journal Européen des Systèmes Automatisés, 37 (2003), pp. 587–605.

[30] M. J. D. Powell, *A new algorithm for unconstrained optimization*, in Nonlinear Programming, J. B. Rosen, O. L. Mangasarian, and K. Ritter, eds., Academic Press, London, 1970, pp. 31–65.

[31] A. Sartenaer, *Armijo-type condition for the determination of a generalized Cauchy point in trust region algorithms using exact or inexact projections on convex constraints*, Belg. J. Oper. Res. Statist. Comput. Sci., 33 (1993), pp. 61–75.

[32] A. Sartenaer, *Automatic determination of an initial trust region in nonlinear programming*, SIAM J. Sci. Comput., 18 (1997), pp. 1788–1803.

[33] Ch. Sebudandi and Ph. L. Toint, *Nonlinear optimization for seismic travel time tomography*, Geophysical Journal International, 115 (1993), pp. 929–940.

[34] J. S. Shahabuddin, *Structured Trust-Region Algorithms for the Minimization of Nonlinear Functions*, Ph.D. thesis, Department of Computer Science, Cornell University, Ithaca, NY, 1996.

[35] T. Steihaug, *The conjugate gradient method and trust regions in large scale optimization*, SIAM J. Numer. Anal., 20 (1983), pp. 626–637.

[36] M. Stein, *Large sample properties of simulations using Latin hypercube sampling*, Technometrics, 29 (1987), pp. 143–151.

[37] Ph. L. Toint, *Towards an efficient sparsity exploiting Newton method for minimization*, in Sparse Matrices and Their Uses, I. S. Duff, ed., Academic Press, London, 1981, pp. 57–88.

[38] J. M. B. Walmag and E. J. M. Delhez, *A note on trust-region radius update*, SIAM J. Optim., 16 (2005), pp. 548–562.

[39] Y. Yuan, *An example of non-convergence of trust region algorithms*, in Advances in Nonlinear Programming, Y. Yuan, ed., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 205–218.

SIAM J. Optim.
Vol. 17, No. 3, pp. 665–687
© 2006 Society for Industrial and Applied Mathematics

BEND MINIMIZATION IN PLANAR ORTHOGONAL DRAWINGS USING INTEGER PROGRAMMING*

PETRA MUTZEL† AND RENÉ WEISKIRCHER‡

Abstract. We consider the problem of minimizing the number of bends in a planar orthogonal graph drawing. While the problem can be solved via network flow for a given planar embedding of a graph, it is NP-hard if we consider all planar embeddings. Our approach for biconnected graphs combines a new integer linear programming (ILP) formulation for the set of all embeddings of a planar graph with the network flow formulation of the bend minimization problem for fixed embeddings. We report on extensive computational experiments with two benchmark sets containing a total of more than 12,000 graphs where we compared the performance of our ILP-based algorithm with a heuristic and a previously published branch & bound algorithm for solving the same problem. Our new algorithm is significantly faster than the previously published approach for the larger graphs of the benchmark graphs derived from industrial applications and almost twice as fast for the benchmark graphs from the artificially generated set of hard instances.

Key words. graph drawing, planar drawing, orthogonal drawing, bend minimization, graph theory applications, mixed integer programming, combinatorial optimization, polyhedral combinatorics, branch-and-bound, branch-and-cut, applications of mathematical programming

AMS subject classifications. 05C90, 68R05, 90C11, 90C27, 90C57, 90C90

DOI. 10.1137/040614086

1. Introduction.

Drawing graphs is important in many scientific and economic areas. Applications include the drawing of UML diagrams in software engineering, business process modeling as well as the design and visualization of databases. A popular way of drawing graphs is representing the vertices as boxes and the edges as sequences of horizontal and vertical line segments connecting the boxes. This drawing style is called *orthogonal* drawing. A point where two segments of an edge meet is called a *bend*. Figure 1 shows an orthogonal drawing of a graph.

A well-known approach for computing orthogonal drawings of general graphs is the topology-shape-metrics method (see, for example, [7]). In the first step, the topology of the drawing is computed. The objective in this phase is to minimize the number of edge crossings. In the second step, the shape of the drawing is defined in terms of bends along the edges and angles around the vertices. The objective is to minimize the number of bends for the given topology. Finally, the metrics of the drawing is computed. Commonly adopted optimization requirements in this step are short edge lengths and small area for the given shape. In this paper, we focus on the bend minimization step (the second step). Given a planar graph, the task is to compute an orthogonal representation with the minimum number of bends.

The infinite set of different planar drawings of a graph can be partitioned into a finite set of equivalence classes called *embeddings* of a graph. An embedding defines

*Received by the editors August 30, 2004; accepted for publication (in revised form) February 4, 2006; published electronically September 19, 2006. Results in this paper appeared in preliminary form in the proceedings of IPCO '99, COCOON '00, and COCOON '02.
http://www.siam.org/journals/siopt/17-3/61408.html
†University of Dortmund, Lehrstuhl für Algorithm Engineering/Experimentelle Algorithmen, Fachbereich Informatik LS11, Joseph-von-Fraunhoferstr. 20, 44227 Dortmund, Germany (petra.mutzel@cs.uni-dortmund.de).
‡CSIRO Mathematical and Information Sciences, Clayton South, VIC 3169, Australia (rene.weiskircher@csiro.au).

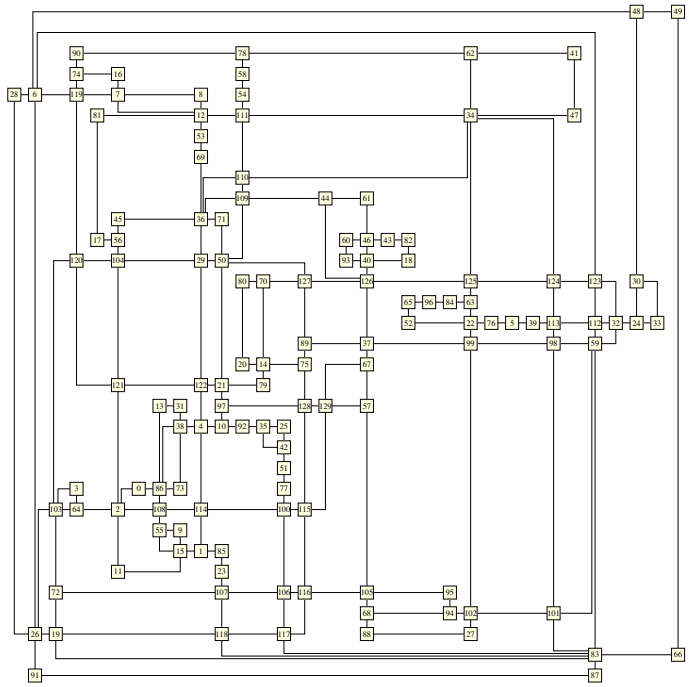


FIG. 1. *An orthogonal drawing of a graph with* 130 *nodes and* 205 *edges.*

the topology of a planar drawing without assigning lengths or shapes to the edges or fixing the shapes and positions of vertices.

A *combinatorial embedding* fixes the sequence of incident edges around each vertex in clockwise order. This also fixes the list of *faces* of a drawing. The faces are the connected regions of the plane defined by a planar drawing. A *planar embedding* additionally defines the outer (unbounded) face of a planar drawing. *Orthogonal representations* are equivalence classes of orthogonal drawings that fix the planar embedding and the bends and angles in an orthogonal drawing.

There are some results in the literature on the topic of optimizing certain functions over the set of all embeddings of a graph. Bienstock and Monma studied the complexity of covering vertices by faces [3] and minimizing certain distance measures on the faces of a graph with respect to the outer face [4, 5]. Tamassia presented the first algorithm for minimizing the number of bends in a planar orthogonal drawing for the case where the embedding is fixed [21]. Garg and Tamassia showed that optimizing the number of bends in an orthogonal drawing over the set of all embeddings of a planar graph is NP-hard [14]. In [9], Di Battista, Liotta, and Vargiu show that the problem is polynomially solvable for series-parallel graphs and 3-planar graphs.

If we consider only orthogonal drawings where the bends of the edges are placed on the same grid as the vertices and all vertices occupy only one grid point, we can only represent graphs with a maximum degree of four (so called *four-graphs*). In [11], Didimo and Liotta present an algorithm that produces planar orthogonal drawings of four-graphs with the minimum number of bends. The running time is exponential only in the number of vertices with degree four.

Bertolazzi, Di Battista, and Didimo [2] used the SPQR-tree data structure to devise a branch & bound algorithm for solving the bend minimization problem over

the set of all embeddings of a planar graph in a more general orthogonal model where a vertex can have degree greater than four. In this paper, we attack the same problem using integer linear programming. To do this, we have developed a new integer linear program describing the set of all combinatorial embeddings of a planar biconnected graph. To our knowledge, this is the first practicable description of the set of all combinatorial embeddings of a planar graph as an integer linear program. We achieve a manageable size of the program by computing the set of variables and constraints recursively using the SPQR-tree data structure. By combining our new integer linear program with a linear program that describes the set of all orthogonal representations of a planar graph with a fixed embedding, we obtain a mixed integer linear program that represents the set of all orthogonal representations for a planar biconnected graph over the set of all embeddings.

We use this new mixed integer linear program to minimize the number of bends in an orthogonal drawing over the set of all embeddings of a planar graph. Like the approach in [2], our new method can only guarantee optimality for biconnected graphs because we also use the SPQR-tree data structure, which is only defined for graphs that have this property. Our algorithm first computes the mixed integer linear program and then uses a commercial solver (CPLEX) to find an optimal solution. This solution is then transformed into an orthogonal representation of the graph.

We tested our approach on two different benchmark sets of graphs. The first consists of graphs derived from industrial applications and the second of graphs computed by a graph generator. The latter set was also used in [2] to measure the performance of the branch & bound algorithm. Our new approach is faster for the larger graphs in the first benchmark set than the branch & bound approach of Bertolazzi, Di Battista, and Didimo [2] and about twice as fast on the graphs in the seconds benchmark set, as our computational results show.

Preliminary descriptions of the computation of the integer linear program that describes all embeddings of a planar biconnected graph can be found in [18] and [19]. A preliminary description of the combination with the integer linear program that describes all orthogonal representations for a fixed embedding and first computational results can be found in [20].

In section 2, we give a short overview of SPQR-trees. We present the four different types of nodes in the tree and the properties of the tree that are important for our approach. Section 3 summarizes the recursive construction of the new integer linear program that describes the combinatorial embeddings of a graph and contains the proof of correctness.

The linear program describing the orthogonal representations of a graph for a fixed embedding is the topic of section 4. This is basically the formulation of a minimum cost flow problem in a special network constructed from the graph and the embedding as a linear program. In section 5, we present the mixed integer linear program that is the result of merging the new integer linear program describing the embeddings of a graph with the linear program that describes the orthogonal representations for a graph where the embedding is fixed.

The topic of section 6 is the algorithm that we use to compute an orthogonal representation of a graph with the minimum number of bends over the set of all embeddings. The computational results we obtained by applying the algorithm to two sets containing a total of more than 12,000 benchmark graphs are given in section 7. We compare the algorithm with a heuristic and with the branch & bound algorithm of Bertolazzi, Di Battista, and Didimo. The conclusion (section 8) summarizes the

main results and contains possible starting points for future work.

2. SPQR-trees. In this section, we give a brief overview of the SPQR-tree data structure for biconnected graphs. A graph is biconnected if it is connected and cannot be disconnected by deleting a single vertex. SPQR-trees have been developed by Di Battista and Tamassia [10]. They represent a decomposition of a biconnected graph into its triconnected components. A connected graph is triconnected if there is no pair of vertices in the graph whose removal splits the graph into two or more components. Such a pair of vertices is called a *split pair*.

An SPQR-tree has four types of nodes and with each node we associate a biconnected graph which is called the *skeleton* of that node. This graph can be seen as a simplified version of the original graph where certain subgraphs are replaced by edges. The vertices in a skeleton are vertices of the original graph. The edges in a skeleton either correspond to edges in the original graph or represent subgraphs. Thus, each node of the SPQR-tree defines a decomposition of the graph. The node types and their skeletons are as follows:

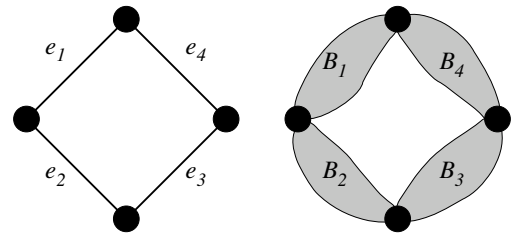
1. Q**-node:** The skeleton consists of two vertices connected by two edges. There is one Q-node for each edge in the graph.
2. S**-node:** The skeleton is a simple cycle with at least three vertices.
3. P**-node:** The skeleton consists of two vertices connected by at least three edges.
4. R**-node:** The skeleton is a triconnected graph with at least four vertices.

All leaves of the SPQR-tree are Q-nodes and all inner nodes S-, P- or R-nodes. When we see the SPQR-tree as an unrooted tree, then it is unique for every biconnected graph. Another important property of these trees is that their size (including the skeletons) is linear in the size of the original graph and that they can be constructed in linear time [17, 15]. See Figure 2 for examples of the skeletons of inner nodes of an SPQR-tree and the decomposition of the graph they define.

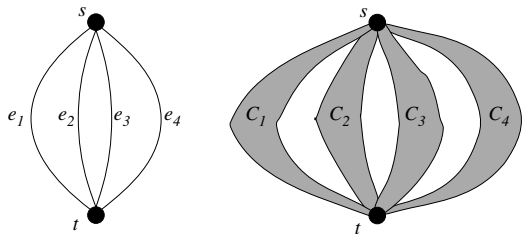
As described in [10], SPQR-trees can be used to represent the set of all combinatorial embeddings of a biconnected planar graph. Every combinatorial embedding of the original graph defines a unique combinatorial embedding for the skeleton of each node in the SPQR-tree. Conversely, when we define an embedding for the skeleton of each node in the SPQR-tree, we define a unique embedding for the original graph. The skeletons of S- and Q-nodes are simple cycles, so they have only one embedding. But the skeletons of the R- and P-nodes have at least two different combinatorial embeddings. This is the reason why they determine the embedding of the graph and we call these nodes the *decision nodes* of the SPQR-tree.

3. The ILP-formulation describing the set of all embeddings. Our new integer linear program (ILP) describing the set of all combinatorial embeddings of a planar graph is constructed recursively using the SPQR-tree data structure. Because SPQR-trees are only defined for biconnected graphs, the same is true for the ILP. We construct the program recursively by splitting the SPQR-tree into smaller SPQR-trees, constructing ILPs for the corresponding smaller graphs, and then merging them into an ILP for the original graph. The basis of the recursive construction are graphs whose SPQR-trees have only one inner node (S-, P- or R-node).

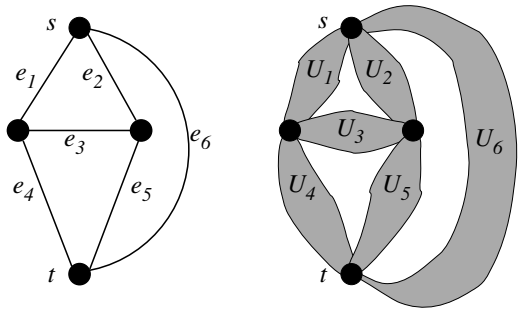
3.1. The ILP for graphs where the SPQR-tree has only one inner node. If the SPQR-tree for a graph G has exactly one inner node μ, then G is isomorphic to the skeleton S_μ of μ. It follows that the graph is either a simple cycle (μ is an S-node), a triconnected graph (μ is an R-node) or consists of two vertices connected



(a) S-node



(b) P-node



(c) R-node

FIG. 2. *The skeletons of the inner nodes of an SPQR-tree together with the decomposition of the graph they define. The grey shapes on the right depict the subgraphs represented by the edges with the same number on the left.*

by at least three edges (μ is a P-node).

In all three cases, the set of combinatorial embeddings is easy to describe. If the graph is a cycle, it has only one combinatorial embedding. If it is a triconnected graph, it has two embeddings that are mirror images of each other. If it consists of two vertices and at least three edges, the embeddings are determined by the different circular permutations of the edges connecting the two vertices.

The variables of the ILP correspond to the set of directed cycles of the graph, that are *face cycles* in at least one embedding. A directed cycle is a face cycle in an embedding, if the area of the plane on the right side of the cycle is empty in every planar drawing that realizes the embedding. First we describe the three possible

cases for graphs whose SPQR-tree has only one inner node together with the ILPs that describe their embeddings. We distinguish the three cases by the type of the only inner node of the SPQR-tree.

3.1.1. The ILP for graphs whose only inner node in the SPQR-tree is an *S*-node. In this case, G is a simple cycle. Therefore, G contains exactly two directed cycles and both are face cycles in the only combinatorial embedding of G. It follows that the corresponding ILP has two variables that are both equal to one in the only solution. So the ILP that describes all embeddings of a graph whose only inner node in its SPQR-tree is an S-node is simply $x_1 = x_2 = 1$.

3.1.2. The ILP for graphs whose only inner node in the SPQR-tree is an *R*-node. In this case the graph G is triconnected. A triconnected graph has two combinatorial embeddings that are mirror images of each other. This means that for any directed cycle c that is a face cycle in the first embedding, the directed cycle $\bar{c}$ passing the same edges as c in the opposite direction is a face cycle of the other embedding.

To find all potential face cycles of the graph, we first compute an arbitrary embedding of G, using for example the linear time planarity test of Hopcroft and Tarjan [16]. The face cycles in this embedding together with their reversal cycles form the set of all cycles in G that are face cycles in at least one embedding.

Let l be the number of faces in an embedding of G (note that $l = m - n + 2$ if m is the number of edges in G and n the number of vertices). Then there are $2l$ directed cycles in G that are face cycles in an embedding. If we write down the two embeddings of G as two vectors $\vec{s_1}, \vec{s_2} \in \{0,1\}^{2l}$, then each of the vectors contain l ones. Any entry with value one in $\vec{s_1}$ has value zero in $\vec{s_2}$ and vice versa. Therefore it is straightforward to describe the embeddings of G as an ILP.

If we order the variables for the face cycles such that the variables with odd numbers represent the face cycles of the first embedding and the variables with even numbers represent the second embedding, the ILP has a very simple structure. Let the variables of the problem be x_1 to x_{2l}. Then the ILP describing the embeddings is the following:

$$\begin{aligned} x_{2i-1} + x_{2l} &= 1 \text{ for } 1 \le i \le l, \\ x_{2i} - x_{2l} &= 0 \text{ for } 1 \le i \le l-1, \\ x_i &\ge 0 \text{ for } 1 \le i \le 2l, \\ x_i &\le 1 \text{ for } 1 \le i \le 2l. \end{aligned}$$

Note that this is a complete description of the polytope that describes the combinatorial embeddings of a triconnected graph. So if we want to optimize a linear function over the set of all embeddings, there is no need to demand that the solutions are integral.

3.1.3. The ILP for graphs whose only inner node in the SPQR-tree is a *P*-node. In this case, the graph G consists of two vertices connected by k edges with $k \ge 3$. Since combinatorial embeddings can be defined by the circular sequence of the edges around each vertex, G has $(k-1)!$ different embeddings. Each pair of edges in G corresponds to two directed cycles and all cycles are face cycles in at least one of the embeddings. Therefore, there are $d = k^2 - k$ variables in the ILP.

We realized that in this case the set of embeddings of G can be interpreted as the set of Hamiltonian tours in a complete directed graph H. The graph H has one

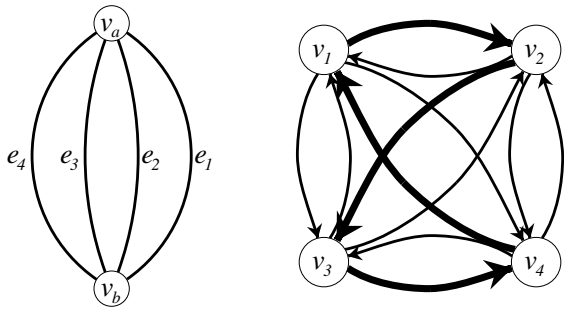


FIG. 3. *A graph where the only inner node in its SPQR-tree is a P-node and the corresponding graph H.*

vertex for every edge of G and one edge for each directed cycle. The edge in H that corresponds to the directed cycle c in G connects the vertices of H that correspond to the two edges in G that form the cycle c.

Figure 3 shows this correspondence. The edges of the graph on the left correspond to the vertices in the complete directed graph H on the right. Edge (v_i, v_j) in H corresponds to the cycle in G that passes edge e_i from v_a to v_b and edge e_j from v_b to v_a. The cycles that are face cycles in the embedding of G on the left correspond to the thick edges in the graph H on the right. Thus, the embedding of G shown on the left corresponds to the tour in H consisting of the thick edges. Note that the sequence of the edges in clockwise order around vertex v_a corresponds to the sequence of the vertices defined by the tour.

Because of the correspondence between the Hamiltonian tours in H and the embeddings of G, we can use the ILP-formulation for the asymmetric traveling salesman problem (ATSP) to describe the set of embeddings of G. We use the formulation of [6], which consists of a linear number of *degree constraints* and an exponential number of *subtour elimination constraints.* In our case, the degree constraints say that each edge in G is contained in exactly two face cycles of each embedding, once for each direction. The subtour elimination constraints say that for each proper subset E' of the edges of G, there must be at least one face cycle that consists of an edge in E' and of an edge in $E \setminus E'$.

Because the number of subtour elimination constraints grows exponentially with the number of edges in G (there is one constraint for each proper subset of the edges of G), we define the ILP for a graph whose SPQR-tree has a P-node as the only inner node just as the set of degree constraints. To cope with the subtour elimination constraints, we use the same approach that is used for the ATSP-problem: We separate the subtour elimination constraints during the optimization process.

So we first compute a solution vector $\vec{s} \in \{0,1\}^d$ for the problem without the subtour elimination constraints and then check if $\vec{s}$ violates any of the subtour elimination constraints. If this is not the case, we have found a valid solution representing a combinatorial embedding of G. Otherwise, we add the violated subtour elimination constraint to the set of constraints and reoptimize. To check if there is a violated subtour elimination constraint, we find a minimum cut in the graph H where the weight of each edge is defined by the corresponding component of the vector $\vec{s}$.

3.2. The ILP for graphs whose SPQR-tree has more than one inner node. If the SPQR-tree of G has more than one inner node, we split the tree at an

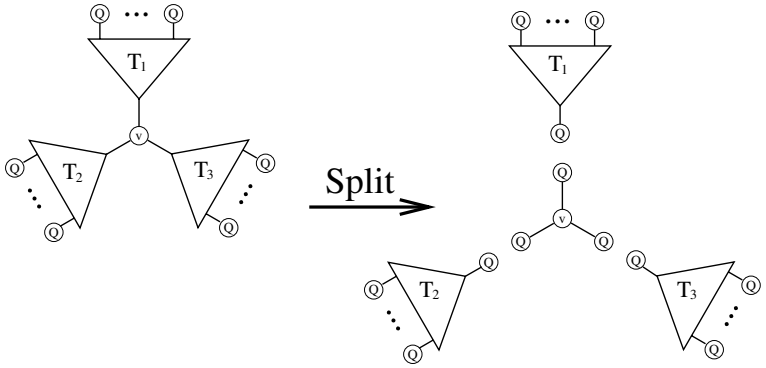


FIG. 4. *Splitting an SPQR-tree at an inner node.*

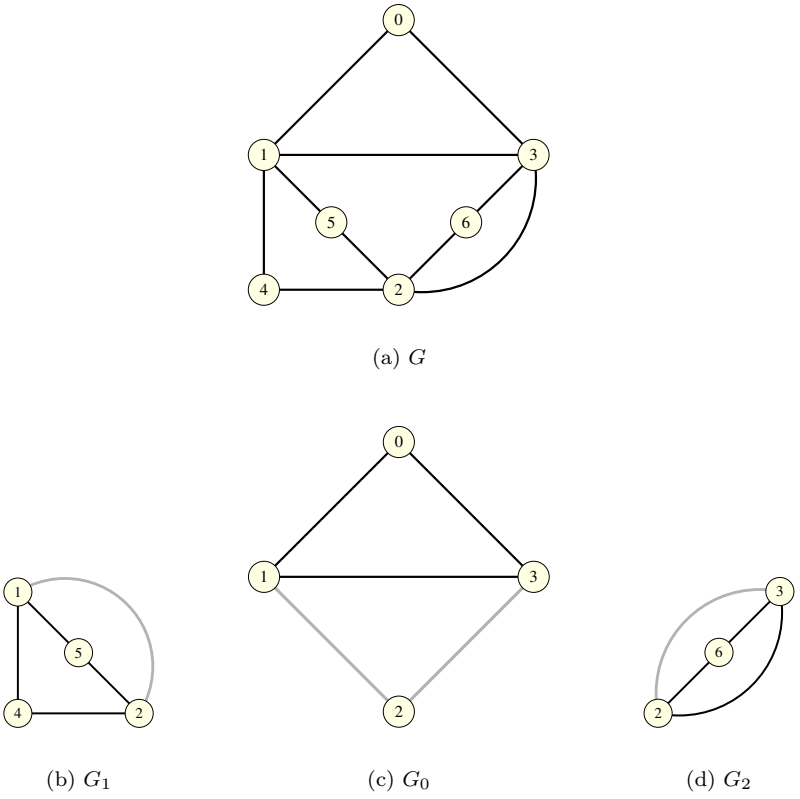
arbitrary decision node (R- or P-node). This is done as shown in Figure 4. In this figure, the split node is v. We split all the edges that connect v to the rest of the tree by inserting two new Q-nodes per edge. This is necessary to ensure that all leaves of the resulting trees are Q-nodes. In this way, we obtain smaller SPQR-trees, the *split trees*.

The graph represented by a split tree is called *split graph*. Each split graph can be obtained from G by replacing subgraphs connected to the rest of G only via a pair of vertices by an edge. These new edges, the *virtual edges*, correspond to the Q-nodes we add in the splitting process. Figure 5 shows a graph together with the split graphs resulting from splitting its SPQR-tree at its R-node. The grey edges are the virtual edges of the split graphs.

One of the resulting split trees contains v as the only inner node and so the ILP for the corresponding split graph has already been defined in the last section. We call this split graph the *center split graph*. In Figure 5, the center split graph is G_0. All other split trees contain at least one decision node less than the original tree, because they do not contain v. We continue the recursive splitting process until we have only SPQR-trees with just one inner node. This is always possible because any SPQR-tree with more than one inner node contains at least one decision node (two S-nodes can never be adjacent).

After recursively computing the ILPs for the split graphs, we merge them into an ILP for the original graph G. To achieve this, we need to lift the constraints computed for the split graphs, since G contains more potential face cycles than the split graphs. Therefore, the ILP describing its embeddings has more variables and thus a higher dimension. The number of variables and thus the number of potential face cycles in the original graph depend to a large degree on v. If it is a P-node, the number of potential face cycles in the original graph is roughly a multiple of the number of face cycles in the split graph where the multiplication factor is the number of face cycles in the skeleton of v. In all other cases, the number of face cycles and thus of variables in the ILP will be about the same as the sum of the number of variables in all split graphs.

The crucial idea in the computation of the merged ILP is to find for each cycle c in a split graph the set L_c of cycles of G that are *represented* by c. We say a cycle c' in G is represented by cycle c in the split graph G_i if the set of edges in c' that are also edges in G_i is identical to the set of nonvirtual edges in c and these edges are passed in the same direction in both cycles.



(a) G

(b) G_1 (c) G_0 (d) G_2

FIG. 5. *A graph G and its split graphs G_0, G_1, and G_2. Grey edges are virtual.*

Let $G_0, \ldots, G_k$ be the split graphs of G and G_0 the center split graph. Let C_i for $0 \leq i \leq k$ be the set of cycles in G_i that are represented by a variable in the corresponding ILP and $C = \bigcup_{i=0}^{k} C_i$. We can split C into two sets: The set C_L of cycles that do not include a virtual edge of a split graph (called *local cycles*, because they contain only nonvirtual edges of one of the G_i) and the set C_M of cycles that contain a virtual edge (called *component cycles*).

The local cycles in C_L are also represented by variables in the ILP for G. We use the component cycles in C_M to construct cycles in G. Every cycle c in $C_i \cap C_M$ with $1 \leq i \leq k$ contains exactly one virtual edge. If we remove this edge, we get a path p in G_i connecting the two vertices of G_i that it shares with G_0. If we take a cycle c of $C_0 \cap C_M$ and replace each virtual edge by a path p in the corresponding split graph obtained from a cycle in C_M, we get a cycle $c' \in L_c$ in G.

Consider, for example, Figure 5. We construct cycles in G by taking a cycle c_0 in the center split graph G_0 and combining it with a cycle c_1 in G_1 and a cycle c_2 in G_2 that both contain the virtual edge (all split graphs except the center split graph contain exactly one virtual edge). This is done by replacing the virtual edges in c_0 by the paths in the split graphs obtained from the cycles in G_1 and G_2 by deleting the virtual edge. We choose the component cycle c_0 given by the vertex sequence $(0, 1, 2, 3)$. We combine c_0 with the component cycle $c_1 = (1, 4, 2)$ in G_1 and the component cycle $c_1 = (2, 6, 3)$ in G_2. The result is the cycle $c_3 = (0, 1, 4, 2, 6, 3)$ in G.

The set L_{c_0} consists of the four different cycles that we can obtain by combining the two possible paths through G_1 and G_2.

The variables in the ILP for G correspond to the cycles in C_L and to the cycles in G that we can construct by combining cycles in C_M. The latter are called *global cycles* because they contain edges from more than one of the split graphs. In general, a component cycle in C_M is used to construct several global cycles in G. We store for each component cycle c the set L_c of global cycles in G constructed by combining c with other cycles. For each cycle in C_L, we define L_c as the cycle itself.

By constructing the variables set in this way, we make sure that the set of variables corresponds exactly to the set of directed cycles that are face cycles in at least one embedding of the graph. This can be shown in a similar way as in the proof of the correctness of the ILP in section 3.3. So we have all the variables necessary to represent all embeddings but significantly less variables than if we introduced a variable for each directed cycle in the graph. The computational results in section 7 show that the number of variables grows only linearly with the size of the benchmark graphs.

Let c be a cycle in a split graph that is represented by variable x_c in the corresponding ILP. The set L_c defines a set L_{x_c} of variables in the ILP for the original graph. We use these sets L_{x_c} to lift the constraints of the ILPs of the split graphs. We simply replace each variable x_c in each constraint computed for a split graph by the sum of the variables in L_{x_c}. We call the constraints generated in this way the *lifted constraints*.

Remember that we have not explicitly computed the subtour elimination constraints for graphs whose only inner node in the SPQR-tree is a P-node. Instead, there is a complete directed graph H for each P-node where we can find violated subtour elimination constraints by computing a minimum cut. When we compute the ILP for G from the ILPs of the split graphs, we have to update the set of cycles that are represented by each edge of H.

If the only inner node of an SPQR-tree is a P-node, every edge e in H represents exactly one directed cycle c of the graph. The weight of e is the value of the corresponding cycle variable x_c in the current solution vector. When we compute the ILP of the original graph from the ILPs of the split graphs, we assume that every edge in H represents a *set* of cycles.

Let us now assume that H is the complete directed graph computed for the separation of subtour elimination constraints in a split graph and e an edge in H representing the set C_e of cycles. After we have computed the global cycles of the original graph, we replace each cycle c in C_e by the set L_c of global cycles generated using c. Thus, the new set of cycles represented by e consists of the set of all the global cycles in G constructed from cycles in C_e.

So each edge e in the complete bidirected graph H corresponds to a set $C(e)$ of cycles in G. To separate the subtour elimination constraints, we define the weight $w(e)$ of each edge e in H as $\sum_{c \in C(e)} x_c$, where x_c is the binary variable of the ILP associated with the cycle c. If there is a proper subset V' of the vertices of H such that the sum of the weights of the edges connecting vertices of V' is greater than $|V' - 1|$, we have found a violated constraint. We call these constraints the *lifted subtour elimination constraints.*

After we have computed the new variables, lifted the constraints of the split graphs, and updated the complete directed graphs computed for each P-node skeleton, we add two types of additional constraints. The first type of constraints says that

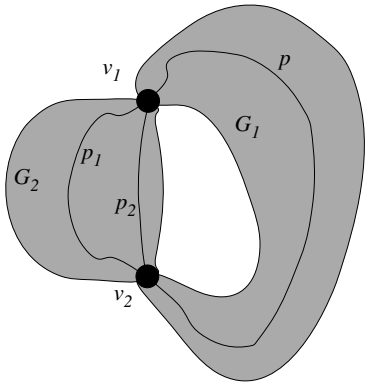


FIG. 6. *Two cycles that pass a split component in the same direction cannot both be face cycles in the same embedding.*

out of all global cycles created using the same component cycle, at most one can be a face cycle in any embedding of the original graph. This is true because there is a split component of a split pair of the graph that all these cycles pass in the same direction. We call this type of constraints the *choice constraints*. If c is a cycle in a split graph, than the corresponding choice constraint is:

$$\sum_{c' \in L_c} x_{c'} \leq 1.$$

To get an intuition why two such cycles can never be face cycles in the same embedding consider Figure 6. We assume that v_1 and v_2 are a split pair of the graph with the split components G_1 and G_2. We also assume that the two directed cycles c_1 and c_2 use the same path p in G_1 but different paths (p_1 and p_2) in G_2. If we assume that both c_1 and c_2 are face cycles in the same embedding Π of G, then we have a contradiction: Since c_1 is a face cycle, the area right of c_1 must be empty in Π and so the edges of p_2 must be left of p_1. But since c_2 is also a face cycle, the area to the right of c_2 must be empty and the path p_1 must be left of p_2.

The second type of constraints fixes the number of global cycles that are face cycles in each embedding. This number must be equal to the number of face cycles in an embedding of the center split graph that contain virtual edges. This constraint is called the *center graph constraint*.

3.3. Correctness of the ILP.

THEOREM 3.1. *Every feasible solution of the ILP corresponds to a combinatorial embedding of the given biconnected planar graph G and vice versa: every combinatorial embedding of G corresponds to a feasible solution for the generated ILP.*

The proof is split into three lemmas. They rely on the fact that by fixing the combinatorial embedding of each decision node in the SPQR-tree of a graph G, we define a unique combinatorial embedding of G. On the other hand, a combinatorial embedding of G defines a unique combinatorial embedding for the skeleton of each decision node in the SPQR-tree of G. From this fact, we can easily derive the following lemma:

LEMMA 3.2. *Let G be a biconnected planar graph and let T be its SPQR-tree. Let μ be a decision node in T, $T_0, \ldots, T_d$ be the split trees of μ (T_0 is the center split tree) and $G_0, \ldots, G_d$ the associated split graphs. Every combinatorial embedding Γ of G*

defines a unique embedding for each G_i. On the other hand, if we fix a combinatorial embedding Γ_i for each G_i, we have defined a unique embedding for G.

To prove the main theorem, we first define the *incidence vector* of a combinatorial embedding. Let C be the set of all directed cycles in the graph that are face cycles in at least one combinatorial embedding of the graph. Then the incidence vector of an embedding Γ is given as a vector in $\{0,1\}^{|C|}$, where the components representing the face cycles in Γ have value one and all other components have value zero.

LEMMA 3.3. *Let Γ be a combinatorial embedding of the biconnected planar graph G. Then the incidence vector χ^Γ satisfies all constraints of the ILP we defined.*

Proof. We prove the lemma using induction over the number n of decision nodes in the SPQR-tree T of G. The value $\chi(x_c)$ is the value of the component in χ associated with the variable for cycle c. We do not consider the case $n = 0$, because G is a simple cycle in this case and has only one combinatorial embedding.

1. $n = 1$:
   No splitting of the SPQR-tree is necessary, the ILP is defined directly by G. The variables are defined as the set of all directed cycles that are face cycles in at least one combinatorial embedding of G. If the decision node in T is an R-node, the constraints we have defined form a complete description of the polytope of all embeddings of G and thus the claim is true.
   Otherwise, the decision node is a P-node and the claim follows from the fact that the embeddings of G correspond to the tours in an ATSP problem and thus satisfy the degree constraints and subtour elimination constraints.
2. $n > 1$:
   From the previous lemma we know that Γ uniquely defines embeddings Γ_i with incidence vectors χ_i for the split graphs G_i. We will use the induction basis to show that χ^Γ satisfies all lifted constraints. We know that the choice constraints are satisfied by χ^Γ because in any embedding there can be only one cycle passing a certain split pair in the same direction (see section 3.2 where we defined the choice constraints).
   Let u be a constraint computed for a split graph G_i and c be a component cycle of G_i that is represented by the variable x_c in u. When we lift u, we replace x_c by the sum $\sum_{c' \in L_c} x_{c'}$, where L_c is the set of global cycles in G generated using c. Since the choice constraints are satisfied, this sum is either 0 or 1. Using the fact that the choice constraints are satisfied and by construction of the χ_i from χ^Γ, we can show that

$$\sum_{c' \in L_c} \chi(x_{c'}) = \chi_i(x_c).$$

   Therefore, all lifted constraints are satisfied. This is also true for the lifted subtour elimination constraints.
   To see that the center graph constraint is satisfied, we observe that we can construct any embedding of G from an embedding Γ_0 of G_0 by replacing edges by the embeddings of subgraphs. The global face cycles in Γ are represented by face cycles in Γ_0 and each local face cycle in G_0 is also a face cycle in Γ_0. Therefore the center graph constraint is satisfied.

It follows that every embedding of G satisfies the constraints of the ILP. □

LEMMA 3.4. *Let G be a biconnected planar graph and $\chi \in \{0,1\}^{|C|}$ a vector satisfying all constraints of the ILP including the lifted subtour elimination constraints. Then χ is the incidence vector of a combinatorial embedding Γ of G.*

Proof. Again, we use induction on the number n of decision nodes in the SPQR-tree T of G and we disregard the case $n = 0$.

1. $n = 1$:
   Like in the previous lemma, our claim holds by definition of the ILP.
2. $n > 1$:
   The proof works in two stages: First we construct vectors χ_i for each split graph from χ and prove that these vectors satisfy the ILPs for the G_i, and are therefore incidence vectors of embeddings Γ_i of the G_i by induction basis. In the second stage, we use the Γ_i to construct an embedding Γ for G and show that χ is the incidence vector of Γ.
   The construction of the χ_i works as follows: When x_c is a variable in the ILP of G_i and the corresponding cycle c is a local cycle, then x_c in the ILP of G_i is defined as the value of x_c in χ. Otherwise, if c is a component cycle, we define the value of x_c as the sum of the values of all variables in χ representing global cycles constructed using c,

$$\chi_i(x_c) := \sum_{c' \in L_c} \chi(x_{c'}) \; .$$

   This value is either 0 or 1 because χ satisfies the choice constraints.
   Because χ satisfies the lifted constraints, each χ_i constructed in this way must satisfy the constraints of the ILP for G_i and by induction basis we know that each χ_i represents an embedding Γ_i of G_i. By combining the embeddings χ_i for the split graphs, we construct an embedding Γ for G.
   To show that χ is the incidence vector of Γ, we define χ^Γ as the incidence vector of Γ and show that χ and χ^Γ are identical. By construction of Γ and χ^Γ, the components in χ^Γ and χ corresponding to local cycles must be equal. The number of global cycles whose variable in χ has value 1 must be equal to the number of faces in Γ consisting of such cycles. This is guaranteed by the center graph constraint. Using the fact that for all face cycles in Γ_0 that contain a virtual edge (the embedding of the center split graph) there must be a represented global cycle in G whose component in χ and in χ^Γ is 1 we can show that both vectors also agree on the values of the variables of the global cycles, and thus must be identical.

It follows that a vector that satisfies all constraints of the ILP is the incidence vector of a combinatorial embedding. □

A more detailed version of the proof for Theorem 3.1 can be found in [22].

4. The linear program describing orthogonal representations for a fixed embedding. Orthogonal representations not only fix the planar embedding of a graph but also the number, type, and sequence of the bends on each edge and the angles between edges incident to the same vertex in an orthogonal drawing. However, they do not fix the lengths of the edge segments in the drawing. The first efficient algorithm for computing an orthogonal representation of a graph with the minimum number of bends for a fixed planar embedding was presented by Tamassia [21]. This algorithm constructs a flow network using the planar embedding and then computes a minimum cost flow in this network. This flow can be translated into an orthogonal representation of the graph with the minimum number of bends for the fixed embedding.

The drawback of the original method of Tamassia is that it cannot deal with vertices of degree greater than four. Some modifications of the algorithm have been

published that overcome this constraint. The approach we use implements the *podevsnef* drawing convention (planar orthogonal drawings with equal vertex size and nonempty faces) defined in [13]. According to this convention, the vertices are drawn as squares of equal size and the edges are positioned on a finer grid than the vertices. Because of this modification, more than one edge can be incident to each of the four sides of a vertex (see Figure 1 on page 666 for an example of a podevsnef drawing).

Bertolazzi, Di Battista, and Didimo describe a minimum cost flow network N that can be used to compute an orthogonal representation in a simplified podevsnef model with the minimum number of bends for a fixed embedding [2]. The model is simplified, because whenever edges run parallel, the first bend of the rightmost edge in the bundle is a bend to the right. Another property of the simplified model is that a vertex with degree at least four always has at least one edge attached to each side. This is the case in Figure 1 on page 666. We have chosen this model for our approach because it is the only podevsnef model we are aware of that can be modeled as a standard min-cost-flow problem.

The network N for G contains one node for every vertex of G (called *v-nodes*) and one node for every face cycle of the given embedding (called *f-nodes*). The basic idea of the network is that the flow on its arcs corresponds to bends on edges and to the angles between neighboring edges incident to the same vertex. One unit of flow corresponds to an angle of 90 degrees. The supply of flow assigned to each node and the capacity of the edges guarantee that a feasible flow corresponds to an orthogonal representation of the graph. The costs of the arcs are defined such that the cost of a flow is equal to the number of bends in the corresponding orthogonal representation.

We used this network and transformed it into a linear program. There is one variable f_e for each arc e in the network that represents the amount of flow routed via e. One constraint for each vertex in the network makes sure that the outgoing amount of flow minus the incoming amount is equal to the supply of the node. Some nodes in the network have negative supply, and thus consume flow. We have one constraint for each arc that says that the flow on the arc must be nonnegative and we also introduce upper bounds on the flow on arcs that start or end in a v-node. The objective function minimizes the sum of the amount of flow over each arc multiplied by the cost of the arc. An optimal solution represents a minimum cost flow in N and thus an orthogonal representation with the minimum number of bends.

LP 1 shows the corresponding linear program (LP). The set E_N is the set of arcs in the network, V the set of vertices in G and F the set of faces in the embedding. The face f_o is the outer face of the embedding. The degree of a face is defined as the number of edges that form the boundary of the face. The variable f_e denotes the flow on arc e. The constraints guarantee that the solution corresponds to an orthogonal representation of G.

LP 1

$$\min \quad \sum_{e \in E_N} cost(e) \cdot f_e$$

subject to

$$\sum_{e=(v,w)\in E_N} f_e - \sum_{e=(w,v)\in E_N} f_e = 4 - deg(b^{-1}(v)) \; \forall \text{ nodes } v \text{ with } b^{-1}(v) \neq f_o$$

$$\sum_{e=(b(f_o),w)\in E_N} f_e - \sum_{e=(w,b(f_o))\in E_N} f_e = \; -4 - deg(f_o)$$

$$
\begin{aligned}
f_e &\leq 4 - deg(b^{-1}(v)) \quad \forall\, e = (v, w) \in E_{vf} \\
f_e &\leq 1 \quad \forall\, e \in E_{fv} \\
f_e &\geq 0 \quad \forall\, e \in E_N
\end{aligned}
$$

Since this linear program is a straightforward formulation of the flow network introduced in [2], the correctness follows directly from the correctness of the flow network.

5. The mixed integer linear program describing the set of all orthogonal representations of a graph. The flow network N of the last section describing the set of orthogonal representations of a graph for a fixed embedding contains one f-node for every face of the embedding. When we want to optimize over the set of all embeddings of a graph, we do not know at the beginning which cycles will be face cycles in an optimal solution. Therefore, we construct a new network N', where we have one c-node for every directed cycle in the graph, that is a face cycle in at least one embedding. These nodes play a similar role to the f-nodes in the linear program for a fixed embedding we presented in the previous section. The set of cycles that are face cycles in at least one embedding of the graph corresponds to the set of variables in our ILP from section 3 that describes the set of all embeddings of a graph.

In a solution of the embedding ILP, the variable corresponding to a cycle has value one if this cycle is a face cycle in the embedding represented by the solution and zero otherwise. The capacities of the arcs in the network N' depend on the values of the cycle variables. If the cycle-variables represent an embedding Γ of the graph, then the set of feasible flows in N' corresponds to the set of feasible flows in the network N constructed for embedding Γ. Let A be the set of arcs incident to the c-node for cycle c in N' and the variable for c in the embedding ILP be zero. Then all arcs in A must have capacity zero. This has the same effect on the flow as removing the c-node from the network.

We first compute the capacities of the arcs and the demand of each c-node analogously to the corresponding values for the f-nodes in the network N. Then we multiply the amount of flow produced or consumed by a c-node with the value of the corresponding variable in the ILP. This ensures that vertices in N' that correspond to cycles in G that are not face cycles do not produce or consume flow.

Any arc that starts or ends at a c-node has capacity zero if the c-node corresponds to a cycle whose ILP-value is zero. If the capacity of the arc is limited even if the corresponding cycle is a face cycle, we can just multiply this limit with the ILP-value of the cycle to achieve this behavior. But the arcs in the network N that connect two f-nodes have unlimited capacity. However, we can easily compute an upper bound f_{max} for the flow produced in the whole network N, by computing the sum of all positive supplies in the network. The resulting value can be used as the upper bound on the flow on any arc. For each arc a in N' connecting two c-nodes v_1 and v_2, we set the capacity to the minimum of $f_{max}x_1$ and $f_{max}x_2$, where x_i for $i \in \{1, 2\}$ is the binary variable in the embedding ILP for the cycle corresponding to node v_i. This guarantees that the flow on a is zero if at least one of the cycles represented by the nodes v_i is not a face cycle in the represented embedding.

In this way, the capacities of the arcs and the amount of flow produced and consumed by the vertices in N' depend on the values of the cycle variables in the ILP. We transform N' into a linear program and merge it with the ILP that represents the embeddings of the graph. To correctly set the supply and demand of the c-nodes,

we need to know the cycle chosen as the outer face cycle. Therefore, we introduce an outer face variable for each cycle and add constraints that guarantee that exactly one of the cycles chosen as face cycles is chosen as the outer face cycle. The result is a mixed integer linear program, where an optimal solution corresponds to an orthogonal representation with the minimum number of bends over the set of *all* embeddings of the input graph.

MILP 1 is the resulting mixed integer linear program (MILP). We omitted the constraints that define the embedding because they are defined recursively. Again, variable f_e denotes the flow on arc e. The set C is the set of cycles in G that are face cycles in at least one embedding. For each of these cycles c, the variable x_c is one if c is a face cycle and variable o_c is one if c is the outer face cycle. The set E_{cc} is the set of arcs that connect two c-nodes. Arcs in E_{vc} start in a v-node and end in a c-node while the arcs in E_{cv} start in a c-node and end in a v-node. The expression $len(c)$ denotes the number of edges in cycle c. The function b is the bijection from the vertices of G to the v-nodes in N' and from the directed cycles in C to the c-nodes of the network.

MILP 1

$$\min \quad \sum_{e \in E_N} cost(e) \cdot f_e$$

subject to

$$\sum_{c \in C} o_c = 1 \tag{5.1}$$

$$x_c - o_c \geq 0 \qquad \forall\, c \in C \tag{5.2}$$

$$\sum_{e=(v,w) \in E_N} f_e - \sum_{e=(w,v) \in E_N} f_e = 4 - deg(v) \qquad \forall\, v \in V \tag{5.3}$$

$$\sum_{e=(c,w) \in E_N} f_e - \sum_{e=(w,c) \in E_N} f_e = x_c(4 - len(c)) - 8o_c \forall\, c \in C \tag{5.4}$$

$$f_e \leq x_c(4 - deg(v)) \qquad \forall\, e = (v,c) \in E_{vc} \tag{5.5}$$

$$f_e \leq x_c \qquad \forall\, e = (c,v) \in E_{cv} \tag{5.6}$$

$$f_e \leq x_{c_1} f_{max} \qquad \forall\, e = (c_1,c_2) \in E_{cc} \tag{5.7}$$

$$f_e \leq x_{c_2} f_{max} \qquad \forall\, e = (c_1,c_2) \in E_{cc} \tag{5.8}$$

$$f_e \geq 0 \qquad \forall\, e \in E_N \tag{5.9}$$

$$x_c, o_c \in \{0,1\} \qquad \forall\, c \in C \tag{5.10}$$

Constraint 5.1 says that there is exactly one cycle chosen as the outer face cycle and the constraints of type 5.2 guarantee that this cycle is also chosen as a face cycle. Constraint 5.3 is the same as in the LP of the previous section because the supply of the nodes representing vertices of G is independent of the chosen embedding. A node in N' corresponding to a cycle that is not chosen as a face cycle does not consume or supply any flow since its cycle variable and its outer face variable are both zero. This is guaranteed by constraint 5.4. If the outer face variable of a cycle is one, the constraint sets its supply to $-4 - len(c)$ and if the cycle variable is one but the outer face variable is zero to $4 - len(c)$.

The constraints of type 5.5, 5.6, 5.7, and 5.8 make sure that the arcs incident to c-vertices where the corresponding cycle is not a face cycle in the chosen embedding have capacity zero. If a cycle is chosen as face cycle, the arcs incident to the corresponding c-node in N' have either the same capacities as in the network N of the previous section or if the capacity in N is unbounded (for the arcs in E_{cc}), the capacity is now set to an upper bound for the flow in the network.

6. The algorithm for minimizing the number of bends. The algorithm first computes the recursive ILP describing the set of all combinatorial embeddings of the graph. This is done by recursively splitting the SPQR-tree into smaller trees and computing the ILPs for the corresponding split-graphs. Then we compute the set of cycle-variables for the original graph by combining the cycles that are represented as variables in the ILPs for the split graph. This also gives us for each cycle variable in the ILPs of the split graphs the list of corresponding cycle variables in the original graph. Using this information, we can lift the constraints of the ILPs, compute the choice constraints, the center graph constraint, and update the complete directed graphs computed for the P-nodes used for the separation of the subtour elimination constraints.

After this step, we use the set of cycle variables computed for the original graph to compute the network N' and the corresponding MILP. We use CPLEX (version 6.5) to compute an integer solution and then check if there are any violated subtour elimination constraints by computing a minimum cut in the complete directed graphs computed for each P-node skeleton. If we find a violated constraint, we add it to the MILP and reoptimize. When we have found a feasible solution, we transform it into an orthogonal representation of the graph.

To improve the performance of the algorithm, we modified the MILP slightly. We realized that we only need outer face variables for half of the cycles. The reason is that for every cycle c represented by a variable in the embedding ILP, the cycle $\bar{c}$ passing the same edges in the opposite direction is also represented by a variable. The orthogonal representations we exclude by introducing outer face variables only for one direction of each undirected cycle are mirror images of other orthogonal representations that can still be represented. Of course, every orthogonal representation has the same number of bends as its mirror image and therefore we do not exclude all optimal solutions.

The second modification is that we hard-coded a complete description of the set of embeddings for P-node skeletons with less than five vertices into our program to reduce the need for separating subtour elimination constraints.

7. Computational results. To test our approach, we used two sets of benchmark graphs. The first was introduced in [8] and consists of 11,529 graphs that either come from industrial applications or were derived from such graphs by introducing small changes. We call this set the *real world set*. The second set consists of randomly generated graphs that were used in [2] to test the performance of the branch & bound approach for minimizing the number of bends. We call this set the *artificial set*.

The graphs in the artificial set are already biconnected and planar, so we can directly apply our new algorithm for minimizing the number of bends and compare it with the branch & bound algorithm from [2]. The majority of the graphs in the real world set are not planar and biconnected. Therefore, we used a standard approach to transform them into planar biconnected graphs whose drawing can be easily transformed back into a drawing of the original graph.

This is done as follows: We first determine the topology of the graph using the standard planarization method implemented in the AGD-library [1]. This method

computes an embedding of a planar subgraph of the original graph and then inserts the missing edges one by one into the current embedding. The crossings produced by inserting an edge are replaced by artificial vertices with degree four. If the resulting graph is not biconnected, we introduce new edges while maintaining planarity using the heuristic presented in [12] that is also implemented in the AGD-library. Note that these operations can increase the number of vertices and edges of a graph considerably.

If we have computed a drawing for the modified graph, it is straightforward to transform it into a drawing for the original graph. We remove all edges from the drawing that were introduced to make the graph biconnected and replace the drawings of the artificial vertices by crossings. Note that this always works because the simple podevsnef standard guarantees that a vertex with degree four has one incident edge on each side of the vertex. Therefore, we can easily replace the drawing of an artificial vertex with an orthogonal crossing. The result is a drawing of the original graph. Note that because of the artificial vertices and edges introduced before the optimization step, this drawing does not necessarily have the minimum number of bends over all embeddings of the original graph.

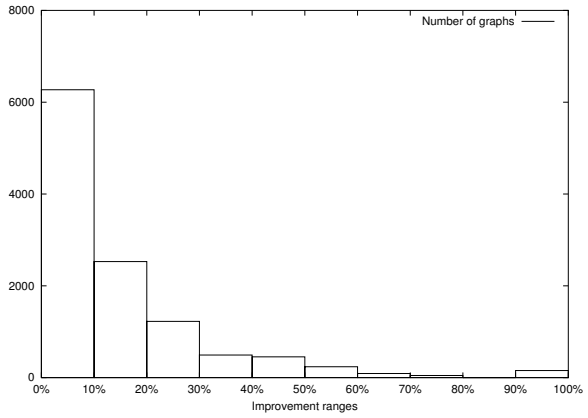
We compared the results produced by our algorithm for minimizing the number of bends over all embeddings to the results computed by the following heuristic: We use the linear planarity test of Hopcroft and Tarjan [16] to compute an embedding for the graph and then compute a minimum cost flow in the network given in section 4. The flow defines an orthogonal representation of the graph with the minimum number of bends for the chosen embedding.

Let h be the number of bends in the orthogonal representation computed by the heuristic and o the number of bends in an orthogonal representation with the minimum number of bends over all embeddings. For each graph in the benchmark sets, we computed the value $\frac{h-o}{h}100\%$. This is the percentage of the improvement we get using an optimal algorithm.

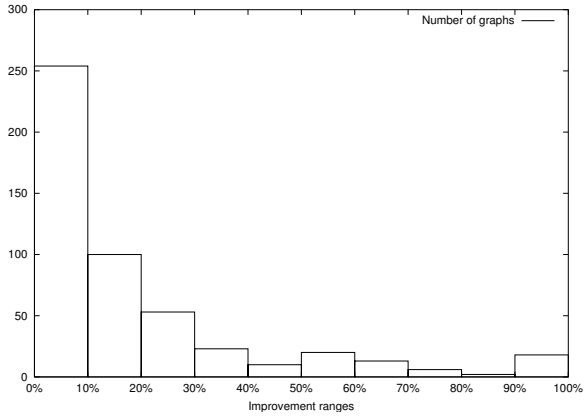
We broke the set of all graphs of each benchmark set into ten subsets. The first subset contained the graphs where the improvement was smaller than ten percent, the second subset contained the graphs where the improvement was at least ten and smaller than 20 percent and so on. The last subset contained the graphs where the improvement was at least 90 percent. Note that the improvement is 100 percent if the heuristic solutions contains bends while the optimum solution contains no bends.

The x-axes in Figure 7 show the improvement ranges while the height of the boxes corresponds to the number of graphs that fall into that range. Figure 7(a) shows the data for the real world graphs while Figure 7(b) shows the data for the artificial graphs. Both diagrams are remarkably similar. For about half of all graphs, optimizing over all embeddings results in a significant reduction of the number of bends (at least 10 percent).

We compared the average running times for graphs with the same number of vertices of our new algorithm (MIX) and the branch & bound algorithm (B&B) from [2]. Both algorithms are written in C++ and were compiled with the flag `-O` using gcc version 2.95.1. The algorithms ran on a Sun Enterprise 450 Model 4400 with four Sun UltraSPARC-II 400 MHz CPUs and 4 GB of memory. Figure 8 shows the corresponding diagrams for the real world graphs (Figure 8(a)) and for the artificial graphs (Figure 8(b)). The x-axes show the number of vertices while the y-axes give the average time in seconds needed for a problem instance. Note that the times given for MIX include the time needed for constructing the recursive ILP, the flow network, and solving the MILP.



(a)



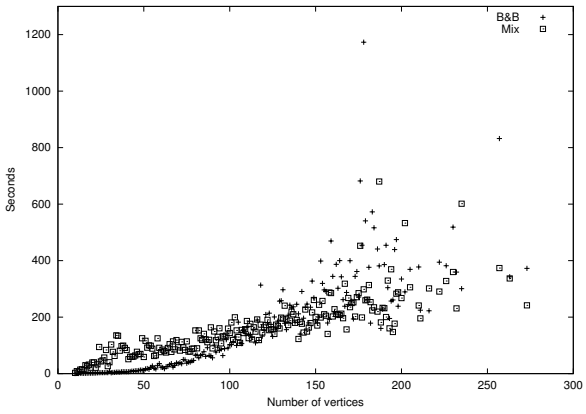
(b)

FIG. 7. *Improvement ranges of our algorithm compared to the heuristic for the real world benchmark set* (a) *and the artificial benchmark set* (b).

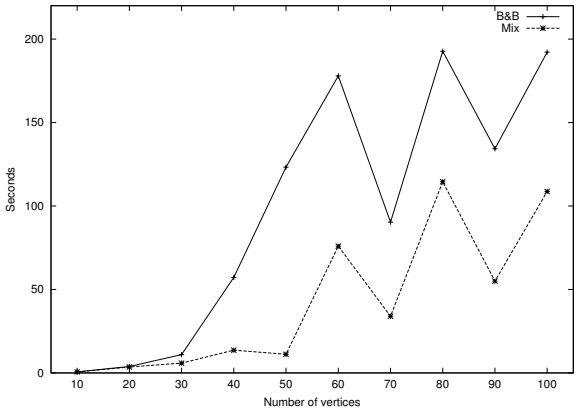
Figure 8(a) shows that B&B is faster for graphs with up to 120 vertices, but for larger graphs our new algorithm is faster. For example, for one graph with 130 vertices and 205 edges, B&B needed 5244 seconds while MIX found an optimal solution in only 108 seconds. The corresponding drawing is shown in Figure 1 on page 666.

There were 197 graphs in the real world benchmark set that B&B could not solve in one hour computation time. Our algorithm could not solve 25 graphs within the same time limit. Another interesting fact is that we only had to add subtour elimination constraints and reoptimize for six graphs out of 11,529. This shows the effect of hard-coding the complete ILP-description for P-nodes with up to four edges. We never had to add more than one subtour elimination constraint.

As Figure 8(b) shows, the speed advantage of our new algorithm is very pronounced for the artificial graphs. The B&B needs on average almost twice as long to



(a)



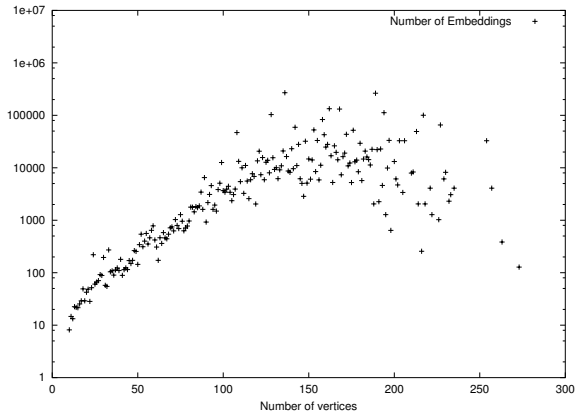
(b)

FIG. 8. *Runtime comparison of our new algorithm with the branch & bound algorithm for the real world benchmark set* (a) *and the artificial benchmark set* (b).

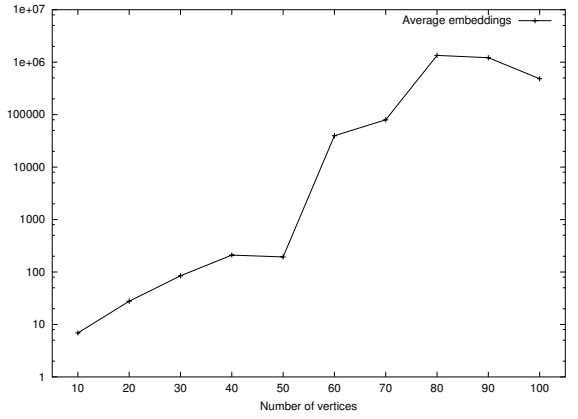
compute an optimal solution compared to our new method.

Figure 9 shows the average number of embeddings for graphs with the same number of vertices for the real world benchmark set and the artificial benchmark set. Again, the x-axes show the number of vertices. The y-axes have a logarithmic scale and show the average number of embeddings for the graphs.

Figure 9(a) shows the number of embeddings of the real world graphs. Until about 150 vertices, the average number of embeddings grows exponentially with the size of the graphs (remember that the y-axis is logarithmic). The reason for the drop in the number of embeddings for larger graphs is the planarization method. Graphs where many edges have to be deleted to obtain a planar graph tend to have large triconnected components after the planarization method is applied because this method replaces crossings with vertices of degree four.



(a)

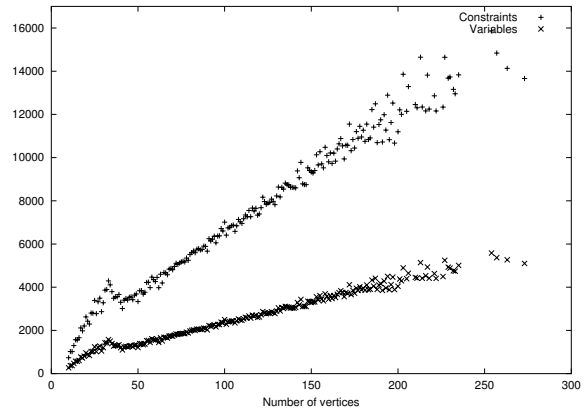


(b)

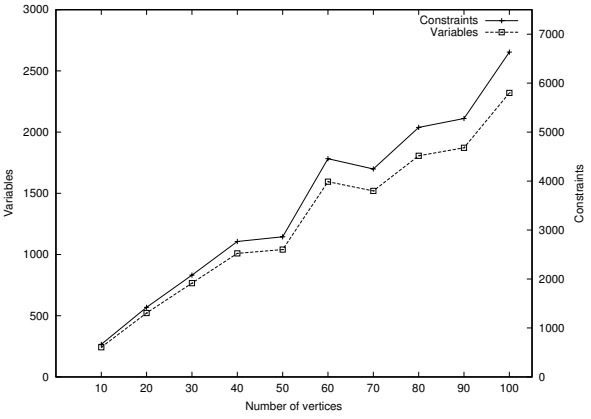
FIG. 9. *The average number of embeddings for graphs in the real world benchmarks set (Figure* 9(a)*) and the artificial benchmark set (Figure* 9(b)*).*

The growth of the average number of embeddings for the artificial graphs is also roughly exponential until about 80 vertices, as Figure 9(b) shows. The number of embeddings varies widely for graphs with the same number of vertices because the set was generated using five different settings for the parameters of the generation algorithm that influence the number of embeddings. The details of the generation algorithm and the parameter settings can be found in [2].

Figure 10 shows the average number of constraints and embeddings in the MILP computed by our algorithm for the real world graphs (Figure 10(a)) and for the artificial graphs (Figure 10(b)). For both benchmarks sets, the number of variables and constraints grows roughly linear with the size of the graphs. This benign growth behavior is in sharp contrast to the exponential growth of the number of embeddings. The spike at 35 vertices in Figure 10(a) mirrors the smaller spike in the number of



(a)



(b)

FIG. 10. *The average number of constraints and variables in the mixed integer linear programs for the real world benchmark set* (a) *and the artificial benchmark set* (b).

embeddings for the real world graphs.

8. Conclusion. Using methods of integer linear programming to minimize the number of bends in an orthogonal drawing seems to be a promising approach. The main drawback is that at the moment, the algorithm can only guarantee optimality for biconnected graphs. The reason is that SPQR-trees are only defined for biconnected graphs. One possible approach to overcome this limitation is to work with the block-cut-tree of biconnected components of the graph. If it can be used to describe the set of all embeddings of a connected graph as an ILP, our approach can be easily extended to deal with any planar graph.

Acknowledgment. We thank Walter Didimo for providing the code of the branch & bound algorithm and the real world benchmark set.

REFERENCES

[1] *AGD User Manual (Version* 1.1*)*, 1999. Universität Wien, Max-Planck-Institut, Saarbrücken, Universität Trier, Trier, Germany, Universität zu Köln, Köln, Germany. See also http://www.mpi-sb.mpg.de/AGD/.

[2] P. Bertolazzi, G. Di Battista, and W. Didimo, *Computing orthogonal drawings with the minimum number of bends*, IEEE Trans. Comput., 49 (2000), pp. 826–840.

[3] D. Bienstock and C. L. Monma, *On the complexity of covering vertices by faces in a planar graph*, SIAM J. Comput., 17 (1988), pp. 53–76.

[4] D. Bienstock and C. L. Monma, *Optimal enclosing regions in planar graphs*, Networks, 19 (1989), pp. 79–94.

[5] D. Bienstock and C. L. Monma, *On the complexity of embedding planar graphs to minimize certain distance measures*, Algorithmica, 5 (1990), pp. 93–109.

[6] G. Carpaneto, M. Dell'Amico, and P. Toth, *Exact solution of large scale asymmetric travelling salesman problems*, ACM Trans. Math. Software, 21 (1995), pp. 394–409.

[7] G. Di Battista, P. Eades, R. Tamassia, and I. G. Tollis, *Graph Drawing*, Prentice Hall, 1999.

[8] G. Di Battista, A. Garg, G. Liotta, R. Tamassia, E. Tassinari, and F. Vargiu, *An experimental comparison of four graph drawing algorithms*, Comput. Geom., 7 (1997), pp. 303–326.

[9] G. Di Battista, G. Liotta, and F. Vargiu, *Spirality and optimal orthogonal drawings*, SIAM J. Comput., 27 (1998), pp. 1764–1811.

[10] G. Di Battista and R. Tamassia, *On-line planarity testing*, SIAM J. Comput., 25 (1996), pp. 956–997.

[11] W. Didimo and G. Liotta, *Computing orthogonal drawings in a variable embedding setting*, in Algorithms and Computation: 9th International Symposium, ISAAC'98, K.-Y. Chwa and O.H. Ibarra, eds., Lecture Notes in Comput. Sci. 1533, Springer-Verlag, Berlin, 1998, pp. 79–88.

[12] S. Fialko and P. Mutzel, *A new approximation algorithm for the planar augmentation problem*, in Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms, San Francisco, CA, 1998, ACM, New York, pp. 260–269.

[13] U. Fössmeier and M. Kaufmann, *Drawing high degree graphs with low bend numbers*, in Graph Drawing (Passau, 1995), F. J. Brandenburg, ed., Lecture Notes in Comput. Sci. 1027, Springer-Verlag, Berlin, 1996, pp. 254–266.

[14] A. Garg and R. Tamassia, *On the computational complexity of upward and rectilinear planarity testing*, SIAM J. Comput., 31 (2001), pp. 601–625.

[15] C. Gutwenger and P. Mutzel, *A linear time implementation of SPQR-trees*, in Graph Drawing (Proc. 2000), J. Marks, ed., Lecture Notes in Comput. Sci. 1984, Springer-Verlag, Berlin, 2001, pp. 77–90.

[16] J. Hopcroft and R. E. Tarjan, *Efficient planarity testing*, J. ACM, 21 (1974), pp. 549–568.

[17] J. E. Hopcroft and R. E. Tarjan, *Dividing a graph into triconnected components*, SIAM J. Comput., 2 (1973), pp. 135–158.

[18] P. Mutzel and R. Weiskircher, *Optimizing over all combinatorial embeddings of a planar graph*, in Proceedings IPCO '99, G. Cornuéjols, R. Burkard, and G. Wöginger, eds., Lecture Notes in Comput. Sci. 1610, Springer-Verlag, Berlin, 1999, pp. 361–376.

[19] P. Mutzel and R. Weiskircher, *Computing optimal embeddings for planar graphs*, in Proceedings COCOON '00, D. Z. Du, P. Eades, V. Estivill-Castro, X. Lin, and A. Sharma, eds., Lecture Notes in Comput. Sci. 1858, Springer-Verlag, Berlin, 2000, pp. 95–104.

[20] P. Mutzel and R. Weiskircher, *Bend minimization in orthogonal drawings using integer programming*, in Proceedings of the 8th Annual International Conference on Computing and Combinatorics (COCOON 2002), O. H. Ibarra and L. Zhang, eds., Lecture Notes in Comput. Sci. 2387, Springer-Verlag, Berlin, 2002, pp. 484–493.

[21] R. Tamassia, *On embedding a graph in the grid with the minimum number of bends*, SIAM J. Comput., 16 (1987), pp. 421–444.

[22] R. Weiskircher, *New Applications of SPQR-Trees in Graph Drawing*, Ph.D thesis, Universität des Saarlandes, Saarbrücken, Germany, 2002.

SIAM J. OPTIM.
Vol. 17, No. 3, pp. 688–710
© 2006 Society for Industrial and Applied Mathematics

A UNIFIED APPROACH AND OPTIMALITY CONDITIONS FOR APPROXIMATE SOLUTIONS OF VECTOR OPTIMIZATION PROBLEMS*

CÉSAR GUTIÉRREZ†, BIENVENIDO JIMÉNEZ‡, AND VICENTE NOVO‡

Abstract. This paper deals with approximate (ε-efficient) solutions of vector optimization problems. We introduce a new ε-efficiency concept which extends and unifies different approximate solution notions introduced in the literature. We obtain necessary and sufficient conditions via nonlinear scalarization, which allow us to study this new class of approximate solutions in a general framework, since any convexity hypothesis is required. Several examples are proposed to show the concepts introduced and the results attained.

Key words. vector optimization, approximate solutions, ε-efficiency, scalarization, gauge functional

AMS subject classifications. 49J52, 90C29, 49M37

DOI. 10.1137/05062648X

1. Introduction.

Approximate solutions are a usual kind of solution used to solve optimization problems. There are two reasons for this statement. First, optimization models are simplified representations of real problems. Second, these models are solved frequently using iterative algorithms or heuristic methods, and these procedures give approximations to the theoretical solution.

In vector optimization, the notion of approximate solution has been defined in several ways. The first and most popular concept was introduced by Kutateladze [12]. This notion has been used to obtain vector variational principles, approximate Kuhn–Tucker type conditions, approximate duality theorems, solution methods, etc. (see [2, 5, 6, 7, 10, 11, 14, 15, 17, 18, 19, 23, 26, 30, 32]).

However, the ε-efficiency set obtained according to Kutateladze's definition is sometimes too large. Thus, several authors have proposed other ε-efficiency concepts (see, for example, [8, 21, 28, 29, 31]). In this paper, all these notions are analyzed through a new concept that allows us to study them simultaneously.

We characterize this new ε-efficiency notion via nonlinear scalarization, i.e., by means of approximate solutions of related nonlinear scalar optimization problems. Necessary conditions are obtained via Minkowski-type functionals, and sufficient conditions are deduced using a new class of monotone functionals.

Our results are general because we do not assume any convexity hypothesis. We consider a nonconvex constrained vector optimization problem and a preference relation to solve it which is not necessarily a preorder relation. This type of preference is usual in economics (see, for example, [20] and references therein).

*Received by the editors March 10, 2005; accepted for publication (in revised form) March 24, 2006; published electronically September 21, 2006. This research was partially supported by the Ministerio de Ciencia y Tecnología (Spain), project BFM2003-02194.

http://www.siam.org/journals/siopt/17-3/62648.html

†Departamento de Matemática Aplicada, Universidad de Valladolid, ETSI Informática, Edificio de Tecnologías de la Información y las Telecomunicaciones, Campus Miguel Delibes, s/n, 47011 Valladolid, Spain (cesargv@mat.uva.es).

‡Departamento de Matemática Aplicada, Universidad Nacional de Educación a Distancia, ETSI Industriales, c/ Juan del Rosal, 12, Ciudad Universitaria, 28040 Madrid, Spain (bjimenez@ind.uned.es, vnovo@ind.uned.es).

The work is structured as follows. In section 2, the vector optimization problem and the preference relation are fixed. Moreover, we describe some notations used in what follows. In section 3, we propose a new ε-efficiency concept and prove some properties of this notion when ε tends to zero. In section 4, we recall several well-known ε-efficiency concepts to show as our concept extends and unifies different notions introduced previously in the literature by Helbig [8], Kutateladze [12], Németh [21], Tanaka [28], and White [31]. In this sense, Helbig's concept is seen as a particular case of a more general notion, which shows the standard nature of our definition. In section 5, we characterize the ε-efficiency set through approximate solutions of scalar optimization problems. The scalarization process is based on the postcomposition of the objective map with a suitable nonlinear scalar functional from the final space. In section 6, the results attained in the previous section are applied to study the finite dimensional case. In particular, several types of ε-efficient solutions are characterized taking into account this special structure. Finally, in section 7, we present some conclusions that summarize this work.

2. Preliminaries. In this work, we consider two topological real linear spaces X and Y. We denote by $\operatorname{int}(C)$, $\operatorname{cl}(C)$, $\operatorname{bd}(C)$, C^c, and $\operatorname{conv}(C)$ the interior, the closure, the boundary, the complement, and the convex hull of a set $C \subset Y$, respectively. The cone generated by a set C is defined as

$$\operatorname{cone}(C) := \bigcup_{\alpha>0} \alpha C,$$

and it is said that $D \subset Y$ is a cone if $\operatorname{cone}(D) = D$. Let us observe that $0 \in \operatorname{cone}(C)$ if and only if $0 \in C$ and thus it is possible that $0 \in D$ or $0 \notin D$ when D is a cone. We say that a set C is solid if $\operatorname{int}(C) \neq \emptyset$, is proper if $\emptyset \neq C \neq Y$, and is pointed if $C \cap (-C) \subset \{0\}$, i.e., if $C \cap (-C) = \{0\}$ when $0 \in C$ and if $C \cap (-C) = \emptyset$ when $0 \notin C$. We denote the nonnegative orthant in $\mathbb{R}^p$ by $\mathbb{R}^p_+$.

The topological dual space of Y is denoted by Y^*. For a cone $D \subset Y$, its positive polar cone (resp., strict positive polar cone) is denoted by D^+ (resp., D^{s+}).

In this paper, we study the vector optimization problem

$$\operatorname{Min}\{f(x) : x \in S\}, \tag{2.1}$$

where $f : X \to Y$ and $S \subset X$, $S \neq \emptyset$. As usual, to solve (2.1) the preference relation $\leq$ defined in Y by a nonempty set $D \subset Y$ is used, which models the preferences stated by the decision-maker:

$$y, z \in Y,\ y \leq z \iff y - z \in -D.$$

We assume that D is a pointed cone. Notice that the relation $\leq$ is not a preorder, since D is not necessarily a convex set.

We recall that $x_0 \in S$ is an efficient solution of (2.1) with respect to D (or an efficient solution for short) if $(f(x_0) - D) \cap f(S) \subset \{f(x_0)\}$. We denote the set of efficient solutions of (2.1) with respect to D by $\operatorname{E}(f, D)$ and with respect to $\operatorname{int}(D)$ by $\operatorname{WE}(f, D)$ (in this case it is assumed that D is solid and these efficient solutions of (2.1) are called weakly efficient solutions).

3. A new concept of approximate efficiency in vector optimization. Next, we introduce a new approximate solution concept for vector optimization problems. This notion is motivated in the following idea: An approximate solution of (2.1)

is every feasible point $x_0 \in S$ such that for every feasible point $x \in S$ whose image $f(x)$ is better than $f(x_0)$, the improvement $f(x_0) - f(x)$ is near zero.

To define this concept, we use a proper pointed co-radiant set $C \subset Y$, i.e., a proper pointed set C such that $\alpha d \in C \; \forall\, d \in C$, $\forall\, \alpha > 1$. Moreover, we assume that C is a solid set and we denote $C(\varepsilon) := \varepsilon C \; \forall\, \varepsilon > 0$ and

$$C(0) := \bigcup_{\varepsilon > 0} C(\varepsilon)\,. \tag{3.1}$$

Let us observe that a pointed cone is a pointed co-radiant set. However, the class of pointed co-radiant sets is wider. For example, given any convex set $A \subset Y$ such that $0 \in A$, the complement A^c is a co-radiant set, and for each continuous linear functional $g \in Y^*$ and $\alpha > 0$, the set $A^c \cap \{y \in Y : g(y) \geq \alpha\}$ is a co-radiant pointed set, which is not a cone. It is easy to check that if C is a nonempty pointed co-radiant set such that $0 \notin \mathrm{cl}(C)$, then $C(\varepsilon) \neq C(0) \; \forall\; \varepsilon > 0$, and $C(0)$ is proper.

LEMMA 3.1.

(i) *$C(\varepsilon)$ is a solid pointed co-radiant set $\forall\, \varepsilon > 0$.*

(ii) *$C(\varepsilon_2) \subset C(\varepsilon_1)$ $\forall\, \varepsilon_1, \varepsilon_2 > 0$, $\varepsilon_1 < \varepsilon_2$.*

(iii) *$C(0)$ is a solid pointed cone.*

Proof. Part (i). As $C(\varepsilon) = \varepsilon C \; \forall\, \varepsilon > 0$, and C is a solid pointed co-radiant set, then $C(\varepsilon)$ is also a solid pointed co-radiant set $\forall\, \varepsilon > 0$.

Part (ii). Let $\varepsilon_1, \varepsilon_2 > 0$, $\varepsilon_1 < \varepsilon_2$, and $y \in C(\varepsilon_2)$. There exists $d \in C$ such that $y = \varepsilon_2 d$. For

$$\alpha := 1 + (\varepsilon_2 - \varepsilon_1)/\varepsilon_1$$

we have that $y = \alpha(\varepsilon_1 d) \in C(\varepsilon_1)$, since $\alpha > 1$ and $C(\varepsilon_1)$ is a co-radiant set. Then, $C(\varepsilon_2) \subset C(\varepsilon_1)$.

Part (iii). It is clear that $C(0) = \mathrm{cone}(C)$, and thus $C(0)$ is a solid cone.

If $y \in C(0) \cap (-C(0))$, then there exist $\delta, \nu > 0$ such that $y \in C(\delta) \cap (-C(\nu))$. Consider $\beta = \min\{\delta, \nu\} > 0$. By parts (i)–(ii) we see that $y \in C(\beta) \cap (-C(\beta)) \subset \{0\}$, and therefore, $C(0)$ is a pointed set. ∎

DEFINITION 3.2. *Let $\varepsilon \geq 0$. We say that a feasible point $x_0 \in S$ is an ε-efficient solution of* (2.1) *with respect to C (or an ε-efficient solution for short) if*

$$(f(x_0) - C(\varepsilon)) \cap f(S) \subset \{f(x_0)\}.$$

We denote by $\mathrm{AE}(f, C, \varepsilon)$ the set of ε-efficient solutions of (2.1) with respect to C.

As C is a solid set, it follows that $\mathrm{int}(C)$ is a nonempty pointed co-radiant set and we can also consider the set of all ε-efficient solutions of (2.1) with respect to $\mathrm{int}(C)$ (or weakly ε-efficient solutions for short):

$$\mathrm{WAE}(f, C, \varepsilon) := \mathrm{AE}(f, \mathrm{int}(C)\,, \varepsilon) = \{x \in S : (f(x) - \mathrm{int}(C)\,(\varepsilon)) \cap f(S) = \emptyset\}.$$

Notice that

$$\mathrm{int}(C)\,(0) := \bigcup_{\varepsilon > 0} \varepsilon\, \mathrm{int}(C) = \bigcup_{\varepsilon > 0} \mathrm{int}(C(\varepsilon)) \tag{3.2}$$

is an open cone, and as C is a proper co-radiant set, it follows that $0 \notin \mathrm{int}(C)\,(\varepsilon)$ $\forall\, \varepsilon \geq 0$.

Remark 3.3. Let us observe that when $\varepsilon = 0$ we have $\mathrm{AE}(f, C, 0) = \mathrm{E}(f, C(0))$ and $\mathrm{WAE}(f, C, 0) = \mathrm{E}(f, \mathrm{int}(C)\,(0))$, since the sets $C(0)$ and $\mathrm{int}(C)\,(0)$ are cones. In

what follows the notations $\mathrm{AE}(f,C,0)$ and $\mathrm{WAE}(f,C,0)$ are used in order to stress that efficient solutions with respect to the preference relations induced in Y by $C(0)$ and $\mathrm{int}(C)\,(0)$ are ε-efficient and weakly ε-efficient solutions of (2.1) with respect to C and precision $\varepsilon = 0$, respectively.

For each $\varepsilon > 0$ it follows that $\mathrm{AE}(f,C,\varepsilon) \subset \mathrm{WAE}(f,C,\varepsilon)$. Moreover, as $\mathrm{int}(C)\,(0) \subset \mathrm{int}(C(0)) \subset C(0)$, we have that

$$\mathrm{AE}(f,C,0) \subset \mathrm{WE}(f,C(0)) \subset \mathrm{WAE}(f,C,0)\,.$$

Theorem 3.4 shows several properties of the family $\{\mathrm{AE}(f,C,\varepsilon)\}_{\varepsilon\geq 0}$. As usual, for a set $K \subset Y$ we denote $f^{-1}(K) = \{x \in X : f(x) \in K\}$.

THEOREM 3.4.

(i) $\mathrm{AE}(f,C,0) \subset \mathrm{AE}(f,C,\varepsilon)\ \forall\,\varepsilon > 0$.

(ii) $\mathrm{AE}(f,C,\varepsilon_1) \subset \mathrm{AE}(f,C,\varepsilon_2)\ \forall\,\varepsilon_1,\varepsilon_2 > 0$, $\varepsilon_1 < \varepsilon_2$.

(iii) $\bigcap_{\varepsilon>0} \mathrm{AE}(f,C,\varepsilon) = \mathrm{AE}(f,C,0)$.

(iv) *Let* $(x_n) \subset S$, $(\varepsilon_n) \subset \mathbb{R}_+$, *and* $y \in Y$ *such that* $x_n \in \mathrm{AE}(f,C,\varepsilon_n)$, $\varepsilon_n \downarrow 0$, *and* $f(x_n) \to y$. *Then* $f^{-1}(y) \cap S \subset \mathrm{WAE}(f,C,0)$.

(v) *Let* $(x_n) \subset S$ *and* $(\varepsilon_n) \subset \mathbb{R}_+$ *such that* $x_n \in \mathrm{AE}(f,C,\varepsilon_n)$ *and* $\varepsilon_n \downarrow 0$. *Consider*

$$K := \bigcap_n (f(x_n) - C(\varepsilon_n)).$$

Then $f^{-1}(K) \cap S \subset \mathrm{AE}(f,C,0)$.

Proof. Part (i). Let $\varepsilon > 0$ and $x \in \mathrm{AE}(f,C,0)$. It follows that

$$(f(x) - C(\varepsilon)) \cap f(S) \subset \left(f(x) - \bigcup_{\delta>0} C(\delta)\right) \cap f(S) = (f(x) - C(0)) \cap f(S) \subset \{f(x)\}$$

and thus $x \in \mathrm{AE}(f,C,\varepsilon)$.

Part (ii). Let $\varepsilon_1,\varepsilon_2 > 0$, $\varepsilon_1 < \varepsilon_2$, and $x \in \mathrm{AE}(f,C,\varepsilon_1)$. By Lemma 3.1(ii) we have that $C(\varepsilon_2) \subset C(\varepsilon_1)$ and we deduce that

$$(f(x) - C(\varepsilon_2)) \cap f(S) \subset (f(x) - C(\varepsilon_1)) \cap f(S) \subset \{f(x)\}.$$

Then $x \in \mathrm{AE}(f,C,\varepsilon_2)$.

Part (iii). From part (i) it follows that

$$\mathrm{AE}(f,C,0) \subset \bigcap_{\varepsilon>0} \mathrm{AE}(f,C,\varepsilon)\,.$$

Conversely, let $x \in \bigcap_{\varepsilon>0} \mathrm{AE}(f,C,\varepsilon)$. Then, for each $\varepsilon > 0$ we have

$$(f(x) - C(\varepsilon)) \cap f(S) \subset \{f(x)\}$$

and

$$(f(x) - C(0)) \cap f(S) = \bigcup_{\varepsilon>0} ((f(x) - C(\varepsilon)) \cap f(S)) \subset \{f(x)\}.$$

Therefore, $x \in \mathrm{AE}(f,C,0)$.

Part (iv). Let $x \in f^{-1}(y) \cap S$ and suppose that there exists $z \in S$ such that $f(z) \in f(x) - \operatorname{int}(C)\,(0)$. From (3.2) it follows that there exists $\varepsilon > 0$ verifying $f(z) \in f(x) - \operatorname{int}(C(\varepsilon))$. As $f(x_n) \to y$ we deduce that there exists $n_0 \in \mathbb{N}$ such that

$$f(z) + y - f(x_n) \in f(x) - \operatorname{int}(C(\varepsilon)) \quad \forall\, n \geq n_0.$$

As $\varepsilon_n \downarrow 0$, it follows from Lemma 3.1(ii) that there exists $n_1 \geq n_0$ such that

$$f(z) \in f(x_n) - \operatorname{int}(C(\varepsilon_n)) \quad \forall\, n \geq n_1,$$

and this relation contradicts the weak ε_n-efficiency of x_n, taking into account that $\mathrm{AE}(f, C, \varepsilon_n) \subset \mathrm{WAE}(f, C, \varepsilon_n)$. This finishes the proof of Part (iv).

Part (v). Consider $x \in f^{-1}(K) \cap S$. As $f(x) \in K$ and $x_n \in \mathrm{AE}(f, C, \varepsilon_n)$, we have that

$$f(x) \in (f(x_n) - C(\varepsilon_n)) \cap f(S) \subset \{f(x_n)\} \quad \forall\, n$$

and we deduce that $f(x) = f(x_n)\ \forall\, n$. Therefore,

$$(f(x) - C(\varepsilon_n)) \cap f(S) = (f(x_n) - C(\varepsilon_n)) \cap f(S) \subset \{f(x_n)\} = \{f(x)\} \quad \forall\, n.$$

Thus, by (3.1) we see that

$$(f(x) - C(0)) \cap f(S) \subset \{f(x)\}$$

and we conclude that $x \in \mathrm{AE}(f, C, 0)$. □

Remark 3.5. From Theorem 3.4(iv) it is clear that, if f is a continuous map at $x_0 \in S$ and there exist $(x_n) \subset S$ and $(\varepsilon_n) \subset \mathbb{R}_+$ such that $x_n \in \mathrm{AE}(f, C, \varepsilon_n)$, $x_n \to x_0$, and $\varepsilon_n \downarrow 0$, then $x_0 \in \mathrm{WAE}(f, C, 0)$.

Remark 3.6. As $\operatorname{int}(C)$ is also a solid pointed co-radiant set, we see that Theorem 3.4 holds if we change C by $\operatorname{int}(C)$ and $\mathrm{AE}(f, C, \varepsilon)$ by $\mathrm{WAE}(f, C, \varepsilon)$. In this case, let us observe that the conclusion in Part (iv) is $f^{-1}(y) \cap S \subset \mathrm{WAE}(f, \operatorname{int}(C)\,, 0) = \mathrm{WAE}(f, C, 0)$ since $\operatorname{int}(\operatorname{int}(C)) = \operatorname{int}(C)$.

4. Relations with other ε-efficiency concepts. In this section, we obtain various well-known ε-efficiency concepts by considering suitable (not necessarily convex) sets C in Definition 3.2. Let us observe that in subsection 4.3 we give a new ε-efficiency concept in the sense of Helbig.

4.1. ε-efficiency in the senses of Kutateladze and Németh. Let $D \subset Y$ be a solid pointed convex cone. Suppose that $0 \in D$ and consider $C := H + D$, where $H \subset D \backslash \{0\}$. C is a pointed set, since $C \subset D$ and D is a pointed cone. For each $q \in H$ it is clear that $q + D$ is a solid co-radiant set, since $\operatorname{int}(q + D) = q + \operatorname{int}(D)$ and

$$\alpha(q + D) = q + ((\alpha - 1)q + \alpha D) \subset q + D \quad \forall\, \alpha > 1.$$

Then, writing the set C as

$$C = \bigcup_{q \in H} (q + D),$$

we see that C is a solid co-radiant set and

$$C(\varepsilon) = \bigcup_{q \in H} \varepsilon(q + D) = \bigcup_{q \in H} (\varepsilon q + D) = \varepsilon H + D \quad \forall\, \varepsilon > 0. \tag{4.1}$$

Thus, from Definition 3.2 an ε-efficiency notion can be deduced by taking $C = H + D$. With this notion, for each $\varepsilon > 0$ the following ε-efficiency set is obtained:

$$x \in \mathrm{AE}(f, C, \varepsilon) \iff x \in S, \quad (f(x) - \varepsilon H - D) \cap f(S) \subset \{f(x)\}. \tag{4.2}$$

As $0 \notin C$, for each $\varepsilon > 0$ statement (4.2) becomes

$$x \in \mathrm{AE}(f, C, \varepsilon) \iff x \in S, \quad (f(x) - \varepsilon H - D) \cap f(S) = \emptyset. \tag{4.3}$$

This notion was introduced by Németh [21]. For each $\varepsilon \geq 0$, the set of all ε-efficient (resp., weakly ε-efficient) solutions in this sense, i.e., with respect to $C = H + D$ (resp., $C = \mathrm{int}(H + D)$), is denoted by $\mathrm{AE}(f, C_N, \varepsilon)$ (resp., $\mathrm{WAE}(f, C_N, \varepsilon)$).

If $H = \{q\}$, $\varepsilon > 0$, and $q \in D\backslash\{0\}$, then from (4.3) we obtain the following ε-efficiency notion:

$$x \in \mathrm{AE}(f, C, \varepsilon) \iff x \in S, \quad (f(x) - \varepsilon q - D) \cap f(S) = \emptyset.$$

This concept was introduced by Kutateladze [12], and it is the most popular notion of ε-efficiency (see [31, 25, 9, 35] for more details about it). We denote the set of ε-efficient (resp., weakly ε-efficient) solutions of (2.1) in this sense by $\mathrm{AE}(f, C_K, \varepsilon)$ (resp., $\mathrm{WAE}(f, C_K, \varepsilon)$).

Some properties of Németh's approximate solutions are collected in Proposition 4.2. The following lemma is necessary.

LEMMA 4.1.

(i) $H \subset \mathrm{int}(D) \iff C(0) = \mathrm{int}(D)$.

(ii) $\mathrm{bd}(D) \cap (D\backslash\{0\}) \subset \mathrm{cone}(H) \Rightarrow C(0) = D\backslash\{0\}$.

(iii) $\mathrm{int}(C)\,(0) = \mathrm{int}(D)$.

Proof. Part (i). Suppose that $H \subset \mathrm{int}(D)$. As D is a solid convex cone, from (4.1) we have

$$C(\varepsilon) = \varepsilon H + D \subset \varepsilon\,\mathrm{int}(D) + D \subset \mathrm{int}(D) \quad \forall\, \varepsilon > 0,$$

and $C(0) \subset \mathrm{int}(D)$. Reciprocally, let $d \in \mathrm{int}(D)$ and consider $q \in H$. Then, there exists $\varepsilon > 0$ such that $d - \varepsilon\, q \in D$. It follows that $d \in \varepsilon q + D \subset \varepsilon H + D$ and

$$\mathrm{int}(D) \subset \bigcup_{\varepsilon > 0} C(\varepsilon) = C(0)\,.$$

Next, consider $C(0) = \mathrm{int}(D)$. Then, taking $\varepsilon = 1$ in (4.1) we deduce that $H \subset H + D \subset C(0) = \mathrm{int}(D)$.

Part (ii). If $\mathrm{bd}(D) \cap (D\backslash\{0\}) = \emptyset$, then $D\backslash\{0\}$ is an open set and $H \subset \mathrm{int}(D)$. Thus, by part (i), we have that $C(0) = \mathrm{int}(D) = D\backslash\{0\}$.

Suppose that $\mathrm{bd}(D) \cap (D\backslash\{0\}) \neq \emptyset$. From (4.1) we see that $C(0) = \bigcup_{\varepsilon > 0} C(\varepsilon) \subset D\backslash\{0\}$, since D is a pointed convex cone.

Reciprocally, by the hypothesis,

$$\mathrm{bd}(D) \cap (D\backslash\{0\}) \subset \mathrm{cone}(H) = \bigcup_{\alpha > 0} \alpha\, H \subset C(0)\,. \tag{4.4}$$

Let $d \in \mathrm{int}(D)$ and take a point $d_1 \in \mathrm{bd}(D) \cap (D\backslash\{0\})$. There exist $d_2 \in D$ and $\lambda \in (0, 1)$ such that $d = \lambda d_1 + (1 - \lambda) d_2$. From (4.4) we deduce that there exist $q \in H$ and $\alpha > 0$ such that $d_1 = \alpha q$ and thus

$$d = \lambda(\alpha q) + (1 - \lambda) d_2 \in \lambda \alpha H + D \subset C(0)\,.$$

Thus, $\operatorname{int}(D) \subset C(0)$. From this inclusion and (4.4) it follows that $D\backslash\{0\} \subset C(0)$.

Part (iii). As $H \subset D$ and D is a solid convex cone,

$$\operatorname{int}(C)(0) = \bigcup_{\varepsilon>0} \varepsilon \operatorname{int}(H + D) \subset \operatorname{int}(D).$$

Let $d \in \operatorname{int}(D)$. Taking a point $q \in H$, there exists $\varepsilon > 0$ such that $d - \varepsilon q \in \operatorname{int}(D)$. Therefore, $d \in \operatorname{int}(\varepsilon q + D) \subset \varepsilon \operatorname{int}(H + D)$, and we conclude that $\operatorname{int}(D) \subset \operatorname{int}(C)(0)$. □

PROPOSITION 4.2.

(i) *If* $H \subset \operatorname{int}(D)$, *then*

$$\bigcap_{\varepsilon>0} \mathrm{AE}(f, C_N, \varepsilon) = \mathrm{WE}(f, D),$$

and if $\operatorname{bd}(D) \cap (D\backslash\{0\}) \subset \operatorname{cone}(H)$, *then*

$$\bigcap_{\varepsilon>0} \mathrm{AE}(f, C_N, \varepsilon) = \mathrm{E}(f, D).$$

(ii) *Let* $(x_n) \subset S$, $(\varepsilon_n) \subset \mathbb{R}_+$, *and* $y \in Y$ *such that* $x_n \in \mathrm{AE}(f, C_N, \varepsilon_n)$, $\varepsilon_n \downarrow 0$, *and* $f(x_n) \to y$. *Then* $f^{-1}(y) \cap S \subset \mathrm{WE}(f, D)$.

(iii) *Let* $(x_n) \subset S$ *and* $(\varepsilon_n) \subset \mathbb{R}_+$ *such that* $x_n \in \mathrm{AE}(f, C_N, \varepsilon_n)$ *and* $\varepsilon_n \downarrow 0$. *Consider*

$$K := \bigcap_n (f(x_n) - \varepsilon_n H - D).$$

If $H \subset \operatorname{int}(D)$, *then* $f^{-1}(K) \cap S \subset \mathrm{WE}(f, D)$ *and if* $\operatorname{bd}(D) \cap (D\backslash\{0\}) \subset \operatorname{cone}(H)$, *then* $f^{-1}(K) \cap S \subset \mathrm{E}(f, D)$.

Proof. If $H \subset \operatorname{int}(D)$, then $C(0) = \operatorname{int}(D)$ by Lemma 4.1(i), and so $\mathrm{AE}(f, C_N, 0) = \mathrm{WE}(f, D)$. Moreover, from Lemma 4.1(iii), it follows that $\operatorname{int}(C)(0) = \operatorname{int}(D)$, and we have $\mathrm{WAE}(f, C_N, 0) = \mathrm{WE}(f, D)$. If $\operatorname{bd}(D) \cap (D\backslash\{0\}) \subset \operatorname{cone}(H)$, we deduce from Lemma 4.1(ii) that $C(0) = D\backslash\{0\}$ and $\mathrm{AE}(f, C_N, 0) = \mathrm{E}(f, D)$. Then properties (i)–(iii) hold by Theorem 3.4(iii)–(v). □

In Proposition 4.2 we have extended several properties proved in the literature for the ε-efficiency set in the sense of Kutateladze (see, for example, [9, Lemma 3.3 and Theorem 3.4]) to the approximate solutions in the sense of Németh. We can deduce these properties by considering $H = \{q\}$ in Proposition 4.2.

4.2. ε-efficiency in the sense of White. Let D be a proper solid convex cone such that $0 \in D$. Assume that $\operatorname{cl}(D)$ is a pointed set, consider $q \in \operatorname{int}(D)$, and define

$$C := D \cap (q - D)^c.$$

LEMMA 4.3.

(i) C *is a solid pointed co-radiant set.*

(ii) $C(\varepsilon) = D \cap (\varepsilon q - D)^c \ \forall\, \varepsilon > 0$.

(iii) $C(0) = D\backslash\{0\}$ *and* $\operatorname{int}(C)(0) = \operatorname{int}(D)$.

Proof. Part (i). Suppose that $\operatorname{int}(C) = \emptyset$. Then $\operatorname{int}(D) \cap \operatorname{int}((q - D)^c) = \emptyset$; hence, $\operatorname{int}(D) \subset [\operatorname{int}((q - D)^c)]^c = q - \operatorname{cl}(D)$ and $\operatorname{int}(D) \subset \operatorname{int}(q - \operatorname{cl}(D)) = q - \operatorname{int}(D)$, since D is a solid convex cone. Thus, $0 \in \operatorname{int}(D)$, and it follows that $D = Y$, which is a contradiction.

C is pointed since $C \subset \mathrm{cl}(D)$ and $\mathrm{cl}(D)$ is a pointed set.

Let $d \in C$ and $\alpha > 1$. It is clear that $\alpha\, d \in D$. Moreover, if $\alpha\, d \in q - D$, then there exists $v \in D$ such that $\alpha\, d = q - v$ and

$$d = q - (v + (\alpha - 1)d).$$

As D is a convex cone and $\alpha > 1$, we have $d \in q - D$, which is a contradiction. Then $\alpha\, d \notin q - D$ and therefore $\alpha\, d \in C$.

Part (ii). Let $\varepsilon > 0$ and $z \in D \cap (\varepsilon q - D)^c$. As D is a cone, we have that $(1/\varepsilon)z \in D$. Moreover, $(1/\varepsilon)z \in (q - D)^c$. Indeed, if $(1/\varepsilon)z \in q - D$, then $z \in \varepsilon q - D$, which is absurd. Thus, $(1/\varepsilon)z \in D \cap (q - D)^c$, and it follows that

$$D \cap (\varepsilon q - D)^c \subset \varepsilon(D \cap (q - D)^c) = \varepsilon C = C(\varepsilon)\,.$$

From here we deduce that

$$\varepsilon(D \cap (q - D)^c) = \varepsilon(D \cap ((1/\varepsilon)(\varepsilon q) - D)^c) \subset D \cap (\varepsilon q - D)^c,$$

and we conclude that $C(\varepsilon) = D \cap (\varepsilon q - D)^c$.

Part (iii). First, let us see that

$$\bigcap_{\varepsilon > 0} (\varepsilon q - D) = -\mathrm{cl}(D)\,. \tag{4.5}$$

As $\mathrm{cl}(-D) + \mathrm{int}(-D) \subset \mathrm{int}(-D)$, it follows that $\mathrm{cl}(-D) - \varepsilon q \subset \mathrm{int}(-D)\ \forall \varepsilon > 0$. Hence,

$$-\mathrm{cl}(D) \subset \varepsilon q + \mathrm{int}(-D) \subset \varepsilon q - D \quad \forall\, \varepsilon > 0.$$

Therefore, $-\mathrm{cl}(D) \subset \bigcap_{\varepsilon > 0}(\varepsilon q - D)$. Now, we prove the reciprocal inclusion. If $y \in \bigcap_{\varepsilon > 0}(\varepsilon q - D)$, then $\forall\, \varepsilon > 0$ there exists $d_\varepsilon \in D$ such that $y = \varepsilon q - d_\varepsilon$. Hence,

$$\lim_{\varepsilon \downarrow 0}(-d_\varepsilon) = \lim_{\varepsilon \downarrow 0}(y - \varepsilon q) = y \in -\mathrm{cl}(D)\,,$$

and we have (4.5).

As $\mathrm{cl}(D)$ is a pointed cone, it follows that $D \backslash \{0\} \subset (-\mathrm{cl}(D))^c$ and, taking into account (4.5),

$$C(0) = \bigcup_{\varepsilon > 0} D \cap (\varepsilon q - D)^c = D \cap \left(\bigcap_{\varepsilon > 0} (\varepsilon q - D) \right)^c = D \cap (-\mathrm{cl}(D))^c = D \backslash \{0\}.$$

Analogously, we have

$$\mathrm{int}(C)\,(0) = \bigcup_{\varepsilon > 0} \mathrm{int}(D \cap (\varepsilon q - D)^c) = \mathrm{int}(D) \cap \left(\bigcup_{\varepsilon > 0} \mathrm{int}((\varepsilon q - D)^c) \right)$$

$$= \mathrm{int}(D) \cap \left(\bigcup_{\varepsilon > 0} (\varepsilon q - \mathrm{cl}(D))^c \right) = \mathrm{int}(D) \cap (-\mathrm{cl}(D))^c = \mathrm{int}(D)\,. \qquad \square$$

By Definition 3.2, for each $\varepsilon \geq 0$, the following ε-efficiency concept is obtained:

$$x \in \mathrm{AE}(f, C, \varepsilon) \iff x \in S,\ (f(x) - (D \cap (\varepsilon q - D)^c)) \cap f(S) = \emptyset$$

$$\iff x \in S,\ f(z) \in f(x) - \varepsilon q + D \qquad \forall\, z \in S \text{ such that } f(z) \in f(x) - D.$$

This notion was introduced by White [31]. We denote the set of all ε-efficient (resp., weakly ε-efficient) solutions in this sense by $\mathrm{AE}(f, C_W, \varepsilon)$ (resp., $\mathrm{WAE}(f, C_W, \varepsilon)$). Its elements satisfy the following properties.

PROPOSITION 4.4.

(i) $\bigcap_{\varepsilon>0} \mathrm{AE}(f, C_W, \varepsilon) = \mathrm{E}(f, D)$.

(ii) *Let* $(x_n) \subset S$, $(\varepsilon_n) \subset \mathbb{R}_+$, *and* $y \in Y$ *such that* $x_n \in \mathrm{AE}(f, C_W, \varepsilon_n)$, $\varepsilon_n \downarrow 0$, *and* $f(x_n) \to y$. *Then* $f^{-1}(y) \cap S \subset \mathrm{WE}(f, D)$.

(iii) *Let* $(x_n) \subset S$ *and* $(\varepsilon_n) \subset \mathbb{R}_+$ *such that* $x_n \in \mathrm{AE}(f, C_W, \varepsilon_n)$, $\varepsilon_n \downarrow 0$, *and*

$$K := \bigcap_n (f(x_n) - (D \cap (\varepsilon_n q - D)^c)).$$

Then $f^{-1}(K) \cap S \subset \mathrm{E}(f, D)$.

Proof. By Lemma 4.3(iii) we deduce that $\mathrm{WAE}(f, C_W, 0) = \mathrm{WE}(f, D)$ and $\mathrm{AE}(f, C_W, 0) = \mathrm{E}(f, D)$. Then, the proposition is a consequence of Theorem 3.4(iii)–(v). □

In [35, Propositions 2.6(2) and 2.8], Yokoyama proved parts (i) and (iii) of Proposition 4.4 by considering approximate elements of a set. In Proposition 4.4, we have extended these results to vector optimization problems and we have given an additional property.

4.3. A new ε-efficiency in the sense of Helbig. Suppose that D is a solid pointed cone not necessarily convex such that $0 \in D$. Let $g_i \in D^+\backslash\{0\}$, $i = 1, 2, \ldots, m$, and

$$g(y) := \max_{1 \le i \le m} \{g_i(y)\}. \tag{4.6}$$

Let $\mathcal{M}^{s+} = \{h : Y \to \mathbb{R} : h(d) > 0 \ \forall\, d \in D\backslash\{0\}\}$. It is clear that $g \in \mathcal{M}^{s+}$ if $g_i \in D^{s+}$ for some i. However, it is possible that $g \in \mathcal{M}^{s+}$ and $g_i \notin D^{s+} \ \forall\, i$. For example, if $Y = \mathbb{R}^2$, $D = \mathbb{R}^2_+$, $m = 2$, $g_1(y_1, y_2) = y_1$, and $g_2(y_1, y_2) = y_2$, it is obvious that $g_i \notin D^{s+}$, $i = 1, 2$, and $g \in \mathcal{M}^{s+}$.

For each $\alpha \in \mathbb{R}$ we denote

$$[g > \alpha] = \{y \in Y : g(y) > \alpha\}.$$

Consider the set $C := D \cap [g > 1]$.

LEMMA 4.5.

(i) C *is a solid pointed co-radiant set.*

(ii) $C(\varepsilon) = D \cap [g > \varepsilon] \ \forall\, \varepsilon \ge 0$.

(iii) $\mathrm{int}(D) = \mathrm{int}(C)\,(0) \subset C(0) \subset D\backslash\{0\}$.

(iv) *If* $g \in \mathcal{M}^{s+}$, *then* $C(0) = D\backslash\{0\}$.

Proof. Part (i). It is obvious that C is a pointed co-radiant set, and we prove just that C is solid. Indeed, as D is a proper solid cone, there exist $d \in \mathrm{int}(D)$ and $\alpha := g(d) > 0$, since $g_i \in D^+\backslash\{0\}$, and thus $g_i(d) > 0 \ \forall\, i = 1, 2, \ldots, m$. Then $(2/\alpha)d \in \mathrm{int}(D) \cap [g > 1] = \mathrm{int}(C)$, and C is solid.

Part (ii). For $\varepsilon > 0$ it follows easily since D is a cone and g is a positively homogeneous functional. For $\varepsilon = 0$ it is clear.

Part (iii). Let $d \in \mathrm{int}(D)$. As $g_i \in D^+\backslash\{0\}$, we see that $g(d) > 0$. Then, there exists $\varepsilon > 0$ such that $d \in [g > \varepsilon]$ and we deduce that $d \in \mathrm{int}(C)\,(\varepsilon)$. Thus, it follows that $\mathrm{int}(D) \subset \mathrm{int}(C)\,(0)$.

By part (ii) it is obvious that $C(0) \subset D\backslash\{0\}$. Moreover, we deduce that $\operatorname{int}(D) = \operatorname{int}(C)(0)$, since $\operatorname{int}(C)(0)$ is open.

Part (iv). If $g \in \mathcal{M}^{s+}$, then $g(d) > 0\ \forall\, d \in D\backslash\{0\}$ and we see that $D\backslash\{0\} \subset C(0)$. By part (iii) we have the reciprocal inclusion and, therefore, $C(0) = D\backslash\{0\}$. □

By Definition 3.2, for each $\varepsilon \geq 0$ we obtain the following ε-efficiency set:

$$\begin{aligned} x \in \mathrm{AE}(f, C, \varepsilon) &\iff x \in S,\ (f(x) - (D \cap [g > \varepsilon])) \cap f(S) = \emptyset \\ &\iff x \in S,\ (f(x) - f(S)) \cap (D \cap [g > \varepsilon]) = \emptyset \\ \iff [x \in S,\ z \in S,\ f(z) \in f(x) - D &\Rightarrow g_i(f(x)) - \varepsilon \leq g_i(f(z)) \forall\, i = 1, 2, \ldots, m]. \end{aligned}$$

If we consider the particular case $m = 1$, then we obtain the ε-efficiency notion introduced by Helbig [8]. By using this generalization, the decision-maker can take several criteria g_i into account in order to solve (2.1). We denote the set of ε-efficient (resp., weakly ε-efficient) solutions in this sense by $\mathrm{AE}(f, C_H, \varepsilon)$ (resp., $\mathrm{WAE}(f, C_H, \varepsilon)$). Its elements satisfy the following properties.

Proposition 4.6.

(i) $\mathrm{E}(f, D) \subset \bigcap_{\varepsilon > 0} \mathrm{AE}(f, C_H, \varepsilon) \subset \mathrm{WE}(f, D)$, *and if* $g \in \mathcal{M}^{s+}$, *then*

$$\bigcap_{\varepsilon > 0} \mathrm{AE}(f, C_H, \varepsilon) = \mathrm{E}(f, D)\,.$$

(ii) *Let* $(x_n) \subset S$, $(\varepsilon_n) \subset \mathbb{R}_+$, *and* $y \in Y$ *such that* $x_n \in \mathrm{AE}(f, C_H, \varepsilon_n)$, $\varepsilon_n \downarrow 0$, *and* $f(x_n) \to y$. *Then* $f^{-1}(y) \cap S \subset \mathrm{WE}(f, D)$.

(iii) *Let* $(x_n) \subset S$ *and* $(\varepsilon_n) \subset \mathbb{R}_+$ *such that* $x_n \in \mathrm{AE}(f, C_H, \varepsilon_n)$ *and* $\varepsilon_n \downarrow 0$. *Consider*

$$K := \bigcap_n (f(x_n) - (D \cap [g > \varepsilon_n])).$$

Then $f^{-1}(K) \cap S \subset \mathrm{WE}(f, D)$ *and if* $g \in \mathcal{M}^{s+}$, *then* $f^{-1}(K) \cap S \subset \mathrm{E}(f, D)$.

Proof. From Lemma 4.5(iii) we deduce that

$$\mathrm{E}(f, D) \subset \mathrm{AE}(f, C_H, 0) \subset \mathrm{WE}(f, D)\,.$$

Moreover, by Lemma 4.5(iii)–(iv), we have that $\mathrm{WAE}(f, C_H, 0) = \mathrm{WE}(f, D)$ and for each $g \in \mathcal{M}^{s+}$ we see that $\mathrm{AE}(f, C_H, 0) = \mathrm{E}(f, D)$. Then, parts (i)–(iii) follow easily from Theorem 3.4(iii)–(v). □

4.4. ε-efficiency in the sense of Tanaka. In this subsection we assume that the final space Y is normed and we consider a solid pointed (not necessarily convex) cone D such that $0 \in D$ and the following set:

$$C := D \cap \mathrm{cl}(\mathrm{B})^c\,,$$

where B denotes the open ball of center 0 and radius 1.

Lemma 4.7.

(i) C *is a solid pointed co-radiant set.*

(ii) $C(\varepsilon) = D \cap \mathrm{cl}(\varepsilon \mathrm{B})^c\ \forall\, \varepsilon > 0$.

(iii) $C(0) = D\backslash\{0\}$ *and* $\operatorname{int}(C)(0) = \operatorname{int}(D)$.

Proof. Part (i). First, we prove that C is a solid set. Indeed, there exists a point $q \in \operatorname{int}(D)$, $q \neq 0$, since D is a solid cone. Then, $(2/\|q\|)q \in \operatorname{int}(D) \cap \mathrm{cl}(\mathrm{B})^c = \operatorname{int}(C)$, and C is a solid set.

Moreover, C is a pointed set, since $C \subset D$ and D is a pointed cone. Let $y \in C$ and $\alpha > 1$. As D is a cone, we have that $\alpha y \in D$. Moreover, $\|\alpha y\| = \alpha\|y\| > 1$ since $\alpha > 1$ and $y \notin \mathrm{cl}(\mathrm{B})$. Then $\alpha y \in C$, and it follows that C is a co-radiant set.

Part (ii). Let $y \in C$ and $\varepsilon > 0$. It is clear that $\varepsilon y \in D$ and $\|\varepsilon y\| = \varepsilon\|y\| > \varepsilon$ since $y \in \mathrm{cl}(\mathrm{B})^c$. It follows that $\varepsilon y \in D \cap \mathrm{cl}(\varepsilon\mathrm{B})^c$ and $C(\varepsilon) \subset D \cap \mathrm{cl}(\varepsilon\mathrm{B})^c$. Similarly, if $y \in D \cap \mathrm{cl}(\varepsilon\mathrm{B})^c$, then $(1/\varepsilon)y \in D \cap \mathrm{cl}(\mathrm{B})^c = C$. Thus, $y \in \varepsilon C$ and $C(\varepsilon) = D \cap \mathrm{cl}(\varepsilon\mathrm{B})^c$.

Part (iii). By part (ii) it is clear that

$$C(0) = \bigcup_{\varepsilon>0} D \cap \mathrm{cl}(\varepsilon\mathrm{B})^c = D \cap \left(\bigcap_{\varepsilon>0} \mathrm{cl}(\varepsilon\mathrm{B})\right)^c = D\backslash\{0\}.$$

Analogously,

$$\mathrm{int}(C)\,(0) = \bigcup_{\varepsilon>0} \mathrm{int}(D) \cap \mathrm{cl}(\varepsilon\mathrm{B})^c = \mathrm{int}(D) \cap \left(\bigcap_{\varepsilon>0} \mathrm{cl}(\varepsilon\mathrm{B})\right)^c = \mathrm{int}(D)\,. \qquad \square$$

We denote by $\mathrm{AE}(f, C_T, \varepsilon)$ (resp., $\mathrm{WAE}(f, C_T, \varepsilon)$) the ε-efficiency (resp., weak ε-efficiency) set with respect to this set C. For each $\varepsilon \geq 0$ it follows that

$$\begin{aligned} x \in \mathrm{AE}(f, C_T, \varepsilon) &\iff x \in S,\ (f(x) - (D \cap \mathrm{cl}(\varepsilon\mathrm{B})^c)) \cap f(S) = \emptyset \\ &\iff x \in S,\ (f(x) - D) \cap f(S) \subset f(x) + \mathrm{cl}(\varepsilon\mathrm{B})\,. \end{aligned}$$

This concept of ε-efficiency was introduced by Tanaka [28]. Next, we give some properties of this notion, which extend a previous property proved by Tanaka in [28, Proposition 3.3]. The proof is omitted since it is similar to the proof of Proposition 4.4.

Proposition 4.8.

(i) $\bigcap_{\varepsilon>0} \mathrm{AE}(f, C_T, \varepsilon) = \mathrm{E}(f, D)$.

(ii) *Let* $(x_n) \subset S$, $(\varepsilon_n) \subset \mathbb{R}_+$, *and* $y \in Y$ *such that* $x_n \in \mathrm{AE}(f, C_T, \varepsilon_n)$, $\varepsilon_n \downarrow 0$, *and* $f(x_n) \to y$. *Then* $f^{-1}(y) \cap S \subset \mathrm{WE}(f, D)$.

(iii) *Let* $(x_n) \subset S$ *and* $(\varepsilon_n) \subset \mathbb{R}_+$ *such that* $x_n \in \mathrm{AE}(f, C_T, \varepsilon_n)$, $\varepsilon_n \downarrow 0$, *and*

$$K := \bigcap_n (f(x_n) - (D \cap \mathrm{cl}(\varepsilon_n\mathrm{B})^c)).$$

Then $f^{-1}(K) \cap S \subset \mathrm{E}(f, D)$.

5. A general scalarization for ε-efficient solutions. In the literature, approximate solutions of (2.1) are usually studied in convex problems via Kutateladze's definition and using scalarization methods which characterize this kind of solutions by means of solutions of related scalar optimization problems (see, for example, [1, Theorem 2.1], [2, Theorem 2.1], [16, Theorems 1 and 2], [17, Lemma 2.1], and [31, Lemma 3.2]). However, scalarization methods for ε-efficiency in nonconvex vector optimization problems or using ε-efficiency notions different from Kutateladze's definition are very limited (see [6, Lemma 3.1], [13], [18, Propositions 3.1 and 3.2], and [34, Lemmas 4.1 and 4.2]).

In this section we deduce several scalarizations for ε-efficient solutions obtained by Definition 3.2 and different solid pointed star-shaped co-radiant sets C. We recall that a set C is star-shaped if there exists $q \in C$ such that

$$\alpha q + (1-\alpha)y \in C \quad \forall\, y \in C, \qquad \forall\, \alpha \in (0,1). \tag{5.1}$$

We denote by $\operatorname{kern}(C)$ the set of all points $q \in C$ such that (5.1) holds. It is easy to prove that if C is a co-radiant set, then $\operatorname{kern}(C)$ is also a co-radiant set.

First, we obtain a necessary condition for ε-efficient solutions through a scalarization process based on the following nonconvex separation theorem due to Göpfert et al. [4, Theorem 2.3.1]. We denote $\mathbb{R}q = \{sq : s \in \mathbb{R}\}$.

THEOREM 5.1. *Let $G \subset Y$ be a proper closed solid set and $q \in Y$ be such that*

$$\begin{aligned} G + (0,\infty)q &\subset \operatorname{int}(G)\,, \\ Y &= \mathbb{R}q - G, \\ \forall\, y \in Y,\ \exists\, s \in \mathbb{R} &\text{ such that } y + sq \notin G. \end{aligned}$$

Then, the functional $\varphi_{q,G} : Y \to \mathbb{R}$ defined by

$$\varphi_{q,G}(y) = \inf\{s \in \mathbb{R} : y \in sq - G\} \tag{5.2}$$

is a continuous functional such that

$$\begin{aligned} \{y \in Y : \varphi_{q,G}(y) < c\} &= cq - \operatorname{int}(G) \quad \forall\, c \in \mathbb{R}, \\ \{y \in Y : \varphi_{q,G}(y) = c\} &= cq - \operatorname{bd}(G) \quad \forall\, c \in \mathbb{R}, \\ \varphi_{q,G}(-G) \le 0, &\qquad \varphi_{q,G}(-\operatorname{bd}(G)) = 0. \end{aligned}$$

We denote by $\operatorname{AMin}(g,\varepsilon)$ the set of ε-approximate solutions of the scalar optimization problem

$$\operatorname{Min}\{g(x) : x \in F\},$$

i.e.,

$$\operatorname{AMin}(g,\varepsilon) = \{x \in F : g(x) - \varepsilon \le g(z)\, \forall\, z \in F\},$$

where $g : X \to \mathbb{R}$, $F \subset X$, $F \neq \emptyset$, and $\varepsilon \ge 0$.

LEMMA 5.2. *Let C be a proper star-shaped co-radiant set such that* $\operatorname{kern}(C)$ *is solid. Then the following hold:*

(i) $d + \lambda q \in C\ \forall\, d \in C,\ \forall\, q \in \operatorname{kern}(C),\ \forall\, \lambda > 0$.
(ii) $d + \lambda q \in C(\varepsilon)\ \forall\, \varepsilon > 0,\ \forall\, d \in C(\varepsilon),\ \forall\, q \in \operatorname{kern}(C),\ \forall\, \lambda > 0$.
(iii) $\operatorname{cl}(C(\varepsilon)) + (0,\infty)q \subset \operatorname{int}(C(\varepsilon))\ \forall\, \varepsilon > 0,\ \forall\, q \in \operatorname{int}(\operatorname{kern}(C))$.
(iv) $Y = \mathbb{R}q - C(\varepsilon)\ \forall\, q \in \operatorname{int}(C),\ \forall\, \varepsilon > 0$.
(v) $Y = \mathbb{R}q + \varepsilon \operatorname{int}(\operatorname{kern}(C))\ \forall\, q \in \operatorname{int}(\operatorname{kern}(C)),\ \forall\, \varepsilon > 0$.
(vi) $\forall\, q \in \operatorname{int}(\operatorname{kern}(C)),\ \forall\, \varepsilon > 0,\ \forall\, y \in Y,\ \exists\, s \in \mathbb{R}$ *such that* $y + sq \notin \operatorname{cl}(C(\varepsilon))$.
(vii) $\operatorname{int}(\operatorname{cl}(C(\varepsilon))) = \operatorname{int}(C(\varepsilon))\ \forall\, \varepsilon > 0$.

Proof. Part (i). Let $d \in C$, $q \in \operatorname{kern}(C)$, and $\lambda > 0$. As C is a star-shaped co-radiant set and $q \in \operatorname{kern}(C)$, it follows that

$$d + \lambda q = (\lambda + 1)\left(\frac{1}{\lambda + 1} d + \left(1 - \frac{1}{\lambda + 1}\right) q\right) \in C.$$

Part (ii). Let $d \in C(\varepsilon)$, $q \in \operatorname{kern}(C)$, and $\lambda > 0$. As $(1/\varepsilon)d \in C$, by part (i) it follows that $(1/\varepsilon)d + (\lambda/\varepsilon)q \in C$. Thus, $d + \lambda q \in \varepsilon C = C(\varepsilon)$.

Part (iii). Let $q \in \operatorname{int}(\operatorname{kern}(C))$ and $\varepsilon > 0$. First, we prove that

$$C(\varepsilon) + (0,\infty)q \subset \operatorname{int}(C(\varepsilon))\,. \tag{5.3}$$

Indeed, consider $d \in C(\varepsilon)$ and $\lambda > 0$. There exists a neighborhood U of 0 such that $q + U \subset \ker(C)$, and by Part (ii) we deduce that $d + \lambda(q + U) \subset C(\varepsilon)$. Thus, it follows that $d + \lambda q \in \operatorname{int}(C(\varepsilon))$.

Next, we prove that $\operatorname{cl}(C(\varepsilon)) + (0,\infty)q \subset \operatorname{int}(C(\varepsilon))$. Let $d \in \operatorname{cl}(C(\varepsilon))$ and $\lambda > 0$. There exists $v \in C(\varepsilon)$ such that $q + (1/\lambda)(d - v) \in \operatorname{int}(\ker(C))$. Then, by (5.3) we see that $d + \lambda q = v + \lambda(q + (1/\lambda)(d - v)) \in \operatorname{int}(C(\varepsilon))$ and Part (iii) holds.

Part (iv). Let $q \in \operatorname{int}(C)$, $\varepsilon > 0$, and $y \in Y$. There exists $\delta \in (0,1)$ such that $q - (\delta/\varepsilon)y \in C$. As $\delta < 1$ and $C(\varepsilon)$ is a co-radiant set by Lemma 3.1(i), it follows that

$$y \in (\varepsilon/\delta)q - (1/\delta)(\varepsilon C) \subset \mathbb{R}q - C(\varepsilon)\,.$$

Part (v). Let $q \in \operatorname{int}(\ker(C))$, $\varepsilon > 0$, and $y \in Y$. As in the proof of Part (iv) we see that there exists $\delta \in (0,1)$ such that $y \in (\varepsilon/\delta)q - \varepsilon((1/\delta)\operatorname{int}(\ker(C)))$, and it follows that $y \in \mathbb{R}q - \varepsilon\operatorname{int}(\ker(C))$, since $\operatorname{int}(\ker(C))$ is a co-radiant set. Therefore, we have that $Y \subset \mathbb{R}q - \varepsilon\operatorname{int}(\ker(C))$, which implies that $Y = \mathbb{R}q + \varepsilon\operatorname{int}(\ker(C))$.

Part (vi). Suppose that there exist $q \in \operatorname{int}(\ker(C))$, $\varepsilon > 0$, and $y \in Y$ such that $y + \mathbb{R}q \subset \operatorname{cl}(C(\varepsilon))$. Then, by Part (v) we deduce that

$$Y = (y + \mathbb{R}q) + \varepsilon\operatorname{int}(\ker(C)) - y \subset \operatorname{cl}(C(\varepsilon)) + \varepsilon\operatorname{int}(\ker(C)) - y.$$

From Part (iii) it follows that $\operatorname{cl}(C(\varepsilon)) + \varepsilon\operatorname{int}(\ker(C)) \subset \operatorname{int}(C(\varepsilon))$. Therefore, $Y \subset \varepsilon\operatorname{int}(C) - y$, which is a contradiction, since C is a proper set.

Part (vii). It is clear that $\operatorname{int}(C(\varepsilon)) \subset \operatorname{int}(\operatorname{cl}(C(\varepsilon)))$. Reciprocally, let $d \in \operatorname{int}(\operatorname{cl}(C(\varepsilon)))$ and $q \in \operatorname{int}(\ker(C))$. There exists $\delta > 0$ such that $d - \delta q \in \operatorname{cl}(C(\varepsilon))$. Then, by Part (iii) we deduce that

$$d \in \operatorname{cl}(C(\varepsilon)) + \delta q \subset \operatorname{int}(C(\varepsilon))\,,$$

and therefore $\operatorname{int}(\operatorname{cl}(C(\varepsilon))) \subset \operatorname{int}(C(\varepsilon))$. □

THEOREM 5.3. *Let C be a proper star-shaped co-radiant set such that $\ker(C)$ is solid, $q \in \operatorname{int}(\ker(C))$, and $\varepsilon > 0$. Then the following holds:*

$$x_0 \in \operatorname{WAE}(f, C, \varepsilon) \Rightarrow x_0 \in \operatorname{AMin}\big(\varphi_{q,C(\varepsilon),f(x_0)} \circ f, \varepsilon\big),$$

where $\varphi_{q,C(\varepsilon),f(x_0)}(y) := \varphi_{q,\operatorname{cl}(C(\varepsilon))}(y - f(x_0))\ \forall\, y \in Y$.

Proof. Let $\varepsilon > 0$ and consider $G := \operatorname{cl}(C(\varepsilon))$. By parts (iii), (iv), and (vi) of Lemma 5.2, we deduce that G satisfies the hypotheses of Theorem 5.1. Then, from this theorem we deduce that

$$\{y \in Y : \varphi_{q,\operatorname{cl}(C(\varepsilon))}(y) < 0\} = -\operatorname{int}(\operatorname{cl}(C(\varepsilon)))\,. \tag{5.4}$$

For each $x_0 \in \operatorname{WAE}(f, C, \varepsilon)$ we have that

$$(f(S) - f(x_0)) \cap -\operatorname{int}(C(\varepsilon)) = \emptyset.$$

Then, by (5.4) and Lemma 5.2(vii) we see that

$$(\varphi_{q,C(\varepsilon),f(x_0)} \circ f)(x) = \varphi_{q,\operatorname{cl}(C(\varepsilon))}(f(x) - f(x_0)) \geq 0 \quad \forall\, x \in S.$$

Moreover, by (5.2) it is clear that

$$(\varphi_{q,C(\varepsilon),f(x_0)} \circ f)(x_0) = \varphi_{q,\operatorname{cl}(C(\varepsilon))}(0) = \inf\{s \in \mathbb{R} : sq \in \operatorname{cl}(C(\varepsilon))\} \leq \varepsilon.$$

Thus, $(\varphi_{q,C(\varepsilon),f(x_0)} \circ f)(x_0) - \varepsilon \leq (\varphi_{q,C(\varepsilon),f(x_0)} \circ f)(x)\ \forall\, x \in S$, and we conclude that $x_0 \in \mathrm{AMin}\big(\varphi_{q,C(\varepsilon),f(x_0)} \circ f, \varepsilon\big)$. ∎

Gerth and Weidner [3, section 3] and Rubinov and Gasimov [22, Theorem 7.2] have obtained necessary conditions for weakly efficient solutions using Minkowski-type functionals similar to $\varphi_{q,C(\varepsilon),f(x_0)}$. In Theorem 5.3 we have extended these necessary conditions to the weak ε-efficiency set.

Necessary conditions for the ε-efficiency set in the sense of Kutateladze have been proved via Minkowski-type functionals by Göpfert et al. [4, section 3.1.1] and Tammer [27, Theorem 1]. In Theorem 5.3 we have extended these results to other ε-efficiency concepts since, in the theorem, we have obtained necessary conditions for the general ε-efficiency notion introduced in Definition 3.2 based on the Minkowski-type functional $\varphi_{q,C(\varepsilon),f(x_0)}$.

Next, we extend Theorem 5.3 for some vectors $q \in \mathrm{cl}(\mathrm{int}(\mathrm{kern}(C)))$.

LEMMA 5.4. *Let $G \subset Y$ and let $(\alpha_n) \subset \mathbb{R}$ be such that $\alpha_n \downarrow 0$. Consider $q \in Y$ and the sequence (q_n), where $q_n := (1+\alpha_n)q\ \forall\, n$. Then, $\varphi_{q_n,G}(y) \to \varphi_{q,G}(y)\ \forall\, y \in Y$, where $\varphi_{q_n,G}$ and $\varphi_{q,G}$ are given by* (5.2).

Proof. Let $y \in Y$. For each $s \in \mathbb{R}$ such that $y \in sq - G$, it follows that $y \in (s/(1+\alpha_n))q_n - G$. Then, $\varphi_{q_n,G}(y) \leq s/(1+\alpha_n)$ and we deduce that

$$(1+\alpha_n)\varphi_{q_n,G}(y) \leq \varphi_{q,G}(y). \tag{5.5}$$

Analogously, for each $s \in \mathbb{R}$ such that $y \in sq_n - G$, it follows that $y \in s(1+\alpha_n)q - G$, and we have that

$$\varphi_{q,G}(y) \leq \varphi_{q_n,G}(y)(1+\alpha_n). \tag{5.6}$$

Combining (5.5) and (5.6) we see that

$$\varphi_{q,G}(y) \leq \varphi_{q_n,G}(y)(1+\alpha_n) \leq \varphi_{q,G}(y)$$

and taking the limit when $n \to \infty$ we conclude that $\varphi_{q_n,G}(y) \to \varphi_{q,G}(y)\ \forall\, y \in Y$. ∎

PROPOSITION 5.5. *Let C be a proper star-shaped co-radiant set such that* $\mathrm{kern}(C)$ *is solid. Consider $\varepsilon > 0$, $q \in Y$, $\alpha_n \downarrow 0$, and a sequence (q_n), where $q_n = (1+\alpha_n)q$ and $q_n \in \mathrm{int}(\mathrm{kern}(C))\ \forall\, n$. If $x_0 \in \mathrm{WAE}(f, C, \varepsilon)$, then $x_0 \in \mathrm{AMin}\big(\varphi_{q,C(\varepsilon),f(x_0)} \circ f, \varepsilon\big)$.*

Proof. Let $x_0 \in \mathrm{WAE}(f, C, \varepsilon)$. From Theorem 5.3 we have that

$$(\varphi_{q_n,C(\varepsilon),f(x_0)} \circ f)(x_0) - \varepsilon \leq (\varphi_{q_n,C(\varepsilon),f(x_0)} \circ f)(x) \quad \forall\, x \in S, \qquad \forall\, n$$

and taking the limit when $n \to \infty$ we deduce from Lemma 5.4 that

$$(\varphi_{q,C(\varepsilon),f(x_0)} \circ f)(x_0) - \varepsilon \leq (\varphi_{q,C(\varepsilon),f(x_0)} \circ f)(x) \quad \forall\, x \in S. \qquad ∎$$

Remark 5.6. Let us observe that Theorem 5.3 and Proposition 5.5 give necessary conditions for ε-efficient solutions since $\mathrm{AE}(f, C, \varepsilon) \subset \mathrm{WAE}(f, C, \varepsilon)$.

In Proposition 5.7 we use the recession cone 0^+C of a star-shaped set C to obtain a simple sufficient condition to have $\mathrm{int}(\mathrm{kern}(C)) \neq \emptyset$.

PROPOSITION 5.7. *Let C be a star-shaped set and $q \in \mathrm{kern}(C)$. Then $q + 0^+C \subset \mathrm{kern}(C)$, and* $\mathrm{kern}(C)$ *is a solid set when 0^+C is a solid cone.*

Proof. Consider $d \in 0^+C$, $y \in C$, and $\alpha \in (0,1)$. It is clear that

$$\alpha(q+d) + (1-\alpha)y = \alpha q + (1-\alpha)y + \alpha d \in C + \alpha 0^+C \subset C$$

and $q + 0^+C \subset \operatorname{kern}(C)$. Moreover, $\operatorname{kern}(C)$ is a solid set when 0^+C is a solid cone since $q + \operatorname{int}(0^+C) \subset \operatorname{int}(\operatorname{kern}(C))$. □

Usually, sufficient conditions for efficient solutions are obtained via scalarization methods based on monotone functionals (see, for example, [33, Theorem 9]). Next, we obtain sufficient conditions for ε-efficient solutions through a scalarization process whose functional verifies the following new monotonicity concept. In the rest of this section we assume that Y is a normed space.

DEFINITION 5.8. *Let $K \subset Y$ be a cone. We say that a functional $g : Y \to \mathbb{R}$ is strictly local K-monotone at $y_0 \in Y$ if there exist $\sigma > 0$ and $\rho > 0$ such that*

$$z \in y_0 - (K \cap \rho \mathrm{B}), \quad y \in z - K \Rightarrow g(y) \leq g(z), \tag{5.7}$$

$$y \in y_0 - (K \cap \rho \mathrm{B}) \Rightarrow g(y) + \sigma \|y - y_0\| \leq g(y_0). \tag{5.8}$$

We will say that g is strictly local K-monotone at y_0 with constants σ and ρ. Condition (5.7) tells us that g is a monotone functional with respect to the preference relation defined by the cone K at points near to y_0 (we will say that g is local K-monotone at y_0). These functionals are frequently used to obtain sufficient conditions for the efficient solutions of vector optimization problems. Let us observe that (5.7) is a "local" condition. For example, consider $Y = \mathbb{R}$, $K = \mathbb{R}_+$, and

$$g(y) = \begin{cases} \cos y & \text{if} \quad y \leq 0, \\ 1 & \text{if} \quad y > 0. \end{cases}$$

For each $y_0 > 0$ it is clear that g verifies (5.7) taking $0 < \rho \leq y_0$. However, g does not verify (5.7) at $y_0 > 0 \; \forall \, \rho > 0$.

Property (5.8) says that y_0 is a strict (or sharp) local maximum of order 1 for the scalar optimization problem $\operatorname{Max}\{g(y) : y \in y_0 - K\}$. Below simple examples of strictly local K-monotone functionals are given. Parts (i)–(iii) are obvious, and their proofs are omitted. In Lemmas 6.2 and 6.3 other examples are provided.

Example 5.9. Let g_1, g_2 be two functionals from Y into $\mathbb{R}$ and assume that $0 \in K$.

(i) If g_1 and g_2 are strictly local K-monotone at y_0, then $\max\{g_1, g_2\}$ and $\min\{g_1, g_2\}$ are strictly local K-monotone at y_0.

(ii) If g_1 is strictly local K-monotone at y_0 and g_2 is local K-monotone at y_0, then $g_1 + g_2$ is strictly local K-monotone at y_0.

(iii) If $g \in K^+$, then for each $y_0 \in Y$ and $\rho > 0$, g is local K-monotone at y_0.

(iv) If $g \in K^{s+}$, K is closed, and Y is finite dimensional, then for each $y_0 \in Y$ and $\rho > 0$, g is strictly local K-monotone at y_0 with constants σ and ρ, where

$$\sigma := \min\{g(d) : d \in K, \|d\| = 1\}.$$

Indeed, as the set $A := \{d \in K : \|d\| = 1\}$ is compact and $g(d) > 0 \; \forall \, d \in A$, by the Weierstrass theorem it follows that $\sigma > 0$. Then $\forall \, y \in y_0 - (K \cap \rho \mathrm{B})$, $y \neq y_0$, as $(y_0 - y)/\|y_0 - y\| \in A$ we have that $\sigma \leq g((y_0 - y)/\|y_0 - y\|)$. By linearity we obtain that statement (5.8) is satisfied. Condition (5.7) holds by part (iii).

THEOREM 5.10. *Let $C \subset Y$ be a nonempty pointed co-radiant set. Consider $x_0 \in S$ and let $g : Y \to \mathbb{R}$ be a strictly local $C(0)$-monotone functional at $f(x_0)$ with constants σ and ρ.*

(i) *If $0 \notin \operatorname{cl}(C)$, $x_0 \in \operatorname{AMin}(g \circ f, \delta)$, and $0 < \delta < \sigma\rho$, then $x_0 \in \operatorname{AE}(f, C, \delta/(\sigma\beta))$ $\forall \, \beta > 0$ such that $\operatorname{cl}(\beta \mathrm{B}) \cap C = \emptyset$.*

(ii) *If $x_0 \in \operatorname{AMin}(g \circ f, 0)$, then $x_0 \in \operatorname{AE}(f, C, 0)$.*

Proof. Part (i). To obtain a contradiction, suppose that there exists $\beta > 0$ such that $\mathrm{cl}(\beta \mathrm{B}) \cap C = \emptyset$ and $x_0 \notin \mathrm{AE}(f, C, \delta/(\sigma\beta))$. Then

$$(f(x_0) - C(\delta/(\sigma\beta))) \cap f(S) \not\subset \{f(x_0)\}$$

and there exist $x \in S$ and $d \in C(\delta/(\sigma\beta))$ such that $f(x_0) - d = f(x)$. As $\mathrm{cl}(\beta \mathrm{B}) \cap C = \emptyset$ and $C(\delta/(\sigma\beta)) = (\delta/(\sigma\beta))C$, it follows that $\|(\sigma\beta/\delta)d\| > \beta$, and so $\|d\| > \delta/\sigma$. Thus, there exists $\nu > 0$ such that

$$\frac{\delta + \nu}{\sigma\|d\|} < 1, \quad \delta + \nu < \sigma\rho. \tag{5.9}$$

As $x_0 \in \mathrm{AMin}(g \circ f, \delta)$ and $x \in S$, it follows that

$$g(f(x_0)) - \delta \leq g(f(x)). \tag{5.10}$$

Choosing the point

$$z := f(x_0) + \frac{\delta + \nu}{\sigma\|d\|}(-d),$$

we see that $z \in f(x_0) - C(0)$ and taking into account (5.9),

$$\|z - f(x_0)\| = \frac{\delta + \nu}{\sigma} < \rho.$$

Thus $z \in f(x_0) - (C(0) \cap \rho \mathrm{B})$. As g satisfies (5.7) with $y_0 = f(x_0)$ and $K = C(0)$, and

$$z = f(x_0) + \frac{\delta + \nu}{\sigma\|d\|}(f(x) - f(x_0)) = f(x) + \left(1 - \frac{\delta + \nu}{\sigma\|d\|}\right) d \in f(x) + C(0)\,,$$

we have that

$$g(f(x)) \leq g(z). \tag{5.11}$$

Then, by (5.10) and (5.11) we deduce that

$$g(f(x_0)) - \delta \leq g(z). \tag{5.12}$$

As $z \in f(x_0) - (C(0) \cap \rho \mathrm{B})$ and g is a strictly local $C(0)$-monotone functional at $f(x_0)$ we deduce from (5.8) that

$$g(f(x_0)) \geq g(z) + \sigma\|z - f(x_0)\| = g(z) + \delta + \nu > g(z) + \delta,$$

contrary to (5.12).

Part (ii). Suppose that $x_0 \notin \mathrm{AE}(f, C, 0)$; then there exist $x \in S$ and $d \in C(0)$, $d \neq 0$, such that $f(x_0) - d = f(x)$. Let us select $\nu > 0$ such that

$$\frac{\nu}{\sigma\|d\|} < 1, \quad \nu < \sigma\rho.$$

Now, we proceed exactly as in the proof of Part (i), after equation (5.9), taking into account that $\delta = 0$. □

Remark 5.11. Let us observe that if $0 \notin \mathrm{cl}(C)$, then parts (i) and (ii) of Theorem 5.10 can be rewritten jointly by the following statement: If $x_0 \in \mathrm{AMin}(g \circ f, \delta)$ and $0 \leq \delta < \sigma\rho$, then $x_0 \in \mathrm{AE}(f, C, \delta/(\sigma\beta))\ \forall\, \beta > 0$ such that $\mathrm{cl}(\beta \mathrm{B}) \cap C = \emptyset$.

In [4, section 3.1.1], [13, section 5], and [27, Theorem 2] the reader can find sufficient conditions for ε-efficient solutions in the sense of Kutateladze. Theorem 5.10 extends these results to the general ε-efficiency notion introduced in Definition 3.2.

6. Scalarizations for approximate solutions in Pareto multiobjective optimization problems. In this section we consider that (2.1) is a nonconvex Pareto multiobjective optimization problem ($Y = \mathbb{R}^p$ and $D = \mathbb{R}^p_+$) and we obtain necessary and sufficient conditions for the ε-efficient solutions of (2.1) in the senses of Kutateladze, Németh, Helbig, White, and Tanaka via Proposition 5.5 and Theorem 5.10. For Helbig's notion we take $m = 1$, and thus, the functional (4.6) is denoted by $\langle g, \cdot \rangle$ with $g \in \mathbb{R}^p_+ \backslash \{0\}$.

We denote $f = (f_1, f_2, \ldots, f_p)$ and the components of a vector $y \in \mathbb{R}^p$ by y_i, $i = 1, 2, \ldots, p$. For each $C \subset \mathbb{R}^p$ and $y \in \mathbb{R}^p$ we denote $d(y, C) = \inf\{\|y - z\| : z \in C\}$.

In what follows, we assume the supremum norm $\| \|_\infty$ and the Euclidean norm $\| \|_2$ in $\mathbb{R}^p$ when we consider Németh's and Helbig's ε-efficiency notions, respectively. Moreover, for White's and Tanaka's definitions, we consider a norm of the family $\{\| \|^\mu_\infty : \mu \in \operatorname{int}(\mathbb{R}^p_+)\}$, where $\|y\|^\mu_\infty = \|(\mu_1 y_1, \mu_2 y_2, \ldots, \mu_p y_p)\|_\infty \ \forall\, y \in \mathbb{R}^p$ (see [24, section 3.4.2] for more detail).

THEOREM 6.1. *Let $x_0 \in S$, $q \in \operatorname{int}(\mathbb{R}^p_+)$, and $\varepsilon > 0$.*

(i) *Consider*

$$k_{f(x_0)}(y) := \max_{1 \le i \le p} \left\{ \frac{y_i - f_i(x_0)}{q_i} \right\}.$$

If $x_0 \in \operatorname{WAE}(f, C_K, \varepsilon)$, then $x_0 \in \operatorname{AMin}(k_{f(x_0)} \circ f, \varepsilon)$.

(ii) *Let $L = \{e_1, e_2, \ldots, e_p\}$ be the standard basis of $\mathbb{R}^p$ and let $H = \operatorname{conv}(L)$. Suppose that $q \in H$ and consider*

$$n_{f(x_0)}(y) := \max \left\{ \max_{1 \le i \le p} \left\{ \frac{y_i - f_i(x_0)}{q_i} \right\}, \sum_{i=1}^{p} (y_i - f_i(x_0)) + \varepsilon \right\}.$$

If $x_0 \in \operatorname{WAE}(f, C_N, \varepsilon)$, then $x_0 \in \operatorname{AMin}(n_{f(x_0)} \circ f, \varepsilon)$.

(iii) *Let $g \in \mathbb{R}^p_+ \backslash \{0\}$ and suppose that $\langle g, q \rangle \ge 1$. Consider*

$$h_{f(x_0)}(y) := \max \left\{ \max_{1 \le i \le p} \left\{ \frac{y_i - f_i(x_0)}{q_i} \right\}, \frac{\langle g, y \rangle - \langle g, f(x_0) \rangle + \varepsilon}{\langle g, q \rangle} \right\}.$$

If $x_0 \in \operatorname{WAE}(f, C_H, \varepsilon)$, then $x_0 \in \operatorname{AMin}(h_{f(x_0)} \circ f, \varepsilon)$.

(iv) *Let*

$$t_{f(x_0)}(y) := \max \left\{ \max_{1 \le i \le p} \left\{ \frac{y_i - f_i(x_0)}{q_i} \right\}, \min_{1 \le i \le p} \left\{ \frac{y_i - f_i(x_0)}{q_i} \right\} + \varepsilon \right\}.$$

If $\mu = (1/q_1, 1/q_2, \ldots, 1/q_p)$ and $x_0 \in \operatorname{WAE}(f, C_T, \varepsilon)$ when the norm $\| \|^\mu_\infty$ is considered, then it follows that $x_0 \in \operatorname{AMin}(t_{f(x_0)} \circ f, \varepsilon)$. Moreover, it follows that $\operatorname{AE}(f, C_T, \varepsilon) = \operatorname{AE}(f, C_W, \varepsilon)$ and $\operatorname{WAE}(f, C_T, \varepsilon) = \operatorname{WAE}(f, C_W, \varepsilon)$.

Proof. Part (i). The elements included in $\operatorname{WAE}(f, C_K, \varepsilon)$ are the weakly ε-efficient solutions of (2.1) with respect to the set $C = q + \mathbb{R}^p_+$. It is clear that C is a proper solid convex co-radiant set. As C is a convex set and $\operatorname{int}(C) = q + \operatorname{int}(\mathbb{R}^p_+)$, it follows that $\operatorname{int}(\operatorname{kern}(C)) = q + \operatorname{int}(\mathbb{R}^p_+)$ and $q_n := (1 + 1/n)q \in \operatorname{int}(\operatorname{kern}(C))$ $\forall\, n$. Therefore, from Proposition 5.5 we deduce that if $x_0 \in \operatorname{WAE}(f, C_K, \varepsilon)$, then $x_0 \in \operatorname{AMin}(\varphi_{q, C(\varepsilon), f(x_0)} \circ f, \varepsilon)$, where $\varphi_{q, C(\varepsilon), f(x_0)}(y) = \varphi_{q, C(\varepsilon)}(y - f(x_0)) \ \forall\, y \in \mathbb{R}^p$ and $C(\varepsilon) = \varepsilon q + \mathbb{R}^p_+$.

From (5.2) we have that

$$\varphi_{q, C(\varepsilon)}(y - f(x_0)) = \inf\{s \in \mathbb{R} : y - f(x_0) \in sq - \varepsilon q - \mathbb{R}^p_+\}.$$

Moreover,

$$y - f(x_0) \in sq - \varepsilon q - \mathbb{R}^p_+ \iff s \geq \frac{y_i - f_i(x_0) + \varepsilon q_i}{q_i} \quad \forall\, i = 1, 2, \ldots, p$$

and

$$\varphi_{q,C(\varepsilon),f(x_0)}(y) = \max_{1 \leq i \leq p} \left\{ \frac{y_i - f_i(x_0)}{q_i} \right\} + \varepsilon = k_{f(x_0)}(y) + \varepsilon.$$

As $x_0 \in \mathrm{AMin}\big(\varphi_{q,C(\varepsilon),f(x_0)} \circ f, \varepsilon\big)$, we have that

$$k_{f(x_0)}(f(x_0)) \leq k_{f(x_0)}(f(x)) + \varepsilon \quad \forall\, x \in S$$

and $x_0 \in \mathrm{AMin}\big(k_{f(x_0)} \circ f, \varepsilon\big)$.

Part (ii). $\mathrm{WAE}(f, C_N, \varepsilon)$ contains all weakly ε-efficient solutions of (2.1) with respect to $C = H + \mathbb{R}^p_+$, which is a proper solid convex co-radiant set, and $\mathrm{int}(C) = H + \mathrm{int}\big(\mathbb{R}^p_+\big)$. As $q \in H \cap \mathrm{int}\big(\mathbb{R}^p_+\big)$, it follows that $q_n := (1 + 1/n)q \in H + \mathrm{int}\big(\mathbb{R}^p_+\big)$. Then, by Proposition 5.5 we deduce that if $x_0 \in \mathrm{WAE}(f, C_N, \varepsilon)$, then $x_0 \in \mathrm{AMin}\big(\varphi_{q,C(\varepsilon),f(x_0)} \circ f, \varepsilon\big)$, where $C(\varepsilon) = \varepsilon H + \mathbb{R}^p_+$ is closed. From (5.2) we see that

$$\varphi_{q,C(\varepsilon)}(y - f(x_0)) = \inf\{s \in \mathbb{R} : y - f(x_0) \in sq - \varepsilon H - \mathbb{R}^p_+\}.$$

As

$$-\varepsilon H - \mathbb{R}^p_+ = \left\{ y \in -\mathbb{R}^p_+ : \sum_{i=1}^p y_i \leq -\varepsilon \right\},$$

it follows that

$$y - f(x_0) \in sq - \varepsilon H - \mathbb{R}^p_+ \iff \begin{cases} y - f(x_0) - sq \in -\mathbb{R}^p_+, \\ \displaystyle\sum_{i=1}^p (y_i - f_i(x_0) - sq_i) \leq -\varepsilon \end{cases}$$

$$\iff \begin{cases} s \geq \dfrac{y_i - f_i(x_0)}{q_i} & \forall\, i = 1, 2, \ldots, p, \\ s \geq \displaystyle\sum_{i=1}^p (y_i - f_i(x_0)) + \varepsilon \end{cases}$$

since $q \in H$ implies $\sum_{i=1}^p q_i = 1$. Thus, $\varphi_{q,C(\varepsilon)}(y - f(x_0)) = n_{f(x_0)}(y)\ \forall\, y \in \mathbb{R}^p$, and we conclude that $x_0 \in \mathrm{AMin}\big(n_{f(x_0)} \circ f, \varepsilon\big)$.

Part (iii). The elements of $\mathrm{WAE}(f, C_H, \varepsilon)$ are weakly ε-efficient solutions of (2.1) with respect to the proper solid convex co-radiant set $C = \mathbb{R}^p_+ \cap [\langle g, \cdot \rangle > 1]$. As C is convex, we have that $\mathrm{kern}(C) = \mathbb{R}^p_+ \cap [\langle g, \cdot \rangle > 1]$ and $(1 + 1/n)q \in \mathrm{int}(\mathrm{kern}(C))$ $\forall\, n$. Thus, by Proposition 5.5 we deduce that if $x_0 \in \mathrm{WAE}(f, C_H, \varepsilon)$, then $x_0 \in \mathrm{AMin}\big(\varphi_{q,C(\varepsilon),f(x_0)} \circ f, \varepsilon\big)$. From (5.2) with $G = \mathrm{cl}(C(\varepsilon)) = \mathbb{R}^p_+ \cap [\langle g, \cdot \rangle \geq \varepsilon]$ we see that

$$\varphi_{q,\mathrm{cl}(C(\varepsilon))}(y - f(x_0)) = \inf\{s \in \mathbb{R} : y - f(x_0) \in sq - (\mathbb{R}^p_+ \cap [\langle g, \cdot \rangle \geq \varepsilon])\}$$

and

$$y - f(x_0) \in sq - (\mathbb{R}^p_+ \cap [\langle g, \cdot\rangle \geq \varepsilon]) \iff \begin{cases} s \geq \dfrac{y_i - f_i(x_0)}{q_i} & \forall\, i = 1, 2, \ldots, p, \\ s \geq \dfrac{\langle g, y\rangle - \langle g, f(x_0)\rangle + \varepsilon}{\langle g, q\rangle}. \end{cases}$$

Therefore, $\varphi_{q,\mathrm{cl}(C(\varepsilon))}(y - f(x_0)) = h_{f(x_0)}(y)\ \forall\, y \in \mathbb{R}^p$, and it follows that $x_0 \in \mathrm{AMin}\big(h_{f(x_0)} \circ f, \varepsilon\big)$.

Part (iv). $\mathrm{WAE}(f, C_T, \varepsilon)$ is the weak ε-efficiency set of (2.1) with respect to $C = \mathbb{R}^p_+ \cap \mathrm{cl}(\mathrm{B})^c$, where the open unit ball B is defined by the norm $\|\ \|^\mu_\infty$. It is clear that C is a proper solid nonconvex co-radiant set. Moreover, C is a star-shaped set since, for example, $(1 + 1/n)q \in \mathrm{kern}(C)\ \forall\, n$. Indeed, $\forall\, y \in C$ it follows that there exists a component $y_i > q_i$ and

$$\alpha(1 + 1/n)q + (1 - \alpha)y \in \mathbb{R}^p_+ \cap \mathrm{cl}(\mathrm{B})^c \quad \forall\, \alpha \in (0, 1), \quad \forall\, n,$$

because $C \subset \mathbb{R}^p_+$ and $\alpha(1 + 1/n)q_i + (1 - \alpha)y_i > q_i$.

We have that $\mathbb{R}^p_+ \subset 0^+C$, since $\forall\, y \in C$, $\forall\, \alpha > 0$, and $\forall\, d \in \mathbb{R}^p_+$ we have that $y + \alpha d \in \mathbb{R}^p_+ + \mathbb{R}^p_+ = \mathbb{R}^p_+$ and

$$\|y + \alpha d\|^\mu_\infty = \max_{1 \leq i \leq p} \left\{ \frac{y_i + \alpha d_i}{q_i} \right\} \geq \max_{1 \leq i \leq p} \left\{ \frac{y_i}{q_i} \right\} = \|y\|^\mu_\infty > 1.$$

Therefore, by Proposition 5.7 we deduce that $(1 + 1/n)q + \mathbb{R}^p_+ \subset \mathrm{kern}(C)\ \forall\, n$, and so $(1 + 1/n)q \in \mathrm{int}(\mathrm{kern}(C))\ \forall\, n$, and by Proposition 5.5 we have that if $x_0 \in \mathrm{WAE}(f, C_T, \varepsilon)$, then $x_0 \in \mathrm{AMin}\big(\varphi_{q,C(\varepsilon),f(x_0)} \circ f, \varepsilon\big)$. From (5.2) we see that

$$\varphi_{q,\mathrm{cl}(C(\varepsilon))}(y - f(x_0)) = \inf\{s \in \mathbb{R} : y - f(x_0) \in sq - (\mathbb{R}^p_+ \cap (\varepsilon \mathrm{B})^c)\}$$

since $\mathrm{cl}(C(\varepsilon)) = \mathbb{R}^p_+ \cap (\varepsilon \mathrm{B})^c$. It follows that

$$y - f(x_0) \in sq - (\mathbb{R}^p_+ \cap (\varepsilon \mathrm{B})^c) \iff \begin{cases} s \geq \dfrac{y_i - f_i(x_0)}{q_i} & \forall\, i = 1, 2, \ldots, p, \\ s \geq \min\limits_{1 \leq i \leq p} \left\{ \dfrac{y_i - f_i(x_0)}{q_i} \right\} + \varepsilon. \end{cases}$$

Thus, $\varphi_{q,\mathrm{cl}(C(\varepsilon))}(y - f(x_0)) = t_{f(x_0)}(y)\ \forall\, y \in \mathbb{R}^p$ and $x_0 \in \mathrm{AMin}\big(t_{f(x_0)} \circ f, \varepsilon\big)$.

Finally, $\mathrm{AE}(f, C_W, \varepsilon)$ is the ε-efficiency set with respect to $\mathbb{R}^p_+ \cap (q - \mathbb{R}^p_+)^c$. As $\mathbb{R}^p_+ \cap (q - \mathbb{R}^p_+)^c = \mathbb{R}^p_+ \cap \mathrm{cl}(\mathrm{B})^c = C$, we conclude that

$$\mathrm{AE}(f, C_T, \varepsilon) = \mathrm{AE}(f, C_W, \varepsilon) \tag{6.1}$$

and

$$\mathrm{WAE}(f, C_T, \varepsilon) = \mathrm{WAE}(f, C_W, \varepsilon)\,. \qquad \square$$

In [34, Proposition 3.2], Yokoyama proved relation (6.1) for $q = (1, 1, \ldots, 1)$. We extend this result to each $q \in \mathrm{int}\big(\mathbb{R}^p_+\big)$.

Next, we deduce sufficient conditions for approximate solutions of Pareto multiobjective optimization problems. The following monotonicity properties of the functions $n_{f(x_0)}$, $h_{f(x_0)}$, and $t_{f(x_0)}$ are necessary, in order to apply Theorem 5.10.

LEMMA 6.2. *The functions $n_{f(x_0)}$ and $t_{f(x_0)}$ are strictly local $\mathbb{R}^p_+$-monotone at $f(x_0)$ by taking $\|\ \|_\infty$, $\sigma = 1$, $\rho = \varepsilon/p$ and $\|\ \|^\mu_\infty$, $\sigma = 1$, $\rho = \varepsilon$ in (5.7)–(5.8), respectively.*

Proof. It is clear that if $y, z \in \mathbb{R}^p$ and $y \in z - \mathbb{R}^p_+$, then $n_{f(x_0)}(y) \le n_{f(x_0)}(z)$ and $t_{f(x_0)}(y) \le t_{f(x_0)}(z)$. Therefore, the functions $n_{f(x_0)}$ and $t_{f(x_0)}$ satisfy condition (5.7) $\forall \rho > 0$ and $y_0 = f(x_0)$.

Let us consider $\mathbb{R}^p$ with the norm $\|\ \|_\infty$ and $y \in f(x_0) - (\mathbb{R}^p_+ \cap (\varepsilon/p)\mathrm{B})$. Then

$$\frac{\varepsilon}{p} > \|y - f(x_0)\|_\infty \ge |y_i - f_i(x_0)| = f_i(x_0) - y_i \quad \forall\, i = 1, 2, \dots, p,$$

and so $\sum_{i=1}^{p}(y_i - f_i(x_0)) + \varepsilon > 0$. As $y - f(x_0) \in -\mathbb{R}^p_+$, it follows that

$$n_{f(x_0)}(y) = \sum_{i=1}^{p}(y_i - f_i(x_0)) + \varepsilon$$

and we see that

$$n_{f(x_0)}(f(x_0)) - n_{f(x_0)}(y) = \varepsilon - \sum_{i=1}^{p}(y_i - f_i(x_0)) - \varepsilon$$

$$= -\sum_{i=1}^{p}(y_i - f_i(x_0)) = \sum_{i=1}^{p}|y_i - f_i(x_0)| \ge \|y - f(x_0)\|_\infty.$$

Therefore, $n_{f(x_0)}$ is a strictly local $\mathbb{R}^p_+$-monotone function at $f(x_0)$ by taking the norm $\|\ \|_\infty$ in $\mathbb{R}^p$ and constants $\sigma = 1$ and $\rho = \varepsilon/p$.

Next, let us consider $\mathbb{R}^p$ with the norm $\|\ \|^\mu_\infty$, $\mu = (1/q_1, 1/q_2, \dots, 1/q_p)$, and $y \in f(x_0) - (\mathbb{R}^p_+ \cap \varepsilon\mathrm{B})$. Then

$$\min_{1\le i\le p}\left\{\frac{y_i - f_i(x_0)}{q_i}\right\} + \varepsilon = -\max_{1\le i\le p}\left\{-\frac{y_i - f_i(x_0)}{q_i}\right\} + \varepsilon = -\|y - f(x_0)\|^\mu_\infty + \varepsilon > 0,$$

and we have that

$$t_{f(x_0)}(f(x_0)) - t_{f(x_0)}(y) = \varepsilon + \|y - f(x_0)\|^\mu_\infty - \varepsilon = \|y - f(x_0)\|^\mu_\infty.$$

Therefore, $t_{f(x_0)}$ is a strictly local $\mathbb{R}^p_+$-monotone function at $f(x_0)$ by taking the norm $\|\ \|^\mu_\infty$ in $\mathbb{R}^p$ and constants $\sigma = 1$ and $\rho = \varepsilon$. □

For each $g \in \operatorname{int}\big(\mathbb{R}^p_+\big)$ we denote

$$m_g = \min\{\langle g, y\rangle : y \in \mathbb{R}^p_+,\ \|y\|_2 = 1\},$$

where $\|\ \|_2$ is the Euclidean norm. Let us observe that $m_g > 0$.

LEMMA 6.3. *Let us consider $\varepsilon > 0$, $g \in \operatorname{int}\big(\mathbb{R}^p_+\big)$, and the Euclidean norm in $\mathbb{R}^p$. Then, the function $h_{f(x_0)}$ is strictly local $\mathbb{R}^p_+$-monotone at $f(x_0)$ satisfying (5.7)–(5.8) with constants $\sigma = m_g/\langle g, q\rangle$ and $\rho = \varepsilon/\|g\|_2$.*

Proof. It is clear that $h_{f(x_0)}$ satisfies (5.7) $\forall\, \rho > 0$ and $y_0 = f(x_0)$. To prove (5.8) consider the Euclidean norm in $\mathbb{R}^p$, $\sigma = m_g/\langle g, q\rangle$, $\rho = \varepsilon/\|g\|_2$, and $y \in f(x_0) - (\mathbb{R}^p_+ \cap \rho\mathrm{B})$. We have that

$$h_{f(x_0)}(y) = \frac{\langle g, y\rangle - \langle g, f(x_0)\rangle + \varepsilon}{\langle g, q\rangle},$$

since $y_i - f_i(x_0) \leq 0 \; \forall\, i = 1, 2, \ldots, p$ and

$$-\langle g, y - f(x_0)\rangle = |\langle g, y - f(x_0)\rangle| \leq \|g\|_2 \, \|y - f(x_0)\|_2 < \|g\|_2 \rho = \varepsilon.$$

Therefore,

$$h_{f(x_0)}(f(x_0)) - h_{f(x_0)}(y) = \frac{\langle g, f(x_0) - y\rangle}{\langle g, q\rangle} \geq \frac{m_g \|f(x_0) - y\|_2}{\langle g, q\rangle} = \sigma \|f(x_0) - y\|_2$$

and so the lemma holds. □

THEOREM 6.4. *Let $L = \{e_1, e_2, \ldots, e_p\}$ be the standard basis of $\mathbb{R}^p$, $H = \operatorname{conv}(L)$, and $0 \leq \delta < \varepsilon/p$ and suppose that $x_0 \in \operatorname{AMin}\big(n_{f(x_0)} \circ f, \delta\big)$. Then $x_0 \in \operatorname{AE}(f, C_N, \delta/\beta) \; \forall\, \beta > 0$ such that $\beta < 1/p$.*

Proof. Consider $C = H + \mathbb{R}^p_+$. As H is the unit simplex, we have that $\operatorname{cone}(H) = \mathbb{R}^p_+ \backslash \{0\}$, and it follows that $\operatorname{bd}\big(\mathbb{R}^p_+\big) \cap (\mathbb{R}^p_+ \backslash \{0\}) \subset \operatorname{cone}(H)$. Then, by Lemma 4.1(ii) we see that $C(0) = \mathbb{R}^p_+ \backslash \{0\}$.

Moreover, $0 \notin \operatorname{cl}(C)$ and as B is defined by the norm $\|\ \|_\infty$, we see that $\operatorname{cl}(\beta \mathrm{B}) \cap C = \emptyset \; \forall\, \beta < 1/p$, since

$$d(0, C) = d(0, H) = 1/p.$$

By Lemma 6.2, we have that $n_{f(x_0)}$ is a strictly local $\mathbb{R}^p_+$-monotone function at $f(x_0)$ with constants $\sigma = 1$ and $\rho = \varepsilon/p$. Therefore, as $x_0 \in \operatorname{AMin}\big(n_{f(x_0)} \circ f, \delta\big)$ and $0 \leq \delta < \varepsilon/p$, by Theorem 5.10 and Remark 5.11 the conclusion follows. □

THEOREM 6.5. *Let $g \in \operatorname{int}\big(\mathbb{R}^p_+\big)$, $\varepsilon > 0$, and $0 \leq \delta < m_g \varepsilon/(\langle g, q\rangle \|g\|_2)$. If a point $x_0 \in \operatorname{AMin}\big(h_{f(x_0)} \circ f, \delta\big)$, then $x_0 \in \operatorname{AE}(f, C_H, \delta \langle g, q\rangle \|g\|_2/m_g)$.*

Proof. From Lemma 6.3 we have that $h_{f(x_0)}$ is a strictly local $\mathbb{R}^p_+$-monotone function at $f(x_0)$ with constants $\sigma = m_g/\langle g, q\rangle$ and $\rho = \varepsilon/\|g\|_2$. By Lemma 4.5(i) and (iv) we see that $C = \mathbb{R}^p_+ \cap [\langle g, \cdot\rangle > 1]$ is a pointed co-radiant set such that $C(0) = \mathbb{R}^p_+ \backslash \{0\}$, since $g \in \operatorname{int}\big(\mathbb{R}^p_+\big)$. Moreover, $0 \notin \operatorname{cl}(C)$ and if $\beta = 1/\|g\|_2$, then

$$\langle g, d\rangle \leq \|g\|_2 \|d\|_2 \leq \|g\|_2 \beta = 1 \quad \forall\, d \in \operatorname{cl}(\beta \mathrm{B})$$

and $\operatorname{cl}(\beta \mathrm{B}) \cap C = \emptyset$. Thus, by Theorem 5.10 and Remark 5.11 it follows that if $x_0 \in \operatorname{AMin}\big(h_{f(x_0)} \circ f, \delta\big)$ and $0 \leq \delta < m_g \varepsilon/(\langle g, q\rangle \|g\|_2)$, then

$$x_0 \in \operatorname{AE}(f, C_H, \delta \langle g, q\rangle \|g\|_2/m_g)\,,$$

which completes the proof. □

THEOREM 6.6. *Consider $0 \leq \delta < \varepsilon$ and suppose that $x_0 \in \operatorname{AMin}\big(t_{f(x_0)} \circ f, \delta\big)$. Then $x_0 \in \operatorname{AE}(f, C_T, \delta)$, where $\mu = (1/q_1, 1/q_2, \ldots, 1/q_p)$.*

Proof. Let $C = \mathbb{R}^p_+ \cap \operatorname{cl}(\mathrm{B})^c$ with B defined by the norm $\|\ \|^\mu_\infty$ and $\mu = (1/q_1, 1/q_2, \ldots, 1/q_p)$. It is clear that $0 \notin \operatorname{cl}(C)$ and $\operatorname{cl}(\beta \mathrm{B}) \cap C = \emptyset \; \forall\, \beta \in (0, 1]$. By Lemma 4.7(iii) we have that $C(0) = \mathbb{R}^p_+ \backslash \{0\}$, and from Lemma 6.2 it follows that $t_{f(x_0)}$ is a strictly local $\mathbb{R}^p_+$-monotone function at $f(x_0)$ by taking the norm $\|\ \|^\mu_\infty$ in $\mathbb{R}^p$ and constants $\sigma = 1$ and $\rho = \varepsilon$. Then, by Theorem 5.10 and Remark 5.11 we deduce that $x_0 \in \operatorname{AE}(f, C_T, \delta/\beta) \; \forall\, \beta \in (0, 1]$. Therefore $x_0 \in \operatorname{AE}(f, C_T, \delta)$. □

7. Conclusions. In this paper we have introduced a new ε-efficiency concept for vector optimization problems, which extends and unifies several different ε-efficiency notions previously defined in the literature.

We prove several properties of this new concept and characterize it via approximate solutions of related scalar optimization problems. These results have been

obtained in a general framework since we do not assume any convexity hypothesis. Thus, our results can be applied in nonconvex vector optimization problems and with preference structures which are not necessarily preorder relations.

As a final conclusion, we think that several results of this work can be useful in order to develop new methods to solve vector optimization problems. In this line, to obtain approximate Kuhn–Tucker conditions, approximate duality assertions and approximate Lagrange multiplier rules for the new ε-efficiency concept could be interesting.

Acknowledgment. The authors are grateful to the anonymous referees for their helpful comments and suggestions.

REFERENCES

[1] S. Deng, *On approximate solutions in convex vector optimization*, SIAM J. Control Optim., 35 (1997), pp. 2128–2136.
[2] J. Dutta and V. Vetrivel, *On approximate minima in vector optimization*, Numer. Funct. Anal. Optim., 22 (2001), pp. 845–859.
[3] C. Gerth and P. Weidner, *Nonconvex separation theorems and some applications in vector optimization*, J. Optim. Theory Appl., 67 (1990), pp. 297–320.
[4] A. Göpfert, H. Riahi, C. Tammer, and C. Zălinescu, *Variational Methods in Partially Ordered Spaces*, Springer-Verlag, New York, 2003.
[5] C. Gutiérrez, *Condiciones de ε-Eficiencia en Optimización Vectorial*, Ph.D. thesis, Universidad Nacional de Educación a Distancia, Madrid, 2004.
[6] C. Gutiérrez, B. Jiménez, and V. Novo, *Multiplier rules and saddle-point theorems for Helbig's approximate solutions in convex Pareto problems*, J. Global Optim., 32 (2005), pp. 367–383.
[7] C. Gutiérrez, B. Jiménez, and V. Novo, *A property of efficient and ε-efficient solutions in vector optimization*, Appl. Math. Lett., 18 (2005), pp. 409–414.
[8] S. Helbig, *On a New Concept for ε-Efficiency*, talk at "Optimization Days 1992," Montreal, 1992.
[9] S. Helbig and D. Pateva, *On several concepts for ε-efficiency*, OR Spectrum, 16 (1994), pp. 179–186.
[10] H. Idrissi, P. Loridan, and C. Michelot, *Approximation of solutions for location problems*, J. Optim. Theory Appl., 56 (1988), pp. 127–143.
[11] G. Isac, *The Ekeland's principle and the Pareto ε-efficiency*, in Multi-Objective Programming and Goal Programming: Theories and Applications, Lecture Notes in Econom. and Math. Systems 432, Springer-Verlag, Berlin, 1996, pp. 148–163.
[12] S. S. Kutateladze, *Convex ε-programming*, Soviet Math. Dokl., 20 (1979), pp. 391–393.
[13] Z. Li and S. Wang, *ε-approximate solutions in multiobjective optimization*, Optimization, 44 (1998), pp. 161–174.
[14] J. C. Liu, *ε-duality theorem of nondifferentiable nonconvex multiobjective programming*, J. Optim. Theory Appl., 69 (1991), pp. 153–167.
[15] J. C. Liu, *ε-Pareto optimality for nondifferentiable multiobjective programming via penalty function*, J. Math. Anal. Appl., 198 (1996), pp. 248–261.
[16] J. C. Liu, *ε-properly efficient solutions to nondifferentiable multiobjective programming problems*, Appl. Math. Lett., 12 (1999), pp. 109–113.
[17] J. C. Liu and K. Yokoyama, *ε-optimality and duality for multiobjective fractional programming*, Comput. Math. Appl., 37 (1999), pp. 119–128.
[18] P. Loridan, *ε-solutions in vector minimization problems*, J. Optim. Theory Appl., 43 (1984), pp. 265–276.
[19] P. Loridan, *ε-duality theorem of nondifferentiable nonconvex multiobjective programming*, J. Optim. Theory Appl., 74 (1992), pp. 565–566.
[20] V. L. Makarov, M. J. Levin, and A. M. Rubinov, *Mathematical Economic Theory: Pure and Mixed Types of Economic Mechanisms*, North-Holland, Amsterdam, 1995.
[21] A. B. Németh, *A nonconvex vector minimization problem*, Nonlinear Anal., 10 (1986), pp. 669–678.
[22] A. M. Rubinov and R. N. Gasimov, *Scalarization and nonlinear scalar duality for vector optimization with preferences that are not necessarily a pre-order relation*, J. Global Optim., 29 (2004), pp. 455–477.

[23] L. G. Ruhe and G. B. Fruhwirth, *ε-optimality for bicriteria programs and its application to minimum cost flows*, Computing, 44 (1990), pp. 21–34.
[24] Y. Sawaragi, H. Nakayama, and T. Tanino, *Theory of Multiobjective Optimization*, Academic Press, Orlando, FL, 1985.
[25] T. Staib, *On two generalizations of Pareto minimality*, J. Optim. Theory Appl., 59 (1988), pp. 289–306.
[26] C. Tammer, *A generalization of Ekeland's variational principle*, Optimization, 25 (1992), pp. 129–141.
[27] C. Tammer, *Stability results for approximately efficient solutions*, OR Spectrum, 16 (1994), pp. 47–52.
[28] T. Tanaka, *A new approach to approximation of solutions in vector optimization problems*, in Proceedings of APORS, 1994, M. Fushimi and K. Tone, eds., World Scientific Publishing, Singapore, 1995, pp. 497–504.
[29] I. Vályi, *Approximate solutions of vector optimization problems*, in Systems Analysis and Simulation, A. Sydow, M. Thoma, and R. Vichnevetsky, eds., Akademie-Verlag, Berlin, 1985, pp. 246–250.
[30] I. Vályi, *Approximate saddle-point theorems in vector optimization*, J. Optim. Theory Appl., 55 (1987), pp. 435–448.
[31] D. J. White, *Epsilon efficiency*, J. Optim. Theory Appl., 49 (1986), pp. 319–337.
[32] D. J. White, *Epsilon-dominating solutions in mean-variance portfolio analysis*, European J. Oper. Res., 105 (1998), pp. 457–466.
[33] A. P. Wierzbicki, *On the completeness and constructiveness of parametric characterizations to vector optimization problems*, OR Spectrum, 8 (1986), pp. 73–87.
[34] K. Yokoyama, *Epsilon approximate solutions for multiobjective programming problems*, J. Math. Anal. Appl., 203 (1996), pp. 142–149.
[35] K. Yokoyama, *Relationships between efficient set and ε-efficient set*, in Proceedings of the International Conference on Nonlinear Analysis and Convex Analysis, World Scientific, River Edge, NJ, 1999, pp. 376–380.

SIAM J. OPTIM.
Vol. 17, No. 3, pp. 711–737
© 2006 Society for Industrial and Applied Mathematics

SOLVING SECOND ORDER CONE PROGRAMMING VIA A REDUCED AUGMENTED SYSTEM APPROACH*

ZHI CAI† AND KIM-CHUAN TOH‡

To the memory of Jos Sturm

Abstract. The standard Schur complement equation-based implementation of interior-point methods for second order cone programming may encounter stability problems in the computation of search directions, and as a consequence, accurate approximate optimal solutions are sometimes not attainable. Based on the eigenvalue decomposition of the $(1,1)$ block of the augmented equation, a reduced augmented equation approach is proposed to ameliorate the stability problems. Numerical experiments show that the new approach can achieve more accurate approximate optimal solutions than the Schur complement equation-based approach.

Key words. second order cone programming, augmented equation, Nesterov–Todd direction, stability

AMS subject classifications. 90C20, 90C22, 90C51, 65K05

DOI. 10.1137/040614797

1. Introduction.

A second order cone programming (SOCP) problem is a linear optimization problem over a cross product of second order convex cones. A wide range of problems can be formulated as SOCP problems; they include linear programming (LP) problems, convex quadratically constrained quadratic programming problems, filter design problems [5, 20], and problems arising from limit analysis of collapses of solid bodies [6]. An extensive list of application problems that can be formulated as SOCP problems can be found in [14]. For a comprehensive introduction to SOCP, we refer the reader to the paper by Alizadeh and Goldfarb [1].

SOCP itself is a subclass of semidefinite programming (SDP). In theory, SOCP problems can be solved as SDP problems. However, it is far more efficient computationally to solve SOCP problems directly. A few interior-point methods (IPMs) have been developed to solve SOCP problems directly [3, 21, 25]. But these IPMs sometimes fail to deliver solutions with satisfactory accuracy. The main objective of this paper is to propose a method that can solve an SOCP to high accuracy but with comparable or moderately higher cost than the standard IPMs employing the Schur complement equation (SCE) approach. We note that global polynomial convergence results for IPMs for SOCP can be found in [15] and the references therein.

Given a column vector x_i, we will write it as $x_i = [x_i^0\,;\,\bar{x}_i]$ with x_i^0 being the first component and $\bar{x}_i$ consisting of the remaining components. Given square matrices P, Q, the notation $[P\,;\,Q]$ means that Q is appended to the last row of P, and $\mathrm{diag}(P,Q)$ denotes the block diagonal matrix with P, Q as its diagonal blocks. Throughout this paper, $\|\cdot\|$ denotes the matrix 2-norm or vector 2-norm, unless otherwise specified. For a given matrix M, we let $\lambda_{\max}(M)$ and $\lambda_{\min}(M)$ be the

*Received by the editors September 9, 2004; accepted for publication (in revised form) March 23, 2006; published electronically September 29, 2006.
http://www.siam.org/journals/siopt/17-3/61479.html

†High Performance Computing for Engineered Systems (HPCES), Singapore-MIT Alliance, 4 Engineering Drive 3, Singapore 117576, Singapore (smap0035@nus.edu.sg).

‡Department of Mathematics, National University of Singapore, 2 Science Drive 2, Singapore 117543, Singapore, and Singapore-MIT Alliance, 4 Engineering Drive 3, Singapore 117576, Singapore (mattohkc@math.nus.edu.sg).

largest and smallest eigenvalues of M in magnitude, respectively. The condition number of a matrix M (not necessarily square) is the number $\kappa(M) = \sigma_{\max}(M)/\sigma_{\min}(M)$, where $\sigma_{\max}(M)$ and $\sigma_{\min}(M)$ are the largest and smallest singular values of M, respectively. For a matrix M^μ which depends on a positive parameter μ, the notation $\|M^\mu\| = O(\mu)$ ($\|M^\mu\| = \Omega(\mu)$) means that there is a positive constant c such that $\|M^\mu\| \le c\mu$ ($\|M^\mu\| \ge c\mu$) as $\mu \downarrow 0$, and $\|M^\mu\| = \Theta(\mu)$ means that there are positive constants c_1, c_2 such that $c_1\mu \le \|M^\mu\| \le c_2\mu$ as $\mu \downarrow 0$. More generally, for a function K^μ depending on a positive parameter μ, the notation $\|M^\mu\| = \Theta(1)K^\mu$ means that there are positive constants c_1, c_2 such that $c_1K^\mu \le \|M^\mu\| \le c_2K^\mu$ for all μ sufficiently small.

Consider the following standard primal and dual SOCP problems:

$$(1)\qquad \begin{aligned} &\text{(P)} \quad \min\left\{\sum_{i=1}^N c_i^T x_i \,:\, \sum_{i=1}^N A_i x_i \,=\, b,\ x_i \succeq 0,\ i=1,\dots,N\right\},\\ &\text{(D)} \quad \max\{b^T y \,:\, A_i^T y + z_i \,=\, c_i,\ z_i \succeq 0,\ i=1,\dots,N\}, \end{aligned}$$

where $A_i \in \mathbb{R}^{m\times n_i}$, $c_i, x_i, z_i \in \mathbb{R}^{n_i}$, $i = 1,\dots,N$, and $y \in \mathbb{R}^m$. The constraint $x_i \succeq 0$ is a second order cone constraint defined by $x_i^0 \ge \|\bar{x}_i\|$. In particular, if the cone dimension n_i is 1, then the constraint is simply the standard nonnegativity constraint $x_i \ge 0$, and such a variable is called a linear variable. For convenience, we define

$$A \,=\, [A_1 \ A_2 \ \cdots \ A_N], \quad c \,=\, [c_1 \,;\, c_2 \,;\, \cdots \,;\, c_N],$$

$$x \,=\, [x_1 \,;\, x_2 \,;\, \cdots \,;\, x_N], \quad z \,=\, [z_1 \,;\, z_2 \,;\, \cdots \,;\, z_N], \quad n \,=\, \sum_{i=1}^N n_i.$$

The notation $x \succeq 0$ ($x \succ 0$) means that each x_i is in (the interior of) the ith second order cone.

In this paper, we will assume that A has full row rank, and that (P) and (D) in (1) are strictly feasible. Under these assumptions, the solutions to the perturbed KKT conditions of (1) form a path (known as the central path) in the interior of the primal-dual feasible region. At each iteration of an IPM, the Newton equation associated with the perturbed KKT conditions needs to be solved. By performing block eliminations, one can solve either a system of linear equations of size $m + n$ or one of size m. These linear systems are known as the augmented equation and the Schur complement equation (SCE), respectively. The SCE has the obvious advantage of being smaller in size as well as being symmetric positive definite. Currently, most implementations of IPMs [3, 21, 24, 25] are based on solving the SCE. However, as we shall see in section 3, the SCE can be severely ill-conditioned when the barrier parameter is close to 0. This typically causes numerical difficulties and imposes a limit on how accurately one can solve an SOCP problem.

In the case of LP, the ill-conditioning of the augmented equation was analyzed by Wright [28, 29]. Under certain assumptions including nondegeneracy, the computed search direction from the augmented equation is shown to be sufficiently accurate for the IPM to converge to high accuracy. The structure of the ill-conditioning of the SCE arising from LP was analyzed in [13]. A stabilization method based on performing Gaussian elimination with a certain pivoting order was also proposed to transform the SCE into a better-conditioned linear system of equations.

In nonlinear conic programming, however, the ill-conditioning of the augmented equation and the SCE is much more complicated than that in LP. The potential

numerical difficulties posed by the ill-conditioned SCE in SOCP were recognized by developers of solvers for SOCP (see, e.g., [3, 4, 23, 25]). It was also recognized by Goldfarb and Scheinberg [9], and that motivated them to propose and analyze a product-form Cholesky factorization for the Schur complement matrix. Subsequently, Sturm [23] implemented the product-form Cholesky factorization [9] in his very popular code SeDuMi to solve the SCE arising at each iteration of a homogeneous self-dual (HSD) IPM. SeDuMi also employed sophisticated techniques to minimize numerical cancellations when computing the SCE and its factorization [23]. These sophisticated techniques typically greatly improve the stability of the SCE approach. However, for certain extreme cases, they do not entirely ameliorate the numerical difficulties caused by the inherently ill-conditioned SCE; see section 4.

The IPM code SeDuMi differs from standard infeasible interior-point methods in that it solves the HSD embedding model. A natural question to ask is whether SeDuMi's unusually good performance arises from the inherent structure of the HSD model itself or from the sophisticated numerical techniques it uses in solving the SCE (or both). For a certain class of SOCP problems with no strictly feasible primal/dual points, we show numerically in section 4 that SeDuMi's superior performance can be explained by the structure of the HSD model itself. For some SOCP problems with strictly feasible points, we shall also see in section 4 that the sophisticated numerical techniques sometimes may offer only limited improvement in the attainable accuracy when compared to simpler techniques used to solve the SCE.

Herein we propose a method to compute the search directions based on a reduced augmented equation (RAE). This RAE is derived by applying block row operations to the augmented equation, together with appropriate partitioning of the eigenspace of its (1,1) block. The RAE is generally much smaller in size compared to the original augmented equation. By their construction, RAE-based IPMs are computationally more expensive than SCE-based IPMs. Fortunately, numerical experiments show that if sparsity in the SOCP data is properly preserved when forming the RAE, it can generally be solved rather efficiently by a judicious choice of a symmetric indefinite system solver.

The RAE-based IPMs are superior to SCE-based IPMs in that the former can usually deliver approximate optimal solutions that are much more accurate than the latter before numerical difficulties are encountered. For example, for the `schedxxx` SOCP problems selected from the DIMACS library [17], our RAE-based IPMs are able to obtain accuracies of 10^{-9} or better, while the SCE-based IPMs (SDPT3 version 3.1 and SeDuMi) can only obtain accuracies of 10^{-3} or 10^{-4} in some cases.

The paper is organized as follows. In section 2, we introduce the augmented and Schur complement equations. In section 3, we analyze the conditioning and the growth in the norm of the Schur complement matrix. We also discuss how the latter affects the primal infeasibility as the interior-point iterates approach optimality. In section 4, we present numerical results obtained from two different SCE-based primal-dual IPMs. In section 5, we derive the RAE. The conditioning of the reduced augmented matrix is analyzed in section 6. In section 7, we discuss major computational issues for efficiently solving the RAE. Numerical results for an RAE-based IPM are presented in section 8. We conclude the paper in section 9.

2. The augmented and Schur complement equations. In this section, we present the linear systems that need to be solved to compute the search direction at each IPM iteration.

For x_i in a second order cone, we define

$$\boldsymbol{aw}(x_i) \;=\; \begin{bmatrix} x_i^0 & \bar{x}_i^T \\ \bar{x}_i & x_i^0 I \end{bmatrix}, \qquad \gamma(x_i) \;=\; \sqrt{(x_i^0)^2 - \|\bar{x}_i\|^2}. \tag{2}$$

For a given barrier parameter ν, the perturbed KKT conditions of (1) in matrix form are

$$Ax \;=\; b, \quad A^T y + z \;=\; c, \quad \boldsymbol{aw}(x)\,\boldsymbol{aw}(z)e^0 \;=\; \nu e^0, \tag{3}$$

where $e^0 = [e_1\,;\, e_2\,;\, \cdots\,;\, e_N]$, with e_i being the first unit vector in $\mathbb{R}^{n_i}$. The matrix $\boldsymbol{aw}(x) \;=\; \mathrm{diag}(\boldsymbol{aw}(x_1), \ldots, \boldsymbol{aw}(x_N))$ is a block diagonal matrix with $\boldsymbol{aw}(x_1), \ldots, \boldsymbol{aw}(x_N)$ as its diagonal blocks. The matrix $\boldsymbol{aw}(z)$ is defined similarly.

For reasons of computational efficiency that we will explain later, in most IPM implementations for SOCP, a block diagonal scaling matrix is usually applied to the last equation in (3). Here, we apply the Nesterov–Todd (NT) scaling matrix [25] to produce the following equation:

$$\boldsymbol{aw}(Fx)\,\boldsymbol{aw}(F^{-1}z)e^0 = \nu e^0, \tag{4}$$

where $F = \mathrm{diag}(F_1, \ldots, F_N)$ is chosen such that $Fx \;=\; F^{-1}z \;=:\; v$. For details on the conditions that F must satisfy and on other scaling matrices, we refer the reader to [15]. Let

$$f_i \;=\; \begin{bmatrix} f_i^0 \\ \bar{f}_i \end{bmatrix} \;:=\; \frac{1}{\sqrt{2\Big(\gamma(x_i)\gamma(z_i) + x_i^T z_i\Big)}} \begin{bmatrix} \dfrac{1}{\omega_i} z_i^0 + \omega_i x_i^0 \\ \dfrac{1}{\omega_i}\bar{z}_i - \omega_i \bar{x}_i \end{bmatrix},$$

where $\omega_i = \sqrt{\gamma(z_i)/\gamma(x_i)}$. (Note that $\gamma(f_i) = 1$.) The precise form of F_i is given by

$$F_i \;=\; \omega_i \begin{bmatrix} f_i^0 & \bar{f}_i^T \\ \bar{f}_i & I + \frac{\bar{f}_i \bar{f}_i^T}{1+f_i^0} \end{bmatrix}. \tag{5}$$

Let $\mu = x^T z/N$ be the normalized complementarity gap. The Newton equation associated with the perturbed KKT conditions (3) with NT scaling is given by

$$A\Delta x \;=\; r_p, \quad A^T\Delta y + \Delta z \;=\; r_d, \quad VF\Delta x + VF^{-1}\Delta z \;=\; r_c, \tag{6}$$

where $V = \boldsymbol{aw}(v)$, $r_p = b - Ax$, $r_d = c - z - A^T y$, $r_c = \sigma\mu e^0 - Vv$. Note that we have chosen ν to be $\nu = \sigma\mu$ for some parameter $\sigma \in (0, 1)$.

The solution $(\Delta x, \Delta y, \Delta z)$ of the Newton equation (6) is referred to as the search direction. At each IPM iteration, solving (6) for the search direction is computationally the most expensive step. Observe that by eliminating Δz, the Newton equation (6) reduces to the so-called augmented equation

$$\begin{bmatrix} -F^2 & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} \;=\; \begin{bmatrix} r_x \\ r_p \end{bmatrix}, \tag{7}$$

where $r_x = r_d - FV^{-1}r_c$. The augmented equation can further be reduced in size by eliminating Δx in (7) to produce the SCE

$$\underbrace{AF^{-2}A^T}_{M}\Delta y \;=\; r_y \;:=\; r_p + AF^{-2}r_x \;=\; r_p + AF^{-2}r_d - AF^{-1}V^{-1}r_c. \tag{8}$$

The coefficient matrix M in (8) is known as the Schur complement matrix. It is symmetric positive definite if $x, z \succ 0$. The search direction corresponding to (6) always exists as long as $x, z \succ 0$. Note that if the scaling matrix F is not applied to the last equation in (3), the corresponding Schur complement matrix would be $A\,\boldsymbol{aw}(z)^{-1}\boldsymbol{aw}(x)\,A^T$, which is a nonsymmetric matrix. This nonsymmetric coefficient matrix is not guaranteed to be nonsingular even when $x, z \succ 0$. Moreover, computing its sparse LU factorization is usually much more expensive than computing the sparse Cholesky factorization of M.

In their simplest form, most current implementations of IPMs compute the search direction $(\Delta x, \Delta y, \Delta z)$ based on the SCE (8) via the following procedure.

Simplified SCE approach:

(i) Compute the Schur complement matrix M and the vector r_y;
(ii) Compute the Cholesky or sparse Cholesky factor of M;
(iii) Compute Δy by solving two triangular linear systems involving the Cholesky factor;
(iv) Compute Δz via $\Delta z = r_d - A^T\Delta y$, and Δx via $\Delta x = F^{-2}(A^T\Delta y - r_x)$.

We should note that various heuristics to improve the numerical stability of the simplified SCE approach are usually incorporated in the actual implementations. We will describe in section 4 variants of the above approach implemented in two publicly available SOCP solvers, SDPT3, version 3.1 [26], and SeDuMi, version 1.05 [22].

The SCE is preferred because it is usually a much smaller system compared to the augmented or Newton equations. Furthermore, the Schur complement matrix has the highly desirable property of being symmetric positive definite. (In contrast, the coefficient matrix in (7) is symmetric indefinite while that of (6) is nonsymmetric.) Consequently, the SCE can be solved very efficiently via Cholesky or sparse Cholesky factorization of M. We should mention that there are highly efficient and machine optimized sparse Cholesky codes readily available in the public domain, the prime example being the sparse Cholesky codes of Ng and Peyton [16]. Comparatively, the state-of-the-art LDL^T factorization codes (an example being the MA47 codes of Duff and Reid [18]) for a sparse symmetric indefinite matrix available in the public domain are less advanced.

3. Conditioning of M and the deterioration of primal infeasibility. Despite the advantages of the SCE approach described in the last section, the SCE is, however, generally severely ill-conditioned when the iterates (x, y, z) approach optimality, and this typically causes numerical difficulties. The most common numerical difficulty one may encounter in practice is that the Schur complement matrix M is numerically indefinite, although in exact arithmetic M is positive definite. Furthermore, the computed solution Δy from (8) may also be very inaccurate in that the residual norm $\|r_y - M\Delta y\|$ is much larger than the machine epsilon, and this typically causes the IPM to stall.

In this section, we will analyze the relationship between the norm $\|M\|$, the residual norm $\|r_y - M\Delta y\|$ of the computed solution Δy, and the primal infeasibility $\|r_p\|$ as the interior-point iterates approach optimality.

3.1. Eigenvalue decomposition of F^2. To analyze the norm $\|M\|$ and the conditioning of M, we need to know the eigenvalue decomposition of F^2. Recall that $F = \mathrm{diag}(F_1, \ldots, F_N)$. Thus it suffices to find the eigenvalue decomposition of F_i^2, where F_i is given in (5). By noting that for cones of dimensions $n_i \geq 2$, F_i^2 can be written as $F_i^2 = \omega_i^2(I + 2(f_i f_i^T - e_i e_i^T))$, the eigenvalue decomposition of F_i^2 can readily be found. (The case where $n_i = 1$ is easy, and $F_i^2 = z_i/x_i$.) Without going

through the algebraic details, the eigenvalue decomposition of F_i^2 is given by

$$F_i^2 = Q_i \Lambda_i Q_i^T, \qquad Q_i = \begin{bmatrix} -\frac{1}{\sqrt{2}} & +\frac{1}{\sqrt{2}} & 0 & \cdots & 0 \\ \frac{1}{\sqrt{2}}\, g_i & \frac{1}{\sqrt{2}}\, g_i & q_i^3 & \cdots & q_i^{n_i} \end{bmatrix}, \tag{9}$$

with $\Lambda_i = \omega_i^2 \operatorname{diag}\big((f_i^0 - \|\bar{f}_i\|)^2,\, (f_i^0 + \|\bar{f}_i\|)^2,\, 1, \ldots, 1\big)$, and

$$g_i := [g_i^0\,;\, \bar{g}_i] = \bar{f}_i / \|\bar{f}_i\| \in \mathbb{R}^{n_i - 1}. \tag{10}$$

Notice that since $\gamma(f_i) = 1$, the first eigenvalue is the smallest, and the second is the largest. The set $\{q_i^3, \ldots, q_i^{n_i}\}$ is an orthonormal basis of the subspace $\{u \in \mathbb{R}^{n_i - 1} : u^T g_i = 0\}$. To construct such an orthonormal basis, one may first construct the $(n_i - 1) \times (n_i - 1)$ Householder matrix H_i [10] associated with the vector g_i; then the last $n_i - 2$ columns of H_i is such an orthonormal basis. The precise form of H_i will be given later in section 6.

3.2. Analysis of $\|M\|$ and the conditioning of M. Recall that M is dependent on the normalized complementarity gap μ. Here we analyze how fast the norm $\|M\|$ and the condition number of M will grow when $\mu \downarrow 0$, i.e., when the interior-point iterates approach an optimal solution (x^*, y^*, z^*). To simplify the analysis, we will assume that strict complementarity holds at the optimal solution. Unless otherwise stated, we assume that $n_i \geq 2$ in this subsection.

Strict complementarity [2] implies that for each pair of the optimal primal and dual solutions, x_i^* and z_i^*, we have $\gamma(x_i^*) + \|z_i^*\|$ and $\gamma(z_i^*) + \|x_i^*\|$ both positive. In other words, (a) either $\gamma(x_i^*) = 0$ or $z_i^* = 0$, but not both; and (b) either $\gamma(z_i^*) = 0$ or $x_i^* = 0$, but not both. Under the strict complementarity assumption, we have the following three types of eigenvalue structures (following the classification in [9]) for F_i^2 when $\boldsymbol{aw}(x_i)\boldsymbol{aw}(z_i)e_i = \mu e_i$ and μ is small. Note that $x_i^T z_i = \mu$.

- Type 1 solution: $x_i^* \succ 0$, $z_i^* = 0$. In this case, $\gamma(x_i) = \Theta(1)$, $\gamma(z_i) = \Theta(\mu)$, and $\omega_i = \Theta(\sqrt{\mu})$. Also, $f_i^0, \|\bar{f}_i\| = \Theta(1)$, implying that all the eigenvalues of F_i^2 are $\Theta(\mu)$.
- Type 2 solution: $x_i^* = 0$, $z_i^* \succ 0$. In this case, $\gamma(x_i) = \Theta(\mu)$, $\gamma(z_i) = \Theta(1)$, and $\omega_i = \Theta(1/\sqrt{\mu})$. Also, $f_i^0, \|\bar{f}_i\| = \Theta(1)$, implying that all the eigenvalues of F_i^2 are $\Theta(1/\mu)$.
- Type 3 solution: $\gamma(x_i^*) = 0$, $\gamma(z_i^*) = 0$, $x_i^*, z_i^* \neq 0$. In this case, $\gamma(x_i), \gamma(z_i) = \Theta(\sqrt{\mu})$, and $\omega_i = \Theta(1)$. This implies that $f_i^0, \|\bar{f}_i\| = \Theta(1/\sqrt{\mu})$. Thus the largest eigenvalue of F_i^2 is $\Theta(1/\mu)$, and by the fact that $\gamma(f_i) = 1$, the smallest eigenvalue of F_i^2 is $\Theta(\mu)$. The rest of the eigenvalues are $\Theta(1)$.

Let D be the diagonal matrix consisting of the eigenvalues of F^2 sorted in ascending order. Then we have $F^2 = QDQ^T$, where the columns of Q are the sorted eigenvectors of F^2. Let D be partitioned into $D = \operatorname{diag}(D_1, D_{2a}, D_{2b})$ such that $\operatorname{diag}(D_1)$ consists of all the small eigenvalues of F^2 of order $\Theta(\mu)$, and $\operatorname{diag}(D_{2a})$ and $\operatorname{diag}(D_{2b})$ consist of the remaining eigenvalues of order $\Theta(1)$ and $\Theta(1/\mu)$, respectively. Note that the three groups of eigenvalues need not all be present. We also partition the matrix Q as $Q = [Q^{(1)}\ ,\ Q^{(2a)}\ ,\ Q^{(2b)}]$. Then $\widetilde{A} := AQ$ is partitioned as $\widetilde{A} = [\widetilde{A}_1\ ,\ \widetilde{A}_{2a}\ ,\ \widetilde{A}_{2b}] = [AQ^{(1)}\ ,\ AQ^{(2a)}\ ,\ AQ^{(2b)}]$. With the above partitions, we can express M as

$$M = \underbrace{\widetilde{A}_1 D_1^{-1} \widetilde{A}_1^T}_{M_1} + \underbrace{\widetilde{A}_{2a} D_{2a}^{-1} \widetilde{A}_{2a}^T}_{M_{2a}} + \underbrace{\widetilde{A}_{2b} D_{2b}^{-1} \widetilde{A}_{2b}^T}_{M_{2b}}. \tag{11}$$

LEMMA 3.1. (a) *For the matrices M_1, M_{2a}, and M_{2b} in* (11), *we have*

$$\|M_1\| = \Theta(1/\mu)\|\widetilde{A}_1\|^2, \quad \|M_{2a}\| = \Theta(\|\widetilde{A}_{2a}\|^2), \quad \|M_{2b}\| = \Theta(\mu)\|\widetilde{A}_{2b}\|^2.$$

(b) *Suppose there are Type* 1 *or Type* 3 *solutions so that $\widetilde{A}_1$ is not a null matrix. Then the following statements hold:* (i) $\|M\| = \Theta(1/\mu)\|\widetilde{A}_1\|^2$. (ii) *If $\widetilde{A}_1$ does not have full row rank, then* $\|M^{-1}\| = (O(1)\|\widetilde{A}_{2a}\|^2 + O(\mu)\|\widetilde{A}_{2b}\|^2)^{-1} = \Omega(1)(\|\widetilde{A}_{2a}\|^2 + \|\widetilde{A}_{2b}\|^2)^{-1}$.

Proof. (a) We shall prove only the first result since the other two can be proved similarly. By the definition of D_1, there are positive constants c_1, c_2 such that $(c_1/\mu)I \preceq D_1^{-1} \preceq (c_2/\mu)I$. Thus $(c_1/\mu)\widetilde{A}_1\widetilde{A}_1^T \preceq M_1 \preceq (c_2/\mu)\widetilde{A}_1\widetilde{A}_1^T$. This implies that $(c_1/\mu)\|\widetilde{A}_1\widetilde{A}_1^T\| \le \|M_1\| \le (c_2/\mu)\|\widetilde{A}_1\widetilde{A}_1^T\|$, and the required result follows by noting that $\|\widetilde{A}_1\widetilde{A}_1^T\| = \|\widetilde{A}_1\|^2$.

(b)(i) From (11), it is clear that $\|M\| = \Theta(1/\mu)\|\widetilde{A}_1\|^2$. (b)(ii) If $\widetilde{A}_1$ does not have full row rank, then the null space $\mathcal{N}(\widetilde{A}_1^T)$ is nontrivial. Let U be a matrix whose columns form an orthonormal basis of $\mathcal{N}(\widetilde{A}_1^T)$ and $W := U^T(M_{2a} + M_{2b})U$. Since $\widetilde{A}_1^T U = 0$, we have $U^T M U = U^T(M_{2a} + M_{2b})U = W$. By the Courant–Fischer theorem, it is clear that $\lambda_{\min}(M) \le \lambda_{\min}(W)$. Thus $\|M^{-1}\| = 1/\lambda_{\min}(M) \ge 1/\lambda_{\min}(W) = \|W^{-1}\| \ge 1/\|W\|$. Since $\|W\| \le \|M_{2a}\| + \|M_{2b}\| = O(1)\|\widetilde{A}_{2a}\|^2 + O(\mu)\|\widetilde{A}_{2b}\|^2 = O(1)(\|\widetilde{A}_{2a}\|^2 + \|\widetilde{A}_{2b}\|^2)$, the required result follows. ∎

Remark 3.1. (a) Lemma 3.1 implies that the growth in $\|M\|$ is caused by F^2 having small eigenvalues of order $\Theta(\mu)$.

(b) If $\widetilde{A}_1$ is present and does not have full row rank, then $\kappa(M) = \Omega(1/\mu)\|\widetilde{A}_1\|^2 (\|\widetilde{A}_{2a}\|^2 + \|\widetilde{A}_{2b}\|^2)^{-1}$. On the other hand, if $\widetilde{A}_1$ has full row rank (which implies that the number of eigenvalues of F^2 of order $\Theta(\mu)$ is at least m), then $\kappa(M) = \Theta(1)\kappa(\widetilde{A}_1)^2$.

(c) If there are only Type 2 solutions (thus $x^* = 0$ and $z \succ 0$), then $M = AQDQ^TA^T$ with $D = \Theta(\mu)$. In this case, we have $\kappa(M) = \Theta(1)\kappa(A)^2$.

Based on the results in [2], we have the following theorem concerning the rank of $\widetilde{A}_1$ and $\widetilde{A}_{2a}$. We refer the reader to [2] for the definitions of primal and dual degeneracies.

THEOREM 3.1. *Suppose that (x^*, y^*, z^*) satisfies strict complementarity. If the primal optimal solution x^* is primal nondegenerate, then $[\widetilde{A}_1\ ,\ \widetilde{A}_{2a}]$ has full row rank when μ is small. If the dual optimal solution (y^*, z^*) is dual nondegenerate, then $\widetilde{A}_1$ has full column rank when μ is small.*

Proof. The result follows from Theorems 20 and 21 in [1]. ∎

Remark 3.2. We should emphasize that while Theorem 3.1 says that $[\widetilde{A}_1,\ \widetilde{A}_{2a}]$ has full row rank when the optimal solution is strictly complementary and primal and dual nondegenerate, the matrix $\widetilde{A}_1$, however, does not necessarily have full row rank under the same condition. In the event that $\widetilde{A}_1$ is present and does not have full row rank, Remark 3.1(b) says that M is ill-conditioned with $\kappa(M) = \Omega(1/\mu)$. Thus even if the optimal solution is strict complementary and primal and dual nondegenerate, M does not necessarily have bounded condition number when $\mu \downarrow 0$. In contrast, for an LP problem (for which all the cones have dimensions $n_i = 1$ and $\widetilde{A}_{2a}$ is absent), primal and dual nondegeneracy ensure that $\widetilde{A}_1$ has full column and row rank, and as a result, M has bounded condition number when $\mu \downarrow 0$.

3.3. Analysis of the deterioration of primal infeasibility. Although Cholesky factorization is stable for any symmetric positive definite matrix, the conditioning

of the matrix may still affect the accuracy of the computed solution of the SCE. It is a common phenomenon that for SOCP, the accuracy of the computed search direction deteriorates as μ decreases due to an increasingly ill-conditioned M. As a result of this loss of accuracy in the computed solution, the primal infeasibility $\|r_p\|$ typically increases or stagnates when the IPM iterates approach optimality.

With the analysis of $\|M\|$ given in the last subsection, we will now analyze why the primal infeasibility may deteriorate or stagnate as interior-point iterations progress.

LEMMA 3.2. *Suppose at the kth iteration, the residual vector in solving the SCE* (8) *is* $\xi = r_y - M\Delta y$. *Assuming that* Δx *is computed exactly via the equation* $\Delta x = F^{-2}(A^T\Delta y - r_x)$, *then the primal infeasibility for the next iterate* $x^+ = x + \alpha\Delta x$, $\alpha \in [0,1]$, *is given by*

$$r_p^+ := b - Ax^+ \;=\; (1-\alpha)r_p \,+\, \alpha\xi.$$

Proof. We have $r_p^+ = (1-\alpha)r_p + \alpha(r_p - A\Delta x)$. Now $A\Delta x = AF^{-2}(A^T\Delta y - r_x) = M\Delta y - r_y + r_p$; thus $r_p - A\Delta x = r_y - M\Delta y = \xi$, and the lemma is proved. ∎

Remark 3.3. (a) In Lemma 3.2, we assume for simplicity that the component direction Δx is computed exactly. In finite precision arithmetic, errors will be introduced in the computation of Δx and that will also worsen the primal infeasibility r_p^+ of the next iterate other than $\|\xi\|$.

(b) Observe that if the SCE is solved exactly, i.e., $\xi = 0$, then $\|r_p^+\| = (1-\alpha)\|r_p\|$, and the primal infeasibility should decrease monotonically.

Lemma 3.2 implies that if the SCE is not solved to sufficient accuracy, then the inaccurate residual vector ξ may worsen the primal infeasibility of the next iterate. By standard perturbation error analysis, the worst-case residual norm of $\|\xi\|$ can be shown to be proportional to $\|M\|\|\Delta y\|$ times the machine epsilon u. The precise statement is given in the next lemma.

LEMMA 3.3. *Let* u *be the machine epsilon. Given a symmetric positive definite matrix* $B \in \mathbb{R}^{n\times n}$ *with* $(n+1)^2u \le 1/3$, *if Cholesky factorization is applied to* B *to solve the linear system* $Bx = b$ *to produce a computed solution* $\widehat{x}$, *then* $(B+\Delta B)\widehat{x} = b$, *for some* ΔB *with* $\|\Delta B\|$ *satisfying the following inequality:* $\|\Delta B\widehat{x}\| \le 3(n+1)^2u\|B\|\|\widehat{x}\|$. *Thus*

$$\|b - B\widehat{x}\| \;=\; \|\Delta B\widehat{x}\| \;=\; O(n^2)u\|B\|\|\widehat{x}\|.$$

Proof. The lemma follows straightforwardly from Theorems 10.3 and 10.4 and their extensions in [12]. ∎

Remark 3.4. Lemma 3.3 implies that if $\|B\|\|\hat{x}\|$ is large, then in the worst-case scenario, the residual norm $\|b - B\widehat{x}\|$ is expected to be proportionately large.

By Lemma 3.2 and the application of Lemma 3.3 to the SCE, we expect in the worst case the primal infeasibility $\|r_p\|$ to grow to some extent that is proportional to $\|M\|\|\Delta y\|u$. We end this section by presenting a numerical example to illustrate the relation between $\|r_p\|$ and $\|M\|\|\Delta y\|u$ in the last few iterations of an SCE-based IPM when solving the SOCP problems `rand200_800_1` and `sched_50_50_orig` (described in section 4).

The IPM we use is the primal-dual path-following method with Mehrotra predictor-corrector implemented in the MATLAB software SDPT3, version 3.1 [26]. But we should mention that to be consistent with the analysis presented in this section, the search directions are computed based on the simplified SCE approach presented in section 2, not the more sophisticated variant implemented in SDPT3.

TABLE 1
The norm of the Schur complement matrix and $\|r_p\|$ associated with the last few IPM iterations for solving the SOCP problems `rand_200_800_1` *and* `sched_50_50_orig`.

Iter	$\|M\|$	$\|M^{-1}\|$	$\mu := x^T z/N$	$\|\Delta y\|$	$\|r_y - M\Delta y\|$	$\|r_p\|$	$\frac{\|r_p\|}{\|M\|\|\Delta y\|u}$
`rand_200_800_1`							
9	1.8e+13	4.9e+02	9.2e-07	2.0e-02	8.7e-06	7.1e-06	8.9e-02
10	2.9e+14	2.3e+02	1.2e-07	2.7e-03	2.5e-05	1.8e-05	1.0e-01
11	3.8e+15	4.0e+01	1.1e-08	5.1e-04	4.9e-05	6.1e-05	1.4e-01
12	1.9e+17	6.8e+00	1.2e-09	9.4e-05	2.2e-04	1.4e-04	3.5e-02
13	1.2e+18	3.8e+01	1.8e-10	2.5e-03	3.7e-02	9.0e-04	1.4e-03
`sched_50_50_orig`							
25	5.0e+08	1.5e+04	1.5e-01	3.3e+02	7.5e-09	1.1e-06	3.0e-02
26	3.0e+09	2.1e+04	4.3e-02	2.2e+02	2.1e-07	1.4e-05	9.8e-02
27	2.7e+10	1.7e+04	9.9e-03	4.8e+01	1.9e-07	1.5e-05	5.4e-02
28	4.9e+11	1.9e+04	1.4e-03	1.0e+01	1.5e-06	1.0e-05	9.3e-03
29	5.3e+12	2.1e+04	3.3e-04	2.1e+00	1.2e-05	2.9e-04	1.2e-01

Table 1 shows the norms $\|M\|$, $\|M^{-1}\|$, $\|r_y - M\Delta y\|$ when solving the SCE (8). For this problem, $\|M\|$ and (hence $\kappa(M)$) grows like $\Theta(1/\mu)$ because its optimal solutions x_i^*, z_i^* are all of Type 3. The fifth and sixth columns in the table show that the residual norm in solving the SCE and $\|r_p\|$ deteriorate as $\|M\|$ increases. This is consistent with the conclusions of Lemmas 3.2 and 3.3. The last column further shows that $\|r_p\|$ increases proportionately to $\|M\|\|\Delta y\|u$, where the machine epsilon u is approximately 2.2×10^{-16}.

Figure 1 illustrates the phenomenon graphically for the SOCP problems `rand200_800_1` and `sched_50_50_orig`. The curves plotted correspond to the relative duality gap (`relgap`), and the relative primal and dual infeasibility (`p-inf` and `d-inf`), defined by

$$\texttt{relgap} = \frac{|c^T x - b^T y|}{1 + (|c^T x| + |b^T y|)/2}, \quad \texttt{p-inf} = \frac{\|r_p\|}{1 + \|b\|}, \quad \texttt{d-inf} = \frac{\|r_d\|}{1 + \|c\|}. \tag{12}$$

4. Computational results of two SCE-based IPMs on solving some SOCP problems. Here we present numerical results for the SCE-based IPMs implemented in the public domain solvers, SDPT3, version 3.1 [26], and SeDuMi, version 1.05 [22]. In this paper, all the numerical results are obtained in MATLAB 6.5 from a Pentium IV 2.4GHz PC with 1G RAM running a Linux operating system.

Before we analyze the performance of the SCE-based IPMs implemented in SDPT3 and SeDuMi, we must describe the methods employed to solve the SCE in both solvers. The IPM in SDPT3 is an infeasible path-following method that attempts to solve the central path equation based on (3), even if this path does not exist. It solves the resulting SCE at each IPM iteration as follows. First it computes the Cholesky or sparse Cholesky factor of the Schur complement matrix M. Then the computed Cholesky factor is used to construct a preconditioner within a preconditioned symmetric quasi-minimal residual Krylov subspace iterative solver employed to solve the SCE for Δy. The computations of Δz and Δx are the same as in the simplified SCE approach presented in section 2.

SeDuMi is a very well implemented SCE-based public domain solver for both SOCP and SDP. The IPM in SeDuMi is based not on the central path for the original

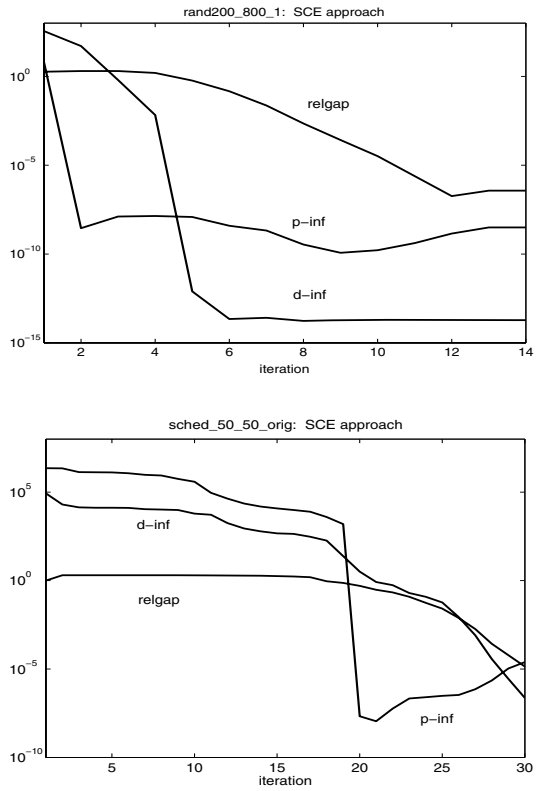


FIG. 1. *Convergence history of the SOCP problems* `rand200_800_1` *and* `sched_50_50_orig` *when solved by the SCE-based IPM in SDPT*3*, version* 3.1. *Notice that the relative primal infeasibility* `p-inf` *deteriorates as interior-point iterates approach optimality, while* `relgap` *may stagnate.*

primal and dual problems (1), but on that of the HSD model of Ye, Todd, and Mizuno [30]. The HSD model has the nice theoretical property that a strictly feasible primal and dual point always exists even if the original problems do not have one, and as a result the central path for the HSD model always exists, which is not necessarily true for the original problems in (1). As a consequence of this nice property, its solution set is always bounded. The same cannot be said for the original problems. For a problem that models an unrestricted variable by the difference of two nonnegative variables, the solution set for the original primal SOCP (P) is unbounded, and the feasible region of (D) has an empty interior, implying that the primal-dual central path does not exist. The HSD model, on the other hand, does not suffer from these defects. Thus the IPM in SeDuMi will not feel the effect of the unbounded solution set and nonexistence of the central path in the original problems in (1), but the effect of the unboundedness of the solution set on the infeasible path-following IPM in SDPT3 can be substantial and often causes serious numerical difficulties.

The computation of the search direction in SeDuMi is based on the SCE associated with the HSD model. But it employs sophisticated numerical techniques to minimize numerical cancellations in its implementation of the SCE approach [23]. It computes the Schur complement matrix in the scaled space (called the v-space) framework and transforms back and forth between quantities in the scaled and original spaces. It employs the sparse Cholesky codes adapted from Ng and Peyton [16] to compute the

factorization. It also employs the product-form Cholesky factorization [9] to handle dense columns. If the computed Cholesky factor is deemed sufficiently stable, SeDuMi will proceed to compute Δy by solving two triangular linear systems involving the Cholesky factor; otherwise, it will solve the SCE by using the preconditioned conjugate gradient iterative method with a preconditioner constructed from the Cholesky factor. Note that the Cholesky factorization has been shown in [9] to produce stable triangular factors for the Schur complement matrix if the iterates are sufficiently close to the central path and strict complementarity holds at optimality. It is important to note, however, that using a stable method to solve the SCE does not necessarily imply that the computed direction $(\Delta x, \Delta y, \Delta z)$ based on the SCE approach will produce a small residual norm with respect to the original linear system (6); see Theorem 3.2 of [11] for the case of SDP.

We tested the SCE-based IPMs in SDPT3 and SeDuMi on the following set of SOCP problems. The statistics for the test problems are shown in Table 2.

(a) The first set consists of 18 SOCP problems in the DIMACS library collected by Pataki and Schmieta [17], available at http://dimacs.rutgers.edu/Challenges/Seventh/Instances/.

(b) The second set consists of 10 SOCP problems from the FIR Filter Optimization Toolbox of Scholnik and Coleman [20], available at http://www.csee.umbc.edu/∼dschol2/opt.html.

(c) The last set consists of 10 randomly generated SOCP problems. These random problems `randxxx` are generated to be feasible and are dominated by Type 3 solutions. For each problem, the constraint matrix A has the form $V_1 \Sigma V_2^T$, where V_1, V_2 are matrices whose columns are orthonormal, and Σ is a diagonal matrix with random diagonal elements drawn from the standard normal distribution, but a few of the diagonal elements are set to 10^5 to make A moderately ill-conditioned.

In our experiments, we stop the IPM iteration in SDPT3 when any of the following situations are encountered: (1) $\max(\texttt{relgap}, \texttt{p-inf}, \texttt{d-inf}) \leq 10^{-10}$; (2) incurable numerical difficulties (such as the Schur complement matrix being numerically indefinite) occur; (3) `p-inf` has deteriorated to the extent that `p-inf` > `relgap`. SeDuMi also has a set of stopping conditions that are similar but based on the variables of the HSD model. In SeDuMi, the dual conic constraints are not strictly enforced; thus the measure `d-inf` for SeDuMi is defined to be $\texttt{d-inf} = \max(\|r_d\|, \|z^-\|)$, where $\|z^-\|$ measures how much the dual conic constraints are violated. We define

$$\phi := \log_{10}(\max\{\texttt{relgap}, \texttt{p-inf}, \texttt{d-inf}\}). \tag{13}$$

Table 2 shows the numerical results for SDPT3 and SeDuMi on 36 SOCP problems. Observe that the accuracy exponents (ϕ) for many of the problems fall short of the target of -10. For the `sched-xxx` problems, the accuracy exponents attained are especially poor, only -3 or -4 in some cases. We should mention that the results shown in Table 2 are not isolated to just the IPMs implemented in SDPT3 or SeDuMi; similar results were also reported in the SCE-based IPM implemented by Andersen, Roos, and Terlaky [3]. For example, for the problem `sched_50_50_orig`, the IPM in [3] reported the values 0.9 and 0.002 for the maximum violation of certain primal bound constraints and the dual constraints, respectively.

From Table 2, we have thus seen the performance of SCE-based IPMs for two rather different implementations in SDPT3 and SeDuMi. It is worthwhile to analyze the performance of these implementations to isolate the factor contributing to the

TABLE 2

Accuracy attained by 2 SCE-based IPMs for solving SOCP problems. The timings reported are in seconds. A number of the form "1.7-4" means 1.7×10^{-4}. An entry of the form "793 × 3" in the "SOC" column means that there are 793 three-dimensional second order cones. The numbers under the "LIN" column are the numbers of linear variables.

				SDPT3				SeDuMi				
Problem	m	SOC	LIN	ϕ	Time	relgap	p-inf	d-inf	ϕ	Time	relgap	p-inf	d-inf
nb	123	793×3	4	-3.5	6.5	2.2-4	3.3-4	8.3 -9	-11.3	12.8	6.5-13	4.8-12	2.8-15
nb-L1	915	793×3	797	-4.9	13.0	7.1-7	1.4-5	9.6-11	-12.2	14.7	6.2-13	1.2-14	1.5-14
nb-L2	123	1×1677 ; 838×3	4	-5.5	10.9	5.1-8	3.1-6	8.9-12	-9.3	33.9	5.4-10	3.1-12	9.7-12
nb-L2-bessel	123	1×123 ; 838×3	4	-6.4	6.2	3.1-7	4.3-7	8.8-11	-10.5	20.1	3.3-11	8.0-14	1.7-13
nql30	3680	900×3	3602	-4.8	5.1	1.7-5	2.6-6	3.6-13	-10.2	2.5	6.8-11	3.4-11	3.4-11
nql60	14560	3600×3	14402	-6.5	21.1	3.4-7	1.3-7	2.8-12	-10.0	11.8	1.0-10	1.1-11	1.1-11
nql180	130080	32400×3	129602	-5.3	278.3	4.9-6	8.5-7	5.7-12	-9.2	229.8	5.8-10	1.9-11	1.9-11
qssp30	3691	1891×4	2	-8.7	4.5	2.0 -9	2.3-10	2.2-14	-11.1	4.5	7.1-13	4.8-12	7.5-12
qssp60	14581	7381×4	2	-7.9	24.1	1.4-8	1.7 -9	3.4-15	-10.6	26.5	3.3-12	1.7-11	2.7-11
qssp180	130141	65341×4	2	-7.6	493.1	2.8-8	4.8-10	1.7-14	-11.2	665.9	7.0-12	1.2-12	1.8-12
sched-50-50-o	2527	1×2474 ; 1×3	2502	-4.5	5.0	1.4-5	3.2-5	2.3-7	-7.0	6.2	1.0-12	1.0-7	4.0-14
sched-100-50-o	4844	1×4741 ; 1×3	5002	-3.7	11.7	1.8-4	4.2-6	9.7 -9	-6.0	14.4	2.9-13	1.0-6	1.4-12
sched-100-100-o	8338	1×8235 ; 1×3	10002	-2.8	20.8	1.2-3	1.6-3	1.5-5	-3.3	31.0	6.6-11	4.6-4	4.1-11
sched-200-100-o	18087	1×17884 ; 1×3	20002	-3.8	77.8	2.6-5	1.7-4	3.5-8	-3.9	66.4	4.8-12	1.2-4	2.3-11
sched-50-50-s	2526	1×2475	2502	-7.2	5.0	6.0-8	8.5 -9	4.5-15	-8.2	7.7	1.0-13	7.0 -9	1.1-14
sched-100-50-s	4843	1×4742	5002	-7.7	11.7	1.0-8	2.1-8	7.2-14	-8.9	21.3	1.1-11	1.3 -9	1.2-11
sched-100-100-s	8337	1×8236	10002	-6.2	21.2	5.5-8	7.0-7	2.8-14	-7.1	34.9	3.2-12	7.8-8	5.3-15
sched-200-100-s	18086	1×17885	20002	-6.5	61.3	3.1-7	3.2-7	2.2-13	-7.8	115.8	1.2-12	1.7-8	3.8-14

TABLE 2
(cont.)

				SDPT3					SeDuMi			
Problem	m	SOC	LIN	ϕ	Time	relgap	p-inf	d-inf	ϕ	Time	relgap	p-inf	d-inf
firL1Linfalph	3074	5844×3		-10.0	235.6	7.3-11	1.1-10	0.8-15	-4.7	286.8	5.2-7	1.8-5	0.0-16
firL1Linfeps	7088	4644×3	1	-9.9	255.5	1.1-10	1.2-10	6.8-16	-10.4	106.5	3.4-13	4.3-11	1.2-14
firL1	6223	5922×3		-10.1	612.6	3.0-11	7.3-11	1.0-15	-9.0	598.4	1.8-11	1.0 -9	8.1-12
firL2a	1002	1×1003		-10.3	35.1	5.0-11	7.2-16	0.8-16	-12.6	21.3	7.8-15	2.7-13	4.5-15
firL2L1alph	5868	1×3845 ; 1922×3	1	-10.1	117.6	8.0-11	1.2-11	6.2-16	-3.3	197.6	6.2-7	5.1-4	4.4-11
firL2L1eps	4124	1×203 ; 3922×3		-10.4	181.3	3.6-11	2.1-11	0.9-15	-9.3	198.9	1.3-11	4.8-10	1.1-11
firL2Linfalph	203	1×203 ; 2942×3		-10.0	127.7	7.8-11	9.3-11	7.4-16	-9.5	237.4	4.0-12	3.5-10	8.1-14
firL2Linfeps	6086	1×5885 ; 2942×3		-10.1	369.7	7.1-11	3.8-11	6.7-16	-9.1	262.2	8.1-10	5.7-10	0.0-16
firL2	102	1×103		-11.3	0.3	5.2-12	4.6-16	1.5-16	-13.1	0.1	1.3-15	7.3-14	3.3-15
firLinf	402	3962×3		-8.9	465.4	1.2 -9	1.1 -9	1.0-15	-9.3	936.5	6.6-13	5.4-10	2.4-13
rand200-300-1	200	20×15		-7.2	2.9	5.6-8	3.0 -9	6.4-15	-6.4	7.6	3.3-7	4.3-7	0.0-16
rand200-300-2	200	20×15		-6.3	3.0	5.6-7	1.4-8	5.6-14	-5.0	12.8	7.8-6	9.0-6	0.0-16
rand200-800-1	200	20×40		-6.1	5.5	8.0-7	1.6 -9	2.0-14	-5.0	25.4	1.0-5	1.1-6	0.0-16
rand200-800-2	200	20×40		-4.7	6.0	1.9-5	1.0-8	6.8-14	-5.8	56.2	8.8-7	1.6-6	0.0-16
rand400-800-1	400	40×20		-6.5	19.0	2.9-7	1.8-8	8.7-12	-5.1	29.0	7.1-6	2.7-6	0.0-16
rand400-800-2	400	40×20		-5.6	17.7	2.6-6	8.2 -9	1.3 -9	-4.5	56.6	3.5-5	2.2-5	0.0-16
rand700-1e3-1	700	70×15		-7.1	74.4	8.1-8	1.8-8	3.0-14	-5.7	142.6	1.6-6	1.9-6	0.0-16
rand700-1e3-2	700	70×15		-5.3	80.4	5.0-6	1.1-7	6.6-14	-4.6	199.4	1.4-5	2.6-5	0.0-16
rand1000-2e3	1000	100×20		-5.7	230.4	1.9-6	1.3-8	9.4-10	-5.0	600.0	7.4-6	9.8-6	0.0-16
rand1500-3e3	1500	150×20		-7.0	812.3	1.2-8	1.0-7	8.7-14	-7.0	2119.6	9.7-8	8.8-8	0.0-16

good performance in one implementation but not the other. On the first 10 SOCP problems, `nbxxx`, `nqlxxx`, and `qsspxxx` in the DIMACS library, SeDuMi performs much better than the IPM in SDPT3 in terms of accuracy. We hypothesize that SeDuMi is able to obtain accurate approximate optimal solutions for these test problems primarily because of nice theoretical properties (existence of a strictly feasible point, and boundedness of solution set) of the HSD model. These problems contain linear variables that are the results of modeling unrestricted variables as the difference of two nonnegative vectors. Consequently, the resulting primal SOCP problems have unbounded solution sets, and the feasible regions of the dual SOCP problems have empty interior. It should come as no surprise that the IPM in SDPT3 has trouble solving such a problem to high accuracy since the ill-conditioning in the Schur complement matrix is made worse by the growing norm of the primal linear variables as the iterates approach optimality. On the other hand, for the IPM in SeDuMi, the ill-conditioning of the Schur complement matrix is not amplified since the norm of the primal variables in the HSD model stays bounded.

To verify the above hypothesis, we solve the `nbxxx`, `nqlxxx`, and `qsspxxx` problems again in SDPT3, but at each IPM iteration, we trim the growth in the primal linear variables, x^u_+, x^u_-, arising from unrestricted variables x_u using the following heuristic [26]:

$$x^u_+ \;:=\; x^u_+ - 0.8\min(x^u_+, x^u_-), \qquad x^u_- \;:=\; x^u_- - 0.8\min(x^u_+, x^u_-). \tag{14}$$

This modification does not change the original variable x^u, but it slows down the growth of x^u_+, x^u_-. After these modified vectors have been obtained, we also modify the associated dual linear variables z^u_+, z^u_- as follows if $\mu \le 10^{-4}$:

$$(z^u_+)_i \;:=\; \frac{0.5\mu}{\max(1,(x^u_+)_i)}, \qquad (z^u_-)_i \;:=\; \frac{0.5\mu}{\max(1,(x^u_-)_i)}. \tag{15}$$

Such a modification in z^u_+, z^u_- ensures that they approach 0 at the same rate as μ, and thus prevents the dual problem from attaining the equality constraints in (D) prematurely.

The results shown in Table 3 support our hypothesis. Observe that with the heuristic in (14) and (15) to control the growth of $(x^u_+)_i/(z^u_+)_i$ and $(x^u_-)_i/(z^u_-)_i$, the

TABLE 3

Performance of the SCE-based IPM in SDPT3 in solving SOCP problems with linear variables coming from unrestricted variables. The heuristics in (14) *and* (15) *are applied at each IPM iteration.*

		SDPT3					SeDuMi			
Problem	ϕ	Time	`p-inf`	`d-inf`	`relgap`	ϕ	Time	`p-inf`	`d-inf`	`relgap`
nb-u	-10.2	14.2	6.4-11	1.1-13	5.2-16	-11.1	13.6	6.5-13	8.4-12	0.0-16
nb-L1-u	-10.0	28.2	9.9-11	1.1-11	2.2-16	-12.2	15.1	6.1-13	1.0-14	1.0-14
nb-L2-u	-10.2	16.9	5.8-11	1.6-11	6.6-16	-9.3	33.8	5.4-10	3.1-12	6.5-12
nb-L2-bessel-u	-10.2	12.9	6.7-11	3.3-11	3.3-16	-10.5	20.6	3.3-11	7.9-14	1.7-13
nql30-u	-10.1	7.1	8.7-11	2.4-12	8.0-13	-10.2	3.5	6.8-11	3.4-11	2.8-11
nql60-u	-10.4	29.9	4.4-11	2.0-11	2.8-13	-10.0	12.0	1.0-10	1.1-11	8.9-12
nql180-u	-9.7	455.7	2.1-10	1.2-11	8.4-14	-9.2	263.8	5.8-10	1.9-11	1.1-11
qssp30-u	-10.0	4.3	7.4-11	9.2-11	2.7-15	-11.3	4.0	7.1-13	4.8-12	5.2-12
qssp60-u	-8.8	21.7	1.4 -9	1.8 -9	4.0-14	-10.8	26.5	3.3-12	1.7-11	1.7-11
qssp180-u	-9.0	560.9	1.1 -9	6.7-10	9.5-15	-11.2	694.2	7.0-12	1.2-12	9.9-13

IPM in SDPT3 can also achieve accurate approximate solutions, just as the IPM based on the HSD model in SeDuMi is able to achieve. It is surprising that such a simple heuristic to control the growth can result in such a dramatic improvement on the achievable accuracy, even though the problems (P) and (D) in (1) do not have a strictly feasible point and the corresponding central path does not exist.

On other problems such as `schedxxx`, `firxxx`, and `randxxx`, the performances of SDPT3 and SeDuMi are quite comparable in terms of accuracy attained, although SeDuMi is generally more accurate on the `schedxxx` problems, while SDPT3 performs somewhat better on the `randxxx` problems. On the `firxxx` problems, SDPT3 seems to be more robust, whereas SeDuMi runs into numerical difficulties quite early when solving `firL1Linfalph` and `firL2L1alph`.

The empirical evidence of Table 2 shows that even though sophisticated numerical techniques used to solve the SCE in SeDuMi can help to achieve better accuracy, sometimes these techniques give limited improvement over simpler techniques employed in SDPT3. On SOCP problems where the two solvers have vastly different performance in terms of accuracy, the difference can be attributed to the inherent IPM models used in the solvers rather than the numerical techniques employed to solve the SCE. The conclusion we may draw here is that the SCE is generally inherently ill-conditioned, and if our wish is to compute the search direction of (6) to higher accuracy, a new approach other than the SCE is necessary.

5. Reduced augmented equation. In this section, we present a new approach to compute the search direction via a potentially better-conditioned linear system of equations. Based on the new approach, the accuracy of the computed search direction is expected to be better than that computed from the SCE when μ is small. In this new approach, we assume that the iterate (x, y, z) is sufficiently close to the central path so that the eigenvalues of F^2 separate into three distinct groups as described in section 3.2.

In this approach, we start with the augmented equation in (7). By using the eigenvalue decomposition $F^2 = QDQ^T$ presented in section 3.1, where $Q = \mathrm{diag}(Q_1, \ldots, Q_N)$ and $D = \mathrm{diag}(\Lambda_1, \ldots, \Lambda_N)$, we can diagonalize the (1,1) block and rewrite the augmented equation (7) as follows:

$$\begin{bmatrix} -D & \widetilde{A}^T \\ \widetilde{A} & 0 \end{bmatrix} \begin{bmatrix} \Delta\widetilde{x} \\ \Delta y \end{bmatrix} = \begin{bmatrix} \widetilde{r} \\ r_p \end{bmatrix}, \tag{16}$$

where

$$\widetilde{A} = AQ, \quad \Delta\widetilde{x} = Q^T\Delta x, \quad \widetilde{r} = Q^T r_x. \tag{17}$$

The augmented equation (16) has dimension $m+n$, which is usually much larger than m, the dimension of the SCE. We can try to reduce its size while overcoming some of the undesirable features of the SCE such as the growth of $\|M\|$ when $\mu \downarrow 0$.

Let the diagonal matrix D be partitioned into two parts as $D = \mathrm{diag}(D_1, D_2)$ with $\mathrm{diag}(D_1)$ consisting of the small eigenvalues of F^2 of order $\Theta(\mu)$ and $\mathrm{diag}(D_2)$ consisting of the remaining eigenvalues of order $\Theta(1)$ or $\Theta(1/\mu)$. We partition the eigenvector matrix Q accordingly as $Q = [Q^{(1)}\ ,\ Q^{(2)}]$. Then $\widetilde{A}$ is partitioned as $\widetilde{A} = [\widetilde{A}_1\ ,\ \widetilde{A}_2] = [AQ^{(1)}\ ,\ AQ^{(2)}]$ and $\widetilde{r} = [\widetilde{r}_1\,;\,\widetilde{r}_2] = [(Q^{(1)})^T r_x\,;\,(Q^{(2)})^T r_x]$. Similarly, $\Delta\widetilde{x}$ is partitioned as $\Delta\widetilde{x} = [\Delta\widetilde{x}_1\,;\,\Delta\widetilde{x}_2] = [(Q^{(1)})^T\Delta x\,;\,(Q^{(2)})^T\Delta x]$.

By substituting the above partitions into (16), and eliminating $\Delta\widetilde{x}_2$, it is easy to show that solving the system (16) is equivalent to solving the following:

$$\begin{bmatrix} \widetilde{A}_2 D_2^{-1} \widetilde{A}_2^T & \widetilde{A}_1 \\ \widetilde{A}_1^T & -D_1 \end{bmatrix} \begin{bmatrix} \Delta y \\ \Delta\widetilde{x}_1 \end{bmatrix} = \begin{bmatrix} r_p + \widetilde{A}_2 D_2^{-1} \widetilde{r}_2 \\ \widetilde{r}_1 \end{bmatrix}, \tag{18}$$

$$\Delta\widetilde{x}_2 \;=\; D_2^{-1}(\widetilde{A}_2^T \Delta y - \widetilde{r}_2) \;=\; D_2^{-1}(Q^{(2)})^T(A^T\Delta y - r_x). \tag{19}$$

By its construction, the coefficient matrix in (18) does not have large elements when $\mu \downarrow 0$. But its (1,1) block is generally singular or nearly singular, especially when μ is close to 0. Since a singular (1,1) block is not conducive for symmetric indefinite factorization of the matrix or the construction of preconditioners for the matrix, we will construct an equivalent system with a (1,1) block that is less likely to be singular. Let E_1 be a given positive definite diagonal matrix with the same dimension as D_1. Throughout this paper, we take $E_1 = I$. Let $S_1 = E_1 + D_1$. By adding $\widetilde{A}_1 S_1^{-1}$ times the second block equation in (18) to the first block equation, we get $\widetilde{A}\,\mathrm{diag}(S^{-1}, D_2^{-1})\widetilde{A}^T\Delta y + \widetilde{A}_1 S_1^{-1} E_1 \Delta\widetilde{x}_1 \;=\; r_p + \widetilde{A}\,\mathrm{diag}(S^{-1}, D_2^{-1})\widetilde{r}$. This, together with the second block equation in (18) but scaled by $S_1^{-1/2}$, we get the following equivalent system:

$$\underbrace{\begin{bmatrix} \widetilde{M} & \widetilde{A}_1 S_1^{-1/2} \\ S_1^{-1/2}\widetilde{A}_1^T & -D_1 E_1^{-1} \end{bmatrix}}_{\mathcal{B}} \begin{bmatrix} \Delta y \\ S_1^{-1/2} E_1 \Delta\widetilde{x}_1 \end{bmatrix} = \begin{bmatrix} q \\ S_1^{-1/2}\widetilde{r}_1 \end{bmatrix}, \tag{20}$$

where

$$\widetilde{M} \;=\; \widetilde{A}\,\mathrm{diag}(S_1^{-1}, D_2^{-1})\widetilde{A}^T, \quad q \;=\; r_p + \widetilde{A}\,\mathrm{diag}(S_1^{-1}, D_2^{-1})\widetilde{r}. \tag{21}$$

We call the system in (20) the *reduced augmented equation* (RAE).

Note that once Δy and $\Delta\widetilde{x}_1$ are computed from (20) and $\Delta\widetilde{x}_2$ is computed from (19), Δx can be recovered through the equation $\Delta x = Q[\Delta\widetilde{x}_1\,;\,\Delta\widetilde{x}_2]$.

Remark 5.1. (a) If the matrix D_1 is null, then the RAE (20) is reduced to the SCE (8).

(b) $\mathcal{B}$ is a quasi-definite matrix [8, 27]. Such a matrix has the nice property that any symmetric reordering $\Pi\mathcal{B}\Pi^T$ has a "Cholesky factorization" $L\Lambda L^T$, where Λ is diagonal with both positive and negative diagonal elements.

Observe that the $(1,1)$ block, $\widetilde{M}$, in (20) has the same structure as the Schur complement matrix $M = \widetilde{A}\,\mathrm{diag}(D_1^{-1}, D_2^{-1})\widetilde{A}^T$. But for $\widetilde{M}$, $\|\mathrm{diag}(S_1^{-1}, D_2^{-1})\| = O(1)$, whereas for M, $\|\mathrm{diag}(D_1^{-1}, D_2^{-1})\| = O(1/\mu)$. Because of this difference, the reduced augmented matrix $\mathcal{B}$ has bounded norm as $\mu \downarrow 0$, but $\|M\|$ is generally unbounded. Under certain conditions, $\mathcal{B}$ can be shown to have a condition number that is bounded independent of the normalized complementarity gap μ. The precise statements are given in the following theorems.

THEOREM 5.1. *Suppose in* (20) *we use a partition such that* $\mathrm{diag}(D_1)$ *consists of all the eigenvalues of* F^2 *of order* $\Theta(\mu)$. *If the optimal solution of* (1) *satisfies strict complementarity, then* $\|\mathcal{B}\|$ *satisfies the inequality* $\|\mathcal{B}\| \;=\; O(1)\,\|A\|^2$. *Thus* $\|\mathcal{B}\|$ *is bounded independent of* μ *(as* $\mu \downarrow 0$*).*

Proof. It is easy to see that

$$\|\mathcal{B}\| \le \sqrt{2}\,\max(\|\widetilde{M}\| + \|\widetilde{A}_1 S_1^{-1/2}\|,\, \|S_1^{-1/2}\widetilde{A}_1^T\| + \|D_1 E_1^{-1}\|).$$

Under the assumption that the optimal solution of (1) satisfies strict complementarity, then as $\mu \downarrow 0$, $\|D_1\| \downarrow 0$, and $\|D_2^{-1}\| = O(1)$, it is possible to find a constant (independent of μ) $\tau \ge 1$ such that $\max(\|S_1^{-1}\|, \|D_2^{-1}\|, \|D_1 E_1^{-1}\|) \le \tau$. Now $\|\widetilde{M}\| \le \|\widetilde{A}\|\max(\|S_1^{-1}\|, \|D_2^{-1}\|)\|\widetilde{A}\| \le \tau\|\widetilde{A}\|^2$ and $\|S_1^{-1/2}\widetilde{A}_1^T\| = \|\widetilde{A}_1 S_1^{-1/2}\| \le \tau\|\widetilde{A}_1\|$; thus we have

$$\|\mathcal{B}\| \;\le\; \tau\sqrt{2}\max(\|\widetilde{A}\|^2 + \|\widetilde{A}_1\|,\, \|\widetilde{A}_1\| + 1) \;\le\; \tau\sqrt{2}(\|A\| + 1)^2.$$

From here, the required result follows. ∎

LEMMA 5.1. *The reduced augmented matrix* $\mathcal{B}$ *in* (20) *satisfies the following inequality:*

$$\|\mathcal{B}^{-1}\| \;\le\; 2\sqrt{2}\max(\|\widetilde{M}^{-1}\|\,,\, \|W^{-1}\|),$$

where $W = B_1^T\widetilde{M}^{-1}B_1 + D_1E_1^{-1}$ *with* $B_1 = \widetilde{A}_1 S_1^{-1/2}$.

Proof. From [19, p. 389], it can be deduced that

$$\mathcal{B}^{-1} \;=\; \begin{bmatrix} \widetilde{M}^{-1/2}(I-P)\widetilde{M}^{-1/2} & \widetilde{M}^{-1}B_1W^{-1} \\ W^{-1}B_1^T\widetilde{M}^{-1} & -W^{-1} \end{bmatrix},$$

where $P = \widetilde{M}^{-1/2}B_1W^{-1}B_1^T\widetilde{M}^{-1/2}$. Note that P satisfies the condition $0 \preceq P \preceq I$; i.e., P and $I - P$ are positive semidefinite. By the definition of W, we have $0 \preceq W^{-1/2}B_1^T\widetilde{M}^{-1}B_1W^{-1/2} \preceq I$, and thus $\|\widetilde{M}^{-1/2}B_1W^{-1/2}\| \le 1$. This implies that

$$\|\widetilde{M}^{-1}B_1W^{-1}\| \le \|\widetilde{M}^{-1/2}\|\,\|\widetilde{M}^{-1/2}B_1W^{-1/2}\|\,\|W^{-1/2}\| \;\le\; \max(\|\widetilde{M}^{-1}\|\,,\, \|W^{-1}\|).$$

It is easy to see that

$$\|\mathcal{B}^{-1}\| \le \sqrt{2}\max(\|\widetilde{M}^{-1/2}(I-P)\widetilde{M}^{-1/2}\| + \|\widetilde{M}^{-1}B_1W^{-1}\|\,,\, \|W^{-1}B_1^T\widetilde{M}^{-1}\| + \|W^{-1}\|).$$

From here, the required result follows. ∎

THEOREM 5.2. *Suppose in* (20) *we use a partition such that* $\mathrm{diag}(D_1)$ *consists of all the eigenvalues of* F^2 *of order* $\Theta(\mu)$. *If the optimal solution of* (1) *satisfies strict complementarity and the primal and dual nondegeneracy conditions defined in* [2], *then the condition number of the coefficient matrix in* (20) *is bounded independent of* μ *(as* $\mu \downarrow 0$*).*

Proof. Let D_2 be further partitioned into $D_2 = \mathrm{diag}(D_{2a}, D_{2b})$, where $\mathrm{diag}(D_{2a})$ and $\mathrm{diag}(D_{2b})$ consist of eigenvalues of F^2 of order $\Theta(1)$ and $\Theta(1/\mu)$, respectively. Let $Q^{(2)}$ and $\widetilde{A}_2$ be partitioned accordingly as $Q^{(2)} = [Q^{(2a)},\ Q^{(2b)}]$ and $\widetilde{A}_2 = [AQ^{(2a)},\ AQ^{(2b)}] =: [\widetilde{A}_{2a},\ \widetilde{A}_{2b}]$. By Theorems 20 and 21 in [1], dual nondegeneracy implies that $\widetilde{A}_1 = AQ^{(1)}$ has full column rank and primal nondegeneracy implies that $[\widetilde{A}_1,\ \widetilde{A}_{2a}]$ has full row rank. Since $\|\widetilde{M} - [\widetilde{A}_1,\ \widetilde{A}_{2a}]\mathrm{diag}(S_1^{-1},\ D_{2a}^{-1})[\widetilde{A}_1\,,\ \widetilde{A}_{2a}]^T\| = O(\mu)$, thus $\sigma_{\min}(\widetilde{M})$ is bounded away from 0 even when $\mu \downarrow 0$. This, together with the fact that $\widetilde{A}_1 S_1^{-1/2}$ has full column rank, implies that the matrix $W := S_1^{-1/2}\widetilde{A}_1^T\widetilde{M}^{-1}\widetilde{A}_1S_1^{-1/2} + D_1E_1^{-1}$ has $\sigma_{\min}(W)$ bounded away from 0 even when $\mu \downarrow 0$. By Lemma 5.1, $\|\mathcal{B}^{-1}\|$ is bounded independent of μ. By Theorem 5.1, $\|\mathcal{B}\|$ is also bounded independent of μ, and the required result follows. ∎

6. Reduced augmented equation and primal infeasibility. Let $[\xi\,;\,\eta]$ be the residual vector for the computed solution of (20).

LEMMA 6.1. *Let u be the machine epsilon and let l be the dimension of $\Delta\widetilde{x}_1$. Suppose $(l+m)u \le 1/2$ and we use Gaussian elimination with partial pivoting (GEPP) to solve* (20) *to get the computed solution $(\Delta y\,;\,\Delta\widetilde{x}_1)$; then the residual vector $[\xi\,;\,\eta]$ for the computed solution satisfies the following inequality:*

$$\|(\xi\,;\,\eta)\|_\infty \le 4(l+m)^3 u\rho\,\|\mathcal{B}\|_\infty \|(\Delta y\,;\,\Delta\widetilde{x}_1)\|_\infty,$$

where ρ is the growth factor associated with GEPP.

Proof. This lemma follows from Theorem 9.5 in [12]. □

Remark 6.1. Theorem 5.1 stated that if strict complementarity holds at the optimal solution, then $\|\mathcal{B}\|_\infty$ will not grow as $\mu \downarrow 0$ in contrast to $\|M\|$, which usually grows proportionately to $\Theta(1/\mu)$. Now because the growth factor ρ for GEPP is usually $O(1)$, Lemma 6.1 implies that the residual norm $\|(\xi\,;\,\eta)\|_\infty$ will be maintained at some level proportional to $u\|A\|^2$ even when $\mu \downarrow 0$.

Now we establish the relationship between the residual norm in solving (20) and the primal infeasibility associated with the search direction computed from the RAE approach. Suppose that in computing $\Delta\widetilde{x}_2$ from (19), a residual vector δ is introduced, i.e.,

$$\Delta\widetilde{x}_2 = D_2^{-1}(Q^{(2)})^T(A^T\Delta y - r) - \delta.$$

Then we have the following lemma for the primal infeasibility of the next iterate.

LEMMA 6.2. *Suppose Δx is computed from the RAE approach. Then the primal infeasibility $\|r_p^+\|$ for the next iterate $x^+ = x + \alpha\Delta x$, $\alpha \in [0,1]$, satisfies the following inequality:*

$$\|r_p^+\| \;\le\; (1-\alpha)\|r_p\| + \alpha\|\xi + \widetilde{A}_2\delta - \widetilde{A}_1 S_1^{-1/2}\eta\|.$$

Proof. The proof is quite routine and we omit it.

Remark 6.2. From Lemma 6.2, we see that if the RAE returns a small residual norm, then the primal infeasibility of the next iterate would not be seriously worsened by the residual norm. From Theorem 5.1 and Lemma 6.1, we expect the residual norm $\|[\xi;\eta]\|$ to be small since the upper bound on $\|\mathcal{B}\|$ is independent of μ. Also, since by its construction, D_2^{-1} does not have large elements, $\|\delta\|$ is expected to be small as well.

Figure 2 shows the convergence behavior of the IPM in SDPT3, but with search directions computed from the RAE (20) for problems `ran200_800_1` and `sched_50_50_orig`. As can be seen from the relative primal infeasibility curves, the RAE approach is more stable than the SCE approach. It is worth noting that under the new approach, the solver is able to deliver 10 digits of accuracy; i.e., $\phi \le -10$. This is significantly better than the accuracy $\phi \approx -6$ attained by the SCE approach. Note that we use a partition such that eigenvalues of F^2 that are smaller than 10^{-3} are put in D_1.

In Table 4, we show the norms $\|\mathcal{B}\|$, $\|\mathcal{B}^{-1}\|$ and the residual norm in solving the RAE (20) for the last few IPM iterations when solving the problems `rand200_800_1` and `sched_50_50_orig`. Observe that $\|\mathcal{B}\|$ and $\kappa(\mathcal{B})$ do not grow when $\mu \downarrow 0$, in contrast to $\|M\|$ and $\kappa(M)$ in Table 1. The residual norm for the computed solution of (20) remains small throughout, and in accordance with Lemma 6.1, the residual norm is approximately equal to $u\|\mathcal{B}\|$ times the norm of the computed solution. By Lemma 6.2, the small residual norm in solving the RAE explains why the primal infeasibility computed from the RAE approach does not deteriorate as in the SCE approach.

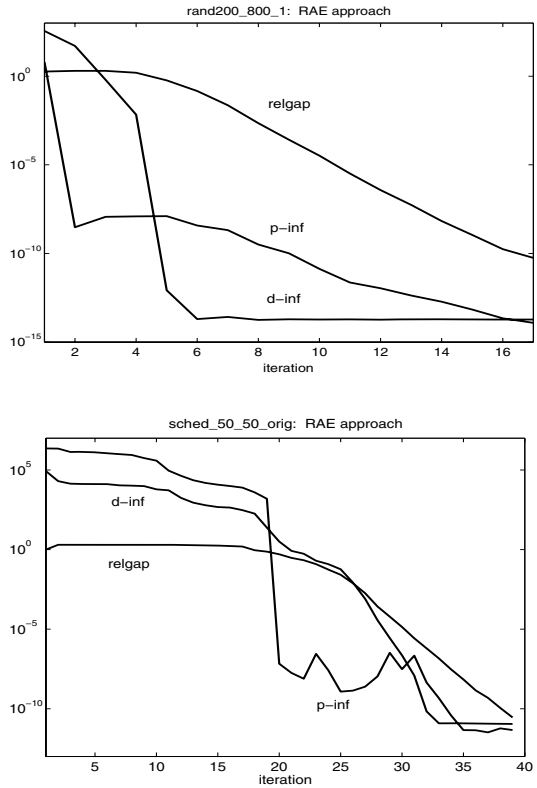


FIG. 2. *Same as Figure* 1 *except for the RAE approach in computing the search directions for the problems* `rand200_800_1` *and* `sched_50_50_orig`*. Notice that the primal infeasibility does not deteriorate when the iterates approach optimality. Both problems are primal and dual nondegenerate, and strict complementarity holds at optimality.*

TABLE 4

Condition number of the reduced augmented matrix $\mathcal{B}$ *associated with the last few IPM iterations for solving the SOCP problems* `rand200_800_1` *and* `sched_50_50_orig`*. The maximum number of columns in* $\widetilde{A}_1$ *for the former problem is* 19*, and that for the latter is* 82*.*

Iter	$\|\mathcal{B}\|$	$\|\mathcal{B}^{-1}\|$	$x^T z/N$	$\|[\Delta y; \Delta\widetilde{x}_1]\|$	Residual norm	$\|r_p\|$	$\dfrac{\|r_p\|}{\|\mathcal{B}\|\|[\Delta y; \Delta\widetilde{x}_1]\|u}$
`rand200_800_1`							
12	3.7e+11	2.7e+02	1.3e-09	4.6e-04	6.9e-09	1.8e-08	4.7e-01
13	3.4e+11	2.8e+02	1.9e-10	2.3e-04	4.3e-09	8.1e-09	4.7e-01
14	2.6e+11	3.7e+02	2.5e-11	1.0e-04	1.2e-09	2.9e-09	4.9e-01
15	2.3e+11	4.2e+02	4.0e-12	3.6e-05	3.4e-10	9.4e-10	5.1e-01
16	2.0e+11	5.2e+02	5.6e-13	1.3e-05	1.4e-10	4.8e-10	8.6e-01
`sched_50_50_orig`							
33	1.1e+08	3.9e+04	7.9e-07	1.4e-02	9.3e-13	1.2e-12	3.6e-03
34	6.2e+07	1.4e+04	1.6e-07	2.1e-03	3.9e-14	1.9e-11	6.8e-01
35	6.2e+07	2.7e+04	3.7e-08	1.7e-03	8.4e-15	1.2e-12	5.0e-02
36	6.2e+07	1.5e+04	7.5e-09	3.8e-04	2.8e-15	2.7e-11	5.1e+0
37	6.2e+07	2.2e+04	2.8e-09	2.2e-04	6.5e-16	2.5e-11	8.1e+0

7. Computational issues. The theoretical analysis in the last section indicates that the RAE approach is potentially more stable than the standard SCE approach, but the trade-off is that the former needs to solve a larger indefinite linear system. Thus, how to efficiently solve (20) is one of our major concerns in the implementation.

In forming the reduced augmented matrix $\mathcal{B}$, those operations involving Q (the eigenvector matrix of F^2) must be handled carefully by exploiting the structure of Q to avoid incurring significant storage and computational cost. Also, the sparsity of AA^T must be properly preserved when computing $\widetilde{M}$.

7.1. Computations involving Q. The operations involving Q in assembling the RAE (20) are as follows:

- Computation of the (1,1) block $\widetilde{M} = AQ\,\mathrm{diag}(S_1^{-1}, D_2^{-1})Q^TA^T$;
- Computation of the (1,2) block $\widetilde{A}_1 = AQ^{(1)}$ and the right-hand side vector $\widetilde{r} = Q^Tr_x$.

To carry out the above operations efficiently, we need to derive an explicit formula for Q to facilitate such calculations. Recall the eigenvector matrix Q_i (9) associated with the ith second order cone.

To get an explicit description of Q_i, we need to construct the Householder matrix H_i explicitly. Without going into the algebraic details, the precise form of H_i is given as follows:

(22)
$$H_i \;=\; I - h_ih_i^T, \quad h_i := \begin{bmatrix} h_i^0 \\ \bar{h}_i \end{bmatrix} = \frac{1}{\tau_i}\begin{bmatrix} \tau_i^2\,\mathrm{sign}(g_i^0) \\ \bar{g}_i \end{bmatrix} \;\in\; \mathbb{R}^{n_i-1}, \quad \tau_i := \sqrt{1+|g_i^0|}.$$

With some algebraic manipulations, the eigenvector Q_i can be rewritten in the form given in the next lemma.

LEMMA 7.1. *Let $\beta_i = -\mathrm{sign}(h_i^0)/\sqrt{2}$. We have $Q_i = \mathrm{diag}(K_i, I) - u_iv_i^T$, where*

$$K_i \;=\; \begin{bmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \beta_i & \beta_i \end{bmatrix}, \quad u_i \;=\; \begin{bmatrix} 0 \\ h_i \end{bmatrix}, \quad v_i \;=\; \begin{bmatrix} \beta_ih_i^0 \\ \beta_ih_i^0 \\ \bar{h}_i \end{bmatrix}. \tag{23}$$

Proof. Note that by construction, the first column of H_i is given by $-\mathrm{sign}(g_i^0)g_i$. Let $\alpha = \frac{1}{\sqrt{2}} + \mathrm{sign}(g_i^0)$. From (9), we have

$$\begin{aligned} Q_i \;&=\; \begin{bmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & \cdots & 0 \\ \frac{1}{\sqrt{2}}g_i & \alpha g_i & 0 & \cdots & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & I - h_ih_i^T \end{bmatrix} \\ &=\; \begin{bmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & \ldots & 0 \\ -\frac{\mathrm{sign}(g_i^0)}{\sqrt{2}} & -\frac{\mathrm{sign}(g_i^0)}{\sqrt{2}} - 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \ldots & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} - \begin{bmatrix} 0 \\ h_i \end{bmatrix}\begin{bmatrix} -\frac{\tau_i}{\sqrt{2}} \\ h_i^0 - \alpha\tau_i \\ \bar{h}_i \end{bmatrix}^T. \end{aligned}$$

It is readily shown that $h_i^0 - \alpha\tau_i = -\tau_i/\sqrt{2} = \beta_ih_i^0$. Now it is easy to see that the required results hold. ∎

Observe that each Q_i is a rank-one perturbation of a highly sparse block diagonal matrix. Based on the above lemma, those operations listed at the beginning of this subsection, except the first one, can be computed straightforwardly. To compute the matrix $\widetilde{M}$, we have to further analyze the structure of the matrix $Q_i \,\mathrm{diag}(S_{1i}^{-1}, D_{2i}^{-1}) Q_i^T$.

Let $G_i = \mathrm{diag}(S_{1i}^{-1}, D_{2i}^{-1})$ and $\Sigma_i = \mathrm{diag}(K_i, I)$; then $Q_i = \Sigma_i - u_i v_i^T$ and $Q_i\, G_i\, Q_i^T = \Sigma_i\, G_i\, \Sigma_i^T - \Sigma_i G_i v_i u_i^T - u_i v_i^T G_i \Sigma_i^T + u_i v_i^T\, G_i\, v_i u_i^T$. By setting $\rho_i = v_i^T G_i v_i$ and $\widetilde{v}_i = \Sigma_i G_i v_i / \sqrt{\rho_i}$, we have $Q_i\, G_i\, Q_i^T = \Sigma_i G_i \Sigma_i^T + l_i l_i^T - \widetilde{v}_i \widetilde{v}_i^T$, where $l_i = \widetilde{v}_i - \sqrt{\rho_i} u_i$. Thus each component matrix $\widetilde{M}_i$ in $\widetilde{M} = \sum_{i=1}^N \widetilde{M}_i$ can be expressed as

$$\widetilde{M}_i \;=\; A_i Q_i G_i Q_i^T A_i^T \;=\; A_i \left(\Sigma_i G_i \Sigma_i^T\right) A_i^T + (A_i l_i)(A_i l_i)^T - (A_i \widetilde{v}_i)(A_i \widetilde{v}_i)^T. \tag{24}$$

Since $\Sigma_i G_i \Sigma_i^T$ is a highly sparse block diagonal matrix, $\widetilde{M}_i$ is a symmetric rank-two perturbation to a sparse matrix if $A_i A_i^T$ is sparse. Hence, the computational complexity of $\widetilde{M}$ is only slightly more expensive than that for the Schur complement matrix M.

7.2. Handling dense columns. Let $\Sigma = \mathrm{diag}(\Sigma_1, \ldots, \Sigma_N)$, and

$$A_l \;=\; [A_1 l_1, \ldots, A_N l_N]\,, \quad A_v \;=\; [A_1 \widetilde{v}_1, \ldots, A_N \widetilde{v}_N]\,.$$

Then it is readily shown that

$$\widetilde{M} \;=\; A\,\Sigma\,\mathrm{diag}(S_1^{-1}, D_2^{-1})\Sigma^T\, A^T + A_l A_l^T - A_v A_v^T. \tag{25}$$

If AA^T is sparse, then the first matrix in (25) is sparse as well. For an SOCP problem where all the cones are low-dimensional, typically the matrices A_l and A_v are also sparse. In that case, the RAE (20) may be solved directly. However, if high-dimensional cones exist, then A_l and A_v invariably contain dense columns. Moreover, when A is sparse but has dense columns, AA^T will also be dense. In order to preserve the sparsity in $\widetilde{M}$, it is necessary to handle the dense columns separately when they exist.

Let P_1 be the dense columns in $A\Sigma\,\mathrm{diag}(S_1^{-1/2}, D_2^{-1/2})$ and A_l, and let P_2 be the dense columns in A_v. Let $\widetilde{M}_s = \widetilde{M} - P_1 P_1^T + P_2 P_2^T$ be the "sparse part" of $\widetilde{M}$. It is well known that by introducing the following auxiliary variables, $t_1 = P_1^T \Delta y$, $t_2 = -P_2^T \Delta y$, the dense columns can be removed from $\widetilde{M}$; see [4]. The precise form of the RAE (20) with dense column handling is as follows:

$$\begin{bmatrix} \widetilde{M}_s & U \\ U^T & -C \end{bmatrix} \begin{bmatrix} \Delta y\,;\, S_1^{-1/2} E_1 \Delta \widetilde{x}_1\,;\, t_1\,;\, t_2 \end{bmatrix} \;=\; \begin{bmatrix} q\,;\, S_1^{-1/2} \widetilde{r}_1\,;\, 0\,;\, 0 \end{bmatrix}, \tag{26}$$

where q is defined in (21) and $U = \left[\widetilde{A}_1 S_1^{-1/2},\ \ P_1,\ \ P_2\right]$, $C = \mathrm{diag}(D_1 E_1^{-1}, I_1, -I_2)$. Here I_1, I_2 are identity matrices.

7.3. Direct solvers for symmetric indefinite systems. Solving the sparse symmetric indefinite system (26) is one of the most expensive steps at each IPM iteration. Thus, it is critical that the solver used must be as efficient as possible.

We consider two methods for solving (26). The first is the Schur complement method, which is also equivalent to the Sherman–Morrison–Woodbury formula. The

second is the LDL^T factorization implemented in MA47 [18]. Each of these methods has its own advantages under different circumstances.

Schur complement method. This method is widely used for dense column handling in IPM implementations; see [4] and the references therein. It uses the sparse matrix $\widetilde{M}_s$ as the pivoting matrix to perform block eliminations in (26). It is readily shown that solving (26) is equivalent to solving the following systems:

$$\left(U^T\widetilde{M}_s^{-1}U+C\right)\left[\begin{array}{c} S_1^{-1/2}E_1\Delta\widetilde{x}_1\,;\,t_1\,;\,t_2 \end{array}\right] = U^T\widetilde{M}_s^{-1}q-\left[\begin{array}{c} S_1^{-1/2}\widetilde{r}_1\,;\,0\,;\,0 \end{array}\right], \tag{27}$$

$$\Delta y = \widetilde{M}_s^{-1}q-\widetilde{M}_s^{-1}U\left[\begin{array}{c} S_1^{-1/2}E_1\Delta\widetilde{x}_1\,;\,t_1\,;\,t_2 \end{array}\right].$$

Note that since $\widetilde{M}$ is symmetric positive definite, its "sparse part," $\widetilde{M}_s$, is typically also positive definite if the number of dense columns removed from $\widetilde{M}$ is small. If $\widetilde{M}_s$ is indeed positive definite, then (27) can be solved by Cholesky or sparse Cholesky factorization. As mentioned before, highly efficient and optimized Cholesky solvers are readily available in the public domain. Another advantage of the Schur complement method is that the symbolic factorization of $\widetilde{M}_s$ and the pivoting order of the Cholesky factorization need only be computed once or twice during the initial phase of the IPM iteration and they can be reused for subsequent IPM iterations even when the partition in D changes.

But the Schur complement method does have a major disadvantage in that the matrix $U^T\widetilde{M}_s^{-1}U+C$ is typically dense. This can lead to a huge computational burden when U has a large number of columns, say, more than a few hundred. Furthermore, the Schur complement method is numerically less stable than a method that solves (26) directly.

Roughly speaking, the Schur complement method is best suited for problems with U having a small number of columns. When U has a large number of columns or when $\widetilde{M}_s$ is not positive definite, we have to solve (26) directly by the second method described below.

MA47. MA47 is a direct solver developed by Reid and Duff [18] for sparse symmetric indefinite systems. This is perhaps the only publicly available state-of-the-art direct solver for sparse symmetric indefinite systems. It does not appear to be as efficient as the sparse Cholesky codes of Ng and Peyton [16].

The MA47 solver implements the multifrontal sparse Gaussian elimination described in [7]. In the algorithm, the pivots used are not limited to only 1×1 diagonal pivots but also include 2×2 block diagonal pivots. The solver performs a prefactorization phase (called symbolic factorization) on the coefficient matrix to determine a pivoting order so as to minimize fill-ins. In the actual factorization process, this pivoting order may be modified to obtain better numerical stability. Note that in sparse Cholesky factorization, the pivoting order is not modified after the symbolic factorization phase. Because significant overhead may be incurred when the pivoting order is modified in the factorization process, running MA47 is sometimes much more expensive than the sparse Cholesky routine of Ng and Peyton on matrices with the same dimensions and sparsity patterns.

The advantage of using MA47 to solve (26) is that it does not introduce a fully dense matrix in the solution process. Thus it is more suitable for SOCP problems with U having a relatively large number of columns.

However, the MA47 method does have a disadvantage in that the symbolic factorization of the reduced augmented matrix needs to be recomputed whenever the partition in D changes.

7.4. Partitioning strategy. As shown in section 6, the RAE approach for computing the search directions has the potential to overcome certain numerical instabilities encountered in the SCE approach. The RAE was derived from the augmented equation (7) by modifying the part of the coefficient matrix involving the small eigenvalues of F^2. Here we will describe the partition we use in $D = \mathrm{diag}(D_1, D_2)$.

The choice of D_1 is dictated by the need to strike a balance between our desire to compute more accurate search directions and to minimize the size of the RAE to be solved. For computational efficiency, it is better to have as few columns in the matrix U (26) as possible, thus suggesting that the threshold for labelling an eigenvalue as "small" should be low. But to obtain better accuracy, it is beneficial to partition eigenvalues that are smaller than, say 10^{-3}, into D_1 to improve the conditioning of the reduced augmented matrix.

With due consideration in balancing the two issues just mentioned, we adopt a hybrid strategy in computing the search direction at each IPM iteration. If $\kappa(F^2) \geq 10^6$, put the eigenvalues of F^2 that are smaller than 10^{-3} in D_1, and the rest in D_2; otherwise, put all the eigenvalues of F^2 in D_2.

Some of our test problems also contain linear blocks (i.e., cones with dimensions $n_i = 1$). In this case, $F_i^2 = z_i/x_i$ is a scalar, and we put F_i^2 in D_1 if it is smaller than 10^{-3}; otherwise, we put it in D_2.

As noted in Remark 3.1, when $\widetilde{A}_1$ has full row rank (for which a necessary condition is that the number of small eigenvalues put into D_1 is at least m), the Schur complement matrix M is not highly ill-conditioned, and it is not necessary to use the RAE approach to compute the search directions. When such a situation occurs, we use the SCE approach.

8. Numerical experiments. The RAE (20) or (26) is more expensive to solve than the SCE (8) because it is larger in size. As we have discussed in the last section, we can try to minimize the additional computational cost by a judicious choice of the solver used. If the number of columns in U is small, then using the Schur complement method to solve (26) should not be much more expensive than solving the SCE. We adopt the following heuristic rule to select the solver used to solve (26). If the number of columns in U is less than 200, we use the Schur complement method; otherwise, we use the MA47 method.

The RAE approach is implemented in MATLAB based on the IPM in SDPT3, version 3.1; see [25]. But the search direction at each iteration is computed based on the RAE (26). We use the same stopping criteria mentioned in section 4. Again, the numerical results are obtained from a Pentium IV 2.4GHz PC with 1G RAM.

We consider the same SOCP problems in section 4. But in order to focus on the comparison between the SCE and RAE approaches without the complication of unbounded primal solution sets, we exclude the `nbxxx`, `nqlxxx`, and `qsspxxx` problems from the numerical experiments in this section. Our major concerns in the experiments are efficiency and accuracy. We measure efficiency by the total CPU time taken, while accuracy is again measured by the accuracy exponent defined in (13).

The numerical results for the RAE-based IPM are presented in Table 5. In the table, T_{iter} denotes the average CPU time taken per iteration. For the RAE-based IPM, the number of IPM iterations taken for each problem is given under the column "iter." The number in each bracket gives the number of iterations using the RAE approach. The total CPU time taken to solve each problem is given under the column "Time." The number in each bracket gives the CPU time taken by the iterations using the RAE approach.

TABLE 5

A comparison between 2 SCE-based IPMs and the RAE-based IPM for solving SOCP problems. The last column in the table gives the maximum number of columns in the matrix U in (26)*. The numbers in round brackets are those taken by the iterations using the RAE approach.*

	SDPT3		SeDuMi		RAE approach							
Problem	ϕ	$T_{\rm iter}$	ϕ	$T_{\rm iter}$	ϕ	$T_{\rm iter}$	iter	Time	`relgap`	`p-inf`	`d-inf`	nc(U)
sched-50-50-o	-4.5	0.18	-7.0	0.17	-10.0	0.26	37 (27)	9.7 (7.2)	10.0-11	1.2-12	1.2-11	86
sched-100-50-o	-4.3	0.39	-6.0	0.37	-10.6	0.90	41 (26)	36.9 (30.3)	2.6-11	1.7-11	1.7-11	494
sched-100-100-o	-2.5	0.76	-3.3	0.71	-9.1	1.94	41 (24)	79.5 (65.2)	7.5-10	7.1-11	1.2-10	245
sched-200-100-o	-4.0	2.08	-3.9	1.60	-10.1	7.30	49 (38)	357.8 (323.4)	8.2-11	4.7-11	1.9-11	578
sched-50-50-s	-6.2	0.17	-8.2	0.24	-10.6	0.28	31 (26)	8.6 (7.3)	2.6-11	4.6-12	4.2-15	85
sched-100-50-s	-7.0	0.41	-8.9	0.41	-10.4	1.15	33 (24)	37.9 (33.2)	4.4-11	9.8-12	6.6-14	495
sched-100-100-s	-5.9	0.76	-7.1	0.63	-9.4	2.12	33 (28)	70.0 (63.7)	3.4-11	4.2-10	2.5-14	254
sched-200-100-s	-6.2	2.23	-7.8	2.33	-10.5	13.32	31 (25)	412.9 (388.8)	2.9-11	3.8-12	2.0-13	578
firL1Linfalph	-9.9	6.92	-4.7	10.71	-9.9	8.23	35 (18)	288.0 (153.8)	4.5-11	1.4-10	0.8-15	10
firL1Linfeps	-10.2	6.67	-10.4	3.83	-10.2	8.22	39 (32)	320.7 (232.9)	5.8-11	1.6-12	7.3-16	646
firL1	-10.1	26.65	-9.0	24.35	-10.1	28.00	23 (0)	644.0 (0.0)	2.3-11	7.3-11	1.0-15	0
firL2a	-10.3	4.37	-12.6	4.29	-10.3	4.57	8 (0)	36.6 (0.0)	5.0-11	0.8-15	0.9-16	2
firL2L1alph	-10.1	5.13	-3.3	5.24	-10.1	5.27	23 (18)	121.1 (72.4)	8.0-11	1.6-11	6.2-16	4
firL2L1eps	-10.4	9.56	-9.3	9.51	-10.5	13.86	19 (13)	263.3 (185.0)	3.4-11	2.4-11	0.9-15	1
firL2Linfalph	-10.1	3.79	-9.5	8.88	-10.1	6.60	34 (29)	224.5 (201.6)	7.8-11	6.4-14	7.3-16	19
firL2Linfeps	-10.2	16.86	-9.1	7.74	-10.1	17.23	22 (0)	379.1 (0.0)	7.1-11	4.2-12	6.7-16	2
firL2	-11.3	0.03	-13.1	0.02	-11.3	0.04	8 (3)	0.3 (0.1)	5.2-12	3.3-16	2.0-16	1
firLinf	-8.9	17.30	-9.3	34.48	-8.9	22.40	29 (18)	649.6 (443.4)	3.6-10	1.3 -9	1.0-15	170
rand200-300-1	-7.6	0.22	-6.4	0.59	-10.5	0.26	14 (5)	3.6 (1.5)	3.3-11	1.9-13	5.9-15	47
rand200-300-2	-6.0	0.23	-5.0	1.07	-10.1	0.28	16 (6)	4.5 (2.0)	8.1-11	4.6-14	5.3-14	62
rand200-800-1	-5.6	0.45	-5.0	2.11	-10.1	0.54	16 (8)	8.7 (4.7)	7.6-11	1.2-14	1.8-14	19
rand200-800-2	-4.9	0.45	-5.8	4.67	-10.1	0.55	18 (8)	9.9 (4.8)	8.1-11	1.0-13	7.6-14	19
rand400-800-1	-5.8	1.60	-5.1	2.63	-10.3	1.80	14 (6)	25.3 (11.5)	5.3-11	5.4-14	2.6-14	40
rand400-800-2	-5.8	1.68	-4.5	5.66	-10.4	1.80	15 (6)	27.0 (11.6)	3.9-11	2.0-13	5.1-14	40
rand700-1e3-1	-6.1	5.67	-5.7	10.93	-10.1	6.42	17 (8)	109.1 (55.1)	7.3-11	4.5-14	2.4-14	123
rand700-1e3-2	-5.3	5.84	-4.6	18.07	-10.1	6.59	20 (10)	131.8 (71.6)	8.6-11	5.2-13	5.8-14	151
rand1000-2e3	-5.7	20.95	-5.0	50.00	-11.3	24.66	14 (6)	345.3 (151.3)	4.7-12	1.6-13	6.9-14	100
rand1500-3e3	-7.0	67.69	-7.0	192.69	-10.1	69.27	15 (6)	1039.0 (388.9)	8.5-11	6.9-13	9.2-14	450

The numerical results in Table 2 show that the SCE-based IPMs may not deliver approximate optimal solutions with small primal infeasibilities. In Table 5, we see that the RAE-based IPM can drive the primal infeasibilities of all the problems to a level of 10^{-9} or smaller. For the `schedxxx` and `randxxx` problem sets, both the SCE-based IPMs in SDPT3 and SeDuMi cannot deliver accurate approximate solutions where the accuracies attained a range from $\phi = -2.5$ to $\phi = -8.9$ for the `schedxxx` set and from $\phi = -4.5$ to $\phi = -7.6$ for the `randxxx` set. The RAE-based IPM, however, can achieve solutions with accuracy $\phi \leq -9.1$ for all the problems in these two sets. The improvement in the attainable accuracy is more than five orders of magnitude in some cases. For the `firxxx` problems, the SCE approach can already produce accurate approximate solutions, and the RAE approach produces comparable accuracies.

The good performance in terms of accuracy of the RAE-based IPM on the `schedxxx` and `randxxx` problem sets is consistent with the theoretical results established in section 6. The SOCP problems in the `schedxxx` set are primal and dual nondegenerate, and strict complementarity holds at optimality. For the `randxxx` set, all the problems are primal nondegenerate, but four of the problems are dual degenerate. It is interesting to note that dual degeneracy does not seem to affect the performance of RAE on these degenerate problems. This fact is consistent with the observation we made in Remark 6.2.

By Theorem 5.2, the condition number of the reduced augmented matrix for the problems in the `schedxxx` set is bounded when $\mu \downarrow 0$. But as noted in Remark 3.2, strict complementarity and primal and dual nondegeneracy in an SOCP do not necessarily imply that the associated Schur complement matrix has bounded condition numbers when $\mu \downarrow 0$. The numerical results produced by the `schedxxx` problems concretely show the difference in numerical stability between the SCE and RAE approaches.

From the average CPU time taken per IPM iteration for the RAE and SCE approaches in Table 5, we see that the RAE approach is reasonably efficient in that the ratio (compared with SDPT3) is at most 6.0 for all the test problems, and 78% of them have ratios between 1.0 and 2.0.

The objective values obtained by the RAE-based IPM are given in Table 6.

TABLE 6
Primal and dual objective values obtained by the IPM using the RAE approach.

Problem	Primal objective	Dual objective	Problem	Primal objective	Dual objective
sched-50-50-o	2.6673000979 4	2.6673000977 4	firL2Linfalph	-7.0591166471 -3	-7.0591167258 -3
sched-100-50-o	1.8188993937 5	1.8188993936 5	firL2Linfeps	-1.4892049051 -3	-1.4892049762 -3
sched-100-100-o	7.1736778669 5	7.1736778615 5	firL2	-3.1186645862 -3	-3.1186645914 -3
sched-200-100-o	1.4136044650 5	1.4136044649 5	firLinf	-1.0068176528 -2	-1.0068176895 -2
sched-50-50-s	7.8520384401 0	7.8520384399 0	rand200-300-1	-1.5094030119 2	-1.5094030119 2
sched-100-50-s	6.7165031103 1	6.7165031100 1	rand200-300-2	-1.2861024800 2	-1.2861024801 2
sched-100-100-s	2.7330785593 1	2.7330785592 1	rand200-800-1	1.8086048337 0	1.8086048335 0
sched-200-100-s	5.1811961028 1	5.1811961027 1	rand200-800-2	-2.3277765222 1	-2.3277765220 1
firL1Linfalph	-3.0673166232 -3	-3.0673166686 -3	rand400-800-1	6.6607764189 0	6.6607764185 0
firL1Linfeps	-2.7112896665 -3	-2.7112897249 -3	rand400-800-2	6.3708631137 1	6.3708631135 1
firL1	-2.9257813804 -4	-2.9257816083 -4	rand700-1e3-1	-7.1501954797 1	-7.1501954802 1
firL2a	-7.1457742547 -4	-7.1457747536 -4	rand700-1e3-2	-5.5374169002 1	-5.5374169007 1
firL2L1alph	-5.7634914619 -5	-5.7634994782 -5	rand1000-2e3	-2.4138366508 4	-2.4138366508 4
firL2L1eps	-8.4481294535 -4	-8.4481297976 -4	rand1500-3e3	1.7396653464 4	1.7396653465 4

TABLE 7

Primal nondegeneracy ("p.n.d") and dual nondegeneracy ("d.n.d") and strict complementarity ("s.c.") information of approximate solutions of some SOCP problems. A "1" means true and a "0" means false. A number of the form (34/35) *in the second column means that at the computed approximate optimal solution, the column rank of* $\widetilde{A}_1$ *is* 34*, and the number of columns in* $\widetilde{A}_1$ *is* 35*.*

Problem	p.n.d	d.n.d	s.c.	Problem	p.n.d	d.n.d	s.c.
sched-50-50-orig	1	1 (79/79)	1	rand200_800_1	1	1 (19/19)	1
sched-50-50-scaled	1	1 (83/83)	1	rand200_800_2	1	1 (19/19)	1
firL2a	1	1 (1/1)	1	rand400_800_1	1	1 (40/40)	1
firL2Linfalph	1	1 (15/15)	1	rand400_800_2	1	1 (40/40)	1
firL2	1	1 (1/1)	1	rand700_1e3_1	1	0 (84/85)	1
rand200_300_1	1	0 (34/35)	1	rand700_1e3_2	1	0 (126/130)	1
rand200_300_2	1	0 (62/65)	1				

As we are able to compute rather accurate approximate solutions for (1), it is worthwhile to gather information such as primal and dual degeneracy and strict complementarity for some of the smaller SOCP problems we have considered in this paper. Such information is given in Table 7. We note that the degeneracies of the problems are determined by computing the numerical row and column rank (via the MATLAB command `rank`) of the matrices in Theorems 20 and 21 in [1], respectively.

9. Conclusion. We analyzed the accuracy of the search direction computed from the SCE approach, and how the residual norm in the computed solution affects the primal infeasibility and hence the achievable accuracy in the approximate optimal solution.

We also discussed the factors contributing to the good numerical performance of the very well implemented SCE-based IPM in the software SeDuMi.

A reduced augmented equation is proposed to compute the search direction at each IPM iteration when the SCE cannot be solved to sufficient accuracy. The proposed RAE approach can improve the robustness of IPM solvers for SOCP. It can be implemented efficiently by carefully preserving the sparsity structure in the problem data. Numerical results show that the new approach can produce more accurate approximate optimal solutions compared to the SCE approach.

Acknowledgments. The authors are grateful to Professor Robert Freund for valuable suggestions on the manuscript. The authors also thank the associate editor, Professor M. J. Todd, and the referees for many helpful comments and suggestions on improving the paper.

REFERENCES

[1] F. ALIZADEH AND D. GOLDFARB, *Second-order cone programming*, Math. Program., 95 (2003), pp. 3–51.

[2] F. ALIZADEH AND S. H. SCHMIETA, *Optimization with Semidefinite, Quadratic and Linear Constraints*, Retcor Research Report 23-97, Rutgers Center for Operations Research, Rutgers University, Piscataway, NJ, 1997; available online from http://rutcor.rutgers.edu/pub/rrr/reports97/23.ps.

[3] E. D. ANDERSEN, C. ROOS, AND T. TERLAKY, *On implementing a primal-dual interior-point method for conic quadratic optimization*, Math. Program., 95 (2003), pp. 249–277.

[4] K. D. ANDERSEN, *A modified Schur complement method for handling dense columns in interior point methods for linear programming*, ACM Trans. Math. Software, 22 (1996), pp. 348–356.

[5] K. M. Tsui, S. C. Chan, and K. S. Yeung, *Design of FIR digital filters with prescribed flatness and peak error constraints using second order cone programming*, IEEE Trans. Circuits Syst. II, 52 (2005), pp. 601–605.
[6] H. Ciria and J. Peraire, *Computation of upper and lower bounds in limit analysis using second-order cone programming and mesh adaptivity*, in Proceedings of the 9th ASCE Specialty Conference on Probabilistic Mechanics and Structural Reliability, 2004.
[7] I. S. Duff and J. K. Reid, *The multifrontal solution of indefinite sparse symmetric linear equations*, ACM Trans. Math. Software, 9 (1983), pp. 302–325.
[8] A. George and KH. Ikramov, *On the condition of symmetric quasi-definite matrices*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 970–977.
[9] D. Goldfarb and K. Scheinberg, *Product-form Cholesky factorization in interior point methods for second-order cone programming*, Math. Program., 103 (2005), pp. 153–179.
[10] G. Golub and C. Van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, MD, 1996.
[11] M. Gu, *Primal-dual interior-point methods for semidefinite programming in finite precision*, SIAM J. Optim., 10 (2000), pp. 462–502.
[12] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 1996.
[13] V. Kovacevic-Vujcic and M. D. Asic, *Stabilization of interior-point methods for linear programming*, Comput. Optim. Appl., 14 (1999), pp. 331–346.
[14] M. S. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret, *Applications of second-order cone programming*, Linear Algebra Appl., 284 (1998), pp. 193–228.
[15] 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.
[16] E. G. Ng and B. W. Peyton, *Block sparse Cholesky algorithms on advanced uniprocessor computers*, SIAM J. Sci. Comput., 14 (1993), pp. 1034–1056.
[17] G. Pataki and S. H. Schmieta, *The DIMACS Library of Mixed Semidefinite-Quadratic-Linear Programs*, http://dimacs.rutgers.edu/Challenges/Seventh/Instances/ (2000).
[18] J. Reid and I. S. Duff, *MA47, A Fortran Code for Direct Solution of Indefinite Sparse Symmetric Linear Systems*, Report RAL-95-001, Rutherford Appleton Laboratory, Oxfordshire, England, 1995.
[19] Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS Publishing, Boston, 1996.
[20] D. Scholnik and J. Coleman, *An FIR Filter Optimization Toolbox for Matlab* 5–7, http://www.csee.umbc.edu/∼dschol2/opt.html (2004).
[21] J. F. Sturm, *Using SeDuMi* 1.02, *a MATLAB toolbox for optimization over symmetric cones*, Optim. Methods Softw., 11/12 (1999), pp. 625–653.
[22] J. F. Sturm, *Implementation of interior point methods for mixed semidefinite and second order cone optimization problems*, Optim. Methods Softw., 17 (2002), pp. 1105–1154.
[23] J. F. Sturm, *Avoiding numerical cancellation in the interior point method for solving semidefinite programs*, Math. Program., 95 (2003), pp. 219–247.
[24] K. C. Toh, M. J. Todd, and R. H. Tütüncü, *SDPT*3*—A MATLAB software package for semidefinite programming*, Optim. Methods Softw., 11/12 (1999), pp. 545–581.
[25] R. H. Tütüncü, K. C. Toh, and M. J. Todd, *Solving semidefinite-quadratic-linear programs using SDPT*3, Math. Program., 95 (2003), pp. 189–217.
[26] K. C. Toh, R. H. Tütüncü, and M. J. Todd, *On the implementation of SDPT*3 *(version* 3.1*)—A MATLAB software package for semidefinite-quadratic-linear programming*, invited paper, 2004 IEEE Conference on Computer-Aided Control System Design, Taipei, Taiwan.
[27] R. J. Vanderbei, *Symmetric quasidefinite matrices*, SIAM J. Optim., 5 (1995), pp. 100–113.
[28] S. J. Wright, *Stability of linear equations solvers in interior-point methods*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 1287–1307.
[29] S. Wright, *Stability of augmented system factorizations in interior-point methods*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 191–222.
[30] Y. Ye, M. J. Todd, and S. Mizuno, *An* $O(\sqrt{n}L)$*-iteration homogeneous and self-dual linear programming algorithm*, Math. Oper. Res., 19 (1994), pp. 53–67.

SIAM J. OPTIM.
Vol. 17, No. 3, pp. 738–775
© 2006 Society for Industrial and Applied Mathematics

MODELING COMPENSATION FOR OPTICAL FIBER COMMUNICATION SYSTEMS*

JOHN ZWECK† AND SUSAN E. MINKOFF†

Abstract. Today the vast majority of telecommunication and Internet messages are sent along fiber optic cables buried underground. Binary data (encoded as a sequence of pulses of light) may travel thousands of kilometers to reach its final destination. The fibers that are used for this data transfer necessarily contain manufacturing impurities that lead to fast and slow polarization states for the propagating signal. This imperfection in the fiber results in a random distortion effect known as polarization-mode dispersion (PMD). As binary data travels along these fibers, the pulses spread, causing the ones to decrease in value and the zeros to increase. Thus, the received message may contain errors. To decrease the likelihood of errors in the received signal, a device known as a compensator can be placed at the receiver. Determining an optimal setting for the compensator involves rotating the fiber in the compensator to best align its slow axis with the fast axis of the transmission fiber. Such a rotation should cancel out some of the effects of PMD. Modeling this system numerically requires that one generate fiber realizations with large amounts of PMD. To measure rotation angle goodness of fit between compensation and transmission fiber requires that one choose a feedback signal for the compensator. We compare the eye opening, spectral line, and degree of polarization ellipsoid feedback signals. While the eye opening feedback mechanism is the most accurate measure, it is difficult to optimize numerically. The degree of polarization and spectral line feedback signals act as smooth surrogates for the eye.

Key words. optical fiber communication, photonics, PDE-constrained optimization, polarization-mode dispersion, Monte Carlo methods

AMS subject classifications. 65C05, 78A48, 78A50, 90B18, 90C90, 94A99

DOI. 10.1137/050632610

1. Introduction.

The vast majority of long-distance telecommunications and Internet traffic is carried by optical fiber communication systems [31]. In an optical communication system binary data is encoded onto a sequence of pulses of light which are then transmitted over long distances through optical fiber. A material property of optical fiber called birefringence causes the pulses to spread and distort as they propagate. This distortion of the optical signal is called polarization-mode dispersion (PMD) and is governed by the linear-PMD equation, which is a special case of the Manakov-PMD equation. The Manakov-PMD equation models the propagation of light through dispersive, nonlinear, birefringent optical fiber and is derived from Maxwell's equations [25], [43]. In the 1980s scientists realized that PMD in optical fibers would have a significant impact on the performance of high-data rate systems [18], [30]. By the time the signal reaches its destination it may not be possible to correctly decode the transmitted binary message; i.e., bit errors may occur. The engineering goal is to ensure that the bit-error ratio, which is the probability that a bit error occurs, is as small as possible (typically 10^{-9}–10^{-15}). A major difficulty in achieving this goal is that the birefringence of optical fiber, and hence the bit-error ratio, varies randomly over time. To reduce the probability of a large bit-error ratio, engineers have proposed using devices called optical PMD compensators.

*Received by the editors May 27, 2005; accepted for publication (in revised form) March 1, 2006; published electronically October 3, 2006. This work was supported by a grant from the NASA Goddard/UMBC Center for Advanced Study of Photonics Research.

http://www.siam.org/journals/siopt/17-3/63261.html

†Department of Mathematics and Statistics, University of Maryland Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250 (zweck@math.umbc.edu, sminkoff@math.umbc.edu).

In birefringent fiber, the speed of the light depends on its polarization state. To first approximation, a given optical fiber has two special principal states of polarization: fast and slow. The delay in the arrival times between light traveling in these two states is called the differential group delay (DGD). If the DGD is too large, bit errors will occur. A PMD compensator is a device that is placed after the transmission fiber just prior to the receiver. It is designed to reduce the deleterious effect that PMD has on the performance of a communication system. A *simple PMD compensator* consists of a device called a polarization controller that is used to change the polarization state of the light, followed by a piece of compensating fiber with a fixed DGD. The polarization controller rotates the fast polarization state of the transmission fiber onto the slow polarization state of the compensation fiber, thereby reducing the total DGD. In this paper, we study the problem of how best to optimize the performance of a simple PMD compensator. For each random realization of the birefringence of the transmission fiber, the goal is to find the rotation of the polarization controller that will minimize the bit-error ratio.

The physical and statistical properties of PMD, the equations that govern the propagation of light through birefringent optical fiber, and Monte Carlo-based models for the random variation in the birefringence of optical fiber have been widely studied over the last 25 years, and several excellent review articles have recently appeared [24], [31], [43].

Because the random fiber realizations that produce unacceptably large PMD are extremely rare, it is very difficult to observe them experimentally. It is also not practical to simulate them using a standard Monte Carlo search of the state space of all possible fiber realizations. However, it is precisely these rare, large-PMD fiber realizations that are most important to consider when assessing the effectiveness of a PMD compensator. Recently, Biondini and Kath [2], [3], [4] developed a multiple importance sampling algorithm that uses biased Monte Carlo simulations to efficiently generate realizations of the fiber which have large amounts of PMD. This advance made it possible to perform simulation studies that more accurately assess the performance of PMD compensators [35], [36], [38], [39], [64].

In recent years several different approaches have been proposed for reducing bit errors due to PMD. Comprehensive reviews of these ideas can be found in [7], [26], [52]. One approach is to install newly designed low-PMD fiber [45]. However, replacing fiber is prohibitively expensive for existing systems. Another approach is to design the shape of the transmitted light pulses to make the signal more resilient to PMD [52]. In addition to these passive methods, active PMD compensation methods have also been proposed. Active compensation techniques can be applied to the optical signal either just before it enters the receiver (optical compensation) or after the optical signal has been converted back into an electrical signal in the receiver (electrical compensation) [39].

We focus in this paper on optical PMD compensation only. Optical PMD compensation is complicated by the fact that the DGD and principal states depend on the frequency of the light. If the PMD in the transmission fiber were actually frequency-independent, then a simple optical PMD compensator like that described above could completely eliminate the effects of PMD. In the realistic case of frequency-dependent PMD it is still theoretically possible (but not experimentally viable) to completely eliminate the effects of PMD via solution of an inverse problem [52]. Compensators that account for at least some frequency-dependent PMD have a relatively large number of degrees of freedom; i.e., their objective functions are defined on a high-dimensional space. However, such devices are costly to build and operate. Since low

cost is desirable, we chose to study the simple optical PMD compensator described above, which has only two degrees of freedom.

An obvious choice for the feedback mechanism to use for compensation is the bit-error ratio. However, it is not actually possible to compute the bit-error ratio in a real system. Instead, for a simple compensator, the most common feedback signals are the power in a spectral line [27] and the degree of polarization (DOP) [33]. The goal is to maximize these feedback signals since small values of the DGD usually result in large values of the monitored signal. In addition, it has been shown that the performance of a compensator can be improved by scrambling the polarization state of the input signal [47]. Therefore, the monitor signals we chose to study for this paper are the spectral line and polarization-scrambled DOP. We compare these two feedback mechanisms to a third—the eye opening [8]—which is highly correlated to the bit-error ratio. Unfortunately, the eye opening monitor is not very practical since it is difficult to build and operate (requiring complex fast electronic circuitry).

Several simulation studies have compared the performance of different PMD compensators. Sunnerud et al. [52], [53] and Buchali and Bülow [7] compared compensators with a few (2–5) degrees of freedom. Their results show that compensators with three or more degrees of freedom are somewhat more effective than the simple compensator. Although they used the spectral line and DOP monitors (without polarization scrambling) they did not study the properties of these objective functions or compare them to the eye opening monitor. Moreover, because they used standard Monte Carlo simulations they were not able to access the rare large-PMD fibers that are of real interest when quantifying the performance of a compensator. I. Lima et al. [39] used multiple importance sampling to study the performance of a simple compensator with a fixed DGD. They showed that the optimal value for the fixed DGD in the compensator is about 2–3 times larger than the mean DGD of the transmission line, averaged over fiber realizations. However, they only used the eye opening objective function in their work. In addition, none of the papers just cited carefully studies how the performance of a compensator depends on the choice of algorithm used to optimize the objective function.

We compare the spectral line, polarization-scrambled DOP, and eye opening objective functions, both for a particular fiber realization and statistically over many fiber realizations. In the special case that the PMD is frequency-independent, we derive analytical formulae for the spectral line and polarization-scrambled DOP objective functions. Each objective function can be regarded as a doubly periodic function on a two-dimensional plane that parametrizes a certain set of rotations of three-dimensional space. Our formulae show that the polarization-scrambled DOP has a single maximum, while the spectral line can have at least two maxima. However, there may be more local maxima in the general case of frequency-dependent PMD.

We also systematically study the combined effect that the choice of feedback signal and optimization algorithm have on the performance of a simple PMD compensator. By using importance sampling with a large number of fiber realizations, we are able to assess performance for the very rare large-PMD realizations of the fiber that are most important to consider when studying PMD compensators. Because the eye opening is so highly correlated to the bit-error ratio, it is very reasonable to assume that the best performance is obtained when the rotation of the polarization controller is given by the global maximum of the eye opening objective function. However, it is difficult to locate the global maximum of the eye opening. On the other hand, we will show that it is much easier to locate the global maxima for the spectral line and polarization-scrambled DOP, which indicates that these two objective functions are

smoother than the eye opening.

The main conclusion of our study is that the spectral line and polarization-scrambled DOP act as smooth surrogates for the eye opening (or bit-error ratio) objective function. This result is obtained by comparing the performance of the compensators with different local and global optimization algorithms, and by statistical studies that compare the location of local and global maxima of the three different objective functions. We show that the best trade-off between computational cost and performance is obtained using the polarization-scrambled DOP objective function with a multilevel optimization algorithm. This algorithm uses a global genetic algorithm followed by a local conjugate gradient algorithm. Preliminary work [64] compared the performance of the simple compensator with the three objective functions but used only local optimization algorithms and did not compare the structure of the objective functions.

In section 2, we review the basic mathematical models for optical fiber communications systems with PMD. In section 3, we describe the PMD compensator we study in this work, and in section 4 we derive formulae for the polarization-scrambled DOP and spectral line objective functions in the special case of only first-order PMD. In section 5 we review the importance sampling algorithm and in section 6 we describe the optimization algorithms we used. Finally, in section 7 we present the results of our optimization case study.

2. Mathematical models for optical communication systems. In this section, we review the linear-PMD equation and the coarse-step algorithm that is used to generate random realizations of the fiber. We also explain how to calculate and measure the performance of an optical communication system.

2.1. The governing equations. In this subsection, we review the Manakov-PMD equation that describes the propagation of an optical signal through birefringent optical fiber. The birefringence of optical fiber, which is the physical cause of PMD in the signal, varies randomly along the fiber and over the course of time due to temperature variations and mechanical vibrations.

Light in an optical fiber propagates in two eigenmodes which are distinguished from each other by their polarization states. To model the propagation of light in an optical fiber, we choose coordinates, (x, y, z), so that the positive z-axis is the propagation direction along the fiber and the (x, y)-plane is orthogonal to the fiber. The electric field of light propagating at a carrier frequency, ω_0, is the real part of the complex-valued vector field, $\mathbf{E}$, which we express as [43]

$$\mathbf{E}(x, y, z, \tau) = \kappa \left[U_1(z, \tau)\mathbf{R}_1 + U_2(z, \tau)\mathbf{R}_2\right] \exp[i\beta(\omega_0, z)z - i\omega_0\tau]. \tag{2.1}$$

Here τ denotes physical time, and the dispersion relation of the fiber is determined by the frequency-dependence of the wavenumber, $\beta(\omega_0, z)$. The vector fields $\mathbf{R}_1$ and $\mathbf{R}_2$ are two eigenfunctions that describe the (x, y)-dependence of the electric field, $\mathbf{E}$. These vector fields are orthogonal to the propagation direction z, and to each other. The functions U_1 and U_2 describe the slow variation of the envelope of the electric field about the rapidly varying carrier wave given by the exponential factor in (2.1). The column vector $\mathbf{U}(z, t) = (U_1(z, t), U_2(z, t))^T$, which is called the *Jones vector* of the light, models the *polarization state* of the light at (z, t). Here we have transformed from physical time, τ, to retarded time, $t = \tau - \frac{\partial\beta}{\partial\omega}(\omega_0, z)z$, which defines a coordinate system that is moving with the group velocity $[\frac{\partial\beta}{\partial\omega}(\omega_0, z)]^{-1}$. The normalization constant κ is chosen so that $|\mathbf{U}|^2 = |U_1|^2 + |U_2|^2$ corresponds to the *optical power* of the

signal. The data is encoded onto the signal by allocating a time slot to each bit and varying the power of the signal so that the power is large in the time slots allocated to the ones and small in the time slots allocated to the zeros.

If the refractive index is perfectly axially symmetric, then the two eigenmodes are equal and the signal is not affected by PMD. However, in real fiber this degeneracy is broken due to imperfections in the fiber. Consequently, real optical fiber has a small *birefringence*: Light propagating in the two different eigenmodes travels at slightly different group velocities. Therefore, if the power of an optical pulse is split between the two polarization eigenmodes, $\mathbf{R}_1$ and $\mathbf{R}_2$, then as it propagates through the fiber, the power will spread in the time domain and can become severely distorted. This phenomenon is called *polarization-mode dispersion* (PMD). As a result, optical power will be transferred between the time slots allocated to the different bits, potentially resulting in bit errors at the receiver. PMD is particularly difficult to combat because it is inherently stochastic in nature. As the distance, z, along the fiber increases, the eigenfunctions $\mathbf{R}_1$ and $\mathbf{R}_2$ rotate rapidly and randomly in the (x, y)-plane but are otherwise unchanged. At each fixed distance, they also rotate randomly over time on a scale of minutes to hours due to temperature variations and mechanical vibrations. Consequently, a PMD compensator must be continually optimized to correct for PMD-induced distortions in the received optical signal.

The equation governing the z-evolution of $\mathbf{U}$ is the *coupled nonlinear Schrödinger equation* (CNLS). The CNLS is derived from Maxwell's equations by averaging over the rapid variations of the carrier wave, $\exp[i\beta(\omega_0, z)z - i\omega_0\tau]$, and over the eigenfunctions, $\mathbf{R}_1$ and $\mathbf{R}_2$ [43]. The CNLS states that

$$\frac{\partial \mathbf{U}}{\partial z} = g\mathbf{U} + i\Delta\mathsf{B}\mathbf{U} - \Delta\mathsf{B}'\frac{\partial \mathbf{U}}{\partial t} - \frac{i}{2}\beta''\frac{\partial^2 \mathbf{U}}{\partial t^2} + i\gamma\left[|\mathbf{U}|^2\mathbf{U} - \frac{1}{3}(\mathbf{U}^\dagger\sigma_2\mathbf{U})\sigma_2\mathbf{U}\right]. \tag{2.2}$$

Here, the scalar coefficient g is the loss coefficient of the fiber. The factor $\Delta\mathsf{B} = \Delta\mathsf{B}(\omega_0, z)$ is the *birefringence matrix*, which is a 2×2 Hermitian matrix that models the anisotropy and asymmetry of the linear dielectric response tensor of the fiber, averaged with respect to the eigenfunctions $\mathbf{R}_1$ and $\mathbf{R}_2$. The matrix $\Delta\mathsf{B}'$ is defined by $\Delta\mathsf{B}' = \frac{\partial \Delta\mathsf{B}}{\partial \omega}(\omega_0, z)$, and the scalar $\beta'' = \frac{\partial^2 \beta}{\partial \omega^2}(\omega_0, z)$ is the chromatic (frequency-dependent) dispersion. The scalar coefficient γ measures the strength of the Kerr nonlinearity in the fiber. The Kerr nonlinearity arises from the fact that the refractive index of optical fiber has a small dependence on the optical power $|\mathbf{U}|^2$ of the light. Finally, $\dagger$ denotes conjugate transpose, and σ_2 is the second of the three Pauli spin matrices

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \qquad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{2.3}$$

Because of the polarization properties of glass [48], optical fiber is *linearly birefringent*, which means that

$$\Delta\mathsf{B} \;=\; \Delta\beta(\cos\theta\sigma_3 + \sin\theta\sigma_1), \tag{2.4}$$

where $\Delta\beta$ is the magnitude of birefringence and θ is an orientation angle whose significance will be explained later. Since θ is a very weak function of frequency, we assume that $\Delta\mathsf{B}' = \Delta\beta'(\cos\theta\sigma_3 + \sin\theta\sigma_1)$. Wai and Menyuk [58] proposed a model for the random variation of the birefringence along the fiber in which $\Delta\beta'\cos\theta$ and $\Delta\beta'\sin\theta$ are independent Gaussian random processes with mean zero and the same

standard deviation [58]. Recently, Galtarossa et al. [19], [20] experimentally validated this model.

The random variations in the fiber birefringence occur on a length scale of 1–100 m and result in very rapid changes in the polarization state of the light as it propagates. However, since optical communication systems are at least several hundred kilometers long, it is not computationally feasible to simulate propagation through birefringent fiber by using a numerical method that takes small enough steps along the fiber to track these rapid changes in the polarization state of the light. This problem can be overcome by transforming the rapid changes in the polarization state of the light at the carrier frequency out of the CNLS to obtain the *Manakov-PMD equation* [42], [43]:

$$\frac{\partial \mathbf{W}}{\partial z} = g\mathbf{W} - \Delta\beta'\bar{\sigma}_3\frac{\partial \mathbf{W}}{\partial t} - \frac{i}{2}\beta''\frac{\partial^2 \mathbf{W}}{\partial t^2} + i\gamma|\mathbf{W}|^2\mathbf{W}. \tag{2.5}$$

Here $\mathbf{W}(z,t) = \mathsf{Q}(z)\mathbf{U}(z,t)$, where $\mathsf{Q}(z)$ is a unitary transformation and $\bar{\sigma}_3 = \mathsf{T}(z)^{-1}\sigma_3\mathsf{T}(z)$ for a matrix $\mathsf{T}(z)$ that is determined by the birefringence parameters $\Delta\beta$ and θ. To explain the rationale for making a transformation of the form $\mathsf{Q}(z)$, we first note that the Fourier conjugate of the retarded time, t, is frequency, ω, measured relative to the carrier frequency, ω_0, of the optical signal. Even though the Fourier transform, $\widehat{\mathbf{U}}(z,\omega)$, of $\mathbf{U}(z,t)$ varies rapidly with the propagation distance z, it only has a very weak dependence on frequency, ω. Therefore, it makes sense to transform $\mathbf{U}$ so that the new coordinates exactly follow the rapid changes of the polarization state of the signal at the center frequency, ω_0, i.e., so that $\widehat{\mathbf{W}}(z,0)$ is constant in z. In this new coordinate system, the Fourier transform, $\widehat{\mathbf{W}}(z,\omega)$, of the solution of the Manakov-PMD equation measures the slow variation of the polarization state of the light at each frequency, ω, with respect to the polarization state of the light at the carrier frequency, ω_0.

2.2. The linear PMD equation. In this subsection we review the linear PMD equation which is a special case of the Manakov-PMD equation. The linear PMD equation is appropriate for studying the statistical behavior of PMD and PMD compensators. We will use this equation to explain how PMD results in the spreading and distortion of optical pulses.

The linear PMD equation is obtained by omitting all but the second term on the right-hand side of (2.5). Another widely studied special case of the Manakov-PMD equation is the scalar nonlinear Schrödinger equation, which is the equation obtained when there is no birefringence ($\Delta\beta' = 0$) and the optical signal at the transmitter is polarized [25], [43]. In many systems, the primary source of bit errors is not the nonlinear effects that are modeled by the scalar nonlinear Schrödinger equation but rather the effects of PMD that are modeled by the linear PMD equation. Moreover, because PMD is a stochastic effect, and because bit errors are so rare, PMD must be studied statistically using a large number of randomly chosen realizations of the birefringence of the optical fiber. Simulations that do not include the Kerr nonlinearity are computationally several orders of magnitude faster than those that do.

We now explain how PMD gives rise to the spreading and distortion of optical pulses, and introduce the concept of differential group delay. The linear PMD equation is most readily analyzed in the frequency domain, where

$$\frac{\partial \widehat{\mathbf{W}}}{\partial z} \;=\; i\Delta\beta'\,\bar{\sigma}_3(z)\,\omega\,\widehat{\mathbf{W}}. \tag{2.6}$$

The solution to this equation can be expressed in the form $\widehat{\mathbf{W}}(z,\omega) = \widehat{f}(\omega)\,\widehat{\mathbf{A}}(z,\omega)$, where $\widehat{f}$ is a real scalar-valued function, and $|\widehat{\mathbf{A}}(z,\omega)|^2 = 1$. As a function of time, the power of the optical signal is given by $|f|^2$, where f is the inverse Fourier transform of $\widehat{f}$. The vector $\widehat{\mathbf{A}}$ is the polarization state of the signal. Suppose that $\mathsf{F} = \mathsf{F}(z,\omega)$ is a matrix such that

$$\frac{\partial \widehat{\mathbf{A}}}{\partial \omega} \;=\; i\mathsf{F}\widehat{\mathbf{A}}. \tag{2.7}$$

It can be shown that F is a Hermitian matrix that is determined by the quantities $\Delta\beta'$ and $\bar{\sigma}_3$ that characterize the birefringence of the optical fiber [43]. The absolute difference of the two real eigenvalues of F is called the *differential group delay* (DGD), and the eigenvectors of F are called the *principal states of polarization*. The DGD and the principal states of polarization depend on both the propagation distance, z, and the frequency, ω.

To see how the DGD is related to pulse spreading, suppose for simplicity that the matrix F is ω-independent, at least over the frequency bandwidth of the signal, $\widehat{\mathbf{W}}$. In this case, we say that the fiber birefringence generates only *first-order PMD*. Diagonalizing F, we see that

$$\widehat{\mathbf{A}}(z,\omega) \;=\; \mathsf{P}e^{i\omega\mathsf{D}(z)}\mathsf{P}^\dagger\widehat{\mathbf{A}}(z,0), \tag{2.8}$$

where $\mathsf{P} = (\mathbf{v}_1, \mathbf{v}_2)$ is unitary, and $\mathsf{D}(z) = \operatorname{diag}\left(-\frac{\tau(z)}{2}, \frac{\tau(z)}{2}\right)$. Here $\tau(z)$ is the DGD. (By ignoring a common phase, we can assume that $\operatorname{trace}(\mathsf{F}) = 0$.) Therefore, the optical signal is given by

$$\mathbf{W}(z,t) \;=\; c_1 f\left(t + \tfrac{\tau(z)}{2}\right)\mathbf{v}_1 \;+\; c_2 f\left(t - \tfrac{\tau(z)}{2}\right)\mathbf{v}_2, \tag{2.9}$$

where c_1 and c_2 are complex constants with $|c_1|^2 + |c_2|^2 = 1$. The DGD is therefore the time delay between light that is launched in the two different principal states of polarization. If the power of an optical pulse is split between the two principal states (i.e., $c_k \neq 0$ for $k = 1, 2$) and the DGD is large, then the power of the pulse is spread out and distorted as a function of time and bit errors are more likely to occur at the receiver.

2.3. The Stokes representation. In our discussion so far, we have modeled the polarization state of light using the Jones representation. In this subsection, we review an alternate approach based on the Stokes representation [5], [24], [43], [44]. Using the Stokes representation, we can regard the propagation of the polarization state of light through birefringent optical fiber as a random walk on a two-dimensional sphere.

With the Jones representation, polarization states are represented as unit vectors $\mathbf{U} \in \mathbf{C}^2$, i.e., as points on the sphere $S^3 \subset \mathbf{C}^2$, while with the Stokes representation they are represented using unit Stokes vectors, $\mathbf{S} = (S_1, S_2, S_3) \in S^2 \subset \mathbf{R}^3$. The sphere of unit Stokes vectors is called the *Poincaré sphere*. The mapping $\psi : S^3 \to S^2$ from Jones space to Stokes space is defined by $\mathbf{S} = \psi(\mathbf{U}) = \mathbf{U}^\dagger\overrightarrow{\sigma}\mathbf{U}$, where $\overrightarrow{\sigma} = (\sigma_3, \sigma_1, -\sigma_2)$ is the Pauli spin vector. For each $\mathbf{S}$, the inverse image $\psi^{-1}(\mathbf{S})$ is the circle in S^3 that consists of all Jones vectors of the form $\mathbf{U}_\varphi = e^{i\varphi}\mathbf{U}$, where $\varphi \in S^1$ and $\psi(\mathbf{U}) = \mathbf{S}$. The angle φ is an overall phase that plays no role in the theory of PMD. Since the matrix acting on the right-hand side of (2.6) is an element of

the Lie algebra $\mathfrak{su}(2)$ of trace-free skew-Hermitian matrices, the solution of the linear-PMD equation is of the form $\widehat{\mathbf{W}}(z,\omega) = \mathsf{T}(z,\omega)\widehat{\mathbf{W}}(0,\omega)$, where the PMD-transmission matrix, T, is an element of the Lie group SU(2) of unitary matrices with determinant 1. The action of an element T of SU(2) on $S^3 \subset \mathbf{C}^2$ is equivalent to the action of an element $\mathtt{R}$ of the special orthogonal group, SO(3), on S^3, where T is mapped to $\mathtt{R}$ by the 2-1 and onto map $\Psi : \mathrm{SU}(2) \to \mathrm{SO}(3)$ that is determined by $\mathtt{R}\overrightarrow{\sigma} = \mathsf{T}^\dagger \overrightarrow{\sigma}\mathsf{T}$. In particular, if $\mathtt{R}$ is a rotation by an angle θ about an axis $\hat{r} \in S^2 \subset \mathbf{R}^3$, then $\mathsf{T} = \pm[\cos(\theta/2)\mathrm{I} - i\sin(\theta/2)\hat{r}\cdot\overrightarrow{\sigma}]$. The Lie algebra $\mathfrak{so}(3)$ of $SO(3)$ consists of all antisymmetric matrices. The induced map between Lie algebras, $\Psi_* : \mathfrak{su}(2) \to \mathfrak{so}(3)$, is defined by $\Psi_*(i\overrightarrow{\beta}\cdot\overrightarrow{\sigma}) = \beta\times$. Here $\overrightarrow{\beta} = (\beta_1,\beta_2,\beta_3) \in \mathbf{R}^3$ and

$$\beta\times = \begin{pmatrix} 0 & -\beta_3 & \beta_2 \\ \beta_3 & 0 & -\beta_1 \\ -\beta_2 & -\beta_1 & 0 \end{pmatrix} \tag{2.10}$$

determines an isomorphism between $\mathfrak{so}(3)$ and $\mathbf{R}^3$. Therefore, if $\widehat{\mathbf{S}}$ is the Stokes vector that is equivalent to the Jones vector $\widehat{\mathbf{A}}$ defined below (2.6), then in Stokes space, the linear PMD (2.6) is given by

$$\frac{\partial\widehat{\mathbf{S}}}{\partial z} \;=\; \overrightarrow{\beta}\times\widehat{\mathbf{S}}, \tag{2.11}$$

where $\overrightarrow{\beta} = \overrightarrow{\beta}(\omega,z)$ is the *local birefringence vector* of the fiber. If the local birefringence vector is constant, then the polarization state $\widehat{\mathbf{S}}$ at a single frequency traces a circle on the Poincaré sphere centered at $\overrightarrow{\beta}$. However, in real linearly birefringent fibers, the local birefringence vector, $\overrightarrow{\beta}$, moves randomly on the equator of the sphere.[1] Therefore, the polarization state of the light, $\widehat{\mathbf{S}}$, moves randomly over the entire sphere. More specifically, given a length of birefringent fiber and an optical signal with a given input polarization state, consider the probability distribution of output polarization states, $\widehat{\mathbf{S}}$, on the Poincaré sphere, where the samples are generated from different realizations of the fiber birefringence. The length scale required for the probability distribution of $\widehat{\mathbf{S}}$ to become uniform on the sphere is on the order of a kilometer, which is short compared to the total length of a communication system. These observations provide the primary motivation for the coarse-step method for modeling PMD, which we will discuss in section 2.4 below.

To explain how the DGD evolves as a function of distance along the fiber, we introduce the *polarization dispersion vector*, $\overrightarrow{\Omega} = \overrightarrow{\Omega}(\omega,z) = (\Omega_1,\Omega_2,\Omega_3)$, which is defined in terms of the Hermitian matrix F in (2.7) by $\mathsf{F} = \mathrm{trace}(\mathsf{F})\mathrm{I} + \frac{1}{2}\overrightarrow{\Omega}\cdot\overrightarrow{\sigma}$. Then the Stokes formulation of (2.7) is

$$\frac{\partial\widehat{\mathbf{S}}}{\partial\omega} \;=\; \overrightarrow{\Omega}\times\widehat{\mathbf{S}}. \tag{2.12}$$

Consequently, the unit vectors $\pm\overrightarrow{\Omega}/|\overrightarrow{\Omega}|$ represent the two principal states of polarization of the fiber on the Poincaré sphere, and the magnitude, $|\overrightarrow{\Omega}|$, is the DGD. Finally, the *dynamical PMD equation*, which describes the evolution of the polarization dispersion vector, is given by [46]

$$\frac{\partial\overrightarrow{\Omega}}{\partial z} \;=\; \frac{\partial\overrightarrow{\beta}}{\partial\omega} \;+\; \overrightarrow{\beta}\times\overrightarrow{\Omega}. \tag{2.13}$$

[1] The angle θ in (2.4) is half the angle between $\overrightarrow{\beta}$ and the positive X-axis.

This equation can be derived by differentiating (2.11) with respect to ω and (2.12) with respect to z.

2.4. The probability space of fiber realizations. Each choice of a set of random local birefringence vectors along the fiber is called a *fiber realization*. In this subsection, we describe the coarse-step method that is used to generate different fiber realizations.

When a PMD compensator is used, each time the fiber realization changes, the compensator needs to be reset via optimization. In reality, the fiber realization can both drift gradually over time due to temperature fluctuations and change abruptly to an unrelated realization in response to large disturbances such as a truck passing overhead [6], [32], [52], [57]. We assume that when the fiber realization changes, there is no correlation between the old and new realizations. Statistically we model PMD by defining an appropriate space of fiber realizations, imposing a probability distribution on this space, and devising a method for randomly sampling the space of fiber realizations. To do so, it is commonly assumed that optical fiber is homogeneous. In other words, the statistical properties of the local birefringence vector $\overrightarrow{\beta}$ do not depend on distance, z, along the fiber. We will consider a space of fiber realizations consisting of fibers of a given length with a prescribed average DGD. In reality, such a space of fiber realizations could correspond either to fibers that are fabricated using the same manufacturing process, or to realizations of a single fiber whose birefringence is randomly varying over time.

The most commonly used method for generating fiber realizations with a given average DGD is the *coarse-step method* [13], [59] which we now describe using the Stokes representation [44]. The coarse-step method can be regarded as a computationally efficient, statistically correct, numerical method for solving the linear-PMD (or Manakov-PMD) equation. In the coarse-step method, the action of a birefringent fiber on the Stokes vector, $\widehat{\boldsymbol{S}}$, of the light is modeled as the concatenation of the action of N segments of fixed-birefringence fiber, each of which is preceded by a random rotation of $\widehat{\boldsymbol{S}}$ on the Poincaré sphere. If $\widehat{\boldsymbol{S}}^{(n)}(\omega)$ denotes the Stokes vector of light at frequency ω after the nth fiber segment, then

$$\widehat{\boldsymbol{S}}^{(n)}(\omega) \;=\; \mathtt{R}(\omega)\,\mathtt{Q}_n\widehat{\boldsymbol{S}}^{(n-1)}(\omega). \tag{2.14}$$

Here, the matrix $\mathtt{R}(\omega)$ is the element of SO(3) that is the rotation about the X-axis through an angle $\Delta\beta'\omega\Delta z$. This rotation models the propagation of light through each of the fixed-birefringence fiber segments. The quantity $\Delta\beta'$ is the magnitude of the frequency derivative of the local birefringence vector of each of the N fiber segments, and is related to the average DGD by $\Delta\beta' = (3\pi N/8)^{1/2}\,\overline{\mathrm{DGD}}/L$, where L is the length of the fiber [42]. The matrix $\mathtt{Q}_n$ is a frequency-independent, random rotation that is chosen using the canonical uniform probability distribution on SO(3). Specifically, $\mathtt{Q}_n$ can be expressed as an Euler-angle rotation matrix [23] of the form $\mathtt{Q}_n = \mathtt{R}_X(\psi_n)\,\mathtt{R}_Y(\theta_n)\,\mathtt{R}_X(\phi_n)$. Here, $\mathtt{R}_X(\psi)$ is a rotation about the X-axis through an angle ψ. The angles ψ_n and ϕ_n are uniformly distributed in $[0, 2\pi]$, and $\cos\theta_n$ is uniformly distributed in $[-1, 1]$. In the special case that $N = 1$, the coarse-step method generates only first-order PMD. For a large enough number of fiber segments, the coarse-step method produces the same statistical properties as are obtained using Wai and Menyuk's model for the randomly varying birefringence [42]. Moreover, the coarse-step method is much more computationally efficient, since it takes steps on the order of a kilometer rather than the meter-long steps required to track the rapid

variations in the birefringence.

The dynamical PMD equation (2.13) can also be solved using a coarse-step approach. If $\overrightarrow{\Omega}^{(n)} = \overrightarrow{\Omega}^{(n)}(z,\omega)$ is the polarization dispersion vector after the nth fiber segment, then

$$\overrightarrow{\Omega}^{(n)} \;=\; \mathtt{R}(\omega)\left[\mathtt{Q}_n\,\overrightarrow{\Omega}^{(n-1)} \;+\; \Delta\overrightarrow{\Omega}_n\right], \tag{2.15}$$

where $\Delta\overrightarrow{\Omega}_n = \Delta\beta'\Delta z\,\overrightarrow{e}_X$ is the polarization dispersion vector of the nth segment. Here $\overrightarrow{e}_X = (1,0,0)^T$, and we regard $\overrightarrow{\Omega}$ as a column vector. In section 3 below, we will use (2.15) to explain the basic idea behind PMD compensation.

There is a large literature on the statistical properties of PMD. The most important result is that in the limit as $N \to \infty$, the DGD is Maxwellian distributed with distribution $f_{\mathrm{DGD}}(x) = \frac{2\pi x^2}{\alpha^3}\exp(-x^2/2\alpha^2)$, where $\alpha = (\pi/8)^{1/2}\,\overline{\mathrm{DGD}}$ [46]. In particular, there is an extremely small probability that the DGD of a fiber realization is significantly larger than the average DGD.

2.5. The receiver and performance evaluation metrics. To evaluate the degree to which a PMD compensator reduces the probability of errors due to PMD, we also need a model of the receiver. The purpose of the receiver is to convert the transmitted optical signal into an electrical current, to determine a *clock time* that is used to set the beginning and ending points of the time intervals for each of the bits being transmitted [55], and finally to make a decision as to whether the voltage of the received electrical current in each of these bit slots is to be received as a one or a zero. This decision is based on a choice of *decision voltage*. If the received voltage is larger than the decision voltage, the bit is declared to be a one. Otherwise a zero is received.

A receiver model should include an algorithm to evaluate the performance of the communication system. There are several ways to measure performance—the most fundamental means is via the *bit-error ratio*, which is given by $\mathrm{BER} = \frac{1}{2}(p_{1|0} + p_{0|1})$. Here $p_{1|0}$ is the probability of receiving a one given that a zero was transmitted, and $p_{0|1}$ is the probability of receiving a zero given that a one was transmitted. In a real system bit errors occur due to the combined effect of PMD and noise from optical amplifiers. Since we did not include noise in our simulations, we measured the performance using a quantity called the *eye opening* [62] rather than the BER. The eye opening of a noise-free signal at the receiver is defined to be the difference between the smallest electrical voltage of a one and the largest electrical voltage of a zero at the clock time.[2] There is a strong correlation between the BER and the eye opening [41], [49], [63]. To study the degree to which PMD degrades the performance of the system, we define the *eye opening penalty* for a particular fiber realization to be the ratio between the eye opening for a version of the system that has just a transmitter and receiver (i.e., no fiber and hence no PMD) and the eye opening for the system with a transmitter, PMD due to the given fiber realization, and a receiver. The more the PMD reduces the eye opening, the larger the eye opening penalty for that fiber realization. System designers typically specify that a *system outage* occurs if the eye opening penalty exceeds a specified threshold, such as 2 dB [52].[3] They require that

[2]The term "eye opening" is used here because in Figure 3.2 the image of each signal under the mapping $t \mapsto t \bmod T$ looks like an eye, where T is the bit period.

[3]A linear factor of x corresponds to $10\log_{10} x$ decibels (dB). So, if the eye opening penalty for a particular fiber realization is 2 dB, then the eye opening is 63% of the value one would see in a system without PMD.

the *outage probability*, which is the probability that such an outage occurs, be on the order of 10^{-3} to 10^{-6}, corresponding to a few minutes to hours of outage per year. A major goal of this paper is to present a numerical model that can be used to assess the degree to which a PMD compensator can reduce the outage probability due to PMD.

3. Polarization-mode dispersion compensation. In this section, we discuss the optical PMD compensator whose performance we study in this paper. As discussed in section 2, imperfections in optical fibers result in two principal states of polarization for the light. Light traveling in the fast principal state arrives at the receiver ahead of the light traveling in the slow state. In practice the power in the optical signal is split between these two principal states, so that the optical pulses used to encode the binary data become spread out and distorted. Consequently, over long distances, the message being transmitted will be corrupted by errors. To compensate for this spread of information, physical devices called optical PMD compensators can be used.

A variety of designs have been proposed for optical PMD compensators [52], [53]. We chose to study a simple compensator that is easy to build and operate. To motivate the design of this compensator, we consider the important special case that the transmission fiber has only first-order PMD, i.e., that the polarization dispersion vector, $\overrightarrow{\Omega}_T$, of the transmission line is frequency independent. Recall from (2.9) that if the polarization state of the signal is not closely aligned with either of the principal states of the transmission fiber, then the larger the DGD, $|\overrightarrow{\Omega}_T|$, the more the signal will be distorted due to PMD, and the greater the probability of a bit error. The simple PMD compensator we study is a device that can at least partially cancel out the DGD of the transmission fiber. The idea is to insert a segment of *compensation fiber* between the transmission fiber and the receiver, and to rotate the polarization state of the signal between the transmission and compensation fibers so as to align the fast principal state $\overrightarrow{\Omega}_T$ of the transmission fiber with the slow principal state $-\overrightarrow{\Omega}_C$ of the compensation fiber. Then by (2.15), the total polarization dispersion vector, $\overrightarrow{\Omega}_R$, at the receiver is given by

$$\overrightarrow{\Omega}_R \;=\; \mathtt{Q}\,\overrightarrow{\Omega}_T \;+\; \overrightarrow{\Omega}_C, \tag{3.1}$$

where $\mathtt{Q}$ is the rotation between the transmission and compensation fibers. Consequently, if the DGD of the compensation fiber were equal to that of the transmission fiber, then the total DGD at the receiver would be zero; i.e., $|\overrightarrow{\Omega}_R| = 0$.

Simple compensators based on this principle have been built, and a diagram of one is shown in Figure 3.1. The optical signal generated in the transmitter propagates through the transmission fiber. The compensator itself is located immediately prior to the receiver and consists of a polarization controller that can be adjusted to transform the polarization state of the fiber by any desired rotation, followed by a short segment of compensation fiber. The compensation fiber is designed to be polarization maintaining in that its principal states of polarization and its DGD are fixed. Notice that since the transmission DGD, $|\overrightarrow{\Omega}_T|$, of a fiber realization may not equal the fixed DGD, $|\overrightarrow{\Omega}_C|$, of the compensator, the total DGD, $|\overrightarrow{\Omega}_R|$, in (3.1) may not be zero. In addition, since this compensator consists of only one polarization controller and one segment of polarization maintaining fiber, it can compensate only for the DGD at the carrier frequency, and not for PMD distortions at all frequencies in the signal.

After passing through the compensation fiber, the optical signal is monitored and a feedback loop is used to adjust the rotation performed by the polarization controller. Given a realization of the birefringence in the transmission fiber, the feedback loop

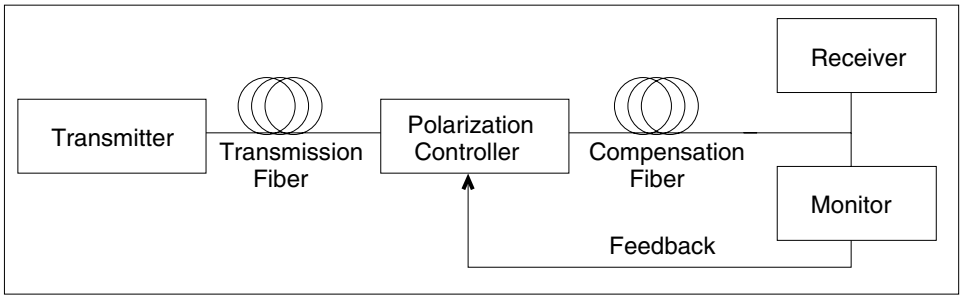


FIG. 3.1. *An optical communication system with a simple PMD compensator.*

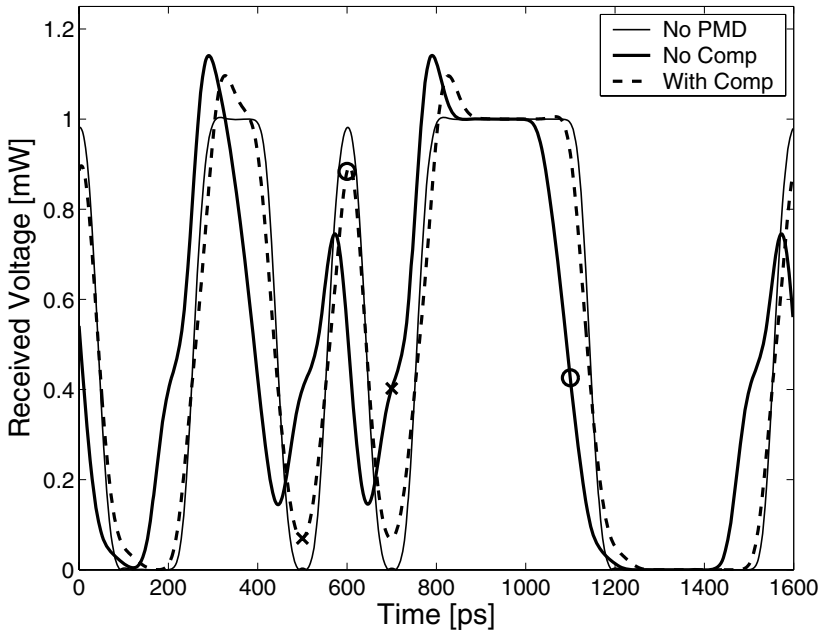


FIG. 3.2. *Compensated versus uncompensated signals.*

can be modeled by optimizing the function from the state space SO(3) of all possible rotations of the polarization controller to **R**, which is given by the monitor.

In Figure 3.2, we show the effect that PMD can have on a signal and how a PMD compensator can decrease the signal distortion due to PMD. The results we show are for a particular fiber realization. Different fiber realizations could have quite different effects on the signal. We plot the received electrical current as a function of time in three cases. The thin solid curve shows the case where there is no PMD in the transmission line and hence no pulse distortion. (This signal is called the *back-to-back* signal as the transmitter and receiver abut each other.) The data pattern 1001101011110000 can be easily recognized in the signal. Notice that the voltage in the signal does not return to zero between two consecutive ones. Signals like this are called *non–return-to-zero* and are commonly used in communication systems. The thick solid curve shows the same signal after it has traveled through the transmission fiber with PMD. The DGD was 75 ps at the carrier frequency, and there was a strong frequency-dependence to the polarization dispersion vector. The signal has clearly

been distorted due to PMD. The circle at 1100 ps shows the minimum-voltage one, and the cross at 700 ps shows the maximum-voltage zero. The eye opening, which is the difference between the height of the circle and the cross, is 0.02 mV. Therefore the eye opening after the transmission fiber is very small compared to the back-to-back eye opening, which is 0.98 mV. Finally, the thick dashed curve shows this same signal after compensation. The circle at 600 ps shows the minimum-voltage one, and the cross at 500 ps shows the maximum-voltage zero. The compensator has increased the eye opening to 0.81 mV, which is a marked improvement over the uncompensated case.

In this paper, we study the performance of this PMD compensator for three choices of monitor. Different monitors and different random realizations of the birefringence in the transmission fiber correspond to different choices of the objective function to be optimized. Since our ultimate goal is to minimize the bit-error ratio, the monitor should be chosen so that the monitored value is strongly correlated (or anticorrelated) to the bit-error ratio, i.e., to the eye opening. The most obvious choice of monitor is the bit-error ratio itself, or the eye opening. However, it is not possible to measure the bit-error ratio in a real system, and it is often not feasible to measure the eye opening in real time.

We will now briefly describe each of the monitors used in our numerical experiments—the eye opening, spectral line, and the DOP ellipsoid. The *eye opening monitor* measures the eye opening of the signal after propagation through the system, relative to the eye opening of the back-to-back signal. Optical signals typically have carrier frequencies of about 200 THz (or about 2×10^{12} Hz) and a bandwidth of about 20 GHz (or 2×10^{10} Hz). The optical signal is sent through a photodetector to convert it from an optical to an electrical signal. The electric signal has a shifted frequency spectrum (shifted relative to the optical signal) in the range $[0, 10]$ GHz. The *spectral line* monitoring technique requires that an electric filter be used to monitor the power in a particular frequency (or tone). In our work we have used a filter to monitor the power in the 5 GHz tone using a window of width 0.5 GHz. The spectral line feedback mechanism attempts to maximize the power in this tone relative to the reference back-to-back signal, which has gone straight from transmitter to receiver without encountering the optical fiber (hence without being affected by PMD).

In the case of only first-order PMD, we will show in (4.17) that the power in the 5 GHz spectral line decreases monotonically with increasing DGD between DGD values of 0 and 100 ps (the width of each bit slot; see also Figure 1 in [51]). Thus it would be hoped that for moderate amounts of DGD, an optimization algorithm which maximizes the power in this spectral line (or frequency) will compensate for the DGD present in the fiber. As has been discussed earlier, the eye opening is correlated to the DGD, and we see now that for first-order PMD, the amount of DGD is correlated to the power in the spectral line. Thus this feedback measure achieves our aim of providing a real-valued function correlated to the eye opening. In the case of higher-order PMD, this correlation becomes more complicated.

We now explain why the functions to be optimized can be regarded as being defined on the 2-sphere, S^2, rather than on the three-dimensional manifold, SO(3). As in our discussion of the coarse-step method in section 2.4, the rotation $\mathtt{Q}$ in (3.1) can be expressed in the form $\mathtt{Q} = \mathtt{R}_X(\psi)\,\mathtt{R}_Y(\theta)\,\mathtt{R}_X(\phi)$, where $\psi, \theta, \phi \in [0, 2\pi]$, and $\mathtt{R}_X(\phi)$ is a rotation by an angle ϕ about the X-axis. We assume that the polarization dispersion vector, $\overrightarrow{\Omega}_C$, of the compensator is parallel to the X-axis. Consequently, $\overrightarrow{\Omega}_R = \mathtt{R}_X(\psi)\left[\mathtt{R}_Y(\theta)\,\mathtt{R}_X(\phi)\,\overrightarrow{\Omega}_T + \overrightarrow{\Omega}_C\right]$. Since the final rotation, $\mathtt{R}_X(\psi)$, does not affect the values of any of the objective functions, we can ignore it and regard $\mathtt{Q} = \mathtt{Q}(\phi, \theta)$

as a function of (ϕ, θ) alone; i.e.,

$$\overrightarrow{\Omega}_R(\phi,\theta) \;=\; \mathtt{R}_Y(\theta)\,\mathtt{R}_X(\phi)\,\overrightarrow{\Omega}_T \;+\; \overrightarrow{\Omega}_C. \tag{3.2}$$

Since $\mathtt{R}_X(\pi) = \mathrm{diag}(1,-1,-1)$, $\mathtt{R}_Y(-\theta)\mathtt{R}_X(\pi) = \mathtt{R}_X(\pi)\mathtt{R}_Y(\theta)$. Therefore, the rotation $\mathtt{Q}(\phi+\pi,-\theta)$ has the same effect as the rotation $Q(\phi,\theta)$. Consequently, we can regard the objective functions as being defined on the rectangle $R = \{(\phi,\theta) \in [-\pi,\pi]\times[0,\pi]\}$. Moreover, ignoring a final rotation about the X-axis, $\mathtt{Q}(\phi,0) = I$ and $\mathtt{Q}(\phi,\pi) = \mathtt{R}_Y(\pi)$ are constants, independent of ϕ. Therefore, the space of rotations performed by the polarization controller is actually diffeomorphic to the sphere, S^2: The rotation $Q(\phi,\theta)$ corresponds to the point on the sphere with spherical coordinates (ϕ,θ), where $\phi = \phi_0$ is a circle of longitude and $\theta = \theta_0$ is a circle of latitude. Consequently, the objective functions are actually defined on S^2.

Next, we examine the structure of the objective function for the eye opening and spectral line monitors. In Figure 3.3, we show a contour plot of the eye opening objective function for a particular fiber realization. In this figure, we have parametrized the sphere in the space of rotations using spherical coordinates (ϕ,θ) so that the horizontal lines $\theta = -\pi$ and $\theta = \pi$ map to the south and north poles, respectively, and the vertical lines $\phi = \pm\pi$ both map to the same great semicircle of longitude. The eye opening objective function in Figure 3.3 has two local maxima located close to $(\phi,\theta) = (-\pi,\pi/2)$ and $(\phi,\theta) = (0,\pi/2)$. The spectral line objective function in Figure 3.4 has a similar structure, although it is smoother than the eye opening objective function. In particular we note that the steep ridge located between $\phi = 0$ and $\phi = -\pi/2$ for the eye objective function has been considerably smoothed out in the spectral line objective function diagram. Thus local optimization techniques would be more effective at finding optima for the spectral line than for the eye in this case.

To motivate the degree of polarization ellipsoid monitor, we observe that when a signal propagates through fiber with PMD there are two basic reasons why the eye opening at the receiver can be large: Either the total DGD is small or the input state of polarization of the signal is aligned with one of the principal states of the fiber [47]. In Figure 3.3, the eye opening is large near $(-\pi,\pi/2)$ because the total DGD is small there, and it is large near $(0,\pi/2)$ because the rotation of the polarization controller is such that the principal state of the entire length of fiber, $\overrightarrow{\Omega}_R$, is parallel to the input state of polarization of the signal. In a real system, the input polarization state can drift over time. Consequently, it is better to operate a PMD compensator near where the DGD is minimized rather than near where the input polarization state of the signal is aligned with a principal state of the fiber [52]. The DOP ellipsoid is one such monitor.

The *DOP ellipsoid* monitor is obtained by polarization scrambling the DOP of the signal. The DOP is used to monitor PMD because polarized signals become depolarized when they propagate through fiber with PMD. The degree to which the signal is depolarized depends on the DGD and on the input state of polarization. Any optical signal can be decomposed as the sum of polarized and unpolarized components [5]. The DOP is the ratio of the power in the polarized component to the total power. If $\mathbf{U}(t) = f(t)\mathbf{U}_0$ denotes the Jones vector of an input polarized signal, then the total power is $S_0 = \int_{-\infty}^{\infty} |\widehat{f}(\omega)|^2\,d\omega$, where $\widehat{f}$ is the Fourier transform of f. The power in the polarized component of the signal is given by the length of the average Stokes vector,

$$\mathbf{S} \;=\; \int_{-\infty}^{\infty} \widehat{\mathbf{S}}(\omega)\,|\widehat{f}(\omega)|^2\,d\omega, \tag{3.3}$$

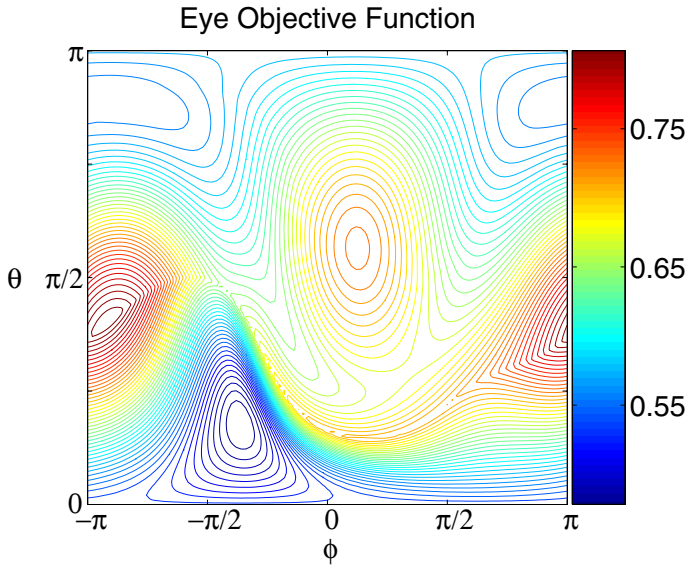


FIG. 3.3. *Eye opening objective function for a typical fiber realization.*

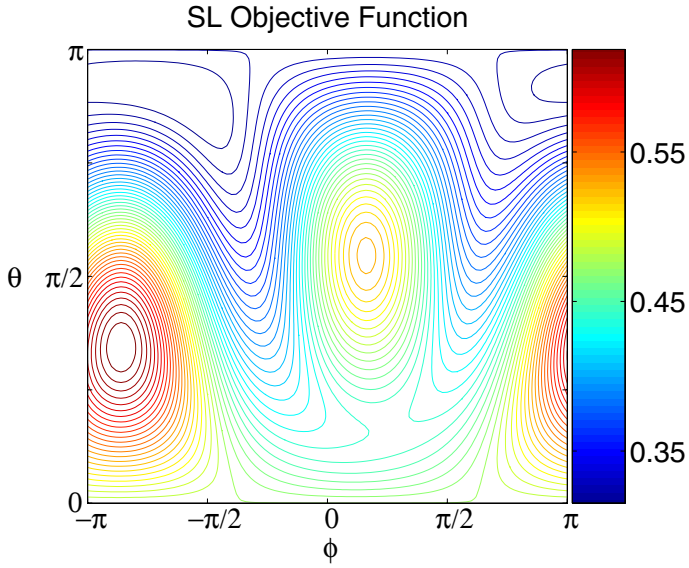


FIG. 3.4. *Spectral line objective function for a typical fiber realization. (Same fiber as is shown in Figure* 3.3*.)*

where $\widehat{\mathbf{S}}(\omega)$ is the Stokes vector at frequency ω defined in section 2.3. The DOP is given by $\text{DOP} = |\mathbf{S}|/S_0$.

To explain why a large DGD results in a low DOP, consider a fiber with only first-order PMD for which the polarization dispersion vector is fixed at the north pole. Then, as a function of frequency, ω, the Stokes vector $\widehat{\mathbf{S}}(\omega)$ traces out an arc of a circle of latitude on the sphere. By (2.12), the larger the DGD, $|\overrightarrow{\Omega}|$, the longer the arc. Suppose, for example, that the input polarization state is such that the circle of

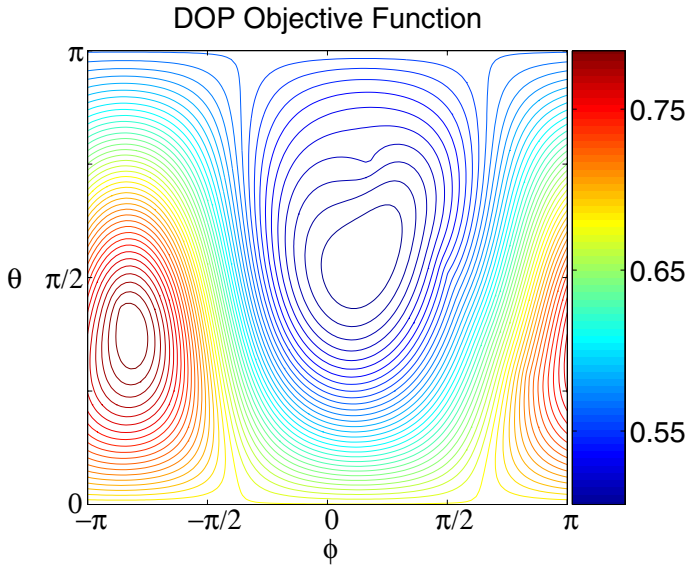


FIG. 3.5. *DOP ellipsoid objective function for a typical fiber realization. (Same fiber as is shown in Figure* 3.3.*)*

latitude is the equator, and the DGD is large enough so that $\widehat{\mathbf{S}}(\omega)$ traces out the entire equatorial circle as ω varies over the bandwidth of the signal. Then, by symmetry, the integral in (3.3) is close to zero and so the DOP is small. More generally, since (3.3) is a weighted average of vectors, the larger the DGD, the smaller the DOP.

The DOP ellipsoid is defined [9], [14], [15], [50], [54] so that for each unit vector $\mathbf{S}_{\text{in}}$, polarized light with Stokes vector $\mathbf{S}_{\text{in}}$ is sent in to a fiber with PMD, and the average output Stokes vector, $\mathbf{S}_{\text{out}}$, is measured. The set of all such vectors $\mathbf{S}_{\text{out}}$ forms the DOP ellipsoid. The length of the shortest principal axis of the DOP ellipsoid is approximately given by [14]

$$\lambda_{\min} \approx 1 - \tfrac{1}{2}\,\overline{\Delta\omega^2}\,|\,\overrightarrow{\Omega}\,|^2. \tag{3.4}$$

Here $\overline{\Delta\omega^2} = (1/S_0)\int_{-\infty}^{\infty} \omega^2 |\widehat{f}(\omega)|^2\, d\omega$ is a measure of the square of the bandwidth of the signal, and $|\,\overrightarrow{\Omega}\,|$ is the DGD at the carrier frequency. Therefore, with the DOP ellipsoid feedback mechanism, we aim to maximize $\lambda_{\min}$ and therefore to minimize the DGD. To fit the ellipsoid we used a Euclidean invariant linear least-squares algorithm that minimized the algebraic distance to 36 output Stokes vectors $\mathbf{S}_{\text{out}}$ [21]. In Figure 3.5, we show a plot of the DOP ellipsoid objective function for the same fiber realization as in Figures 3.3 and 3.4. This objective function has a single maximum that is located near the maximum of the eye opening objective function (and which corresponds to minimizing the DGD). However, unlike the case of the eye opening and spectral line objective functions, it does not have a second local maximum corresponding to the case that the input state of polarization is aligned with one of the principal states of the fiber. The shortest axis of the ellipsoid corresponds to the worst possible choice of input polarization state. The worst state is the one for which the power in the signal is evenly split between the two principal states of polarization of the fiber, rather than being aligned with one of them. Therefore the DOP ellipsoid objective function depends on the DGD but not on the input polarization state of the signal.

4. Analysis of the objective functions for first-order PMD. In section 3 we saw that for a particular fiber realization the eye opening and spectral line objective functions had two local maxima, whereas the DOP ellipsoid objective function had only one local maximum which corresponded to minimizing the DGD. In this section we show that this behavior is typical by deriving formulae for the DOP ellipsoid and spectral line objective functions in the special case of first-order PMD, i.e., in the case that the PMD vector $\overrightarrow{\Omega}_T$ of the transmission fiber is frequency independent. In section 7 we will use numerical simulation to quantify the performance of a PMD compensator, and we will use the analysis in this section to help explain those results. In section 7 we also provide statistical evidence that for an arbitrary fiber realization, we can regard the objective function as being a perturbation of an objective function for a fiber with only first-order PMD. This statistical result is to be expected since for many fiber realizations $\overrightarrow{\Omega}_T$ has only a weak dependence on frequency across the bandwidth of the signal.

As we explained in section 3, the domain for the two-dimensional optimization problem we wish to solve is the unit sphere, $S^2 \subset \mathbf{R}^3$. In other words, we want to solve the problem

$$\max_{p \in S^2} J(p), \tag{4.1}$$

where $J : S^2 \to \mathbf{R}$ is an objective function defined by one of the three feedback mechanisms discussed earlier. In section 7, we will in fact solve an unconstrained optimization problem of the form

$$\max_{(\phi,\theta) \in \mathbf{R}^2} J(\phi, \theta), \tag{4.2}$$

where (ϕ, θ) are spherical coordinates and J is now regarded as a doubly periodic function on $\mathbf{R}^2$. This unconstrained form of the problem is easier to solve computationally.

For our analysis of the DOP ellipsoid, we assume that the polarization dispersion vector $\overrightarrow{\Omega}_R$ of the entire system from transmitter to receiver, including the compensation fiber, is frequency independent. Let τ_T and τ_C be the DGD of the transmission and compensation fibers, respectively. We can assume that $\overrightarrow{\Omega}_C = \tau_C \overrightarrow{e}_X$. If we let Ψ be the angle between $\overrightarrow{\Omega}_T$ and $\overrightarrow{\Omega}_C$, then in spherical coordinates, $\overrightarrow{\Omega}_T = \tau_T(\cos\Psi, \cos\beta \sin\Psi, \sin\beta \sin\Psi)^T$ for some angle β. Substituting (3.2) into (3.4), we find that the DOP ellipsoid objective function is given by

$$J(\phi, \theta) = 1 - \tfrac{1}{2} \overline{\Delta\omega^2} \left[\tau_C^2 + \tau_T^2 + 2\tau_C \tau_T H(\phi, \theta)\right], \tag{4.3}$$

where

$$H(\phi, \theta) = \cos\Psi \cos\theta + \sin\Psi \sin\theta \sin(\phi - \beta). \tag{4.4}$$

If $\Psi \neq 0, \pi$, then the optimization problem (4.2) has six critical points (ϕ, θ) in $[-\pi, \pi] \times [0, \pi]$. The global maximum of H has value 1 and is located at $(\phi, \theta) = (\beta + \pi/2, \Psi)$, and the global minimum is -1 at $(\phi, \theta) = (\beta - \pi/2,\ \pi - \Psi)$. There are saddle points at $(\beta, 0)$, $(\beta + \pi, 0)$, (β, π), and $(\beta + \pi, \pi)$. These saddle points lie on the singularity set $\{\theta = 0\} \cup \{\theta = \pi\}$ of the spherical coordinate mapping from $[-\pi, \pi] \times [0, \pi]$ to S^2, and are mapped to $(0, 0, \pm 1) \in S^2$. However, they do not correspond to critical points for the optimization problem (4.1) on S^2, since $(0, 0, \pm 1)$

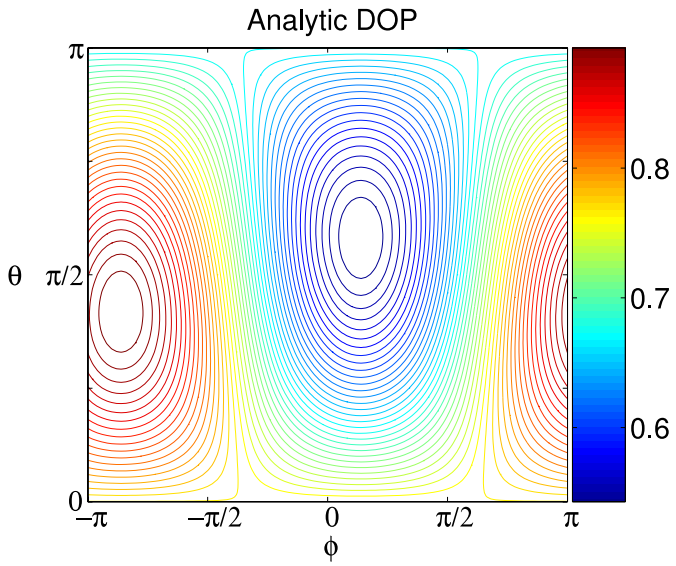


FIG. 4.1. *DOP ellipsoid objective function obtained using* (4.3) *for the first-order PMD approximation of the fiber in Figure* 3.5.

are not critical points of

$$\max_{(x,y,z)\in\mathbf{R}^3} H(x,y,z) = \cos(\Psi)z + \sin(\Psi)y \qquad \text{subject to} \quad x^2+y^2+z^2 = 1. \tag{4.5}$$

If $\Psi = 0$ or π, then $H(\phi,\theta) = \pm\cos\theta$ has global optima at $\theta = 0, \pi$, and these critical points are also global optima for the problem (4.1) on the sphere. To summarize, in the case of only first-order PMD, the DOP ellipsoid objective function on S^2 has one maximum and one minimum which is antipodal to the maximum.

Our analysis agrees well with the numerically computed objective function in Figure 3.5. Given a fiber realization, we can obtain an associated fiber realization with only first-order PMD by computing the polarization dispersion vector, $\overrightarrow{\Omega}_T$, at the carrier frequency of the signal. For the fiber realization in Figure 3.5, the parameters in the formula for $\overrightarrow{\Omega}_T$ are $\tau_C = 30$ ps, $\tau_T = 2.68\,\tau_C$, $\Psi = 105°$, and $\beta = -65°$. We chose $\sqrt{\Delta\omega^2} = 8.75 \times 10^9$ Hz, so as to fit the result in Figure 3.5, as was done in [14]. In Figure 4.1, we show the DOP ellipsoid objective function given by (4.3) with these parameters. The close agreement between the analytical and numerical results is noteworthy since the PMD of the fiber realization used for Figure 3.5 depends strongly on frequency. (In fact, the second-order PMD, $|\overrightarrow{\Omega}_\omega|$, which is the length of the frequency derivative of the polarization dispersion vector, is 3.1 times its mean value.)

For the analysis of the spectral line objective function, we work in Jones space with matrices in SU(2) acting on $\mathbf{C}^2$, rather than in Stokes space. Suppose that the input optical signal at the transmitter is of the form $f(t)\mathbf{u}_0$, where f is a real-valued scalar function, and $\mathbf{u}_0 \in \mathbf{C}^2$ is constant. Let R be the element of SU(2) that corresponds to the product $\prod_n \mathsf{Q}_n$ of the random rotations used in the coarse-step method to model a realization of the transmission fiber. If we approximate the transmission fiber by a fiber with only first-order PMD, then by (2.8) the output optical signal v after the PMD compensator is the $\mathbf{C}^2$-valued function whose Fourier transform is given by

$$\widehat{v}(\omega;\phi,\theta) = \widehat{f}(\omega)\mathsf{B}_C(\omega)\mathsf{Q}(\phi,\theta)\mathsf{P}e^{i\omega\mathsf{D}}\mathsf{P}^\dagger\mathsf{R}\mathbf{u}_0, \tag{4.6}$$

where $\mathsf{P} = (\mathbf{v}_1, \mathbf{v}_2) \in \mathrm{SU}(2)$ is the matrix of principal states of the transmission fiber and $\mathsf{D} = \mathrm{diag}(-\tau_T/2, \tau_T/2)$. In addition,

$$\mathsf{Q}(\phi,\theta) = \begin{pmatrix} e^{-i\phi/2}\cos\theta/2 & -ie^{i\phi/2}\sin\theta/2 \\ -ie^{-i\phi/2}\sin\theta/2 & e^{i\phi/2}\cos\theta/2 \end{pmatrix} \tag{4.7}$$

is the element of SU(2) that models the action of the polarization controller, and

$$\mathsf{B}_C(\omega) = \begin{pmatrix} e^{-i\omega\tau_C/2} & 0 \\ 0 & e^{i\omega\tau_C/2} \end{pmatrix} \tag{4.8}$$

models the DGD in the compensation fiber. Let $\sigma = \frac{\tau_T+\tau_C}{2}$ and $\eta = \frac{\tau_T-\tau_C}{2}$. Then

$$v(t;\phi,\theta) \;=\; \begin{pmatrix} \mathsf{A}_{11}f(t+\sigma) + \mathsf{A}_{12}f(t-\eta) \\ \mathsf{A}_{21}f(t+\eta) + \mathsf{A}_{22}f(t-\sigma) \end{pmatrix}, \tag{4.9}$$

where $\mathsf{A} = \mathsf{Q}(\phi,\theta)(c_1\mathbf{v}_1, c_2\mathbf{v}_2)$. Here $c_k = \mathbf{v}_k^*\mathsf{R}\mathbf{u}_0 \in \mathbf{C}$ is the projection onto the principal state $\mathbf{v}_k$ of the Jones vector of the signal at the central frequency after the transmission fiber.

If we ignore the optical filter and model the electrical filter as a Dirac function centered at frequency ω_{SL}, then the spectral line objective function is given by

$$J(\phi,\theta) = \left|\widehat{P}(\omega_{\mathrm{SL}};\phi,\theta)\right|^2, \qquad \text{where} \quad P(t;\phi,\theta) = |v(t;\phi,\omega)|^2 \tag{4.10}$$

is the received optical power.

To calculate an explicit formula for J we first express quadratic functions of A in terms of the generalized Stokes vectors

$$\mathbf{S}_{jk} \;=\; \mathbf{v}_j^\dagger\overrightarrow{\sigma}\mathbf{v}_k, \qquad j,k \in \{1,2\}, \tag{4.11}$$

where $\overrightarrow{\sigma}$ is the Pauli spin 3-vector defined in section 2.3 whose components are elements of SU(2). Notice that $\mathbf{S}_{11}$ and $\mathbf{S}_{22} = -\mathbf{S}_{11} \in \mathbf{R}^3$ are the standard Stokes vectors of $\mathbf{v}_1$ and $\mathbf{v}_2$, and that $\overline{\mathbf{S}}_{21} = -\mathbf{S}_{12} \in \mathbf{C}^3$. If Q_k denotes the kth row of $\mathsf{Q}(\phi,\theta)$, then $\mathsf{Q}_k^\dagger\mathsf{Q}_k = \frac{1}{2}(\mathrm{I} + (-1)^{k-1}\mathbf{r}(\phi,\theta)\cdot\overrightarrow{\sigma})$ in SU(2), where

$$\mathbf{r}(\phi,\theta) \;=\; (\cos\theta, \sin\phi\sin\theta, -\cos\phi\sin\theta) \tag{4.12}$$

parametrizes a sphere. Then,

$$|\mathsf{A}_{22}|^2 \;=\; \tfrac{1}{2}|c_{22}|^2(1-\mathbf{r}\cdot\mathbf{S}_{22}) \qquad \text{and} \qquad \Re(\mathsf{A}_{22}\overline{\mathsf{A}}_{21}) \;=\; -\tfrac{1}{2}\mathbf{r}\cdot\Re(c_1\bar{c}_2\mathbf{S}_{21}), \tag{4.13}$$

where $\Re$ denotes the real part, with similar formulae for other quadratic functions of A. Finally, let

$$g_{a,b}(t) := f(t+a)f(t+b) \qquad \text{and} \qquad G_{a,b} := \widehat{g}_{a,b}(\omega_{\mathrm{SL}}) \in \mathbf{C}. \tag{4.14}$$

Combining (4.6)–(4.14), the objective function for the optimization problem (4.2) on the plane is of the form

$$J(\phi,\theta) \;=\; \tfrac{1}{4}|\alpha + \mathbf{r}(\phi,\theta)\cdot\mathbf{s}|^2, \tag{4.15}$$

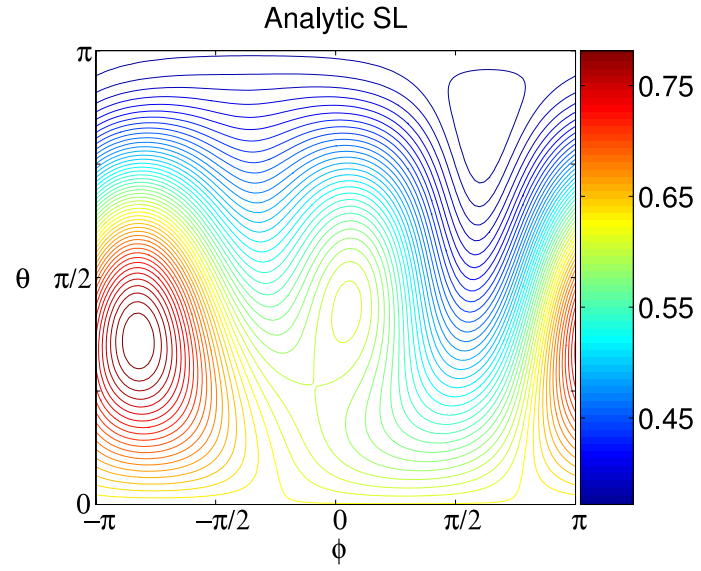


FIG. 4.2. *Spectral line objective function obtained using* (4.15) *for the first-order PMD approximation of the fiber in Figure* 3.4.

where $\alpha \in \mathbf{C}$ and $\mathbf{s} \in \mathbf{C}^3$ are defined by

$$\begin{aligned} \alpha &= |c_1|^2(G_{\sigma,\sigma} + G_{\eta,\eta}) \;+\; |c_2|^2(G_{-\sigma,-\sigma} + G_{-\eta,-\eta}), \\ \mathbf{s} &= \left\{|c_1|^2(G_{\sigma,\sigma} - G_{\eta,\eta}) + |c_2|^2(G_{-\sigma,-\sigma} - G_{-\eta,-\eta})\right\}\mathbf{S}_{11} \\ &\quad + \; 2(G_{\sigma,-\eta} - G_{-\sigma,\eta})\Re(c_1\bar{c}_2\mathbf{S}_{21}). \end{aligned} \tag{4.16}$$

In the special case that there is no compensation ($\tau_C = 0$, $(\phi, \theta) = (0, 0)$), we obtain the well-known formula for the electrical power P_{SL} in frequency ω_{SL} as a function of the DGD τ [28]:

$$P_{\mathrm{SL}}(\tau) = \left[1 - 4\gamma(1-\gamma)\sin^2\left(\tfrac{\omega_{\mathrm{SL}}\tau}{2}\right)\right] P_{\mathrm{SL}}(0), \tag{4.17}$$

where $\gamma = |c_1|^2$ is the power splitting factor and

$$P_{\mathrm{SL}}(0) = \left| \int |f(t)|^2 e^{i\omega_{\mathrm{SL}}t} dt \right|^2. \tag{4.18}$$

Notice the correlation between the power P_{SL} and the DGD τ: For the 5 GHz spectral line with $\gamma = \frac{1}{2}$, $P_{\mathrm{SL}}(\tau)/P_{\mathrm{SL}}(0)$ decreases from 1 to 0 as τ increases from 0 to 100 ps.

If we reformulate the optimization problem to be of the form (4.1) and make an orthogonal change of coordinates, we obtain

$$\max_{\mathbf{x}\in\mathbf{R}^3} J(\mathbf{x}) \;=\; (\mathbf{x} - \mathbf{x}_0)^T \Lambda (\mathbf{x} - \mathbf{x}_0) \qquad \text{subject to} \;\; \|\mathbf{x}\| = 1, \tag{4.19}$$

where $\mathbf{x}_0 \in \mathbf{R}^3$ and Λ is a real diagonal matrix. Using the method of Lagrange multipliers, we find that in the nondegenerate case there are no more than six critical points on S^2, including a global maximum and global minimum. However, it is not possible to obtain explicit formulae for the critical points since they are the zeros of a degree six polynomial.

In Figure 4.2 we plot the analytical objective function given by formula (4.15).

This plot corresponds to an approximation of the fiber when only first-order PMD is present, and it should be compared to the objective function shown in Figure 3.4. The difference between the two objective functions is due to the large amount of second-order PMD present in the fiber realization used for Figure 3.4. Note, however, that the global maximum is in approximately the same location in both figures. There are four critical points of the objective function on S^2: two maxima, one minimum, and a saddle point. As in the case of the DOP ellipsoid, the saddle points on $\{\theta = 0\} \cup \{\theta = \pi\}$ are not critical points of (4.19). Notice that the eye opening objective function in Figure 3.3 has five critical points on S^2.

5. Importance sampling for PMD. In this subsection, we review the importance sampling algorithm we used to accurately compute outage probabilities due to PMD. Importance sampling increases the computational efficiency of Monte Carlo sampling from the space of fiber realizations.

The problem of evaluating the performance of a PMD compensator is quite challenging because system designers require compensators to maintain a high degree of integrity: The system should lose accuracy for no more than a few minutes per year. It is too time consuming to gather enough samples to accurately measure such low probabilities in a laboratory experiment. It is also not feasible to accurately evaluate the performance of a PMD compensator using numerical simulations based on standard Monte Carlo sampling.

PMD-induced outages occur when the eye opening is small. In a system without a PMD compensator, small eye openings are correlated to large DGD values, which occur very rarely since large DGD values correspond to sampling from the tail of a Maxwellian distribution. As we discussed in section 3, in systems with PMD compensators, the DGD in the transmission line can be at least partially canceled out by the DGD of the compensator at the carrier frequency. Therefore, after the compensator, large DGD values *at that frequency* are exceedingly rare. In general, however, the polarization dispersion vector, and hence the DGD, depends on frequency. It is useful to quantify the strength of this dependence using *second-order PMD* (SOPMD), which is defined to be the length, $|\overrightarrow{\Omega}_\omega|$, of the partial derivative of the polarization dispersion vector with respect to frequency. After the signal has traversed the compensation fiber, the DGD may be small at the carrier frequency, but the SOPMD may still be large enough that the eye opening will be small. To summarize, outages tend to occur only in the very rare case that either the DGD or the SOPMD is large relative to its average value. Recently, variance reduction techniques have been developed to greatly increase the computational efficiency of Monte Carlo simulations of PMD. Variance reduction techniques, which have been successfully applied in many contexts [16], [29], [34], simulate low probability events of interest by concentrating Monte Carlo simulations in those regions of the probability state space that are most likely to give rise to these events. More specifically, in the context of PMD [4], let $\boldsymbol{\theta}$ denote a particular fiber realization in the space Θ of all possible fiber realizations, and let p_θ be the probability density function (pdf) on Θ. Let $X : \Theta \to \mathbf{R}^K$ be a random variable on Θ, such as the eye opening or the two-dimensional quantity, $(|\overrightarrow{\Omega}|, |\overrightarrow{\Omega}_\omega|)$. Let $I : \mathbf{R}^K \to \{0, 1\}$ be the indicator function for a prescribed range of values $R \subset \mathbf{R}^K$, i.e., $I(x) = 1$ if $x \in R$ and $I(x) = 0$ otherwise. In practice, R could be a bin in the histogram of X. Our goal is compute the probability, P, that $X(\boldsymbol{\theta})$ lies in R,

$$P = \int_\Theta I(X(\boldsymbol{\theta}))\, p_\theta(\boldsymbol{\theta})\, d\boldsymbol{\theta}. \tag{5.1}$$

Using a standard Monte Carlo simulation, an estimator, $\hat{P}$, for P is given by

$$\hat{P} = \frac{1}{M}\sum_{m=1}^{M} I(X(\boldsymbol{\theta}_m)), \tag{5.2}$$

where we have drawn M samples $\boldsymbol{\theta}_m$ according to p_θ. If the events that lie in the region R are very rare, i.e., $P \ll 1$, then the relative variance of the Monte Carlo estimator $\hat{P}$ is $\sigma_{\hat{P}}/\hat{P} \sim (MP)^{-1/2}$. So, for example, if $P = 10^{-6}$, as is typically the case for an outage probability, then about $M = 10^8$ samples are required to ensure that the relative variance of $\hat{P}$ is on the order of 10%.

If a variance reduction technique is used, instead of drawing samples according to p_θ, we draw them according to a *biasing distribution*, p^*_θ, chosen so that $p^*_\theta(\boldsymbol{\theta})$ is relatively large when $X(\boldsymbol{\theta}) \in R$. Then the probability P can be expressed in the form

$$P = \int_\Theta I(X(\boldsymbol{\theta}))\, L(\boldsymbol{\theta})\, p^*_\theta(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \tag{5.3}$$

where $L = p_\theta/p^*_\theta$ is called the *likelihood ratio*. If we now use a Monte Carlo simulation to draw samples $\boldsymbol{\theta}^*_m$ from Θ according to the biasing distribution p^*_θ, then an estimator $\hat{P}^*$ for P is given by

$$\hat{P}^* = \frac{1}{M}\sum_{m=1}^{M} I(X(\boldsymbol{\theta}^*_m))L(\boldsymbol{\theta}^*_m). \tag{5.4}$$

If the biasing distribution p^*_θ is chosen appropriately, many more of the samples $\boldsymbol{\theta}_m$ will fall into the region R and contribute to the sum in (5.4). To ensure that $\hat{P}^*$ is computed correctly, each sample is weighted by its likelihood ratio, which is small where p^*_θ is large relative to p_θ. If the biasing distribution is chosen appropriately, then the relative variance of P can be much smaller than for an unbiased Monte Carlo simulation.

One important question to address is how the biasing distribution should be determined. When using importance sampling, the researcher must use a combination of physical intuition and mathematical analysis to determine appropriate biasing distributions. Recently, two different variance reduction techniques—an importance sampling algorithm and a multicanonical Monte Carlo method—have been developed for simulations of PMD. These two methods take different approaches to solving the problem of finding an appropriate biasing distribution. The multicanonical Monte Carlo method of Berg and Neuhaus [1] is an iterative method that was adapted for simulations of PMD by Yevick [60], [61] and later by A. Lima [37]. At the nth iteration of the method, samples are drawn from a biasing distribution $p^{*,n}_\theta$, and at the end of each iteration, $p^{*,n}_\theta$ is updated in such a way that as n increases there is approximately an equal number of hits in each bin of the histogram of the eye opening for that iteration. Consequently, after sufficiently many iterations, the relative variance between the bins of the eye opening histogram will be small (even for the low-probability, small eye opening bins).

The second technique (importance sampling) was developed by Biondini and Kath [2], [3], [4], [17] to generate large values of first- and second-order PMD. Building on the work of I. Lima [38], [39] and A. Lima [35], we used this algorithm to generate the results in this paper. To accurately compute outage probabilities for a PMD compensator, it is necessary to sample a large region in the $(|\overrightarrow{\Omega}|, |\overrightarrow{\Omega}_\omega|)$-plane. Since

this cannot be done efficiently using a single choice of biasing distribution [39], several biasing distributions, p_j^*, are used, each of which targets a different region of the $(|\overrightarrow{\Omega}|, |\overrightarrow{\Omega}_\omega|)$-plane. To compute the probability P in (5.3), we associate a weight function $w_j : \Theta \to \mathbf{R}$ to the distribution p_j^* and define a multiple importance sampling Monte Carlo estimator, $\hat{P}$, for P by

$$\hat{P} \;=\; \sum_{j=1}^{J} \frac{1}{M_j} \sum_{m=1}^{M_j} w_j(\boldsymbol{\theta}_{j,m})\, I(X(\boldsymbol{\theta}_{j,m}))\, L(\boldsymbol{\theta}_{j,m}), \tag{5.5}$$

where M_j samples are drawn using the jth distribution, and $\boldsymbol{\theta}_{j,m}$ is the mth such sample. A formula for the relative variance of $\hat{P}$ which was used for the results in this paper is given in [4].

The choices of the biasing distributions and the weights can have a large effect on the relative variance of the estimator $\hat{P}$ in (5.5). For this paper, we used the simple and efficient choice of weights that is given by the *balance heuristic* [56]. With this heuristic,

$$w_j(\boldsymbol{\theta}) \;=\; \frac{M_j p_j^*(\boldsymbol{\theta})}{\sum_{j'=1}^{J} M_{j'} p_{j'}^*(\boldsymbol{\theta})}; \tag{5.6}$$

i.e., the weight $w_j(\boldsymbol{\theta})$ is the probability of realizing the sample $\boldsymbol{\theta}$ using p_j^*, relative to the probability of realizing this sample using all J biasing distributions. Therefore, the distribution p_j^* is weighted most heavily in those regions of the sample space Θ where it is largest.

To define a biasing distribution that targets a region of the $(|\overrightarrow{\Omega}|, |\overrightarrow{\Omega}_\omega|)$-plane, Fogal, Biondini, and Kath [17] first determined fiber realizations that maximize a specified linear combination of $|\overrightarrow{\Omega}|$ and $|\overrightarrow{\Omega}_\omega|$. Consider, for example, the simplest case of maximizing the DGD. From (2.15) we see that to maximize the DGD, the rotation matrices $\mathtt{Q}_n$ should be chosen so that the vector $\mathtt{Q}_n \overrightarrow{\Omega}^{(n-1)}$ is aligned with $\Delta\overrightarrow{\Omega}_n$. This set of rotations, $\{\mathtt{Q}_n\}_{n=1}^N$, defines a particular fiber realization. Once this fiber realization has been determined, a biasing distribution is chosen that preferentially selects nearby fiber realizations. In this way, a family of biasing distributions can be chosen, each of which targets a different region of the $(|\overrightarrow{\Omega}|, |\overrightarrow{\Omega}_\omega|)$-plane. For the results in this paper, we used the ten biasing distributions described in [39]. This multiple importance sampling algorithm has been successfully used to simulate low-probability regions in the $(|\overrightarrow{\Omega}|, |\overrightarrow{\Omega}_\omega|)$-plane. For example, the joint pdf of $(|\overrightarrow{\Omega}|, |\overrightarrow{\Omega}_\omega|)$ has been calculated down to probability levels of 10^{-8} or less with a relative variance of less than 10% using a total of only 6×10^5 samples [39].

To evaluate the performance of a PMD compensator, we actually need to generate fiber realizations that produce low-probability, small values of the eye opening. Although the importance sampling algorithm is not explicitly designed to generate small eye opening values, A. Lima [36] demonstrated that there is a strong correlation between small eye openings and large values of $(|\overrightarrow{\Omega}|, |\overrightarrow{\Omega}_\omega|)$. She reached this conclusion in a study of PMD compensators by comparing results obtained using importance sampling with those obtained using the multicanonical Monte Carlo algorithm. The multiple importance sampling algorithm has several advantages over the multicanonical Monte Carlo algorithm, at least for simulations of PMD: The relative variance is easier to calculate, the algorithm is more computationally efficient, and it can be

easily parallelized by using different processors to draw samples from the different biasing distributions.

6. Optimization. The original optimization studies we performed were carried out using an object-oriented package that performed local optimization only. The Hilbert Class Library (HCL) code was developed at Rice University [22], and preliminary results obtained using this optimization software are discussed in [64]. As described in that paper, we used HCL's limited-memory BFGS (LMBFGS) [40] algorithm with line search [10] to determine an appropriate rotation for the polarization controller. While these early results were intriguing and allowed us to compare the performance of the optimization algorithm with the spectral line and DOP ellipsoid feedback mechanisms, the engineering problem is to find the "best" rotation for the polarization controller, i.e., to find the global optimum. While LMBFGS is a fast (Newton-based) technique that has been "globalized" to accommodate arbitrary initial guesses via the line search feature of HCL, the technique is not guaranteed to find the global optimum. Therefore, after completing the previous study, we decided to incorporate into our optics simulator an object-oriented optimization package that contains a variety of optimization tools (including some global techniques). The Design Analysis Kit for Optimization and Terascale Applications (DAKOTA) is the optimization package we used to obtain the results presented in this paper. The package was developed at Sandia National Labs to allow users to optimize their (generally PDE-based) simulation models for purposes of engineering design.

Our optimization problem is an unconstrained problem of the form (4.2) (although simple bound constraints may be placed on the rotation angles from periodicity), and our goal is to find the global optimum given reasonable computer time limitations. The design variables are continuous, and no analytic gradient information is available.

The experiments described in this paper fall into three categories of optimization jobs—local optimization, multistart jobs which use local optimization repeatedly on the same problem, and global hybrid optimization runs. The local optimization was performed using DAKOTA's *OPT++* library [11]. The *OPT++* library contains mostly gradient-based nonlinear programming algorithms for constrained or unconstrained optimization. In this paper we chose *OPT++*'s conjugate gradient (CG) method, which is appropriate for unconstrained optimization. The optimization in all these experiments is performed over two rotation angles (ϕ and θ). Numerical (finite difference) gradients are used in the CG algorithm with a relative finite difference step size of 0.0001. The gradient stopping tolerance is also set to the default value of 10^{-4}.

The second set of numerical experiments invokes DAKOTA's multistart capabilities. In each multistart job, a series of local optimization runs are completed, each using a different starting point. The multistart jobs also use *OPT++*'s CG routine. Numerical finite difference gradients are used with the same tolerances as in the single starting point local optimization experiments. These multistart jobs were run using either two, four, or nine equally spaced starting points. Since the ϕ values range over the interval $[-\pi, \pi]$, and θ takes values over the interval from $[0, \pi]$, the nine starting points were chosen to be the equally spaced points $(\phi, \theta) = \{(-\pi, 0), (-\pi/3, 0), (\pi/3, 0), (-\pi, \pi/3), (-\pi/3, \pi/3), (\pi/3, \pi/3), (-\pi, 2\pi/3), (-\pi/3, 2\pi/3), (\pi/3, 2\pi/3)\}$. Note that periodicity implies that these equally spaced points cover the (ϕ, θ)-domain uniformly. (The edges of the domain wrap.) For the four-point runs, the initial guesses are $(\phi, \theta) = \{(-\pi, 0), (-\pi, \pi/2), (0, 0), (0, \pi/2)\}$. Finally, the two-point runs use $\{(-\pi, 0), (0, \pi/2)\}$ as starting guesses.

The third type of optimization jobs are hybrid multilevel optimization schemes

which use a global optimization method initially and then switch to a local Newton-based scheme once the optimization is close enough to the solution to ensure fast convergence to the optimum. In our case we chose a genetic algorithm (GA) for the global method. GAs are based on Darwin's theory of survival of the fittest [12]. The GA starts with a random selection of points called a "population." The values of the parameters being optimized over form a string of mathematical "DNA" (a conglomeration of values for the parameters being optimized) which then is adapted to a best fit (or optimal configuration) by a process of natural selection, breeding, and mutation [12]. GAs are convenient when there are multiple local optima or when gradients cannot be computed easily. In those cases, GAs can be used to determine regions of the solution space where the global optimum may be located [12]. Global methods such as GAs, however, are slow to complete convergence to a minimizer and are best used in conjunction with a fast local method. Typical GA behavior shows a rapid decrease in the objective function initially, but then a steady slowing of progress towards the minimum. Often only a few initial GA iterations suffice to move the focus of the optimization to the region of interest.

In our GA runs, we chose a population size of five points (representing five (ϕ, θ) pairs) and ran the GA for a maximum of 25 function evaluations. The stopping tolerance for convergence was chosen to be (a loose) 5×10^{-2}. Once the GA run is finished, control is passed to a local method (in our case we chose the Fletcher–Reeves conjugate gradient algorithm from the *CONMIN* package in DAKOTA). The CONMIN package contains both constrained and unconstrained minimization algorithms similar to *OPT++*. The same default stopping tolerances were used for the local optimization part of the multilevel jobs as for the purely local *OPT++* jobs.

7. Numerical results. We now describe a suite of numerical experiments that we ran to investigate the relative performance of the three compensation feedback mechanisms (DOP ellipsoid, spectral line, and eye opening). The experiments used the local and global optimization routines from DAKOTA discussed in section 6. Specifically, for each of the three feedback mechanisms we ran five optimization jobs involving 200,000 fiber realizations (or Monte Carlo simulations) each. For each feedback mechanism we optimized the compensator using local optimization (conjugate gradients), multilevel global optimization, and a multistart global routine involving varying numbers of starting guesses. The aim in all cases was to compensate for the transmission fiber DGD. As discussed in section 6, the multilevel strategy starts with a small number of GA iterations. After narrowing the optimization search area via the GA, the algorithm passes control to a fast local gradient-based optimization routine (CG in our case) to find the (hopefully) global optima. For the multistart runs, local optimization is "globalized" by running the local optimization to completion from a few different starting points. We ran three sets of multistart jobs for each feedback mechanism. In the first multistart job we used two initial points. In the second we used four starting guesses, and in the last simulation we used nine points. The multistart algorithm is an ad hoc globalization of local methods but suffers from the fact that the starting points are chosen at random. No prior (or current) knowledge of the objective function surface is used.

The goal is to drive the outage probability down to a small acceptable level. System designers typically require a probability of no more than 10^{-3}–10^{-6} that the eye is 40% closed relative to its original back-to-back value (i.e., 2 dB down from the ideal case of no PMD). Note that an eye corresponding to a 1 dB decrease from the back-to-back signal is 79% open. An eye that is 2 dB down is 63% open, and 3 dB

down corresponds to an eye that is only 50% open. Values of 1–2 dB reduction are of greatest interest. A value of 3 dB down corresponds to an eye that is so closed as to indicate system failure.

Parameters in the optical communications simulator which were fixed in these simulations include specifications about the transmitted signal, transmission fiber, and receiver. We modeled the transmitted signal by allocating a time interval of duration 100 ps to each binary digit (corresponding to a data rate of 10 Gb/s). We used the 16-bit data pattern 1001101011110000, and we encoded the binary data onto the amplitude of the signal by setting $\mathbf{U}(0,t) = (\chi_1(t), 0)^T$. Here $\chi_1(t) = 1$ when t is in the time interval of a one, and $\chi_1(t) = 0$ when t is in the time interval of a zero. The signal was polarized since $U_2(0,t) = 0$. We then applied a Gaussian filter to $\mathbf{U}$ to produce a smooth signal. The width of the filter was chosen so that the time required for the signal power to increase from 10% to 90% of its maximum value of 1 milliWatt was 30 ps. The transmission fiber was modeled using the coarse-step method (described in section 2.4) with 80 fiber segments. The average DGD of the transmission fiber was set at 30 ps, and the DGD of the compensation fiber was fixed at 30 ps. We modeled the receiver using a 60 GHz Gaussian-shaped optical filter, a photodetector that converts the optical power to electrical voltage, and a fifth-order electrical Bessel filter with half-width of 8 GHz.

Our first conclusion from these numerical experiments is that the shape of the objective function depends strongly on the feedback mechanism. The eye opening feedback signal is highly correlated to the bit-error ratio, which ultimately is the quantity to optimize. Unfortunately the eye feedback mechanism results in a fairly rough objective function. On the other hand, both the spectral line and DOP feedback mechanisms are smoother than the eye and so are easier to optimize. Figures 3.3, 3.4, and 3.5 are plots of the objective functions for the eye, spectral line, and DOP feedback mechanisms for a typical fiber realization. We see that for this example, the eye objective function has a steep ridge. Features such as this can hinder the progress of local methods towards the maximum. The DOP and spectral line objective functions are considerably smoother.

In Figure 7.1 we show results for the eye opening feedback mechanism. We plot outage probability as a function of the eye opening penalty (in dB down from the back-to-back signal) for six cases: (1) uncompensated signal, (2) compensated case using local optimization, (3) compensated using the global multilevel strategy, and (4–6) compensated using the global multistart strategy with varying numbers of initial guesses. We note that for the majority of fiber realizations, the eye diagram is mostly open and so the eye opening penalty is small. In other words, there is a large probability of a small eye opening penalty. In Figure 7.1, the outage probability is the probability that the eye opening penalty exceeds the value on the horizontal axis. The eye opening penalty will exceed 0 dB, whenever the eye is more closed with PMD than without it, which occurs almost all the time. Consequently, in Figure 7.1 the outage probability is approximately 1.0 when the eye opening penalty is 0 dB. It is only for the very rare fiber realizations with large DGD that the eye opening penalty is large, i.e., that the eye is mostly closed. Therefore, there is a very small outage probability that the eye opening penalty exceeds 2 dB. Clearly, if the typical case resulted in a partially or fully closed eye, the state of optical communications would be considerably more dire. The need for importance sampling to bias the Monte Carlo simulations towards the "bad cases" reflects that in most cases, the DGD of the fiber is small. As expected, the local CG method does compensate for the DGD by reducing the outage probability from 10^{-2} to about 5×10^{-4} at an eye opening penalty

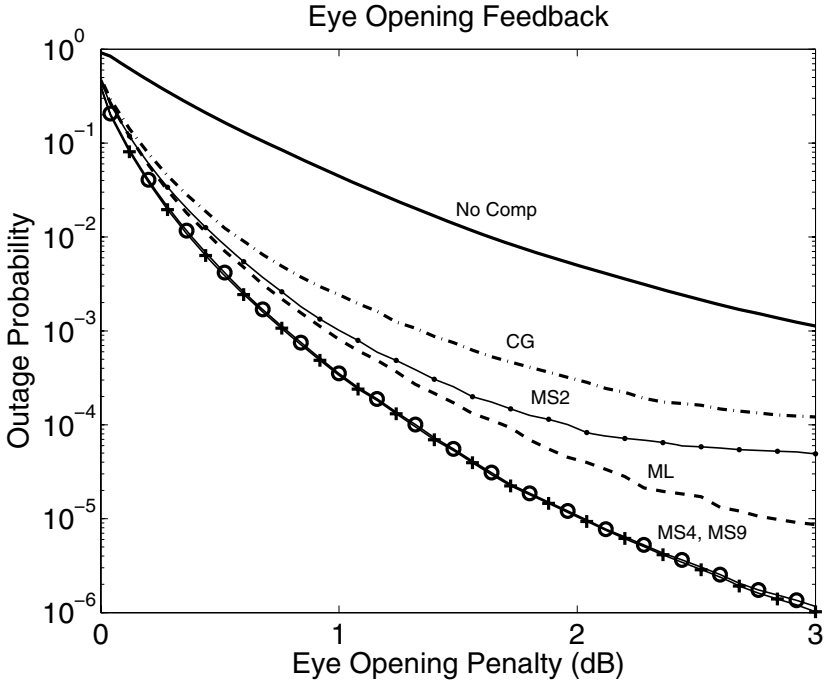


FIG. 7.1. *Outage probability as a function of eye opening penalty for different optimization algorithms applied to the eye opening feedback mechanism. The curves shown include the no compensation case (thick solid line), local (CG) optimization (dot-dash line), the multilevel hybrid optimization scheme (thick dashed line), and three curves for the multistart method (thin dotted line for multistart with two starting guesses; thin line with circles for multistart with four initial guesses; and thin line with crosses for multistart with nine starting points for the local optimization runs). Note that there is no difference in optimization results for the multistart scheme using four and nine starting points in the case of the eye opening feedback mechanism.*

TABLE 7.1

Average number of function evaluations per optimization routine and specified feedback mechanism. The average is taken over 200,000 *MC simulations.*

Average number of function evaluations					
	CG	Multilevel	Multistart 2 pts	Multistart 4 pts	Multistart 9 pts
SL	53	64	95	191	454
DOP	52	65	99	199	458
Eye	52	68	99	194	459

of 2 dB. However, the best global method in this case (multistart with four points) further decreases the outage probability to about 10^{-5}. We believe that multistart with four points is finding the global optimum for the eye feedback signal since the multistart algorithm with nine points is unable to reduce the outage probability further. Consequently, this result represents the best possible performance for this fixed DGD compensator.

As Table 7.1 indicates, the cost of the multistart jobs is high. Multistart four- and nine-point schemes require 3–7 times as many function evaluations as the multilevel technique (200 or 450 iterations, respectively, versus about 65). In the case of the eye feedback mechanism, multilevel optimization reduces the outage probability to about 4×10^{-5} at 2 dB. In general the number of function evaluations required (for

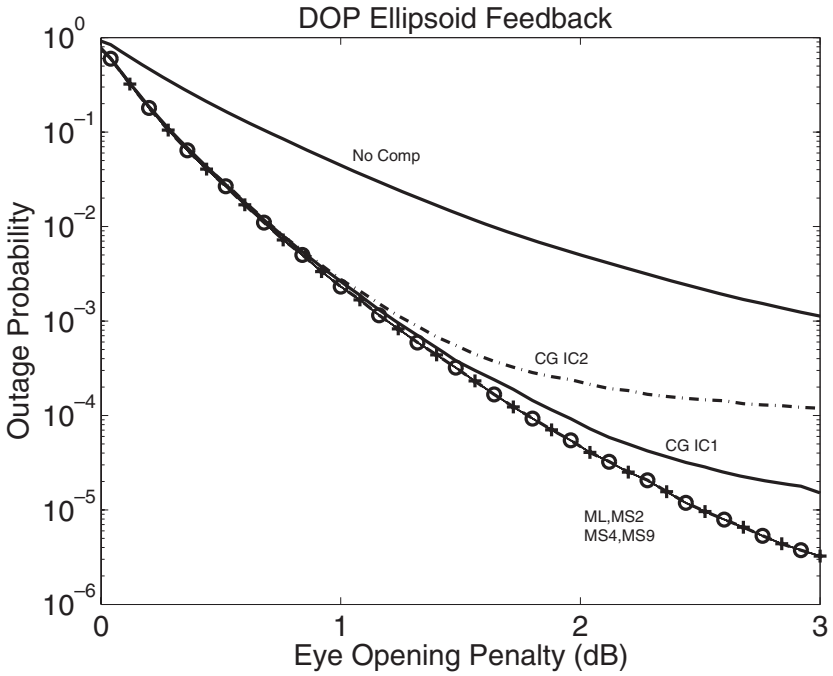


FIG. 7.2. *Outage probability as a function of eye opening penalty for different optimization algorithms applied to the DOP feedback mechanism. The curves shown include the no compensation case (thick solid line), local (CG) optimization (also a thick solid line) starting from the initial guess* $(\phi, \theta) = (0, \pi/2)$*, and local optimization (CG) (dot-dashed line) starting from the initial guess* $(\phi, \theta) = (0, 0)$*. The curves for the multilevel hybrid optimization scheme (thick dashed line), multistart method with two starting guesses (thin dotted line), multistart with four initial guesses (thin line with circles), and multistart with nine initial guesses (thin line with crosses) lie on top of each other.*

all feedback mechanisms and all levels of DGD) increases from least expensive to most costly in the following order: local conjugate gradient method, the multilevel technique, and, finally, multistart. The multistart job cost increases approximately linearly with the number of starting points used. Table 7.1 shows the number of function evaluations for all the algorithms and feedback mechanisms at a fixed eye opening penalty (2 dB).

In Figures 7.2 and 7.3 we show the outage probability versus eye opening penalty for the local and global optimization methods and the DOP and spectral line feedback mechanisms, respectively. These figures include curves for two different CG simulations. The first run (CG IC1) used a starting point of $(\phi, \theta) = (0, \frac{\pi}{2})$, and the second (CG IC2) started from $(0, 0)$. For the DOP ellipsoid, the local CG algorithm (CG IC1) took an average of 50 function evaluations and resulted in an outage probability of 7×10^{-5} at 2 dB. The multilevel scheme took 65 function evaluations to arrive at an outage probability of 3×10^{-5}. Multistart takes considerably more function evaluations to arrive at the same outage probability value (see Table 7.1). The lack of model-based knowledge built into the multistart scheme is clearly to blame for this dramatic increase in the number of function evaluations without a corresponding outage probability reduction.

In sections 3 and 4 we saw that the DOP ellipsoid is very smooth and usually has

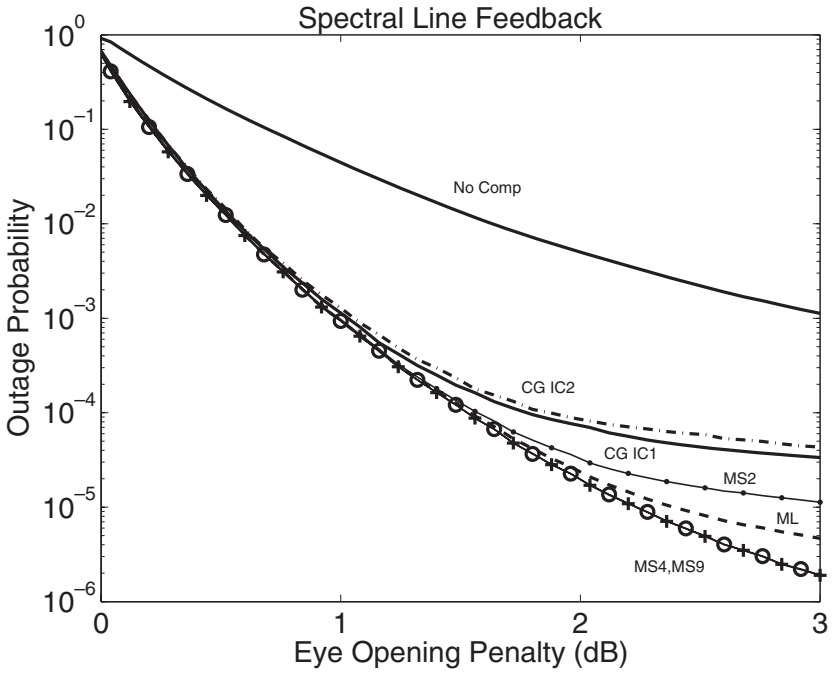


FIG. 7.3. *Outage probability as a function of eye opening penalty for different optimization algorithms applied to the spectral line feedback mechanism. The curves shown include the no compensation case (thick solid line), local (CG) optimization (also a thick solid line) starting from the initial guess* $(\phi, \theta) = (0, \pi/2)$*, and local optimization (CG) (dot-dashed line) starting from the initial guess* $(\phi, \theta) = (0, 0)$*. The curve for the multilevel hybrid optimization scheme is denoted by a thick dashed line, and the multistart method with two starting guesses is indicated by a thin dotted line. The multistart method with four initial guesses (thin line with circles) and multistart with nine initial guesses (thin line with crosses) coincide for this feedback mechanism.*

only one maximum whereas the spectral line can have at least two maxima. Starting from the south pole (IC2), there is a small probability that the CG algorithm will stall near a saddle point since the objective function can be very flat due to distortions inherent in the spherical coordinate mapping. For the DOP ellipsoid, starting from the equator (IC1), the CG algorithm will usually head straight to the top of the only hill. Since this maximum is between the equator and the saddle points at the south pole, there is very little chance that the algorithm will go via these saddle points and hence very little chance that it will get stuck there. Therefore, the outage probability is lower with IC1 than with IC2. Different initial conditions lead to outage probability curves which lie between the curves for IC1 and IC2.

Figure 7.4 compares the best local and best global schemes for the three feedback mechanisms. For each of the pdfs used to determine the curves in this figure, importance sampling produces a relative variation in each bin of the histogram of less than 10%. One can surmise the relative smoothness of the objective functions for the three feedback mechanisms by noting the distance between the local and global curves in the figure. For the eye feedback signal, the local method is least effective at reducing outage probability. Yet the global scheme is more effective for this objective function than it is for either DOP or spectral line. In other words, for the eye feedback mechanism, the outage probability for the best global optimization algorithm is over

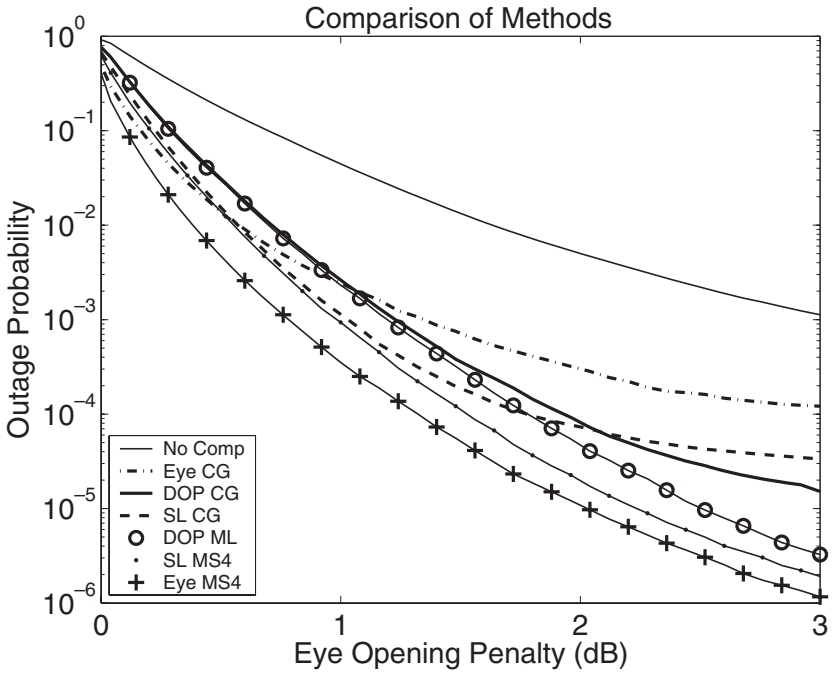


FIG. 7.4. *A comparison of the local and best global optimization results for the three feedback mechanisms. For local optimization of the spectral line and DOP ellipsoid we show the results for the starting point IC*1 $(\phi, \theta) = (0, \pi/2)$.

an order of magnitude smaller than for local optimization. A much smaller decrease occurs for the spectral line feedback signal. (All these comparisons were made at the 2 dB point in the plots.)

The DOP and spectral line feedback signals act as smooth surrogate approximations to the more realistic but bumpier eye signal. The DAKOTA reference manual [11] describes surrogate-based optimization as an iterative process that periodically recalibrates an approximate model via data from a true model. The DOP and spectral line functions are not true surrogates, as in this work we fix the objective functions used throughout a particular optimization run. No updates to the shape of objective function occur during the course of the optimization. Finally, we note (Table 7.1) that for the spectral line, the multilevel scheme does a better job of reducing the objective function than the two-point multistart scheme, but at 50% less cost. Our conclusions are that the simpler smooth DOP and spectral line feedback mechanisms do a good job of approximating the realistic (but bumpy) eye. Moreover, multilevel appears to do the best job of minimizing the number of function evaluations while still reducing the outage probability.

In Figure 7.5 we assess the statistical effect that higher-order PMD has on the performance of the spectral line and DOP ellipsoid feedback mechanisms. For each fiber realization we chose the setting of the polarization controller by optimizing the analytical objective functions given by (4.3) and (4.15) that we obtained using the first-order PMD approximation of the fiber. Given the solution to this surrogate optimization problem, we then computed the performance of the compensator using the original all-order PMD fiber realization. For the DOP ellipsoid we calculated

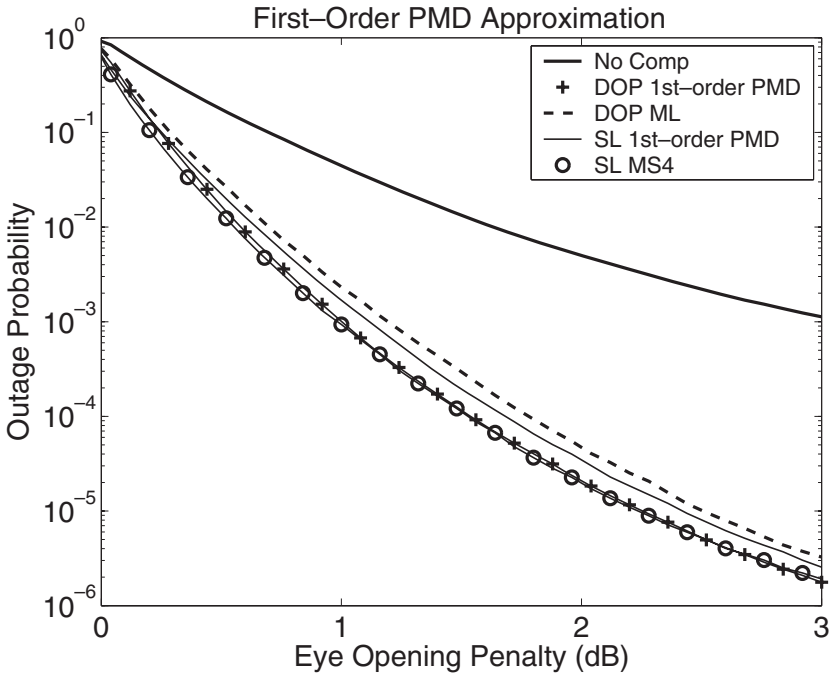


FIG. 7.5. *A comparison of outage probability for the first-order PMD approximation of the objective functions coming from the DOP ellipsoid (thin line with pluses) and the spectral line (thin line) versus the all-order PMD objective functions for the DOP ellipsoid (thick dashed line) and the spectral line (thin line with circles). These results were obtained using global optimization. The result without compensation is shown as a thick line.*

the global maximum analytically using the formula given above (4.5), while for the spectral line we optimized the analytical objective function numerically using the multistart strategy with four starting points. In Figure 7.5, for the DOP ellipsoid, we compare the "analytical" outage probability curve (thin line with pluses) to the one we obtained numerically using the all-order PMD objective functions (thick dashed line). We also compare the analytic and numerical outage probability curves for the spectral line, which we show with a thin line and a thin line with circles, respectively. For both the DOP ellipsoid and the spectral line, the fairly close agreement between the analytic and numerical curves confirms that we can regard the all-order PMD objective functions as being perturbations of objective functions for fibers with only first-order PMD.

Next, we explain the relative performance of the different methods in Figure 7.4 by comparing the relative location of the local and global maxima of the different objective functions. This discussion will also quantify the degree to which the spectral line and DOP ellipsoid objective functions act as smooth surrogates of the eye opening. We begin by discussing the three feedback mechanisms optimized via global methods. First, we observe that across the entire range of eye opening penalty values in Figure 7.4, the outage probability is always larger for the DOP ellipsoid than for the spectral line, and that the eye opening feedback mechanism has the lowest outage probability. The primary reason for the poorer performance with the DOP ellipsoid is that the distance between the global maxima of the DOP ellipsoid and eye objective

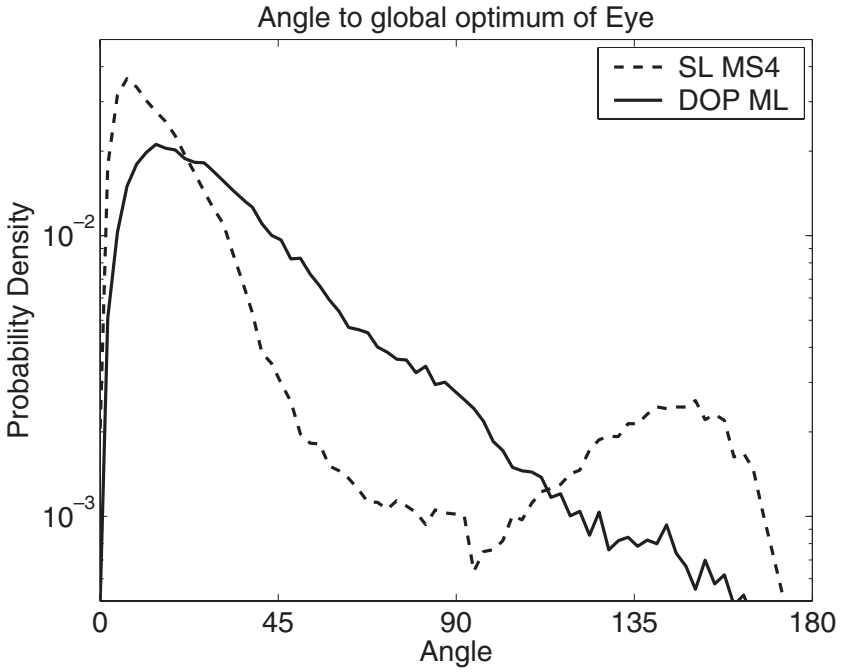


FIG. 7.6. *The pdf of the angle between the global maximum of the eye objective function and the maximum obtained for the spectral line using multistart with four starting points (thick dashed curve) and for the DOP ellipsoid using the multilevel algorithm (thick solid curve).*

functions tends to be larger than between the global maxima of the spectral line and eye objective functions.

To verify this observation, we gathered statistics of the distances between the global maxima for the different feedback mechanisms. To define a physically meaningful notion of distance between two rotations of the polarization controller in a compensator, we begin by recalling that any rotation of S^2 can be expressed as a rotation $R_{\overrightarrow{r}}(\Psi)$ by an angle $\Psi \in [0, \pi]$ about an axis $\overrightarrow{r} \in S^2$. Given two rotations R_1 and R_2, consider the rotation E_0 such that $R_2 = E_0 R_1$, and let Ψ_0 be the angle such that $E_0 = R_{\overrightarrow{r}_0}(\Psi_0)$. In our compensator model the rotations R_1 and $R_X(\pi)R_1$ have the same effect on the signal. Therefore, we also consider the rotation E_π defined by $R_2 = E_\pi R_X(\pi) R_1$ and let Ψ_π be the angle such that $E_\pi = E_0 R_X(\pi) = R_{\overrightarrow{r}_\pi}(\Psi_\pi)$. Finally, we define an angle $\Psi \in [0, \pi]$ by $\Psi = \min\{\Psi_0, \Psi_\pi\}$. The angle Ψ is our measure of distance between two rotations R_1 and R_2 performed by the polarization controller. (Note though that the distance function given by Ψ does not satisfy the triangle inequality.) In the case of only first-order PMD, the angle between global maximum and global minimum of the DOP ellipsoid objective function given by (4.3) is 180°.

In Figure 7.6 we plot the pdf of the angle between the numerically computed global maxima of the spectral line and the eye opening (thick dashed curve) and between the global maxima of the DOP ellipsoid and eye opening (thick solid curve). The most likely angle is 7° for the spectral line and 14° for the DOP ellipsoid. We observe that the probability that the angle is between 30° and 90° degrees is significantly greater for the DOP ellipsoid than for the spectral line. Therefore the global maximum

of the spectral line objective function is usually closer to that of the eye opening than is the global maximum of the DOP ellipsoid. Consequently, the eye opening tends to be somewhat larger for the spectral line than for the DOP ellipsoid. This observation explains why the outage probability is smaller for the spectral line than for the DOP ellipsoid when global optimization is used. One of the physical reasons for this performance difference is that the DOP ellipsoid objective function is defined by minimizing the output DOP over all possible input polarization states. However, when we computed the eye opening penalties for the outage probability curves, we chose the input polarization state to be $(S_1, S_2, S_3) = (1, 0, 0)$ rather than choosing it to be the state which resulted in the smallest DOP at the receiver. If we maximize the DOP for the input polarization state with the smallest output DOP, we do not obtain the same global maximum as we would if we maximized the DOP (or spectral line) when the input polarization state is $(1, 0, 0)$.

We also observe in Figure 7.6 that the probability that the angle is between 130° and 170° is much larger for the spectral line than for the DOP ellipsoid. The physical reason for this feature can be explained using the formulae for the objective functions we derived in section 4: In the case of only first-order PMD the DOP ellipsoid has only one maximum, whereas the spectral line objective function can have at least two maxima. Therefore, in a small proportion of cases the global maximum of the spectral line can be far from that of the eye opening. For example, suppose as in Figures 3.3 and 3.4 that the eye opening has local maxima at (ϕ_1, θ_1) and (ϕ_2, θ_2), and that the spectral line also has local maxima located near these two points. It could happen that the global maximum for the eye opening is at (ϕ_1, θ_1) whereas the global maximum for the spectral line is located near (ϕ_2, θ_2).

We also found that if we gather statistics over only those fiber realizations for which the second-order PMD is large relative to the DGD, then the most likely angle between the global maxima of the DOP ellipsoid and the eye opening feedback increases markedly. In contrast, the most likely angle between the global maxima of the spectral line and the eye opening is unchanged. (We do not show these results.) These results suggest that when higher-order PMD is introduced, the global maxima of the DOP ellipsoid and the eye opening tend to move apart from each other, whereas the global maximum of the spectral line remains closer to that of the eye opening. This difference in behavior with higher-order PMD provides a second explanation for why the outage probability is smaller for the spectral line than for the DOP ellipsoid when global optimization is used.

To summarize, when global optimization is used, it is important to choose an objective function with the property that the global maximum is close to that of the eye opening. Therefore, with global optimization, the spectral line is a better surrogate for the eye opening than is the DOP ellipsoid.

Finally, we explain the relative performance with the three feedback signals when local optimization is used. Looking again at Figure 7.4, when the eye opening penalty is larger than 2 dB, the eye opening feedback does not perform as well as the spectral line with local optimization. The performance with the DOP ellipsoid is comparable to that with the spectral line, but as we saw in Figure 7.2, it can depend significantly on the choice of starting point for the local optimization. To explain the poorer performance with the eye opening feedback, for each feedback mechanism we plot the pdf of the angle, Ψ, between the local maxima reached with conjugate gradients and the global maximum (see Figure 7.7). First, we observe that the DOP ellipsoid and the spectral line have slightly higher probabilities of a very small angle ($\Psi \approx 0$) between the local and global maxima than is the case for the eye opening feedback

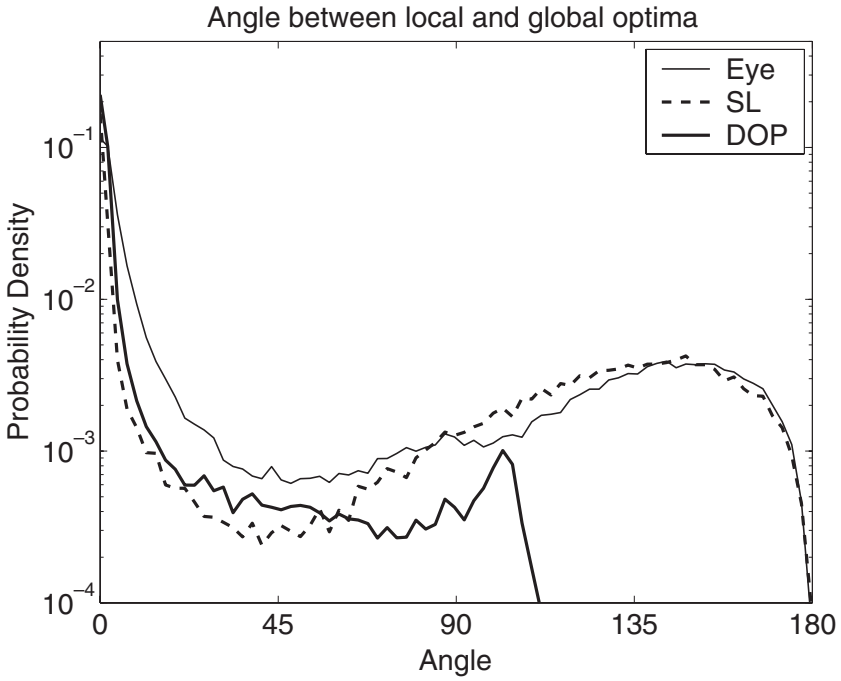


FIG. 7.7. *The pdf of the angle between the local and global maxima for the eye opening (thin solid curve), spectral line (dashed curve), and DOP ellipsoid (thick solid curve). The local and global optimization algorithms are the ones shown in Figure* 7.4.

mechanism. However, there is a significantly higher probability that the angle is between 10° and 50° for the eye opening. This high probability is not seen for the DOP or spectral line objective functions. These observations provide further evidence of the roughness of the eye opening objective function and help to explain why the eye objective function is the worst feedback mechanism for local optimization when the eye opening penalties are large. The other significant feature in Figure 7.7 is that the probability that the angle exceeds 100° is extremely small for the DOP ellipsoid, but it is relatively large for the other two objective functions. This feature is also present in the pdfs if we gather statistics over only those rare fiber realizations for which the eye opening penalty for the DOP ellipsoid (or spectral line) is larger than 1 dB.

We found that the curves in Figure 7.7 have approximately the same shape when we gather statistics over fibers with large second-order PMD. Consequently, even with higher-order PMD, the DOP ellipsoid objective function tends to be very smooth and to have only one maximum, just as the analysis in section 4 showed for fibers with only first-order PMD. Why then for the DOP ellipsoid is the outage probability at 3 dB smaller with global than with local optimization in Figure 7.4? One possible reason is that for some fiber realizations we observed that the DOP ellipsoid objective function has suboptimal, wide flat regions. Also, when second-order PMD is large enough, there can be small bumps in the bottom of the valleys which may present a problem for a local algorithm.

8. Conclusions. In an optical fiber communication system, binary data is transmitted through optical fiber using a sequence of pulses of light. The birefringence of the fiber causes the pulses to spread and distort as they propagate and increases the

probability that bit errors will occur. A simple optical PMD compensator can be used to reduce these distortions and errors. Since the birefringence varies randomly over time, the compensator must be continually optimized with the aid of a feedback signal. To evaluate the performance of a compensator, an optimization problem must be solved for a large number of random realizations of the birefringence. In each case, the goal is to locate the operating point at which the bit-error ratio is smallest. Since it is not possible to measure the bit-error ratio in a real system, we studied three commonly used feedback signals: the eye opening, spectral line, and DOP ellipsoid. To adequately sample the very rare fiber realizations that result in a large uncompensated bit-error ratio, we performed Monte Carlo simulations with multiple importance sampling. We quantified the degree to which the performance of a compensator depends on the choice of feedback signal and optimization algorithm by computing the probability that the eye opening penalty exceeds a given threshold, i.e., that the bit-error ratio is large.

Although the eye opening is highly correlated to the bit-error ratio, its objective function is quite rough and is therefore hard to optimize. Our results show that the spectral line and DOP objective functions act as smooth surrogate approximations to the rougher eye opening. In the special case of first-order PMD, we proved that the spectral line objective function can have as many as six critical points on the sphere, whereas the DOP ellipsoid has only one maximum and one minimum. We verified that these conclusions also hold statistically over a wide range of fiber realizations with higher order PMD. Since the spectral line objective function is similar to the eye opening, the performance is somewhat better with the spectral line than with the DOP ellipsoid when global optimization is used. However, the DOP ellipsoid objective function is smoother and easier to optimize than the spectral line. In conclusion, since it is most desirable to have a low outage probability for large eye opening penalties, we suggest that multilevel optimization with the DOP ellipsoid feedback gives a good trade-off between the requirements of high performance, computational cost, and complexity of the feedback mechanism.

Acknowledgments. We thank Paul Leo (YAFO Networks), Curtis Menyuk, Aurenice Lima, Ivan Lima, Brian Marks (all at UMBC), and Michael Eldred and Tony Giunta (both at Sandia National Labs) for generously sharing their knowledge with us. We also thank the reviewers for helpful comments.

REFERENCES

[1] B. A. Berg and T. Neuhaus, *Multicanonical algorithms for first order phase transitions*, Phys. Lett. B, 267 (1991), pp. 249–253.

[2] G. Biondini and W. L. Kath, *PMD emulation with Maxwellian length sections and importance sampling*, IEEE Photon. Technol. Lett., 16 (2004), pp. 789–791.

[3] G. Biondini, W. L. Kath, and C. R. Menyuk, *Importance sampling for polarization-mode dispersion*, IEEE Photon. Technol. Lett., 14 (2002), pp. 310–312.

[4] G. Biondini, W. L. Kath, and C. R. Menyuk, *Importance sampling for polarization-mode dispersion: Techniques and applications*, J. Lightwave Technol., 22 (2004), pp. 1201–1215.

[5] M. Born and E. Wolf, *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light*, 7th ed., Cambridge University Press, Cambridge, UK, 1999.

[6] M. Brodsky, P. Magill, and N. Frigo, *Polarization-mode dispersion of installed recent vintage fiber as a parametric function of temperature*, IEEE Photon. Technol. Lett., 16 (2004), pp. 209–211.

[7] F. Buchali and H. Bülow, *Adaptive PMD compensation by electrical and optical techniques*, J. Lightwave Technol., 22 (2004), pp. 1116–1126.

[8] F. Buchali, S. Lanne, J.-P. Thiery, W. Baumert, and H. Bülow, *Fast eye monitor for* 10 *Gbit/s and its application for optical PMD compensation*, in Proceedings of the Optical Fiber Communication Conference, Anaheim, CA, 2001, paper TuP5.

[9] P. C. Chou, J. M. Fini, and H. A. Haus, *Real-time principal state characterization for use in PMD compensators*, IEEE Photon. Technol. Lett., 13 (2001), pp. 568–570.

[10] J. Dennis and R. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1983.

[11] M. Eldred, A. Giunta, B. van Bloemen Waanders, S. F. Wojtkiewicz, Jr., W. Hart, and M. Alleva, *DAKOTA, a Multilevel Parallel Object-Oriented Framework for Design Optimization, Parameter Estimation, Uncertainty Quantification, and Sensitivity Analysis. Version* 3.0 *Reference Manual*, Technical report SAND2001-3515, Sandia National Laboratories, Albuquerque, NM, 2002.

[12] M. Eldred, A. Giunta, B. van Bloemen Waanders, S. F. Wojtkiewicz, Jr., W. Hart, and M. Alleva, *DAKOTA, a Multilevel Parallel Object-Oriented Framework for Design Optimization, Parameter Estimation, Uncertainty Quantification, and Sensitivity Analysis. Version* 3.0 *Users Manual*, Technical report SAND2001-3796, Sandia National Laboratories, Albuquerque, NM, 2002.

[13] S. G. Evangelides, Jr., L. F. Mollenauer, J. P. Gordon, and N. S. Bergano, *Polarization multiplexing with solitons*, J. Lightwave Technol., 10 (1992), pp. 28–25.

[14] J. M. Fini, *Coherent Multi-photon Interference and Compensation of Polarization Dispersion*, Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, 2001.

[15] J. M. Fini, P. C. Chou, and H. A. Haus, *Estimation of polarization dispersion parameters for compensation with reduced feedback*, in Proceedings of the Optical Fiber Communication Conference, Anaheim, CA, 2001, paper WAA6.

[16] G. S. Fishman, *Monte Carlo: Concepts, Algorithms and Applications*, Springer-Verlag, New York, 1996.

[17] S. L. Fogal, G. Biondini, and W. L. Kath, *Multiple importance sampling for first- and second-order polarization-mode dispersion*, IEEE Photon. Technol. Lett., 14 (2002), pp. 1273–1275.

[18] G. J. Foschini and C. D. Poole, *Statistical theory of polarization dispersion in single mode optical fibers*, J. Lightwave Technol., 9 (1991), pp. 1439–1456.

[19] A. Galtarossa, L. Palmieri, M. Schiano, and T. Tambosso, *Measurement of birefringence correlation length in long, single-mode fibers*, Opt. Lett., 26 (2001), pp. 962–964.

[20] A. Galtarossa, L. Palmieri, M. Schiano, and T. Tambosso, *Statistical characterization of fiber random birefringence*, Opt. Lett., 25 (2000), pp. 1322–1324.

[21] W. Gander, G. H. Golub, and R. Strebel, *Least squares fitting of circles and ellipses*, BIT, 34 (1994), pp. 558–578.

[22] M. S. Gockenbach, M. J. Petro, and W. W. Symes, *C++ classes for linking optimization with complex simulations*, ACM Trans. Math. Software, 25 (1999), pp. 191–212.

[23] H. Goldstein, *Classical Mechanics*, Addison-Wesley, Reading, MA, 1980.

[24] J. P. Gordon and H. Kogelnik, *PMD fundamentals: Polarization-mode dispersion in optical fibers*, Proc. Natl. Acad. Sci. USA, 97 (2000), pp. 4541–4550.

[25] A. Hasegawa and F. Tappert, *Transmission of stationary nonlinear optical pulses in dispersive dielectric fibers.* I. *Anomalous dispersion*, Appl. Phys. Lett., 23 (1973), pp. 142–144.

[26] H. F. Haunstein, W. Sauer-Greff, A. Dittrich, K. Sticht, and R. Urbansky, *Principles for electronic equalization of polarization-mode dispersion*, J. Lightwave Technol., 22 (2004), pp. 1169–1182.

[27] F. Heismann, D. A. Fishman, and D. L. Wilson, *Automatic compensation of first-order polarization mode dispersion in a* 10 *Gb/s transmission system*, in Proceedings of the European Conference on Optical Communication, Vol. 1, Madrid, Spain, 1998, pp. 529–530.

[28] G. Ishikawa and H. Ooi, *Polarization-mode dispersion sensitivity and monitoring in* 40 *Gb/s OTDM and* 10 *Gb/s NRZ transmission experiments*, in Proceedings of the Optical Fiber Communication Conference, San Jose, CA, 1998, paper WC5.

[29] M. C. Jeruchim, P. Balaban, and K. S. Shamugan, *Simulation of Communication Systems*, Plenum, New York, 1992.

[30] I. P. Kaminow, *Polarization in optical fibers*, IEEE J. Quantum Electron., QE-17 (1981), pp. 15–22.

[31] I. P. Kaminow and T. Li, *Optical Fiber Telecommunications* IV-B: *Systems and Impairments*, Academic Press, San Diego, CA, 2002.

[32] M. Karlsson, J. Brentel, and P. A. Andrekson, *Long-term measurement of PMD and polarization drift in installed fiber*, J. Lightwave Technol., 18 (2000), pp. 941–951.

[33] N. Kikuchi, *Analysis of signal degree of polarization degradation used as control signal for optical polarization mode dispersion compensation*, J. Lightwave Technol., 19 (2001), pp. 480–486.

[34] D. P. Landau and K. Binder, *A Guide to Monte Carlo Simulations in Statistical Physics*, Cambridge University Press, New York, 2000.

[35] A. O. Lima, I. T. Lima, Jr., C. R. Menyuk, G. Biondini, B. S. Marks, and W. L. Kath, *Statistical analysis of the performance of PMD compensators using multiple importance sampling*, IEEE Photon. Technol. Lett., 15 (2003), pp. 1716–1718.

[36] A. O. Lima, I. T. Lima, Jr., C. R. Menyuk, and J. Zweck, *Performance evaluation of single-section and three-section PMD compensators using extended Monte Carlo methods*, in Proceedings of the Optical Fiber Communication Conference, Anaheim, CA, 2005, paper OME27.

[37] A. O. Lima, I. T. Lima, Jr., J. Zweck, and C. R. Menyuk, *Efficient computation of PMD-induced penalties using multicanonical Monte Carlo simulations*, in Proceedings of the European Conference on Optical Communication, Rimini, Italy, IEEE, 2003, paper We364, pp. 538–539.

[38] I. T. Lima, G. Biondini, B. S. Marks, W. L. Kath, and C. R. Menyuk, *Analysis of PMD compensators with fixed DGD using importance sampling*, IEEE Photon. Technol. Lett., 14 (2002), pp. 627–629.

[39] I. T. Lima, Jr., A. O. Lima, G. Biondini, C. R. Menyuk, and W. L. Kath, *A comparative study of single-section polarization-mode dispersion compensators*, J. Lightwave Technol., 22 (2004), pp. 1023–1032.

[40] D. Liu and J. Nocedal, *On the limited memory BFGS method for large scale optimization*, Math. Programming, 45 (1989), pp. 503–528.

[41] D. Marcuse, *Derivation of analytical expressions for the bit-error probability in lightwave systems with optical amplifiers*, J. Lightwave Technol., 8 (1990), pp. 1816–1823.

[42] D. Marcuse, C. R. Menyuk, and P. K. A. Wai, *Application of the Manakov-PMD equation to studies of signal propagation in optical fibers with randomly varying birefringence*, J. Lightwave Technol., 15 (1997), pp. 1735–1745.

[43] C. R. Menyuk, *Application of multiple-length-scale methods to the study of optical fiber transmission*, J. Engrg. Math., 36 (1999), pp. 113–136.

[44] C. R. Menyuk, B. S. Marks, I. T. Lima, J. Zweck, Y. Sun, and G. M. Carter, *Polarization effects in long-haul undersea systems*, in Undersea Fibre Communication Systems, J. Chesnoy, ed., Elsevier Press, San Diego, CA, 2002, pp. 270–306.

[45] D. A. Nolan, X. Chen, and M. J. Li, *Fibers with low polarization-mode dispersion*, J. Lightwave Technol., 22 (2004), pp. 1066–1077.

[46] C. D. Poole, J. H. Winters, and J. A. Nagel, *Dynamical equation for polarization dispersion*, Opt. Lett., 16 (1991), pp. 372–374.

[47] H. Y. Pua, K. Peddanarappagari, B. Zhu, C. Allen, K. Demarest, and R. Hui, *An adaptive first-order polarization-mode dispersion compensation system aided by polarization scrambling: Theory and demonstration*, J. Lightwave Technol., 18 (2000), pp. 832–841.

[48] S. C. Rashleigh, *Origins and control of polarization effects in single-mode fibers*, J. Lightwave Technol., 1 (1983), pp. 312–331.

[49] J. L. Rebola and A. V. T. Cartaxo, *Q-factor estimation and impact of spontaneous-spontaneous beat noise on the performance of optically preamplified systems with arbitrary optical filtering*, J. Lightwave Technol., 21 (2003), pp. 87–95.

[50] H. Rosenfeldt, C. Knothe, R. Ulrich, E. Brinkmeyer, U. Feiste, C. Schubert, J. Berger, R. Ludwig, H. G. Weber, and A. Ehrhardt, *Automatic PMD compensation at* 40 *Gb/s and* 80 *Gb/s using a* 3*-dimensional DOP evaluation for feedback*, in Proceedings of the Optical Fiber Communication Conference, Anaheim, CA, 2001, paper PD27.

[51] D. Sandel, M. Yoshida-Dierolf, R. Noé, A. Schöpflin, E. Gottwald, and G. Fischer, *Automatic polarization mode dispersion compensation in* 40 *Gb/s optical transmission system*, Electron. Lett., 34 (1998), pp. 2258–2259.

[52] H. Sunnerud, M. Karlsson, C. Xie, and P. A. Andrekson, *Polarization-mode dispersion in high-speed fiber-optic transmission systems*, J. Lightwave Technol., 20 (2002), pp. 2204–2219.

[53] H. Sunnerud, C. Xie, M. Karlsson, R. Samuelsson, and P. A. Andrekson, *A comparison between different PMD compensation techniques*, J. Lightwave Technol., 20 (2002), pp. 368–378.

[54] P. M. Sylla, C. J. K. Richardson, M. VanLeeuwen, M. Saylors, and J. Goldhar, *DOP ellipsoids for systems with frequency-dependent principal states*, IEEE Photon. Technol. Lett., 13 (2001), pp. 1310–1312.

[55] P. R. Trischitta and E. L. Varma, *Jitter in Digital Transmission Systems*, Artech House, Boston, 1989.

[56] E. Veach, *Robust Monte Carlo Methods for Light Transport Simulation*, Ph.D. thesis, Stanford University, Stanford, CA, 1997.

[57] D. Waddy, L. Chen, and X. Bao, *Theoretical and experimental study of the dynamics of polarization-mode dispersion*, IEEE Photon. Technol. Lett., 14 (2002), pp. 468–470.

[58] P. K. A. Wai and C. R. Menyuk, *Polarization mode dispersion, decorrelation, and diffusion in optical fibers with randomly varying birefringence*, J. Lightwave Technol., 14 (1996), pp. 148–157.

[59] P. K. A. Wai, C. R. Menyuk, and H. H. Chen, *Stability of solitons in randomly varying birefringent fibers*, Opt. Lett., 16 (1991), pp. 1231–1233.

[60] D. Yevick, *Multicanonical communication system modeling—application to PMD statistics*, IEEE Photon. Technol. Lett., 14 (2002), pp. 1512–1514.

[61] D. Yevick, *The accuracy of multicanonical system models*, IEEE Photon. Technol. Lett., 15 (2003), pp. 224–226.

[62] J. Zhou and M. J. O'Mahony, *Optical transmission system penalties due to fiber polarization mode dispersion*, IEEE Photon. Technol. Lett., 6 (1994), pp. 1265–1267.

[63] J. Zweck, I. T. Lima, Jr., Y. Sun, A. O. Lima, C. R. Menyuk, and G. M. Carter, *Modeling receivers in optical communication systems with polarization effects*, Opt. Photon. News, 14 (2003), pp. 30–35.

[64] J. Zweck, S. E. Minkoff, A. O. Lima, I. T. Lima, Jr., and C. R. Menyuk, *A comparative study of feedback controller sensitivity to all orders of PMD for a fixed DGD compensator*, in Proceedings of the Optical Fiber Communication Conference, Atlanta, 2003, paper ThY.

SIAM J. OPTIM.
Vol. 17, No. 3, pp. 776–794
© 2006 Society for Industrial and Applied Mathematics

SUFFICIENT SECOND-ORDER OPTIMALITY CONDITIONS FOR AN ELLIPTIC OPTIMAL CONTROL PROBLEM WITH POINTWISE CONTROL-STATE CONSTRAINTS*

A. RÖSCH† AND F. TRÖLTZSCH‡

Abstract. An optimal control problem for a semilinear elliptic equation is investigated, where pointwise constraints are given on control and state. The state constraints are of mixed (bottleneck) type, where associated Lagrange multipliers can be chosen as bounded and measurable functions. Based on this property, a second-order sufficient optimality condition is established that takes into account strongly active constraints.

Key words. optimal control, elliptic differential equation, sufficient second-order optimality condition, pointwise mixed control-state constraints

AMS subject classifications. 49K20, 90C48

DOI. 10.1137/050625850

1. Introduction.

In this paper we consider the optimal control problem to minimize

$$F(y,u) = \int_\Omega f(x,y(x))\,dx + \int_\Gamma g(x,y(x),u(x))\,ds(x) \tag{1.1}$$

subject to the state equations

$$\begin{aligned} Ay + y &= 0 && \text{in } \Omega, \\ \partial_{n_A} y &= b(x,y,u) && \text{on } \Gamma, \end{aligned} \tag{1.2}$$

the control constraints

$$0 \le u(x) \quad \text{for } x \in \Gamma, \tag{1.3}$$

and to the mixed control-state constraints

$$c(x) \le u(x) + \gamma(x)y(x) \quad \text{for } x \in \Gamma. \tag{1.4}$$

The main task of our paper is to establish second-order sufficient optimality conditions that are close to the associated necessary ones. For control-constrained problems, this issue was discussed quite completely in literature for semilinear elliptic and parabolic equations. Specifically, we mention Bonnans [4], Casas, Tröltzsch, and Unger [9], Goldberg and Tröltzsch [12], and Heinkenschloss and Tröltzsch [13].

The main difficulty in our problem is the presence of the pointwise control-state constraint $c(x) \le u(x) + \gamma(x)y(x)$ in (1.4). If pointwise state constraints are given, then the theory of sufficient second-order conditions is faced with specific difficulties

*Received by the editors March 3, 2005; accepted for publication (in revised form) April 14, 2006; published electronically October 3, 2006.
http://www.siam.org/journals/siopt/17-3/62585.html
†Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Altenbergerstraße 69, A-4040 Linz, Austria (arnd.roesch@oeaw.ac.at).
‡Technische Universität Berlin, Fakultät II – Mathematik und Naturwissenschaften, Str. des 17. Juni 136, D-10623 Berlin, Germany (troeltzsch@math.tu-berlin.de).

that are still far from being solved. In particular, these problems arise for pointwise state constraints of the type $c(x) \le y(x)$. Here, the Lagrange multipliers associated with the state constraints are Borel measures so that the associated adjoint state exhibits low regularity; cf. Casas [5], [6] or Alibert and Raymond [1]. This fact causes specific difficulties in the discussion for second-order sufficient optimality conditions. We refer to Casas, Tröltzsch, and Unger [10] and Raymond and Tröltzsch [15] or to Casas and Mateos [7], who consider the case of finitely many state constraints.

In our problem (1.1)–(1.4), the situation is slightly simpler, since the constraint (1.4) is a *mixed* control-state constraint of bottleneck type. In the associated parabolic case, the Lagrange multipliers are more regular. They can be assumed to be bounded and measurable functions; see Bergounioux and Tröltzsch [3] and Arada and Raymond [2]. The existence of bounded and measurable Lagrange multipliers for linear-quadratic elliptic optimal control problems is proved in Tröltzsch [21]. The semilinear elliptic case is investigated in Rösch and Tröltzsch [16].

Higher regularity of the multipliers is the main advantage enabling us to establish second-order conditions. The second-order conditions should require minimal assumptions, i.e., they should be as close as possible to associated necessary conditions. Usually, this task is accomplished by considering strongly active sets (see [11] for control-constrained optimal control of ordinary differential equations). Here, we apply this technique to our case of mixed constraints. The analysis shows that this is not an easy task. It indicates that pointwise state constraints of more general type will give rise to even more difficult techniques.

Our paper extends the results of [17], [18], where second-order conditions are derived for a weakly singular integral state equation and for parabolic equations, respectively. Let us shortly sketch the main difference between these papers and our new discussion: In [17], [18] the proof of sufficiency is based on the nonnegativity of some inverse operators related to the Fréchet derivative of the control-to-state mapping. It is this nonnegativity that cannot, in general, be expected for elliptic problems. Therefore, here we abstain from such an assumption. We only require a solvability property of an auxiliary elliptic problem. Based on similarly weak assumptions, also the regularity of Lagrange multipliers has been shown in [16], [21].

Moreover, in our paper the definition of strongly active sets associated with the mixed constraints is a more natural way than the one in [17], [18].

Let us remark that the inequality constraints in our problem differ from the inequality constraints considered in [17], [18], where $u \le c + \gamma y$ is investigated instead of (1.4). It would be easy to adapt our theory to the inequality constraints in [17], [18]. However, even for parabolic problems, the methods in [17], [18] cannot be applied to the inequality constraints (1.3) and (1.4) since certain inverse operators do not preserve the nonnegativity.

The paper is organized as follows: In section 2 we formulate first- and second-order optimality conditions and state the main result. Section 3 contains auxiliary results. The proof that our second-order conditions are sufficient for local optimality is presented in section 4.

In the paper, we use the following notations: By $b'(x,y,u)$ and $b''(x,y,u)$ we denote the gradient and the Hessian matrix of b with respect to (y,u):

$$b'(x,y,u) = \begin{pmatrix} b_y(x,y,u) \\ b_u(x,y,u) \end{pmatrix}, \quad b''(x,y,u) = \begin{pmatrix} b_{yy}(x,y,u) & b_{yu}(x,y,u) \\ b_{yy}(x,y,u) & b_{uu}(x,y,u) \end{pmatrix}.$$

Here, the notations $b_y(x,y,u) = D_y b(x,y,u)$, $b_{yy}(x,y,u) = D_{yy} b(x,y,u)$, etc. are

used for the partial derivatives. The norms $|b'|$, $|b''|$ are defined by adding the absolute values of all entries of b' and b'', respectively. By ∂_{n_A} we denote the conormal derivative.

We adapt the following assumptions from [9]:

(A1) For each $x \in \Omega$ or Γ, respectively, the functions $f = f(x,y)$, $g = g(x,y,u)$, and $b = b(x,y,u)$ are of class C^2 with respect to (y,u). For fixed (y,u) they are Lebesgue measurable with respect to $x \in \Omega$ or $x \in \Gamma$, respectively.

(A2) In this assumption, $p > N-1$, s, and r denote fixed parameters that depend on the dimension N of the domain Ω. The constants s and r express the regularities $y|_\Gamma \in L^s(\Gamma)$ and $y \in L^r(\Omega)$ in the linearized system associated to (1.2). As usual, r' and s' denote mutually conjugate numbers. For instance, s' is defined by $1/s' + 1/s = 1$.
For all $M > 0$, there are constants $C_M > 0$, functions $\psi_f^M \in L^{(r/2)'}(\Omega)$, $\psi_f^{M,1} \in L^{(s/2)'}(\Gamma)$, $\psi_f^{M,2} \in L^{(s/2)'}(\Gamma)$, $\psi_f^{M,3} \in L^\infty(\Gamma)$, and a continuous, monotone increasing function $\eta \in C(\mathbb{R}^+ \cup \{0\})$ with $\eta(0) = 0$ such that

(i) $b(.,0,0) \in L^p(\Gamma)$, for some $p > N-1$,

$$b_y(x,y,u) \le 0 \qquad \text{for a.e. } x \in \Gamma \quad \text{and} \quad \forall (y,u) \in \mathbb{R}^2, \tag{1.5}$$

$$|b'(x,y,u)| + |b''(x,y,u)| \le C_M,$$
$$|b''(x,y_1,u_1) - b''(x,y_2,u_2)| \le C_M \eta(|y_1 - y_2| + |u_1 - u_2|)$$

for almost all $x \in \Gamma$ and all $|y|, |u|, |y_i|, |u_i| \le M$, $i = 1,2$.

(ii) $f(.,0) \in L^1(\Omega)$, $f_y(.,0) \in L^{r'}(\Omega)$, $f_{yy}(.,0) \in L^{(r/2)'}(\Omega)$, and

$$|f_{yy}(x,y_1) - f_{yy}(x,y_2)| \le \psi_f^M(x)\eta(|y_1 - y_2|)$$

for almost all $x \in \Omega$ and all $|y_i| \le M$, $i = 1,2$.

(iii) $g(.,0,0) \in L^1(\Gamma)$, $g_y(.,0,0) \in L^{s'}(\Gamma)$, $g_u \in L^2(\Gamma)$, $g_{yy}(.,0,0) \in L^{(s/2)'}(\Gamma)$, $g_{yu}(.,0,0) \in L^{2(s/2)'}(\Gamma)$, $g_{uu}(.,0,0) \in L^\infty(\Gamma)$, and

$$|g_{yy}(x,u_1,y_1) - g_{yy}(x,u_2,y_2)| \le \psi_f^{M,1}(x)\eta(|y_1 - y_2| + |u_1 - u_2|)$$
$$|g_{yu}(x,u_1,y_1) - g_{yu}(x,u_2,y_2)| \le \psi_f^{M,2}(x)\eta(|y_1 - y_2| + |u_1 - u_2|)$$
$$|g_{yu}(x,u_1,y_1) - g_{yu}(x,u_2,y_2)| \le \psi_f^{M,3}(x)\eta(|y_1 - y_2| + |u_1 - u_2|)$$

for almost all $x \in \Omega$ and all $|y_i| \le M$, $|u_i| \le M$, $i = 1,2$.

Other estimates of b, f, g and their first derivatives can be derived from (A1), (A2) by the mean value theorem.

(A3) We assume that $c, \gamma \in C(\Gamma)$, and $\gamma(x) \ge 0 \quad \forall x \in \Gamma$.

(A4) The domain $\Omega \subset \mathbb{R}^N$ is bounded and has a Lipschitz boundary Γ. The Lebesgue surface measure induced on Γ is denoted by $ds(x)$. The elliptic operator A is defined by

$$Ay(x) = -\sum_{i,j=1}^m D_i(a_{ij}(x)D_j y(x)),$$

where $a_{ij} \in L^\infty(\Omega)$ satisfy, for some positive m_0, the condition of uniform ellipticity

$$\sum_{i,j=1}^m a_{ij}(x)\xi_i\xi_j \ge m_0|\xi|^2.$$

2. First- and second-order optimality conditions. We are looking for a control in the space $U = L^\infty(\Gamma)$, while the state is defined as a weak solution of (1.2) in the state space $Y = C(\bar\Omega) \cap H^1(\Omega)$ by

$$\int_\Omega \Big(\sum_{i,j=1}^m a_{ij} D_j y D_i v + yv \Big)\, dx = \int_\Gamma b(.,y,u)v \; ds(x) \qquad \forall v \in H^1(\Omega). \tag{2.1}$$

We endow Y with the norm $\|y\|_Y = \|y\|_{C(\bar\Omega)} + \|y\|_{H^1(\Omega)}$. It can be shown that, for each $u \in L^\infty(\Gamma)$, the elliptic equation (1.2) admits a unique weak solution $y = y(u) \in Y$, see [8]. Moreover, Casas and Tröltzsch [8] have proved that the solution mapping $G:\ u \mapsto y$ from $L^\infty(\Gamma)$ into Y is of class C^2.

In this paper, we discuss sufficient conditions for a local minimum. Therefore, we investigate a candidate $\bar u$ for the local optimum and an ε-neighborhood of $\bar u$:

$$B_\varepsilon(\bar u) = \{u \in L^\infty(\Gamma) :\ \|u - \bar u\|_{L^\infty(\Gamma)} < \varepsilon\}.$$

For any fixed $\varepsilon > 0$ and arbitrary $\bar u \in U$, there exists a constant $M = M(\varepsilon)$ such that

$$\|y(u)\|_Y \le M \qquad \forall y \in B_\varepsilon(\bar u).$$

The boundary values of y are of particular importance for us. Thus we define the mapping $S : L^\infty(\Gamma) \to C(\bar\Gamma)$ with $S = \tau G$ that assigns to u the boundary values of y. Here, τ denotes the trace operator. Clearly, the Fréchet differentiability of the operator G implies the differentiability of S. The application of $S'(\bar u)$ to an element $h \in U$ is given by the boundary values of the solution z of the elliptic problem

$$\begin{aligned} Az + z &= 0 && \text{in } \Omega, \\ \partial_{n_A} z - \bar b_y z &= \bar b_u h && \text{on } \Gamma, \end{aligned} \tag{2.2}$$

i.e., $S'(\bar u)h = z|_\Gamma$. Here we have used the abbreviations $\bar b_y = b_y(x, \bar y(x), \bar u(x))$ and $\bar b_u = b_u(x, \bar y(x), \bar u(x))$. The operator $S'(\bar u)$ is extended to a linear continuous operator in $\mathcal{L}(L^2(\Gamma))$. From now on, $S'(\bar u)$ will be understood in this way. For the remainder term in the first-order Taylor expansion of $y(\bar u + h)$, we obtain the property

$$\frac{\|y(\bar u + h) - y(\bar u) - z(\bar u, h)\|_{L^2(\Gamma)}}{\|h\|_{L^2(\Gamma)}} \to 0 \quad \text{as } \|h\|_{L^\infty(\Gamma)} \to 0$$

using a known result of Maurer [14].

Next, we introduce the L^2-adjoint operator $S'(\bar u)^* \in \mathcal{L}(L^2(\Gamma))$. This operator is given by $S'(\bar u)^* \mu = \varphi|_\Gamma$, where φ is the solution of the elliptic problem

$$\begin{aligned} A^*\varphi + \varphi &= 0 && \text{in } \Omega, \\ \partial_{n_{A^*}} \varphi - \bar b_y\ \varphi &= \bar b_u\ \mu && \text{on } \Gamma, \end{aligned} \tag{2.3}$$

and A^* is the formal adjoint operator to A. In all that follows, let $(\bar y, \bar u)$ be a candidate for a local solution of (1.1)–(1.4). Let us set up the associated first-order necessary optimality conditions in form of a Karush–Kuhn–Tucker type theorem. To this aim, we introduce the Lagrange functional $L : Y \times L^\infty(\Gamma) \times Y \times L^\infty(\Gamma)^2 \to \mathbb{R}$,

$$\begin{aligned} L(y,u,p,\mu_1,\mu_2) = F(y,u) &+ \int_\Omega \Big(\sum_{i,j=1}^m a_{ij} D_j y D_i p + yp \Big)\, dx - \int_\Gamma b\, p \; ds(x) \\ &- \int_\Gamma \mu_1 u \; ds(x) - \int_\Gamma (u + \gamma y - c)\mu_2 \; ds(x). \end{aligned}$$

Let us comment on this choice for L. The elliptic equation (1.2) is considered in Y, while the inequality constraints (1.3) are posed in $L^\infty(\Gamma)$. Knowing the general Karush–Kuhn–Tucker theory in Banach spaces, one expects associated Lagrange multipliers $p \in Y^*$ and $\mu_i \in (L^\infty(\Gamma))^*$, together with a related quite complicated Lagrange functional. However, special techniques for optimal control problems of bottleneck type allow to show that, under natural assumptions, the Lagrange multipliers can be expressed by regular functions, i.e., $p \in Y$ and $\mu_i \in L^\infty(\Gamma)$; we refer to Tröltzsch [21] and Rösch and Tröltzsch [16]. This well-known advantage of bottleneck type problems is our key idea to establish special second-order sufficient optimality conditions, which can hardly be expected for $\mu_i \in (L^\infty(\Gamma))^*$. The existence of such regular multipliers can be shown under a Slater type condition and the assumption $\gamma(x) \geq 0$. Here, the nonnegativity of γ plays a crucial role.

Therefore, we are justified to *assume* that an adjoint state $\bar{p} \in Y$ and Lagrange multipliers $\bar{\mu}_i \in L^\infty(\Gamma)$ exist such that $(\bar{y}, \bar{u}, \bar{p}, \bar{\mu}_1, \bar{\mu}_2)$ satisfies the following first-order necessary optimality system (FON):

$$(FON)\quad \begin{cases} D_y L(\bar{y}, \bar{u}, \bar{p}, \bar{\mu}_1, \bar{\mu}_2) &= 0 \\ D_u L(\bar{y}, \bar{u}, \bar{p}, \bar{\mu}_1, \bar{\mu}_2) &= 0 \\ \text{and for almost all } x \in \Gamma & \\ \bar{\mu}_1(x) &\geq 0 \\ \bar{\mu}_2(x) &\geq 0 \\ \bar{u}(x)\bar{\mu}_1(x) &= 0 \\ (\bar{u}(x) + \gamma(x)\bar{y}(x) - c(x))\bar{\mu}_2(x) &= 0. \end{cases}$$

Note that the Lagrange multipliers may not be unique. The last two conditions of (FON) are the well-known *complementary slackness conditions*. They imply $\bar{\mu}_1(x) > 0 \;\Rightarrow \bar{u}(x) = 0$ and $\bar{\mu}_2(x) > 0 \;\Rightarrow c(x) = \bar{u}(x) + \gamma(x)\bar{y}(x)$. Let us express these optimality conditions also in terms of the partial differential equation. As it is well known, the first equation of (FON) is equivalent to the adjoint equation

$$\begin{aligned} A^*\bar{p} + \bar{p} &= f_y(x, \bar{y}) && \text{in } \Omega, \\ \partial_{n_{A^*}}\bar{p} - b_y(x, \bar{y}, \bar{u})\bar{p} &= g_y(x, \bar{y}, \bar{u}) - \gamma\bar{\mu}_2 && \text{on } \Gamma. \end{aligned} \tag{2.4}$$

The second equation of (FON) is equivalent to

$$g_u(x, \bar{y}, \bar{u}) + b_u(x, \bar{y}, \bar{u})\bar{p} - \bar{\mu}_1 - \bar{\mu}_2 = 0. \tag{2.5}$$

Next, we discuss a sufficient second-order optimality condition (SSC). For this purpose, following Dontchev et al. [11], we define *strongly active sets* and the associated *critical subspace*. Assume that $(\bar{y}, \bar{u}, \bar{p}, \bar{\mu}_1, \bar{\mu}_2)$ fulfils (FON).

DEFINITION 2.1. *Let $\delta_1, \delta_2 > 0$ be real numbers and $\bar{\mu}_1, \bar{\mu}_2 \in L^\infty(\Gamma)$ be Lagrange multipliers introduced in (FON). The sets*

$$A_1(\delta_1) := \{x \in \Gamma : \bar{\mu}_1(x) \geq \delta_1\}, \tag{2.6}$$

$$A_2(\delta_2) := \{x \in \Gamma \setminus A_1(\delta_1) : \bar{\mu}_2(x) \geq \delta_2\} \tag{2.7}$$

are called strongly active sets.

All further arguments hold true for an arbitrary choice of δ_1 and δ_2. Later, these numbers will be chosen such that a second-order sufficient optimality condition is

satisfied. To shorten the notation, we will drop the dependence of the active sets on these parameters in the proofs, but we will use the detailed notation for the statements of the main results.

DEFINITION 2.2. *We say that* $(y,u) \in C(\bar\Omega) \times L^\infty(\Gamma)$ *belongs to the critical subspace, if*

$$u = 0 \qquad \text{on } A_1(\delta_1), \tag{2.8}$$

$$u + \gamma y|_\Gamma = 0 \qquad \text{on } A_2(\delta_2), \tag{2.9}$$

and

$$\begin{aligned} Ay + y &= 0 && \text{in } \Omega, \\ \partial_{n_A} y - \bar b_y\ y &= \bar b_u\ u && \text{on } \Gamma. \end{aligned} \tag{2.10}$$

Notice that (2.10) implies $y|_\Gamma = S'(\bar u)u$. This critical subspace is larger than really needed. A smaller critical convex cone is discussed at the end of the paper.

Before we formulate the second-order sufficient optimality condition, let us find the explicit expression of $L''_{(u,y)}(\bar y,\bar u,\bar p,\bar\mu_1,\bar\mu_2)[h_y,h_u]^2$:

$$\begin{aligned} L''_{(u,y)}(\bar y,\bar u,\bar p,\bar\mu_1,\bar\mu_2)[h_y,h_u]^2 &= \int_\Omega f_{yy}h_y^2\,dx + \int_\Gamma (g_{yy}h_y^2 + 2g_{yu}h_yh_u + g_{uu}h_u^2)\,ds(x) \\ &\quad + \int_\Gamma (\bar b_{yy}h_y^2 + 2\bar b_{yu}h_yh_u + \bar b_{uu}h_u^2)\bar p\,ds(x). \end{aligned} \tag{2.11}$$

Here, $h_y \in C(\bar\Omega)$, $h_u \in L^\infty(\Gamma)$ denote arbitrary increments of y and u, respectively. Now we state the second-order sufficient optimality condition.

(SSC): There exist positive numbers δ, δ_1, δ_2 such that the definiteness condition

$$L''_{(u,y)}(\bar y,\bar u,\bar p,\bar\mu_1,\bar\mu_2)[h_y,h_u]^2 \ge \delta\|h_u\|^2_{L^2(\Gamma)} \tag{2.12}$$

holds true for all (h_y,h_u) belonging to the critical subspace defined upon δ_1, δ_2.

In our further analysis, the boundary value problem

$$\begin{aligned} Av + v &= 0 && \text{in } \Omega, \\ \partial_{n_A} v + (-\bar b_y + \chi_{A_2(\delta_2)}\bar b_u\gamma)v &= \phi && \text{on } \Gamma \end{aligned} \tag{2.13}$$

plays a basic role. We require the following regularity assumption:

(R) *For* $\phi = 0$*, the problem* (2.13) *has only the trivial solution* $v = 0$.

For instance, this assumption is fulfilled if

$$-\bar b_y + \gamma\chi_{A_2(\delta_2)}\bar b_u \ge 0 \qquad \text{a.e. on } \Gamma. \tag{2.14}$$

Here χ_{A_2} denotes the characteristic function of the set $A_2(\delta_2)$. Thanks to (1.5) and (A3), this condition is fulfilled if $\bar b_u \ge 0$ holds. Now, we state the main result of the paper.

THEOREM 2.3 (second-order sufficiency). *Assume that* $(\bar y,\bar u,\bar p,\bar\mu_1,\bar\mu_2)$ *fulfils the first-order optimality system (FON) and the regularity condition (R) holds. If the second-order condition (SSC) is satisfied, then there exist* $\delta_s > 0$ *and* $\varepsilon > 0$ *such that the quadratic growth condition*

$$F(y,u) - F(\bar y,\bar u) \ge \delta_s\|u-\bar u\|^2_{L^2(\Gamma)} \tag{2.15}$$

holds for all admissible pairs (y,u) *with* $\|u-\bar u\|_{L^\infty(\Gamma)} < \varepsilon$*. Therefore,* $\bar u$ *is a locally optimal control in the norm of* $L^\infty(\Gamma)$.

The proof is carried out in section 4.

3. Auxiliary results.

LEMMA 3.1. *Let $\beta \in L^\infty(\Gamma)$ be a fixed function that is almost everywhere nonnegative. Then for all $\phi \in L^p(\Gamma)$ with $p > N-1$, the weak solution $v \in H^1(\Omega)$ of*

$$\begin{aligned} Av + v &= 0 && \text{in } \Omega, \\ \partial_{n_A} v + \beta(\cdot)\, v &= \phi && \text{on } \Gamma \end{aligned} \tag{3.1}$$

belongs to $C(\bar\Omega)$ and satisfies the estimate

$$\|v\|_{C(\bar\Omega)} \le c_p \, \|\phi\|_{L^p(\Gamma)} \tag{3.2}$$

with a positive constant c_p that does not depend on ϕ.

For this classical result, we refer to [6] and the arguments in [1] concerning the case of Lipschitz domains.

For $N = 2$ the trace of $v \in H^1(\Omega)$ belongs to $L^r(\Gamma)$, with any $r < \infty$. For $N > 2$, $v|_\Gamma$ belongs to $L^{\frac{2(N-1)}{N-2}}(\Gamma)$. It implies that $v|_\Gamma \in L^{2+s}(\Gamma)$ with $s = 2/(N-2) > 0$ (arbitrary $s > 0$ for $N = 2$) so that the mapping $\phi \mapsto v|_\Gamma$ is continuous from $L^2(\Gamma)$ to $L^{2+s}(\Gamma)$ and from $L^p(\Gamma)$ to $L^\infty(\Gamma)$, in particular also to $L^{p+s}(\Gamma)$. By classical interpolation, cf. Triebel [20, 1.18.7, Thm. 1], there exists a positive constant δ (independent of s), such that the mapping $\phi \mapsto v|_\Gamma$ satisfies

$$\|v\|_{L^{s+\delta}(\Gamma)} \le c_s \, \|\phi\|_{L^s(\Gamma)} \qquad \forall\, s \ge 2. \tag{3.3}$$

provided that $\beta(\cdot) \ge 0$. Here and in what follows, we write $\|v\|_{L^2(\Gamma)}$ rather than $\|v|_\Gamma\|_{L^2(\Gamma)}$. We can dispense with this sign condition on β, if a regularity condition is fulfilled. To show that, we consider (3.1) for an arbitrary $\beta \in L^\infty(\Gamma)$ and assume that the associated homogeneous equation (3.1) has only the trivial solution. Then the mapping $S_\beta : L^2(\Gamma) \to L^2(\Gamma)$ that assigns to ϕ the trace of the solution v of (3.1) is well defined and continuous.

To verify this, we consider also the shifted equation

$$\begin{aligned} Av + v &= 0 && \text{in } \Omega, \\ \partial_{n_A} v + (\|\beta\|_{L^\infty(\Gamma)} + \beta(\cdot))\, v &= \phi && \text{on } \Gamma. \end{aligned} \tag{3.4}$$

Clearly, the associated mapping $\tilde S_\beta : L^2(\Gamma) \to L^2(\Gamma)$, $\tilde S_\beta : \phi \mapsto v|_\Gamma$, is well defined and compact. By the Fredholm theory, it has only countably many eigenvalues. A number $\lambda \in \mathbb{R}$ is an eigenvalue of $\tilde S_\beta$, if $\tilde S_\beta v = \lambda v$ holds with a nontrivial v. This means that λv solves (3.4) with boundary data $\phi = v$, i.e., after dividing by λ, if the boundary condition

$$\partial_{n_A} v + (\|\beta\|_{L^\infty(\Gamma)} + \beta(\cdot))\, v = \lambda^{-1} v \tag{3.5}$$

is satisfied with some nontrivial v. Obviously, we have a one-to-one correspondence between the eigenvalues of $\tilde S_\beta$ and those of S_β. The boundary condition (3.5) holds for nontrivial v iff the condition $\partial_{n_A} v + \beta(\cdot)\, v = (\lambda^{-1} - \|\beta\|_{L^\infty(\Gamma)})\, v$ is fulfilled so that $1/(\lambda^{-1} - \|\beta\|_{L^\infty(\Gamma)})$ is an eigenvalue of S_β.

In view of this, the assumption on the homogeneous equation (3.1) implies that (3.1) is uniquely solvable for all $\phi \in L^2(\Gamma)$ and that S_β is continuous in $L^2(\Gamma)$.

LEMMA 3.2. *Assume that the homogeneous equation* (3.1) *has only the trivial solution and that $\phi \in L^\infty(\Gamma)$ is given arbitrarily. Let $v \in H^1(\Omega)$ be the solution of*

(3.1). *Then there exists a constant c_β not depending on ϕ such that the following estimates hold true:*

$$\begin{array}{rcll} \|v\|_{L^2(\Gamma)} & \le & c_\beta \|\phi\|_{L^2(\Gamma)}, & \\ \|v\|_{C(\Gamma)} & \le & c_\beta \|\phi\|_{L^p(\Gamma)} & \forall\, p > N-1, \\ \|v\|_{L^1(\Gamma)} & \le & c_\beta \|\phi\|_{L^1(\Gamma)}. & \end{array} \tag{3.6}$$

Proof. (i) The first estimate is a simple consequence of the continuity of S_β in $L^2(\Gamma)$. It is only stated for convenience.

(ii) The second inequality follows by bootstrapping: The solution v solves $Av+v = 0$ subject to the boundary condition

$$\partial_{n_A} v = \phi - \beta(\cdot)\, v.$$

By (3.3) and the first estimate, we find with some $s > 0$ and some generic constant c that

$$\begin{array}{rcl} \|v\|_{L^{2+s}(\Gamma)} & \le & c\,(\|\phi\|_{L^2(\Gamma)} + \|\beta\|_{L^\infty(\Gamma)} \|v\|_{L^2(\Gamma)}) \\ & \le & c\,\|\phi\|_{L^2(\Gamma)} \le c\,\|\phi\|_{L^{2+s}(\Gamma)}. \end{array}$$

Repeating this estimate, we get from the one in $L^{2+s}(\Gamma)$ that

$$\begin{array}{rcl} \|v\|_{L^{2+2\,s}(\Gamma)} & \le & c\,(\|\phi\|_{L^{2+s}(\Gamma)} + \|\beta\|_{L^\infty(\Gamma)} \|v\|_{L^{2+s}(\Gamma)}) \\ & \le & c\,\|\phi\|_{L^{2+\,s}(\Gamma)} \le c\,\|\phi\|_{L^{2+2\,s}(\Gamma)}. \end{array}$$

After finitely many steps, the estimate

$$\|v\|_{L^p(\Gamma)} \le c\,\|\phi\|_{L^p(\Gamma)} \tag{3.7}$$

can be derived for some $p > N-1$. In view of (3.2), boundary data from $L^p(\Gamma)$ are transformed to continuous solutions for $p > N-1$. Therefore, by (3.3), it follows that

$$\begin{array}{rcl} \|v\|_{C(\Gamma)} & \le & c_p\,(\|\phi\|_{L^p(\Gamma)} + \|\beta\|_{L^\infty(\Gamma)} \|v\|_{L^p(\Gamma)}) \\ & \le & c_p\,(\|\phi\|_{L^p(\Gamma)} + \|\beta\|_{L^\infty(\Gamma)} c_p\,\|\phi\|_{L^p(\Gamma)}) \le c\,\|\phi\|_{L^p(\Gamma)}. \end{array}$$

(iii) To show the last estimate, we proceed by duality. The operator S_β is self-adjoint. Moreover, roughly speaking, we have by (ii) that its restriction $S_{\beta,p}$ to $L^p(\Gamma)$ is continuous from $L^p(\Gamma)$ to $C(\Gamma)$. We can assume $p \ge 2$. The adjoint operator $S^*_{\beta,p}$ is continuous from $C(\Gamma)^*$ to $L^{p'}(\Gamma)$, where p' is conjugate to p. Therefore, it is in particular continuous in $L^1(\Gamma)$. Finally, it can be shown that $S^*_{\beta,p}\,\phi = S^*_\beta\,\phi = S_\beta\,\phi$ for all $\phi \in L^2(\Gamma)$. This shows that S_β is continuous in $L^1(\Gamma)$ so that the third estimate is true. These facts are explained more precise and slightly more detailed in [16]. ∎

We should remark that the third estimate is not surprising. If $\beta \ge 0$, the estimate follows from the results by Casas [5] and Alibert and Raymond [1]. They have shown in this case that the boundary value problem (3.1) with given regular Borel measure ϕ admits a unique solution $v \in W^{1,\sigma}(\Omega)$ for all $\sigma < N/(N-1)$. Clearly, this implies the L^1-estimate. More or less, the result for arbitrary β is a natural extension. We have presented these details for the convenience of the reader.

As a corollary of the preceding lemma, we obtain for $\beta := -\bar b_y + \gamma \chi_{A_2} \bar b_u$ the following result.

LEMMA 3.3. *Suppose that the regularity condition (R) is satisfied. Then, for all $\phi \in L^2(\Gamma)$, the boundary value problem*

$$\begin{array}{rll} Av + v = 0 & & \text{in } \Omega, \\ \partial_{n_A} v + (-\bar b_y + \chi_{A_2} \bar b_u \gamma) v = \phi & & \text{in } \Gamma \end{array} \tag{3.8}$$

has a unique solution $v \in H^1(\Omega)$. Moreover, the estimate

$$\|v\|_{L^1(\Gamma)} \le c_1 \|\phi\|_{L^1(\Gamma)} \tag{3.9}$$

is fulfilled with some constant c_1 that does not depend on ϕ.

To perform our analysis, we repeatedly need controls u defined as follows:

$$u(x) = \begin{cases} \phi(x) & \text{on } \Gamma \setminus A_2, \\ \phi(x) - \gamma(x)(S'(\bar u)u)(x) & \text{on } A_2. \end{cases} \tag{3.10}$$

The next lemma shows that this setting is correct.

LEMMA 3.4. *Assume that the regularity condition (R) is fulfilled. Then, there is exactly one function $u \in L^\infty(\Gamma)$ that satisfies condition* (3.10)*. Moreover, the estimates*

$$\|u\|_{L^1(\Gamma)} \le c_1 \|\phi\|_{L^1(\Gamma)}, \tag{3.11}$$

$$\|u\|_{L^2(\Gamma)} \le c_2 \|\phi\|_{L^2(\Gamma)}, \tag{3.12}$$

$$\|u\|_{L^\infty(\Gamma)} \le c_\infty \|\phi\|_{L^\infty(\Gamma)} \tag{3.13}$$

hold with certain constants c_1, c_2, c_∞ that do not depend on ϕ.

Proof. Suppose that $u \in L^\infty(\Gamma)$ satisfies (3.10). Put $v := G'(\bar u)u$. Then v satisfies the elliptic problem with the boundary condition

$$\partial_{n_A} v - \bar b_y v = \begin{cases} \bar b_u \phi & \text{on } \Gamma \setminus A_2, \\ \bar b_u(\phi - \gamma v) & \text{on } A_2, \end{cases} \tag{3.14}$$

that is

$$\partial_{n_A} v + (-\bar b_y + \chi_{A_2} \bar b_u \gamma) v = \bar b_u \phi \quad \text{on } \Gamma. \tag{3.15}$$

This is exactly the boundary condition of (3.8). Consequently, the solution v is unique. Therefore, if u satisfies (3.10), then $v = G'(\bar u)u$ is unique, hence u is unique, because of

$$u = \begin{cases} \phi & \text{on } \Gamma \setminus A_2, \\ \phi - \gamma v|_\Gamma & \text{on } A_2. \end{cases} \tag{3.16}$$

On the other hand, starting from ϕ, the solution v of the elliptic equation with the boundary condition (3.15) is well defined, and the function u given by (3.16) satisfies (3.10), since, by definition of v, $u = S'(\bar u)v|_\Gamma$.

The estimate (3.11) is obtained by Lemma 3.3. Estimate (3.12) follows by standard arguments. The Stampacchia method [19] delivers estimate (3.13). ∎

To prove the main result, we later have to compare the reference pair $(\bar y, \bar u)$ with another admissible pair (y,u), where $y = G(u)$. In this case, we estimate the difference

$$y|_\Gamma - \bar y|_\Gamma = S(u) - S(\bar u) = S'(\bar u)(u - \bar u) + r_1(\bar u, u - \bar u), \tag{3.17}$$

where r_1 stands for the associated first-order remainder term of S. In the following, if there is no risk of notational confusion, we denote for short the remainder $r_1(\bar u, u - \bar u)$ and the derivative $S'(\bar u)$ by r_1 and S', respectively.

Before continuing our analysis of second-order sufficiency, let us discuss the main difficulties and our main ideas to resolve them. We start without the pointwise control-state constraints. On A_1, we have $\bar{u}(x) \equiv 0$, hence $u - \bar{u} \geq 0$ on A_1. The associated term in the Lagrange functional can be estimated as

$$\int_{A_1} \bar{\mu}_1(u - \bar{u})\, ds(x) \geq \int_{A_1} \delta_1(u - \bar{u})\, ds(x) = \delta_1 \|u - \bar{u}\|_{L^1(A_1)}. \tag{3.18}$$

In the proof of the sufficiency theorem, the L^1-norms on the right-hand side will compensate for the lack of coercivity, since (2.12) does not contribute to definiteness on $A_1 \cup A_2$.

However, we cannot expect such a property for the mixed control-state constraints. It can happen that $\int_{A_2} \bar{\mu}_2(u+y-\bar{u}-\bar{y})\, ds(x) = 0$ although $\|u-\bar{u}\|_{L^1(A_2)} > 0$ holds simultaneously.

To overcome this difficulty, we represent u in the form $u = u_1 + u_2$, where the component u_1 is chosen in such a way that an estimate similar to (3.18) holds. The u_2-part stands for the additional margin of freedom that is caused by subtracting the values of u and $\bar{u}$ outside of A_2. This splitting is performed by

$$\begin{array}{llll} u_1 = \bar{u}, & & u_2 = u - \bar{u} & \text{on } \Gamma \setminus A_2, \\ u_2 = -\gamma(S'u_2 + r_1), & & u_1 = u - u_2 & \text{on } A_2. \end{array} \tag{3.19}$$

The functions u_1 and u_2 are well defined. To see this, we write u_2 in the form

$$u_2 = \begin{cases} \phi & \text{on } \Gamma \setminus A_2, \\ \phi - \gamma S'u_2 & \text{on } A_2, \end{cases}$$

where $\phi = u - \bar{u}$ on $\Gamma \setminus A_2$, $\phi = \gamma r_1$ on A_2. Then u_2 is well defined by Lemma 3.4. Note that $S'(\bar{u})u_2 = S'(\bar{u})(\chi_{\Gamma\setminus A_2}(u - \bar{u}) + \chi_{A_2}u_2)$. From (3.13) and (3.19) we easily get

$$\|u_2\|_{L^\infty(\Gamma)} \leq c_3(\|u - \bar{u}\|_{L^\infty(\Gamma)} + \|r_1\|_{L^\infty(\Gamma)}).$$

The Fréchet differentiability of S in $L^\infty(\Gamma)$ implies

$$\|r_1\|_{L^\infty(\Gamma)} \leq \|u - \bar{u}\|_{L^\infty(\Gamma)}$$

for sufficiently small $\|u - \bar{u}\|_{L^\infty(\Gamma)}$.

Therefore, it holds by $u_1 = u - u_2$ that

$$\begin{aligned} \|u_1 - \bar{u}\|_{L^\infty(A_2)} &\leq \|u - \bar{u}\|_{L^\infty(A_2)} + \|u_2\|_{L^\infty(A_2)} \\ &\leq c_4\|u - \bar{u}\|_{L^\infty(\Gamma)}. \end{aligned} \tag{3.20}$$

LEMMA 3.5. *Assume that $(\bar{y}, \bar{u}, \bar{p}, \bar{\mu}_1, \bar{\mu}_2)$ fulfills the first-order optimality system (FON) and the regularity condition (R) holds. Then there exists a positive constant c_A such that, for all $\varepsilon > 0$ the estimates*

$$\int_\Gamma (u - \bar{u})\bar{\mu}_1\, ds(x) \geq \frac{\delta_1}{\varepsilon}\|u - \bar{u}\|^2_{L^2(A_1(\delta_1))}, \tag{3.21}$$

$$\int_\Gamma (u - \bar{u} + \gamma(y - \bar{y}))\bar{\mu}_2\, ds(x) \geq c_A \cdot \frac{\delta_2}{\varepsilon}\|u_1 - \bar{u}\|^2_{L^2(A_2(\delta_2))} \tag{3.22}$$

are valid for all admissible pairs (u, y) satisfying $\|u - \bar{u}\|_{L^\infty(\Gamma)} < \varepsilon$.

Proof. (i) Because of (FON), $\bar\mu_1(x) > 0$ can only hold where $\bar u(x) = 0$. If $\bar u(x) > 0$, then $\bar\mu_1(x) = 0$. Moreover, u is admissible, hence $u \ge 0$ and we have almost everywhere

$$(u - \bar u)\bar\mu_1 \ge 0.$$

Therefore we get by (2.6)

$$\int_\Gamma (u - \bar u)\bar\mu_1 \, ds(x) \ge \int_{A_1} (u - \bar u)\bar\mu_1 \, ds(x) \ge \delta_1 \|u - \bar u\|_{L^1(A_1)}.$$

By our assumption, we have $\|u - \bar u\|_{L^\infty(\Gamma)} < \varepsilon$. In particular, this inequality implies $\|u - \bar u\|_{L^\infty(A_1)} < \varepsilon$. Consequently,

$$\int_\Gamma (u - \bar u)\bar\mu_1 \, ds(x) \ge \delta_1 \|u - \bar u\|_{L^1(A_1)} \frac{\|u - \bar u\|_{L^\infty(A_1)}}{\varepsilon} \ge \frac{\delta_1}{\varepsilon} \|u - \bar u\|^2_{L^2(A_1)},$$

and (3.21) is proven.

(ii) Next, we discuss the integral in (3.22). Because of (FON), $\bar\mu_2(x) > 0$ can hold only if $\bar u(x) + \gamma(x)\bar y(x) = c(x)$. In addition, (y, u) is admissible, hence in particular $c(x) \le u(x) + \gamma(x)y(x)$. Therefore, we obtain almost everywhere

$$(u - \bar u + \gamma(y - \bar y))\bar\mu_2 \ge 0$$

and

$$\begin{aligned} \int_\Gamma (u - \bar u + \gamma(y - \bar y))\bar\mu_2 \, ds(x) &\ge \int_{A_2} (u - \bar u + \gamma(y - \bar y))\bar\mu_2 \, ds(x) \\ &\ge \delta_2 \|u - \bar u + \gamma(y - \bar y)\|_{L^1(A_2)} \end{aligned} \tag{3.23}$$

by definition (2.7). Let us discuss this integral more detailed. Expressing $y - \bar y$ in terms of the controls by (3.17),

$$u - \bar u + \gamma(y - \bar y) = u - \bar u + \gamma(S'(\bar u)(u - \bar u) + r_1) \tag{3.24}$$

is found. Since $u = u_1 + u_2$ and $u_2 + \gamma S' u_2 + \gamma r_1 = 0$ on A_2 hold by Definition (3.19), we find

$$u + \gamma(S'u + r_1) = u_1 + u_2 + \gamma S'u_1 + \gamma S'u_2 + \gamma r_1 = u_1 + \gamma S'u_1 \quad \text{on } A_2. \tag{3.25}$$

Consequently, (3.23) and (3.24) yield

$$\begin{aligned} \int_\Gamma (u - \bar u + \gamma(y - \bar y))\bar\mu_2 \, ds(x) &\ge \delta_2 \|u_1 - \bar u + \gamma S'(\bar u)(u_1 - \bar u)\|_{L^1(A_2)} \\ &= \delta_2 \|w + \gamma S' w\|_{L^1(A_2)} \end{aligned} \tag{3.26}$$

with $w := u_1 - \bar u$. Notice, that $w = 0$ on $\Gamma \setminus A_2$. Moreover, we set $v = G'w$ and

$$z = \begin{cases} 0 & \text{on } \Gamma \setminus A_2, \\ w + \gamma v|_\Gamma & \text{on } A_2. \end{cases}$$

By this definition, we have

$$\|w + \gamma S' w\|_{L^1(A_2)} = \|z\|_{L^1(A_2)}$$

and therefore

$$\int_\Gamma (u-\bar u+\gamma(y-\bar y))\bar\mu_2\, ds(x) \ge \delta_2\|z\|_{L^1(A_2)}. \tag{3.27}$$

Then we find

$$\begin{aligned} Av+v &= 0 && \text{in } \Omega,\\ \partial_{n_A} v-\bar b_y v &= \bar b_u w && \text{on } \Gamma. \end{aligned} \tag{3.28}$$

On A_2 we have

$$\bar b_u w = \bar b_u(z-\gamma v|_\Gamma) = \bar b_u z - \chi_{A_2}\bar b_u \gamma v|_\Gamma.$$

Because of $z=w=0$ on $\Gamma\backslash A_2$, this equation is also correct on $\Gamma\backslash A_2$ and consequently it holds that

$$\begin{aligned} Av+v &= 0 && \text{in } \Omega,\\ \partial_{n_A} v+(-\bar b_y+\chi_{A_2}\bar b_u\gamma)v &= \bar b_u z && \text{on } \Gamma. \end{aligned} \tag{3.29}$$

Applying Lemma 3.3, we obtain

$$\|v\|_{L^1(\Gamma)} \le c\|z\|_{L^1(\Gamma)} = c\|z\|_{L^1(A_2)}. \tag{3.30}$$

Setting $\bar\gamma = \|\gamma\|_{C(\Gamma)}$, we get

$$\begin{aligned} \|w\|_{L^1(A_2)} &= \|z-\gamma v\|_{L^1(A_2)}\\ &\le \|z\|_{L^1(A_2)} + \bar\gamma\|v\|_{L^1(A_2)}\\ &\le \|z\|_{L^1(A_2)} + \bar\gamma c\|z\|_{L^1(A_2)} \end{aligned}$$

or

$$\|z\|_{L^1(A_2)} \ge \frac{1}{1+\bar\gamma c}\|w\|_{L^1(A_2)}. \tag{3.31}$$

Combining (3.27) and (3.31), we find

$$\int_\Gamma (u-\bar u+\gamma(y-\bar y))\bar\mu_2\, ds(x) \ge \frac{\delta_2}{1+\bar\gamma c}\|w\|_{L^1(A_2)} = \frac{\delta_2}{1+\bar\gamma c}\|u_1-\bar u\|_{L^1(A_2)}. \tag{3.32}$$

Invoking again $\|u-\bar u\|_{L^\infty(\Gamma)} < \varepsilon$ and (3.32), we obtain

$$\begin{aligned} \int_{A_2} (u-\bar u+\gamma(y-\bar y))\bar\mu_2\, ds(x) &\ge \frac{\delta_2}{1+\bar\gamma c}\|u_1-\bar u\|_{L^1(A_2)}\cdot\frac{\|u-\bar u\|_{L^\infty(A_2)}}{\varepsilon}\\ &\ge \frac{\delta_2}{c_4\varepsilon(1+\bar\gamma c)}\|u_1-\bar u\|^2_{L^2(A_2)}, \end{aligned}$$

implying inequality (3.22) with $c_A = \frac{1}{c_4(1+\bar\gamma c)}$. □

If $A_1\cup A_2=\Gamma$, then the critical subspace contains only the function $(y,u)=(0,0)$. Then the assumptions of Theorem 2.3 are trivially fulfilled. In this case, (3.21) and (3.22) imply the so-called *first-order sufficient optimality conditions*.

4. Second-order sufficient optimality condition. Here, we outline the proof of the sufficiency Theorem 2.3. This part is very similar to the discussion in [18]. Nevertheless, for the convenience of the reader, we present the main steps of the proof.

We select an arbitrary admissible control u in a sufficiently small L^∞-neighborhood of $\bar u$ and have to show that $F(y,u) \ge F(\bar y,\bar u)$. Let us introduce the increments $\delta u := u - \bar u$ and $\delta y := G'(\bar u)\delta u$. We split $\delta u = u_0 + u_+$, where

$$\begin{array}{lll} u_0 = 0, & u_+ = \delta u & \text{on } A_1, \\ u_0 = \delta u, & u_+ = 0 & \text{on } \Gamma \setminus (A_1 \cup A_2), \\ u_0 = -\gamma S'(\bar u)u_0, & u_+ = \delta u - u_0 & \text{on } A_2. \end{array} \tag{4.1}$$

Notice that $u_0 + \gamma S'(\bar u)u_0 = 0$ on A_2. This setting is justified again by Lemma 3.4: It holds

$$u_0 = \begin{cases} \phi & \text{on } \Gamma \setminus A_2, \\ \phi - \gamma S' u_0 & \text{on } A_2, \end{cases}$$

where ϕ is defined by

$$\phi = \begin{cases} 0 & \text{on } A_1 \cup A_2, \\ \delta u & \text{on } \Gamma \setminus (A_2 \cup A_1). \end{cases}$$

The part u_0 belongs to the critical subspace, while u_+ is the part of δu that accounts for the effects of first-order sufficiency. Furthermore, we define $y_0 := G'u_0$ and $y_+ := G'u_+$. By the linearity of G', we have $\delta y = y_0 + y_+$.

Below, we estimate the difference $L(y,u,\bar p,\bar\mu_1,\bar\mu_2) - L(\bar y,\bar u,\bar p,\bar\mu_1,\bar\mu_2)$. Let us write for short $L(y,u) - L(\bar y,\bar u)$, since $(\bar p,\bar\mu_1,\bar\mu_2)$ remains fixed in all the following considerations. We also do not explicitly indicate the point $(\bar y,\bar u,\bar p,\bar\mu_1,\bar\mu_2)$, where all derivatives are taken, i.e., we write $L_u u$ instead of $D_u L(\bar y,\bar u,\bar p,\bar\mu_1,\bar\mu_2)u$.

LEMMA 4.1. *Under the assumptions of Theorem* 2.3,

$$L(y,u) - L(\bar y,\bar u) \ge \frac{\delta}{4}\|u_0\|^2_{L^2(\Gamma)} - \frac{c_s}{2}\|u_+\|^2_{L^2(\Gamma)} + r_2 + \tilde r_2 \tag{4.2}$$

holds, where r_2, $\tilde r_2$ are second-order remainder terms with

$$\frac{|r_i|}{\|u-\bar u\|^2_{L^2(\Gamma)}} \to 0 \quad \textit{as } \|u-\bar u\|_{L^\infty(\Gamma)} \to 0.$$

Proof. Using Taylor's expansion, in view of (FON) we get

$$\begin{aligned} L(y,u) - L(\bar y,\bar u) &= L_u[u-\bar u] + L_y[y-\bar y] + \frac{1}{2}(L_{uu}[u-\bar u]^2 \\ &\quad + 2L_{uy}[u-\bar u, y-\bar y] + L_{yy}[y-\bar y]^2) + r_2 \\ &= \frac{1}{2}(L_{uu}[u-\bar u]^2 + 2L_{uy}[u-\bar u, y-\bar y] + L_{yy}[y-\bar y]^2) + r_2. \end{aligned} \tag{4.3}$$

The following property of the remainder is known:

$$\frac{|r_2(\bar u,h)|}{\|h\|^2_{L^2(\Gamma)}} \to 0 \quad \text{as } \|h\|_{L^\infty(\Gamma)} \to 0.$$

For the proof we refer to [22]. According to the notation of Lemma 3.4, we get $y-\bar{y}=\delta y+r_1$. Replacing $y-\bar{y}$ by δy in (4.3), another second-order remainder term is needed

$$\begin{aligned}\tilde{r}_2 := &\frac{1}{2}(L_{uu}[u-\bar{u}]^2+2L_{uy}[u-\bar{u},y-\bar{y}]+L_{yy}[y-\bar{y}]^2)\\ &-\frac{1}{2}(L_{uu}[\delta u]^2+2L_{uy}[\delta u,\delta y]+L_{yy}[\delta y]^2).\end{aligned}$$

It is easy to show that

$$\frac{|\tilde{r}_2|}{\|u-\bar{u}\|^2_{L^2(\Gamma)}}\to 0 \quad \text{as } \|u-\bar{u}\|_{L^\infty(\Gamma)}\to 0.$$

With these notations, (4.3) admits the form

$$L(y,u)-L(\bar{y},\bar{u})=\frac{1}{2}(L_{uu}[\delta u]^2+2L_{uy}[\delta u,\delta y]+L_{yy}[\delta y]^2)+r_2+\tilde{r}_2. \tag{4.4}$$

We continue by splitting the Lagrange functional in terms of u_0 and u_+,

$$\begin{aligned}L_{uu}[\delta u]^2+2L_{uy}[\delta u,\delta y]+L_{yy}[\delta y]^2 = &L_{uu}[u_0]^2+2L_{uy}[u_0,y_0]+L_{yy}[y_0]^2\\ &+L_{uu}[u_+]^2+2L_{uy}[u_+,y_+]+L_{yy}[y_+]^2\\ &+2L_{uu}[u_0,u_+]+2L_{uy}[u_0,y_+]\\ &+2L_{uy}[u_+,y_0]+2L_{yy}[y_0,y_+].\end{aligned}$$

As u_0 belongs to the critical subspace, the SSC yields

$$L''[u_0,y_0]^2=L_{uu}[u_0]^2+2L_{uy}[u_0,y_0]+L_{yy}[y_0]^2\ge\delta\|u_0\|^2_{L^2(\Gamma)}.$$

The other terms are easily estimated by $\|y_0\|^2_{L^2(\Gamma)}\le\|S'\|^2\|u_0\|^2_{L^2(\Gamma)}$, $\|y_+\|^2_{L^2(\Gamma)}\le\|S'\|^2\|u_+\|^2_{L^2(\Gamma)}$, and by means of Young's inequality,

$$\begin{aligned}|&L_{uu}[u_+]^2+2L_{uy}[u_+,y_+]+L_{yy}[y_+]^2\\ &+2L_{uu}[u_0,u_+]+2L_{uy}[u_0,y_+]\\ &+2L_{uy}[u_+,y_0]+2L_{yy}[y_0,y_+]|\le\frac{\delta}{2}\|u_0\|^2_{L^2(\Gamma)}+c_s\|u_+\|^2_{L^2(\Gamma)}.\end{aligned}$$

In this setting, c_s is a certain (large) constant. Combining the last two results, we arrive at

$$L_{uu}[\delta u]^2+2L_{uy}[\delta u,\delta y]+L_{yy}[\delta y]^2\ge\frac{\delta}{2}\|u_0\|^2_{L^2(\Gamma)}-c_s\|u_+\|^2_{L^2(\Gamma)}.$$

Returning to (4.4), we end up with

$$L(y,u)-L(\bar{y},\bar{u})\ge\frac{\delta}{4}\|u_0\|^2_{L^2(\Gamma)}-\frac{c_s}{2}\|u_+\|^2_{L^2(\Gamma)}+r_2+\tilde{r}_2,$$

which is exactly the assertion. □

In the next lemma, the term $\|u_+\|^2_{L^2(\Gamma)}$ in (4.2) is estimated.

LEMMA 4.2. *Under the assumptions of Theorem* 2.3,

$$\left(\frac{c_s}{2}+\frac{\delta}{4}\right)\|u_+\|^2_{L^2(\Gamma)}\le c_5\|u_1-\bar{u}\|^2_{L^2(A_2)}+c_6\|r_1\|^2_{L^2(\Gamma)}+c_7\|u-\bar{u}\|^2_{L^2(A_1)} \tag{4.5}$$

holds with certain positive constants c_5, c_6, and c_7.

Proof. First, we get on A_1

$$\|u_+\|_{L^2(A_1)} = \|\delta u\|_{L^2(A_1)} = \|u - \bar{u}\|_{L^2(A_1)}.$$

On the whole set Γ we have

$$u_+ + u_0 = \delta u = u - \bar{u}.$$

We apply the operator $I + \gamma S'$ to this equation and consider the image only on the set A_2. Using $u_0 = -\gamma S' u_0$ on A_2, we find

$$u_+ + \gamma S' u_+ = u + \gamma S' u - (\bar{u} + \gamma S' \bar{u}) \text{ on } A_2. \tag{4.6}$$

Now, u is again replaced by $u_1 + u_2$; see (3.19) to obtain on A_2,

$$u_+ + \gamma S' u_+ = u_1 + \gamma S' u_1 + u_2 + \gamma S' u_2 - (\bar{u} - \gamma S' \bar{u}).$$

By definition (3.19), the equation $u_2 + \gamma S' u_2 = -r_1$ is satisfied on A_2. Therefore, here we are able to continue by

$$u_+ + \gamma S' u_+ = u_1 - \bar{u} + (\gamma S'(\bar{u})(u_1 - \bar{u})) - r_1 \text{ on } A_2. \tag{4.7}$$

Due to (4.1), $u_+ = \delta u = u - \bar{u}$ holds on A_1. In addition, u_+ vanishes on $\Gamma \setminus (A_1 \cup A_2)$. Therefore, we find by (4.1) and (4.7),

$$u_+ = \begin{cases} u_1 - \bar{u} - \gamma S'(\bar{u})(u_+ - u_1 + \bar{u}) - r_1 & \text{on } A_2, \\ u - \bar{u} & \text{on } A_1, \\ 0 & \text{on } \Gamma \setminus (A_1 \cup A_2). \end{cases}$$

Again we have a construction that was investigated in Lemma 3.4. Applying (3.12), we get the inequality

$$\|u_+\|_{L^2(\Gamma)} \le c_2 \|\phi\|_{L^2(\Gamma)},$$

where ϕ is defined by

$$\phi = \begin{cases} -r_1 + (u_1 - \bar{u}) + \gamma S'(\bar{u})(u_1 - \bar{u}) & \text{on } A_2, \\ u - \bar{u} & \text{on } A_1, \\ 0 & \text{on } \Gamma \setminus (A_1 \cup A_2). \end{cases}$$

Therefore, we obtain

$$\|u_+\|_{L^2(\Gamma)} \le c_2(\|u - \bar{u}\|_{L^2(A_1)} + c_8 \|u_1 - \bar{u}\|_{L^2(\Gamma)} + \|r_1\|_{L^2(A_2)}),$$

where the positive constant c_8 depends on $\|S'\|$. In view of (3.19), it holds that $\|u_1 - \bar{u}\|_{L^2(\Gamma)} = \|u_1 - \bar{u}\|_{L^2(A_2)}$, and hence we get

$$\|u_+\|_{L^2(\Gamma)} \le c_9 \|u_1 - \bar{u}\|_{L^2(A_2)} + c_2 \|r_1\|_{L^2(A_2)} + c_2 \|u - \bar{u}\|_{L^2(A_1)}.$$

Young's inequality yields

$$\|u_+\|^2_{L^2(\Gamma)} \le 3c_9 \|u_1 - \bar{u}\|^2_{L^2(A_2)} + 3c_2 \|r_1\|^2_{L^2(\Gamma)} + 3c_2 \|u - \bar{u}\|^2_{L^2(A_1)}.$$

A multiplication by $(\frac{c_s}{2} + \frac{\delta}{4})$,

$$\left(\frac{c_s}{2} + \frac{\delta}{4}\right) \|u_+\|^2_{L^2(\Gamma)} \le c_5 \|u_1 - \bar{u}\|^2_{L^2(A_2)} + c_6 \|r_1\|^2_{L^2(\Gamma)} + c_7 \|u - \bar{u}\|^2_{L^2(A_1)},$$

concludes the proof of the lemma. ∎

Now we are able to prove our main result Theorem 2.3.

Proof of Theorem 2.3. Inserting (4.5) in (4.2),

$$\begin{aligned} L(y,u) - L(\bar{y},\bar{u}) \ge \frac{\delta}{4}(\|u_0\|^2_{L^2(\Gamma)} + \|u_+\|^2_{L^2(\Gamma)}) + r_2 + \tilde{r}_2 \\ -c_7 \|u - \bar{u}\|^2_{L^2(A_1)} - c_5 \|u_1 - \bar{u}\|^2_{L^2(A_2)} - c_6 \|r_1\|^2_{L^2(\Gamma)} \end{aligned}$$

is obtained. Next, we return to the objective F,

$$\begin{aligned} L(y,u) - L(\bar{y},\bar{u}) = F(y,u) - F(\bar{y},\bar{u}) - \int_\Gamma \bar{\mu}_1 (u - \bar{u})\, ds(x) - \int_\Gamma (u - \bar{u} \\ + \gamma(y - \bar{y}))\bar{\mu}_2\, ds(x). \end{aligned}$$

Using Lemma 3.5 we find

$$\begin{aligned} F(y,u) - F(\bar{y},\bar{u}) \ge \frac{\delta}{4}(\|u_0\|^2_{L^2(\Gamma)} + \|u_+\|^2_{L^2(\Gamma)}) + r_2 + \tilde{r}_2 \\ + \left(\frac{\delta_1}{\varepsilon} - c_7\right) \|u - \bar{u}\|^2_{L^2(A_1)} + \left(c_A \frac{\delta_2}{\varepsilon} - c_5\right) \|u_1 - \bar{u}\|^2_{L^2(A_2)} \\ -c_6 \|r_1\|^2_{L^2(\Gamma)}. \end{aligned} \tag{4.8}$$

Next, $\|\delta u\|_{L^2(\Gamma)} = \|u_0 + u_+\|^2_{L^2(\Gamma)} \le 2\|u_0\|^2_{L^2(\Gamma)} + 2\|u_+\|^2_{L^2(\Gamma)}$ is applied to continue by

$$\begin{aligned} F(y,u) - F(\bar{y},\bar{u}) \ge \frac{\delta}{8} \|\delta u\|^2_{L^2(\Gamma)} + r_2 + \tilde{r}_2 \\ + \left(\frac{\delta_1}{\varepsilon} - c_7\right) \|u - \bar{u}\|^2_{L^2(A_1)} + \left(c_A \frac{\delta_2}{\varepsilon} - c_5\right) \|u_1 - \bar{u}\|^2_{L^2(A_2)} \\ -c_6 \|r_1\|^2_{L^2(\Gamma)}. \end{aligned} \tag{4.9}$$

Now take ε sufficiently small, such that

$$\frac{\delta_1}{\varepsilon} - c_7 \ge 0 \quad \text{and} \quad c_A \frac{\delta_2}{\varepsilon} - c_5 \ge 0.$$

Then we can omit the associated terms in (4.9),

$$F(y,u) - F(\bar{y},\bar{u}) \ge \frac{\delta}{8} \|\delta u\|^2_{L^2(\Gamma)} + r_2 + \tilde{r}_2 - c_6 \|r_1\|^2_{L^2(\Gamma)}. \tag{4.10}$$

Due to the discussions during the proof, all terms on the right-hand side (except the first one) are small with respect to $\|u - \bar{u}\|^2_{L^2(\Gamma)}$. Therefore

$$F(y,u) - F(\bar{y},\bar{u}) \ge \frac{\delta}{16} \|u - \bar{u}\|^2_{L^2(\Gamma)} \tag{4.11}$$

holds if $\|u - \bar{u}\|_{L^\infty(\Gamma)} < \varepsilon$ and ε is sufficiently small. The quadratic growth condition is proven. We can now choose $\delta_s = \delta/16$. ∎

5. Generalizations. In this section, we discuss weaker assumptions and possible generalizations. The second-order sufficient optimality condition can be weakened. Let us define the weakly active control constraints

$$A_1^{weak} := \{x \in \Gamma \setminus (A_1 \cup A_2) : \ \bar\mu_1(x) > 0\}.$$

On A_1^{weak} we have almost everywhere $\bar u(x) = 0$. The control constraints imply $u(x) - \bar u(x) \geq 0$ a.e. on A_1^{weak} for all admissible controls u. Therefore, it is enough to consider only those elements of the critical subspace, which satisfy the condition

$$u(x) \geq 0 \qquad \forall x \in A_1^{weak}.$$

COROLLARY 5.1. *Suppose that the following weakened second-order sufficient optimality condition is satisfied: The condition (SSC) is required to be fulfilled only for those elements (y, u) belonging to the critical subspace (defined in Definition* 2.2*) which satisfy the condition*

$$u \geq 0 \quad \text{on } A_1^{weak}.$$

This weaker assumption ensures the result of Theorem 2.3*, too.*

The presented techniques can be extended to derive sufficient second-order optimality conditions for other types of optimal control problems. For instance, it applies to distributed elliptic control problems with mixed control constraints considered in Ω. Moreover, it works for two-sided constraints on the control, where we minimize (1.1) subject to (1.2),

$$u_a \leq u(x) \leq u_b \quad \text{for } x \in \Gamma, \tag{5.1}$$

together mixed control-state constraints

$$c(x) \leq u(x) + \gamma(x)y(x) \quad \text{for } x \in \Gamma. \tag{5.2}$$

In this case, the Lagrange functional is

$$\begin{aligned} L(y,u,p,\mu_1,\mu_2,\mu_3) = F(y,u) &+ \int_\Omega \Big(\sum_{i,j=1}^m a_{ij} D_j y D_i p + yp \Big)\, dx - \int_\Gamma b\, p\, ds(x) \\ &- \int_\Gamma \mu_1(u - u_a)\, ds(x) + \int_\Gamma \mu_3(u - u_b)\, ds(x) \\ &- \int_\Gamma (u + \gamma y - c)\mu_2\, ds(x). \end{aligned}$$

In this definition, we tacitly assume that the Lagrange multiplier μ_2 for the constraint (5.2) is a bounded and measurable function. In contrast to the former sections, we have not been able to show this. Then, the necessary first-order optimality conditions

are

$$(FONBOX)\quad\begin{cases} D_y L(\bar y,\bar u,\bar p,\bar\mu_1,\bar\mu_2,\bar\mu_3) &= 0\\ D_u L(\bar y,\bar u,\bar p,\bar\mu_1,\bar\mu_2,\bar\mu_3) &= 0\\ \text{and for almost all } x\in\Gamma & \\ \bar\mu_1(x) &\geq 0\\ \bar\mu_3(x) &\geq 0\\ \bar\mu_2(x) &\geq 0\\ (\bar u(x)-u_a)\bar\mu_1(x) &= 0\\ (\bar u(x)-u_b)\bar\mu_3(x) &= 0\\ (\bar u(x)+\gamma(x)\bar y(x)-c(x))\bar\mu_2(x) &= 0. \end{cases}$$

DEFINITION 5.2. *The strongly active sets for problem* (1.1), (1.2), (5.1), (5.2) *are*

$$A_1(\delta_1) := \{x\in\Gamma : \bar\mu_1(x)\geq\delta_1\}, \tag{5.3}$$

$$A_3(\delta_3) := \{x\in\Gamma : \bar\mu_3(x)\geq\delta_3\}, \tag{5.4}$$

$$A_2(\delta_2) := \{x\in\Gamma\setminus(A_1(\delta_1)\cup A_3(\delta_3)) : \bar\mu_2(x)\geq\delta_2\}. \tag{5.5}$$

A pair $(y,u)\in C(\bar\Omega)\times L^\infty(\Gamma)$ *belongs to the critical subspace, if*

$$u = 0 \qquad \text{on } A_1(\delta_1)\cup A_3(\delta_3), \tag{5.6}$$

$$u+\gamma y|_\Gamma = 0 \qquad \text{on } A_2(\delta_2), \tag{5.7}$$

and

$$\begin{aligned} Ay + y &= 0 && \text{in } \Omega,\\ \partial_{n_A} y - \bar b_y\ y &= \bar b_u\ u && \text{in } \Gamma. \end{aligned} \tag{5.8}$$

Again, (5.8) implies $y|_\Gamma = S'(\bar u)u$.

(SSCBOX): There exist positive numbers δ, δ_1, δ_2, δ_3 such that the definiteness condition

$$L''_{(u,y)}(\bar y,\bar u,\bar p,\bar\mu_1,\bar\mu_2)[h_y,h_u]^2 \geq \delta\|h_u\|^2_{L^2(\Gamma)} \tag{5.9}$$

holds true for all (h_y,h_u) belonging to the critical subspace defined upon δ_1, δ_2, δ_3.

THEOREM 5.3 (second-order sufficiency for box constraints and mixed constraints). *Assume that* $(\bar y,\bar u,\bar p,\bar\mu_1,\bar\mu_2,\bar\mu_3)$ *fulfills the first-order optimality system (FONBOX) and the regularity condition (R) is satisfied. If the second-order condition (SSCBOX) is satisfied, then there exist* $\delta_s>0$ *and* $\varepsilon>0$ *such that the quadratic growth condition*

$$F(y,u)-F(\bar y,\bar u)\geq\delta_s\|u-\bar u\|^2_{L^2(\Gamma)} \tag{5.10}$$

holds for all admissible pairs (y,u) *with* $\|u-\bar u\|_{L^\infty(\Gamma)}<\varepsilon$. *Therefore,* $\bar u$ *is a locally optimal control in the norm of* $L^\infty(\Gamma)$.

This result can be shown along the lines of the sections 3 and 4—with minor modifications. For the upper control constraint, we find another estimate of the type (3.21). In section 4, A_1 has to be replaced by $A_1\cup A_3$ and δ_1 by $\min(\delta_1,\delta_3)$.

REFERENCES

[1] J.-J. Alibert and J.-P. Raymond, *Boundary control of semilinear elliptic equations with discontinuous leading coefficients and unbounded controls*, Numer. Funct. Anal. Optim., 18 (1997), pp. 235–250.

[2] N. Arada and J. P. Raymond, *Optimal control problems with mixed control-state constraints*, SIAM J. Control Optim., 39 (2000), pp. 1391–1407.

[3] M. Bergounioux and F. Tröltzsch, *Optimal control of semilinear parabolic equations with state-constraints of bottleneck type*, ESAIM Control Optim. Calc. Var., 4 (1999), pp. 595–608.

[4] J. F. Bonnans, *Second-order analysis for control constrained optimal control problems of semilinear elliptic equations*, Appl. Math. Optim., 38 (1998), pp. 303–325.

[5] E. Casas, *Control of an elliptic problem with pointwise state constraints*, SIAM J. Control Optim., 24 (1986), pp. 1309–1318.

[6] E. Casas, *Boundary control of semilinear elliptic equations with pointwise state constraints*, SIAM J. Control Optim., 31 (1993), pp. 993–1006.

[7] E. Casas and M. Mateos, *Second order sufficient optimality conditions for semilinear elliptic control problems with finitely many state constraints*, SIAM J. Control Optim., 40 (2002), pp. 1431–1454.

[8] 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.

[9] E. Casas, F. Tröltzsch, and A. Unger, *Second order sufficient optimality conditions for a nonlinear elliptic control problem*, Z. Anal. Anwendungen, 15 (1996), pp. 687–707.

[10] 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.

[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] H. Goldberg and F. Tröltzsch, *Second order sufficient optimality conditions for a class of non–linear parabolic boundary control problems*, SIAM J. Control Optim., 31 (1993), pp. 1007–1025.

[13] 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.

[14] H. Maurer, *First and second order sufficient optimality conditions in mathematical programming and optimal control*, Math. Programming Stud., 14 (1981), pp. 163–177.

[15] 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.

[16] A. Rösch and F. Tröltzsch, *Existence of regular Lagrange multipliers for a nonlinear elliptic optimal control problem with pointwise control-state constraints*, SIAM J. Control Optim., 45 (2006), pp. 548–564.

[17] A. Rösch and F. Tröltzsch, *Sufficient second order optimality conditions for a state-constrained optimal control problem of a weakly singular integral equation*, Numer. Funct. Anal. Optim., 23 (2002), pp. 173–193.

[18] A. Rösch and F. Tröltzsch, *Sufficient second order optimality conditions for a parabolic optimal control problem with pointwise control-state constraints*, SIAM J. Control Optim., 42 (2003), pp. 138–154.

[19] G. Stampacchia, *Le problème de Dirichlet pour les équations elliptiques du second ordre à coefficients discontinus*, Ann. Inst. Fourier (Grenoble), 15 (1965), pp. 189–258.

[20] H. Triebel, *Interpolation Theory, Function Spaces, Differential Operators*, Johann Ambrosius Barth, Heidelberg, 1995.

[21] F. Tröltzsch, *Regular Lagrange multipliers for control problems with mixed pointwise control-state constraints*, SIAM J. Optim., 15 (2005), pp. 616–634.

[22] F. Tröltzsch, *Approximation of nonlinear parabolic boundary control problem by the Fourier method—convergence of optimal controls*, Optimization, 22 (1991), pp. 83–98.

SIAM J. OPTIM.
Vol. 17, No. 3, pp. 795–805
© 2006 Society for Industrial and Applied Mathematics

SECOND-ORDER SUFFICIENT CONDITIONS FOR ERROR BOUNDS IN BANACH SPACES*

YIRAN HE† AND JIE SUN‡

Abstract. Recently, Huang and Ng presented second-order sufficient conditions for error bounds of continuous and Gâteaux differentiable functions in Banach spaces. Wu and Ye dropped the assumption of Huang and Ng on Gâteaux differentiability but required the space to be a Hilbert space. We carry on this research in two directions. First we extend Wu and Ye's result to some non-Hilbert spaces; second, same as Huang and Ng, we work on Banach spaces but provide different second-order sufficient conditions that may allow the function to be non-Gâteaux differentiable.

Key words. error bound, Hölder smooth subdifferential, proximal subdifferential, nonsmooth analysis

AMS subject classifications. 46B20, 49J52, 90C26, 90C31

DOI. 10.1137/040621661

1. Introduction. We consider error bounds for lower semicontinuous functions in Banach spaces. Let f be a proper lower semicontinuous function on a Banach space X. Our goal is to study conditions that guarantee the existence of positive constants γ and m such that

$$\operatorname{dist}^m(x, S) \le \gamma f(x)_+ \quad \text{for all } x \in X, \tag{1.1}$$

where $S := f^{-1}(-\infty, 0]$ and $f(x)_+ := \max\{f(x), 0\}$. We call (1.1) an error bound of order m. If (1.1) holds for $m = 1$, then the error bound is of Lipschitz type, which has been much discussed in the literature; see [7, 10, 11, 12, 13, 14, 15] and the book [5]. If the function f is convex, then there exist many equivalent characterizations for error bounds in terms of the first-order directional derivative or first-order subdifferential of function f. However, if the function is not convex, one usually gives only sufficient conditions in terms of various first-order generalized subdifferentials or first-order generalized directional derivatives [7, 8, 12, 14].

The first-order conditions used in the nonconvex case require that the generalized subdifferentials of f for all $x \notin S$ are bounded away from zero. Specifically, let ∂ be a certain generalized subdifferential of f and let

$$P(\alpha) := \{x \in X : x \notin S, \partial f(x) \cap B(0, \alpha) \neq \emptyset\},$$

where $B(x, \alpha)$ denotes a closed ball centered at x with radius α. In order to establish error bounds for nonconvex functions, it is usually assumed that $P(\alpha)$ is empty for some $\alpha > 0$; in other words, there exists a positive scalar α such that $\|\xi\| \ge \alpha$ for

*Received by the editors December 28, 2004; accepted for publication (in revised form) October 10, 2005; published electronically October 4, 2006. This research was partially supported by grant R-314-000-042/057-112 from the National University of Singapore and Singapore-MIT Alliance.
http://www.siam.org/journals/siopt/17-3/62166.html

†Department of Mathematics, Sichuan Normal University, Chengdu, Sichuan, People's Republic of China (yiranhe@hotmail.com). The work of this author was supported by NSFC and Sichuan Youth Science and Technology Foundation (06ZQ026-013).

‡Department of Decision Sciences and Singapore-MIT Alliance, National University of Singapore, Republic of Singapore (jsun@nus.edu.sg).

all $\xi \in \partial f(X \setminus S)$. This assumption is quite restrictive. One naturally asks whether there are certain conditions for error bound to hold, provided that

$$P(\alpha) \neq \emptyset \quad \text{for every } \alpha > 0. \tag{1.2}$$

If f is sufficiently smooth such that $\partial f(x)$ is a singleton and equals the derivative $f'(x)$ of f for every $x \notin S$, then (1.2) is equivalent to the existence of a sequence $\{x_n\}$ in $X \setminus S$ satisfying that $\lim_{n\to\infty} f'(x_n) = 0$.

Recently, some researchers have considered second-order sufficient conditions for error bounds of lower semicontinuous functions. Huang and Ng [7] proved that if f is Gâteaux differentiable and continuous in a Banach space, then an error bound of Lipschitz type holds under an assumption on certain second-order directional derivatives. Wu and Ye [15] removed this assumption and established a similar result. However, their result requires the space to be a Hilbert space. In this paper we present results that extend Wu and Ye's result to non-Hilbert spaces and results that extend Huang and Ng's work to possibly non-Gâteaux differentiable functions in Banach spaces.

2. Smoothness and subdifferentials. Let X be a Banach space. $B(x, r)$ and $B_r(x)$ denote the closed and the open ball centered at x with radius $r > 0$, respectively.

DEFINITION 2.1 (see [9]). *The modulus of smoothness $\rho_X(\tau)$, $\tau > 0$, of X is defined as*

$$\rho_X(\tau) := \sup\{(\|x + y\| + \|x - y\|)/2 - 1 : x, y \in X, \|x\| = 1, \|y\| = \tau\}.$$

X is said to be uniformly smooth if $\lim_{\tau\to 0+} \rho_X(\tau)/\tau = 0$. A uniformly smooth Banach space is said to have modulus of smoothness of power p *if for some $s > 0$,*

$$\rho_X(\tau) \leq s\tau^p \quad \textit{for all } \tau \geq 0. \tag{2.1}$$

Consider the example of $X = L_p$ $(p > 1)$. For $\tau \geq 0$,

$$\rho_{L_p}(\tau) \leq \begin{cases} \tau^p/p, & p \in (1, 2), \\ (p-1)\tau^2/2, & p \in [2, \infty). \end{cases}$$

Thus, L_p is uniformly smooth for $p > 1$ and has modulus of smoothness of power p for $p \in (1, 2)$ and of power 2 for $p \geq 2$. Let

$$J_p(x) := \{\xi \in X^* : \langle \xi, x\rangle = \|\xi\| \|x\|, \|\xi\| = \|x\|^{p-1}\}.$$

It is known that every uniformly smooth Banach space is reflexive, and if X is a reflexive Banach space, then $J_p(x)$ is the subdifferential of the convex function $x \mapsto \|x\|^p / p$. That is, $\xi \in J_p(x)$ if and only if

$$\|y\|^p / p - \|x\|^p / p \geq \langle \xi, y - x\rangle \quad \text{for all } y \in X.$$

In general, $J_p(x)$ is not necessarily a singleton; however, X is uniformly smooth if and only if $J_p(x)$ is single valued and uniformly continuous on bounded sets [4].

LEMMA 2.2. *Let X be a uniformly smooth Banach space, $x, y \in X$, and $m > 1$. Then*

$$\|y\|^m - \|x\|^m \geq m \langle J_m(x), y - x\rangle .$$

Proof. This is obvious from the definition of subdifferential inequality of convex functions. □

LEMMA 2.3. *Let X be a uniformly smooth Banach space and $x, y \in X$. If X has modulus of smoothness of power m for some $m > 1$, then there exists a constant $L > 0$ such that*

$$\langle J_m(x) - J_m(y), x - y\rangle \le L \|x - y\|^m \quad \text{for all } x, y \in X. \tag{2.2}$$

Proof. See Theorem 2 and Remarks 4 and 5 in [16]. □

Let $f : X \to \mathbb{R} \cup \{\infty\}$ be a proper lower semicontinuous function with

$$\operatorname{dom} f := \{x \in X : f(x) < \infty\} \neq \emptyset.$$

Let us recall several well-known subdifferentials. Let $x \in \operatorname{dom} f$.

- The *Hölder-smooth subdifferential* of order $p > 1$ of f at x is defined as (see [2])

$$\partial_p^{HS} f(x) := \left\{\xi \in X^* : \liminf_{\|v\|\to 0} \frac{f(x+v) - f(x) - \langle \xi, v\rangle}{\|v\|^p} > -\infty\right\}.$$

  When $p = 2$, $\partial_p^{HS} f(x)$ is just the *Lipschitz-smooth subdifferential* $\partial^{LS} f(x)$ of f at x [1]:

$$\partial^{LS} f(x) := \left\{\xi \in X^* : \liminf_{\|v\|\to 0} \frac{f(x+v) - f(x) - \langle \xi, v\rangle}{\|v\|^2} > -\infty\right\}. \tag{2.3}$$

  When X is a Hilbert space and $p = 2$, $\partial_p^{HS} f(x)$ coincides with the proximal subdifferential $\partial^P f(x)$ [3]. Note that $\xi \in \partial^P f(x)$ if and only if there exist $\eta > 0$ and $\sigma > 0$ such that

$$f(x+v) - f(x) \ge \langle \xi, v\rangle - \sigma \|v\|^2 \quad \text{for all } v \in B(0, \eta).$$

- The *Fréchet subdifferential* of f at x is the set

$$\partial^F f(x) := \left\{\xi \in X^* : \liminf_{\|v\|\to 0} \frac{f(x+v) - f(x) - \langle \xi, v\rangle}{\|v\|} \ge 0\right\}.$$

- The *Clarke–Rockafellar subdifferential* of f at x is the set

$$\partial^{CR} f(x) := \left\{\xi \in X^* : \langle \xi, v\rangle \le \sup_{\varepsilon>0} \limsup_{y \to^f x\, t\downarrow 0} \inf_{u \in B_\varepsilon(v)} \frac{f(y+tu) - f(y)}{t},\ \forall v \in X\right\},$$

  where $y \to^f x$ means $y \to x$ and $f(y) \to f(x)$; when f is locally Lipschitz at x, the Clarke–Rockafellar subdifferential coincides with the *Clarke subdifferential*

$$\partial^C f(x) := \left\{\xi \in X^* : \langle \xi, v\rangle \le \limsup_{(y,t)\to(x,0^+)} \frac{f(y+tv) - f(y)}{t}, \forall v \in X\right\}.$$

- The *Hadamard subdifferential* of f at x is the set

$$\partial^H f(x) := \left\{\xi \in X^* : \langle \xi, v\rangle \le \liminf_{(u,t)\to(v,0^+)} \frac{f(x+tu) - f(x)}{t}, \forall v \in X\right\}.$$

  When f is locally Lipschitz at x, the Hadamard subdifferential coincides with the *Gâteaux subdifferential*

$$\partial^G f(x) := \left\{\xi \in X^* : \langle \xi, v\rangle \le \liminf_{t\to 0^+} \frac{f(x+tv) - f(x)}{t}, \forall v \in X\right\}.$$

It is straightforward to verify that for $p > 1$,

$$\partial_p^{HS} f(x) \subset \partial^F f(x) \subset \partial^H f(x) \subset \partial^{CR} f(x). \tag{2.4}$$

PROPOSITION 2.4. *Let g be a continuous function on a Banach space X. Suppose that $\partial_p^{HS} g(x)$ and $\partial_p^{HS}(-g)(x)$ are both nonempty. Then $\partial_p^{HS} g(x)$ is equal to $-\partial_p^{HS}(-g)(x)$ and $\partial_p^{HS} g(x)$ is a singleton.*

Proof. Let $\xi \in \partial_p^{HS} g(x)$ and $x^* \in \partial_p^{HS}(-g)(x)$. From the definition of the Hölder-smooth subdifferential, there exist $\sigma > 0$ and $\eta > 0$ such that for all $v \in B(0, \eta)$,

$$\begin{aligned} g(x+v) - g(x) &\geq \langle \xi, v \rangle - (\sigma/2) \|v\|^p, \\ -g(x+v) + g(x) &\geq \langle x^*, v \rangle - (\sigma/2) \|v\|^p. \end{aligned}$$

Adding these two expressions together, we have

$$\langle \xi + x^*, v \rangle \leq \sigma \|v\|^p \quad \text{for all } v \in B(0, \eta),$$

which implies that $\xi + x^* = 0$ as $p > 1$. Since $\xi \in \partial_p^{HS} g(x)$ and $x^* \in \partial_p^{HS}(-g)(x)$ are arbitrary, $\partial_p^{HS} g(x)$ is equal to $-\partial_p^{HS}(-g)(x)$ and is a singleton. ∎

PROPOSITION 2.5. *The subdifferential ∂_p^{HS} has the following properties:*

(P1) *$\partial_p^{HS} f(x)$ coincides with the subdifferential in the sense of convex analysis whenever f is convex;*

(P2) *$0 \in \partial_p^{HS} f(x)$ whenever $x \in \operatorname{dom} f$ is a local minimum of f;*

(P3) *$\partial_p^{HS}(f+g)(x) \subset \partial_p^{HS} f(x) + \partial_p^{HS} g(x)$ whenever g is a continuous function with the property that $\partial_p^{HS} g(x)$ and $\partial_p^{HS}(-g)(x)$ are both nonempty.*

Proof. (P1) Let g be a convex function and $x \in \operatorname{dom} g$. Just observe that for a convex function the Clarke–Rockafellar subdifferential and the usual (Fenchel) subdifferential in convex analysis coincide for lower semicontinuous functions and that the Fenchel subdifferential is obviously contained in $\partial_p^{HS} g(x)$. The conclusion follows immediately from (2.4).

(P2) It is obvious from the definition of ∂_p^{HS}.

(P3) Note that

$$\partial_p^{HS} f(x) = \partial_p^{HS}(f+g-g)(x) \supset \partial_p^{HS}(f+g)(x) + \partial_p^{HS}(-g)(x), \tag{2.5}$$

where the inclusion relation is from the definition of the Hölder-smooth subdifferential. Since g is continuous and $\partial_p^{HS} g(x)$ and $\partial_p^{HS}(-g)(x)$ are both nonempty, by virtue of Proposition 2.4, $\partial_p^{HS}(-g)(x)$ is a singleton and $\partial_p^{HS}(-g)(x) = -\partial_p^{HS} g(x)$. This together with (2.5) yield the conclusion. ∎

PROPOSITION 2.6. *If X is a uniformly smooth Banach space which has modulus of smoothness of power p for some $p > 1$ and $x \neq 0$, then the Hölder-smooth subdifferential of order p of the functions $\|x\|^p / p$ and $-\|x\|^p / p$ are nonempty and $\partial_p^{HS}(-\|\cdot\|^p / p)(x) = -J_p(x)$.*

Proof. Since X is uniformly smooth, the function $\|\cdot\|$ and hence the convex function $\|\cdot\|^p / p$ are Fréchet differentiable at x. Therefore $\partial_p^{HS}(\|\cdot\|)(x)$ is nonempty by Proposition 2.5. Now we prove that $\partial_p^{HS}(-\|\cdot\|^p / p)(x)$ is nonempty. Since $J_p(x)$ is the subdifferential of $\|x\|^p / p$ in the sense of convex analysis, for $v \neq 0$,

$$\frac{-\|x+v\|^p / p + \|x\|^p / p - \langle -J_p(x), v \rangle}{\|v\|^p} \geq \frac{\langle J_p(x) - J_p(x+v), v \rangle}{\|v\|^p} \geq -L,$$

where the last inequality follows from Lemma 2.3 and L is the constant that appeared in Lemma 2.3. This proves that $-J_p(x)$ belongs to $\partial_p^{HS}(-\|\cdot\|^p / p)(x)$ by the definition of the Hölder-smooth subdifferential. ∎

3. Error bounds in smooth Banach spaces. The following result generalizes the second-order sufficient condition for error bounds established in [15] from the Hilbert space to smooth Banach spaces.

THEOREM 3.1. *Let X be a uniformly smooth Banach space which has modulus of smoothness of power m for some $m > 1$, and let $f : X \to \mathbb{R} \cup \{\infty\}$ be a proper lower semicontinuous function. Suppose that there exists $\delta > 0$ such that for all $x \in f^{-1}(0, \infty)$,*

$$\liminf_{\|u\| \to 1, t \downarrow 0} \frac{f(x+tu) - f(x) - t\langle \xi, u\rangle}{t^m} < -\delta \quad \text{for each } \xi \in \partial_m^{HS} f(x). \tag{3.1}$$

Then

$$\operatorname{dist}^m(x, S) \le (mL/\delta)\, f(x)_+ \quad \text{for all } x \in X, \tag{3.2}$$

where L is the constant that appeared in (2.2).

Proof. Write γ for mL/δ. Suppose that the conclusion does not hold: there exists some u with $f(u) > 0$ such that

$$\operatorname{dist}^m(u, S) > \gamma\, f(u).$$

Then we can find $t > 1$ such that $\operatorname{dist}^m(u, S) > t\gamma\, f(u)$, and hence

$$f(u) = f(u)_+ < \inf_{x \in X} f(x)_+ + \gamma^{-1} c, \tag{3.3}$$

where $c := t\gamma\, f(u)$. Applying the Borwein–Preiss smooth variational principle [2], we obtain the existence of some $v \in X$ such that

$$\|u - v\| < \sqrt[m]{c} \quad \text{and} \tag{3.4}$$

$$f(v)_+ + \gamma^{-1}\Delta_m(v) \le f(x)_+ + \gamma^{-1}\Delta_m(x) \quad \text{for all } x \in X, \tag{3.5}$$

where $\Delta_m(x) := \sum_{k=1}^{\infty} \mu_k \|x - v_k\|^m$ for some sequence $\{v_k\}$ converging to v and some sequence $\{\mu_k\}$ satisfying $\mu_k > 0$ and $\sum_{k=1}^{\infty} \mu_k = 1$.

It follows from (3.4) and the choice of u that $v \notin S$. Hence v is a global minimizer of the function $f(x) + \gamma^{-1}\Delta_m(x)$ and hence a global minimizer of the function $\gamma m^{-1} f(x) + m^{-1}\Delta_m(x)$ over the open set $X \setminus S$. In view of the definition of Hölder-smooth subdifferential ∂_m^{HS}, it follows that

$$0 \in \partial_m^{HS}(\gamma m^{-1} f + m^{-1}\Delta_m)(v). \tag{3.6}$$

Clearly $m^{-1}\Delta_m(x)$ is a real valued continuous convex function. Hence $\partial_m^{HS}(m^{-1}\Delta_m)(v)$ coincides with the subdifferential in the sense of convex analysis by Proposition 2.5 and so is nonempty. Since the space X is uniformly smooth, it follows that for every x, $J_m(x - v_k)$ is a singleton for each k and the sequence $\{J_m(x - v_k)\}_{k=1}^{\infty}$ is bounded. Thus, $m^{-1}\Delta_m(x)$ is Fréchet differentiable with its Fréchet derivative $(m^{-1}\Delta_m)'(x) = \sum_{k=1}^{\infty} \mu_k J_m(x - v_k)$. Since $\partial_m^{HS}(m^{-1}\Delta_m)(v)$ is nonempty, it follows that

$$\partial_m^{HS}(m^{-1}\Delta_m)(v) = \left\{(m^{-1}\Delta_m)'(v)\right\}. \tag{3.7}$$

We claim that $\partial_m^{HS}(-m^{-1}\Delta_m)(v)$ contains $-(m^{-1}\Delta_m)'(v)$ and hence is nonempty. This together with (3.6), Propositions 2.4 and 2.5, and (3.7) yields that

$$\xi := -m\gamma^{-1} \sum_{k=1}^{\infty} \mu_k J_m(v - v_k) \in \partial_m^{HS} f(v). \tag{3.8}$$

Indeed,

$$\begin{aligned}
&\liminf_{h\to 0} \frac{(-m^{-1}\Delta_m)(v+h)-(-m^{-1}\Delta_m)(v)-\left\langle(-m^{-1}\Delta_m)'(v),h\right\rangle}{\|h\|^m}\\
&=\liminf_{h\to 0} \frac{\left\langle(-m^{-1}\Delta_m)'(v+\theta(h)h),h\right\rangle-\left\langle(-m^{-1}\Delta_m)'(v),h\right\rangle}{\|h\|^m} \quad (0<\theta(h)<1)\\
&=\liminf_{h\to 0} \frac{\sum_{k=1}^{\infty}\mu_k\left\langle J_m(v-v_k)-J_m(v+\theta(h)h-v_k),h\right\rangle}{\|h\|^m}\\
&\geq \liminf_{h\to 0} -L\theta(h)^{m-1}\geq -L>-\infty,
\end{aligned}$$

where the first equality is from the mean value theorem and the first inequality follows from Lemma 2.3 and the facts of $\mu_k>0$ and $\sum_{k=1}^{\infty}\mu_k=1$. In view of the definition of Hölder-smooth subdifferential ∂_m^{HS}, it follows that $-(m^{-1}\Delta_m)'(v)\in\partial_m^{HS}(-m^{-1}\Delta_m)(v)$.

By (3.8) and the assumption (3.1), there exist sequences $t_n\to 0+$ and $\|u_n\|\to 1$ such that

$$\lim_{n\to\infty}\frac{f(v+t_nu_n)-f(v)-t_n\left\langle\xi,u_n\right\rangle}{t_n^m}<-\delta=-mL\gamma^{-1}. \tag{3.9}$$

Since $X\setminus S$ is an open set as f is lower semicontinuous, we have $f(v+t_nu_n)>0$ for sufficiently large n. It follows from (3.5) that

$$\begin{aligned}
&\frac{f(v+t_nu_n)-f(v)-t_n\left\langle\xi,u_n\right\rangle}{t_n^m}\\
&=\frac{f(v+t_nu_n)-f(v)+m\gamma^{-1}t_n\sum_{k=1}^{\infty}\mu_k\left\langle J_m(v-v_k),u_n\right\rangle}{t_n^m}\\
&\geq\frac{\sum_{k=1}^{\infty}\mu_k\left\{\|v-v_k\|^m-\|v+t_nu_n-v_k\|^m\right\}+m\sum_{k=1}^{\infty}\mu_k\left\langle J_m(v-v_k),t_nu_n\right\rangle}{\gamma t_n^m}\\
&\geq m\gamma^{-1}t_n^{-m}\sum_{k=1}^{\infty}\mu_k\left\langle J_m(v-v_k)-J_m(v+t_nu_n-v_k),t_nu_n\right\rangle\\
&\geq -mL\gamma^{-1}\|u_n\|^m\to -mL\gamma^{-1}=-\delta\ (\text{ as } n\to\infty),
\end{aligned}$$

where the second inequality follows from Lemma 2.2 and the third inequality follows from Lemma 2.3. This contradicts (3.9). ∎

In view of the assumption (3.1), it is straightforward to see that if the $\partial_p^{HS}f(x)$ is replaced by a larger set such as $\partial^F f(x)$, $\partial^H f(x)$, or $\partial^{CR}f(x)$ (see (2.4)), then the condition becomes more stringent. In other words, our requirement on the subdifferential is fairly weak.

COROLLARY 3.2. *Let X be a uniformly smooth Banach space which has modulus of smoothness of power 2, and let $f:X\to\mathbb{R}\cup\{\infty\}$ be a proper lower semicontinuous function. If there exists $\delta>0$ such that for all $x\in f^{-1}(0,\infty)$ and all $\xi\in\partial^{LS}f(x)$,*

$$\liminf_{\|u\|\to 1,t\downarrow 0}\frac{f(x+tu)-f(x)-t\left\langle\xi,u\right\rangle}{t^2}<-\delta,$$

then

$$\operatorname{dist}^2(x,S)\leq(2L/\delta)\,f(x)_+\quad \textit{for all } x\in X.$$

Proof. Since $p = 2$, the Hölder subdifferential ∂_p^{HS} coincides with the Lipschitz-smooth subdifferential ∂^{LS}. The conclusion thus follows immediately from Theorem 3.1. ∎

Remark 3.1. Since all Hilbert spaces are uniformly smooth with modulus of smoothness of power 2 (see [9]) and since when X is a Hilbert space $\partial^{LS} f(x)$ coincides with the proximal subdifferential $\partial^P f(x)$, Corollary 3.2 generalizes Theorem 3.1 in [15] for its $\varepsilon = \infty$. Moreover, there exist Banach spaces, say $L^p(\mu)$ for $p \geq 2$, which are uniformly smooth with modulus of smoothness of power 2 but are not Hilbert spaces [9]. Therefore Corollary 3.2 is applicable to a broader class of spaces than [15, Theorem 3.1]. The same as what was done in [15], our results can also be verified for general $\varepsilon > 0$. We omit the details for brevity.

From the argument of Theorem 3.1, it can be seen that one can replace the Hölder smooth subdifferential ∂_m^{HS} of f by some other classes of subdifferentials. Let us define an abstract subdifferential in the following.

DEFINITION 3.3 (see [1]). *An abstract subdifferential, denoted by ∂, is any operator that associates a subset $\partial f(x) \subset X^*$ to a lower semicontinuous function $f : X \to \mathbb{R} \cup \{\infty\}$ and a point $x \in X$, satisfying the following properties:*

(P1) *$\partial f(x)$ coincides with the subdifferential in the sense of convex analysis whenever f is convex;*

(P2) *$0 \in \partial f(x)$ whenever $x \in \operatorname{dom} f$ is a local minimum of f;*

(P3) *$\partial(f + g)(x) \subset \partial f(x) + \partial g(x)$ whenever g is a real valued convex continuous function which satisfies $\partial g(x)$ and $\partial(-g)(x)$ are both nonempty.*

Paper [1] provides various classes of subdifferentials satisfying the above properties (P1)–(P3)—for example, the Hadamard subdifferential, the Gâteaux subdifferential, the Fréchet subdifferential, and the Clarke–Rockafellar subdifferential.

For $p > 1$, we denote by Γ_p all the functions of the form

$$\Gamma(x) := \frac{1}{p} \sum_{k=1}^{\infty} \mu_k \|x - u_k\|^p \quad \text{for all } x \in X, \tag{3.10}$$

where $\{u_k\}$ is any convergent sequence in X and $\{\mu_k\}$ is any sequence of nonnegative scalars satisfying $\sum_{k=1}^{\infty} \mu_k = 1$. Clearly, each function in Γ_p is a real valued continuous convex function.

THEOREM 3.4. *Let X be a uniformly smooth Banach space which has modulus of smoothness of power $m > 1$ and let $f : X \to \mathbb{R} \cup \{\infty\}$ be a proper lower semicontinuous function. Let ∂ be an abstract subdifferential satisfying properties* (P1)–(P3) *in Definition* 3.3 *and an additional property:*

(P4) *$\partial(-\Gamma)(x)$ is nonempty for each $\Gamma \in \Gamma_m$.*

If there exists $\delta > 0$ such that for all $x \in f^{-1}(0, \infty)$ and all $\xi \in \partial f(x)$,

$$\liminf_{\|u\| \to 1, t \downarrow 0} \frac{f(x + tu) - f(x) - t \langle \xi, u \rangle}{t^m} < -\delta,$$

then

$$\operatorname{dist}^m(x, S) \leq (mL/\delta)\, f(x)_+ \quad \text{for all } x \in X.$$

Proof. After checking the proof of Theorem 3.1, we know the key role played by the subdifferential is the part from (3.6) to (3.8). Since each $\Gamma(x)$ is a continuous real valued convex function, $\partial\Gamma(x)$ is nonempty. In view of the property (P4) and (P3),

one can establish (3.8) in a similar way. The remaining proof is similar to the proof of Theorem 3.1. □

The above theorems establish m-order error bounds for lower semicontinuous functions in certain classes of Banach spaces. As a corollary of Theorem 3.1, we give an error bound of order one whose proof is similar to that of Theorem 3.3 in [15]. Recall that S is the set $f^{-1}(-\infty, 0]$, and define

$$P(\alpha) := \{x \in X \setminus S : \partial_m^{HS} f(x) \cap B(0, \alpha) \neq \emptyset\} \quad \text{for } \alpha > 0. \tag{3.11}$$

THEOREM 3.5. *Let X be a uniformly smooth Banach space which has modulus of smoothness of power m for some $m > 1$ and let $f : X \to \mathbb{R} \cup \{\infty\}$ be a proper lower semicontinuous function. Suppose that the following two conditions hold.*

(i) *$P(\alpha) \subset f^{-1}(\beta, \infty)$ for some $\alpha > 0$ and some $\beta > 0$.*

(ii) *There exists $\delta > 0$ such that for all $x \in f^{-1}(\beta, \infty)$ and all $\xi \in \partial_m^{HS} f(x)$,*

$$\liminf_{\|u\| \to 1, t \downarrow 0} \frac{f(x + tu) - f(x) - t \langle \xi, u \rangle}{t^m} < -\delta.$$

Then there exists $c > 0$ such that

$$\operatorname{dist}(x, S) \leq c\, f(x)_+ \quad \text{for all } x \in X.$$

4. Error bounds in general Banach spaces. In the last section, we have established second-order sufficient conditions for error bounds of lower semicontinuous functions in smooth Banach spaces. In what follows we will provide different second-order sufficient conditions for error bounds in general Banach spaces. The result of this section generalizes that in [7], which gives second-order sufficient conditions for error bounds in general Banach spaces but requires the function to be Gâteaux differentiable. Our results show that the assumption of Gâteaux differentiability can be removed. Before that, we need to define second-order directional derivative. Let X be a Banach space and $f : X \to \mathbb{R} \cup \{\infty\}$ be a proper lower semicontinuous function. For $x, u, v \in X$, we define respectively Hadamard directional derivative and a second-order directional derivative:

$$f'_-(x; u) := \liminf_{v \to u,\, t \downarrow 0} \frac{f(x + tv) - f(x)}{t};$$

$$d^2_- f(x; u, v) := \liminf_{t \to 0+} \frac{f(x + tu + t^2 v) - f(x) - t f'(x) u}{t^2}$$

whenever f is Gâteaux differentiable.

It can be seen that if f is Gâteaux differentiable at x with $f'(x)$ being the Gâteaux derivative, then $f'_-(x; u) \leq f'(x)u$ for every u; the equality holds if in addition f is locally Lipschitz at x. If f is twice continuously differentiable, then

$$d^2_- f(x; u, 0) = (1/2)\nabla^2 f(x)(u, u),$$

where $\nabla^2 f(x)$ denotes the second-order derivative of f at x.

For $\varepsilon > 0$, we define a set

$$D(\varepsilon) := \left\{x \in X : x \notin S \text{ and } \inf_{\|u\|=1} f'_-(x; u) \geq -\varepsilon\right\}.$$

If f is Gâteaux differentiable on X, then

$$D(\varepsilon) \subset \{x \in X : x \notin S \text{ and } \|f'(x)\| \leq \varepsilon\} =: \mathfrak{D}(\varepsilon),$$

where the set $\mathfrak{D}(\varepsilon)$ is introduced and used in [7] for studying second-order sufficient conditions for continuous and Gâteaux differentiable functions to have error bounds.

The following lemma [12, Lemma 2.3] is a straightforward consequence of Theorem 2(ii) in [6].

LEMMA 4.1. *Let X be a Banach space and $f : X \to \mathbb{R} \cup \{\infty\}$ be a proper lower semicontinuous function. If there exists $\gamma > 0$ such that for every $x \in f^{-1}(0, \infty)$ there is $y \in f^{-1}[0, \infty)$ such that*

$$f(x) - f(y) \geq \gamma \|x - y\| > 0,$$

then $\operatorname{dist}(x, S) \leq \gamma^{-1} f(x)_+$ *for all* $x \in X$.

THEOREM 4.2. *Let $f : X \to \mathbb{R}$ be a continuous function. Suppose that there exist positive scalars r, ρ, and δ such that the following conditions hold:*

(i) $D(\rho) \subset f^{-1}(r, \infty)$;

(ii) $\limsup_{t \to 0+} \sup_{x \in D(\rho)} \inf_{\|u\|=1} \frac{f(x+tu) - f(x) + t f'_-(x; -u)}{t^2} < -\delta$.

Then there exists $\gamma > 0$ such that $\operatorname{dist}(x, S) \leq \gamma^{-1} f(x)_+$ *for all* $x \in X$.

Proof. We need to consider only those points x not in S. In view of the assumption (ii), there exists $\beta \in (0, 1/2]$ such that for every $t \in (0, \beta)$ and every $x \in D(\rho)$, a unit vector u (dependent on t and x) exists and satisfies that

$$\frac{f(x + tu) - f(x) + t f'_-(x; -u)}{t^2} < -\delta. \tag{4.1}$$

Take $\varepsilon = \min\{\rho, \beta\delta/4\}$ and $\gamma = \min\{r, \varepsilon/2\}$.

Let $x \in D(\varepsilon)$ be such that $\operatorname{dist}(x, S) \geq 1$. Put $\lambda = \beta/2$. It follows that $x + \lambda u \notin S$ for any unit vector u. Since $\varepsilon \leq \rho$, $x \in D(\varepsilon) \subset D(\rho)$, it follows from (4.1) that there exists a unit vector u_λ such that

$$f(x + \lambda u_\lambda) - f(x) + \lambda f'_-(x; -u_\lambda) < -\lambda^2 \delta.$$

In view of the definition of $D(\varepsilon)$, $x \in D(\varepsilon)$ implies that $f'_-(x; -u_\lambda) \geq -\varepsilon$. Therefore,

$$f(x) - f(x + \lambda u_\lambda) \geq \lambda^2 \delta - \lambda\varepsilon \geq \gamma\lambda = \gamma \|x - (x + \lambda u_\lambda)\|.$$

For $x \in D(\varepsilon)$ and $\operatorname{dist}(x, S) < 1$, there exists $y \in S$ such that $\|x - y\| < 1$. Since f is continuous, y can be chosen to satisfy $f(y) = 0$. Since $x \in D(\varepsilon)$ and $D(\varepsilon) \subset f^{-1}(r, \infty)$, one has $f(x) > r$. It follows that

$$f(x) - f(y) \geq r > r \|x - y\| \geq \gamma \|x - y\| > 0.$$

For $x \notin D(\varepsilon)$, we have $f'_-(x; u) < -\varepsilon$ for some unit vector u. It follows that there exist a sequence of positive scalars $\{t_n\}$ converging to zero and a sequence $\{u_n\}$ converging to u such that for sufficiently large n,

$$f(x + t_n u_n) - f(x) < -\varepsilon t_n.$$

Since $\gamma < \varepsilon$ and $\|u_n\| \to 1$, $\gamma \|u_n\| < \varepsilon$ for large enough n. This implies that

$$f(x) - f(x + t_n u_n) > \varepsilon t_n \geq \gamma t_n \|u_n\| = \gamma \|x - (x + t_n u_n)\|$$

for sufficiently large n.

Thus, we have shown that for each $x \notin S$, there exists $y \in f^{-1}[0,\infty)$ such that $f(x) - f(y) \geq \gamma \|x - y\|$. Then, by applying Lemma 4.1, we obtain the desired conclusion. □

Huang and Ng [7] considered error bounds in general Banach spaces for a function which is Gâteaux differentiable and continuous. Besides the assumption (i) of Theorem 4.2, [7] requires another condition: There exist $\beta > 0$ and $\delta > 0$ such that for all $x \in \mathfrak{D}(\rho)$,

$$\inf_{\|u\|=1} \sup_{t\in[0,\beta)} d^2_- f(x + tu; u, 0) < -\delta. \tag{4.2}$$

Because f is Gâteaux differentiable and continuous, $f'_-(x; -u) \leq -f'(x)u$. It follows from [7, Theorem 3.1] that the condition (4.2) implies the existence of $\beta > 0$ and $\delta > 0$ such that for all $x \in \mathfrak{D}(\rho)$,

$$\inf_{\|u\|=1} \sup_{t\in(0,\beta)} \frac{f(x + tu) - f(x) + tf'_-(x; -u)}{t^2} < -\delta. \tag{4.3}$$

Note that our assumption (ii) in Theorem 4.2 is that there exist $\beta > 0$ and $\delta > 0$ such that for all $x \in D(\rho)$,

$$\sup_{t\in(0,\beta)} \inf_{\|u\|=1} \frac{f(x + tu) - f(x) + tf'_-(x; -u)}{t^2} < -\delta. \tag{4.4}$$

Since $D(\rho) \subset \mathfrak{D}(\rho)$, it is straightforward that (4.3) and hence (4.2) imply (4.4). The latter is a restatement of our assumption (ii), which is therefore less restrictive than the assumption (4.2) as our assumption (ii) also allows f to be non-Gâteaux differentiable.

REFERENCES

[1] D. Aussel, J.-N. Corvellec, and M. Lassonde, *Mean value property and subdifferential criteria for lower semicontinuous functions*, Trans. Amer. Math. Soc., 347 (1995), pp. 4147–4161.

[2] J. M. Borwein and D. Preiss, *A smooth variational principle with applications to subdifferentiability and to differentiability of convex functions*, Trans. Amer. Math. Soc., 303 (1987), pp. 517–527.

[3] F. H. Clarke, Y. S. Ledyaev, R. J. Stern, and P. R. Wolenski, *Nonsmooth Analysis and Control Theory*, Springer-Verlag, New York, 1998.

[4] K. Deimling, *Nonlinear Functional Analysis*, Springer-Verlag, Berlin, 1985.

[5] F. Facchinei and J.-S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Springer-Verlag, New York, 2003.

[6] A. Hamel, *Remarks to an equivalent formulation of Ekeland's variational principle*, Optimization, 31 (1994), pp. 233–238.

[7] L. R. Huang and K. F. Ng, *On first- and second-order conditions for error bounds*, SIAM J. Optim., 14 (2004), pp. 1057–1073.

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

[9] J. Lindenstrauss and L. Tzafriri, *Classical Banach Spaces*, Vol. II, Ergeb. Math. Grenzgeb. 97, Springer-Verlag, Berlin, 1979.

[10] K. F. Ng and W. H. Yang, *Error bounds for abstract linear inequality systems*, SIAM J. Optim., 13 (2002), pp. 24–43.

[11] K. F. Ng and X. Y. Zheng, *Global error bounds with fractional exponents*, Math. Program., 88 (2000), pp. 357–370.

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

[13] Z. Wu and J. J. Ye, *Sufficient conditions for error bounds*, SIAM J. Optim., 12 (2001), pp. 421–435.
[14] Z. Wu and J. J. Ye, *On error bounds for lower semicontinuous functions*, Math. Program., 92 (2002), pp. 301–314.
[15] Z. Wu and J. J. Ye, *First-order and second-order conditions for error bounds*, SIAM J. Optim., 14 (2003), pp. 621–645.
[16] Z. B. Xu and G. F. Roach, *Characteristic inequalities of uniformly convex and uniformly smooth Banach spaces*, J. Math. Anal. Appl., 157 (1991), pp. 189–210.

SIAM J. OPTIM.
Vol. 17, No. 3, pp. 806–821
© 2006 Society for Industrial and Applied Mathematics

ALL LINEAR AND INTEGER PROGRAMS ARE SLIM 3-WAY TRANSPORTATION PROGRAMS*

JESÚS A. DE LOERA† AND SHMUEL ONN‡

Abstract. We show that any rational convex polytope is polynomial-time representable as a 3-way line-sum transportation polytope of "slim" $(r, c, 3)$ format. This universality theorem has important consequences for linear and integer programming and for confidential statistical data disclosure. We provide a polynomial-time embedding of arbitrary linear programs and integer programs in such slim transportation programs and in bitransportation programs. Our construction resolves several standing problems on 3-way transportation polytopes. For example, it demonstrates that, unlike the case of 2-way contingency tables, the range of values an entry can attain in any slim 3-way contingency table with specified 2-margins can contain arbitrary gaps. Our smallest such example has format $(6, 4, 3)$. Our construction provides a powerful automatic tool for studying concrete questions about transportation polytopes and contingency tables. For example, it automatically provides new proofs for some classical results, including a well-known "real-feasible but integer-infeasible" $(6, 4, 3)$-transportation polytope of M. Vlach, and bitransportation programs where any feasible bitransportation must have an arbitrarily large prescribed denominator.

Key words. integer programming, linear programming, combinatorial optimization, convex polytopes, transportation problems, multicommodity flows, strongly polynomial time, contingency tables, multiway table, statistical table, data security, privacy, approximation algorithms, Markov basis, toric ideal, cofinetiality, disclosure

AMS subject classifications. 90C11, 52B55, 90B06, 68R05, 15A39, 62H17

DOI. 10.1137/040610623

1. Introduction.

Transportation polytopes, their integer points (called contingency tables by statisticians), and their projections have been used and studied extensively in the operations research and mathematical programming literature (see, e.g., [1, 2, 5, 17, 20, 23, 24, 29, 30] and references therein) and in the context of secure statistical data management by agencies such as the U.S. Census Bureau [28] (see, e.g., [3, 4, 9, 10, 13, 18, 22] and references therein).

We start right away with the statement of the main theorem of this article. Its proof will be the subject of section 3. Some of the many implications of the main theorem for linear and integer programming, combinatorial optimization, and confidential statistical data disclosure will be discussed in section 2. The consequences include the solution of several long-standing open questions stated by Vlach in 1986 [29]. Following a common convention we denote by $\mathbb{R}_{\geq 0}$ the nonnegative reals. In what follows, a 3-way transportation polytope is *slim* if one of its dimensions has depth three.

*Received by the editors June 27, 2004; accepted for publication (in revised form) April 12, 2006; published electronically October 4, 2006.

http://www.siam.org/journals/siopt/17-3/61062.html

†University of California at Davis, Davis, CA 95616 (deloera@math.ucdavis.edu, http://www.math.ucdavis.edu/∼deloera). The work of this author was supported in part by NSF grant 0309694, a 2003 UC-Davis Chancellor's fellow award, the Alexander von Humboldt foundation, and the American Institute of Mathematics.

‡Technion–Israel Institute of Technology, 32000 Haifa, Israel (onn@ie.technion.ac.il, http://ie.technion.ac.il/∼onn). The work of this author was supported in part by a grant from the Israel Science Foundation, the American Institute of Mathematics, the Technion President Fund, and the Jewish Communities of Germany Research Fund.

THEOREM 1.1. *Any rational polytope* $P = \{y \in \mathbb{R}^n_{\geq 0} : Ay = b\}$ *is polynomial-time representable as a slim* 3*-way transportation polytope:*

$$T = \left\{ x \in \mathbb{R}^{r\times c\times 3}_{\geq 0} \ : \ \sum_i x_{i,j,k} = w_{j,k}\,,\ \sum_j x_{i,j,k} = v_{i,k}\,,\ \sum_k x_{i,j,k} = u_{i,j} \right\}.$$

By saying that a polytope $P \subset \mathbb{R}^p$ is *representable* as a polytope $Q \subset \mathbb{R}^q$ we mean in the strong sense that there is an injection $\sigma : \{1,\ldots,p\} \longrightarrow \{1,\ldots,q\}$ such that the coordinate-erasing projection

$$\pi : \mathbb{R}^q \longrightarrow \mathbb{R}^p \ : \ x = (x_1,\ldots,x_q) \mapsto \pi(x) = (x_{\sigma(1)},\ldots,x_{\sigma(p)})$$

provides a bijection between Q and P and between the sets of integer points $Q\cap\mathbb{Z}^q$ and $P\cap\mathbb{Z}^p$. In particular, if P is representable as Q, then P and Q are isomorphic in any reasonable sense: They are linearly equivalent, and hence all linear programming related problems over the two are polynomial-time equivalent; they are combinatorially equivalent and hence have the same facial structure; and they are integer equivalent, and therefore all integer programming and integer counting related problems over the two are polynomial-time equivalent as well. The polytope T in the theorem is a 3-way transportation polytope with specified *line-sums* $(u_{i,j}), (v_{i,k}), (w_{j,k})$ (2*-margins* in the statistical context to be elaborated upon below). The arrays in T are of size $(r, c, 3)$; that is, they have r rows, c columns, and "slim" depth 3, which is the best possible: 3-way line-sum transportation polytopes of depth ≤ 2 are equivalent to ordinary 2-way transportation polytopes which are not universal.

An appealing feature of Theorem 1.1 is that the defining system of T has only $\{0,1\}$-valued coefficients and depends only on r and c. Thus, every rational polytope has a representation by one such system, where all information enters through the right-hand side $(u_{i,j}), (v_{i,k}), (w_{j,k})$.

We have also proved a second universality theorem about the following *bitransportation problems*: Given *supply* vectors $s^1, s^2 \in \mathbb{R}^r_{\geq 0}$, *demand* vectors $d^1, d^2 \in \mathbb{R}^c_{\geq 0}$, and *capacity* matrix $u \in \mathbb{R}^{r\times c}_{\geq 0}$, find a pair of nonnegative "transportations" $x^1, x^2 \in \mathbb{R}^{r\times c}_{\geq 0}$ satisfying supply and demand requirements $\sum_j x^k_{i,j} = s^k_i$, $\sum_i x^k_{i,j} = d^k_j$, $k = 1, 2$, and capacity constraints $x^1_{i,j} + x^2_{i,j} \leq u_{i,j}$. In other words, find $x^1, x^2 \geq 0$ such that x^k has row-sum s^k and column-sum d^k for $k = 1, 2$, and $x^1 + x^2 \leq u$.

THEOREM 1.2. *Any rational polytope* $P = \{y \in \mathbb{R}^n_{\geq 0} : Ay = b\}$ *is polynomial-time representable as a bitransportation polytope*

$$F = \left\{ (x^1, x^2) \in \mathbb{R}^{r\times c}_{\geq 0} \oplus \mathbb{R}^{r\times c}_{\geq 0} \ : \ x^1_{i,j} + x^2_{i,j} \leq u_{i,j}\,, \right.$$

$$\left. \sum_j x^k_{i,j} = s^k_i\,,\ \sum_i x^k_{i,j} = d^k_j\,,\quad k = 1,2 \right\}.$$

The proof is an easy adjustment of part of the proof of Theorem 1.1 (i.e., Theorem 3.3) and is presented in section 3.5. The theorem remains valid if we take all supplies to have the same value $s^k_i = U$, $i = 1,\ldots,r$, $k = 1, 2$; further, all capacities $u_{i,j}$ can be taken to be $\{0, U\}$-valued, giving a stronger statement.

The bitransportation problem gives at once a very simple two-commodity flow network as follows: start with the directed bipartite graph with vertex set $I \uplus J$,

$|I| = r$, $|J| = c$, and arc set $I \times J$ with capacities $u_{i,j}$, and augment it with two sources a_1, a_2 and two sinks b_1, b_2 and with arcs (a_k, i), $i \in I$, (j, b_k), $j \in J$, $k = 1, 2$ with capacities $u(a_k, i) := s_i^k$, $u(j, b_k) := d_j^k$. The feasible bitransportations are then precisely the two-commodity flows of maximal total value. This implies a result first obtained by A. Itai [19]: every linear program is polynomially equivalent to a two-commodity flow problem. It is worth noting that our transformation is in fact much simpler than Itai's. In particular, the above network is exceedingly special: every dipath has length three and is of the form (a_k, i, j, b_k) for some $k \in \{1, 2\}$, $i \in I$, and $j \in J$ and involves only one "interesting" arc ij. Further, each such arc ij carries flow of each commodity on precisely one path.

To demonstrate the concrete nature of our transformations, the procedures that convert any given data A, b to data to the representations of Theorems 1.1 and 1.2 have been implemented in a computer program which is available on-line (see [27]).

2. The consequences of the main results. We now discuss some consequences of Theorems 1.1 and 1.2. A few of them were first presented in [7].

2.1. Universality of transportation polytopes: Solution of Vlach's problems. As mentioned above, there is a large body of literature on the structure of various transportation polytopes. In particular, in the comprehensive paper [29], M. Vlach surveys some ten families of necessary conditions published over the years by several authors (including Schell, Haley, Smith, Morávek, and Vlach) on the line-sums $(u_{i,j}), (v_{i,j}), (w_{i,j})$ for a transportation polytope to be nonempty, and raises several concrete problems regarding these polytopes. Specifically, [29, Problems 4, 7, 9, 10] ask about the sufficiency of some of these conditions. Our results say that transportation polytopes (in fact already of slim, $(r, c, 3)$, arrays) are universal and include all polytopes. This indicated that the answer to each of Problems 4, 7, 9, and 10 has to be negative. Indeed we have already verified this.

Example 2.1 (Smith II conditions are not sufficient). Using our encoding, in particular, applying the algorithm of Theorem 3.2 to the infeasible polyhedron $P = \{(x, y) : x + y = 1, x + y = 2,\ x, y \geq 0\}$, with 2 as an upper bound on its entry values, we obtained concrete 2-margins (below). These 2-margins satisfy conditions (8.1)–(8.3) on page 72 of [29] while giving an infeasible system; thus the example solves open problem 7 in [29]. Note for reference that for the given matrices the top-left corners are the margin values $u_{1,1}, v_{1,1}, w_{1,1}$.

$$\begin{pmatrix} 2 & 0 & 0 & 0 & 2 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 & 2 & 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 & 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 2 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 & 0 & 2 & 0 & 0 & 0 & 0 & 2 \\ 0 & 2 & 0 & 0 & 0 & 0 & 2 & 2 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 & 2 & 0 & 0 & 2 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 2 \end{pmatrix}, \quad \begin{pmatrix} 2 & 2 & 2 \\ 2 & 0 & 2 \\ 2 & 2 & 2 \\ 2 & 0 & 2 \\ 2 & 0 & 2 \\ 2 & 2 & 2 \\ 2 & 0 & 2 \\ 2 & 2 & 2 \\ 2 & 2 & 2 \\ 2 & 2 & 2 \\ 2 & 2 & 2 \end{pmatrix}, \quad \begin{pmatrix} 2 & 2 & 0 \\ 2 & 2 & 0 \\ 2 & 2 & 0 \\ 2 & 2 & 0 \\ 7 & 0 & 1 \\ 6 & 0 & 2 \\ 3 & 0 & 5 \\ 0 & 2 & 4 \\ 0 & 2 & 4 \\ 0 & 2 & 4 \\ 0 & 2 & 4 \end{pmatrix}.$$

Similarly, Problem 12 on page 76 of [29] asks whether all dimensions can occur as that of a suitable transportation polytope: the affirmative answer, given very recently in

[15], follows also at once from our universality result. Our construction also provides a powerful tool for studying concrete questions about transportation polytopes and their integer points, by allowing us to write down simple systems of equations that encode desired situations and lifting them up. Here is an example to this effect.

Example 2.2 (Vlach's rational-nonempty integer-empty transportation). Using our construction, we automatically recover the smallest known example, first discovered by Vlach [29], of a rational-nonempty integer-empty transportation polytope, as follows. We start with the polytope $P = \{y \geq 0 : 2y = 1\}$ in one variable, containing a (single) rational point but no integer point. Our construction represents it as a transportation polytope T of $(6, 4, 3)$-arrays with line-sums given by the three matrices below; by Theorem 1.1, T is integer equivalent to P and hence also contains a (single) rational point but no integer point.

$$\begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}.$$

Returning to the Vlach problems, [29, Problem 13] asks for a characterization of those line-sums margins that guarantee an integer point in a 3-way transportation polytope T. In [18], Irving and Jerrum showed that deciding whether $T \cap \mathbb{Z}^{r \times c \times h} \neq \emptyset$ is NP-complete, and hence an efficient such characterization cannot exist unless $NP = coNP$. An immediate corollary of Theorem 1.1 strengthens this result to hold for slim arrays:

COROLLARY 2.3. *Deciding if a slim,* $(r, c, 3)$*, transportation polytope has an integer point is NP-complete.*

A comprehensive complexity classification of this decision problem under various assumptions on the array size and on the input, as well as of the related lattice point counting problem and other variants, appeared in [6].

The last Vlach problem [29, Problem 14] asks whether there is a *strongly polynomial-time* algorithm for deciding the (real) feasibility $T \neq \emptyset$ of a transportation polytope. Since the system defining T is $\{0, 1\}$-valued, the results of Tardos [26] provide an affirmative answer. However, the existence of a strongly polynomial-time algorithm for linear programming in general is open and of central importance; our construction embeds any linear program in an $(r, c, 3)$ transportation program in polynomial-time, but unfortunately this process is *not* strongly polynomial. Nonetheless, our construction may shed some light on the problem and may turn out useful in sharpening the boundary (if any) between strongly and weakly polynomial-time solvable linear programs.

2.2. Universality for approximations. The representation manifested by Theorem 1.1 allows us to represent an arbitrary integer programming problem $\min\{cy : y \in \mathbb{N}^n, Ay = b\}$ as a problem of finding minimum cost integer transportation,

$$\min \left\{ \sum_{i,j,k} p_{i,j,k} x_{i,j,k} \ : \ x \in \mathbb{N}^{r \times c \times 3}, \ \sum_i x_{i,j,k} = w_{j,k}, \right.$$

$$\left. \sum_j x_{i,j,k} = v_{i,k}, \ \sum_k x_{i,j,k} = u_{i,j} \right\},$$

by simply extending the cost vector c by zeros to a cost array p. In particular, the feasible (integer) solutions y to the original problem are in *cost-preserving* bijection with the feasible (integer) transportations x (that is, $cy = px$ for any corresponding pair). This shows that the representation preserves *approximations*, and that minimum cost transportation problems of slim format $(r, c, 3)$ are universal for approximation as well. In particular, any nonapproximability result—say, for the maximum clique problem [14]—lifts at once to the slim minimum cost transportation problem: just start with an integer programming formulation of the maximum clique problem with $\{0, 1\}$-valued right-hand-side vector b, and lift it up. We get the following hardness-of-approximation result.

COROLLARY 2.4. *Under the assumption $P \neq NP$, there is an $\epsilon > 0$ such that there is no polynomial-time $(rc)^{\epsilon}$-approximation algorithm for the minimum cost slim $(r, c, 3)$ line-sum transportation problem.*

We do not attempt here to provide the largest possible ϵ. Note, of course, that in particular, unless $P = NP$, there is no constant ratio approximation for the 3-way transportation problem (the problem is not in the class APX).

2.3. Confidential statistical data disclosure: Entry-range. Next, we briefly discuss some of the applications to statistical model theory: a comprehensive treatment can be found in [8]. A central goal of statistical data management by agencies such as the U.S. Census Bureau is to allow public access to information on their data base while protecting confidentiality of individuals whose data is in the base. A common practice [10], taken in particular by the Bureau [28], is to allow the release of some margins of tables in the base but not the individual entries themselves. The security of an entry is closely related to the range of values it can attain in any table with the fixed released collection of margins: if the range is "simple," then the entry may be exposed, whereas if it is "complex" the entry may be assumed secure.

In this subsection only, we use the following notation, which is common in statistical applications. A *d-table of size $n = (n_1, \ldots, n_d)$* is an array of nonnegative integers $x = (x_{i_1,\ldots,i_d})$, $1 \leq i_j \leq n_j$. For any $0 \leq k \leq d$ and any k-subset $J \subseteq \{1, \ldots, d\}$, the *$k$-margin* of x corresponding to J is the k-table $x^J := (x^J_{i_j : j \in J}) := (\sum_{i_j : j \notin J} x_{i_1,\ldots,i_d})$ obtained by summing the entries over all indices *not in* J. For instance, the 2-margins of a 3-table $x = (x_{i_1,i_2,i_3})$ are its *line-sums* x^{12}, x^{13}, x^{23} such as $x^{13} = (x^{13}_{i_1,i_3}) = (\sum_{i_2} x_{i_1,i_2,i_3})$, and its 1-margins are its *plane-sums* x^1, x^2, x^3 such as $x^2 = (x^2_{i_2}) = (\sum_{i_1,i_3} x_{i_1,i_2,i_3})$.

A *statistical model* is a triple $\mathcal{M} = (d, \mathcal{J}, n)$, where $\mathcal{J}$ is a set of subsets of $\{1, \ldots, d\}$ none containing the other and $n = (n_1, \ldots, n_d)$ is a tuple of positive integers. The model dictates the collection of margins for d-tables of size n to be specified. Our results concern the models $(3, \{12, 13, 23\}, (r, c, 3))$, that is, slim, $(r, c, 3)$-tables, with all three of their 2-margins specified.

For any model $\mathcal{M} = (d, \mathcal{J}, n)$ and any specified collection of margins $u = (u^J : J \in \mathcal{J})$ under the model $\mathcal{M}$, the corresponding set of *contingency tables* with collection of margins u is

$$C(\mathcal{M}; u) := \{\, x \in \mathbb{N}^{n_1 \times \cdots \times n_d} \;:\; x^J = u^J \,,\; J \in \mathcal{J} \,\} \,.$$

Clearly, this set is precisely the set of integer points in the corresponding transportation polyhedron.

Finally, we define entry-ranges. Permuting coordinates, we may always consider the first entry $x_{\mathbf{1}}$, where $\mathbf{1} := (1, \ldots, 1)$. The *entry-range* of a collection of margins

u under a model $\mathcal{M}$ is the set $R(\mathcal{M};u) := \{x_{\mathbf{1}} : x \in C(\mathcal{M};u)\} \subset \mathbb{N}$ of values $x_{\mathbf{1}}$ can attain in any table with these margins.

Often, the entry-range is an interval and hence "simple" and vulnerable, that is, for some $a, b \in \mathbb{N}$, $R(\mathcal{M};u) = \{r \in \mathbb{N} : a \le r \le b\}$. For instance, as shown in [8], this indeed is the case for any 1-margin model $\mathcal{M} = (d, \{1,2,\dots,d\}, (n_1,\dots,n_d))$ and any collection of margins $u = (u^1,\dots,u^d)$ under $\mathcal{M}$.

In striking contrast with this situation and with recent attempts by statisticians to better understand entry behavior of slim 3-tables (cf. [3, 4, 10]), we have the following surprising consequence of Theorem 1.1, implying that entry-ranges of 2-margined slim 3-table models consist of all finite sets of nonnegative integers and hence are "complex" and presumably secure. For the proof, see [8].

COROLLARY 2.5 (universality of entry-range). *For any finite set $D \subset \mathbb{N}$ of nonnegative integers, there are $r, c,$ and 2-margins for $(r, c, 3)$-tables such that the set of values occurring in a fixed entry in all possible tables with these margins is precisely D.*

Example 2.6 (Gap in entry-range of 2-margined 3-tables). Applying our automatic universal generator [27] to the polytope $P = \{y \ge 0 : y_0 - 2y_1 = 0,\ y_1 + y_2 = 1\}$ in three variables, we obtain the following 2-margins for $(16, 11, 3)$-tables giving entry-range $D = \{0, 2\}$,

$$\begin{pmatrix}
2&0&0&0&0&0&2&0&0&2&0&0&0&0&0&0\\
0&0&0&0&0&2&0&0&0&0&0&0&0&0&0&2\\
2&0&0&0&0&0&0&0&0&0&2&0&0&0&0&2\\
2&2&2&2&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&2&2&2&2&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&2&2&2&2&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&2&2&2&2\\
2&0&0&0&2&0&0&0&2&0&0&0&2&0&0&0\\
0&2&0&0&0&2&0&0&0&2&0&0&0&2&0&0\\
0&0&2&0&0&0&2&0&0&0&2&0&0&0&2&0\\
0&0&0&2&0&0&0&2&0&0&0&2&0&0&0&2
\end{pmatrix},$$

$$\begin{pmatrix}
2&2&2&2&2&2&2&2&2&2&2&2&2&2&2&2\\
4&0&0&0&0&2&2&0&0&2&2&0&0&0&0&4\\
2&2&2&2&2&2&2&2&2&2&2&2&2&2&2&2
\end{pmatrix},$$

$$\begin{pmatrix}
4&1&3&6&6&6&6&0&0&0&0\\
2&3&3&0&0&0&0&2&2&2&2\\
0&0&0&2&2&2&2&6&6&6&6
\end{pmatrix};$$

with a suitable "human" short cut it is possible to get it down to the following (possibly smallest) collection of margins for $(6, 4, 3)$-tables, giving again the entry-range $D = \{0, 2\}$ with a gap,

$$\begin{pmatrix}
2&1&2&0&2&0\\
1&0&2&0&0&2\\
1&0&0&2&2&0\\
0&1&0&2&0&2
\end{pmatrix},\quad
\begin{pmatrix}
2&1&2&3&0&0\\
2&1&0&0&2&1\\
0&0&2&1&2&3
\end{pmatrix},\quad
\begin{pmatrix}
2&3&2\\
2&1&2\\
2&1&2\\
2&1&2
\end{pmatrix}.$$

Further applications of Theorem 1.1 to statistical model theory are discussed in [8]; these include important consequences for *Markov bases* of 2-margined slim 3-way models. (Recall that a Markov basis is a set of moves that connects any pair of tables in the model that have the same set of margins, and is needed for the design of a random walk on the space of tables with fixed margins to address the problems of *sampling* and *estimating* various statistics on this space; see [8] for more details.)

2.4. Universality of the bitransportation problem. Our construction for Theorem 1.2 allows automatic generation of bitransportation programs with integer supplies, demands and capacities, where any feasible bitransportation must have an arbitrarily large prescribed denominator, in contrast with Hu's celebrated half-integrality theorem for the undirected case [16].

Example 2.7 (Bitransportations with arbitrarily large denominator). Fix any positive integer q. Start with the polytope $P = \{y \geq 0 : qy = 1\}$ in one variable containing the single point $y = \frac{1}{q}$. Our construction represents it as a bitransportation polytope F with integer supplies, demands and capacities, where y is embedded as the transportation $x^1_{1,1}$ of the first commodity from supply vertex $1 \in I$ to demand vertex $j \in J$. By Theorem 1.2, F contains a single bitransportation with $x^1_{1,1} = y = \frac{1}{q}$. For instance, for $q = 3$ we get the bitransportation problem with the data

$$u = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{pmatrix}, \quad s^1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \quad s^2 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix},$$

$$d^1 = (1 \;\; 1 \;\; 1 \;\; 1 \;\; 1 \;\; 1 \;\; 0 \;\; 0), \quad d^2 = (0 \;\; 0 \;\; 0 \;\; 1 \;\; 1 \;\; 1 \;\; 2 \;\; 1),$$

which has the following unique, $\{0, \frac{1}{3}, \frac{2}{3}\}$-valued, bitransportation solution:

$$x^1 = \frac{1}{3}\begin{pmatrix} 1 & 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 2 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}, \quad x^2 = \frac{1}{3}\begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 & 1 \end{pmatrix}.$$

By Theorem 1.2, any (say, feasibility) linear programming problem can be encoded as such a bitransportation problem (unbounded programs can also be treated by adding to the original system a single equality $\sum_{j=0}^{n} y_j = U$ with y_0 a new "slack" variable and U derived from the Cramer's rule bound of Theorem 10.3 [25]). Thus, any (hopefully combinatorial) algorithm for the bitransportation problem will give an algorithm for general linear programming. There has been much interest lately (A. Levin [21]) in combinatorial approximation algorithms for (fractional) multiflows, e.g., [11, 12]; these yield, via Theorem 1.2, approximation algorithms for general linear programming, which might prove a useful and fast solution strategy in practice. Details of this will appear elsewhere.

3. The three-stage construction. Our construction consists of three stages which are independent of each other as reflected in Lemma 3.1 and Theorems 3.2

and 3.3 below. Stage one, in section 3.1, is a simple preprocessing based on standard scaling ideas, in which a given polytope is represented as another whose defining system involves only small, $\{-1,0,1,2\}$-valued coefficients, at the expense of increasing the number of variables. This enables us to make the entire construction run in time polynomial in the size of the input. However, for systems with small coefficients, such as in the examples above, this may result in unnecessary blow-up and can be skipped. Stage two, in section 3.2, represents any rational polytope as a 3-way transportation polytope with specified *plane-sums* and *forbidden-entries*. In the last stage, in section 3.3, any plane-sum transportation polytope with upper-bounds on the entries is represented as a slim 3-way line-sum transportation polytope. In section 3.4 these three stages are integrated to give Theorem 1.1, and a complexity estimate is provided to close the presentation. Theorem 1.2 is a result of an easy modification of Theorem 3.2, and it is the content of section 3.5.

3.1. Preprocessing: Coefficient reduction. Let $P = \{y \geq 0 : Ay = b\}$ where $A = (a_{i,j})$ is an integer matrix and b is an integer vector. We represent it as a polytope $Q = \{x \geq 0 : Cx = d\}$, in polynomial-time, with a $\{-1,0,1,2\}$-valued matrix $C = (c_{i,j})$ of coefficients, as follows. Consider any variable y_j and let $k_j := \max\{\lfloor \log_2 |a_{i,j}| \rfloor : i = 1,\ldots,m\}$ be the maximum number of bits in the binary representation of the absolute value of any $a_{i,j}$. We introduce variables $x_{j,0},\ldots,x_{j,k_j}$, and relate them by the equations $2x_{j,s} - x_{j,s+1} = 0$. The representing injection σ is defined by $\sigma(j) := (j,0)$, embedding y_j as $x_{j,0}$. Consider any term $a_{i,j}\,y_j$ of the original system. Using the binary expansion $|a_{i,j}| = \sum_{s=0}^{k_j} t_s 2^s$ with all $t_s \in \{0,1\}$, we rewrite this term as $\pm \sum_{s=0}^{k_j} t_s x_{j,s}$. To illustrate, consider a system consisting of the single equation $3y_1 - 5y_2 + 2y_3 = 7$. Then the new system is

$$
\begin{array}{rrrrrrrrr}
2x_{1,0} & -x_{1,1} & & & & & & = & 0,\\
 & & 2x_{2,0} & -x_{2,1} & & & & = & 0,\\
 & & & 2x_{2,1} & -x_{2,2} & & & = & 0,\\
 & & & & & 2x_{3,0} & -x_{3,1} & = & 0,\\
x_{1,0} & +x_{1,1} & -x_{2,0} & & -x_{2,2} & & +x_{3,1} & = & 7.
\end{array}
$$

It is easy to see that this procedure provides the sought representation, and we get the following.

LEMMA 3.1. *Any rational polytope $P = \{y \geq 0 : Ay = b\}$ is polynomial-time representable as a polytope $Q = \{x \geq 0 : Cx = d\}$ with $\{-1,0,1,2\}$-valued defining matrix C.*

3.2. Representing polytopes as plane-sum entry-forbidden transportation polytopes. The next stage of construction we are about to explain will normally be applied to the output $Q = \{x \geq 0 : Cx = d\}$ of stage one, but we present the construction for a general polyhedron P since the construction holds in that generality. Let $P = \{y \geq 0 : Ay = b\}$, where $A = (a_{i,j})$ is an $m \times n$ integer matrix and b is an integer vector: we assume that P is bounded and hence a (possibly empty) polytope, with an integer upper bound U on the value of any coordinate y_j of any $y \in P$ (U can be derived efficiently from Cramer's rule as explained in Theorem 10.3 of [25]).

For each variable y_j, let r_j be the maximum of the sum of the positive coefficients of y_j over all equations and the sum of absolute values of the negative coefficients of y_j over all equations:

$$r_j := \max\left(\sum_k \{a_{k,j} : a_{k,j} > 0\}\,,\ \sum_k \{|a_{k,j}| : a_{k,j} < 0\}\right)\ .$$

Let $r := \sum_{j=1}^n r_j$, $R := \{1,\ldots,r\}$, $h := m+1$, and $H := \{1,\ldots,h\}$. We now describe how to construct vectors $u, v \in \mathbb{Z}^r, w \in \mathbb{Z}^h$, and a set $E \subset R \times R \times H$ of triples—the "enabled," non-"forbidden" entries—such that the polytope P is represented as the corresponding transportation polytope of $r \times r \times h$ arrays with plane-sums u, v, w and only entries indexed by E enabled:

$$T = \left\{ x \in \mathbb{R}_{\geq 0}^{r\times r\times h} \ :\ x_{i,j,k} = 0 \ \text{ for all } \ (i,j,k) \notin E\,, \ \text{ and } \right.$$

$$\left. \sum_{i,j} x_{i,j,k} = w_k\,,\ \sum_{i,k} x_{i,j,k} = v_j\,,\ \sum_{j,k} x_{i,j,k} = u_i \right\}\ .$$

We also indicate the injection $\sigma : \{1,\ldots,n\} \longrightarrow R \times R \times H$ giving the desired embedding of the coordinates y_j as the coordinates $x_{i,j,k}$ and the representation of P as T (see paragraph following Theorem 1.1).

Basically, each equation $k = 1,\ldots,m$ will be encoded in a "horizontal plane" $R \times R \times \{k\}$ (the last plane $R \times R \times \{h\}$ is included for consistency and its entries can be regarded as "slacks"); and each variable y_j, $j = 1,\ldots,n$, will be encoded in a "vertical box" $R_j \times R_j \times H$, where $R = \biguplus_{j=1}^n R_j$ is the natural partition of R with $|R_j| = r_j$, namely with $R_j := \{1 + \sum_{l<j} r_l, \ldots, \sum_{l\leq j} r_l\}$.

Now, all "vertical" plane-sums are set to the same value U, that is, $u_j := v_j := U$ for $j = 1,\ldots,r$. All entries not in the union $\biguplus_{j=1}^n R_j \times R_j \times H$ of the variable boxes will be forbidden. We now describe the enabled entries in the boxes; for simplicity we discuss the box $R_1 \times R_1 \times H$, the others being similar. We distinguish between the two cases $r_1 = 1$ and $r_1 \geq 2$. In the first case, $R_1 = \{1\}$; the box, which is just the single line $\{1\}\times\{1\}\times H$, will have exactly two enabled entries $(1,1,k^+), (1,1,k^-)$ for suitable k^+, k^- to be defined later. We set $\sigma(1) := (1,1,k^+)$, namely embed $y_1 = x_{1,1,k^+}$. We define the *complement* of the variable y_1 to be $\bar{y}_1 := U - y_1$ (and likewise for the other variables). The vertical sums u, v then force $\bar{y}_1 = U - y_1 = U - x_{1,1,k^+} = x_{1,1,k^-}$, so the complement of y_1 is also embedded. Next, consider the case $r_1 \geq 2$. For each $s = 1,\ldots,r_1$, the line $\{s\} \times \{s\} \times H$ (respectively, $\{s\} \times \{1 + (s \bmod r_1)\} \times H$) will contain one enabled entry $(s, s, k^+(s))$ (respectively, $(s, 1 + (s \bmod r_1), k^-(s))$. All other entries of $R_1 \times R_1 \times H$ will be forbidden. Again, we set $\sigma(1) := (1,1,k^+(1))$, namely embed $y_1 = x_{1,1,k^+(1)}$; it is then not hard to see that, again, the vertical sums u, v force $x_{s,s,k^+(s)} = x_{1,1,k^+(1)} = y_1$ and $x_{s,1+(s \bmod r_1),k^-(s)} = U - x_{1,1,k^+(1)} = \bar{y}_1$ for each $s = 1,\ldots,r_1$. Therefore, both y_1 and $\bar{y}_1$ are each embedded in r_1 distinct entries.

To clarify the above description it is helpful to visualize the $R \times R$ matrix $(x_{i,j,+})$ whose entries are the vertical line-sums $x_{i,j,+} := \sum_{k=1}^h x_{i,j,k}$. For instance, if we have three variables with $r_1 = 3, r_2 = 1, r_3 = 2$ then $R_1 = \{1,2,3\}, R_2 = \{4\}, R_3 = \{5,6\}$,

and the line-sums matrix $x = (x_{i,j,+})$ is

$$\begin{pmatrix} x_{1,1,+} & x_{1,2,+} & 0 & 0 & 0 & 0 \\ 0 & x_{2,2,+} & x_{2,3+} & 0 & 0 & 0 \\ x_{3,1,+} & 0 & x_{3,3,+} & 0 & 0 & 0 \\ 0 & 0 & 0 & x_{4,4,+} & 0 & 0 \\ 0 & 0 & 0 & 0 & x_{5,5,+} & x_{5,6,+} \\ 0 & 0 & 0 & 0 & x_{6,5,+} & x_{6,6,+} \end{pmatrix} = \begin{pmatrix} y_1 & \bar{y}_1 & 0 & 0 & 0 & 0 \\ 0 & y_1 & \bar{y}_1 & 0 & 0 & 0 \\ \bar{y}_1 & 0 & y_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & U & 0 & 0 \\ 0 & 0 & 0 & 0 & y_3 & \bar{y}_3 \\ 0 & 0 & 0 & 0 & \bar{y}_3 & y_3 \end{pmatrix}.$$

We now encode the equations by defining the horizontal plane-sums w and the indices $k^+(s), k^-(s)$ mentioned above as follows. For $k = 1, \ldots, m$, consider the kth equation $\sum_j a_{k,j} y_j = b_k$. Define the index sets $J^+ := \{j : a_{k,j} > 0\}$ and $J^- := \{j : a_{k,j} < 0\}$, and set $w_k := b_k + U \cdot \sum_{j \in J^-} |a_{k,j}|$. The last coordinate of w is set for consistency with u, v to be $w_h = w_{m+1} := r \cdot U - \sum_{k=1}^{m} w_k$. Now, with $\bar{y}_j := U - y_j$ the complement of variable y_j as above, the kth equation can be rewritten as

$$\sum_{j \in J^+} a_{k,j} y_j + \sum_{j \in J^-} |a_{k,j}| \bar{y}_j = \sum_{j=1}^{n} a_{k,j} y_j + U \cdot \sum_{j \in J^-} |a_{k,j}| = b_k + U \cdot \sum_{j \in J^-} |a_{k,j}| = w_k.$$

We encode this equation by setting, for each $j \in J_+$, $k^+(s) = k$ for $|a_{k,j}|$ many different values of s (respectively, for each $j \in J_-$ we set $k^-(s) = k$ for enough values of s). By suitably setting $k^+(s) := k$ or $k^-(s) := k$, this has the effect of pulling enough copies of the variables y_j or $\bar{y}_j$ to the corresponding kth horizontal plane. Of course, once a variable is used at a certain horizontal level it cannot be used in others. By the choice of r_j there are sufficiently many copies of variables y_j $\bar{y}_j$, possibly with a few redundant copies which are absorbed in the last hyperplane by setting $k^+(s) := m + 1$ or $k^-(s) := m + 1$. For instance, if $m = 8$, the first variable y_1 has $r_1 = 3$ as above, its coefficient $a_{4,1} = 3$ in the fourth equation is positive, its coefficient $a_{7,1} = -2$ in the seventh equation is negative, and $a_{k,1} = 0$ for $k \neq 4, 7$, then we set $k^+(1) = k^+(2) = k^+(3) := 4$ (so $\sigma(1) := (1, 1, 4)$ embedding y_1 as $x_{1,1,4}$), $k^-(1) = k^-(2) := 7$, and $k^-(3) := h = 9$. This way, all equations are suitably encoded, and we obtain the following theorem.

THEOREM 3.2. *Any rational polytope $P = \{y \in \mathbb{R}^n_{\geq 0} : Ay = b\}$ is polynomial-time representable as a plane-sum entry-forbidden* 3*-way transportation polytope*

$$T = \left\{ x \in \mathbb{R}^{r \times r \times h}_{\geq 0} \ : \ x_{i,j,k} = 0 \ \textit{for all} \ (i,j,k) \notin E\,, \quad \textit{and} \right.$$

$$\left. \sum_{i,j} x_{i,j,k} = w_k\,,\ \sum_{i,k} x_{i,j,k} = v_j\,,\ \sum_{j,k} x_{i,j,k} = u_i \right\}\,.$$

Here E denotes the set of enabled, nonforbidden entries.

Proof. The proof follows from the construction outlined above and Lemma 3.1. □

3.3. Representing plane-sum entry-bounded as slim line-sum entry-free. Here we start with a transportation polytope of plane-sums and *upper-bounds*

$e_{i,j,k}$ on the entries,

$$P = \left\{ y \in \mathbb{R}^{l\times m\times n}_{\geq 0} : \sum_{i,j} y_{i,j,k} = c_k \, , \sum_{i,k} y_{i,j,k} = b_j \, , \sum_{j,k} y_{i,j,k} = a_i \, , \; y_{i,j,k} \leq e_{i,j,k} \right\} .$$

Clearly, this is a more general form than that of T appearing in Theorem 3.2 above; the forbidden entries can be encoded by setting a "forbidding" upper-bound $e_{i,j,k} := 0$ on all forbidden entries $(i,j,k) \notin E$ and an "enabling" upper-bound $e_{i,j,k} := U$ on all enabled entries $(i,j,k) \in E$. Thus, by Theorem 3.2, any rational polytope is representable also as such a plane-sum entry-bounded transportation polytope P. We now describe how to represent, in turn, such a P as a slim line-sum (unrestricted-entry) transportation polytope of the form of Theorem 1.1,

$$T \; = \; \left\{ x \in \mathbb{R}^{r\times c\times 3}_{\geq 0} \; : \; \sum_{I} x_{I,J,K} = w_{J,K} \, , \; \sum_{J} x_{I,J,K} = v_{I,K} \, , \; \sum_{K} x_{I,J,K} = u_{I,J} \right\} .$$

This stage of our construction was first presented in [6] while studying the complexity of deciding if T has an integer point; we include the details for completeness of the presentation. We give explicit formulas for $u_{I,J}, v_{I,K}, w_{J,K}$ in terms of a_i, b_j, c_k, and $e_{i,j,k}$ as follows. Put $r := l \cdot m$ and $c := n + l + m$. The first index I of each entry $x_{I,J,K}$ will be a pair $I = (i,j)$ in the r-set

$$\{(1,1), \ldots, (1,m), (2,1), \ldots, (2,m), \ldots, (l,1), \ldots, (l,m)\} \, .$$

The second index J of each entry $x_{I,J,K}$ will be a pair $J = (s,t)$ in the c-set

$$\{(1,1), \ldots, (1,n), (2,1), \ldots, (2,l), (3,1), \ldots, (3,m)\} \, .$$

The last index K will simply range in the 3-set $\{1,2,3\}$. We represent P as T via the injection σ given explicitly by $\sigma(i,j,k) := ((i,j),(1,k),1)$, embedding each variable $y_{i,j,k}$ as the entry $x_{(i,j),(1,k),1}$. Let U now denote the minimum between the two values $\max\{a_1, \ldots, a_l\}$ and $\max\{b_1, \ldots, b_m\}$. The 2-margins entries will be

$$u_{(i,j),(1,t)} = e_{i,j,t}, \quad u_{(i,j),(2,t)} = \begin{cases} U & \text{if } t = i, \\ 0 & \text{otherwise,} \end{cases} \quad u_{(i,j),(3,t)} = \begin{cases} U & \text{if } t = j, \\ 0 & \text{otherwise,} \end{cases}$$

$$v_{(i,j),t} = \begin{cases} U & \text{if } t = 1, \\ e_{i,j,+} & \text{if } t = 2, \\ U & \text{if } t = 3, \end{cases}$$

$$w_{(i,j),1} = \begin{cases} c_j & \text{if } i = 1, \\ m \cdot U - a_j & \text{if } i = 2, \\ 0 & \text{if } i = 3. \end{cases} \qquad w_{(i,j),2} = \begin{cases} e_{+,+,j} - c_j & \text{if } i = 1, \\ 0 & \text{if } i = 2, \\ b_j & \text{if } i = 3. \end{cases}$$

$$w_{(i,j),3} = \begin{cases} 0 & \text{if } i = 1, \\ a_j & \text{if } i = 2, \\ l \cdot U - b_j & \text{if } i = 3. \end{cases}$$

THEOREM 3.3. *Any rational plane-sum entry-bounded* 3*-way transportation polytope*

$$P = \left\{ y \in \mathbb{R}^{l\times m\times n}_{\geq 0} : \sum_{i,j} y_{i,j,k} = c_k \,, \sum_{i,k} y_{i,j,k} = b_j \,, \sum_{j,k} y_{i,j,k} = a_i \,,\ y_{i,j,k} \leq e_{i,j,k} \right\}$$

is strongly-polynomial-time representable as a line-sum slim transportation polytope

$$T = \left\{ x \in \mathbb{R}^{r\times c\times 3}_{\geq 0} \ :\ \sum_{I} x_{I,J,K} = w_{J,K} \,,\ \sum_{J} x_{I,J,K} = v_{I,K} \,,\ \sum_{K} x_{I,J,K} = u_{I,J} \right\}.$$

Proof. We outline the proof; complete details appeared in [6]. First, consider any $y = (y_{i,j,k}) \in P$; we claim the embedding via σ of $y_{i,j,k}$ in $x_{(i,j),(1,k),1}$ can be extended uniquely to $x = (x_{I,J,K}) \in T$. First, the entries $x_{I,(3,t),1}$, $x_{I,(2,t),2}$ and $x_{I,(1,t),3}$ for all $I = (i,j)$ and t are zero since so are the line-sums $w_{(3,t),1}$, $w_{(2,t),2}$ and $w_{(1,t),3}$. Next, consider the entries $x_{I,(2,t),1}$: since all entries $x_{I,(3,t),1}$ are zero, examining the line-sums $u_{I,(2,t)}$ and $v_{I,1} = U$, we find $x_{(i,j),(2,i),1} = U - \sum_{t=1}^{n} x_{(i,j),(1,t),1} = U - y_{i,j,+} \geq 0$ whereas for $t \neq i$ we get $x_{(i,j),(2,t),1} = 0$. This also gives the entries $x_{I,(2,t),3}$: we have $x_{(i,j),(2,i),3} = U - x_{(i,j),(2,i),1} = y_{i,j,+} \geq 0$ whereas for $t \neq i$ we have $x_{(i,j),(2,t),3} = 0$. Next, consider the entries $x_{I,(1,t),2}$: since all entries $x_{I,(1,t),3}$ are zero, examining the line-sums $u_{(i,j),(1,k)} = e_{i,j,k}$ we find $x_{(i,j),(1,k),2} = e_{i,j,k} - y_{i,j,k} \geq 0$ for all i,j,k. Next consider the entries $x_{I,(3,t),2}$: since all entries $x_{I,(2,t),2}$ are zero, examining the line-sums $u_{(i,j),(3,t)}$ and $v_{(i,j),2} = e_{i,j,+}$, we find $x_{(i,j),(3,j),2} = e_{i,j,+} - \sum_{k=1}^{l} x_{(i,j),(1,k),2} = y_{i,j,+} \geq 0$ whereas for $t \neq j$ we get $x_{(i,j),(3,t),2} = 0$. This also gives the entries $x_{I,(3,t),3}$: we have $x_{(i,j),(3,j),3} = U - x_{(i,j),(3,j),2} = U - y_{i,j,+} \geq 0$ whereas for $t \neq j$ we get $x_{(i,j),(3,t),3} = 0$. Using the relations established above, one can easily check that all line-sums are correct.

Conversely, given any $x = (x_{I,J,K}) \in T$, let $y = (y_{i,j,k})$ with $y_{i,j,k} := x_{(i,j),(1,k),1}$. Since x is nonnegative, so is y. Further, $e_{i,j,k} - y_{i,j,k} = x_{(i,j),(1,k),2} \geq 0$ for all i,j,k and hence y obeys the entry upper-bounds. Finally, using the relations established above $x_{(i,j),(3,t),2} = 0$ for $t \neq j$, $x_{(i,j),(2,t),3} = 0$ for $t \neq i$, and $x_{(i,j),(3,j),2} = x_{(i,j),(2,i),3} = y_{i,j,+}$, we obtain

$$\sum_{i,j} y_{i,j,k} \ = \ \sum_{i,j} x_{(i,j),(1,k),1} \ = \ w_{(1,k),1} \ = \ c_k \,, \qquad 1 \leq k \leq n \ ;$$

$$\sum_{i,k} y_{i,j,k} \ = \ \sum_{i} x_{(i,j),(3,j),2} \ = \ w_{(3,j),2} \ = \ b_j \,, \qquad 1 \leq j \leq m \ ;$$

$$\sum_{j,k} y_{i,j,k} \ = \ \sum_{j} x_{(i,j),(2,i),3} \ = \ w_{(2,i),3} \ = \ a_i \,, \qquad 1 \leq i \leq l \ .$$

This shows that y satisfies the plane-sums as well and hence is in P. Since integrality is also preserved in both directions, this completes the proof. □

3.4. The main theorem and a complexity estimate. Call a class $\mathcal{P}$ of rational polytopes *polynomial-time representable* in a class $\mathcal{Q}$ if there is a polynomial-time

algorithm that represents any given $P \in \mathcal{P}$ as some $Q \in \mathcal{Q}$. The resulting binary relation on classes of rational polytopes is clearly transitive. Thus, the composition of Theorem 3.2 (which incorporates Lemma 3.1) and Theorem 3.3 gives at once Theorem 1.1 stated in the introduction. Working out the details of our three-stage construction, we can give the following estimate on the number of rows r and columns c in the resulting representing transportation polytope, in terms of the input. The computational complexity of the construction is also determined by this bound, but we do not dwell on the details here.

THEOREM 1.1 (with complexity estimate). *Any polytope* $P=\{y \in \mathbb{R}^n_{\geq 0} : Ay = b\}$ *with integer* $m \times n$ *matrix* $A = (a_{i,j})$ *and integer* b *is polynomial-time representable as a slim transportation polytope*

$$T = \left\{ x \in \mathbb{R}^{r\times c\times 3}_{\geq 0} \ : \ \sum_i x_{i,j,k} = w_{j,k}\,,\ \sum_j x_{i,j,k} = v_{i,k}\,,\ \sum_k x_{i,j,k} = u_{i,j} \right\}\,,$$

with $r = O(m^2(n+L)^2)$ *rows and* $c = O(m(n+L))$ *columns, where*

$$L := \sum_{j=1}^{n} \max_{i=1}^{m} \lfloor \log_2 |a_{i,j}| \rfloor .$$

3.5. Proof of the universality of the bitransportation problem. We conclude with the modification of the proof of Theorem 3.3 that establishes Theorem 1.2.

THEOREM 1.2. *Any rational polytope* $P = \{y \in \mathbb{R}^n_{\geq 0} : Ay = b\}$ *is polynomial-time representable as a bipartite bitransportation polytope*

$$F = \left\{ (x^1, x^2) \in \mathbb{R}^{r\times c}_{\geq 0} \oplus \mathbb{R}^{r\times c}_{\geq 0} \ : \ x^1_{i,j} + x^2_{i,j} \leq u_{i,j}\,, \right.$$

$$\left. \sum_j x^k_{i,j} = s^k_i\,,\ \sum_i x^k_{i,j} = d^k_j\,, \quad k = 1,2 \right\}\,.$$

Here r, c are the same values as presented in Theorem 1.1 above. Moreover, the statement remains valid with all supplies s^k_i having the same value U and all capacities $u_{i,j}$ being 0 or U for some suitable nonnegative integer U.

Proof. We do an easy adjustment of the proof of Theorem 3.3 above: We essentially need to describe the capacities, demands and supplies (for each of two commodities) for a bipartite network with $l \cdot m$ nodes for the first part and $n + l + m$ nodes in the second part, with $l \cdot m \cdot (n + l + m)$ arcs. Take the capacities of the arcs to be $u_{i,j}$ as defined in section 3.3; take the supplies to be $s^1_i := v_{i,1} = U$ and $s^2_i := v_{i,3} = U$ for all i, and take the demands to be $d^1_j := w_{j,1}$ and $d^2_j := w_{j,3}$ for all j. Note that by taking s^2_i and d^2_j to be $v_{i,3}$ and $w_{j,3}$ instead of $v_{i,2}$ and $w_{j,2}$ we can guarantee that all supplies have the same value U. Moreover, since the proof follows by the composition of Theorem 3.2 and Theorem 3.3, and the former makes use of forbidden entries only, rather than upper bounds, it is easy to see that we can take all upper bounds $e_{i,j,k}$ in the latter (and hence all $u_{i,j}$) to be either 0 or U, proving the stronger statement. More visually, the data can also be described in matrix form as follows:

$$u = \begin{pmatrix}
e_{1,1,1} & e_{1,1,2} & \cdots & e_{1,1,n} & U & 0 & \cdots & 0 & U & 0 & \cdots & 0 \\
e_{1,2,1} & e_{1,2,2} & \cdots & e_{1,2,n} & U & 0 & \cdots & 0 & 0 & U & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
e_{1,m,1} & e_{1,m,2} & \cdots & e_{1,m,n} & U & 0 & \cdots & 0 & 0 & 0 & \cdots & U \\
\\
e_{2,1,1} & e_{2,1,2} & \cdots & e_{2,1,n} & 0 & U & \cdots & 0 & U & 0 & \cdots & 0 \\
e_{2,2,1} & e_{2,2,2} & \cdots & e_{2,2,n} & 0 & U & \cdots & 0 & 0 & U & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
e_{2,m,1} & e_{2,m,2} & \cdots & e_{2,m,n} & 0 & U & \cdots & 0 & 0 & 0 & \cdots & U \\
\\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\\
e_{l,1,1} & e_{l,1,2} & \cdots & e_{l,1,n} & 0 & 0 & \cdots & U & U & 0 & \cdots & 0 \\
e_{l,2,1} & e_{l,2,2} & \cdots & e_{l,2,n} & 0 & 0 & \cdots & U & 0 & U & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
e_{l,m,1} & e_{l,m,2} & \cdots & e_{l,m,n} & 0 & 0 & \cdots & U & 0 & 0 & \cdots & U
\end{pmatrix},$$

$$s^1 = s^2 = \begin{pmatrix}
U \\ U \\ \vdots \\ U \\ \\ U \\ U \\ \vdots \\ U \\ \\ \vdots \\ \\ U \\ U \\ \vdots \\ U
\end{pmatrix},$$

$$d^1 = (\, c_1, \quad c_2, \quad \ldots, \quad c_n, \quad m \cdot U - a_1, \quad m \cdot U - a_2, \ldots, \quad m \cdot U - a_l, \quad 0, \quad 0, \quad \ldots, \quad 0\,),$$
$$d^2 = (\, 0, \quad 0, \quad \ldots, \quad 0, \quad a_1, \quad a_2, \quad \ldots, a_l, \quad l \cdot U - b_1, \quad l\cdot, U - b_2, \ldots, \quad l \cdot U - b_m\,). \qquad \square$$

Acknowledgments. The authors are grateful to G. Rote, F. Santos, S. Hoşten, G. Ziegler, and the anonymous referees for their comments and suggestions.

REFERENCES

[1] M. Baiöu and M. L. Balinski, *The stable allocation (or ordinal transportation) problem*, Math. Oper. Res., 27 (2002), pp. 485–503.

[2] M. L. Balinski and F. J. Rispoli, *Signature classes of transportation polytopes*, Math. Program Ser. A, 60 (1993), pp. 127–144.

[3] L. H. Cox, *Bounds on entries in* 3*-dimensional contingency tables*, in Inference Control in Statistical Databases: From Theory to Practice, J. Domingo-Ferrer, ed., Lecture Notes in Comput. Sci. 2316, Springer, New York, 2002, pp. 21–33.

[4] L. H. Cox, *On properties of multi-dimensional statistical tables*, J. Stat. Plann. Inference, 117 (2003), pp. 251–273.

[5] M. Cryan, M. Dyer, H. Müller, and L. Stougie, *Random walks on the vertices of transportation polytopes with constant number of sources*, in Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms (Baltimore, MD), ACM, New York, 2003, pp. 330–339.

[6] J. De Loera and S. Onn, *The complexity of three-way statistical tables*, SIAM J. Comput., 33 (2004), pp. 819–836.

[7] J. De Loera and S. Onn, *All rational polytopes are transportation polytopes and all polytopal integer sets are contingency tables*, in Integer Programming and Combinatorial Optimization, Proceedings of the 10th International IPCO Conference, Lecture Notes in Comput. Sci. 3064, Springer, New York, 2004, pp. 338–351.

[8] J. De Loera and S. Onn, *Markov bases of three-way tables are arbitrarily complicated*, J. Symbolic Comput., 41 (2006), pp. 173–181.

[9] P. Diaconis and A. Gangolli, *Rectangular arrays with fixed margins* in Discrete Probability and Algorithms (Minneapolis, MN, 1993), D. Aldous, P. Diaconis, J. Spencer, and J. M. Steele, eds., IMA Vol. Math. Appl. 72, Springer, New York, 1995, pp. 15–41.

[10] G. T. Duncan, S. E. Fienberg, R. Krishnan, R. Padman, and S. F. Roehrig, *Disclosure limitation methods and information loss for tabular data*, in Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, P. Doyle, J. I. Land, J. M. Theeuwes, and L. V. Zayatz, eds., North-Holland, Amsterdam, 2001, pp. 135–166.

[11] L. K. Fleischer, *Approximating fractional multicommodity flow independent of the number of commodities*, SIAM J. Discrete Math., 13 (2000), pp. 505–520.

[12] N. Garg and J. Könemann, *Faster and simpler algorithms for multicommodity flow and other fractional packing problems*, in Proceedings of the 39th Annual IEEE Symposium on Foundations of Computer Science (New York), IEEE, New York, 1998, pp. 300–309.

[13] D. Gusfield, *A graph theoretic approach to statistical data security*, SIAM J. Comput., 17 (1988), pp. 552–571.

[14] J. Håstad, *Clique is hard to approximate within* $n^{1-\epsilon}$, Acta Math., 182 (1999), pp. 105–142.

[15] S. Hoşten and S. Sullivant, *Gröbner bases and polyhedral geometry of reducible and cyclic models*, J. Combin. Theory Ser. A, 100 (2002), pp. 277–301.

[16] T. C. Hu, *Multi-commodity network flows*, Oper. Res., 11 (1963), pp. 344–360.

[17] F. K. Hwang, S. Onn, and U. G. Rothblum, *A polynomial time algorithm for shaped partition problems*, SIAM J. Optim., 10 (1999), pp. 70–81.

[18] R. Irving and M. R. Jerrum, *Three-dimensional statistical data security problems*, SIAM J. Comput., 23 (1994), pp. 170–184.

[19] A. Itai, *Two-commodity flow*, J. Assoc. Comput. Mach., 25 (1978), pp. 596–611.

[20] V. Klee and C. Witzgall, *Facets and vertices of transportation polytopes*, in Mathematics of the Decision Sciences, Part I (Stanford, CA, 1967), AMS, Providence, RI, 1968, pp. 257–282.

[21] A. Levin, *Personal communication*, 2004.

[22] C. R. Mehta and N. R. Patel, *A network algorithm for performing Fisher's exact test in* $r \times c$ *contingency tables*, J. Amer. Statist. Assoc., 78 (1983), pp. 427–434.

[23] S. Onn and U. G. Rothblum, *Convex combinatorial optimization*, Discrete Comput. Geom., 32 (2004), pp. 549–566.

[24] S. Onn and L. J. Schulman, *The vector partition problem for convex objective functions*, Math. Oper. Res., 26 (2001), pp. 583–590.

[25] A. Schrijver, *Theory of Linear and Integer Programming*, John Wiley & Sons, Ltd., Chichester, 1986.

[26] E. Tardos, *A strongly polynomial algorithm to solve combinatorial linear programs*, Oper. Res., 34 (1986), pp. 250–256.

[27] Universal Generator, http://www.math.ucdavis.edu/∼deloera, http://ie.technion.ac.il/∼onn.

[28] U.S. Bureau of the Census, *American FactFinder*, interactive information accessible at http://factfinder.census.gov/servlet/BasicFactsServlet.

[29] M. Vlach, *Conditions for the existence of solutions of the three-dimensional planar transportation problem*, Discrete Appl. Math., 13 (1986), pp. 61–78.

[30] V. A. Yemelichev, M. M. Kovalev, and M. K. Kravtsov, *Polytopes, Graphs and Optimisation*, Cambridge University Press, Cambridge, UK, 1984.

SIAM J. OPTIM.
Vol. 17, No. 3, pp. 822–843
© 2006 Society for Industrial and Applied Mathematics

CONVERGENT SDP-RELAXATIONS IN POLYNOMIAL OPTIMIZATION WITH SPARSITY*

JEAN B. LASSERRE†

Abstract. We consider a polynomial programming problem $\mathbf{P}$ on a compact basic semialgebraic set $\mathbf{K} \subset \mathbb{R}^n$, described by m polynomial inequalities $g_j(X) \geq 0$, and with criterion $f \in \mathbb{R}[X]$. We propose a hierarchy of semidefinite relaxations in the spirit of those of Waki et al. [*SIAM J. Optim.*, 17 (2006), pp. 218–242]. In particular, the SDP-relaxation of order r has the following two features: (a) The number of variables is $O(\kappa^{2r})$, where $\kappa = \max[\kappa_1, \kappa_2]$ with κ_1 (resp., κ_2) being the maximum number of variables appearing in the monomials of f (resp., appearing in a single constraint $g_j(X) \geq 0$). (b) The largest size of the linear matrix inequalities (LMIs) is $O(\kappa^r)$. This is to compare with the respective number of variables $O(n^{2r})$ and LMI size $O(n^r)$ in the original SDP-relaxations defined in [J. B. Lasserre, *SIAM J. Optim.*, 11 (2001), pp. 796–817]. Therefore, great computational savings are expected in case of sparsity in the data $\{g_j, f\}$, i.e., when κ is small, a frequent case in practical applications of interest. The novelty with respect to [H. Waki, S. Kim, M. Kojima, and M. Maramatsu, *SIAM J. Optim.*, 17 (2006), pp. 218–242] is that we prove convergence to the global optimum of $\mathbf{P}$ when the sparsity pattern satisfies a condition often encountered in large size problems of practical applications, and known as the *running intersection property* in graph theory. In such cases, and as a by-product, we also obtain a new representation result for polynomials positive on a compact basic semialgebraic set, a *sparse* version of Putinar's Positivstellensatz [M. Putinar, *Indiana Univ. Math. J.*, 42 (1993), pp. 969–984].

Key words. real algebraic geometry, positive polynomials, sum of squares, semidefinite programming

AMS subject classifications. 12E05, 12Y05, 90C22

DOI. 10.1137/05064504X

1. Introduction.

In this paper we consider the polynomial programming problem

$$\mathbf{P}: \quad \inf_{x \in \mathbb{R}^n} \{ f(x) \mid \quad x \in \mathbf{K} \}, \tag{1.1}$$

where $f \in \mathbb{R}[X]$, and $\mathbf{K} \subset \mathbb{R}^n$ is the basic closed semialgebraic set defined by

$$\mathbf{K} := \{ x \in \mathbb{R}^n \mid \quad g_j(x) \geq 0, \quad j = 1, \ldots, m \}, \tag{1.2}$$

for some polynomials $\{g_j\}_{j=1}^m \subset \mathbb{R}[X]$.

The hierarchy of semidefinite programming (SDP) relaxations introduced in [11] provides a sequence of SDPs of increasing size, whose associated sequence of optimal values converges to the global minimum of $\mathbf{P}$. Moreover, as proved by Schweighofer in [17], convergence to a global minimizer of $\mathbf{P}$ (if unique) also holds. For more details, the reader is referred to [5, 11, 17] and the many references therein. In addition, practice reveals that convergence is usually fast, and often *finite* (up to machine precision); see, e.g., Henrion and Lasserre [5].

However, despite these nice features, the size of the SDP-relaxations grows rapidly with the size of the original problem. Typically, the kth SDP-relaxation has to handle

*Received by editors November 12, 2005; accepted for publication (in revised form) April 13, 2006; published electronically October 16, 2006. This research was sponsored by the French National Research Agency (ANR) under grant NT-05-3-41612.

http://www.siam.org/journals/siopt/17-3/64504.html

†LAAS-CNRS and Institute of Mathematics, LAAS 7 Avenue du Colonel Roche, 31077 Toulouse Cédex 4, France (lasserre@laas.fr).

at least one linear matrix inequality (LMI) of size $\binom{n+k}{n}$ and $\binom{n+2k}{n}$ variables, which clearly limits the applicability of the methodology to problems with small to medium size only. Therefore, validation of the above methodology for larger size problems (and even more, for large scale problems) is a real challenge of practical importance.

One way to extend the applicability of the methodology to problems of larger size is to take into account *sparsity* in the original data, frequently encountered in practical cases. Indeed, as is typical in many applications of interest, f as well as the polynomials $\{g_j\}$ that describe $\mathbf{K}$ are sparse, i.e., each monomial of f and each polynomial g_j are only concerned with a small subset of variables. This is the approach taken in Waki et al. [9] (extending Kim, Kojima, and Waki [7] and Kojima, Kim, and Waki [8]), where the authors have built up a hierarchy of SDP-relaxations in the spirit of those in [11], but where sparsity is taken into account. Sometimes, a sparsity pattern can be "read" from the data of $\mathbf{P}$ but not always, and in [9] the authors have proposed a systematic procedure to detect and structure sparsity in $\mathbf{P}$, via the so-called *chordal extension* of the *correlation sparsity pattern graph* (csp graph); the csp graph has as many nodes as variables, and a link between two nodes (i.e., variables) means that these two variables both appear in a monomial of the objective function or in some inequality constraint $g_j \geq 0$ of $\mathbf{P}$. Once a sparsity pattern has been detected, they define a simplified "sparse" version of the SDP-relaxations of [11]; briefly, in the dual, the sum of squares (s.o.s.) multiplier associated with a constraint is now a polynomial in only those variables appearing in that constraint. In doing so, they have obtained impressive gains in the size of the resulting SDP-relaxations, as well as in the computational time needed for obtaining an optimal solution. As a matter of fact, they were even able to solve problems that could not be handled with the original SDP-relaxations. However, and despite good approximations are obtained in most problems in their sample of experiments, convergence to the global minimum is *not* guaranteed.

Contribution. Our contribution is twofold: We first propose a hierarchy of SDP-relaxations $\{\mathbf{Q}_r\}$ in the spirit of the original SDP-relaxations [11] and close to those defined in [9]. They are valid for arbitrary polynomial programming problems, and have the following three appealing features:

(a) In the SDP-relaxation $\mathbf{Q}_r$ of order r, the number of variables is $O(\kappa^{2r})$, where $\kappa = \max[\kappa_1, \kappa_2]$ with κ_1 (resp., κ_2) being the maximum number of variables appearing in each monomial of f (resp., in a single constraint $g_j(X) \geq 0$).

(b) The largest size of the LMIs is $O(\kappa^r)$.

This is to compare with the respective number of variables $O(n^{2r})$ and LMI size $O(n^r)$ in the original SDP-relaxations defined in [11].

(c) Under a certain condition on the sparsity pattern, the resulting sequence of their optimal value *converges* to the global minimum of $\mathbf{P}$.

So in view of (a) and (b), and when κ is small ($\kappa \ll n$), i.e., when sparsity is present, dramatic computational savings can be expected. In other words, these new SDP-relaxations are inherently exploiting sparsity in the data $\{f, g_j\}$ when present. Moreover, the size of the SDP-relaxation $\mathbf{Q}_r$ is in a sense *minimal*, at least when considering such types of SDP-relaxations, because one should at least handle moments involving κ variables, whenever some monomial of κ variables appears in the data $\{f, g_j\}$.

The condition under which such SDP-relaxations converge to the global minimum of $\mathbf{P}$ is easy to describe, and reflects a sparsity pattern frequently encountered in large scale problems. Namely, let $\{1, \ldots, n\}$ be the union $\bigcup_{k=1}^{p} I_k$ of subsets $I_k \subset \{1, \ldots, n\}$.

Every polynomial g_j in the definition (1.2) of $\mathbf{K}$ is only concerned with variables $\{X_i \,|\, i \in I_k\}$ for some k. Next, $f \in \mathbb{R}[X]$ can be written $f = f_1 + \cdots + f_p$, where each f_k uses only variables $\{X_i \,|\, i \in I_k\}$. In cases where the subsets $\{I_k\}$ are not so easy to detect, one may use the procedure of Waki et al. [9] via the chordal extension of the csp graph.

Finally, the collection $\{I_1, \ldots, I_p\}$ should obey the following condition: For every $k = 1, \ldots, p-1$,

$$I_{k+1} \cap \bigcup_{j=1}^{k} I_j \subseteq I_s \quad \text{for some } s \leq k. \tag{1.3}$$

Notice that (1.3) is always satisfied when $p = 2$. Property (1.3) depends on the ordering and so can be satisfied possibly after some relabelling of the I_k's. Moreover, if not satisfied, one may enforce (1.3) but at the price of enlarging some of the sets I_k. If $I_1, \ldots, I_p$ are the maximal cliques of a chordal graph, then (1.3) is satisfied possibly after some reordering of the cliques, and is known as the *running intersection property*; for more details on chordal graphs, the reader is referred to Fukuda et al. [4] and Nakata et al. [15].

In particular, (1.3) is naturally satisfied in a number of applications, in particular, in what we call *strong* and *weak* coupling. In the former, we have $I_k \cap I_{k+j} = \emptyset$ whenever $j > 1$, so that (1.3) holds. In the latter, there is a set of *coupling variables* with index set $I_0' \subset \{1, \ldots, n\}$, and a partition of $\{1, \ldots, n\} \setminus I_0'$ into p disjoint subsets of *independent variables* I_k', $k = 1, \ldots, p$. In this case one has $I_k := I_0' \cup I_k'$, $k = 1, \ldots, p$, and so $I_k \cap I_j = I_0'$ for all $j \neq k$, which in turn implies that (1.3) holds.

At last, and as a by-product of the property (1.3) of the sparsity pattern, we also obtain a new *sparse representation* result for polynomials, nonnegative on a basic closed semialgebraic set, a *sparse* version of Putinar's Positivstellensatz [16].

Link with related literature. As already mentioned, our work is closely related to the recent work of Kojima, Kim, and Waki [8] and Waki et al. [9], in which they were the first to exploit sparsity of data and modify (or simplify) in an appropriate way the original SDP-relaxations defined in [11]. Our SDP-relaxations are very close to those defined in [9], but handle p additional quadratic constraints. These p additional constraints, together with condition (1.3), are crucial to prove our convergence result. To summarize, our result implies that by a slight modification of the SDP-relaxations defined in [9], convergence is now guaranteed when the sparsity pattern satisfies (1.3).

The paper is organized as follows. After introducing notation and definitions, our main result is presented in section 3, and for clarity of exposition, some proofs are postponed to section 4, whereas auxiliary results needed in some proofs are postponed to an appendix section.

2. Notation and definitions. As common in algebra, variables of polynomials are denoted with capitals (e.g., X) whereas points in $\mathbb{R}^n$ are denoted with small letters (e.g., x). For a real symmetric matrix $A \in \mathbb{R}^{n \times n}$, the notation $A \succeq 0$ (resp., $A \succ 0$) stands for A is positive definite (resp., semidefinite), and for a vector x, let x' denote its transpose.

Let $\mathbb{R}[X]$ denote the ring of real polynomials in the variables $X_1, \ldots, X_n$. In the usual canonical basis $v_\infty(X) = \{X^\alpha \,|\, \alpha \in \mathbb{N}^n\}$ of monomials, a polynomial $g \in \mathbb{R}[X]$ is written

$$g(X) = \sum_{\alpha \in \mathbb{N}^n} g_\alpha X^\alpha, \tag{2.1}$$

for some real vector $\mathbf{g} = \{g_\alpha\}$ with finitely many nonzero coefficients.

With $\alpha \in \mathbb{N}^n$, let $|\alpha| := \sum_i \alpha_i$, and let $\mathbb{R}_r[X] \subset \mathbb{R}[X]$ be the $\mathbb{R}$-vector space of polynomials of degree at most r, with usual canonical basis of monomials $v_r(X) = \{X^\alpha \,|\, \alpha \in \mathbb{N}^n;\, |\alpha| \le r\}$.

Let $I_0 := \{1, \ldots, n\}$ be the union $\cup_{k=1}^p I_k$ of p subsets I_k, $k = 1, \ldots, p$, with cardinal denoted n_k. Let $\mathbb{R}[X(I_k)]$ denote the ring of polynomials in the n_k variables $X(I_k) = \{X_i \,|\, i \in I_k\}$, and so $\mathbb{R}[X(I_0)] = \mathbb{R}[X]$.

For each $k = 0, 1, \ldots, p$, let $\mathcal{I}_k$ be the set of all subsets of I_k. Next, for every $\alpha \in \mathbb{N}^n$, let $\mathrm{supp}\,(\alpha) \in \mathcal{I}_0$ be the support of α, i.e.,

$$\mathrm{supp}\,(\alpha) := \{\, i \in \{1, \ldots, n\} \,:\quad \alpha_i \neq 0 \,\}, \qquad \alpha \in \mathbb{N}^n.$$

For instance, with $n = 6$ and $\alpha := (004020)$, $\mathrm{supp}\,(\alpha) = \{3, 5\}$. Next, define

$$S_k := \{\, \alpha \in \mathbb{N}^n \,:\quad \mathrm{supp}\,(\alpha) \,\in\, \mathcal{I}_k \,\}, \qquad k = 1, \ldots, p. \tag{2.2}$$

A polynomial $h \in \mathbb{R}[X(I_k)]$ can be viewed as a member of $\mathbb{R}[X]$, and is written

$$h(X) \,=\, h(X(I_k)) \,=\, \sum_{\alpha \in S_k} h_\alpha\, X^\alpha \tag{2.3}$$

for some real vector $\mathbf{h} = \{h_\alpha\}$ with finitely many nonzero coefficients.

2.1. Moment matrix. Let $y = (y_\alpha)_{\alpha \in \mathbb{N}^n}$ (i.e., a sequence indexed in the canonical basis $v_\infty(X)$), and define the linear functional $L_y : \mathbb{R}[X] \to \mathbb{R}$ to be

$$g \mapsto L_y(g) \,:=\, \sum_{\alpha \in \mathbb{N}^n} g_\alpha\, y_\alpha, \tag{2.4}$$

whenever g is as in (2.1).

As already presented in [11], given a sequence $y = (y_\alpha)_{\alpha \in \mathbb{N}^n}$, the *moment* matrix $M_r(y)$ associated with y, is the matrix with rows and columns indexed in $v_r(X)$, and such that

$$M_r(y)(\alpha, \beta) \,:=\, L_y(X^\alpha X^\beta) \,=\, y_{\alpha+\beta} \quad \forall\, \alpha, \beta \in \mathbb{N}^n \text{ with } |\alpha|, |\beta| \le r.$$

A sequence y is said to have a representing measure μ on $\mathbb{R}^n$ if

$$y_\alpha \,=\, \int_{\mathbb{R}^n} X^\alpha\, \mu(dX) \qquad \forall\, \alpha \in \mathbb{N}^n.$$

Let $s(r) := \binom{n+r}{r}$ be the dimension of vector space $\mathbb{R}_r[X]$. For a vector $\mathbf{u} \in \mathbb{R}^{s(r)}$, let $u \in \mathbb{R}[X]$ be the polynomial $u(X) = \langle \mathbf{u}, v_r(X) \rangle$. Then, one has

$$\langle \mathbf{u}, M_r(y)\mathbf{u} \rangle \,=\, L_y(u^2) \qquad \forall\, \mathbf{u} \in \mathbb{R}^{s(r)}.$$

Therefore, if y has a representing measure μ, then

$$\langle \mathbf{u}, M_r(y)\mathbf{u} \rangle \,=\, L_y(u^2) \,=\, \int_{\mathbb{R}^n} u(X)^2\, \mu(dX) \,\ge\, 0,$$

which implies $M_r(y) \succeq 0$ (as $\mathbf{u} \in \mathbb{R}^{s(r)}$ was arbitrary).

Of course, in general, not every sequence y such that $M_r(y) \succeq 0$ for all $r \in \mathbb{N}$ has a representing measure. The $\mathbf{K}$-moment problem is precisely concerned with finding conditions on the sequence y, to ensure it is the moment sequence of some measure μ, with support contained in $\mathbf{K} \subset \mathbb{R}^n$.

2.2. Localizing matrix. Let $h \in \mathbb{R}[X]$ be a given polynomial

$$h(X) \,=\, \sum_{\gamma \in \mathbb{N}^n} h_\gamma \, X^\gamma,$$

and let $y = (y_\alpha)_{\alpha \in \mathbb{N}^n}$ be given. The *localizing* matrix $M_r(h\,y)$ associated with h and y is the matrix with rows and columns indexed in $v_r(X)$, obtained from the moment matrix $M_r(y)$ by

$$M_r(h\,y)(\alpha,\beta) \,:=\, L_y(h(X)\,X^\alpha X^\beta) \,=\, \sum_{\gamma \in \mathbb{N}^n} h_\gamma \, y_{\gamma+\alpha+\beta}$$

for all $\alpha, \beta \in \mathbb{N}^n$, with $|\alpha|, |\beta| \le r$.

As before, let $\mathbf{u} \in \mathbb{R}^{s(r)}$, and let $u := \langle \mathbf{u}, v_r(X)\rangle \in \mathbb{R}_r[X]$. Then

$$\langle \mathbf{u}, M_r(h\,y)\mathbf{u}\rangle \,=\, L_y(h\,u^2) \qquad \forall\, \mathbf{u} \in \mathbb{R}^{s(r)},$$

and if y has a representing measure μ with support contained in the set $\{x \in \mathbb{R}^n \,:\, h(x) \ge 0\}$, then

$$\langle \mathbf{u}, M_r(h\,y)\mathbf{u}\rangle \,=\, L_y(h\,u^2) \,=\, \int_{\mathbb{R}^n} h(X)\,u(X)^2\,\mu(dX) \,\ge 0,$$

which implies $M_r(h\,y) \succeq 0$ (as $\mathbf{u} \in \mathbb{R}^{s(r)}$ was arbitrary).

Next, with $k \in \{1,\ldots,p\}$ fixed, and $h \in \mathbb{R}[X(I_k)]$, let $M_r(y, I_k)$ (resp., $M_r(h\,y, I_k)$) be the moment (resp., localizing) submatrix obtained from $M_r(y)$ (resp., $M_r(h\,y)$) by retaining only those rows (and columns) $\alpha \in \mathbb{N}^n$ of $M_r(y)$ (resp., $M_r(h\,y)$) with $\mathrm{supp}\,(\alpha) \in \mathcal{I}_k$.

In doing so, $M_r(y, I_k)$ and $M_r(h\,y, I_k)$ can be viewed as moment and localizing matrices with rows and columns indexed in the canonical basis $v_r(X(I_k))$ of $\mathbb{R}_r[X(I_k)]$. Indeed, $M_r(y, I_k)$ contain only variables y_α with $\mathrm{supp}\,(\alpha) \in \mathcal{I}_k$, and so does $M_r(h\,y, I_k)$ because $h \in \mathbb{R}[X(I_k)]$. And for every polynomial $u \in \mathbb{R}_r[X(I_k)]$, with coefficient vector $\mathbf{u}$ in the basis $v_r(X(I_k))$, we also have

$$\begin{aligned}
\langle \mathbf{u}, M_r(y, I_k)\mathbf{u}\rangle &= L_y(u^2) \qquad \forall\, u \in \mathbb{R}_r[X(I_k)],\\
\langle \mathbf{u}, M_r(h\,y, I_k)\mathbf{u}\rangle &= L_y(h\,u^2) \qquad \forall\, u \in \mathbb{R}_r[X(I_k)],
\end{aligned}$$

and therefore,

$$M_r(y, I_k) \,\succeq\, 0 \Leftrightarrow L_y(u^2) \ge 0 \qquad \forall\, u \in \mathbb{R}_r[X(I_k)], \tag{2.5}$$

$$M_r(h\,y, I_k) \,\succeq\, 0 \Leftrightarrow L_y(h\,u^2) \ge 0 \qquad \forall\, u \in \mathbb{R}_r[X(I_k)]. \tag{2.6}$$

3. Main result. Consider problem **P** as defined in (1.1), and recall that $I_0 = \{1,\ldots,n\} = \bigcup_{k=1}^p I_k$ for some subsets $I_k \subset \{1,\ldots,n\}$, $k = 1,\ldots,p$. The subsets $\{I_k\}$ may be read directly from the data or may have been obtained by some procedure, e.g., the one described in Waki et al. [9].

With $\|x\|_\infty$ (resp., $\|x\|$) denoting the usual sup-norm (resp., Euclidean norm) of a vector $x \in \mathbb{R}^n$, we make the following assumption.

Assumption 1. *Let* $\mathbf{K} \subset \mathbb{R}^n$ *be as in* (1.2). *Then, there is* $M > 0$ *such that* $\|x\|_\infty < M$ *for all* $x \in \mathbf{K}$.

In view of Assumption 1, one has $\|X(I_k)\|^2 \leq n_k M^2$, $k = 1, \ldots, p$, and therefore, in the definition (1.2) of $\mathbf{K}$, we add the p redundant quadratic constraints

$$g_{m+k}(X) := n_k M^2 - \|X(I_k)\|^2 \geq 0, \quad k = 1, \ldots, p, \tag{3.1}$$

and set $m' = m + p$, so that $\mathbf{K}$ is now defined by

$$\mathbf{K} := \{x \in \mathbb{R}^n \mid \quad g_j(x) \geq 0, \quad j = 1, \ldots, m'\}. \tag{3.2}$$

Notice that $g_{m+k} \in \mathbb{R}[X(I_k)]$, for every all $k = 1, \ldots, p$.

Assumption 2. Let $\mathbf{K} \subset \mathbb{R}^n$ be as in (3.2). The index set $J = \{1, \ldots, m'\}$ is partitioned into p disjoint sets J_k, $k = 1, \ldots, p$, and the collections $\{I_k\}$ and $\{J_k\}$ satisfy the following:

(i) For every $j \in J_k$, $g_j \in \mathbb{R}[X(I_k)]$, that is, for every $j \in J_k$, the constraint $g_j(X) \geq 0$ is only concerned with the variables $X(I_k) = \{X_i \mid i \in I_k\}$. Equivalently, viewing g_j as a polynomial in $\mathbb{R}[X]$, $g_{j\alpha} \neq 0 \Rightarrow \mathrm{supp}\,(\alpha) \in \mathcal{I}_k$.

(ii) The objective function $f \in \mathbb{R}[X]$ can be written

$$f = \sum_{k=1}^{p} f_k, \quad \text{with } f_k \in \mathbb{R}[X(I_k)], \quad k = 1, \ldots, p. \tag{3.3}$$

Equivalently, $f_\alpha \neq 0 \Rightarrow \mathrm{supp}\,(\alpha) \in \cup_{k=1}^{p} \mathcal{I}_k$.

(iii) Property (1.3) holds.

As already mentioned, (1.3) always holds when $p \leq 2$.

Example 3.1. When $n = 6$ and $m = 6$, let

$$g_1(X) = X_1X_2 - 1; \quad g_2(X) = X_1^2 + X_2X_3 - 1; \quad g_3(X) = X_2 + X_3^2X_4,$$

and

$$g_4(X) = X_3 + X_5; \quad g_5(X) = X_3X_6; \quad g_6(X) = X_2X_3.$$

Then one may choose $p = 4$ with

$$I_1 = \{1, 2, 3\};\ I_2 = \{2, 3, 4\};\ I_3 = \{3, 5\};\ I_4 = \{3, 6\},$$

and $J_1 = \{1, 2, 6\}, J_2 = \{3\}, J_3 = \{4\}, J_4 = \{4\}$. So in Example 3.1, the objective function $f \in \mathbb{R}[X]$ should be a sum of polynomials in $\mathbb{R}[X_1, X_2, X_3]$, $\mathbb{R}[X_2, X_3, X_4]$, $\mathbb{R}[X_3, X_5]$, and $\mathbb{R}[X_3, X_6]$ (also considered as polynomials in $\mathbb{R}[X_1, \ldots, X_6]$).

Remark 3.1. For every $k = 1, \ldots, p$, let

$$\mathbf{K}_k := \{x \in \mathbb{R}^{n_k} : g_j(x) \geq 0 \quad \forall\, j \in J_k\}. \tag{3.4}$$

For every $k = 1, \ldots, p$, the set $\mathbf{K}_k \subset \mathbb{R}^{n_k}$ satisfies Putinar's condition, that is, there exists $u \in \mathbb{R}[X(I_k)]$ which can be written $u = u_0 + \sum_{l \in J_k} u_l\, g_l$ for some s.o.s. polynomials $\{u_0, u_l\} \subset \mathbb{R}[X(I_k)]$, and such that the level set $\{x \in \mathbb{R}^{n_k} : u \geq 0\}$ is compact. (Take $u = g_{m+k}$.) When satisfied, Putinar's condition has the important consequences stated in Theorem 4.1.

3.1. Convergent SDP-relaxations. For each $j = 1, \ldots, m'$, and depending on its parity, write $\deg g_j = 2r_j - 1$ or $2r_j$. Next, with $2r \geq 2r_0 := \max[\deg f, \max_j 2r_j]$, consider the following SDP:

$$\mathbf{Q}_r : \quad \left\{ \begin{array}{lrl} \inf\limits_{y} & L_y(f) & \\ \text{s.t.} & M_r(y, I_k) \succeq 0, & k = 1, \ldots, p \\ & M_{r-r_j}(g_j\, y, I_k) \succeq 0, & j \in J_k; \quad k = 1, \ldots, p \\ & y_0 = 1, & \end{array} \right. \tag{3.5}$$

where the moment and localizing matrices $M_r(y, I_k)$, $M_r(g_j\, y, I_k)$ have been defined at the end of section 2.2. Denote the optimal value of $\mathbf{Q}_r$ by $\inf \mathbf{Q}_r$, and $\min \mathbf{Q}_r$ if the infimum is attained.

Notice that $\mathbf{Q}_r$ is well defined under Assumption 2(i)–(ii). Assumption 2(iii) is only useful to show convergence in Theorem 3.1 below.

The SDP $\mathbf{Q}_r$ is a relaxation of $\mathbf{P}$. Indeed, with $x \in \mathbb{R}^n$ being a feasible solution of $\mathbf{P}$, the moment vector $y = \{y_\alpha\}$ of the Dirac measure $\mu = \delta_x$ at x is feasible for $\mathbf{Q}_r$, with value $L_y(f) = \int f d\mu = f(x)$.

Under Assumption 2, and from the definition of $M_r(y, k)$ and $M_r(g_j\, y, k)$ in section 2.2, the SDP-relaxation $\mathbf{Q}_r$ contains only variables y_α with α in the set

$$\Gamma_r := \left\{ \alpha \in \mathbb{N}^n : \quad \text{supp}\,(\alpha) \in \bigcup_{k=1}^{p} \mathcal{I}_k\,; \quad |\alpha| \leq 2r \right\}. \tag{3.6}$$

Remark 3.2. (i) Maximality of the I_k''s is not required, i.e., one may have $I_j \subset I_k$ for some pair (j, k). In this case, the LMI constraint $M_r(y, I_j) \succeq 0$ is redundant. However, if nondesirable in theory, in practice it may be more convenient to allow for nonmaximality.

(ii) Comparing with the SDP-relaxations of Waki et al. [9]. When the sets $\{I_k\}$ are just the *cliques* $\{C_k\}$ obtained from the chordal extension of the csp graph as defined in [9], then the SDP-relaxations (3.5) are basically the same as those defined in (32) in [9]. The only difference is in the definition of the feasible set $\mathbf{K}$ of $\mathbf{P}$, where we have now included the p redundant quadratic constraints (3.1). In this case, the SDP-relaxations (3.5) are thus stronger than (32) in [9], because they are more constrained.

In view of the definition of the moment matrix $M_r(y, I_k)$, write

$$M_r(y, I_k) \;=\; \sum_{\alpha \in \mathbb{N}^n} y_\alpha B_\alpha^k, \qquad k = 1, \ldots, p,$$

for appropriate symmetric matrices $\{B_\alpha^k\}$, and notice that for every $k = 1, \ldots, p$, one has $B_\alpha^k = 0$ whenever $\text{supp}\,(\alpha) \notin \mathcal{I}_k$. Similarly, for every $k = 1, \ldots, p$, and $j \in J_k$, write

$$M_{r-r_j}(g_j\, y, I_k) \;=\; \sum_{\alpha \in \mathbb{N}^n} y_\alpha C_\alpha^{jk},$$

for appropriate symmetric matrices $\{C_\alpha^{jk}\}$, and notice that $C_\alpha^{jk} = 0$ whenever $\text{supp}\,(\alpha) \notin \mathcal{I}_k$.

The dual SDP $\mathbf{Q}_r^*$ of $\mathbf{Q}_r$, reads

$$\left\{\begin{array}{ll} \sup\limits_{\Omega_k, Z_{jk}, \lambda} & \lambda \\ \text{s.t.} & \sum\limits_{k:\,\mathrm{supp}\,(\alpha)\in\mathcal{I}_k}\left[\langle\Omega_k, B_\alpha^k\rangle + \sum\limits_{j\in J_k}\langle Z_{jk}, C_\alpha^{jk}\rangle\right] + \lambda\,\delta_{\alpha 0} \;=\; f_\alpha \\ & \forall \alpha \in \Gamma_r \\ & \Omega_k, Z_{jk} \succeq 0, \qquad j \in J_k, \quad k = 1,\ldots,p, \end{array}\right. \tag{3.7}$$

where Γ_r is defined in (3.6) and $\delta_{\alpha 0}$ is the usual Kronecker symbol. From an arbitrary feasible solution $(\lambda, \Omega_k, Z_{jk})$ of $\mathbf{Q}_r^*$, multiplying each side of the constraint in (3.7) with X^α for all $\alpha \in \Gamma_r$, and summing up yields

$$\sum_{\alpha\in\Gamma_r}\left[\sum_{k:\,\mathrm{supp}\,(\alpha)\in\mathcal{I}_k}\left(\langle\Omega_k, B_\alpha^k X^\alpha\rangle + \sum_{j\in J_k}\langle Z_{jk}, C_\alpha^{jk} X^\alpha\rangle\right)\right] \;=\; f(X) - \lambda,$$

which, denoting $\Gamma_{kr} := \{\alpha \in \mathbb{N}^n : \mathrm{supp}\,(\alpha) \in \mathcal{I}_k;\ |\alpha| \le 2r\}$, can be rewritten

$$\sum_{k=1}^{p}\left[\left\langle \Omega_k, \sum_{\alpha\in\Gamma_{kr}} B_\alpha^k X^\alpha\right\rangle + \sum_{j\in J_k}\left\langle Z_{jk}, \sum_{\alpha\in\Gamma_{kr}} C_\alpha^{jk} X^\alpha\right\rangle\right] \;=\; f(X) - \lambda. \tag{3.8}$$

Proceeding as in Lasserre [11], and using the spectral decomposition of matrices $\Omega_k, Z_{jk} \succeq 0$, write

$$\Omega_k = \sum_l \mathbf{q}_{kl}\mathbf{q}_{kl}', \quad Z_{jk} = \sum_t \mathbf{q}_{jkt}\mathbf{q}_{jkt}', \quad j \in J_k, \quad k = 1,\ldots,p,$$

for some vectors $\{\mathbf{q}_{kl}, \mathbf{q}_{jkt}\}$. Next, notice that

$$\sum_{\alpha\in\Gamma_{kr}} B_\alpha^k X^\alpha \;=\; v_r(X(I_k))\, v_r(X(I_k))', \qquad k = 1,\ldots,p \tag{3.9}$$

(recall that $v_r(X(I_k))$ is the canonical basis of $\mathbb{R}_r[X(I_k)]$). Similarly, for every $k = 1,\ldots,p$, and $j \in J_k$,

$$\sum_{\alpha\in\Gamma_{kr}} C_\alpha^{jk} X^\alpha \;=\; g_j(X)\, v_{r-r_j}(X(I_k))\, v_{r-r_j}(X(I_k))'. \tag{3.10}$$

In view of the dimension of the matrix Ω_k (resp., Z_{jk}), one may identify $\mathbf{q}_{kl}$ (resp., $\mathbf{q}_{jkt}$) with the vector of coefficients of a polynomial $q_{kl} \in \mathbb{R}_r[X(I_k)]$ (resp., $q_{jkt} \in \mathbb{R}_{r-r_j}[X(I_k)]$), and so for every l, t

$$\langle v_r(X(I_k)), \mathbf{q}_{kl}\rangle \;=\; q_{kl}(X), \quad k = 1,\ldots,p,$$

$$\langle v_{r-r_j}(X(I_k)), \mathbf{q}_{jkt}\rangle \;=\; q_{jkt}(X), \quad j \in J_k, \quad k = 1,\ldots,p.$$

Combining the latter with (3.8)–(3.10), one may rewrite (3.8) as

$$\sum_{k=1}^{p}\left[\sum_l q_{kl}(X)^2 + \sum_{j\in J_k} g_j(X)\sum_t q_{jkt}(X)^2\right] \;=\; f(X) - \lambda.$$

In other words,

$$f-\lambda \;=\; \sum_{k=1}^{p}\left(\,q_k+\sum_{j\in J_k} q_{jk}\,g_j\,\right), \tag{3.11}$$

for some s.o.s. polynomials $q_k, q_{jk}\in\mathbb{R}[X(I_k)]$, $k=1,\ldots,p$, a *sparse* version of Putinar's representation [16] for the polynomial $f-\lambda$, nonnegative on $\mathbf{K}$.

Finally, in view of what precedes, the dual $\mathbf{Q}_r^*$ also reads

$$\left\{\begin{array}{lll} \sup\limits_{q_k,q_{jk},\lambda} & \lambda & \\ \text{s.t.} & f-\lambda \;=\; \sum\limits_{k=1}^{p}\left(\,q_k+\sum\limits_{j\in J_k} q_{jk}\,g_j\,\right) & \\ & q_k,\,q_{jk}\in\mathbb{R}[X(I_k)] \text{ and s.o.s.}, & j\in J_k,\quad k=1,\ldots,p \\ & \deg q_k,\,\deg q_{jk}g_j\le 2r, & j\in J_k,\quad k=1,\ldots,p. \end{array}\right. \tag{3.12}$$

THEOREM 3.1. *Let* $\mathbf{P}$ *be as defined in* (1.1)*, with global minimum denoted* $\min\mathbf{P}$*, and let Assumptions* 1 *and* 2 *hold. Let* $\{\mathbf{Q}_r\}$ *be the hierarchy of SDP-relaxations defined in* (3.5)*. Then the following hold:*

(a) $\inf\mathbf{Q}_r\uparrow\min\mathbf{P}$ *as* $r\to\infty$.

(b) *If* $\mathbf{K}$ *has a nonempty interior, then there is no duality gap between* $\mathbf{Q}_r$ *and its dual* $\mathbf{Q}_r^*$*, and* $\mathbf{Q}_r^*$ *is solvable for sufficiently large* r*, i.e.,* $\inf\mathbf{Q}_r=\max\mathbf{Q}_r^*$.

(c) *Let* y^r *be a nearly optimal solution of* $\mathbf{Q}_r$*, with, e.g.,*

$$L_{y^r}(f)\;\le\;\inf\mathbf{Q}_r+\frac{1}{r}\qquad\forall\, r\ge r_0,$$

and let $\widehat{y}^r:=\{y^r_\alpha:\;|\alpha|=1\}$*. If* $\mathbf{P}$ *has a unique global minimizer* $x^*\in\mathbf{K}$*, then*

$$\widehat{y}^r\;\to\;x^*\qquad\text{as } r\to\infty. \tag{3.13}$$

For a proof, see section 4.1. Theorem 3.1 establishes convergence of the hierarchy of SDP-relaxations to the global minimum $\min\mathbf{P}$, as well as convergence to a global minimizer $x^*\in\mathbf{K}$ (if unique).

3.2. Computational complexity. The number of variables for the SDP-relaxation $\mathbf{Q}_r$ defined in (3.5) is bounded by $\sum_{k=1}^{p}\binom{n_k+2r}{2r}$, and so, if all n_k's are *close* to each other, say $n_k\approx n/p$ for all k, then one has at most $O(p(\frac{n}{p})^{2r})$ variables, a big saving when compared with $O(n^{2r})$ in the original SDP-relaxations defined in [11] and implemented in [5].

In addition, one also has p LMI constraints of size $O((\frac{n}{p})^r)$ and $m+p$ LMI constraints of size $O((\frac{n}{p})^{r-r'})$ (where $2r'$ is the largest degree of the polynomials g_j's), to be compared with a single LMI constraint of size $O(n^r)$ and m LMI constraints of size $O(n^{r-r'})$ in [5, 11]. So, for instance, when using an interior point method, it is definitely better to handle p LMIs, each of size $(n/p)^r$, rather than a single LMI of size n^r.

For illustration purposes, consider the following elementary example. Let $n = 4$, and consider the optimization problem

$$\mathbf{P}: \quad \begin{cases} \inf\limits_{x} & x_1x_2 + x_1x_3 + x_1x_4 \\ \text{s.t.} & x_1^2 + x_2^2 \le a_{12} \\ & x_1^2 + x_3^2 \le a_{13} \\ & x_1^2 + x_4^2 \le a_{14}. \end{cases}$$

Hence, $I_1 = \{1,2\}$, $I_2 = \{1,3\}$, $I_3 = \{1,4\}$. The first SDP-relaxation $\mathbf{Q}_1$ in the hierarchy is obtained with $r = 1$ and reads

$$\inf_y \; y_{1100} + y_{1010} + y_{1001}$$

$$\begin{bmatrix} 1 & y_{1000} & y_{0100} \\ y_{1000} & y_{2000} & y_{1100} \\ y_{0100} & y_{1100} & y_{0200} \end{bmatrix}, \begin{bmatrix} 1 & y_{1000} & y_{0010} \\ y_{1000} & y_{2000} & y_{1010} \\ y_{0010} & y_{1010} & y_{0020} \end{bmatrix}, \begin{bmatrix} 1 & y_{1000} & y_{0001} \\ y_{1000} & y_{2000} & y_{1001} \\ y_{0001} & y_{1001} & y_{0002} \end{bmatrix} \succeq 0$$

$$a_{12} - y_{2000} - y_{0200} \ge 0;\; a_{13} - y_{2000} - y_{0020} \ge 0;\; a_{14} - y_{2000} - y_{0002} \ge 0.$$

3.3. Extraction of solutions. As for the standard SDP-relaxations of [11], one may also detect global optimality, i.e., when $\min \mathbf{Q}_{s_0} = \min \mathbf{P}$ for some s_0, in which case *finite* convergence occurs, and the SDP-relaxation $\mathbf{Q}_{s_0}$ is said to be *exact*. Recall that for the standard SDP-relaxations [11], one has defined a rank-test to detect finite convergence (see, e.g., Lasserre [12]), as well as an *extraction procedure* (applied to the moment matrix of an exact SDP-relaxation) to obtain one or several global minimizers $x^* \in \mathbb{R}^n$ of $\mathbf{P}$; for more details, see Henrion and Lasserre [5, 6].

For all j,k with $I_{jk} := I_j \cap I_k \neq \emptyset$, denote by $\mathcal{I}_{jk}$ the set of subsets of I_{jk}. Let $M_r(y, I_{jk})$ be the submatrix obtained from $M_r(y, I_j)$ or $M_r(y, I_k)$, by selecting only those rows and columns $\alpha \in \mathbb{N}^n$, with $\operatorname{supp}(\alpha) \in \mathcal{I}_{jk}$ and $|\alpha| \le r$.

THEOREM 3.2. *Let Assumption* 2(i)–(ii) *hold, and let* $\{\mathbf{Q}_r\}$ *be the hierarchy of SDP-relaxations defined in* (3.5). *Let* $a_k := \max_{j \in J_k}[r_j]$, *for all* $k = 1, \ldots, p$, *and assume that* y *is an optimal solution of* $\mathbf{Q}_{s_0}$ *for some* s_0.

The SDP-relaxation $\mathbf{Q}_{s_0}$ *is exact, i.e.,* $\min \mathbf{Q}_{s_0} = \min \mathbf{P}$, *if*

$$\operatorname{rank} M_{s_0}(y, I_k) \;=\; \operatorname{rank} M_{s_0 - a_k}(y, I_k), \qquad k = 1, \ldots, p, \tag{3.14}$$

and if $\operatorname{rank} M_{s_0}(y, I_{jk}) = 1$ *for all pairs* (j,k) *with* $I_j \cap I_k \neq \emptyset$.

Moreover, let $\Delta_k := \{x^*(k)\} \subset \mathbb{R}^{n_k}$ *be a set of solutions obtained from the extraction procedure applied to each moment matrix* $M_{s_0}(y, I_k)$, $k = 1, \ldots, p$. *Then every* $x^* \in \mathbb{R}^n$ *obtained by* $(x_i^*)_{i \in I_k} = x^*(k)$ *for some* $x^*(k) \in \Delta_k$ *is an optimal solution of* $\mathbf{P}$.

For a proof, see section 4.2.

Remark 3.3. In Theorem 3.2, Assumption 2(iii) is not needed. In addition, it also holds even if the SDP-relaxations are defined with the original set $\mathbf{K}$ defined in (1.2) instead of (3.2), i.e., without the additional quadratic constraints (3.1). And so, Theorem 3.2 is also valid for noncompact sets $\mathbf{K}$, provided Assumption 2(i)–(ii) holds true.

3.4. A sparse representation result. As a by-product of Theorem 3.1, we obtain the following representation result.[1]

[1] In a recent note [10], Kojima and Maramatsu have improved Corollary 3.3 and show the same result without assuming that $\mathbf{K}$ has a nonempty interior.

COROLLARY 3.3. *Let* $\mathbf{K}$ *be as in* (3.2) *with the additional quadratic constraints* (3.1)*, and with nonempty interior. Let Assumption* 2 *hold. If* $f \in \mathbb{R}[X]$ *is strictly positive on* $\mathbf{K}$*, then*

$$f = \sum_{k=1}^{p} \left(q_k + \sum_{j \in J_k} q_{jk}\, g_j \right), \tag{3.15}$$

for some s.o.s. polynomials $q_k, q_{jk} \in \mathbb{R}[X(I_k)]$*,* $k = 1, \ldots, p$*.*

Proof. Let $f \in \mathbb{R}[X]$ be strictly positive on $\mathbf{K}$, and let $f^* > 0$ be its global minimum on $\mathbf{K}$. From Theorem 3.1(a)–(b), we have $\inf \mathbf{Q}_r = \max \mathbf{Q}_r^* \uparrow f^*$, as $r \to \infty$. Therefore, let $r \in \mathbb{N}$ be such that $\max \mathbf{Q}_r^* \geq f^*/2 > 0$, and as $\mathbf{Q}_r^*$ is solvable, let (q_k, q_{jk}, λ) be an arbitrary optimal solution, so that $\max \mathbf{Q}_r^* = \lambda > 0$. From that solution, one obtains (3.11), i.e.,

$$f - \lambda = \sum_{k=1}^{p} \left(q_k + \sum_{j \in J_k} q_{jk}\, g_j \right),$$

for some s.o.s. polynomials $q_k, q_{jk} \in \mathbb{R}[X(I_k)]$, $k = 1, \ldots, p$ (associated with the optimal solution (q_k, q_{jk}, λ) of $\mathbf{Q}_r^*$. But then,

$$f = \lambda + \sum_{k=1}^{p} \left(q_k + \sum_{j \in J_k} q_{jk}\, g_j \right)$$

becomes the desired result (by adding $\lambda > 0$ to one of the s.o.s. polynomials q_k). □

Observe that (3.15) is a sparse version of Putinar's representation for polynomials strictly positive on $\mathbf{K}$; see Theorem 4.1. Indeed, (3.15) is a certificate of nonnegativity of f on $\mathbf{K}$. Finally, Corollary 3.3 also holds if $\mathbf{K}$ is such that for every $k = 1, \ldots, p$, $\mathbf{K}_k$ satisfies Putinar's condition (so that there is no need of the quadratic constraints (3.1)).

3.5. Examples. We provide here some examples considered in Waki et al. [9].

Example 3.2. The chained singular function. With n a multiple of 4,

$$I_k = \{k, k+1, k+2, k+3\}, \quad k = 1, \ldots, n-3,$$

and the sparsity pattern satisfies (1.3). One has $\kappa = 4$.

Example 3.3. The Broyden banded function. In this case,

$$I_k = \{k, k+1, \ldots, \min[k+6, n]\}, \quad k = 1, \ldots, n,$$

and the sparsity pattern also satisfies (1.3). One has $\kappa = 7$.

Example 3.4. The Broyden tridiagonal function. In this case

$$I_k = \{k, k+1, \min[n, k+2]\}, \quad k = 1, \ldots, n,$$

and the sparsity pattern also satisfies (1.3). One has $\kappa = 3$.

Example 3.5. The chained Wood function. In this case, with n a multiple of 4,

$$I_k = \{k, k+1, k+2, k+3\}, \quad k = 1, \ldots, n-3,$$

and the sparsity pattern also satisfies (1.3). One has $\kappa = 4$.

Example 3.6. The generalized Rosenbrock function. In this case,

$$I_k = \{k, k-1\}, \quad k = 2, \ldots, n,$$

and the sparsity pattern also satisfies (1.3).

Example 3.7. The optimal control problem (38) considered in [9]. In this case,

$$I_k = \{\{y_{k,j}\}_{j=1}^{n_y}, \{x_{k,l}\}_{l=1}^{n_x}\}, \quad k = 1, \ldots, M-1,$$

$I_M = \{\{y_{M,j}\}_{j=1}^{n_y}\}$, and the sparsity pattern also satisfies (1.3). One has $\kappa = n_x \times n_y$. Example 3.7 is typical of what we call *strong coupling*, always the case in discrete-time optimal control problems. Indeed, the *control* variables at each period are *independent*, whereas the coupling of periods is done through the *state* equations (i.e., the dynamics) and via the *state* variables.

In view of Remark 3.2, the SDP-relaxations (3.5) are stronger than (32) in [9], when the sets $\{I_k\}$ are the same as the cliques $\{C_k\}$ in [9], which is the case in all the previous examples, for which Waki et al. [9] report excellent numerical results; in particular, problems of large size that could not be handled via the standard SDP-relaxations of [11] have been solved relatively easily.

Indeed, for instance, in Examples 3.4, 3.5, and 3.6, they have solved problems with up to $n = 500$ variables, a remarkable result! For the interested reader, more details and numerical results can be found in [9].

4. Proofs. We first restate Putinar's theorem, which is crucial in the proof of Theorem 3.1 below.

THEOREM 4.1 (see Putinar [16]). *Let $\mathbf{K} \subset \mathbb{R}^n$ be a compact basic semialgebraic set as defined in* (1.2)*, and let $y = (y_\alpha)_{\alpha \in \mathbb{N}^n}$ be given. Let $M_r(y)$ and $M_r(g_j\, y)$ be the moment and localizing matrices defined in section* 2*. Assume that there exists $u \in \mathbb{R}[X]$ such that $u = u_0 + \sum_{j=1}^m u_j\, g_j$ for some s.o.s. polynomials $\{u_j\}_{j=0}^m \subset \Sigma^2$, and such that the level set $\{x : u(x) \geq 0\}$ is compact.*

(a) *If $h \in \mathbb{R}[X]$ is strictly positive on $\mathbf{K}$, then $h = h_0 + \sum_{j=1}^m h_j\, g_j$ for some s.o.s. polynomials $\{h_j\}_{j=0}^m \subset \Sigma^2$.*

(b) *If $M_r(y) \succeq 0$ and $M_r(g_j\, y) \succeq 0$ for all $j = 1, \ldots, m$, and all $r = 0, 1, \ldots$, then y has a representing measure μ with support contained in $\mathbf{K}$.*

4.1. Proof of Theorem 3.1. (a) We first prove that $\mathbf{Q}_r$ has a feasible solution. Recall the definitions

$$\begin{aligned}
\Gamma_{kr} &:= \{\, \alpha \in \mathbb{N}^n \,:\, \mathrm{supp}\,(\alpha) \in \mathcal{I}_k; \quad |\alpha| \leq 2r \,\}, \quad k = 1, \ldots, p,\\
\Gamma_r &:= \bigcup_{k=1}^p \Gamma_{kr} \,=\, \left\{ \alpha \in \mathbb{N}^n \,:\, \mathrm{supp}\,(\alpha) \in \bigcup_{k=1}^p \mathcal{I}_k; \quad |\alpha| \leq 2r \right\},\\
\Gamma &:= \bigcup_{r \in \mathbb{N}} \Gamma_r \,=\, \left\{ \alpha \in \mathbb{N}^n \,:\, \mathrm{supp}\,(\alpha) \in \bigcup_{k=1}^p \mathcal{I}_k \right\}.
\end{aligned}$$

Let $\nu := \delta_x$ be the Dirac measure at a feasible solution $x \in \mathbf{K}$ of $\mathbf{P}$, and let

$$y_\alpha = \int X^\alpha\, d\nu \qquad \forall\, \alpha \in \Gamma_r.$$

Recalling the definition of $M_r(y, I_k)$ and $M_{r-r_j}(g_j\, y, I_k)$ in section 2.2, one has $M_r(y, I_k) \succeq 0$ and $M_{r-r_j}(g_j\, y, I_k) \succeq 0$; therefore, y is an obvious feasible solution of $\mathbf{Q}_r$. Next we prove that $\inf \mathbf{Q}_r > -\infty$ for all sufficiently large r.

Recall that $2r_0 \geq \max[\deg f, \max_j \deg r_j]$. In view of Assumption 1 and from the definition of the set $\mathbf{K}_k$ in (3.4), there exists N such that $N \pm X^\alpha > 0$ on $\mathbf{K}_k$ for all $\alpha \in \Gamma_{kr_0}$, and all $k = 1, \ldots, p$. Therefore, for every $k = 1, \ldots, p$ and $\alpha \in \Gamma_{kr_0}$, the polynomial $N \pm X^\alpha$ belongs to the quadratic module $Q_k \subset \mathbb{R}[X(I_k)]$ generated by $\{g_j\}_{j \in J_k} \subset \mathbb{R}[X(I_k)]$, i.e.,

$$Q_k := \left\{ \sigma_0 + \sum_{j \in J_k} \sigma_j \, g_j \; : \quad \sigma_j \text{ s.o.s. in } \in \mathbb{R}[X(I_k)] \quad \forall \, j \in \{0\} \cup J_k \right\}.$$

But there is even some $l(r_0)$ such that $N \pm X^\alpha \in Q_k(l(r_0))$ for all $\alpha \in \Gamma_{kr_0}$ and $k = 1, \ldots, p$, where $Q_k(t) \subset Q_k$ is the set of elements of Q_k which have a representation $\sigma_0 + \sum_{j \in J_k} \sigma_j \, g_j$ for some s.o.s. $\{\sigma_j\} \subset \mathbb{R}[X(I_k)]$ with $\deg \sigma_0 \leq 2t$ and $\deg \sigma_j g_j \leq 2t$ for all $j \in J_k$. Of course we also have $N \pm X^\alpha \in Q_k(l)$ for all $\alpha \in \Gamma_{kr_0}$, whenever $l \geq l(r_0)$. Therefore, let us take $l(r_0) \geq r_0$.

For every feasible solution y of $\mathbf{Q}_{l(r_0)}$ one has

$$| \, L_y(X^\alpha) \, | \leq N, \qquad \alpha \in \Gamma_{kr_0}; \quad k = 1, \ldots, p.$$

This follows from $y_0 = 1$, $M_{l(r_0)}(y, I_k) \succeq 0$ and $M_{l(r_0)-r_j}(g_j \, y, I_k) \succeq 0$, which implies

$$L_y(N \pm X^\alpha) \; = \; L_y(\sigma_0) + \sum_{j \in J_k} L_y(\sigma_j \, g_j) \geq 0$$

because the σ_j's are s.o.s. (see (2.5) and (2.6)).

As $2r_0 \geq \deg f$, it follows that $L_y(f) \geq -N \sum_\alpha |f_\alpha|$. This is because by Assumption 2(ii), $f_\alpha \neq 0 \Rightarrow \alpha \in \Gamma_{r_0}$. Hence $\inf \mathbf{Q}_{l(r_0)} > -\infty$.

So from what precedes, and with $s \in \mathbb{N}$ arbitrary, let $l(s) \geq s$ be such that

$$N_s \pm X^\alpha \in Q_k(l(s)) \qquad \forall \, \alpha \in \Gamma_{ks}; \quad k = 1, \ldots, p, \tag{4.1}$$

for some N_s. Next, let $r \geq l(r_0)$ (so that $\inf \mathbf{Q}_r > -\infty$), and let y^r be a nearly optimal solution of $\mathbf{Q}_r$ with value

$$\inf \mathbf{Q}_r \; \leq \; L_{y^r}(f) \; \leq \; \inf \mathbf{Q}_r + \frac{1}{r} \quad \left(\leq \min \mathbf{P} + \frac{1}{r} \right). \tag{4.2}$$

Fix $s \in \mathbb{N}$. Notice that from (4.1), for all $r \geq l(s)$, one has

$$| \, L_{y^r}(X^\alpha) \, | \; \leq \; N_s \qquad \forall \, \alpha \in \Gamma_s.$$

Therefore, for all $r \geq r_0$,

$$|y^r_\alpha| \; = \; | \, L_{y^r}(X^\alpha) \, | \; \leq \; N'_s \quad \forall \, \alpha \in \Gamma_s, \tag{4.3}$$

where $N'_s = \max[N_s, V_s]$, with

$$V_s := \max \{ |y^r_\alpha| \; : \quad \alpha \in \Gamma_s; \; r_0 \leq r \, < \, l(s) \}.$$

Complete each y^r with zeros to make it an infinite vector in l_∞, indexed in the canonical basis $v_\infty(X)$ of $\mathbb{R}[X]$. Notice that $y^r_\alpha \neq 0$ only if $\alpha \in \Gamma$.

In view of (4.3), one has

$$|y^r_\alpha| \leq N'_s \qquad \forall \, \alpha \in \Gamma; \quad 2s - 1 \leq |\alpha| \leq 2s \tag{4.4}$$

for all $s = 1, 2, \ldots$.

Hence, define the new sequence $\widehat{y}^r \in l_\infty$ defined by $\widehat{y}_0 := 1$, and

$$\widehat{y}^r_\alpha := \frac{y^r_\alpha}{N'_s} \qquad \forall\, \alpha \in \Gamma, \quad 2s-1 \le |\alpha| \le 2s$$

for all $s = 1, 2, \ldots$, and in l_∞, consider the sequence $\{\widehat{y}^r\}$ as $r \to \infty$.

Obviously, the sequence $\{\widehat{y}^r\}$ is in the unit ball B_1 of l_∞, and so, by the Banach–Alaoglu theorem (see, e.g., Ash [1, Thm. 3.5.16]), there exists $\widehat{y} \in B_1$ and a subsequence $\{r_i\}$, such that $\widehat{y}^{r_i} \to \widehat{y}$ as $i \to \infty$ for the weak $\star$ topology $\sigma(l_\infty, l_1)$ of l_∞. In particular, pointwise convergence holds, that is,

$$\lim_{i\to\infty} \widehat{y}^{r_i}_\alpha \to \widehat{y}_\alpha, \qquad \alpha \in \mathbb{N}^n.$$

Notice that $\widehat{y}_\alpha \neq 0$ only if $\alpha \in \Gamma$. Next, define $y_0 := 1$ and

$$y_\alpha := \widehat{y}_\alpha \times N'_s, \quad 2s-1 \le |\alpha| \le 2s, \quad s = 1, 2, \ldots .$$

The pointwise convergence $\widehat{y}^{r_i} \to \widehat{y}$ implies the pointwise convergence $y^{r_i} \to y$, i.e.,

$$\lim_{i\to\infty} y^{r_i}_\alpha \to y_\alpha \qquad \forall\, \alpha \in \Gamma. \tag{4.5}$$

Let $s \in \mathbb{N}$ be fixed. From the pointwise convergence (4.5), we deduce that

$$\lim_{i\to\infty} M_s(y^{r_i}, I_k) \,=\, M_s(y, I_k) \succeq 0, \qquad k = 1, \ldots, p.$$

Similarly

$$\lim_{i\to\infty} M_s(g_j\, y^{r_i}, I_k) \,=\, M_s(g_j\, y, I_k) \succeq 0, \quad j \in J_k, \qquad k = 1, \ldots, p.$$

As s was arbitrary, we obtain that for all $k = 1, \ldots, p$,

$$M_r(y, I_k) \succeq 0; \quad M_r(g_j\, y, I_k) \succeq 0, \quad j \in J_k; \quad r = 0, 1, 2, \ldots . \tag{4.6}$$

Introduce the subsequence y^k obtained from y by

$$y^k := \{\, y_\alpha \,:\, \mathrm{supp}\,(\alpha) \,\in\, \mathcal{I}_k \,\} \qquad \forall\, k = 1, \ldots, p. \tag{4.7}$$

Recall that $M_r(y, I_k)$ (resp., $M_r(g_j\, y, I_k)$) is also the moment matrix $M_r(y^k)$ (resp., the localizing matrix $M_r(g_j\, y^k)$) for the sequence y^k indexed in the canonical basis $v_\infty(X(I_k))$ of $\mathbb{R}[X(I_k)]$; see section 2.2.

Therefore, by Remark 3.1, (4.6) implies that y^k has a representing measure ν_k with support contained in $\mathbf{K}_k$, $k = 1, \ldots, p$; see Theorem 4.1. As $y^k_0 = 1$, ν_k is a probability measure on $\mathbf{K}_k$ for all $k = 1, \ldots, p$.

Next, let j, k be such that $I_{jk} := I_j \cap I_k \neq \emptyset$, and recall that $\mathcal{I}_{jk}$ is the set of all subsets of I_{jk}. Let $m_{jk} := \mathrm{card}\,(I_j \cup I_k)$ and let $n_{jk} := \mathrm{card}\,(I_j \cap I_k)$. Define $\pi_j : \mathbb{R}^{m_{jk}} \to \mathbb{R}^{n_j}$, $\pi_k : \mathbb{R}^{m_{jk}} \to \mathbb{R}^{n_k}$, and $\pi_{jk} : \mathbb{R}^{m_{jk}} \to \mathbb{R}^{n_{jk}}$, the natural projections with respect to the variables $\{X_i \,|\, i \in I_j\}$, $\{X_i \,|\, i \in I_k\}$, and $\{X_i \,|\, i \in I_j \cap I_k\}$, respectively. Let $\mathbf{K}_{j\vee k} \subset \mathbb{R}^{m_{jk}}$ and $\mathbf{K}_{j\wedge k} \subset \mathbf{K}_{j\vee k}$ be the compact sets

$$\mathbf{K}_{j\vee k} := \{\, x \in \mathbb{R}^{m_{jk}} \,:\, \pi_j(x) \,\in\, \mathbf{K}_j; \quad \pi_k(x) \,\in\, \mathbf{K}_k \,\}; \quad \mathbf{K}_{j\wedge k} := \pi_{jk}(\mathbf{K}_{j\vee k}).$$

The probability measures ν_j and ν_k can be understood as probability measures on $\mathbf{K}_{j\vee k}$, supported on $\mathbf{K}_j = \pi_j(\mathbf{K}_{j\vee k})$ and $\mathbf{K}_k = \pi_k(\mathbf{K}_{j\vee k})$, respectively.

Observe that from the definition (4.7) of y^j and y^k, one has

$$y^j_\alpha \,=\, y^k_\alpha \qquad \forall\, \alpha \text{ with } \mathrm{supp}\,(\alpha) \in \mathcal{I}_{jk},$$

and as measures on compact sets are moment determinate, it follows that the marginal probability measures of ν_j and ν_k on $\mathbf{K}_{j\wedge k}$ (i.e., with respect to the variables $X = \{X_i \,|\, i \in I_{jk}\}$) are the *same* probability measure, denoted ν_{jk}. That is,

$$y^k_\alpha \,=\, y^j_\alpha \,=\, \int X^\alpha \, d\nu_{jk} \qquad \forall\, \alpha \text{ with } \mathrm{supp}\,(\alpha) \in \mathcal{I}_{jk}.$$

From Lemma 6.4, there exists a probability measure μ on $\mathbf{K}$, constructed from the ν_k's, and with marginal ν_k on $\mathbf{K}_k$ for all $k = 1, \ldots, p$. In particular, this implies

$$y_\alpha \,=\, \int X^\alpha \, d\mu \qquad \forall\, \alpha \in \Gamma. \tag{4.8}$$

Recall that by Assumption 2, $f_\alpha \neq 0 \Rightarrow \alpha \in \Gamma$, and so $L_y(f) = \int f\, d\mu$. On the other hand, from (4.2) and the pointwise convergence (4.5),

$$\min \mathbf{P} \,\geq\, \lim_{i\to\infty} \inf \mathbf{Q}_{r_i} \,=\, \lim_{i\to\infty} L_{y^{r_i}}(f) \,=\, L_y(f) \,=\, \int f\, d\mu.$$

But as μ is supported on $\mathbf{K}$, we necessarily have $\int f\, d\mu \,\geq\, f^* \,=\, \min \mathbf{P}$, and so $\min \mathbf{P} \,=\, \int f\, d\mu$. Therefore, we have proved that $\lim_{i\to\infty} \inf \mathbf{Q}_{r_i} \,=\, \min \mathbf{P}$, and so $\inf \mathbf{Q}_r \,\uparrow\, \min \mathbf{P}$ follows because the sequence $\{\inf \mathbf{Q}_r\}$ is monotone nondecreasing. This completes the proof of (a).

(b) In the feasible solution ν that we have constructed at the beginning of the proof of (a), choose now ν to be *uniform* on $\mathbf{K}$, and let $y = \{y_\alpha\}_{\alpha\in\mathbb{N}^n}$ be the vector of all its moments, well defined because $\mathbf{K}$ is compact. As $\mathbf{K}$ has a nonempty interior, the probability measure ν satisfies $M_r(y) \succ 0$ and $M_r(g_j y) \succ 0$, for all $j = 1, \ldots, m$, and all $r = 0, 1, \ldots$.

Then, obviously, $M_r(y, I_k) \succ 0$ (resp., $M_r(g_j\, y, I_k) \succ 0$, $j \in J_k$) as a submatrix of $M_r(y) \succ 0$ (resp., $M_r(g_j\, y) \succ 0$), for all $k = 1, \ldots, p$.

Hence, the feasible solution y is now strictly feasible, i.e., Slater's condition holds for $\mathbf{Q}_r$. This implies the absence of a duality gap between $\mathbf{Q}_r$ and its dual $\mathbf{Q}^*_r$, and as $\inf \mathbf{Q}_r > -\infty$ for sufficiently large r, $\mathbf{Q}^*_r$ is solvable, i.e., $\inf \mathbf{Q}_r = \sup \mathbf{Q}^*_r = \max \mathbf{Q}^*_r$. This completes the proof of (b).

(c) Finally, let $x^* \in \mathbf{K}$ be the unique global minimizer of $\mathbf{P}$, and let y^r be as in Theorem 3.1(c). From (a) there exists a subsequence y^{r_i} for which we have the pointwise convergence $y^{r_i} \,\to\, y$ (see (4.5)), where y is the moment sequence of a probability measure μ on $\mathbf{K}$. In particular, (4.8) holds and $\min \mathbf{P} = \int f\, d\mu$. From uniqueness of the global minimizer $x^* \in \mathbf{K}$, it follows that $\mu = \delta_{x^*}$ (the Dirac measure at $x^* \in \mathbf{K}$). But then (4.8) yields

$$\lim_{i\to\infty} y^{r_i}_\alpha \,=\, y_\alpha \,=\, \int X^\alpha \, d\mu \,=\, (x^*)^\alpha \qquad \forall \alpha \in \Gamma.$$

Taking $\alpha \in \Gamma$ with $|\alpha| = 1$ yields $\widehat{y}^{r_i} \,\to\, x^*$, and as the converging subsequence was arbitrary, it follows that the whole sequence $\widehat{y}^r$ converges to $x^* \in \mathbf{K}$, the desired result. □

4.2. Proof of Theorem 3.2. Let $\gamma_k := \operatorname{rank} M_{s_0}(y, I_k)$, $k = 1, \ldots, p$. From (3.14) the vector $y^k = \{y^k_\alpha\}$ defined in (4.7) (with $|\alpha| \le 2s_0$) is the vector of moments (up to order $2s_0$) of a γ_k-atomic probability measure ν_k supported on $\mathbf{K}_k \subset \mathbb{R}^{n_k}$, with $\mathbf{K}_k$ being defined in (3.4), $k = 1, \ldots, p$. This follows from a result of Curto and Fialkow [3, Thm. 1.6] already used in Lasserre [12] to prove finite convergence of SDP-relaxations for 0-1 programs; see also Laurent [14] for a shorter proof of [3, Thm. 1.6], and related comments.

Therefore, when applying the extraction procedure defined in [6] to the moment matrix $M_{s_0}(y^k) = M_{s_0}(y, I_k)$, $k = 1, \ldots, p$, one obtains sets of vectors $\Delta_k := \{x^l(k)\}_{l=1}^{\gamma_k} \subset \mathbf{K}_k$ for all $k = 1, \ldots, p$.

With $\delta_\bullet$ denoting the Dirac measure at $\bullet$, one may thus write

$$\nu_k \;=\; \sum_{l=1}^{\gamma_k} p_{kl}\, \delta_{x^l(k)}, \qquad \text{for some} \qquad p_{kl} > 0 \quad \forall\, l; \quad \sum_{l=1}^{\gamma_k} p_{kl} = 1$$

for all $k = 1, \ldots, p$.

But then, pick *any* solution $x^{l_k}(k) \in \Delta_k$, for some l_k, $k = 1, \ldots, p$, and define $x^* \in \mathbb{R}^n$ to be the vector such that

$$x^*(k) := \{x^*_i\}_{i \in I_k} \;=\; x^{l_k}(k); \quad k = 1, \ldots, p. \tag{4.9}$$

There is no ambiguity for x^*_i when $i \in I_j \cap I_k \neq \emptyset$ for some $j, k \in \{1, \ldots, p\}$, because in this case, from $\operatorname{rank} M_{s_0}(y, I_{jk}) = 1$, we deduce that $y^{jk} = \{y_\alpha\}$ with $\operatorname{supp}(\alpha) \in \mathcal{I}_j \cap \mathcal{I}_k$, is the vector of moments (up to order $2s_0$) of some Dirac measure ν_{jk}. As in the proof of (a), ν_{jk} is the marginal of ν_k and ν_j on $\mathbf{K}_{j \wedge k}$ (i.e., with respect to the variables $\{X_i : i \in I_j \cap I_k\}$), and so the Dirac measure at some point denoted $x(j \wedge k) \in \mathbf{K}_{j \wedge k}$.

Hence, for any two choices $x^{l_j}(j) \in \Delta_j$ and $x^{l_k}(k) \in \Delta_k$, the point $x^* \in \mathbb{R}^n$ defined in (4.9) is in $\mathbf{K}$. We can thus construct $s := \prod_{k=1}^p \gamma_k$ solutions $\{x^\omega\}_{\omega=1}^s \subset \mathbf{K}$, each associated with the probability $p_\omega := \prod_{k=1}^p p_{kl_k}$ if $x^\omega(k) = x^{l_k}(k) \in \Delta_k$, for some $l_k \in \{1, \ldots, \gamma_k\}$, $k = 1, \ldots, p$. But then, by construction, the probability measure μ on $\mathbb{R}^n$, defined by

$$\mu := \sum_{\omega=1}^{s} p_\omega \delta_{x^\omega},$$

is supported on $\mathbf{K}$, and its marginal probability measure on $\mathbf{K}_k$ is ν_k for all $k = 1, \ldots, p$. Therefore,

$$\min \mathbf{P} \;\ge\; \min \mathbf{Q}_{s_0} \;=\; L_y(f) \;=\; \int f\, d\mu \;=\; \sum_{\omega=1}^{s} p_\omega f(x^\omega),$$

which implies that $f(x^\omega) = \min \mathbf{P}$ for all $\omega = 1, \ldots, s$, because $x^\omega \in \mathbf{K}$ for all $\omega = 1, \ldots, s$. Therefore, we have proved that $\min \mathbf{P} = \min \mathbf{Q}_{s_0}$. In addition, each $x^\omega \in \mathbf{K}$ is an optimal solution of $\mathbf{P}$. □

5. Conclusion. We have provided a hierarchy of SDP-relaxations when the polynomial optimization problem $\mathbf{P}$ has some structured sparsity (which can be detected as in Waki et al. [9]). This hierarchy is of the same flavor (in fact a minor modification) as that in Waki et al. [9], for which excellent numerical results have been reported. Our contribution was to prove convergence of the optimal values to

the global minimum of **P** when the sparsity pattern satisfies the condition (1.3), called the *running intersection property* in graph theory, and frequently encountered in practice. Therefore, this result together with [9] opens the door for the applicability of the general approach of SDP-relaxations to medium (and even large) scale polynomial optimization problems, at least when a certain sparsity pattern is present.

6. Appendix. We state some auxiliary results needed in the proof of Theorem 3.1 in section 4.1.

For a topological space Y let $\mathcal{B}(Y)$ denote the usual Borel σ-algebra associated with Y, and let $P(Y)$ denote the space of probability measures on Y. A Borel space is a Borel subset of a complete separable metric space. Let Y, Z be two Borel spaces. A stochastic kernel $q(dy\,|\,z)$ on Y given Z is defined by

- $q(dy\,|\,z) \in P(Y)$ for all $z \in Z$,
- the function $z \mapsto q(B\,|\,z)$ is $\mathcal{B}(Z)$-measurable for all $B \in \mathcal{B}(Y)$.

6.1. Disintegration of a Borel probability measure. The following result states that one may decompose or *disintegrate* a probability measure on a product of Borel spaces into a marginal and a stochastic kernel (also called *conditional probability* when dealing with distributions of random variables).

PROPOSITION 6.1. *Let Y, Z be two Borel spaces, and let μ be a probability measure on $Y \times Z$. Then there exists a probability measure $\nu \in P(Z)$ and a stochastic kernel $q(dy\,|\,z)$ on Y given Z, such that*

$$\mu(A \times B) \;=\; \int_B q(A\,|\,z)\,\nu(dz) \qquad \forall\, A \in \mathcal{B}(Y),\; B \in \mathcal{B}(Z). \tag{6.1}$$

(Proposition 6.1 can be extended to the Cartesian product of an arbitrary number of Borel spaces.) The probability measure ν is called the *marginal* of μ on Z. One also has the converse.

PROPOSITION 6.2. *Let Y, Z be two Borel spaces, and let ν be a probability measure on Z, and $q(dy\,|\,z)$ a stochastic kernel on Y given Z. Then there exists a unique probability measure μ on $Y \times Z$ such that*

$$\mu(A \times B) \;=\; \int_B q(A\,|\,z)\,\nu(dz) \qquad \forall\, A \in \mathcal{B}(Y),\; B \in \mathcal{B}(Z). \tag{6.2}$$

(See, e.g., Ash [1, sect. 6] and Bertsekas and Schreve [2, pp. 139–141].)

Let μ (resp., ν) be a finite Borel probability measure on $\mathbb{R}^n \times \mathbb{R}^m$ (resp., $\mathbb{R}^m \times \mathbb{R}^p$) with all moments $y = (y_{\alpha\beta})_{\alpha\in\mathbb{N}^n,\beta\in\mathbb{N}^m}$ (resp., $z = (z_{\beta\gamma})_{\beta\in\mathbb{N}^m,\gamma\in\mathbb{N}^p}$) finite. Let μ_1 and ν_1 be the respective marginals of μ and ν on $\mathbb{R}^m$, hence with moments

$$\int X^\beta\, d\mu_1(X) \;=\; \int Y^0 X^\beta\, d\mu(Y,X) \;=\; y_{0\beta} \qquad \forall \beta \in \mathbb{N}^m,$$

$$\int X^\beta\, d\nu_1(X) \;=\; \int X^\beta Z^0\, d\mu(X,Z) \;=\; z_{\beta 0} \qquad \forall \beta \in \mathbb{N}^m.$$

If both μ and ν have compact support and $y_{0\beta} = z_{\beta 0}$ for all $\beta \in \mathbb{N}^m$, then $\mu_1 = \nu_1$. This is because measures with compact support are *moment determinate*, i.e., if two measures on a compact subset of $\mathbb{R}^m$ have all same moments, they must coincide.

6.2. Probability measures with given marginals. Case $p = 2$. Let $I_0 := \{1, \ldots, n\}$, and let $I_0 = I_1 \cup I_2$ with $I_1 \cap I_2 \neq \emptyset$. Let $n_k = \operatorname{card} I_k$, for $k = 1, 2$, and $n_{12} = \operatorname{card} I_1 \cap I_2$. For $k = 1, 2$, let $\pi_k : \mathbb{R}^n \to \mathbb{R}^{n_k}$ be the natural projection with respect to I_k, that is,

$$x \mapsto \pi_k(x) = \{x_i \,:\, i \in I_k\}, \qquad x \in \mathbb{R}^n,$$

and let $\pi_{12} : \mathbb{R}^{n_1} \to \mathbb{R}^{n_{12}}$, $\pi_{21} : \mathbb{R}^{n_2} \to \mathbb{R}^{n_{12}}$ be the projections with respect to $I_1 \cap I_2$, that is,

$$\begin{aligned} x &\mapsto \pi_{12}(x) = \{x_i \,:\, i \in I_1 \cap I_2\}, \qquad x \in \mathbb{R}^{n_1}, \\ x &\mapsto \pi_{21}(x) = \{x_i \,:\, i \in I_1 \cap I_2\}, \qquad x \in \mathbb{R}^{n_2}, \end{aligned}$$

and one also extends π_{12} and π_{21} to $\mathbb{R}^n$ in the obvious way.

Next, for $k = 1, 2$, let $\mathbf{K}_k \in \mathcal{B}(\mathbb{R}^{n_k})$ be given, and let $\nu_k \in P(\mathbf{K}_k)$. Denote by ν_{12} and ν_{21} the respective marginals of ν_1 and ν_2 on $\mathbb{R}^{n_{12}}$ (i.e., with respect to the variables $\{X_i,\, i \in I_1 \cap I_2\}$). That is, letting $Z := \mathbb{R}^{n_{12}}$,

$$\begin{aligned} \nu_{12}(B) &= \nu_1(\pi_{12}^{-1}(B) \cap \mathbf{K}_1) \qquad \forall\, B \in \mathcal{B}(Z), \\ \nu_{21}(B) &= \nu_2(\pi_{21}^{-1}(B) \cap \mathbf{K}_2) \qquad \forall\, B \in \mathcal{B}(Z), \end{aligned}$$

and we have

$$\nu_{12}(\pi_{12}(\mathbf{K}_1)) = \nu_{21}(\pi_{21}(\mathbf{K}_2)) = 1. \tag{6.3}$$

Let $\mathbf{K} \subset \mathbb{R}^n$ be the set

$$\mathbf{K} := \{x \in \mathbb{R}^n : \quad \pi_k(x) \in \mathbf{K}_k, \qquad k = 1, 2\}, \tag{6.4}$$

and view the sets $\mathbf{K}_k$, $k = 1, 2$ as naturally embedded in $\mathbb{R}^n$, with $\mathbf{K}_k = \pi_k(\mathbf{K})$, for every $k = 1, 2$.

LEMMA 6.3. *For $k = 1, 2$, let $\mathbf{K}_k \in \mathcal{B}(\mathbb{R}^{n_k})$ be given, and let $\nu_k \in P(\mathbf{K}_k)$ be such that $\nu_{12} = \nu_{21} =: \nu$. Then there exists a probability measure μ on $\mathbf{K}$ with marginals ν_k on $\mathbf{K}_k = \pi_k(\mathbf{K})$, $k = 1, 2$, and marginal ν on $\pi_{12}(\mathbf{K})$.*

Proof. For $k = 1, 2$, let π'_k be the natural projection with respect to $I_k \setminus I_1 \cap I_2$, i.e.,

$$x \mapsto \pi'_k(x) = \{x_i \,:\, i \in I_k \setminus I_1 \cap I_2\}, \quad x \in \mathbb{R}^{n_k},\ k = 1, 2,$$

and define $Y_k \in \mathcal{B}(\mathbb{R}^{n_k - n_{12}})$ to be the Borel set $\{\pi'_k(x) \,:\, x \in \mathbf{K}_k\}$, $k = 1, 2$.

Then, for $k = 1, 2$, one may view ν_k as a probability measure on the Cartesian product $Y_k \times Z$. By Proposition 6.1, and from $\nu_{12} = \nu_{21} =: \nu$, for $k = 1, 2$, one may disintegrate ν_k as

$$\nu_k(A \times B) = \int_B q_k(A \,|\, z)\, \nu(dz) \qquad \forall\, A \in \mathcal{B}(Y_k),\ B \in \mathcal{B}(Z),$$

for some stochastic kernels q_k, $k = 1, 2$. Next, let μ be the measure on $Y_1 \times Z \times Y_2$, defined by

$$\mu(A \times B \times C) = \int_B q_1(A \,|\, z)\, q_2(C \,|\, z)\, \nu(dz),$$

for every Borel rectangle

$$A \times B \times C \in \mathcal{B}(Y_1) \times \mathcal{B}(Z) \times \mathcal{B}(Y_2).$$

Taking $A = Y_1$ yields $q_1(A \,|\, z) = 1$, ν-a.e. and so

$$\mu(Y_1 \times B \times C) = \int_B q_2(C \,|\, z)\, \nu_{12}(dz) = \nu_2(B \times C).$$

Therefore, ν_2 is the marginal of μ on $Z \times Y_2$ (and so on $\mathbf{K}_2$). With similar argument, ν_1 is the marginal of μ on $Y_1 \times Z$ (and so on $\mathbf{K}_1$). Finally, taking $A = Y_1$, $C = Y_2$ and using $q_k(Y_k \,|\, z) = 1$, ν-a.e., yields

$$\mu(Y_1 \times B \times Y_2) = \int_B \nu(dz) = \nu(B),$$

which shows that ν is the marginal of μ on Z, i.e., with respect to the variables X_i, $i \in I_1 \cap I_2$. It remains to prove that $\mu(\mathbf{K}) = 1$. But notice that from the definitions of $\mathbf{K}_1, \mathbf{K}_2$, and ν,

$$q_1(\{\, y \,:\, (y,z) \in \mathbf{K}_1\} \,|\, z) = q_2(\{\, y' \,:\, (z,y') \in \mathbf{K}_2\} \,|\, z) = 1, \qquad \nu\text{-a.e.}$$

So, writing (6.4) as

$$\mathbf{K} = \{(y,z,y') \in \mathbb{R}^n \,:\quad (y,z) \in \mathbf{K}_1\,;\, (z,y') \in \mathbf{K}_2\,\}$$

yields

$$\mu(\mathbf{K}) = \int_Z q_1(\{y \,:\, (y,z) \in \mathbf{K}_1\} \,|\, z)\, q_2(\{y' \,:\, (z,y') \in \mathbf{K}_2\} \,|\, z)\, \nu(dz) = 1.$$

Therefore, ν_k is the marginal of μ on $\mathbf{K}_k = \pi_k(\mathbf{K})$ for $k = 1, 2$, and ν is the marginal of μ on $\pi_{12}(\mathbf{K})$. □

6.3. Probability measures with given marginals. General case. Let I_k, J_k, $k = 1, \ldots, p$, be as in section 2, and let $\mathbf{K} \subset \mathbb{R}^n$ be as defined in (1.2), with $\mathbf{K}_k \subset \mathbb{R}^{n_k}$ as in (3.4), $k = 1, \ldots, p$. Let ν_k be a given probability measure on $\mathbf{K}_k$, $k = 1, \ldots, p$.

Given a set $I \subset I_k$ denote by $X(I)$ the vector of variables $\{X_i\}_{i \in I} \in \mathbb{R}^{|I|}$, and denote by ν_{kI} the marginal of ν_k on $\mathbb{R}^{|I|}$ (i.e., with respect to the variables X_i, $i \in I$), so that ν_k can be disintegrated into $q_k(. \,|\, z) d\nu_{kI}(dz)$ for a stochastic kernel q on $\mathbb{R}^{n_k - |I|}$ given $\mathbb{R}^{|I|}$ (see Proposition 6.1).

We say that the family of probability measures $\{\nu_k\}_{k=1}^p$ is *consistent* with respect to marginals, if whenever $l, k \in \{1, \ldots, p\}$ and $I_k \cap I_l \neq \emptyset$,

$$I \subseteq I_k \cap I_l \Rightarrow \nu_{kI} = \nu_{lI}.$$

Equivalently, when ν_k and ν_l have compact support,

$$\int X^\alpha \, d\nu_k = \int X^\alpha \, d\nu_l \qquad \forall\, \alpha \,:\, \sup(\alpha) \subseteq I_k \cap I_l.$$

For every $k = 1, \ldots, p$, let $W_k := \bigcup_{l=1}^k I_l$, $s_k := |W_k|$, and

$$\Omega_k := \left\{ x \in \mathbb{R}^{s_k} \,|\quad g_j(x) \geq 0, \quad j \in \bigcup_{l=1}^k J_l \right\}. \tag{6.5}$$

Notice that $\Omega_n \equiv \mathbf{K} \subset \mathbb{R}^n$.

LEMMA 6.4. *Let ν_k be a probability measure on $\mathbf{K}_k \subset \mathbb{R}^{n_k}$, $k = 1, \ldots, p$, and assume that the family $\{\nu_k\}_{k=1}^p$ is consistent with respect to marginals. If (1.3) holds, then there is the following:*

(a) *There exists a probability measure μ on $\mathbb{R}^n$ such that ν_k is the marginal of μ with respect to I_k for all $k = 1, \ldots, p$.*

(b) *μ is supported on $\mathbf{K} \subset \mathbb{R}^n$.*

Proof. The proof is by induction on p. With $p = 1$ it is trivial. Let $p = 2$. Observe that the condition (1.3) is automatically satisfied. If $I_1 \cap I_2 = \emptyset$, just let $\mu := \nu_1 \otimes \nu_2$, the product measure on $\mathbf{K}_1 \times \mathbf{K}_2$, i.e.,

$$\mu(A \times B) \,=:\, \nu_1(A)\,\nu_2(B) \qquad \forall\,(A, B) \in \mathcal{B}(\mathbf{K}_1) \times \mathcal{B}(\mathbf{K}_1).$$

If $I_1 \cap I_2 \neq \emptyset$, then the result follows from Lemma 6.3.

Next, suppose that the results holds for $1 \leq m < p$. That is, let Ω_m be as in (6.5), and let ν_k be given probability measures on $\mathbf{K}_k$, $k = 1, \ldots, m$, consistent with marginals, i.e., whenever $l, k \in \{1, \ldots, m\}$, and $I_l \cap I_k \neq \emptyset$,

$$I \subseteq I_k \cap I_l \Rightarrow \nu_{lI} \,=\, \nu_{kI}.$$

Then there exists a probability measure μ_m on Ω_m, such that ν_k is the marginal of μ_m on $\mathbf{K}_k$ (i.e., with respect to the variables X_i, $i \in I_k$), for every $k = 1, \ldots, m$. We next show that it holds true for $m + 1$.

Set $\Delta := I_{m+1} \cap W_m$. If $\Delta = \emptyset$, then just take $\mu_{m+1} := \mu_m \otimes \nu_{m+1}$, the product measure on $\Omega_m \times \mathbf{K}_{m+1}$, and the induction is trivially satisfied for $m+1$. (As $\Delta = \emptyset$, one has $\Omega_{m+1} = \Omega_m \times \mathbf{K}_{m+1}$.)

Consider the case $\Delta \neq \emptyset$, and let $\delta := |\Delta|$, $s_{m+1} := |W_{m+1}|$. Let $\pi_\Delta : \Omega_m \to \mathbb{R}^\delta$, and $\pi'_\Delta : \mathbf{K}_{m+1} \to \mathbb{R}^\delta$ be the natural projection with respect to the variables X_i, $i \in \Delta$. Similarly, let $\pi_{\Delta^c} : \Omega_m \to \mathbb{R}^{s_m - \delta}$, and $\pi'_{\Delta^c} : \mathbf{K}_{m+1} \to \mathbb{R}^{n_{m+1} - \delta}$ be the natural projections with respect to the variables X_i, $i \in W_m \setminus \Delta$, and X_i, $i \in I_{m+1} \setminus \Delta$, respectively. So consider μ_m and ν_{m+1} as probability measures on the Borel spaces

$$Y \times Z \,:=\, \pi_{\Delta^c}(\Omega_m) \times \pi_\Delta(\Omega_m), \quad \text{and} \quad Z' \times Y' \,:=\, \pi'_\Delta(\mathbf{K}_{m+1}) \times \pi'_{\Delta^c}(\mathbf{K}_{m+1}),$$

respectively. Next, consider the marginals $\mu_{m\Delta}$ and $\nu_{(m+1)\Delta}$ of μ_m and ν_{m+1} on Z and Z', respectively, and the corresponding disintegrations,

$$\mu_m \,=\, q_m(.\,|\,z)\,\mu_{m\Delta}(dz)\,; \quad \nu_{m+1} \,=\, q'_m(.\,|\,z)\,\nu_{(m+1)\Delta}(dz).$$

From (1.3), $\Delta \subseteq I_s$ for some $s \in \{1, \ldots, m\}$. Therefore, $\nu_{(m+1)\Delta} = \nu_{s\Delta}$ because $\{\nu_k\}_{k=1}^{m+1}$ are consistent with marginals, and $\mu_{m\Delta} = \nu_{s\Delta} =: \nu$ by the induction hypothesis. Hence, one may take $Z = Z'$, and notice that

$$q_m(Y\,|\,z) \,=\, q'_m(Y'\,|\,z) \,=\, 1, \quad \nu\text{-a.e.} \tag{6.6}$$

Then define the probability measure μ_{m+1} on $Y \times Z \times Y' \subset \mathbb{R}^{s_{m+1}}$ by

$$\mu_{m+1}(A \times B \times C) \,:=\, \int_B q_m(A\,|z)\,q'_m(C\,|z)\,\nu(dz), \tag{6.7}$$

for all Borel rectangles $A \times B \times C \in \mathcal{B}(Y) \times \mathcal{B}(Z) \times \mathcal{B}(Y')$.

We claim that μ_{m+1} has the required properties of the induction hypothesis. First consider the marginal $\mu_{(m+1)I_{m+1}}$ of μ_{m+1} on $Z \times Y'$. It is obtained from (6.7) with $A = Y$. But from (6.6),

$$\begin{aligned}\mu_{(m+1)I_{m+1}}(B \times C) \;=\; \mu_{m+1}(Y \times B \times C) &= \int_B q'_m(C\,|z)\,\nu(dz)\\ &= \int_B q'_m(C\,|z)\,\nu_{(m+1)\Delta}(dz)\\ &= \nu_{m+1}(B \times C)\end{aligned}$$

for all $B \times C$ in $\mathcal{B}(Z) \times \mathcal{B}(Y')$, which proves that $\mu_{(m+1)I_{m+1}} = \nu_{m+1}$, the desired result. Next, consider the marginal $\mu_{(m+1)W_m}$ of μ_{m+1} with respect to the variables X_i, $i \in W_m$. It is obtained from (6.7) with $C = Y'$. So, using (6.6) again,

$$\begin{aligned}\mu_{(m+1)W_m}(A \times B) \;=\; \mu_{m+1}(A \times B \times Y') &= \int_B q_m(A\,|z)\,\nu(dz)\\ &= \int_B q_m(A\,|z)\,\mu_{m\Delta}(dz)\\ &= \mu_m(A \times B)\end{aligned}$$

for all $A \times B$ in $\mathcal{B}(Y) \times \mathcal{B}(Z)$, which proves that $\mu_{(m+1)W_m} = \mu_m$. But then, $\mu_{(m+1)I_k} = \mu_{mI_k}$ for all $k \le m$, and so, by the induction hypothesis, $\mu_{(m+1)I_k} = \mu_{mI_k} = \nu_k$ for all $k \le m$.

Hence, we have constructed a probability measure μ_{m+1} on $Y \times Z \times Y'$, such that for all $k = 1, \ldots, m+1$, ν_k is the marginal of $\mu_{(m+1)I_k}$ with respect to the variables X_i, $i \in I_k$. It remains to show that $\mu_{m+1}(\Omega_{m+1}) = 1$.

But from the definition of $\mathbf{K}_{m+1}$, Y', ν, and $\nu_{m+1}(\mathbf{K}_{m+1}) = 1$,

$$q'_m(B(z)\,|\,z) = 1 \quad \nu\text{-a.e. with } B(z) := \{y \,:\, g_j(z,y) \ge 0\;\forall\, j \in J_{m+1}\}.$$

Similarly, from the definitions of Ω_m, Y, ν, and $\mu_m(\Omega_m) = 1$,

$$q_m(A(z)\,|\,z) = 1 \quad \nu\text{-a.e. with } A(z) := \{y \,:\, g_j(y,z) \ge 0\;\forall\, j \in \cup_{k=1}^m J_k\}.$$

Therefore, (6.7), together with the definition (6.5) of Ω_{m+1}, yields

$$\mu_{m+1}(\Omega_{m+1}) \;=\; \int_Z q_m(A(z)\,|\,z)\,q'_m(B(z)\,|\,z) \times \nu(dz) \;=\; 1.$$

Therefore, the induction hypothesis is also true for $m+1$.

(b) From $\mu(\Omega_n) = 1$, and $\Omega_n = \mathbf{K}$, we obtain $\mu(\mathbf{K}) = 1$, the desired result. □

Acknowledgments. The author is indebted to Professor M. Kojima for very interesting and helpful discussions on the topic of sparse SDP-relaxations. He also wishes to thank T. Netzer and M. Schweighofer from Konstanz University (Germany), who indicated a way to simplify the original SDP-relations of the author in an earlier version, so as to yield the SDP-relaxations of this paper. Finally, the author wishes to thank anonymous referees for helpful remarks and suggestions to improve the initial version of the paper.

REFERENCES

[1] R. B. Ash, *Real Analysis and Probability*, Academic Press, New York, 1972.

[2] D. P. Bertsekas and S. .E. Schreve, *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, New York, 1978.

[3] R. E. Curto and L. A. Fialkow, *The truncated complex K-moment problem*, Trans. Amer. Math. Soc., 352 (2000), pp. 2825–2855.

[4] M. Fukuda, M. Kojima, K. Murota, and K. Nakata, *Exploiting sparsity in semidefinite programming via matrix completion* I: *General framework*, SIAM J. Optim., 11 (2001), pp. 647–674.

[5] D. Henrion and J. B. Lasserre, *GloptiPoly: Global optimization over polynomials with Matlab and SeDuMi*, ACM Trans. Math. Software, 29 (2003), pp. 165–194.

[6] D. Henrion and J. B. Lasserre, *Detecting global optimality and extracting solutions in GloptiPoly*, in Positive Polynomials in Control, Lecture Notes in Control and Inform. Sci. 312, D. Henrion and A. Garulli, eds., Springer-Verlag, Berlin, 2005, pp. 293–310.

[7] S. Kim, M. Kojima, and H. Waki, *Generalized Lagrangian duals and sums of squares relaxations of sparse polynomial optimization problems*, SIAM J. Optim., 15 (2005), pp. 697–719.

[8] M. Kojima, S. Kim, and H. Waki, *Sparsity in sums of squares of polynomials*, Math. Program., 103 (2005), pp. 45–62.

[9] H. Waki, S. Kim, M. Kojima, and M. Maramatsu, *Sums of squares and semidefinite programming relaxations for polynomial optimization problems with structured sparsity*, SIAM J. Optim., 17 (2006), pp. 218–242.

[10] M. Kojima and M. Maramatsu, *A Note on Sparse SOS and SDP-relaxations for Polynomial Optimization Problems over Symmetric Cones*, Technical report, Department of Mathematical and Computing Sciences, Tokyo Institute of Technology, Tokyo, Japan, 2006.

[11] J. B. Lasserre, *Global optimization with polynomials and the problem of moments*, SIAM J. Optim., 11 (2001), pp. 796–817.

[12] J. B. Lasserre, *An explicit equivalent positive semidefinite program for nonlinear* 0-1 *programs*, SIAM J. Optim., 12 (2002), pp. 756–769.

[13] J. B. Lasserre, *A sum of squares approximation of nonnegative polynomials*, SIAM J. Optim., 16 (2006), pp. 751–765.

[14] M. Laurent, *Revisiting two theorems of Curto and Fialkow on moment matrices*, Proc. Amer. Math. Soc., 133 (2005), pp. 2965–2976.

[15] K. Nakata, K. Fujisawa, M. Fukuda, M. Kojima, and K. Murota, *Exploiting sparsity in semidefinite programming via matrix completion* II: *Implementation and numerical results*, Math. Program., 95 (2003), pp. 303–327.

[16] M. Putinar, *Positive polynomials on compact semialgebraic sets*, Indiana Univ. Math. J., 42 (1993), pp. 969–984.

[17] M. Schweighofer, *Optimization of polynomials on compact semialgebraic sets*, SIAM J. Optim., 15 (2005), pp. 805–825.

SIAM J. OPTIM.
Vol. 17, No. 3, pp. 844–860
© 2006 Society for Industrial and Applied Mathematics

STRONG DUALITY IN NONCONVEX QUADRATIC OPTIMIZATION WITH TWO QUADRATIC CONSTRAINTS*

AMIR BECK† AND YONINA C. ELDAR‡

Abstract. We consider the problem of minimizing an indefinite quadratic function subject to two quadratic inequality constraints. When the problem is defined over the complex plane we show that strong duality holds and obtain necessary and sufficient optimality conditions. We then develop a connection between the image of the real and complex spaces under a quadratic mapping, which together with the results in the complex case lead to a condition that ensures strong duality in the *real* setting. Preliminary numerical simulations suggest that for random instances of the extended trust region subproblem, the sufficient condition is satisfied with a high probability. Furthermore, we show that the sufficient condition is always satisfied in two classes of nonconvex quadratic problems. Finally, we discuss an application of our results to robust least squares problems.

Key words. quadratic programming, nonconvex optimization, strong duality, quadratic mappings

AMS subject classifications. 90C20, 90C26, 90C46

DOI. 10.1137/050644471

1. Introduction.

In this paper we consider quadratic minimization problems with two quadratic constraints both in the real and the complex domain:

$$(QP_{\mathbb{C}}) \quad \min_{\boldsymbol{z}\in\mathbb{C}^n} \{f_0(\boldsymbol{z}) : f_1(\boldsymbol{z}) \geq 0, f_2(\boldsymbol{z}) \geq 0\}, \tag{1}$$

$$(QP_{\mathbb{R}}) \quad \min_{\boldsymbol{x}\in\mathbb{R}^n} \{f_0(\boldsymbol{x}) : f_1(\boldsymbol{x}) \geq 0, f_2(\boldsymbol{x}) \geq 0\}. \tag{2}$$

In the real case each function $f_j : \mathbb{R}^n \to \mathbb{R}$ is defined by $f_j(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{A}_j\boldsymbol{x} + 2\boldsymbol{b}_j^T\boldsymbol{x} + c_j$ with $\boldsymbol{A}_j = \boldsymbol{A}_j^T \in \mathbb{R}^{n\times n}, \boldsymbol{b}_j \in \mathbb{R}^n$, and $c_j \in \mathbb{R}$. In the complex setting, $f_j : \mathbb{C}^n \to \mathbb{R}$ is given by $f_j(\boldsymbol{z}) = \boldsymbol{z}^*\boldsymbol{A}_j\boldsymbol{z} + 2\Re(\boldsymbol{b}_j^*\boldsymbol{z}) + c_j$, where $\boldsymbol{A}_j = \boldsymbol{A}_j^*$ are Hermitian matrices, $\boldsymbol{b}_j \in \mathbb{C}^n$, and $c_j \in \mathbb{R}$. The problem $(QP_{\mathbb{R}})$ appears as a subproblem in some trust region algorithms for constrained optimization [6, 10, 26] where the original problem is to minimize a general nonlinear function subject to equality constraints. The subproblem, often referred to as the *two trust region problem* [1] or the *extended trust region problem* [35], has the form

$$(TTRS) \quad \min_{\boldsymbol{x}\in\mathbb{R}^n} \left\{\boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x} + 2\boldsymbol{g}^T\boldsymbol{x} : \|\boldsymbol{x}\| \leq \Delta, \|\boldsymbol{A}^T\boldsymbol{x} + \boldsymbol{c}\| \leq \xi\right\}. \tag{3}$$

More details on trust region algorithms can be found in [8, 23, 36, 37, 10]. A simpler (nonconvex) quadratic problem than (TTRS) is the *trust region subproblem*, which appears in trust region algorithms for *unconstrained* optimization:

$$(TR) \quad \min_{\boldsymbol{x}\in\mathbb{R}^n} \{\boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x} + 2\boldsymbol{g}^T\boldsymbol{x} : \|\boldsymbol{x}\|^2 \leq \delta\}. \tag{4}$$

*Received by the editors November 7, 2005; accepted for publication (in revised form) April 25, 2006; published electronically October 16, 2006.
http://www.siam.org/journals/siopt/17-3/64447.html

†Department of Industrial Engineering and Management, Technion—Israel Institute of Technology, Haifa 32000, Israel (becka@ie.technion.ac.il).

‡Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel (yonina@ee.technion.ac.il).

Problem (TR) has been studied extensively in the literature; see, e.g., [5, 12, 20, 21, 22, 30, 31] and references therein; it enjoys many useful and attractive properties. In particular, it is known that (TR) admits no duality gap and that the semidefinite relaxation (SDR) of (TR) is tight. Moreover, the solution of (TR) can be extracted from the dual solution. A necessary and sufficient condition for $\bar{\boldsymbol{x}}$ to be optimal for (TR) is that there exists $\bar{\alpha} \geq 0$ such that [15, 30]

$$(\boldsymbol{B} + \bar{\alpha}\boldsymbol{I})\bar{\boldsymbol{x}} + \boldsymbol{g} = \boldsymbol{0}, \tag{5}$$

$$\|\bar{\boldsymbol{x}}\|^2 \leq \delta, \tag{6}$$

$$\bar{\alpha}(\|\bar{\boldsymbol{x}}\|^2 - \delta) = 0, \tag{7}$$

$$\boldsymbol{B} + \bar{\alpha}\boldsymbol{I} \succeq \boldsymbol{0}. \tag{8}$$

Unfortunately, in general these results cannot be extended to the (TTRS) problem, or to ($QP_{\mathbb{R}}$). Indeed, it is known that the SDR of ($QP_{\mathbb{R}}$) is not necessarily tight [35, 36]. An exception is when the functions f_0, f_1, f_2 are homogeneous quadratic functions and there exists a positive definite linear combination of the matrices $\boldsymbol{A}_j$ [35]. Another interesting result obtained in [35], based on the dual cone representation approach [33], is that if f_1 is concave and f_2 is linear, then, although the SDR is *not necessarily tight*, ($QP_{\mathbb{R}}$) can be solved efficiently.

If the original nonlinear constrained problem has complex variables, then instead of ($QP_{\mathbb{R}}$) one should consider the complex variant ($QP_{\mathbb{C}}$). Optimization problems with complex variables appear naturally in many engineering applications. For example, if the estimation problem is posed in the Fourier domain, then typically the parameters to be estimated will be complex [24, 28]. In the context of digital communications, many signal constellations are modelled as complex valued. Another area where complex variables naturally arise is narrowband array processing [9].

Of course, every complex quadratic problem of dimension n can be written as a *real* quadratic problem of dimension $2n$ by decomposing the complex vector $\boldsymbol{z}$ as $\boldsymbol{z} = \boldsymbol{x} + i\boldsymbol{y}$, where $\boldsymbol{x} = \Re(\boldsymbol{z})$ and $\boldsymbol{y} = \Im(\boldsymbol{z})$ are real. Then $f_j(\boldsymbol{z})$ can be written as $f_j(\boldsymbol{z}) = \boldsymbol{w}^T\mathbf{Q}_j\boldsymbol{w} + 2\boldsymbol{d}_j^T\boldsymbol{w} + c_j$, with

$$\boldsymbol{w} = \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix} \in \mathbb{R}^{2n}, \mathbf{Q}_j = \begin{pmatrix} \Re(\boldsymbol{A}_j) & -\Im(\boldsymbol{A}_j) \\ \Im(\boldsymbol{A}_j) & \Re(\boldsymbol{A}_j) \end{pmatrix}, \boldsymbol{d} = \begin{pmatrix} \Re(\boldsymbol{b}_j) \\ \Im(\boldsymbol{b}_j) \end{pmatrix}.$$

However, the opposite claim is false: not every real quadratic problem of dimension $2n$ can be formulated as an n-dimensional complex quadratic problem. Evidently, the family of complex quadratic problems is a special case of real quadratic problems. *Why then consider the complex setting separately?* The answer to this question is that, as we shall see, there are stronger results for complex problems than for their real counterparts (cf. section 2).

In this paper we discuss both the complex and real settings. Our interest in the complex case is two-fold: First, as noted above, in certain applications we naturally deal with complex variables. Second, our derivations in the complex setting will serve as a basis for the results in the real case. In section 2, we use an extended version of the S-lemma [13] to show that under some mild conditions strong duality holds for the complex valued problem ($QP_{\mathbb{C}}$) and that the SDR is tight. We then develop optimality conditions similar to those known for the TR problem (4), and present a method for calculating the optimal solution of ($QP_{\mathbb{C}}$) from the dual solution. Thus, all the results known for (TR) can essentially be extended to ($QP_{\mathbb{C}}$). Section 3 treats the real setting. After a discussion of the complex relaxation of ($QP_{\mathbb{R}}$), which is an

alternative lifting procedure to the popular SDP relaxation, we present a sufficient condition that ensures zero duality gap (and tightness of the SDR) for $(QP_{\mathbb{R}})$. Our result is based on the connection between the image of the real and complex spaces under a quadratic mapping. The advantage of our condition is that it is expressed via the dual optimal solution and therefore can be validated in polynomial-time. Furthermore, this condition can be used to establish strong duality in some general classes of problems. As we show, an example where a problem of this form arises naturally is in robust least squares design where the uncertainty set is described by two norm constraints. In addition, preliminary numerical experiments suggest that for random instances of the TTRS problem (3), our condition is often satisfied.

Throughout the paper, the following notation is used: For simplicity, instead of inf/sup we use min/max; however, this does not mean that we assume that the optimum is attained and/or finite. Vectors are denoted by boldface lowercase letters; e.g., $\boldsymbol{y}$, and matrices by boldface uppercase letters; e.g., $\boldsymbol{A}$. For two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, $\boldsymbol{A} \succ \boldsymbol{B}$ ($\boldsymbol{A} \succeq \boldsymbol{B}$) means that $\boldsymbol{A} - \boldsymbol{B}$ is positive definite (semidefinite). $\mathcal{S}_+^n = \{\boldsymbol{A} \in \mathbb{R}^{n\times n} : \boldsymbol{A} \succeq \boldsymbol{0}\}$ is the set all real valued $n \times n$ symmetric positive semidefinite matrices and $\mathcal{H}_n^+ = \{\boldsymbol{A} \in \mathbb{C}^{n\times n} : \boldsymbol{A} \succeq \boldsymbol{0}\}$ is the set of all complex valued $n \times n$ Hermitian positive semidefinite matrices. $\boldsymbol{I}_n$ is the identity matrix of order n. The real and imaginary part of scalars, vectors, or matrices are denoted by $\Re(\cdot)$ and $\Im(\cdot)$. The value of the optimal objective function of an optimization problem

$$(\mathrm{P}) : \min / \max\{f(\boldsymbol{x}) : \boldsymbol{x} \in C\}$$

is denoted by val(P). We use some standard abbreviations such as SDP (semidefinite programming), SDR (semidefinite relaxation), and LMI (linear matrix inequalities).

2. The complex case. We begin by treating the complex valued problem $(QP_{\mathbb{C}})$. Using an extended version of the S-lemma we prove a strong duality result, and then develop necessary and sufficient optimality conditions, similar to those known for the TR problem (4) (conditions (5)–(8)). Finally, we discuss how to extract a solution for $(QP_{\mathbb{C}})$, given a dual optimal point.

2.1. Strong duality for $(QP_{\mathbb{C}})$. The fact that strong duality in (nonconvex) quadratic optimization problems is equivalent in some sense to the existence of a corresponding S-lemma has already been exhibited by several authors [13, 25]. For example, strong duality for quadratic problems with a single constraint can be shown to follow from the nonhomogeneous S-lemma [13], which states that if there exists $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ such that $\bar{\boldsymbol{x}}^T\boldsymbol{A}_2\bar{\boldsymbol{x}} + 2\boldsymbol{b}_2^T\bar{\boldsymbol{x}} + c_2 > 0$, then the following two conditions are equivalent:

1. $\boldsymbol{x}^T\boldsymbol{A}_1\boldsymbol{x} + 2\boldsymbol{b}_1^T\boldsymbol{x} + c_1 \geq 0$ for every $\boldsymbol{x} \in \mathbb{R}^n$ such that $\boldsymbol{x}^T\boldsymbol{A}_2\boldsymbol{x} + 2\boldsymbol{b}_2^T\boldsymbol{x} + c_2 \geq 0$.
2. There exists $\lambda \geq 0$ such that

$$\begin{pmatrix} \boldsymbol{A}_1 & \boldsymbol{b}_1 \\ \boldsymbol{b}_1^T & c_1 \end{pmatrix} \succeq \lambda \begin{pmatrix} \boldsymbol{A}_2 & \boldsymbol{b}_2 \\ \boldsymbol{b}_2^T & c_2 \end{pmatrix}.$$

Generalizations of the S-lemma in the real case are in general not true. For example, the natural extension to the case of two quadratic inequalities that imply a third quadratic inequality does not hold in general (see the example in [4]). However, the following theorem of Fradkov and Yakubovich [13, Theorem 2.2] extends the S-lemma to the complex case. This result will be the key ingredient in proving strong duality.

THEOREM 2.1 (extended S-lemma [13]). *Let*

$$f_j(\boldsymbol{z}) = \boldsymbol{z}^*\boldsymbol{A}_j\boldsymbol{z} + 2\Re(\boldsymbol{b}_j^*\boldsymbol{z}) + c_j, \quad \boldsymbol{z} \in \mathbb{C}^n, j = 0, 1, 2,$$

where $\boldsymbol{A}_j$ *are* $n \times n$ *Hermitian matrices,* $\boldsymbol{b}_j \in \mathbb{C}^n$*, and* $c_j \in \mathbb{R}$*. Suppose that there exists* $\tilde{\boldsymbol{z}} \in \mathbb{C}^n$ *such that* $f_1(\tilde{\boldsymbol{z}}) > 0, f_2(\tilde{\boldsymbol{z}}) > 0$*. Then the following two claims are equivalent:*

1. $f_0(\boldsymbol{z}) \geq 0$ *for every* $\boldsymbol{z} \in \mathbb{C}^n$ *such that* $f_1(\boldsymbol{z}) \geq 0$ *and* $f_2(\boldsymbol{z}) \geq 0$*.*
2. *There exists* $\alpha, \beta \geq 0$ *such that*

$$\begin{pmatrix} \boldsymbol{A}_0 & \boldsymbol{b}_0 \\ \boldsymbol{b}_0^* & c_0 \end{pmatrix} \succeq \alpha \begin{pmatrix} \boldsymbol{A}_1 & \boldsymbol{b}_1 \\ \boldsymbol{b}_1^* & c_1 \end{pmatrix} + \beta \begin{pmatrix} \boldsymbol{A}_2 & \boldsymbol{b}_2 \\ \boldsymbol{b}_2^* & c_2 \end{pmatrix}.$$

The Lagrangian dual of $(QP_{\mathbb{C}})$ can be shown to have the following form:[1]

$$(D_{\mathbb{C}}) \quad \max_{\alpha \geq 0, \beta \geq 0, \lambda} \left\{ \lambda \;\middle|\; \begin{pmatrix} \boldsymbol{A}_0 & \boldsymbol{b}_0 \\ \boldsymbol{b}_0^* & c_0 - \lambda \end{pmatrix} \succeq \alpha \begin{pmatrix} \boldsymbol{A}_1 & \boldsymbol{b}_1 \\ \boldsymbol{b}_1^* & c_1 \end{pmatrix} + \beta \begin{pmatrix} \boldsymbol{A}_2 & \boldsymbol{b}_2 \\ \boldsymbol{b}_2^* & c_2 \end{pmatrix} \right\}. \tag{9}$$

Problem $(D_{\mathbb{C}})$ is sometimes called Shor's relaxation [29]. Theorem 2.2 states that if problem $(QP_{\mathbb{C}})$ is finite and strictly feasible, then $\mathrm{val}(QP_{\mathbb{C}}) = \mathrm{val}(D_{\mathbb{C}})$.

THEOREM 2.2 (strong duality for complex valued quadratic problems). *Suppose that problem* $(QP_{\mathbb{C}})$ *is strictly feasible, i.e., there exists* $\tilde{\boldsymbol{z}} \in \mathbb{C}^n$ *such that* $f_1(\tilde{\boldsymbol{z}}) > 0, f_2(\tilde{\boldsymbol{z}}) > 0$*. If val*$(QP_{\mathbb{C}})$ *is finite, then the maximum of problem* $(D_{\mathbb{C}})$ *is attained and val*$(QP_{\mathbb{C}}) =$ *val*$(D_{\mathbb{C}})$*.*

Proof. Since $\mathrm{val}(QP_{\mathbb{C}})$ is finite then clearly

$$\mathrm{val}(QP_{\mathbb{C}}) = \max_{\lambda} \{\lambda : \mathrm{val}(QP_{\mathbb{C}}) \geq \lambda\}. \tag{10}$$

Now, the statement $\mathrm{val}(QP_{\mathbb{C}}) \geq \lambda$ holds true if and only if the implication

$$f_1(\boldsymbol{z}) \geq 0, f_2(\boldsymbol{z}) \geq 0 \Rightarrow f_0(\boldsymbol{z}) \geq \lambda$$

is valid. By Theorem 2.1 this is equivalent to

$$\exists \alpha, \beta \geq 0 \quad \begin{pmatrix} \boldsymbol{A}_0 & \boldsymbol{b}_0 \\ \boldsymbol{b}_0^* & c_0 - \lambda \end{pmatrix} \succeq \alpha \begin{pmatrix} \boldsymbol{A}_1 & \boldsymbol{b}_1 \\ \boldsymbol{b}_1^* & c_1 \end{pmatrix} + \beta \begin{pmatrix} \boldsymbol{A}_2 & \boldsymbol{b}_2 \\ \boldsymbol{b}_2^* & c_2 \end{pmatrix}. \tag{11}$$

Therefore, by replacing the constraint in (10) with the LMI (11), we obtain that $\mathrm{val}(QP_{\mathbb{C}}) = \mathrm{val}(D_{\mathbb{C}})$. The maximum of $(D_{\mathbb{C}})$ is attained at $(\bar{\lambda}, \bar{\alpha}, \bar{\beta})$, where $\bar{\lambda}$ is the (finite) value $\mathrm{val}(QP_{\mathbb{C}})$ and $\bar{\alpha}, \bar{\beta}$ are the corresponding nonnegative constants that satisfy the LMI (11) for $\lambda = \bar{\lambda}$. ∎

One referee pointed us to a recent related paper [18] from June 2005, which was posted to a web site after we submitted our paper. In [18], the strong duality result of Theorem 2.2 is derived by using an interesting new rank-one decomposition, while our proof is a direct consequence of the classical extended S-lemma of Fradkov and Yakubovich.

It is interesting to note that the dual problem to $(D_{\mathbb{C}})$ is the so-called SDR of $(QP_{\mathbb{C}})$:

$$(SDR_{\mathbb{C}}) \quad \min_{\boldsymbol{Z}} \{\mathrm{Tr}(\boldsymbol{Z}\boldsymbol{M}_0) : \mathrm{Tr}(\boldsymbol{Z}\boldsymbol{M}_1) \geq 0, \mathrm{Tr}(\boldsymbol{Z}\boldsymbol{M}_2) \geq 0, Z_{n+1,n+1} = 1, \boldsymbol{Z} \in \mathcal{H}_{n+1}^+\}, \tag{12}$$

where

$$\boldsymbol{M}_j = \begin{pmatrix} \boldsymbol{A}_j & \boldsymbol{b}_j \\ \boldsymbol{b}_j^* & c_j \end{pmatrix}.$$

[1] This formulation can be found in [34].

By the conic duality theorem (see, e.g., [4]), it follows that if both problems $(QP_{\mathbb{C}})$ and $(D_{\mathbb{C}})$ are strictly feasible, then they attain their solutions and $\mathrm{val}(QP_{\mathbb{C}}) = \mathrm{val}(D_{\mathbb{C}}) = \mathrm{val}(SDR_{\mathbb{C}})$. Finally, we note that strict feasibility of the dual problem $(D_{\mathbb{C}})$ is equivalent to saying that there exist $\tilde{\alpha} \geq 0, \tilde{\beta} \geq 0$ such that $\boldsymbol{A}_0 \succ \tilde{\alpha}\boldsymbol{A}_1 + \tilde{\beta}\boldsymbol{A}_2$. This condition is automatically satisfied when at least one of the constraints or the objective function is strictly convex (see also [35, Proposition 2.1]), an assumption that is true in many practical scenarios, for example in the TTRS problem (3).

2.2. Optimality conditions. Theorem 2.3 will be very useful in section 2.3, where a method for extracting the optimal solution of $(QP_{\mathbb{C}})$ from the optimal dual solution of $(D_{\mathbb{C}})$ will be described.

THEOREM 2.3. *Suppose that both problems $(QP_{\mathbb{C}})$ and $(D_{\mathbb{C}})$ are strictly feasible, and let $(\bar{\alpha}, \bar{\beta}, \bar{\lambda})$ be an optimal solution of $(D_{\mathbb{C}})$. Then $\bar{\boldsymbol{z}}$ is an optimal solution of $(QP_{\mathbb{C}})$ if and only if*

$$(\boldsymbol{A}_0 - \bar{\alpha}\boldsymbol{A}_1 - \bar{\beta}\boldsymbol{A}_2)\bar{\boldsymbol{z}} + \boldsymbol{b}_0 - \bar{\alpha}\boldsymbol{b}_1 - \bar{\beta}\boldsymbol{b}_2 = \boldsymbol{0}, \tag{13}$$

$$f_1(\bar{\boldsymbol{z}}), f_2(\bar{\boldsymbol{z}}) \geq 0, \tag{14}$$

$$\bar{\alpha} f_1(\bar{\boldsymbol{z}}) = \bar{\beta} f_2(\bar{\boldsymbol{z}}) = 0. \tag{15}$$

Proof. The proof follows from the strong duality result (Theorem 2.2) and from saddle point optimality conditions (see, e.g., [2, Theorem 6.2.5]). ∎

Note that a direct consequence of Theorem 2.3 is that the linear system (13) is consistent.

We now develop necessary and sufficient optimality conditions for $(QP_{\mathbb{C}})$ assuming strict feasibility, which are a natural generalization of the optimality conditions (5)–(8) for the trust region subproblem. Notice that for the complex version of the (TTRS), strict feasibility of $(D_{\mathbb{C}})$ is always satisfied since the norm constraint is strictly convex.

THEOREM 2.4. *Suppose that both problems $(QP_{\mathbb{C}})$ and $(D_{\mathbb{C}})$ are strictly feasible. Then $\bar{\boldsymbol{z}}$ is an optimal solution of $(QP_{\mathbb{C}})$ if and only if there exist $\alpha, \beta \geq 0$ such that*

(i) $(\boldsymbol{A}_0 - \alpha\boldsymbol{A}_1 - \beta\boldsymbol{A}_2)\bar{\boldsymbol{z}} + \boldsymbol{b}_0 - \alpha\boldsymbol{b}_1 - \beta\boldsymbol{b}_2 = 0$;

(ii) $f_1(\bar{\boldsymbol{z}}), f_2(\bar{\boldsymbol{z}}) \geq 0$;

(iii) $\alpha f_1(\bar{\boldsymbol{z}}) = \beta f_2(\bar{\boldsymbol{z}}) = 0$;

(iv) $\boldsymbol{A}_0 - \alpha\boldsymbol{A}_1 - \beta\boldsymbol{A}_2 \succeq 0$.

Proof. The necessary part is trivial since $\bar{\boldsymbol{z}}$, $\bar{\alpha}$, and $\bar{\beta}$ of Theorem 2.3 satisfy conditions (i)–(iv). Suppose now that conditions (i)–(iv) are satisfied. Then by (ii), $\bar{\boldsymbol{z}}$ is feasible and therefore $f_0(\bar{\boldsymbol{z}}) \geq \mathrm{val}(QP_{\mathbb{C}})$. To prove the reverse inequality ($f_0(\bar{\boldsymbol{z}}) \leq \mathrm{val}(QP_{\mathbb{C}})$), consider the unconstrained minimization problem:

$$\min_{\boldsymbol{z} \in \mathbb{C}^n} \{f_0(\boldsymbol{z}) - \bar{\alpha} f_1(\boldsymbol{z}) - \bar{\beta} f_2(\boldsymbol{z})\}. \tag{16}$$

We have

$$\begin{aligned} \mathrm{val}((16)) &\leq \min_{\boldsymbol{z} \in \mathbb{C}^n} \{f_0(\boldsymbol{z}) - \bar{\alpha} f_1(\boldsymbol{z}) - \bar{\beta} f_2(\boldsymbol{z}) : f_1(\boldsymbol{z}) \geq 0, f_2(\boldsymbol{z}) \geq 0\} \\ &\leq \min_{\boldsymbol{z} \in \mathbb{C}^n} \{f_0(\boldsymbol{z}) : f_1(\boldsymbol{z}) \geq 0, f_2(\boldsymbol{z}) \geq 0\} = \mathrm{val}(QP_{\mathbb{C}}). \end{aligned} \tag{17}$$

Conditions (i) and (iv) imply that $\bar{\boldsymbol{z}}$ is an optimal solution of (16) so that

$$f_0(\bar{\boldsymbol{z}}) - \bar{\alpha} f_1(\bar{\boldsymbol{z}}) - \bar{\beta} f_2(\bar{\boldsymbol{z}}) = \mathrm{val}((16)) \leq \mathrm{val}(QP_{\mathbb{C}}), \tag{18}$$

where the latter inequality follows from (17). By condition (iii) we have that $f_0(\bar{\boldsymbol{z}}) = f_0(\bar{\boldsymbol{z}}) - \bar{\alpha} f_1(\bar{\boldsymbol{z}}) - \bar{\beta} f_2(\bar{\boldsymbol{z}})$. Combining this with (18) we conclude that $f_0(\bar{\boldsymbol{z}}) \leq \mathrm{val}(QP_{\mathbb{C}})$. ∎

2.3. Finding an explicit solution of ($QP_{\mathbb{C}}$). Theorem 2.3 can be used to find an explicit solution to ($QP_{\mathbb{C}}$) from the solution of the dual ($D_{\mathbb{C}}$). Specifically, in section 2.3.1 we show that given the optimal dual solution, ($QP_{\mathbb{C}}$) reduces to a quadratic feasibility problem, whose solution is described in section 2.3.2.

2.3.1. Reduction to a quadratic feasibility problem. Suppose that both ($QP_{\mathbb{C}}$) and ($D_{\mathbb{C}}$) are strictly feasible. From Theorem 2.3, $\bar{\boldsymbol{z}}$ is an optimal solution if it satisfies (13), (14), and (15). If $\boldsymbol{A}_0 - \bar{\alpha}\boldsymbol{A}_1 - \bar{\beta}\boldsymbol{A}_2 \succ \boldsymbol{0}$, then the (unique) solution to the primal problem ($QP_{\mathbb{C}}$) is given by

$$\bar{\boldsymbol{z}} = -(\boldsymbol{A}_0 - \bar{\alpha}\boldsymbol{A}_1 - \bar{\beta}\boldsymbol{A}_2)^{-1}(\boldsymbol{b}_0 - \bar{\alpha}\boldsymbol{b}_1 - \bar{\beta}\boldsymbol{b}_2).$$

Next, suppose that $\boldsymbol{A}_0 - \bar{\alpha}\boldsymbol{A}_1 - \bar{\beta}\boldsymbol{A}_2$ is positive semidefinite but not positive definite. In this case (13) can be written as $\boldsymbol{z} = \boldsymbol{B}\boldsymbol{w} + \boldsymbol{a}$, where the columns of $\boldsymbol{B}$ form a basis for the null space of $\boldsymbol{A}_0 - \bar{\alpha}\boldsymbol{A}_1 - \bar{\beta}\boldsymbol{A}_2$ and $\boldsymbol{a} = -(\boldsymbol{A}_0 - \bar{\alpha}\boldsymbol{A}_1 - \bar{\beta}\boldsymbol{A}_2)^{\dagger}(\boldsymbol{b}_0 - \bar{\alpha}\boldsymbol{b}_1 - \bar{\beta}\boldsymbol{b}_2)$ is a solution of (13). It follows that $\bar{\boldsymbol{z}} = \boldsymbol{B}\bar{\boldsymbol{w}} + \boldsymbol{a}$ is an optimal solution to ($QP_{\mathbb{C}}$) if and only if conditions (14) and (15) of Theorem 2.3 are satisfied, i.e.,

$$g_1(\bar{\boldsymbol{w}}) \geq 0, g_2(\bar{\boldsymbol{w}}) \geq 0, \bar{\alpha}g_1(\bar{\boldsymbol{w}}) = 0, \bar{\beta}g_2(\bar{\boldsymbol{w}}) = 0, \quad (g_j(\boldsymbol{w}) \equiv f_j(\boldsymbol{B}\boldsymbol{w} + \boldsymbol{a})). \tag{19}$$

We are left with the problem of finding a vector which is a solution of a system of two quadratic equalities or inequalities as described in Table 1. This problem will be called the *quadratic feasibility problem.*

TABLE 1
Cases of the quadratic feasibility problem.

No.	Case	Feasibility problem
I	$\bar{\alpha} = 0, \bar{\beta} = 0$	$g_1(\boldsymbol{w}) \geq 0$ and $g_2(\boldsymbol{w}) \geq 0$
II	$\bar{\alpha} > 0, \bar{\beta} = 0$	$g_1(\boldsymbol{w}) = 0$ and $g_2(\boldsymbol{w}) \geq 0$
III	$\bar{\alpha} = 0, \bar{\beta} > 0$	$g_1(\boldsymbol{w}) \geq 0$ and $g_2(\boldsymbol{w}) = 0$
IV	$\bar{\alpha} > 0, \bar{\beta} > 0$	$g_1(\boldsymbol{w}) = 0$ and $g_2(\boldsymbol{w}) = 0$

Note that since $(\bar{\lambda}, \bar{\alpha}, \bar{\beta})$ is an optimal solution of the dual problem ($D_{\mathbb{C}}$), we must have $\boldsymbol{A}_0 - \bar{\alpha}\boldsymbol{A}_1 - \bar{\beta}\boldsymbol{A}_2 \succeq \boldsymbol{0}$. Thus, the first case is possible only when $\boldsymbol{A}_0 \succeq \boldsymbol{0}$.

We summarize the above discussion in the following theorem.

THEOREM 2.5. *Suppose that both problems ($QP_{\mathbb{C}}$) and ($D_{\mathbb{C}}$) are strictly feasible and let $(\bar{\alpha}, \bar{\beta}, \bar{\lambda})$ be an optimal solution of problem ($D_{\mathbb{C}}$). Then*

1. *if $\boldsymbol{A}_0 - \bar{\alpha}\boldsymbol{A}_1 - \bar{\beta}\boldsymbol{A}_2 \succ \boldsymbol{0}$, then the (unique) optimal solution of ($QP_{\mathbb{C}}$) is given by*

$$\bar{\boldsymbol{z}} = -(\boldsymbol{A}_0 - \bar{\alpha}\boldsymbol{A}_1 - \bar{\beta}\boldsymbol{A}_2)^{-1}(\boldsymbol{b}_0 - \bar{\alpha}\boldsymbol{b}_1 - \bar{\beta}\boldsymbol{b}_2),$$

2. *if $\boldsymbol{A}_0 - \bar{\alpha}\boldsymbol{A}_1 - \bar{\beta}\boldsymbol{A}_2 \succeq \boldsymbol{0}$ but not positive definite, then the solutions of ($QP_{\mathbb{C}}$) are $\boldsymbol{z} = \boldsymbol{B}\boldsymbol{w} + \boldsymbol{a}$, where the columns of $\boldsymbol{B} \in \mathbb{C}^{n\times d}$ form a basis for $\mathcal{N}(\boldsymbol{A}_0 - \bar{\alpha}\boldsymbol{A}_1 - \bar{\beta}\boldsymbol{A}_2)$, $\boldsymbol{a}$ is a solution of (13), and $\boldsymbol{w} \in \mathbb{C}^d$ ($d = \dim\big(\mathcal{N}(\boldsymbol{A}_0 - \bar{\alpha}\boldsymbol{A}_1 - \bar{\beta}\boldsymbol{A}_2)\big)$) is any solution of (19).*

2.3.2. Solving the quadratic feasibility problem. We now develop a method for solving all cases of the quadratic feasibility problem described in Table 1, under the condition that f_1 is strictly concave, i.e., $\boldsymbol{A}_1 \prec \boldsymbol{0}$ (so that the corresponding constraint is strictly convex).[2] The strict concavity of $g_1(\boldsymbol{w}) = f_1(\boldsymbol{B}\boldsymbol{w} + \boldsymbol{a})$ follows

[2]Note that this assumption readily implies that problem ($D_{\mathbb{C}}$) is strictly feasible.

immediately. By applying an appropriate linear transformation on g_1, we can assume without loss of generality that $g_1(\boldsymbol{w}) = \gamma - \|\boldsymbol{w}\|^2$ ($\gamma \geq 0$). Our approach will be to use solutions of at most two (TR) problems.

We split our analysis according to the different cases.

Case I+II. A solution to the feasibility problem in Case I (II) is any solution to the problem $\max\{g_2(\boldsymbol{w}) : \|\boldsymbol{w}\|^2 \leq \gamma\}$ ($\max\{g_2(\boldsymbol{w}) : \|\boldsymbol{w}\|^2 = \gamma\}$)

Case III. We first calculate $\boldsymbol{w}^0, \boldsymbol{w}^1$ given by

$$\boldsymbol{w}^0 \in \operatorname{argmin}\{g_2(\boldsymbol{w}) : \|\boldsymbol{w}\|^2 \leq \gamma\}, \quad \boldsymbol{w}^1 \in \operatorname{argmax}\{g_2(\boldsymbol{w}) : \|\boldsymbol{w}\|^2 \leq \gamma\}.$$

A solution to the feasibility problem is then given by $\bar{\boldsymbol{w}} = \boldsymbol{w}^0 + \eta(\boldsymbol{w}^1 - \boldsymbol{w}^0)$, where η is a solution to the scalar quadratic problem $g_2(\boldsymbol{w}^0 + \eta(\boldsymbol{w}^1 - \boldsymbol{w}^0)) = 0$ with $\eta \in [0, 1]$.

Case IV. Let $\boldsymbol{w}^0$ and $\boldsymbol{w}^1$ be defined by

$$\boldsymbol{w}^0 \in \operatorname{argmin}\{g_2(\boldsymbol{w}) : \|\boldsymbol{w}\|^2 = \gamma\}, \quad \boldsymbol{w}^1 \in \operatorname{argmax}\{g_2(\boldsymbol{w}) : \|\boldsymbol{w}\|^2 = \gamma\}.$$

The case in which $\boldsymbol{w}^0$ and $\boldsymbol{w}^1$ are linearly dependent can be analyzed in the same way as Case III. If $\boldsymbol{w}^0$ and $\boldsymbol{w}^1$ are linearly independent we can define

$$\boldsymbol{u}(\eta) = \boldsymbol{w}^0 + \eta(\boldsymbol{w}^1 - \boldsymbol{w}^0), \quad \boldsymbol{w}(\eta) = \sqrt{\gamma}\frac{\boldsymbol{u}(\eta)}{\|\boldsymbol{u}(\eta)\|}, \quad \eta \in [0, 1].$$

A solution to the feasibility problem is given by $w(\eta)$, where η is any root of the scalar equation $g_2(\boldsymbol{w}(\eta)) = 0,\ \eta \in [0, 1]$. The latter equation can be written (after some elementary algebraic manipulation) as the following *quartic* scalar equation:

$$\left(\gamma \boldsymbol{u}(\eta)^* \boldsymbol{A}_2 \boldsymbol{u}(\eta) + c_2 \|\boldsymbol{u}(\eta)\|^2\right)^2 = 4\gamma \|\boldsymbol{u}(\eta)\|^2 (\Re(\boldsymbol{b}_2^* \boldsymbol{u}(\eta)))^2. \tag{20}$$

Notice that (20) has at most four solutions, which have explicit algebraic expressions.

An alternative procedure for finding an explicit solution of $(QP_{\mathbb{C}})$ is described in [18]. The dominant computational effort in both methods is the solution of the SDP $(SDR_{\mathbb{C}})$ or its dual $(D_{\mathbb{C}})$, which can be solved by a primal dual interior point method that requires $O(n^{3.5})$ operations per accuracy digit (see, e.g., [4, section 6.6.1]).

3. The real case. We now treat the problem $(QP_{\mathbb{R}})$ in which the data and variables are assumed to be real valued. The dual problem to $(QP_{\mathbb{R}})$ is

$$(D_{\mathbb{R}}) \quad \max_{\alpha \geq 0, \beta \geq 0, \lambda} \left\{ \lambda \;\middle|\; \begin{pmatrix} \boldsymbol{A}_0 & \boldsymbol{b}_0 \\ \boldsymbol{b}_0^T & c_0 - \lambda \end{pmatrix} \succeq \alpha \begin{pmatrix} \boldsymbol{A}_1 & \boldsymbol{b}_1 \\ \boldsymbol{b}_1^T & c_1 \end{pmatrix} + \beta \begin{pmatrix} \boldsymbol{A}_2 & \boldsymbol{b}_2 \\ \boldsymbol{b}_2^T & c_2 \end{pmatrix} \right\}. \tag{21}$$

Note that this is *exactly the same* as problem $(D_{\mathbb{C}})$ (problem (9)), where here we use the fact that the data is real and therefore $\boldsymbol{b}_j^* = \boldsymbol{b}_j^T$. The SDR in this case is given by

$$(SDR_{\mathbb{R}}) \quad \min_{\boldsymbol{X}} \{\operatorname{Tr}(\boldsymbol{X}\boldsymbol{M}_0) : \operatorname{Tr}(\boldsymbol{X}\boldsymbol{M}_1) \geq 0, \operatorname{Tr}(\boldsymbol{X}\boldsymbol{M}_2) \geq 0, X_{n+1,n+1} = 1, \boldsymbol{X} \in \mathcal{S}_+^{n+1}\}. \tag{22}$$

In contrast to the complex case, strong duality is generally not true for $(QP_{\mathbb{R}})$. Nonetheless, in this section we use the results obtained for $(QP_{\mathbb{C}})$ in order to establish several results on $(QP_{\mathbb{R}})$. In section 3.1 we show that if the constraints of $(QP_{\mathbb{R}})$ are convex, then $(QP_{\mathbb{C}})$, considered as a relaxation of $(QP_{\mathbb{R}})$, can produce an approximate solution. In section 3.2 we relate the image of the real and complex space under a

quadratic mapping, which will enable us to bridge between the real and complex case. Using the latter result, a sufficient condition for zero duality gap is proved in section 3.3. The condition is expressed via the optimal dual variables and thus can be verified in polynomial time. Preliminary numerical results suggest that for the TTRS problem (3) this condition is often satisfied. Moreover, we identify two general classes of problems with zero duality gap, based on this condition. As we show in section 3.4, these results can be applied to the robust least-squares problem in order to obtain a polynomial time algorithm in the presence of uncertainty sets described by two norm constraints.

3.1. The complex relaxation. As already mentioned, val($QP_{\mathbb{R}}$) is not necessarily equal to val($D_{\mathbb{R}}$). However, the *complex counterpart* ($QP_{\mathbb{C}}$) does satisfy val($QP_{\mathbb{C}}$) = val($D_{\mathbb{R}}$) and we can always find a *complex valued* solution to ($QP_{\mathbb{C}}$) that attains the bound val($D_{\mathbb{R}}$). Therefore, we can consider ($QP_{\mathbb{C}}$) as a tractable relaxation (*the complex relaxation*) of the real valued problem ($QP_{\mathbb{R}}$). The following example, whose data is taken from Yuan [36, p. 59], illustrates this fact.

Example. Consider the following real valued quadratic optimization problem:

$$\min_{x_1,x_2\in\mathbb{R}} \{-2x_1^2 + 2x_2^2 + 4x_1 : x_1^2 + x_2^2 - 4 \le 0, x_1^2 + x_2^2 - 4x_1 + 3 \le 0\}, \tag{23}$$

which is a special case of ($QP_{\mathbb{R}}$) with

$$\boldsymbol{A}_0 = \begin{pmatrix} -2 & 0 \\ 0 & 2 \end{pmatrix}, \boldsymbol{A}_1 = \boldsymbol{A}_2 = -\boldsymbol{I}, \boldsymbol{b}_0(2;0), \boldsymbol{b}_1 = \boldsymbol{0}, \boldsymbol{b}_2 = (2;0), c_0 = 0, c_1 = 4, c_2 = -3.$$

The solution to the dual problem is given by $\bar{\alpha} = 1, \bar{\beta} = 1$, and $\bar{\lambda} = -1$. It is easy to see that the optimal solution to ($QP_{\mathbb{R}}$) is given by $x_1 = 2, x_2 = 0$ and its corresponding optimal solution is 0. The duality gap is thus 1. By the strong duality result of Theorem 2.2, we can find a complex valued solution to the complex counterpart

$$\min_{z_1,z_2\in\mathbb{C}} \{-2|z_1|^2 + 2|z_2|^2 + 4\Re(z_1) : |z_1|^2 + |z_2|^2 - 4 \le 0, |z_1|^2 + |z_2|^2 - 4\Re(z_1) + 3 \le 0\}. \tag{24}$$

with value equal to that of the dual problem (that is, equal to -1). Using the techniques described in section 2.3 we obtain that the solution of problem (24) is $z_1 = 7/4 + \sqrt{15/16}i, z_2 = 0$ with function value -1.

The following theorem states that if the constraints of ($QP_{\mathbb{C}}$) are convex (as in the two trust region problem), then we can extract an *approximate real solution* that is feasible from the optimal complex solution $\bar{\boldsymbol{z}}$ by taking $\bar{\boldsymbol{x}} = \Re(\bar{\boldsymbol{z}})$.

THEOREM 3.1. *Suppose that both* ($QP_{\mathbb{R}}$) *and* ($D_{\mathbb{R}}$) *are strictly feasible. Let* $\boldsymbol{A}_1, \boldsymbol{A}_2 \in \mathbb{R}^{n\times n}$ *be negative definite matrices,* $\boldsymbol{A}_0 = \boldsymbol{A}_0^T \in \mathbb{R}^{n\times n}$, $\boldsymbol{b}_j \in \mathbb{R}^n$, *and* $c_j \in \mathbb{R}$. *Let* $\bar{\boldsymbol{z}}$ *be an optimal complex valued solution of* ($QP_{\mathbb{C}}$) *and let* $\bar{\boldsymbol{x}} = \Re(\bar{\boldsymbol{z}})$. *Then* $\bar{\boldsymbol{x}}$ *is a feasible solution of* ($QP_{\mathbb{R}}$) *and*

$$f_0(\bar{\boldsymbol{x}}) - val(QP_{\mathbb{R}}) \le -\Im(\bar{\boldsymbol{z}})^T \boldsymbol{A}_0 \Im(\bar{\boldsymbol{z}}).$$

Proof. To show that $\bar{\boldsymbol{x}}$ is a feasible solution of ($QP_{\mathbb{R}}$) note that for $\boldsymbol{z} \in \mathbb{C}^n, j = 1, 2$ one has

$$\begin{aligned} 0 \le f_j(\boldsymbol{z}) &= \boldsymbol{z}^* \boldsymbol{A}_j \boldsymbol{z} + 2\Re(\boldsymbol{b}_j^* \boldsymbol{z}) + c_j \\ &= \Re(\boldsymbol{z})^T \boldsymbol{A}_j \Re(\boldsymbol{z}) + \Im(\boldsymbol{z})^T \boldsymbol{A}_j \Im(\boldsymbol{z}) + 2\boldsymbol{b}_j^T \Re(\boldsymbol{z}) + c_j \\ &\le \Re(\boldsymbol{z})^T \boldsymbol{A}_j \Re(\boldsymbol{z}) + 2\boldsymbol{b}_j^T \Re(\boldsymbol{z}) + c_j = f_j(\Re(\boldsymbol{z})), \end{aligned}$$

where the last inequality follows from $\boldsymbol{A}_j \prec \boldsymbol{0}$. Thus, since $\bar{\boldsymbol{z}}$ is feasible, so is $\Re(\bar{\boldsymbol{z}})$. Finally,

$$f_0(\Re(\bar{\boldsymbol{z}}))-\mathrm{val}(QP_{\mathbb{R}}) \leq f_0(\Re(\bar{\boldsymbol{z}}))-\mathrm{val}(QP_{\mathbb{C}}) = f_0(\Re(\bar{\boldsymbol{z}}))-f_0(\bar{\boldsymbol{z}}) = -\Im(\bar{\boldsymbol{z}})^T\boldsymbol{A}_0\Im(\bar{\boldsymbol{z}}). \qquad \square$$

In our example the approximate solution is $(7/4, 0)$ and its function value is equal to 0.875.

The extension from real to complex variables can be considered as *lifting*. A very popular lifting procedure is the SDR in which a nonconvex quadratic optimization problem defined over $\mathbb{R}^n$ is lifted to the corresponding SDR, which is defined over the space of $n \times n$ positive semidefinite matrices $\mathcal{S}_+^n$. This approach has been studied in various contexts such as approximation of combinatorial optimization problems (see [4] and references therein), polynomial inequalities [19], and more. The lifting procedure we suggest is relevant only in the context of quadratic optimization problems with *two* quadratic constraints. Our method is based on extending the real number field $\mathbb{R}$ into the complex number field $\mathbb{C}$. The *value* of the convex relaxation $\mathrm{val}(QP_{\mathbb{C}})$ is equal to the value of the SDR $\mathrm{val}(SDR_{\mathbb{R}})$. The main difference between the two strategies is in the "projection" stage onto $\mathbb{R}^n$. In our strategy, the projection is simple and natural: we take the real part of the vector. If the constraints are convex, then we have obtained a feasible point. In contrast, the choice of projection of the SDR solution, which is an $n \times n$ matrix, is not obvious. There are well established methods for specific instances (such as Max-Cut problems), but it is not clear how to extract a "good" approximate and feasible solution for general convex quadratic constraints. Another advantage to our method is that the procedure for finding a solution to $(QP_{\mathbb{C}})$ defined in section 2.3 can be manipulated so that it will output a real valued optimal solution in the case where strong duality indeed holds. In contrast, projection of the SDR solution may no longer be optimal, even in the case of strong duality.

3.2. The image of the complex and real space under a quadratic mapping. One of the key ingredients in proving the sufficient condition in section 3.3 is a result (Theorem 3.3) on the image of the spaces $\mathbb{C}^n$ and $\mathbb{R}^n$ under a quadratic mapping, composed from two nonhomogeneous quadratic functions. Results on the image of quadratic mappings play an important role in nonconvex quadratic optimization (see, e.g., [17, 25, 27] and references therein). We begin with the following theorem due to Polyak [25, Theorem 2.2], which is very relevant to our analysis.

THEOREM 3.2 (see [25]). *Let $\boldsymbol{A}_1, \boldsymbol{A}_2 \in \mathbb{R}^{n\times n}, (n \geq 2)$ be symmetric matrices for which the following condition is satisfied:*

$$\exists \alpha, \beta \in \mathbb{R} \text{ such that } \alpha\boldsymbol{A}_1 + \beta\boldsymbol{A}_2 \succ \boldsymbol{0}. \tag{25}$$

Let $\boldsymbol{b}_1, \boldsymbol{b}_2 \in \mathbb{R}^n$ and $c_1, c_2 \in \mathbb{R}$, and define $f_j(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{A}_j\boldsymbol{x} + 2\boldsymbol{b}_j^T\boldsymbol{x} + c_j$. Then the set

$$W = \{(f_1(\boldsymbol{x}), f_2(\boldsymbol{x})) : \boldsymbol{x} \in \mathbb{R}^n\}$$

is closed and convex.

The following theorem states that the images of $\mathbb{C}^n$ and $\mathbb{R}^n$ under the quadratic mapping defined in Theorem 3.2 are the same.

THEOREM 3.3. *Consider the setup of Theorem 3.2, and let $f_j(\boldsymbol{z}) = \boldsymbol{z}^*\boldsymbol{A}_j\boldsymbol{z} + 2\Re(\boldsymbol{b}_j^*\boldsymbol{z}) + c_j$. Then the sets*

$$F = \{(f_1(\boldsymbol{z}), f_2(\boldsymbol{z})) : \boldsymbol{z} \in \mathbb{C}^n\}, \quad W = \{(f_1(\boldsymbol{x}), f_2(\boldsymbol{x})) : \boldsymbol{x} \in \mathbb{R}^n\}$$

are equal. The proof of Theorem 3.3 relies on the following lemma.

LEMMA 3.4. *Let* $\boldsymbol{A}$ *be a real* $n \times n$ *symmetric matrix,* $\boldsymbol{b} \in \mathbb{R}^n, c \in \mathbb{R}$, *and* $\beta \geq 0$. *Then*

(26)
$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \{\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} + 2\boldsymbol{b}^T\boldsymbol{x} + c : \|\boldsymbol{x}\|^2 = \beta\} = \min_{\boldsymbol{z}\in\mathbb{C}^n} \{\boldsymbol{z}^*\boldsymbol{A}\boldsymbol{z} + 2\Re(\boldsymbol{b}^*\boldsymbol{z}) + c : \|\boldsymbol{z}\|^2 = \beta\}.$$

Proof. First note that (26) is obvious for $\beta = 0$. Suppose that $\beta > 0$. The value of the first problem in (26) is equal to

$$\max_{\mu}\{\mu : \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} + 2\boldsymbol{b}^T\boldsymbol{x} + c \geq \mu \text{ for every } \boldsymbol{x} \in \mathbb{R}^n \text{ such that } \|\boldsymbol{x}\|^2 = \beta\}. \tag{27}$$

Similarly, the value of the second problem is equal to

$$\max_{\mu}\{\mu : \boldsymbol{z}^*\boldsymbol{A}\boldsymbol{z} + 2\Re(\boldsymbol{b}^*\boldsymbol{z}) + c \geq \mu \text{ for every } \boldsymbol{z} \in \mathbb{C}^n \text{ such that } \|\boldsymbol{z}\|^2 = \beta\}. \tag{28}$$

By Theorem A.2 (note that condition (45) is satisfied for $f_1(\boldsymbol{x}) \equiv \|\boldsymbol{x}\|^2 - \beta$ with $\beta > 0$), the value of both problems is equal to the value of

$$\begin{array}{ll} \max_{\mu,\lambda} & \mu \\ \text{s.t.} & \begin{pmatrix} \boldsymbol{A} & \boldsymbol{b} \\ \boldsymbol{b}^T & c-\mu \end{pmatrix} \succeq \lambda \begin{pmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & -\beta \end{pmatrix}, \end{array}$$

and therefore these values are the same. ∎

We are now ready to prove Theorem 3.3.

Proof of Theorem 3.3. By Theorem 3.2 both W and F are convex. Obviously $W \subseteq F$. To prove the opposite, we first assume without loss of generality that $f_1(\boldsymbol{x}) = \|\boldsymbol{x}\|^2$. The latter assumption is possible since (25) is satisfied. Suppose that $(a, b) \in F$, i.e., $a = \|\boldsymbol{z}\|^2, b = f_2(\boldsymbol{z})$ for some $\boldsymbol{z} \in \mathbb{C}^n$, and let

$$b_{min} = \min\{f_2(\boldsymbol{z}) : \|\boldsymbol{z}\|^2 = a\} \text{ and } b_{max} = \max\{f_2(\boldsymbol{z}) : \|\boldsymbol{z}\|^2 = a\}.$$

By Lemma 3.4, there must be two *real* vectors $\boldsymbol{x}^0, \boldsymbol{x}^1 \in \mathbb{R}^n$ such that $\|\boldsymbol{x}^0\|^2 = \|\boldsymbol{x}^1\|^2 = a$ and $f_2(\boldsymbol{x}^0) = b_{min} \leq b \leq b_{max} = f_2(\boldsymbol{x}^1)$. Therefore, $(a, b_{min}), (a, b_{max}) \in W$. Since W is convex we conclude that (a, b), being a convex combination of (a, b_{min}) and (a, b_{max}), also belongs to W. ∎

3.3. A sufficient condition for zero duality gap of $(QP_{\mathbb{R}})$.

3.3.1. The condition. We now use the results on the complex valued problem $(QP_{\mathbb{C}})$ in order to find a sufficient condition for zero duality gap and tightness of the SDR of the real valued problem $(QP_{\mathbb{R}})$. Our derivation is based on the fact that if an optimal solution of $(QP_{\mathbb{C}})$ is *real valued*, then $(QP_{\mathbb{R}})$ admits no gap with its dual problem $(D_{\mathbb{R}})$.

THEOREM 3.5. *Suppose that both problems* $(QP_{\mathbb{R}})$ *and* $(D_{\mathbb{R}})$ *are strictly feasible and that*

$$\exists \hat{\alpha}, \hat{\beta} \in \mathbb{R} \text{ such that } \hat{\alpha}\boldsymbol{A}_1 + \hat{\beta}\boldsymbol{A}_2 \succ \boldsymbol{0}. \tag{29}$$

Let $(\bar{\lambda}, \bar{\alpha}, \bar{\beta})$ *be an optimal solution of the dual problem* $(D_{\mathbb{R}})$. *If*

$$d = \dim\left(\mathcal{N}(\boldsymbol{A}_0 - \bar{\alpha}\boldsymbol{A}_1 - \bar{\beta}\boldsymbol{A}_2)\right) \neq 1, \tag{30}$$

then val($QP_{\mathbb{R}}$) = val($D_{\mathbb{R}}$) = val($SDR_{\mathbb{R}}$) and there exists a real valued solution to ($QP_{\mathbb{C}}$).

Proof. Since both ($SDR_{\mathbb{R}}$) and ($D_{\mathbb{R}}$) are strictly feasible, $\mathrm{val}(D_{\mathbb{R}}) = \mathrm{val}(SDR_{\mathbb{R}})$. Now, suppose that $d = 0$. Then by (13), a solution to ($QP_{\mathbb{C}}$) is given by the real valued vector

$$\bar{\boldsymbol{x}} = -(\boldsymbol{A}_0 - \bar{\alpha}\boldsymbol{A}_1 - \bar{\beta}\boldsymbol{A}_2)^{-1}(\boldsymbol{b}_0 - \bar{\alpha}\boldsymbol{b}_1 - \bar{\beta}\boldsymbol{b}_2).$$

Since ($QP_{\mathbb{C}}$) has a real valued solution it follows that $\mathrm{val}(QP_{\mathbb{R}}) = \mathrm{val}(QP_{\mathbb{C}}) = \mathrm{val}(D_{\mathbb{R}})$, where the last equality follows from Theorem 2.2.

Next, suppose that $d \geq 2$. By Theorem 2.5, any optimal solution $\bar{\boldsymbol{z}}$ of ($QP_{\mathbb{C}}$) has the form $\bar{\boldsymbol{z}} = \boldsymbol{B}\bar{\boldsymbol{w}} + \boldsymbol{a}$, where $\bar{\boldsymbol{w}} \in \mathbb{C}^d$ is a solution of

$$g_1(\boldsymbol{w}) \geq 0, g_2(\boldsymbol{w}) \geq 0, \bar{\alpha}g_1(\boldsymbol{w}) = 0, \bar{\beta}g_2(\boldsymbol{w}) = 0, \quad (g_j(\boldsymbol{w}) \equiv f_j(\boldsymbol{B}\boldsymbol{w} + \boldsymbol{a})). \tag{31}$$

Both the matrix $\boldsymbol{B}$ and the vector $\boldsymbol{a}$ are chosen to be real valued; such a choice is possible since the columns of $\boldsymbol{B}$ form a basis for the null space of the *real valued* matrix $\boldsymbol{A}_1 - \bar{\alpha}\boldsymbol{A}_1 - \bar{\beta}\boldsymbol{A}_2$ and $\boldsymbol{a}$ is an arbitrary solution of a real valued linear system. Now, obviously $(g_1(\bar{\boldsymbol{w}}), g_2(\bar{\boldsymbol{w}})) \in S_1$, where $S_1 = \{(g_1(\boldsymbol{w}), g_2(\boldsymbol{w})) : \boldsymbol{w} \in \mathbb{C}^d\}$. Since $\boldsymbol{B}$ has full column rank, if (29) is satisfied, then

$$\hat{\alpha}\boldsymbol{B}^T\boldsymbol{A}_1\boldsymbol{B} + \hat{\beta}\boldsymbol{B}^T\boldsymbol{A}_2\boldsymbol{B} \succ \boldsymbol{0}. \tag{32}$$

The LMI (32) together with the fact that $d \geq 2$ imply that the conditions of Theorem 3.3 are satisfied and thus $S_1 = S_2$, where $S_2 = \{(g_1(\boldsymbol{x}), g_2(\boldsymbol{x})) : \boldsymbol{x} \in \mathbb{R}^d\}$. Therefore, there exists $\bar{\boldsymbol{x}} \in \mathbb{R}^d$ such that $g_j(\bar{\boldsymbol{w}}) = g_j(\bar{\boldsymbol{x}})$ and as a result, (31) has a real valued solution. To conclude, $\bar{\boldsymbol{z}} = \boldsymbol{B}\bar{\boldsymbol{x}} + \boldsymbol{a} \in \mathbb{R}^n$ is a real valued vector which is an optimal solution to ($QP_{\mathbb{C}}$). ∎

A more restrictive sufficient condition than (30) is

$$\dim(\mathcal{N}(\boldsymbol{A}_0 - \bar{\alpha}\boldsymbol{A}_1 - \bar{\beta}\boldsymbol{A}_2)) = 0.$$

This condition, as opposed to condition (30), can be directly derived from complementarity conditions of ($SDR_{\mathbb{R}}$) and its dual ($D_{\mathbb{R}}$).

We note that although a direct verification of the sufficient condition (30) requires the solution of the dual problem ($D_{\mathbb{R}}$), we will show that it is possible to use this condition in order to prove strong duality is *always* satisfied for certain classes of structured nonconvex quadratic problems (see section 3.3.3).

3.3.2. Numerical experiments. To demonstrate the fact that for the TTRS problem (3), the sufficient condition of Theorem 3.5 often holds for random problems, we considered different values of m and n (the number of constraints and the number of variables in the original nonlinear problem) and randomly generated 1000 instances of $\boldsymbol{B}, \boldsymbol{g}, \boldsymbol{A}$, and $\boldsymbol{c}$. We chose $\Delta = 0.1$ and $\xi = \|\boldsymbol{A}^T(-\alpha\boldsymbol{A}\boldsymbol{c}) + \boldsymbol{c}\|$, with

$$\alpha = \min\left\{\frac{\Delta}{\|\boldsymbol{A}\boldsymbol{c}\|}, \frac{\boldsymbol{c}^T(\boldsymbol{A}^T\boldsymbol{A})\boldsymbol{c}}{\boldsymbol{c}^T(\boldsymbol{A}^T\boldsymbol{A})^2\boldsymbol{c}}\right\},$$

as suggested in the trust region algorithm of [6]. The SDP problems were solved by SeDuMi [32]. The results are given in Table 2.

In the table, *distribution* is the distribution from which the coefficients of $\boldsymbol{B}, \boldsymbol{g}, \boldsymbol{A}$, and $\boldsymbol{c}$ are generated. There are two possibilities: uniform distribution ($U[0, 1]$) or

TABLE 2
Results for TTRS.

n	m	distribution	N_{suf}	mean	sd
10	1	Normal	997	5.50	2.34
10	1	Uniform	1000	1.61	0.62
10	10	Normal	1000	5.04	2.31
10	10	Uniform	1000	1.60	0.61
100	1	Normal	1000	13.15	2.65
100	1	Uniform	1000	3.75	0.64
100	100	Normal	1000	12.54	2.31
100	100	Uniform	1000	3.71	0.65

standard normal distribution ($N(0,1)$). N_{suf} is the number of problems satisfying the sufficient condition (30) out of 1000. *mean* and *sd* are the mean and standard deviation of the minimal eigenvalue of the matrix $\boldsymbol{A}_0 - \bar{\alpha}\boldsymbol{A}_1 - \bar{\beta}\boldsymbol{A}_2$. Numerically, the dimension of the null space in condition (30) was determined by the number of eigenvalues of the matrix $\boldsymbol{A}_0 - \bar{\alpha}\boldsymbol{A}_1 - \bar{\beta}\boldsymbol{A}_2$ whose absolute value was less than 10^{-8}. It is interesting to note that *almost all* the instances satisfied condition (30) except for 3 cases when $n = 10, m = 1$ with data generated from the normal distribution. Of course, these experiments reflect the situation in random problems and the results might be different (for better or for worse) if the data is generated differently.

3.3.3. Two classes of problems with zero duality gap. We will now present two classes of nonconvex quadratic problems for which the sufficient condition of Theorem 3.5 is always satisfied.

First class. Consider the problem of minimizing an indefinite quadratic function subject to a norm constraint and a linear inequality constraint:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \{\boldsymbol{x}^T\mathbf{Q}\boldsymbol{x} + 2\boldsymbol{b}^T\boldsymbol{x} : \|\boldsymbol{x}\|^2 \le \delta, \boldsymbol{a}^T\boldsymbol{x} \le \xi\}. \tag{33}$$

This problem was treated in [33, 35], where it was shown that the SDR is not always tight, although a polynomial-time algorithm for solving this problem was presented. We will find a condition on the data $(\mathbf{Q}, \boldsymbol{a}, \boldsymbol{b})$ that will be sufficient for zero duality gap.

THEOREM 3.6. *Suppose that problem* (33) *is strictly feasible and* $n \ge 2$. *If the dimension of* $\mathcal{N}(\mathbf{Q} - \lambda_{min}(\mathbf{Q})\boldsymbol{I}_n)$ *is at least* 2, *then strong duality holds for problem* (33).

Proof. Let $(\bar{\lambda}, \bar{\alpha}, \bar{\beta})$ be an optimal solution of the dual problem to (33). From the feasibility of the dual problem it follows that $\mathbf{Q}+\bar{\alpha}\boldsymbol{I}_n \succeq \mathbf{0}$. Now, either $\bar{\alpha} > -\lambda_{min}(\mathbf{Q})$ and in that case $\mathbf{Q} + \bar{\alpha}\boldsymbol{I}_n$ is nonsingular and thus the dimension of $\mathcal{N}(\mathbf{Q} + \bar{\alpha}\boldsymbol{I}_n)$ is 0 or $\bar{\alpha} = -\lambda_{min}(\mathbf{Q})$ and in this case $\mathcal{N}(\mathbf{Q} + \bar{\alpha}\boldsymbol{I}_n)$ is of dimension at least 2 by the assumptions. The result follows now from Theorem 3.5. ∎

Second class. Consider problem $(QP_{\mathbb{R}})$ with matrices $\boldsymbol{A}_i$ of the following form:

$$\boldsymbol{A}_i = \boldsymbol{I}_r \otimes \mathbf{Q}_i, \quad i = 0, 1, 2, \tag{34}$$

where $\mathbf{Q}_i = \mathbf{Q}_i^T \in \mathbb{R}^{m\times m}, r > 1$, and $n = rm$. Here $\otimes$ denotes the Kronecker product. In section 3.4 we will show that this class of problems naturally arises in unstructured robust least squares problems. The following theorem, which is a direct consequence of the sufficient condition (30), states that under some mild conditions (such as strict feasibility), strong duality holds.

THEOREM 3.7. *Suppose that both problems* $(QP_{\mathbb{R}})$ *and* $(D_{\mathbb{R}})$ *are strictly feasible and that* $\boldsymbol{A}_i$ *is given by* (34). *Moreover, suppose that there exist* $\hat{\alpha}$ *and* $\hat{\beta}$ *such that*

$$\hat{\alpha}\mathbf{Q}_1 + \hat{\beta}\mathbf{Q}_2 \succ \mathbf{0}. \tag{35}$$

Then strong duality holds for $(QP_{\mathbb{R}})$.

Proof. The validity of condition (35) readily implies that (29) holds true. Moreover, by the premise of the theorem, both problems $(QP_{\mathbb{R}})$ and $(D_{\mathbb{R}})$ are strictly feasible. We are thus left with the task of proving that condition (30) is satisfied. Indeed, let $(\bar{\lambda}, \bar{\alpha}, \bar{\beta})$ be an optimal solution of the dual problem $(D_{\mathbb{R}})$. Then the matrix $\boldsymbol{A}_0 - \bar{\alpha}\boldsymbol{A}_1 - \bar{\beta}\boldsymbol{A}_2$ is equal to $\boldsymbol{I}_r \otimes (\mathbf{Q}_0 - \bar{\alpha}\mathbf{Q}_1 - \bar{\beta}\mathbf{Q}_2)$. Using properties of eigenvalues of Kronecker products [14], we conclude that the multiplicities of the eigenvalues of the latter matrix must be multiplicities of r, i.e., $r, 2r, \ldots$. The dimension of $\mathcal{N}(\boldsymbol{A}_0 - \bar{\alpha}\boldsymbol{A}_1 - \bar{\beta}\boldsymbol{A}_2)$ is the multiplicity of the eigenvalue 0, which by the fact that $r > 1$, cannot be equal to 1. Hence, by Theorem 3.5, strong duality holds. ∎

3.4. Application to unstructured robust least squares. The *robust least squares* (RLS) problem was introduced and studied in [16, 7]. Consider a linear system $\boldsymbol{A}\boldsymbol{x} \approx \boldsymbol{b}$ where $\boldsymbol{A} \in \mathbb{R}^{r\times n}, \boldsymbol{b} \in \mathbb{R}^r$, and $\boldsymbol{x} \in \mathbb{R}^n$. Assume that the matrix and right-hand side vector $(\boldsymbol{A}, \boldsymbol{b})$ are not fixed but rather given by a family of matrices[3] $(\boldsymbol{A}, \boldsymbol{b}) + \boldsymbol{\Delta}^T$, where $(\boldsymbol{A}, \boldsymbol{b})$ is a known nominal value and $\boldsymbol{\Delta} \in \mathbb{R}^{(n+1)\times r}$ is an unknown perturbation matrix known to reside in a compact uncertainty set $\mathcal{U}$. The RLS approach to this problem is to seek a vector $\boldsymbol{x} \in \mathbb{R}^n$ that minimizes the worst case data error with respect to all possible values of $\boldsymbol{\Delta} \in \mathcal{U}$:

$$\min_{\boldsymbol{x}} \max_{\boldsymbol{\Delta}\in\mathcal{U}} \left\| \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} + \boldsymbol{\Delta}^T \begin{pmatrix} \boldsymbol{x} \\ -1 \end{pmatrix} \right\|^2. \tag{36}$$

In [16] the uncertainty set $\mathcal{U}$ in the unstructured case was chosen to contain all matrices $\boldsymbol{\Delta}$ satisfying a simple Frobenius norm constraint, i.e.,

$$\mathrm{Tr}(\boldsymbol{\Delta}^T\boldsymbol{\Delta}) \le \rho. \tag{37}$$

The RLS problem is considered difficult in the case when the uncertainty set $\mathcal{U}$ is given by an *intersection* of ellipsoids; see the related problem[4] of finding a robust counterpart of a conic quadratic problem [3]. Nonetheless, we will now show that a byproduct of our results is that as long as $r > 1$, the RLS problem with uncertainty set given by an intersection of two ellipsoids is tractable. Specifically, we consider an uncertainty set $\mathcal{U}$ given by two norm constraints:

$$\mathcal{U} = \{\boldsymbol{\Delta} \in \mathbb{R}^{(n+1)\times r} : \mathrm{Tr}(\boldsymbol{\Delta}^T\boldsymbol{B}_i\boldsymbol{\Delta}) \le \rho_i, i = 1, 2\}, \tag{38}$$

where $\boldsymbol{B}_i = \boldsymbol{B}_i^T \in \mathbb{R}^{(n+1)\times(n+1)}$ and $\rho_i > 0$. We also assume that

$$\exists \gamma_1 \ge 0, \gamma_2 \ge 0 \text{ such that } \gamma_1\boldsymbol{B}_1 + \gamma_2\boldsymbol{B}_2 \succ \mathbf{0}. \tag{39}$$

The above condition will ensure strict feasibility of the dual problem to the inner maximization problem of (36).

[3] The perturbation matrix appears in a transpose form for the sake of simplicity of notation.

[4] Note that finding a tractable formulation to the RLS problem is the key ingredient in deriving a robust counterpart of a conic quadratic constraint of the form $\|\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}\| \le \boldsymbol{c}^T\boldsymbol{x} + d$.

The form of the uncertainty set (38) is more general than the standard single-constraint form (37) and it can thus be used to describe more complicated scenarios of uncertainties. Using some simple algebraic manipulations the objective function in (36) can be written as

$$\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}+\boldsymbol{\Delta}^T\tilde{\boldsymbol{x}}\|^2=\mathrm{Tr}(\boldsymbol{\Delta}^T\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^T\boldsymbol{\Delta})+2\mathrm{Tr}((\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})\tilde{\boldsymbol{x}}^T\boldsymbol{\Delta})+\mathrm{Tr}((\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})^T),$$

where we denoted

$$\tilde{\boldsymbol{x}}=\begin{pmatrix}\boldsymbol{x}\\-1\end{pmatrix}. \tag{40}$$

Relying on the identities

$$\mathrm{Tr}(\boldsymbol{A}^T\boldsymbol{B}\boldsymbol{A})=\mathrm{vec}(\boldsymbol{A})^T(\boldsymbol{I}_r\otimes\boldsymbol{B})\,\mathrm{vec}(\boldsymbol{A}),\quad \mathrm{Tr}(\boldsymbol{A}^T\mathbf{C})=\mathrm{vec}(\boldsymbol{A})^T\,\mathrm{vec}(\mathbf{C}) \tag{41}$$

for every $\boldsymbol{A},\mathbf{C}\in\mathbb{R}^{p\times r}\boldsymbol{B}\in\mathbb{R}^{p\times p}$, where for a matrix $\boldsymbol{M}$, $\mathrm{vec}(\boldsymbol{M})$ denotes the vector obtained by stacking the columns of $\boldsymbol{M}$, the inner maximization problem in (36) takes the following form:

$$\max\{\mathrm{vec}(\boldsymbol{\Delta})^T\mathbf{Q}\,\mathrm{vec}(\boldsymbol{\Delta})+2\boldsymbol{f}^T\,\mathrm{vec}(\boldsymbol{\Delta})+c:\boldsymbol{\Delta}\in\mathcal{U}\}, \tag{42}$$

where $\mathbf{Q}=\boldsymbol{I}_r\otimes\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^T$, $\boldsymbol{f}=\mathrm{vec}(\tilde{\boldsymbol{x}}(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})^T)$, and $c=\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|^2$. By the first identity of (41) it follows that $\mathcal{U}$ can be written as

$$\mathcal{U}=\{\boldsymbol{\Delta}\in\mathbb{R}^{(n+1)\times r}:\mathrm{vec}(\boldsymbol{\Delta})^T(\boldsymbol{I}_r\otimes\boldsymbol{B}_i)\,\mathrm{vec}(\boldsymbol{\Delta})\le\rho_i,i=1,2\}.$$

Therefore, all the matrices in the inner maximization problem (42) are of the form $\boldsymbol{I}_r\otimes G$. Noting that all the other conditions of Theorem 3.7 are satisfied (strict feasibility of the primal and dual problems and (35)), we conclude that strong duality holds for (42) and its value is thus equal to the value of the dual problem given by

$$\min_{\alpha\ge0,\beta\ge0,\lambda}\left\{-\lambda\left|\begin{pmatrix}-\mathbf{Q}+\boldsymbol{I}_r\otimes(\alpha\boldsymbol{B}_1+\beta\boldsymbol{B}_2) & -\boldsymbol{f}\\ -\boldsymbol{f}^T & -c-\lambda-\alpha\rho_1-\beta\rho_2\end{pmatrix}\succeq\mathbf{0}\right.\right\}.$$

Now, using the following identities (see [14]):

$$\begin{aligned}\mathbf{Q}&=\boldsymbol{I}_r\otimes\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^T=(\boldsymbol{I}_r\otimes\tilde{\boldsymbol{x}})(\boldsymbol{I}_r\otimes\tilde{\boldsymbol{x}})^T,\\ \boldsymbol{f}&=\mathrm{vec}(\tilde{\boldsymbol{x}}(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})^T)=(\boldsymbol{I}_r\otimes\tilde{\boldsymbol{x}})(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})\end{aligned}$$

the dual problem is transformed to

$$\min_{\alpha\ge0,\beta\ge0,\lambda}\left\{-\lambda\left|\begin{pmatrix}-(\boldsymbol{I}_r\otimes\tilde{\boldsymbol{x}})(\boldsymbol{I}_r\otimes\tilde{\boldsymbol{x}}^T)+\boldsymbol{I}_r\otimes(\alpha\boldsymbol{B}_1+\beta\boldsymbol{B}_2) & -(\boldsymbol{I}_r\otimes\tilde{\boldsymbol{x}})(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})\\ -(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})^T(\boldsymbol{I}_r\otimes\tilde{\boldsymbol{x}})^T & -\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|^2-\lambda-\alpha\rho_1-\beta\rho_2\end{pmatrix}\succeq\mathbf{0}\right.\right\},$$

which, by Schur complement, can be written as

$$\min_{\alpha\ge0,\beta\ge0,\lambda}\left\{-\lambda\left|\begin{pmatrix}\boldsymbol{I}_r & (\boldsymbol{I}_r\otimes\tilde{\boldsymbol{x}})^T & \boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\\ \boldsymbol{I}_r\otimes\tilde{\boldsymbol{x}} & \boldsymbol{I}_r\otimes(\alpha\boldsymbol{B}_1+\beta\boldsymbol{B}_2) & \mathbf{0}\\ (\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})^T & \mathbf{0} & -\lambda-\alpha\rho_1-\beta\rho_2\end{pmatrix}\succeq\mathbf{0}\right.\right\}.$$

Finally, we arrive at the following SDP formulation of the RLS problem (36):

(43)
$$\min_{\alpha\geq 0,\beta\geq 0,\lambda,x}\left\{-\lambda \left| \begin{pmatrix} \boldsymbol{I}_r & (\boldsymbol{I}_r\otimes\tilde{\boldsymbol{x}})^T & \boldsymbol{A}\boldsymbol{x}-\boldsymbol{b} \\ \boldsymbol{I}_r\otimes\tilde{\boldsymbol{x}} & \boldsymbol{I}_r\otimes(\alpha\boldsymbol{B}_1+\beta\boldsymbol{B}_2) & \boldsymbol{0} \\ (\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})^T & \boldsymbol{0} & -\lambda-\alpha\rho_1-\beta\rho_2 \end{pmatrix}\succeq \boldsymbol{0}\right.\right\}.$$

We summarize the discussion in this section in the following theorem.

THEOREM 3.8. *Consider the RLS problem* (36)*, where the uncertainty set* $\mathcal{U}$ *is given by* (38)*,* $r>1$*, and* $\boldsymbol{B}_1,\boldsymbol{B}_2$ *satisfy condition* (39)*. Let* $(\alpha,\beta,\lambda,\boldsymbol{x})$ *be a solution to the SDP problem* (43)*, where* $\tilde{\boldsymbol{x}}$ *is given in* (40)*. Then* $\boldsymbol{x}$ *is the optimal solution of the RLS problem* (36)*.*

Appendix. Extended Finsler's theorem.

THEOREM A.1 (Finsler's theorem [11, 21]). *Let* $\mathbb{F}$ *be one of the fields* $\mathbb{R}$ *or* $\mathbb{C}$ *and let* $\boldsymbol{A},\boldsymbol{B}\in\mathbb{R}^{n\times n}$ *be symmetric matrices. Suppose that there exist* $\boldsymbol{x}_1,\boldsymbol{x}_2\in\mathbb{F}^n$ *such that* $\boldsymbol{x}_1^*\boldsymbol{A}\boldsymbol{x}_1>0$ *and* $\boldsymbol{x}_2^*\boldsymbol{A}\boldsymbol{x}_2<0$*. Then*

$$\boldsymbol{z}^*\boldsymbol{B}\boldsymbol{z}\geq 0 \text{ for every } \boldsymbol{z}\in\mathbb{F}^n \text{ such that } \boldsymbol{z}^*\boldsymbol{A}\boldsymbol{z}=0$$

if and only if there exists $\alpha\in\mathbb{R}$ *such that* $\boldsymbol{B}-\alpha\boldsymbol{A}\succeq\boldsymbol{0}$*.*

We note that the complex case can be reduced to the real case by using

$$\boldsymbol{z}^*\boldsymbol{A}\boldsymbol{z}=(\boldsymbol{x}^T\boldsymbol{y}^T)\begin{pmatrix}\boldsymbol{A} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{A}\end{pmatrix}\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{pmatrix}$$

for all $\boldsymbol{z}=\boldsymbol{z}+i\boldsymbol{y},\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^n$, where $\boldsymbol{A}\in\mathbb{R}^{n\times n}$ is symmetric.

While Finsler's theorem deals with *homogeneous* quadratic forms, the extended version considers *nonhomogeneous* quadratic functions.

THEOREM A.2 (extended Finsler's theorem). *Let* $\mathbb{F}$ *be one of the fields* $\mathbb{R}$ *or* $\mathbb{C}$ *and let* $\boldsymbol{A}_1,\boldsymbol{A}_2\in\mathbb{R}^{n\times n}$ *be symmetric matrices such that*

$$\boldsymbol{A}_2\succeq\eta\boldsymbol{A}_1 \text{ for some } \eta\in\mathbb{R}. \tag{44}$$

Let $f_j:\mathbb{F}^n\to\mathbb{R}, f_j(\boldsymbol{x})=\boldsymbol{x}^*\boldsymbol{A}_j\boldsymbol{x}+2\Re(\boldsymbol{b}_j^T\boldsymbol{x})+c_j$*, where* $\boldsymbol{b}_j\in\mathbb{R}^n$ *and* c_j *is a real scalar.*[5] *Suppose that*

$$\exists\boldsymbol{x}_1,\boldsymbol{x}_2\in\mathbb{F}^n \text{ such that } f_1(\boldsymbol{x}_1)>0, f_1(\boldsymbol{x}_2)<0. \tag{45}$$

Then the following two statements are equivalent:

(i) $f_2(\boldsymbol{x})\geq 0$ *for every* $\boldsymbol{x}\in\mathbb{F}^n$ *such that* $f_1(\boldsymbol{x})=0$*.*

(ii) *There exists* $\lambda\in\mathbb{R}$ *such that*

$$\begin{pmatrix}\boldsymbol{A}_2 & \boldsymbol{b}_2\\ \boldsymbol{b}_2^T & c_2\end{pmatrix}\succeq\lambda\begin{pmatrix}\boldsymbol{A}_1 & \boldsymbol{b}_1\\ \boldsymbol{b}_1^T & c_1\end{pmatrix}.$$

Proof. (ii) $\Rightarrow$ (i) is a trivial implication. Now, suppose that (i) is satisfied. Making the change of variables $\boldsymbol{x}=(1/t)\boldsymbol{y}$ $(\boldsymbol{y}\in\mathbb{F}^n,t\neq 0)$ and multiplying f_1 and f_2 by $|t|^2$, (i) becomes

$$g_2(\boldsymbol{y},t)\geq 0 \text{ for every } \boldsymbol{y}\in\mathbb{F}^n,t\neq 0 \text{ such that } g_1(\boldsymbol{y},t)=0, \tag{46}$$

[5] In the case $\mathbb{F}=\mathbb{R}$, f_j can be written as $f_j(\boldsymbol{x})=\boldsymbol{x}^T\boldsymbol{A}_j\boldsymbol{x}+2\boldsymbol{b}_j^T\boldsymbol{x}+c_j$.

where $g_j(\boldsymbol{y}, t) = \boldsymbol{y}^* \boldsymbol{A}_j \boldsymbol{y} + 2\Re(\boldsymbol{b}_j^T \boldsymbol{y} t) + c_j |t|^2$. Notice that if t would not be restricted to be nonzero, then by Theorem A.1, statement (ii) is true (g_1 and g_2 are homogeneous quadratic functions). Thus, all is left to prove is that (46) is true for $t = 0$. However, by replacing $t \neq 0$ with $t = 0$, (46) reduces to

$$\boldsymbol{y}^* \boldsymbol{A}_2 \boldsymbol{y} \geq 0 \text{ for every } \boldsymbol{y} \in \mathbb{F}^n \text{ such that } \boldsymbol{y}^* \boldsymbol{A}_1 \boldsymbol{y} = 0,$$

which, by Theorem A.1, is equivalent to condition (44). □

The condition in Theorem A.2 holds true, for instance, if $\boldsymbol{A}_2$ is positive definite or if $\boldsymbol{A}_1$ is definite. The case in which $\boldsymbol{A}_1$ is definite was already proven for the real case in [33, Corollary 6].

REFERENCES

[1] K. Anstreicher and H. Wolkowicz, *On Lagrangian relaxation of quadratic matrix constraints*, SIAM J. Matrix Anal. Appl., 22 (2000), pp. 41–55.

[2] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming: Theory and Algorithms*, 2nd ed., John Wiley and Sons, New York, 1993.

[3] A. Ben-Tal and A. Nemirovski, *Robust convex optimization*, Math. Oper. Res., 23 (1998), pp. 769–805.

[4] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*, MPS-SIAM Ser. Optim. 2, SIAM, Philadelphia, 2001.

[5] A. Ben-Tal and M. Teboulle, *Hidden convexity in some nonconvex quadratically constrained quadratic programming*, Math. Programming, 72 (1996), pp. 51–63.

[6] M. R. Celis, J. E. Dennis, and R. A. Tapia, *A trust region strategy for nonlinear equality constrained optimization*, in Numerical Optimization, 1984 (Boulder, CO, 1984), SIAM, Philadelphia, 1985, pp. 71–82.

[7] S. Chandrasekaran, G. H. Golub, M. Gu, and A. H. Sayed, *Parameter estimation in the presence of bounded data uncertainties*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 235–252.

[8] X. Chen and Y.-X. Yuan, *On local solutions of the Celis–Dennis–Tapia subproblem*, SIAM J. Optim., 10 (1999), pp. 359–383.

[9] P. Chevalier and B. Picinbono, *Complex linear-quadratic systems for detection and array processing*, IEEE Trans. Signal Process., 44 (1996), pp. 2631–2634.

[10] A. R. Conn, N. I. M. Gold, and P. L. Toint, *Trust-Region Methods*, MPS/SIAM Ser. Optim. 1, SIAM, Philadelphia, 2000.

[11] P. Finsler, *Uber das vorkommen definiter und semi-definiter formen in scharen quadratische formen*, Commentarii Mathematici Helvetici, 9 (1937), pp. 188–192.

[12] C. Fortin and H. Wolkowicz, *The trust region subproblem and semidefinite programming*, Optim. Methods Softw., 19 (2004), pp. 41–67.

[13] A. L. Fradkov and V. A. Yakubovich, *The S-procedure and the duality relation in convex quadratic programming problems*, Vestnik Leningrad. Univ., 1 (1973), pp. 81–87, 155.

[14] D. M. Gay, *Computing optimal locally constrained steps*, SIAM J. Sci. Statist. Comput., 2 (1981), pp. 186–197.

[15] D. M. Gay, *A trust-region approach for linearly constrained optimization*, in Proceedings of the Dundee Biennial Conference on Numerical Analysis, Dundee, Austalia, 1983, Lecture Notes in Math. 1066, Springer, Berlin, 1984, pp. 72–105.

[16] L. El Ghaoui and H. Lebret, *Robust solution to least-squares problems with uncertain data*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 1035–1064.

[17] J. B. Hiriart-Urruty and M. Torki, *Permanently going back and forth between the "quadratic world" and the "convexity world" in optimization*, Appl. Math. Optim., 45 (2002), pp. 169–184.

[18] Y. Huang and S. Zhang, *Complex Matrix Decomposition and Quadratic Programming*, Technical report SEEM2005-02, Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong, 2005.

[19] J. B. Lasserre, *Semidefinite programming versus LP relaxations for polynomial programming*, Math. Oper. Res., 27 (2002), pp. 347–360.

[20] J. M. Martínez, *Local minimizers of quadratic functions on Euclidean balls and spheres*, SIAM J. Optim., 4 (1994), pp. 159–176.

[21] J. J. Moré, *Generalizations of the trust region subproblem*, Optim. Methods Softw. 2, (1993), pp. 189–209.
[22] J. J. Moré and D. C. Sorensen, *Computing a trust region step*, SIAM J. Sci. Statist. Comput., 4 (1983), pp. 553–572.
[23] J.-M. Peng and Y. Yuan, *Optimality conditions for the minimization of a quadratic with two quadratic constraints*, SIAM J. Optim., 7 (1997), pp. 579–594.
[24] B. Picinbono and P. Chevalier, *Widely linear estimation with complex data*, IEEE Trans. Signal Process., 43 (1995), pp. 2030–2033.
[25] B. T. Polyak, *Convexity of quadratic transformations and its use in control and optimization*, J. Optim. Theory Appl., 99 (1998), pp. 553–583.
[26] M. J. D. Powell and Y. Yuan, *A trust region algorithm for equality constrained optimization*, Math. Programming, 49 (1990/91), pp. 189–211.
[27] M. V. Ramana and A. J. Goldman, *Some geometric results in semidefinite programming*, J. Global Optim., 7 (1995), pp. 33–50.
[28] P. J. Schreier, L. L. Scharf, and C. T. Mullis, *Detection and estimation of improper complex random signals*, IEEE Trans. Inform. Theory, 51 (2005), pp. 306–312.
[29] N. Z. Shor, *Quadratic optimization problems*, Izv. Akad. Nauk SSSR Tekhn. Kibernet., 1 (1987), pp. 128–139, 222.
[30] D. C. Sorensen, *Newton's method with a model trust region modification*, SIAM J. Numer. Anal., 19 (1982), pp. 409–426.
[31] R. J. Stern and H. Wolkowicz, *Indefinite trust region subproblems and nonsymmetric eigenvalue perturbations*, SIAM J. Optim., 5 (1995), pp. 286–313.
[32] J. F. Sturm, *Using SeDuMi* 1.02*, a MATLAB toolbox for optimization over symmetric cones*, Optim. Methods Softw., 11/12 (1999), pp. 625–653.
[33] J. F. Sturm and S. Zhang, *On cones of nonnegative quadratic functions*, Math. Oper. Res., 28 (2003), pp. 246–267.
[34] L. Vandenberghe and S. Boyd, *Semidefinite programming*, SIAM Rev., 38 (1996), pp. 40–95.
[35] Y. Ye and S. Zhang, *New results on quadratic minimization*, SIAM J. Optim., 14 (2003), pp. 245–267.
[36] Y. Yuan, *On a subproblem of trust region algorithms for constrained optimization*, Math. Programming, 47 (1990), pp. 53–63.
[37] Y. Zhang, *Computing a Celis–Dennis–Tapia trust-region step for equality constrained optimization*, Math. Programming, 55 (1992), pp. 109–124.

SIAM J. OPTIM.
Vol. 17, No. 3, pp. 861–870
© 2006 Society for Industrial and Applied Mathematics

DISCONTINUOUS BUT WELL-POSED OPTIMIZATION PROBLEMS*

JACQUELINE MORGAN† AND VINCENZO SCALZO‡

Abstract. Using classes of *sequentially pseudocontinuous* functions, recently introduced by the authors, the aim of the paper is to investigate Tikhonov and parametric well-posedness for optimization problems when the objective functions are not necessarily sequentially lower semicontinuous. Sequential pseudocontinuity is a property more general than sequential semicontinuity and finds motivations in choice theory, since the continuity of preference relations on first countable topological spaces is characterized by the sequential pseudocontinuity of any utility function. Examples show that it is not possible to improve the results with other well-known classes of discontinuous functions.

Key words. Tikhonov well-posedness, parametric well-posedness, minimum problems, parametric minimum problems, sequentially pseudocontinuous functions

AMS subject classifications. 49K27, 49K40, 90C31

DOI. 10.1137/050636358

1. Introduction.

The problem of minimizing an extended real valued function f is declared *well-posed* if any minimizing sequence clusters to a minimizer. That is, if $f(y_n) \longrightarrow \inf f$ and f has only one minimum point, then $(y_n)_n$ converges to the minimum point, while if f has more than one minimum point, then some subsequence $(y_{n_k})_k$ converges to a minimizer. This crucial property—most informative when the minimum point is unique and also interesting from a numerical point of view—was introduced by Tikhonov in [12], and it was the objective of numerous studies: see, for example, [3], [13], and [6]. It is commonly assumed there that f is at least sequentially lower semicontinuous. There are, however, some important instances in which even less than sequentially lower semicontinuity prevails. To wit, in the framework of choice theory when an abstract preference relation $\succeq$ is represented by a utility function, to assume $\succeq$ *continuous* on a first countable topological space (see [4] and [2]) amounts to requiring that every utility function representing $\succeq$ be *sequentially pseudocontinuous* (Proposition 1 in [9]). A sequentially pseudocontinuous function is not necessarily sequentially lower semicontinuous; see [7]. Broadly, sequential pseudocontinuity requires that strict inequalities be preserved along approximating sequences (see Proposition 2.3). Moreover, if $\succeq$ is a *weakly continuous* preference relation [1] represented by utility functions, then any such utility satisfies a property introduced in [8] which is called *sequential weak pseudocontinuity* in the following. The sequential weak pseudocontinuity generalizes the sequential pseudocontinuity; see [8]. The sequential (weak) pseudocontinuity allows us to extend several well-known results already obtained for functions at least sequentially lower semicontinuous: the Weierstrass theorem, convergence results for minimum points and for social Nash equilibria, existence results for MinSup and MinInf problems; see [7] and [8]. Finally, the sequential (weak) pseudocontinuity is related to monotone functions: any strictly monotone function is a sequentially pseudocontinuous function and any monotone function is a

*Received by the editors July 19, 2005; accepted for publication (in revised form) May 13, 2006; published electronically October 16, 2006.
http://www.siam.org/journals/siopt/17-3/63635.html
†Dipartimento di Matematica e Statistica, Università di Napoli Federico II, via Cinthia, 80126 Napoli, Italy (morgan@unina.it).
‡Dipartimento di Matematica e Applicazioni "R. Caccioppoli," Università di Napoli Federico II, via Cinthia, 80126 Napoli, Italy (scalzo@unina.it).

sequentially weakly pseudocontinuous function; see [7] and [8] for finite dimensional spaces.

The aim of the paper is to generalize the results already obtained for Tikhonov well-posedness and parametric well-posedness using just the sequentially pseudocontinuous or the sequentially weakly pseudocontinuous functions. The paper is organized as follows. In section 2 we present some properties and a useful characterization for these classes of functions. In section 3 we obtain new sufficient conditions for Tikhonov well-posedness of unconstrained optimization problems. In section 4 we consider the case of constrained and parametric problems and compare our results with the previous ones (see [6]). Finally, examples show that it is not possible to further improve our results.

2. Pseudocontinuous functions. This section recalls some well-known concepts and introduces a few classes of functions.

For simplicity let all spaces be metric. However, we point out that all the results of the paper could be proved, using the same arguments, in the more general framework of *sequential convergence spaces* (see Kuratowski [5]).

Let f be an extended real valued function defined on a metric space Z. The function f is *sequentially lower semicontinuous* at $z \in Z$ if

$$f(z) \leq \liminf_{n\to\infty} f(z_n) \ \forall\ z_n \longrightarrow z \text{ in } Z,$$

and f is *sequentially upper semicontinuous* at z if $-f$ is sequentially lower semicontinuous at z.

Let $(A_n)_n$ be a sequence of subsets of Z; then

- $z \in \operatorname{Liminf} A_n$ (see [5], *inner limit* of $(A_n)_n$ in [10]) if and only if there exists a sequence $(z_n)_n$ converging in Z to z and such that $z_n \in A_n$ for n sufficiently large ($n \in \mathbb{N}$);
- $z \in \operatorname{Limsup} A_n$ (see [5], *outer limit* of $(A_n)_n$ in [10]) if and only if there exists a subsequence (A_{n_k}) of $(A_n)_n$ and a sequence $(z_k)_k$ converging to z in Z such that $z_k \in A_{n_k}$ for each $k \in \mathbb{N}$.

Let K be a set-valued function from X to Y, two metric spaces.

- K is *sequentially lower semicontinuous* at a point $x \in X$ if $K(x) \subseteq \operatorname{Liminf} K(x_n)$ for all $x_n \longrightarrow x$.
- K is *sequentially closed* at a point $x \in X$ if $\operatorname{Limsup} K(x_n) \subseteq K(x)$ for all $x_n \longrightarrow x$.

Let us remind the reader about some definitions introduced in [7] and [8].

Definition 2.1 (Definition 2.4 in [7]). *Let f be an extended real valued function defined on Z and $z \in Z$.*

- *f is said to be* sequentially lower pseudocontinuous at z *if*

$$f(y) < f(z) \Rightarrow \begin{cases} f(y) < \liminf\limits_{n\to\infty} f(z_n) \\ \forall\ z_n \longrightarrow z. \end{cases}$$

- *f is said to be* sequentially upper pseudocontinuous at z *if $-f$ is sequentially lower pseudocontinuous at z.*
- *f is said to be* sequentially pseudocontinuous at z *if it is both sequentially lower and upper pseudocontinuous at z.*

Definition 2.2 (Definition 3.1 in [8]). *Let f be an extended real valued function defined on Z and $z \in Z$.*

• *f is said to be* sequentially lower weakly pseudocontinuous at z *(sequentially lower quasi-continuous in* [8]*) if*

$$f(y) < f(z) \Rightarrow \begin{cases} f(y) \leq \liminf\limits_{n\to\infty} f(z_n) \\ \forall\ z_n \longrightarrow z. \end{cases}$$

• *f is said to be* sequentially upper weakly pseudocontinuous at z *(sequentially upper quasi-continuous in* [8]*) if $-f$ is sequentially lower weakly pseudocontinuous at z.*

• *f is said to be* sequentially weakly pseudocontinuous at z *(sequentially quasi-continuous in* [8]*) if it is both sequentially lower and upper weakly pseudocontinuous at z.*

Trivially a sequentially lower pseudocontinuous function is also sequentially lower weakly pseudocontinuous. Conversely, however, the well-known Dirichlet function (which is equal to 0 on all rational numbers and equal to 1 on all irrational numbers) is sequentially lower weakly pseudocontinuous, but it is not sequentially lower pseudocontinuous. Moreover, the class of sequentially lower pseudocontinuous functions strictly includes the class of sequentially lower semicontinuous function (see [7, Example 2.1]). Characterizations of the sequential lower pseudocontinuity are presented in [7]. For simplicity, from now on, we omit the term *sequentially*.

Finally, we prove a new characterization of pseudocontinuous functions, useful in the following.

PROPOSITION 2.3. *Let f be an extended real valued function defined on a metric space Z. Then f is pseudocontinuous on Z if and only if the following holds:*

$$\left.\begin{array}{c} f(x) < f(y) \\ x_n \longrightarrow x \\ y_n \longrightarrow y \end{array}\right\} \Longrightarrow \limsup_{n\to\infty} f(x_n) < \liminf_{n\to\infty} f(y_n). \tag{2.1}$$

Proof. First, assume that f is pseudocontinuous on Z. Let $f(x) < f(y)$, $x_n \longrightarrow x$, and $y_n \longrightarrow y$. We set $\mathrm{Im}(f) = \{f(z) \,/\, z \in Z\}$.

If there exists a value $f(z) \in]f(x), f(y)[$, then one has

$$\limsup_{n\to\infty} f(x_n) < f(z) < \liminf_{n\to\infty} f(y_n).$$

Otherwise, let $]f(x), f(y)[\cap \mathrm{Im}(f) = \emptyset$. Since f is upper pseudocontinuous at x, one has

$$\limsup_{n\to\infty} f(x_n) < f(y).$$

Now, if $f(x) < \limsup_{n\to\infty} f(x_n)$, then $]f(x), f(y)[\cap \mathrm{Im}(f) \neq \emptyset$, which is in conflict with our assumption. So, $\limsup_{n\to\infty} f(x_n) \leq f(x) < f(y)$. Similarly, f being lower pseudocontinuous at y, one gets $f(y) \leq \liminf_{n\to\infty} f(y_n)$ and then

$$\limsup_{n\to\infty} f(x_n) \leq f(x) < f(y) \leq \liminf_{n\to\infty} f(y_n),$$

that is, the property (2.1).

Finally, assume that the property (2.1) is satisfied. Let $f(x) < f(y)$, $x_n \longrightarrow x$, and $(y_n)_n$ such that $y_n = y$ for all n. Then

$$\limsup_{n\to\infty} f(x_n) < f(y);$$

that is, f is upper pseudocontinuous at x. Similarly, one can prove that property (2.1) implies the lower pseudocontinuity of f at y. □

Moreover, pseudocontinuity is connected with monotonicity. In fact, the following result extends the results given in [7] and [8] for finite dimensional spaces.

PROPOSITION 2.4. *Let f be an extended real valued function defined on a normed space V, and let $\mathcal{C}$ be a convex and pointed cone in V with its apex at the origin and nonempty interior. If f is strictly monotone (resp., monotone) with respect to $\mathcal{C}$, that is, strictly increasing (resp., increasing),*

$$y \in x + \operatorname{int}\mathcal{C} \Longleftrightarrow f(x) < f(y) \ \ (\textit{resp.}, \ \ f(x) \leq f(y)),$$

or strictly decreasing (resp., decreasing),

$$y \in x + \operatorname{int}\mathcal{C} \Longleftrightarrow f(x) > f(y) \ \ (\textit{resp.}, \ \ f(x) \geq f(y)),$$

then f is pseudocontinuous (resp., weakly pseudocontinuous) on V.

Proof. Suppose that f is strictly decreasing with respect to $\mathcal{C}$.

We first prove that f is lower pseudocontinuous. Let z and y be such that $f(y) < f(z)$, and let $z_n \longrightarrow z$. Then we have $y \in z + \operatorname{int}\mathcal{C}$. So, there exists an element $y' \in y - \operatorname{int}\mathcal{C}$ and an open neighborhood $\mathcal{A}$ of z such that $y' \in u + \operatorname{int}\mathcal{C}$ for all $u \in \mathcal{A}$. Since $z_n \longrightarrow z$, we have that $y' \in z_n + \operatorname{int}\mathcal{C}$ for n sufficiently large. Then

$$f(y) < f(y') \leq \liminf_{n\to\infty} f(z_n).$$

Now we prove that f is upper pseudocontinuous. Let z and y be such that $f(z) < f(y)$, and let $z_n \longrightarrow z$. Then $z \in y + \operatorname{int}\mathcal{C}$. Moreover, there exists $y' \in z - \operatorname{int}\mathcal{C}$ and an open neighborhood $\mathcal{B}$ of z such that $u \in y' + \operatorname{int}\mathcal{C}$ for all $u \in \mathcal{B}$. Consequently, $f(z_n) < f(y')$ for n sufficiently large. So,

$$\limsup_{n\to\infty} f(z_n) \leq f(y') < f(y).$$

Analogously, we obtain that f is pseudocontinuous if it is strictly increasing.

With similar arguments, one can prove that any monotone function is weakly pseudocontinuous. □

3. Well-posed unconstrained unparametric optimization. Let Y be a metric space and f be a proper function defined on Y with values in $]-\infty, +\infty]$. We recall that the minimum problem

$$\mathcal{M}: \ \min_{y\in Y} f(y)$$

is *Tikhonov well-posed* (see [12], [3]) if $-\infty < \inf f$ and there exists a unique global minimum point $\hat{y}$ and any sequence $(y_n)_n$ such that

$$f(y_n) - \min_{y\in Y} f(y) \longrightarrow 0$$

(called a *minimizing sequence*) is converging to $\hat{y}$. Moreover, $\mathcal{M}$ is said to be *Tikhonov well-posed in the generalized sense* (see [12], [3]) if $-\infty < \inf f$, $\operatorname{argmin}(Y, f) = \{y' \in Y \,/\, f(y') \leq f(y) \text{ for all } y \in Y\}$ is nonempty, and any minimizing sequence has at least a subsequence converging to a global minimum point.

In the following, for simplicity, we refer to well-posed and well-posed in the generalized sense problems.

Sufficient conditions for well-posedness and well-posedness in the generalized sense of the problem $\mathcal{M}$ have been obtained for lower semicontinuous functions (see Chapter 1 in [3]).

In this section, we generalize the previous results using *lower weakly pseudocontinuous* and *lower pseudocontinuous* functions. In fact, we have the following theorem.

THEOREM 3.1. *Let Y be a compact metric space, and let $f : Y \longrightarrow]-\infty, +\infty]$ be a proper function. If* $\operatorname{argmin}(Y, f) = \{\hat{y}\}$ *and f is lower weakly pseudocontinuous on Y, then $\mathcal{M}$ is well-posed.*

Proof. Let $(y_n)_n$ be a minimizing sequence of $\mathcal{M}$ which does not converge to $\hat{y}$. Since Y is compact, there exists a subsequence $(y_{n_k})_k$ converging to a point $\bar{y} \neq \hat{y}$. Hence, $\lim_{k\to\infty} f(y_{n_k}) = f(\hat{y}) < f(\bar{y})$, and we have $f(y_{n_k}) \in [f(\hat{y}), f(\bar{y})[$ for k sufficiently large. If $f(y_{n_k}) = f(\hat{y})$ for k sufficiently large, then $y_{n_k} = \hat{y}$, which is impossible since $y_{n_k} \longrightarrow \bar{y}$. So, there exists $y' \in Y$ such that $f(y') \in]f(\hat{y}), f(\bar{y})[$. Since f is lower weakly pseudocontinuous at $\bar{y}$, we have $f(y') \leq \liminf_{k\to\infty} f(y_{n_k}) = f(\hat{y})$, and we get a contradiction. □

About well-posedness in the generalized sense, sufficient conditions are obtained using lower pseudocontinuous functions.

THEOREM 3.2. *Let Y be a compact metric space, and let $f : Y \longrightarrow]-\infty, +\infty]$ be a proper function. If f is lower pseudocontinuous on Y, then $\mathcal{M}$ is well-posed in the generalized sense.*

Proof. First, in light of [8, Corollary 3.1], we have that $\operatorname{argmin}(Y, f)$ is nonempty. Assume a minimizing sequence $(y_n)_n$ has a subsequence $(y_{n_k})_k$ that converges to $\bar{y} \notin \operatorname{argmin}(Y, f)$. So, there exists $y \in Y$ such that $f(y) < f(\bar{y})$. Since f is lower pseudocontinuous at $\bar{y}$, we have $f(y) < \liminf_{k\to\infty} f(y_{n_k}) = \min_{z\in Y} f(z)$, and we get a contradiction. □

In order to obtain sufficient conditions for well-posedness, lower weak pseudocontinuity (used in Theorem 3.1) cannot be weakened using the minimal conditions for the existence of minimum points in a sequential setting (see [8] for general sequential convergence spaces and [11] for topological spaces). In fact, in Example 3.1, a *transfer lower continuous* function (see [11]) on a compact space determines a minimum problem which is not well-posed.

Example 3.1. Let $Y = [0, 2]$ and $f : Y \longrightarrow \mathbb{R}$ be such that

$$f(y) = \begin{cases} (y-1)^2 & \text{if } y \in [0, 1[, \\ 2 - y & \text{if } y \in [1, 2]. \end{cases}$$

The function f is not lower weakly pseudocontinuous at $y = 1$, but it is transfer lower continuous on $[0, 2]$. Now, if $y_n \longrightarrow 1^-$, we have $f(y_n) \longrightarrow 0 = \min f$, but $(y_n)_n$ does not converge to the unique minimum point $\hat{y} = 2$. So, the associate minimum problem is not well-posed.

As shown by Example 3.2, lower pseudocontinuity (used in Theorem 3.2) cannot be substituted by lower weak pseudocontinuity in order to obtain a minimum problem well-posed in the generalized sense.

Example 3.2. Let $Y = [0, 2]$ and $f : Y \longrightarrow \mathbb{R}$ be such that

$$f(y) = \begin{cases} 0 & \text{if } y \in [0, 2]\backslash\{1\}, \\ 1 & \text{if } y = 1. \end{cases}$$

The function f is not lower pseudocontinuous at $y = 1$, but it is lower weakly pseudocontinuous on $[0, 2]$. If $y_n \longrightarrow 1^-$, we have $f(y_n) \longrightarrow 0 = \min f$, but 1 is not a

minimum point. So, the associate minimum problem is not well-posed in the generalized sense.

In order to obtain well-posedness, compactness of Y can be replaced by completeness of Y and a suitable diameter assumption (already considered in [3]). For any $\varepsilon > 0$, let

$$M(\varepsilon) = \left\{ y \in Y \,/\, f(y) - \inf_{z \in Y} f(z) < \varepsilon \right\},$$

and let $diam[M(\varepsilon)]$ be the diameter of $M(\varepsilon)$.

THEOREM 3.3. *Let $f : Y \longrightarrow]-\infty, +\infty]$ be a proper function. If $\mathcal{M}$ is well-posed, then*

$$\lim_{\varepsilon \downarrow 0} diam[M(\varepsilon)] = 0. \tag{3.1}$$

Moreover, if Y is complete, f is lower pseudocontinuous and bounded from below on Y, and (3.1) *is satisfied, then $\mathcal{M}$ is well-posed.*

Proof. The first part of the thesis is given in [3, Theorem 11]. Here we prove that (3.1) is a sufficient condition for well-posedness.

Let $(y_n)_n$ be a minimizing sequence of $\mathcal{M}$. In light of (3.1), $(y_n)_n$ is a Cauchy sequence. Y being complete, $(y_n)_n$ converges to a point $\hat{y} \in Y$. If $\hat{y} \notin \operatorname{argmin}(Y, f)$, there exists $y \in Y$ such that $f(y) < f(\hat{y})$. Since f is lower pseudocontinuous at $\hat{y}$, we get the following contradiction:

$$f(y) < \liminf_{n \to \infty} f(y_n) = \inf_{z \in Y} f(z).$$

So, $\hat{y} \in \operatorname{argmin}(Y, f)$. Again from (3.1), it follows that $\operatorname{argmin}(Y, f) = \{\hat{y}\}$, and the proof is concluded. ∎

4. Well-posed constrained parametric optimization. Given X, Y, two metric spaces, let $f : X \times Y \longrightarrow]-\infty, +\infty]$ be a proper function and K be a set-valued function defined on X with nonempty values in Y. For any $x \in X$, we consider the following parametric minimum problem:

$$\mathcal{M}(x) : \quad \min_{y \in K(x)} f(x, y).$$

Let $\boldsymbol{M} = \{\mathcal{M}(x) \,/\, x \in X\}$. Following [13], the family $\boldsymbol{M}$ is said to be *parametrically well-posed at $x \in X$* if

(i) $-\infty < \inf\{f(x, y) \,/\, y \in K(x)\}$ and there exists a unique global solution to $\mathcal{M}(x)$;

(ii) if $x_n \longrightarrow x$, any sequence $(y_n)_n \subseteq Y$, with $y_n \in K(x_n)$ for n sufficiently large and such that

$$f(x_n, y_n) - \inf_{z \in K(x_n)} f(x_n, z) \longrightarrow 0,$$

is converging to the unique global solution to $\mathcal{M}(x)$.

If $x_n \longrightarrow x$, a sequence $(y_n)_n \subseteq Y$ which satisfies the above condition (ii) is said to be an *approximating sequence* of $\mathcal{M}(x)$ (with respect to $(x_n)_n$).

Moreover, if $\operatorname{argmin}(K(x), f(x, \cdot))$ is nonempty, $\boldsymbol{M}$ is said to be *parametrically well-posed in the generalized sense at $x \in X$* if, for every sequence $x_n \longrightarrow x$, any

approximating sequence (with respect to $(x_n)_n$) has some subsequence converging to a point of $\operatorname{argmin}(K(x), f(x,\cdot))$.

The following theorem gives sufficient conditions for the parametric well-posedness of the family $\boldsymbol{M}$, explicit on any data and weaker than continuity of the objective function. Let ν be the marginal function defined on X by $\nu(x) = \inf_{y\in K(x)} f(x,y)$.

THEOREM 4.1. *Let $x \in X$. If Y is compact and*

(i) *the function f is pseudocontinuous at (x,y), for any $y \in K(x)$, and*

(ii) *the set-valued function K is closed and lower semicontinuous at x,*

then the family $\boldsymbol{M}$ is parametrically well-posed in the generalized sense at x. Moreover, if $\mathcal{M}(x)$ has only one solution, then $\boldsymbol{M}$ is parametrically well-posed at x.

Proof. Since $K(x)$ is compact and f is lower pseudocontinuous at (x,y) for all $y \in K(x)$, in light of [8, Corollary 3.1], $\operatorname{argmin}(K(x), f(x,\cdot))$ is nonempty. Now we prove that $\boldsymbol{M}$ is parametrically well-posed in the generalized sense at x. Let $x_n \longrightarrow x$ and $(y_n)_n$ be an approximating sequence (with respect to $(x_n)_n$) such that any converging subsequence does not converge to an element of $\operatorname{argmin}(K(x), f(x,\cdot))$. Let $(y_{n_k})_k$ converge to a point y which is not a global minimum point of $f(x,\cdot)$ over $K(x)$. Since K is closed at x, we have that $y \in K(x)$ and there exists $z \in K(x)$ such that $f(x,z) < f(x,y)$. K being lower semicontinuous at x, there exists a sequence $z_k \longrightarrow z$ such that $z_k \in K(x_{n_k})$ for k sufficiently large. From Proposition 2.3, one has

$$\limsup_{k\to\infty} \nu(x_{n_k}) \le \limsup_{k\to\infty} f(x_{n_k}, z_k) < \liminf_{k\to\infty} f(x_{n_k}, y_{n_k}). \tag{4.1}$$

Let α be a real number such that

$$\limsup_{k\to\infty} \nu(x_{n_k}) < \alpha < \liminf_{k\to\infty} f(x_{n_k}, y_{n_k}). \tag{4.2}$$

Therefore, there exists $k_o \in \mathbb{N}$ such that

$$\nu(x_{n_k}) - f(x_{n_k}, y_{n_k}) < \alpha - f(x_{n_k}, y_{n_k}) \tag{4.3}$$

for all $k \ge k_o$. So, we obtain

$$0 = \lim_{k\to\infty} [\nu(x_{n_k}) - f(x_{n_k}, y_{n_k})] \le \alpha - \liminf_{k\to\infty} f(x_{n_k}, y_{n_k}) < 0,$$

and we get a contradiction.

Assume now that $\operatorname{argmin}(K(x), f(x,\cdot)) = \{\hat{y}\}$. Let $(y_n)_n$ be an approximating sequence (with respect to $(x_n)_n$) which does not converge to $\hat{y}$. By compactness of Y, there exists a subsequence $(y_{n_k})_k$ of $(y_n)_n$ converging to a point $y \in Y\backslash\{\hat{y}\}$. Now the thesis follows using the same arguments as in the previous case. ∎

Sufficient conditions, weaker than continuity of f, for the parametric well-posedness of $\boldsymbol{M}$ at a point x are given in [6]. More precisely, a lower semicontinuous function f, whose marginal function ν is assumed to be upper semicontinuous, is considered. Obviously, the assumptions of Theorem 4.1 are not connected with the assumptions used in [6]. Moreover, we note that the result obtained in [6] can be improved replaying the lower semicontinuity of f with the lower pseudocontinuity. In fact, we have the following theorem.

THEOREM 4.2. *Let $x \in X$. If Y is compact and*

(i) *the function f is lower pseudocontinuous at (x,y), for any $y \in K(x)$,*

(ii) *the function ν is upper semicontinuous at x, and*

(iii) *the set-valued function K is closed at x,*

then the family $\boldsymbol{M}$ is parametrically well-posed in the generalized sense at x. Moreover, if $\mathcal{M}(x)$ has only one solution, then $\boldsymbol{M}$ is parametrically well-posed at x.

Proof. Let $x_n \longrightarrow x$. Assume an approximating sequence and $(y_n)_n$ (with respect to $(x_n)_n$) have a subsequence $(y_{n_k})_k$ that converges to $y \notin \operatorname{argmin}(K(x), f(x,\cdot))$ (nonempty in light of [8, Corollary 3.1]). So, one has $f(x,z) < f(x,y)$ for some $z \in K(x)$, and f being lower pseudocontinuous at (x,y), one gets

$$\nu(x) \leq f(x,z) < \liminf_{k\to\infty} f(x_{n_k}, y_{n_k}).$$

Since ν is upper semicontinuous at x, one obtains (4.2) for some real number α. So, as in the proof of Theorem 4.1, it follows that $\boldsymbol{M}$ is parametrically well-posed in the generalized sense at x.

If $\mathcal{M}(x)$ has only one solution, similarly one can obtain that $\boldsymbol{M}$ is parametrically well-posed at x. □

Note that the assumptions on the function f in Theorem 4.1 are not connected with those in Theorem 4.2, as shown by Examples 4.1 and 4.2.

Example 4.1. Let $f : [0,1] \times [0,1] \longrightarrow \mathbb{R}$ be defined by

$$f(x,y) = \begin{cases} 2(1-x) & \text{if } (x,y) \in [0,1[\times [0,1/2], \\ 2y(1-x) & \text{if } (x,y) \in [0,1[\times]1/2,1], \\ -1 & \text{if } (x,y) \in \{1\} \times [0,1] \end{cases}$$

and $K(x) = [0,1]$ for any $x \in [0,1]$.

The function f is pseudocontinuous at $(1,y)$ for all $y \in [0,1]$, but the marginal function ν is not upper semicontinuous at $x = 1$.

Example 4.2. Let $f : [0,1] \times [0,1] \longrightarrow \mathbb{R}$ be defined by

$$f(x,y) = \begin{cases} x(y-1) & \text{if } (x,y) \in]0,1] \times [0,1[, \\ -1 & \text{if } (x,y) \in [0,1] \times \{1\}, \\ 0 & \text{if } (x,y) \in \{0\} \times [0,1[\end{cases}$$

and $K(x) = [0,1]$ for any $x \in [0,1]$.

All assumptions of Theorem 4.2 are satisfied at $x = 0$, but f is not upper pseudocontinuous at $(0,1)$.

Moreover, the assumption of pseudocontinuity used in Theorem 4.1 cannot be weakened with weak pseudocontinuity. In fact, Example 4.3 shows a parametric minimum problem with a weakly pseudocontinuous objective function which is not parametrically well-posed in the generalized sense.

Example 4.3. Let $f : [0,1] \times [0,3] \longrightarrow \mathbb{R}$ be the function defined as below:

- if $x \in [0,1[$,

$$f(x,y) = \begin{cases} 1-y & \text{if } y \in [0,1[, \\ 0 & \text{if } y \in [1,2], \\ (x-1)y - 1 & \text{if } y \in]2,3]. \end{cases}$$

- if $x = 1$,

$$f(x,y) = \begin{cases} 1-y & \text{if } y \in [0,1], \\ -1 & \text{if } y \in]1,3], \end{cases}$$

and let K be the set-valued function from $[0,1]$ to $[0,3]$ defined by $K(x) = [0,2x]$. The function f is weakly pseudocontinuous at $(1,y)$ for all $y \in [0,3]$, but it is not

pseudocontinuous at $(1,2)$, and K satisfies assumption (ii) of Theorem 4.1. If we consider the sequences $(x_n)_n = (1-1/n)_n$ and $(y_n)_n = (1-1/n)_n$, then $(y_n)_n$ is an approximating sequence of $\mathcal{M}(1)$ (with respect to $(x_n)_n$), and it converges to $1 \notin \operatorname{argmin}(K(1), f(1,\cdot)) =]1,2]$. Hence, $\boldsymbol{M}$ is not parametrically well-posed in the generalized sense at $x = 1$.

In order to obtain parametric well-posedness, the compactness of Y can be replaced by completeness and a diameter assumption. Let

$$M(x,\varepsilon) = \big\{y \in K(x) \;/\; f(x,y) - \nu(x) < \varepsilon\big\}.$$

The assumptions of Theorem 4.1 (or Theorem 4.2) are sufficient conditions for parametric well-posedness together with the following condition (already considered in [6]):

$$\lim_{\varepsilon \downarrow 0} diam[\cup_{u \in B(x,\varepsilon)} M(u,\varepsilon)] = 0, \tag{4.4}$$

where $B(x,\varepsilon)$ is the open ball with center x and ray ε. In fact, we have the following result.

THEOREM 4.3. *If $\boldsymbol{M}$ is parametrically well-posed at x, then* (4.4) *holds. Moreover, if Y is complete, f is bounded from below, and* (i) *and* (ii) *of Theorem* 4.1 *(or* (i), (ii), *and* (iii) *of Theorem* 4.2*) are satisfied, then* (4.4) *implies that $\boldsymbol{M}$ is parametrically well-posed at x.*

Proof. Assume that $\boldsymbol{M}$ is parametrically well-posed at x and $(\varepsilon_n)_n$ is a decreasing sequence of positive real numbers converging to 0. Let $A(\varepsilon) = \cup_{u \in B(x,\varepsilon)} M(u,\varepsilon)$. If (4.4) is not true, there exists a positive number ℓ and $n_o \in \mathbb{N}$ such that

$$\ell < diam[A(\varepsilon_n)] \quad \forall\, n \geq n_o.$$

So, for any $n \geq n_o$, there exist $y_n, z_n \in A(\varepsilon_n)$ such that $\ell < d(y_n, z_n)$. Consequently, for any $n \geq n_o$, there exist x_n^1 and x_n^2 belonging to $B(x,\varepsilon_n)$ such that

$$y_n \in M(x_n^1, \varepsilon_n) \quad \text{and} \quad z_n \in M(x_n^2, \varepsilon_n).$$

Now the sequences $(x_n^1)_n$ and $(x_n^2)_n$ converge to x, and the sequences $(y_n)_n$ and $(z_n)_n$ are approximating sequences of $\mathcal{M}(x)$ (with respect to $(x_n^1)_n$ and $(x_n^2)_n$, respectively). Since $\boldsymbol{M}$ is parametrically well-posed at x, the sequences $(y_n)_n$ and $(z_n)_n$ converge to the same point (that is the unique solution to $\mathcal{M}(x)$), and we get the following contradiction:

$$0 < \ell \leq \lim_{n \to \infty} d(y_n, z_n) = 0.$$

Assume now that $x_n \longrightarrow x$ and $(y_n)_n$ is an approximating sequence (with respect to $(x_n)_n$). From (4.4), it follows that $(y_n)_n$ is a Cauchy sequence, so it converges to a point $\hat{y}$. Since K is closed at x, one has $\hat{y} \in K(x)$. If $\hat{y}$ is not a solution to $\mathcal{M}(x)$, using the same arguments of the proof of Theorem 4.1, we get a contradiction. Hence $\hat{y}$ is a solution to $\mathcal{M}(x)$. Again from (4.4), $\operatorname{argmin}(K(x), f(x,\cdot)) = \{\hat{y}\}$, and the proof is concluded. ∎

Acknowledgment. The authors thank two anonymous referees for their valuable comments and suggestions on the previous versions of this paper.

REFERENCES

[1] D. E. Campbell and M. Walker, *Optimization with weak continuity*, J. Econom. Theory, 50 (1990), pp. 459–464.

[2] G. Debreu, *Continuity property of Paretian utility*, Internat. Econom. Rev., 5 (1964), pp. 285–293.

[3] A. L. Dontchev and T. Zolezzi, *Well-Posed Optimization Problems*, Springer-Verlag, Berlin, Heidelberg, New York, 1993.

[4] S. Eilenberg, *Order topological spaces*, Amer. J. Math., 63 (1941), pp. 39–45.

[5] C. Kuratowski, *Topology*, Academic Press, New York, 1966.

[6] R. Lucchetti and T. Zolezzi, *On well-posedness and stability analysis in optimization*, in Mathematical Programming with Data Perturbations, Lecture Notes in Pure and Appl. Math. 195, Dekker, New York, 1998, pp. 223–251.

[7] J. Morgan and V. Scalzo, *Pseudocontinuity in optimization and nonzero sum games*, J. Optim. Theory Appl., 120 (2004), pp. 181–197.

[8] J. Morgan and V. Scalzo, *New results on value functions and applications to MaxSup and MaxInf problems*, J. Math. Anal. Appl., 300 (2004), pp. 68–78.

[9] J. Morgan and V. Scalzo, *Asymptotical behavior of finite and possible discontinuous economies*, Econom. Theory, to appear.

[10] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, Heidelberg, 1998.

[11] G. Tian and J. Zhou, *Transfer continuities, generalizations of the Weierstrass and maximum theorems: A full characterization*, J. Math. Econom., 24 (1995), pp. 281–303.

[12] A. N. Tikhonov, *On the stability of the functional optimization problem*, U.S.S.R. Comput. Math. and Math. Phys., 6 (1966), pp. 26–33.

[13] T. Zolezzi, *Well-posedness criteria in optimization with application to the calculus of variations*, Nonlinear Anal., 25 (1995), pp. 437–453.

SIAM J. OPTIM.
Vol. 17, No. 3, pp. 871–890
© 2006 Society for Industrial and Applied Mathematics

CORRECTOR-PREDICTOR METHODS FOR SUFFICIENT LINEAR COMPLEMENTARITY PROBLEMS IN A WIDE NEIGHBORHOOD OF THE CENTRAL PATH*

XING LIU† AND FLORIAN A. POTRA†

Abstract. A higher order corrector-predictor interior-point method is proposed for solving sufficient linear complementarity problems. The algorithm produces a sequence of iterates in the $\mathcal{N}_\infty^-$ neighborhood of the central path. The algorithm does not depend on the handicap κ of the problem. It has $O((1+\kappa)\sqrt{n}L)$ iteration complexity and is superlinearly convergent even for degenerate problems.

Key words. linear complementarity, interior-point, path-following, corrector-predictor, wide neighborhood

AMS subject classifications. 90C51, 90C33

DOI. 10.1137/050623723

1. Introduction.

The Mizuno–Todd–Ye (MTY) predictor-corrector algorithm proposed by Mizuno, Todd, and Ye [9] is a typical representative of a large class of MTY-type predictor-corrector methods, which play a very important role among primal-dual interior-point methods. It was the first algorithm for linear programming that had both polynomial complexity and superlinear convergence. This result was extended to monotone linear complementarity problems that are nondegenerate, in the sense that they have a strictly complementarity solution [6, 23]. It turned out that the nondegeneracy assumption is not restrictive, since according to [10] a large class of interior-point methods, which contains MTY, can have only linear convergence if this assumption is violated. However, it is possible to obtain arbitrarily high order of convergence for degenerate problems by using higher order information of the central path [19, 21].

The existence of a central path is crucial for interior-point methods. An important result of the 1991 monograph of Kojima et al. [7] shows that the central path exists for any P_* linear complementarity problem, provided that the relative interior of its feasible set is nonempty. We recall that every P_* linear complementarity problem is a $P_*(\kappa)$ problem for some $\kappa \geq 0$, i.e.,

$$P_* = \bigcup_{\kappa \geq 0} P_*(\kappa).$$

The class of sufficient matrices was introduced by Cottle, Pang, and Stone [3] in connection with the linear complementarity problems. A matrix $M \in R^{n\times n}$ is said to be column sufficient if

$$z_i(Mz)_i \leq 0 \quad \forall i \quad \text{implies} \quad z_i(Mz)_i = 0 \quad \forall i.$$

The matrix M is called row sufficient if its transpose is column sufficient. The matrix M is sufficient if it is both column and row sufficient. It is proved in the same book

*Received by the editors February 1, 2005; accepted for publication (in revised form) April 28, 2006; published electronically October 20, 2006. This work was supported in part by National Science Foundation grant 0139701.

http://www.siam.org/journals/siopt/17-3/62372.html

†Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD 21250 (liu2@math.umbc.edu, potra@math.umbc.edu).

that the class of sufficient matrices is closely related to the existence of the solution of the linear complementarity problems and the convexity of the solution set. A surprising result given by Väliaho [22] showed that the class of P_* matrices coincides with the class of sufficient matrices. Therefore, every P_* linear complementarity problem is a sufficient linear complementarity problem, and vice versa. The class of sufficient linear complementarity problems is a very general framework for studying interior-point methods. In 1995 Miao [8] extended the MTY predictor-corrector algorithm for $P_*(\kappa)$ linear complementarity problems. His algorithm has $O((1+\kappa)\sqrt{n}L)$ iteration complexity and is quadratically convergent for nondegenerate problems. However, the constant κ is explicitly used in the construction of the algorithm, which implies that the algorithm cannot be used for sufficient linear complementarity problems. Potra and Sheng [17] extended the MTY predictor-corrector algorithm further for sufficient complementarity problems. While the algorithms of [17] do not depend on the constant κ, their computational complexity does: if the problem is a $P_*(\kappa)$ linear complementarity problem, they terminate in at most $O((1+\kappa)\sqrt{n}L)$ iterations. Moreover, the algorithms may attain arbitrarily high orders of convergence on nondegenerate problems. Predictor-corrector algorithms with arbitrarily high order of convergence for degenerate sufficient linear complementarity problems were given in [19]. The algorithms, as shown in [18], have $O((1+\kappa)\sqrt{n}L)$ iteration complexity for $P_*(\kappa)$ linear complementarity problems.

All the above algorithms operate in l_2 neighborhoods, also known as the small neighborhoods, of the central path. It is well known, however, that primal-dual interior-point methods have better practical performances in wide neighborhoods of the central path. Unfortunately, the iteration complexity of the predictor-corrector methods that use wide neighborhoods is worse than the complexity of the corresponding methods for small neighborhoods. Moreover, as shown in [2, 4], it is more difficult to develop and analyze predictor-corrector methods in wide neighborhoods. The best iteration complexity achieved by any known interior-point method for monotone linear complementarity problems in the wide neighborhoods using first order information is $O(nL)$. By using a large neighborhood defined by a suitable self-regular proximity measure, Peng, Terlaky, and Zhao [12] have obtained a predictor-corrector method with $O(\log n\sqrt{n}L)$ iteration complexity which is superlinearly convergent on nondegenerate problems. It turns out that the complexity result can be improved by using higher order information. The algorithms described in [11, 5, 24] have $O(\sqrt{n}L)$ iteration complexity. However, these algorithms are not of a predictor-corrector type, and they are not superlinearly convergent. The algorithm described in [20] operates in the δ_∞^- neighborhood and is superlinear convergent for sufficient linear complementarity problems, but no complexity results have been proved for this algorithm. A predictor-corrector method for monotone linear complementarity problems using wide neighborhoods of the central path was proposed in [14]. The algorithm has $O(\sqrt{n}L)$ iteration complexity by using a higher order predictor, and it is superlinear convergent even for degenerate problems. In a recent paper, Potra and Liu [16] extended the algorithm in [14] to sufficient linear complementarity problems. Two algorithms are analyzed in [16]. Both algorithms are of predictor-corrector type acting in between two wide neighborhoods of the central path. The radii of those neighborhoods have to satisfy an inequality that depends on the handicap κ of the problems. The first algorithm in [16] depends also on κ, while the second does not. The second algorithm uses the first algorithm by assigning $\kappa = 1$ and then doubles κ until a certain criterion is satisfied. Both algorithms have

$O((1+\kappa)^{1+1/m}\sqrt{n}L)$ iteration complexity and are superlinearly convergent even for degenerate problems.

The traditional predictor-corrector algorithms operate between two neighborhoods of the central path. The predictor step aims to decrease the duality gap while keeping the point in the outer neighborhood. It is followed by a corrector step, which brings the point back into the inner neighborhood so that the next predictor-corrector iteration can be performed. As analyzed in a recent paper [15], the centering direction is not as efficient in the wide neighborhoods as in the small neighborhoods, so that a line search on the centering direction is always needed in the corrector step using wide neighborhoods. Moreover, since the pure centering direction is anyhow inefficient in the wide neighborhoods, a corrector-predictor method was proposed in [15], where the corrector is used to improve both optimality and centrality. In the present paper, we generalize this algorithm to sufficient linear complementarity problems. By using higher order information, the algorithm has $O((1+\kappa)\sqrt{n}L)$ iteration complexity, which matches the best iteration complexity obtained in the small neighborhoods. Moreover, our algorithm is superlinearly convergent even for degenerate problems. More precisely, by using a predictor with order $m_p > 1$, we show that the duality gap converges to zero with Q-order $m_p + 1$ in the nondegenerate case and with Q-order $(m_p + 1)/2$ in the degenerate case. Our algorithm improves considerably the results of [16]. First, the algorithm is a corrector-predictor interior-point method so that it uses only one wide neighborhood of the central path, whose radius can be any number between 0 and 1, and therefore does not depend on κ. Second, its iteration complexity is improved ($O((1+\kappa)\sqrt{n}L)$ versus $O((1+\kappa)^{1+1/m}\sqrt{n}L)$). Finally, by contrast with the algorithms of [16] the present algorithm reduces the duality gap both in the corrector and the predictor steps, and therefore it is more efficient. In the present paper we work on horizontal linear complementarity problems (HLCP), which is a slight generalization of the standard linear complementarity problem. Equivalence results of different variants of linear complementarity problems can be found in [1]. We choose to work on HLCP because of its symmetry.

Conventions. We denote by $\mathbb{N}$ the set of all nonnegative integers. $\mathbb{R}$, $\mathbb{R}_+$, and $\mathbb{R}_{++}$ denote the set of real, nonnegative real, and positive real numbers, respectively. For any real number κ, $\lceil\, \kappa \,\rceil$ denotes the smallest integer greater than or equal to κ. Given a vector x, the corresponding uppercase symbol denotes, as usual, the diagonal matrix X defined by the vector. The symbol e represents the vector of all ones, with dimension given by the context.

We denote componentwise operations on vectors by the usual notations for real numbers. Thus, given two vectors u, v of the same dimension, uv, u/v, etc. will denote the vectors with components $u_i v_i$, u_i/v_i, etc. This notation is consistent as long as componentwise operations always have precedence in relation to matrix operations. Note that $uv \equiv Uv$ and if A is a matrix, then $Auv \equiv AUv$, but in general $Auv \neq (Au)v$. Also if f is a scalar function and v is a vector, then $f(v)$ denotes the vector with components $f(v_i)$. For example if $v \in \mathbb{R}^n_+$, then $\sqrt{v}$ denotes the vector with components $\sqrt{v_i}$, and $1 - v$ denotes the vector with components $1 - v_i$. Traditionally the vector $1 - v$ is written as $e - v$, where e is the vector of all ones. Inequalities are to be understood in a similar fashion. For example if $v \in \mathbb{R}^n$, then $v \geq 3$ means that $v_i \geq 3$, $i = 1, \ldots, n$. Traditionally this is written as $v \geq 3\, e$. If $\|\,.\,\|$ is a vector norm on $\mathbb{R}^n$ and A is a matrix, then the operator norm induced by $\|\,.\,\|$ is defined by $\|\,A\,\| = \max\{\|\,Ax\,\| \;;\; \|\,x\,\| = 1\}$. As a particular case we note that if U is the diagonal matrix defined by the vector u, then $\|\,U\,\|_2 = \|\,u\,\|_\infty$.

We use the notations $O(\cdot)$, $\Omega(\cdot)$, $\Theta(\cdot)$, and $o(\cdot)$ in the standard way to express asymptotic relationships between functions. The most common usage will be associated with a sequence $\{x^k\}$ of vectors and a sequence $\{\tau_k\}$ of positive real numbers. In this case $x^k = O(\tau_k)$ means that there is a constant K (dependent on problem data) such that for every $k \in \mathbb{N}$, $\left\| x^k \right\| \leq K\tau_k$. Similarly, if $x^k > 0$, $x^k = \Omega(\tau_k)$ means that $(x^k)^{-1} = O(1/\tau_k)$. If we have both $x^k = O(\tau_k)$ and $x^k = \Omega(\tau_k)$, we write $x^k = \Theta(\tau_k)$.

If $x, s \in \mathbb{R}^n$, then the vector $z \in \mathbb{R}^{2n}$ obtained by concatenating x and s will be denoted by $\lceil\, x,\, s \,\rfloor$, i.e.,

$$z = \lceil\, x,\, s \,\rfloor = \begin{bmatrix} x \\ s \end{bmatrix} = \left[\, x^T,\, s^T \,\right]^T. \tag{1.1}$$

Throughout this paper the mean value of xs will be denoted by

$$\mu(z) = \frac{x^T s}{n}. \tag{1.2}$$

2. The $P_*(\kappa)$ horizontal linear complementarity problem. Given two matrices $Q, R \in \mathbb{R}^{n\times n}$ and a vector $b \in \mathbb{R}^n$, the horizontal linear complementarity problem (HLCP) consists in finding a pair of vectors $z = \lceil\, x,\, s \,\rfloor$ such that

$$\begin{aligned} xs &= 0, \\ Qx + Rs &= b, \\ x, s &\geq 0. \end{aligned} \tag{2.1}$$

The standard (monotone) linear complementarity problem (SLCP or simply LCP) is obtained by taking $R = -I$, and Q positive semidefinite. Let $\kappa \geq 0$ be a given constant. We say that (2.1) is a $P_*(\kappa)$ HLCP if

$$Qu + Rv = 0 \text{ implies } (1+4\kappa)\sum_{i\in\mathcal{I}^+} u_i v_i + \sum_{i\in\mathcal{I}^-} u_i v_i \geq 0 \quad \text{for any } u, v \in \mathbb{R}^n,$$

where $\mathcal{I}^+ = \{i : u_i v_i > 0\}$ and $\mathcal{I}^- = \{i : u_i v_i < 0\}$. If the above condition is satisfied, we say that (Q, R) is a $P_*(\kappa)$ pair and write $(Q, R) \in P_*(\kappa)$. In case $R = -I$, $(Q, -I)$ is a $P_*(\kappa)$ pair if and only if Q is a $P_*(\kappa)$ matrix in the sense that

$$(1+4\kappa)\sum_{i\in\hat{\mathcal{I}}^+} x_i[Qx]_i + \sum_{i\in\hat{\mathcal{I}}^-} x_i[Qx]_i \geq 0 \quad \forall x \in \mathbb{R}^n,$$

where $\hat{\mathcal{I}}^+ = \{i : x_i[Qx]_i > 0\}$ and $\hat{\mathcal{I}}^- = \{i : x_i[Qx]_i < 0\}$. Problem (2.1) is then called a $P_*(\kappa)$ LCP and is extensively discussed in [7]. If (Q, R) belongs to the class

$$P_* = \bigcup_{\kappa\geq 0} P_*(\kappa),$$

then we say that (Q, R) is a P_* pair and (2.1) is a P_* HLCP.

The class of sufficient matrices was defined by Cottle, Pang, and Stone [3]. The appropriate generalization to sufficient pair [18, 19] is in terms of the null space of the matrix $[Q\ R] \in \mathbb{R}^{n\times 2n}$

$$\Phi := \mathcal{N}([Q\ R]) = \{\lceil\, u,\, v \,\rfloor | Qu + Rv = 0\} \tag{2.2}$$

and its orthogonal space

$$\Phi^{\perp} = \left\{ \lceil\, u,\, v \,\rfloor | u = Q^T x,\, v = R^T x \ \text{ for some } x \in \mathbb{R}^n \right\}. \tag{2.3}$$

(Q, R) is called column sufficient if

$$\lceil\, u,\, v \,\rfloor \in \Phi, \quad uv \leq 0 \text{ implies } uv = 0,$$

and row sufficient if

$$\lceil\, u,\, v \,\rfloor \in \Phi^{\perp}, \quad uv \geq 0 \text{ implies } uv = 0.$$

(Q, R) is a sufficient pair if it is both column and row sufficient. The corresponding results of row and column sufficient matrices in [3] can be extended to row and column sufficient pairs (see, for example, [20]): (Q, R) is a sufficient pair if and only if for any b, the HLCP (2.1) has a convex (perhaps empty) solution set and every KKT point of

$$\begin{array}{cc} \min_{x,s}, & x^T s \\ \text{s.t.} & Qx + Rs = b, \\ & x, s \geq 0, \end{array}$$

is a solution of (2.1).

Väliaho's result [22] states that a matrix is sufficient if and only if it is a $P_*(\kappa)$ matrix for some $\kappa \geq 0$. The result can be extended to sufficient pairs by using the equivalence results from [1] (see also [20]): (Q, R) is a sufficient pair if and only if there is a finite $\kappa \geq 0$ so that (Q, R) is a $P_*(\kappa)$ pair. By extension, a P_* HLCP will be called a sufficient HLCP and a P_* pair will be called a sufficient pair.

Let us note that if (Q, R) is a sufficient pair, then the matrix $[Q\ R]$ is full rank. In fact, we have the following slightly stronger result.

THEOREM 2.1. *Given two matrices $Q, R \in \mathbb{R}^{n\times n}$, if the pair (Q, R) is column sufficient, the matrix $[Q\ R]$ is full rank.*

Proof. Let r be the rank of Q, and the LU factorization of Q can be written as

$$PQ = L \begin{bmatrix} F_1 & F_2 \\ 0 & 0 \end{bmatrix},$$

where P is a permutation matrix, $L \in \mathbb{R}^{n\times n}$ is an invertible lower triangular matrix, $F_1 \in \mathbb{R}^{n-r\times n-r}$ is an invertible upper triangular matrix, and F_2 and the zeros are matrices with the correct dimensions.

Let us denote by G

$$G = L^{-1}PR = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix},$$

where $G_{11} \in \mathbb{R}^{n-r\times n-r}$, and G_{12}, G_{21}, and G_{22} are with the correct dimensions.

Since permutation matrices are invertible, and L is invertible, we have

$$\operatorname{rank}([Q\ R]) = \operatorname{rank}(L^{-1}P[Q\ R]) = \operatorname{rank}\left(\begin{bmatrix} F_1 & F_2 & G_{11} & G_{12} \\ 0 & 0 & G_{21} & G_{22} \end{bmatrix} \right).$$

We denote by u_1 and u_2 the components of u in the first r and last $n - r$ indices, respectively, and similarly for v. Therefore, $\lceil\, u,\, v \,\rfloor \in \Phi$ is equivalent to

$$\begin{cases} F_1 u_1 + F_2 u_2 + G_{11} v_1 + G_{12} v_2 = 0, \\ \qquad\qquad\qquad\quad G_{21} v_1 + G_{22} v_2 = 0. \end{cases}$$

For any $v_2 \in \mathrm{Ker}(G_{22})$, we construct a pair of vectors u and v such that

$$v_1 = 0, \quad u_2 = -v_2, \quad u_1 = -F_1^{-1}(F_2 u_2 + G_{12} v_2).$$

Clearly, we have $\lceil\, u,\, v \,\rfloor \in \Phi$. Moreover, we also obtain

$$u_1 v_1 \le 0 \quad \text{and} \quad u_2 v_2 = -v_2^2 \le 0.$$

Because (Q, R) is column sufficient, we have

$$u_1 v_1 = 0 \quad \text{and} \quad u_2 v_2 = 0.$$

We thus have that $v_2 \in \mathrm{Ker}(G_{22})$ implies $v_2 = 0$. Therefore G_{22} is invertible, and

$$\mathrm{rank}([Q\ R]) = \mathrm{rank}\left(\begin{bmatrix} F_1 & F_2 & G_{11} & G_{12} \\ 0 & 0 & G_{21} & G_{22} \end{bmatrix}\right) = n. \qquad \square$$

It is interesting to remark that row sufficiency alone does not imply the full rank property. For example, take

$$Q = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad R = \begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix};$$

it is easily seen that (Q, R) is row sufficient, but $\mathrm{rank}([Q\ R]) = 1$. We also note that in [19, 18, 20], the full rank property was given as an assumption, which in fact always hold because of the above theorem.

We denote the set of all feasible points of HLCP by

$$\mathcal{F} = \{z = \lceil\, x,\, s \,\rfloor \in \mathbb{R}^{2n}_{+} \ : \quad Qx + Rs = b\}$$

and its solution set by

$$\mathcal{F}^* = \{z^* = \lceil\, x^*,\, s^* \,\rfloor \in \mathcal{F} \ : \quad x^* s^* = 0\}.$$

The relative interior of $\mathcal{F}$, which is also known as the set of strictly feasible points or the set of interior points, is given by

$$\mathcal{F}^0 = \mathcal{F} \bigcap \mathbb{R}^{2n}_{++}.$$

It is known (see, for example, [7]) that if $\mathcal{F}^0$ is nonempty, then the nonlinear system

$$\begin{aligned} xs &= \tau e, \\ Qx + Rs &= b \end{aligned}$$

has a unique positive solution for any $\tau > 0$. The set of all such solutions defines the central path $\mathcal{C}$ of the HLCP, that is,

$$\mathcal{C} = \{z \in \mathbb{R}^{2n}_{++} \ : \ F_\tau(z) = 0,\, \tau > 0\},$$

where

$$F_\tau(z) = \begin{bmatrix} xs - \tau e \\ Qx + Rs - b \end{bmatrix}.$$

If $F_\tau(z) = 0$, then it is easy to see that $\tau = \mu(z)$, where $\mu(z)$ is given by (1.2). The wide neighborhood $\mathcal{N}_\infty^-(\alpha)$ is defined as

$$\mathcal{N}_\infty^-(\alpha) = \{z \in \mathcal{F}^0 \ : \ \delta_\infty^-(z) \le \alpha\},$$

where $0 < \alpha < 1$ is a given parameter and

$$\delta_\infty^-(z) := \left\| \left[\frac{xs}{\mu(z)} - e \right]^- \right\|_\infty$$

is a proximity measure of z to the central path. Alternatively, if we denote

$$\mathcal{D}(\beta) = \{z \in \mathcal{F}^0 : xs \ge \beta\mu(z)\},$$

then the neighborhood $\mathcal{N}_\infty^-(\alpha)$ can also be written as

$$\mathcal{N}_\infty^-(\alpha) = \mathcal{D}(1-\alpha).$$

It is well known (see, for example, the proof in [15]) that

$$\lim_{\alpha \downarrow 0} \mathcal{N}_\infty^-(\alpha) = \lim_{\beta \uparrow 1} \mathcal{D}(\beta) = \mathcal{C}, \quad \lim_{\alpha \uparrow 1} \mathcal{N}_\infty^-(\alpha) = \lim_{\beta \downarrow 0} \mathcal{D}(\beta) = \mathcal{F}.$$

3. A higher order corrector-predictor algorithm. The higher order corrector and predictor use higher derivatives of the central path. Given a point $z = \lceil x, s \rfloor \in \mathcal{D}(\beta)$, we consider the curve given by an mth order vector valued polynomial of the form

$$z(\theta) = z + \sum_{i=1}^{m} w^i \theta^i, \tag{3.1}$$

where the vectors $w^i = \lceil u^i, v^i \rfloor$ are obtained as solutions of the following linear systems:

$$\begin{aligned}
&\begin{cases} su^1 + xv^1 = \gamma\mu e - (1+\epsilon)xs, \\ Qu^1 + Rv^1 = 0, \end{cases} \\
&\begin{cases} su^2 + xv^2 = \epsilon xs - u^1 v^1, \\ Qu^2 + Rv^2 = 0, \end{cases} \\
&\begin{cases} su^i + xv^i = -\sum_{j=1}^{i-1} u^j v^{i-j}, \\ Qu^i + Rv^i = 0, \end{cases} \qquad i = 3, \ldots, m.
\end{aligned} \tag{3.2}$$

In a corrector step we choose $\epsilon = 0$ and $\gamma \in [\underline{\gamma}, \overline{\gamma}]$, where $0 < \underline{\gamma} < \overline{\gamma} < 1$ are given parameters, while in a predictor step we take

$$\gamma = 0 \quad \text{and} \quad \epsilon = \begin{cases} 0 & \text{if HLCP is nondegenerate,} \\ 1 & \text{if HLCP is degenerate.} \end{cases} \tag{3.3}$$

We note that in the corrector step, where we have $\epsilon = 0$, w^1 is the affine scaling direction if $\gamma = 0$ and the classical centering direction if $\gamma = 1$. In system (3.2), w^1 is a convex combination of the affine scaling and the centering directions. The directions w^i are related to the higher derivatives of the central path [19]. We note that the central path passing through z is analytic in μ when HLCP is nondegenerate and in $\sqrt{\mu}$ when HLCP is degenerate.

As the m linear systems in (3.2) have the same left-hand matrix, only one matrix factorization and m backsolves are needed. Therefore it involves $O(n^3) + O(m\,n^2)$ arithmetic operations. We take $m = m_c$ in the corrector step, and $m = m_p$ in the predictor step. From (3.1) and (3.2) it follows that

$$x(\theta)s(\theta) = (1-\theta)^{1+\epsilon}xs + \gamma\theta\mu e + \sum_{i=m+1}^{2m} \theta^i h^i,$$

$$\mu(\theta) = (1-\theta)^{1+\epsilon}\mu + \gamma\theta\mu + \sum_{i=m+1}^{2m} \theta^i (e^T h^i/n),$$

$$\text{where } h^i = \sum_{j=i-m}^{m} u^j v^{i-j}. \tag{3.4}$$

In the development of our algorithm, we want to preserve positivity of each iterated point. We thus give an upper bound θ_0 for the step-length taken both in the predictor and the corrector step:

$$\theta_0 = \sup\{\hat{\theta}_0 \ :\ x(\theta) > 0,\ s(\theta) > 0 \quad \forall \theta \in [0\,, \hat{\theta}_0]\}. \tag{3.5}$$

We introduce the following notation, which will be used in describing both the corrector and predictor steps:

$$p(\theta) = \frac{x(\theta)s(\theta)}{\mu(\theta)}, \quad f(\theta) = \min_{i=1,\ldots,n} p_i(\theta). \tag{3.6}$$

The corrector. The corrector step is obtained by taking $\epsilon = 0$, and $0 < \gamma < 1$ in (3.1)–(3.2). The main purpose of the corrector step is to increase proximity to the central path. However, we also improve the normalized complementarity gap $\mu(\theta)$ at the same time. We choose $\sigma \in [\underline{\sigma}, \overline{\sigma}]$, where $0 < \underline{\sigma} < \overline{\sigma} < 1$ are given parameters, and define

$$\theta_1 = \sup\{\hat{\theta}_1 \ :\ 0 \le \hat{\theta}_1 \le \theta_0,\ \mu(\theta) \le (1 - \sigma(1-\gamma)\theta)\mu \quad \forall \theta \in [0, \hat{\theta}_1]\}. \tag{3.7}$$

The step-length of the corrector is obtained as

$$\theta_c = \operatorname{argmax}\{f(\theta) \ :\ \theta \in [\,0, \theta_1]\,\}. \tag{3.8}$$

As a result of the corrector step we obtain the point

$$\overline{z} = \lceil\, \overline{x}\,,\, \overline{s}\,\rfloor := z(\theta_c). \tag{3.9}$$

We have clearly $\overline{z} \in \mathcal{D}(\beta_c)$ with $\beta_c > \beta$. While the parameter β is fixed during the algorithm, the positive quantity β_c varies from iteration to iteration. However, we will prove that there is a constant $\beta_c^* > \beta$, such that $\beta_c > \beta_c^*$ in all iterations.

The predictor. The predictor is obtained by taking $z = \overline{z}$, where $\overline{z}$ is the result of the corrector step, and $\gamma = 0$ in (3.1)–(3.2). The aim of the predictor step is to decrease the normalized complementarity gap as much as possible while keeping the iterate in $\mathcal{D}(\beta)$. We define the predictor step-length as

$$\theta_p = \operatorname{argmin}\{\mu(\theta) \ :\ \theta \in [\,0, \theta_2\,]\,\}, \tag{3.10}$$

where

$$\theta_2 = \max\{\hat{\theta}_2 \ : \ z(\theta) \in \mathcal{D}(\beta) \quad \forall \theta \in [0, \hat{\theta}_2]\}. \tag{3.11}$$

A standard continuity argument can be used to show that $z(\theta) > 0 \ \forall \theta \in [0, \theta_2]$. As a result of the predictor step, we obtain a point

$$z^+ = \lceil x^+, s^+ \rfloor := z(\theta_p). \tag{3.12}$$

By construction we have $z^+ \in \mathcal{D}(\beta)$, so that a new corrector step can be applied. Summing up we can formulate the following iterative procedure.

ALGORITHM 1.

Given real parameters $0 < \beta < 1$, $0 < \underline{\gamma} < \overline{\gamma} < 1$, $0 < \underline{\sigma} < \overline{\sigma} < 1$, integers m_c, $m_p \geq 1$, and a vector $z^0 \in \mathcal{D}(\beta)$:

Set $k \leftarrow 0$;
repeat
(corrector step)
Set $z \leftarrow z^k$;
Choose $\gamma \in [\underline{\gamma}, \overline{\gamma}]$ and set $m = m_c$;
Compute directions $w^i = \lceil u^i, v^i \rfloor$, $i = 1, \ldots, m$, by solving (3.2);
Compute θ_0 from (3.5)
If HLCP is skew-symmetric, set $\sigma = 1$ and $\theta_1 = \theta_0$;
Else, choose $\sigma \in [\underline{\sigma}, \overline{\sigma}]$, and compute θ_1 from (3.7);
Compute corrector step-length θ_c from (3.8);
Compute $\overline{z}$ from (3.9);
Set $\overline{z}^k \leftarrow \overline{z}$, $\overline{\mu}_k \leftarrow \overline{\mu} = \mu(\overline{z})$.
(predictor step)
Set $z \leftarrow \overline{z}^k$, $\gamma = 0$, and $m = m_p$;
Compute directions $w^i = \lceil u^i, v^i \rfloor$, $i = 1, \ldots, m$, by solving (3.2);
Compute θ_p from (3.10);
Compute z^+ from (3.12);
Set $z^{k+1} \leftarrow z^+$, $\mu_{k+1} \leftarrow \mu^+ = \mu(z^+)$, $k \leftarrow k + 1$.
continue

The computation of the exact values of θ_c and θ_p is quite involved, so that in practice good estimates of θ_c and θ_p are obtained by appropriate line search procedures. In particular, by adopting the line search procedure from [15] we can preserve both the computational complexity and the superlinear convergence of the theoretical algorithm. In fact the convergence properties can be proved by using the explicit lower bounds in the next section.

4. Polynomial complexity. We analyze in this section the computational complexity of Algorithm 1. In the proof of the complexity results, we will use the following lemmas, which were proved in [16].

LEMMA 4.1. *Assume that HLCP* (2.1) *is* $P_*(\kappa)$, *let* $w = \lceil u, v \rfloor$ *be the solution of the linear system*

$$\begin{aligned} su + xv &= a, \\ Qu + Rv &= 0, \end{aligned}$$

where $z = \lceil x, s \rfloor \in \mathbb{R}^{2n}_{++}$ *and* $a \in \mathbb{R}^n$ *are given vectors, and consider the index sets:*

$$\mathcal{I}^+ = \{i : u_i v_i > 0\}, \quad \mathcal{I}^- = \{i : u_i v_i < 0\}.$$

Then the following inequalities are satisfied:

$$\frac{1}{1+4\kappa}\,\|\,u\,v\,\|_\infty \le \sum_{i\in\mathcal{I}_+} u_i v_i \le \frac{1}{4}\left\|\,(xs)^{-1/2}\,a\,\right\|_2^2.$$

LEMMA 4.2. *Assume that HLCP* (2.1) *is* $P_*(\kappa)$*, and let* $w = \lceil\, u,\, v \,\rfloor$ *be the solution of the linear system*

$$\begin{aligned} su + xv &= a,\\ Qu + Rv &= 0, \end{aligned}$$

where $z = \lceil\, x,\, s\,\rfloor \in \mathbb{R}^{2n}_{++}$ *and* $a \in \mathbb{R}^n$ *are given vectors. Then the following inequality holds:*

$$u^T v \ge -\kappa \left\|\,(xs)^{-1/2}\,a\,\right\|_2^2. \tag{4.1}$$

Let us denote

$$\eta_i = \left\|\,Du^i + D^{-1}v^i\,\right\|_2, \text{ where } D = X^{-1/2}S^{1/2}. \tag{4.2}$$

The following lemma is a slight improvement over the corresponding results in [16] and a generalization to sufficient HLCP of the corresponding results in [15].

LEMMA 4.3. *If HLCP* (2.1) *is sufficient and* $z = \lceil\, x,\, s\,\rfloor \in \mathcal{D}(\beta)$*, then for* $n \ge 8$*, the solution of* (3.2) *satisfies*

$$\frac{1}{\sqrt{1+2\kappa}}\sqrt{\|\,Du^i\,\|_2^2 + \|\,D^{-1}v^i\,\|_2^2} \le \eta_i \le \frac{2}{1+2\kappa}\alpha_i\sqrt{\beta\mu}\left(\frac{(1+2\kappa)\tau}{4}\sqrt{n}\right)^i, \tag{4.3}$$

where

$$\tau = \frac{2\sqrt{\beta(1+\epsilon-\gamma)^2 + (1-\beta)\gamma^2}}{\beta}, \tag{4.4}$$

and the sequence

$$\alpha_i = \frac{1}{i}\binom{2i-2}{i-1} \le \frac{1}{i}4^i$$

is the solution of the following recurrence scheme:

$$\alpha_1 = 1, \quad \alpha_i = \sum_{j=1}^{i-1}\alpha_j\,\alpha_{i-j}.$$

Proof. The first part of the inequality follows immediately, since by using (3.2) and Lemma 4.2 we have

$$\left\|\,Du^i + D^{-1}v^i\,\right\|_2^2 = \left\|\,Du^i\,\right\|_2^2 + 2u^{i\,T}v^i + \left\|\,D^{-1}v^i\,\right\|_2^2$$

$$\ge \left\|\,Du^i\,\right\|_2^2 + \left\|\,D^{-1}v^i\,\right\|_2^2 - 2\kappa\left\|\,Du^i + D^{-1}v^i\,\right\|_2^2.$$

By multiplying the first equations of (3.2) with $(xs)^{-1/2}$ we obtain

$$\begin{aligned} Du^1 + D^{-1}v^1 &= -\big((1+\epsilon)(xs)^{1/2} - \gamma\mu(xs)^{-1/2}\big), \\ Du^2 + D^{-1}v^2 &= -\big(\epsilon(xs)^{1/2} - (xs)^{-1/2}u^1v^1\big), \\ Du^i + D^{-1}v^i &= -(xs)^{-1/2}\sum_{j=1}^{i-1} Du^j D^{-1}v^{i-j}, \quad 3 \le i \le m. \end{aligned}$$

Because $z \in \mathcal{D}(\beta)$ we have $(xs)^{-1/2} \le (1/\sqrt{\beta\mu})e$, and we deduce that

$$\eta_1 = \|(1+\epsilon)(xs)^{1/2} - \gamma\mu(xs)^{-1/2}\|, \quad \eta_2 = \|\epsilon(xs)^{1/2} - (xs)^{-1/2}u^1v^1\|,$$

and

$$\eta_i \le \frac{1}{\sqrt{\beta\mu}}\sum_{j=1}^{i-1} \| Du^j \|_2 \, \| D^{-1}v^{i-j} \|_2, \quad 3 \le i \le m. \tag{4.5}$$

We have

$$\begin{aligned} \eta_1^2 = \|(1+\epsilon)(xs)^{1/2} - \gamma\mu(xs)^{-1/2}\|^2 &= \sum_{j=1}^{n}\left((1+\epsilon)^2 x_j s_j - 2(1+\epsilon)\gamma\mu + \frac{\gamma^2\mu^2}{x_j s_j}\right) \\ &= ((1+\epsilon)^2 - 2(1+\epsilon)\gamma)\mu n + \gamma^2\mu^2\sum_{j=1}^{n}\frac{1}{x_j s_j} \\ &\le \mu n\left((1+\epsilon)^2 - 2(1+\epsilon)\gamma + \frac{\gamma^2}{\beta}\right) = \frac{\beta\mu n\tau^2}{4}, \end{aligned}$$

which shows that the second inequality in (4.3) is satisfied for $i = 1$. We next show that the inequality also holds for $i = 2$; i.e., we want to prove that

$$\eta_2^2 \le \left(\frac{1+2\kappa}{8}\right)^2 \beta\mu n^2\tau^4 \; = \frac{1}{128}\beta\mu n^2\tau^4(2 + 8\kappa + 8\kappa^2).$$

Using Lemma 4.2, Corollary 2.3 of [13], and the fact $z \in \mathcal{D}(\beta)$, we have

$$\begin{aligned} \eta_2^2 = \|\epsilon(xs)^{1/2} - (xs)^{-1/2}u^1v^1\|^2 &= \sum_{j=1}^{n}\left(\epsilon^2 x_j s_j - 2\epsilon u_i^1 v_i^1 + \frac{(u_i^1 v_i^1)^2}{x_j s_j}\right) \\ \le \epsilon^2 n\mu + 2\epsilon\kappa\eta_1^2 + \frac{\eta_1^4}{8\beta\mu}(1 + 4\kappa + 8\kappa^2) &\le \epsilon^2 n\mu + \frac{\epsilon\kappa\beta\mu n\tau^2}{2} + \frac{\beta\mu n^2\tau^4}{128}(1 + 4\kappa + 8\kappa^2). \end{aligned}$$

Therefore, it remains to show that

$$\epsilon^2 n\mu + \frac{\epsilon\kappa\beta\mu n\tau^2}{2} \le \frac{\beta\mu n^2\tau^4}{128}(1 + 4\kappa),$$

which holds trivially for $\epsilon = 0$. The inequality holds for $\epsilon = 1$, provided

$$\beta n\tau^4 \ge 128, \quad n\tau^2 \ge 16.$$

Using the definition of τ (4.4), this reduces to

$$\frac{n(\beta(2-\gamma)^2 + (1-\beta)\gamma^2)^2}{\beta^3} \ge 8, \quad \frac{n(\beta(2-\gamma)^2 + (1-\beta)\gamma^2)}{\beta^2} \ge 4.$$

Since the minimum over $0 \le \beta, \gamma \le 1$ of both left-hand side functions is attained at $\beta = \gamma = 1$, we conclude that the second inequality in (4.3) is satisfied for $i = 2$ whenever $n \ge 8$. For $i \ge 3$ and $1 \le j < i$ we use the first inequality in (4.2) to obtain

$$\begin{aligned}
&\| Du^j \|_2 \, \| D^{-1}v^{i-j} \|_2 + \| Du^{i-j} \|_2 \, \| D^{-1}v^j \|_2 \\
&\quad \le \left(\| Du^j \|_2^2 + \| D^{-1}v^j \|_2^2 \right)^{1/2} \left(\| Du^{i-j} \|_2^2 + \| D^{-1}v^{i-j} \|_2^2 \right)^{1/2} \\
&\quad \le (1+2\kappa)\eta_j \, \eta_{i-j}.
\end{aligned}$$

From (4.5) it follows that

$$\eta_i \le \frac{1+2\kappa}{2\sqrt{\beta\mu}} \sum_{j=1}^{i-1} \eta_j \, \eta_{i-j}, \quad i = 2, \ldots, m.$$

The required inequalities are then easily proved by mathematical induction. ∎

By virtue of Lemma 4.3 we obtain the following bound for $\|h^i\|$.

LEMMA 4.4. *If HLCP* (2.1) *is sufficient and* $z = \lceil x, s \rfloor \in \mathcal{D}(\beta)$, *then for* $n \ge 8$, *the directions computed in* (3.4) *satisfy*

$$\zeta_i := \|h^i\| \le \frac{2\beta\mu}{(1+2\kappa)i} ((1+2\kappa)\tau\sqrt{n}\,)^i, \quad i = m+1, \ldots, 2m. \tag{4.6}$$

Proof. For any $m+1 \le i \le 2m$, we have

$$\begin{aligned}
\| h^i \|_2 &\le \sum_{j=i-m}^{m} \| Du^j \|_2 \, \| D^{-1}v^{i-j} \|_2 \le \sum_{j=1}^{i-1} \| Du^j \|_2 \, \| D^{-1}v^{i-j} \|_2 \\
&= \frac{1}{2} \sum_{j=1}^{i-1} \left(\| Du^j \|_2 \, \| D^{-1}v^{i-j} \|_2 + \| Du^{i-j} \|_2 \, \| D^{-1}v^j \|_2 \right) \\
&\le \frac{1}{2} \sum_{j=1}^{i-1} \sqrt{\| Du^j \|_2^2 + \| D^{-1}v^j \|_2^2} \sqrt{\| Du^{i-j} \|_2^2 + \| D^{-1}v^{i-j} \|_2^2} \\
&\le \frac{1+2\kappa}{2} \sum_{j=1}^{i-1} \eta_j \, \eta_{i-j} \le \frac{2\beta\mu}{1+2\kappa} \left(\frac{(1+2\kappa)\tau\sqrt{n}}{4} \right)^i \sum_{j=1}^{i-1} \alpha_j \, \alpha_{i-j} \\
&= \frac{2\beta\mu}{1+2\kappa} \left(\frac{(1+2\kappa)\tau\sqrt{n}}{4} \right)^i \alpha_i \le \frac{2\beta\mu}{(1+2\kappa)i} \left((1+2\kappa)\tau\sqrt{n} \right)^i,
\end{aligned}$$

where the last inequality follows from the fact that $\alpha_i \le \frac{1}{i} 4^i$. ∎

From the above lemmas we obtain the following result.

COROLLARY 4.5. *If HLCP* (2.1) *is sufficient and* $z = \lceil x, s \rfloor \in \mathcal{D}(\beta)$, *then the following relations hold for any* $\alpha > 0$, $\kappa \ge 0$, *and* $n \ge 8$*:*

$$\frac{\alpha}{\mu} \sum_{i=m+1}^{2m} \theta^i \| h^i \|_2 < 1 \quad \forall\, 0 \le \theta \le \frac{1}{(1+2\kappa)\tau\sqrt{n}} \min \left\{ 1, \left(\frac{1.4\,\alpha\beta}{1+2\kappa} \right)^{\frac{-1}{m+1}} \right\}, \tag{4.7}$$

$$\frac{\alpha}{\mu\sqrt{n}} \sum_{i=m+1}^{2m} \theta^i \| h^i \|_2 < \theta \quad \forall\, 0 \le \theta \le \frac{1}{(1+2\kappa)\tau\sqrt{n}} \min \left\{ 1, \left(1.4\,\alpha\beta\tau \right)^{\frac{-1}{m}} \right\}. \tag{4.8}$$

Proof. For any $t \in (0,1]$, we have

$$\sum_{i=m+1}^{2m} \frac{t^i}{i} \le t^{m+1} \sum_{i=m+1}^{2m} \frac{1}{i} < t^{m+1} \int_m^{2m} \frac{du}{u} = t^{m+1} \log 2 < .7\, t^{m+1}.$$

Using Lemma 4.4 and the above inequality, we obtain

$$\sum_{i=m+1}^{2m} \theta^i \left\| h^i \right\|_2 < \frac{1.4\beta\mu}{1+2\kappa}((1+2\kappa)\tau\sqrt{n}\theta)^{m+1} \quad \forall\, \theta \in \left(0,\ \frac{1}{(1+2\kappa)\tau\sqrt{n}}\right]. \tag{4.9}$$

Therefore,

$$\frac{\alpha}{\mu} \sum_{i=m+1}^{2m} \theta^i \left\| h^i \right\|_2 < \frac{1.4\alpha\beta}{1+2\kappa}((1+2\kappa)\tau\sqrt{n}\theta)^{m+1} \le 1$$
$$\forall\, \theta \in \left(0,\ \frac{1}{(1+2\kappa)\tau\sqrt{n}} \min\left\{1\,,\ \left(\frac{1.4\,\alpha\beta}{1+2\kappa}\right)^{\frac{-1}{m+1}}\right\}\right].$$

Equation (4.9) also implies that

$$\frac{\alpha}{\mu\sqrt{n}} \sum_{i=m+1}^{2m} \theta^i \left\| h^i \right\|_2 < \frac{1.4\alpha\beta}{(1+2\kappa)\sqrt{n}}((1+2\kappa)\tau\sqrt{n}\theta)^{m+1}$$
$$= 1.4\alpha\beta\tau\theta\ ((1+2\kappa)\tau\sqrt{n}\theta)^m \le \theta$$
$$\forall\, \theta \in \left(0,\ \frac{1}{(1+2\kappa)\tau\sqrt{n}} \min\left\{1\,,\ (1.4\,\alpha\beta\tau)^{\frac{-1}{m}}\right\}\right]. \qquad \square$$

From the definition of τ (4.4) it follows that

$$\frac{2(1+\epsilon)\sqrt{1-\beta}}{\sqrt{\beta}} \le \tau \le 2\max\left\{\frac{1+\epsilon}{\sqrt{\beta}}\,,\ \frac{\sqrt{1-\beta+\beta\epsilon^2}}{\beta}\right\} < \frac{2(1+\epsilon)}{\beta}. \tag{4.10}$$

In the corrector step we take $\epsilon = 0$; therefore we will use the bound $\tau < 2/\beta$ in the analysis below.

THEOREM 4.6. *If HCLP* (2.1) *is sufficient, then Algorithm* 1 *is well defined and the following relations hold for any* $\kappa \ge 0$ *and* $n \ge 8$*:*

$$\overline{z}^{\,k}, z^k \in \mathcal{D}(\beta),$$

$$\mu_{k+1} \le \left(1 - \frac{\chi}{(1+2\kappa)n^{\frac{1}{2}+\frac{m_c+1}{2m_c(m_p+1)}}}\right)\overline{\mu}_{\,k}, \quad k = 0, 1, \ldots,$$

$$\overline{\mu}_{\,k+1} \le \left(1 - \frac{\overline{\chi}}{(1+2\kappa)n^{\frac{1}{2}+\upsilon}}\right)\overline{\mu}_{\,k}, \quad k = 0, 1, \ldots,$$

where χ, $\overline{\chi}$ *are constants depending only on* $\beta, \underline{\gamma}, \overline{\gamma}\,, \underline{\sigma}, \overline{\sigma}\,$, *and*

$$\upsilon := \min\left\{\frac{1}{2m_c},\ \frac{m_c+1}{2m_c(m_p+1)}\right\}. \tag{4.11}$$

Proof.

Analysis of the corrector. On the corrector we have $m = m_c$, $\epsilon = 0$, $0 < \underline{\gamma} < \gamma < \overline{\gamma} < 1$, $0 < \underline{\sigma} < \gamma < \overline{\sigma} < 1$, and $\tau < 2/\beta$.

First, we prove that if $z \in \mathcal{D}(\beta)$, then the quantities θ_0 defined in (3.5) satisfy

$$\theta_0 \geq \theta_3 := \frac{\beta}{2(1+2\kappa)\sqrt{n}} \left(\frac{2.8}{1+2\kappa} \right)^{-\frac{1}{m_c+1}}. \tag{4.12}$$

This can be shown by using (4.7) with $\alpha = 2/\beta$ and the fact that $\theta_3 < 1/2$,

$$\frac{x(\theta)s(\theta)}{\mu} > (1-\theta)\frac{xs}{\mu} + \frac{1}{\mu}\sum_{i=m+1}^{2m} \theta^i h^i \geq \frac{\beta}{2}e - \frac{1}{\mu}\sum_{i=m+1}^{2m} \theta^i \left\| h^i \right\|_2 e > 0 \quad \forall \theta \in [0,\, \theta_3].$$

Since $x(0) > 0, s(0) > 0$, we can use a standard continuity argument to show that $x(\theta) > 0, s(\theta) > 0 \ \forall\, \theta \in [0, \theta_3]$, which proves that $\theta_0 \geq \theta_3$.

Next, we show that the quantities θ_1 defined in (3.7) satisfy

$$\theta_1 \geq \theta_4 := \frac{\beta}{2(1+2\kappa)\sqrt{n}} \left(\frac{(1-\overline{\sigma}\,)(1-\overline{\gamma}\,)}{2.8} \right)^{\frac{1}{m_c}}. \tag{4.13}$$

By using (4.8) with $\alpha = 1/\left((1-\overline{\sigma}\,)(1-\overline{\gamma}\,)\right)$, we deduce that the following inequalities hold for any $\theta \in [0,\, \theta_4]$:

$$\begin{aligned} \frac{\mu(\theta) - (1-\sigma(1-\gamma)\theta)\mu}{\mu} &= -(1-\sigma)(1-\gamma)\theta + \frac{1}{\mu\, n}\sum_{i=m+1}^{2m} \theta^i e^T h^i \\ &\leq -(1-\overline{\sigma}\,)(1-\overline{\gamma}\,)\theta + \frac{1}{\mu\sqrt{n}}\sum_{i=m+1}^{2m} \theta^i \left\| h^i \right\|_2 \leq 0, \end{aligned}$$

which shows that $\theta_1 \geq \theta_4$.

At last, we show that if $z \in \mathcal{D}(\beta)$, then

$$f(\theta) \geq \beta + \frac{1}{2}(1-\beta)\gamma\,\theta \geq \beta + \frac{1}{2}(1-\beta)\underline{\gamma}\,\theta \quad \forall\, \theta \in [0, \theta_5]\,, \tag{4.14}$$

where

$$\theta_5 := \min\left\{ \theta_4,\ \frac{\beta}{2(1+2\kappa)\, n^{\frac{1}{2}+\frac{1}{2m_c}}} \left(\frac{(1-\beta)\underline{\gamma}}{5.6} \right)^{\frac{1}{m_c}} \right\} \geq \frac{\chi_5}{(1+2\kappa)n^{\frac{1}{2}+\frac{1}{2m_c}}}, \tag{4.15}$$

$$\chi_5 := \frac{\beta}{2}\min\left\{ \frac{(1-\overline{\sigma}\,)(1-\overline{\gamma}\,)}{2.8}\,,\ \frac{(1-\beta)\underline{\gamma}}{5.6} \right\}. \tag{4.16}$$

It is easily seen that

$$\begin{aligned} p(\theta) = \frac{x(\theta)s(\theta)}{\mu(\theta)} &= \frac{(1-\theta)xs + \gamma\theta\mu e + \sum_{i=m+1}^{2m} \theta^i h^i}{(1-\theta)\mu + \gamma\theta\mu + \sum_{i=m+1}^{2m} \theta^i e^T h^i/n} \\ &\geq \frac{(1-\theta)\beta\mu e + \gamma\theta\mu e + \sum_{i=m+1}^{2m} \theta^i h^i}{(1-\theta)\mu + \gamma\theta\mu + \sum_{i=m+1}^{2m} \theta^i e^T h^i/n} \end{aligned} \tag{4.17}$$

$$= \beta e \frac{(1-\beta)\gamma\theta\mu e + \beta\sum_{i=m+1}^{2m}\theta^i(h^i - (e^Th^i/n)e) + (1-\beta)\sum_{i=m+1}^{2m}\theta^i h^i}{(1-\theta)\mu + \gamma\theta\mu + \sum_{i=m+1}^{2m}\theta^i e^T h^i/n}$$

$$\geq \beta e + \frac{(1-\beta)\gamma\theta\mu - \sum_{i=m+1}^{2m}\theta^i \left\| h^i \right\|_2}{\mu(\theta)}\, e$$

$$\geq \beta e + (1-\beta)\gamma\theta - \frac{1}{\mu}\sum_{i=m+1}^{2m}\theta^i \left\| h^i \right\|_2 e \quad \forall\, \theta \in [0,1].$$

The last inequality follows from the fact that

$$\mu(\theta) \leq (1-\sigma(1-\gamma)\theta)\mu \leq \mu \quad \forall\, \theta \in [0,\, \theta_4].$$

According to (4.8), with α replaced by $2\sqrt{n}/((1-\beta)\gamma)$ and τ replaced by $2/\beta$, we have

$$\frac{1}{\mu}\sum_{i=m+1}^{2m}\theta^i \left\| h^i \right\|_2 < \frac{1}{2}(1-\beta)\gamma\,\theta \quad \forall\theta \leq \frac{\beta}{2(1+2\kappa)\sqrt{n}}\left(\frac{(1-\beta)\gamma}{5.6\sqrt{n}}\right)^{\frac{1}{m}},$$

and (4.14) follows from the above inequality and (4.17).

Relation (4.14) shows that if $z \in \mathcal{D}(\beta)$, then the point $\overline{z}$ obtained in the corrector step of Algorithm 1 belongs to $\mathcal{D}(\beta+\delta)$, where

$$\delta = \frac{1}{2}(1-\beta)\gamma\,\theta_5. \tag{4.18}$$

As we mentioned before, the main purpose of the corrector is to increase proximity to the central path. However, it turns out that if the corrector step-length θ_c is large enough, then we also obtain a significant reduction of the duality gap during the corrector step. In what follows we find a lower bound for θ_c in case the point $z \in \mathcal{D}(\beta)$ is not very well centered. More precisely we show that

$$\exists j \quad \text{such that } p_j := \frac{x_j s_j}{\mu} \leq \beta + .44\,\delta \Rightarrow \theta_c > .2\,\theta_5\,. \tag{4.19}$$

Let us denote

$$\lambda = .44\,\delta = .22(1-\beta)\gamma\,\theta_5,\ q^i = \frac{h^i}{\mu}, \quad i = m+1, \ldots, 2m.$$

For any $\theta \in [0,\, 1]$, we have

$$p_j(\theta) = \frac{x_j(\theta)s_j(\theta)}{\mu(\theta)} = \frac{(1-\theta)p_j + \gamma\theta + \sum_{i=m+1}^{2m}\theta^i q_j^i}{(1-\theta) + \gamma\theta + \sum_{i=m+1}^{2m}\theta^i e^T q^i/n}$$

$$< \frac{(1-\theta)(\beta+\lambda) + \gamma\theta + \sum_{i=m+1}^{2m}\theta^i q_j^i}{(1-\theta) + \gamma\theta + \sum_{i=m+1}^{2m}\theta^i e^T q^i/n}$$

$$= \beta + \lambda + \frac{\gamma(1-\beta-\lambda)\theta - (\beta+\lambda)\sum_{i=m+1}^{2m}\theta^i e^T q^i/n + \sum_{i=m+1}^{2m}\theta^i q_j^i}{1-(1-\gamma)\theta + \sum_{i=m+1}^{2m}\theta^i e^T q^i/n}$$

$$\leq \beta + \lambda + \frac{\gamma(1-\beta-\lambda)\theta + (1+\frac{\beta+\lambda}{\sqrt{n}})\sum_{i=m+1}^{2m}\theta^i \left\| q^i \right\|_2}{1-(1-\gamma)\theta - \frac{1}{\sqrt{n}}\sum_{i=m+1}^{2m}\theta^i \left\| q^i \right\|_2}$$

$$\leq \beta + \lambda + \frac{\gamma(1-\beta)\theta + 2\sum_{i=m+1}^{2m}\theta^i \left\| q^i \right\|_2}{1-(1-\gamma)\theta - \sum_{i=m+1}^{2m}\theta^i \left\| q^i \right\|_2}.$$

Assume now that $\theta \in [0, .2\,\theta_5]$ and set $\theta = .2\phi$. Since $\phi \in [0,\, \theta_5]$, by virtue of (4.8), we can write

$$\sum_{i=m+1}^{2m} \theta^i \left\| q^i \right\|_2 = \sum_{i=m+1}^{2m} .2^i \phi^i \left\| q^i \right\|_2 \leq .2^{m+1} \sum_{i=m+1}^{2m} \phi^i \left\| q^i \right\|_2$$
$$< \frac{.2^{m+1}}{2} \gamma(1-\beta)\phi = \frac{.2^m}{2} \gamma(1-\beta)\theta \leq .1\gamma(1-\beta)\theta.$$

Using the fact that $\theta_5 < .5\ \forall\, n \geq 1$, we obtain

$$p_j(\theta) < \beta + \lambda + \frac{1.2\gamma(1-\beta)\theta}{1-(1-\gamma+.1\,\gamma(1-\beta))\,\theta} \leq \beta + \lambda + \frac{1.2\gamma(1-\beta)\theta}{1-\theta}$$
$$< \beta + \lambda + 1.4\gamma(1-\beta)\theta \leq\ \beta + \lambda + .28\gamma(1-\beta)\theta_5 = \beta + \delta \quad \forall\, \theta \in [0, .2\,\theta_5].$$

It follows that $f(\theta_c) \geq \beta + \delta > \max_{0 \leq \theta \leq .2\,\theta_5} f(\theta)$, wherefrom we deduce that $\theta_c > .2\,\theta_5$.

Analysis of the predictor. In the predictor step we have $\gamma = 0$ and $m = m_p$. From (4.4) it follows that $\tau = 2(1+\epsilon)/\sqrt{\beta} \leq 4/\sqrt{\beta}$. Since the predictor step follows a corrector step, we have $z \in \mathcal{D}(\beta + \delta) \subset \mathcal{D}(\beta)$.

First, we study the behavior of the normalized duality gap in the predictor step. We start by proving that

$$(1 - 2.5\,\theta)\mu \leq \mu(\theta) \leq (1 - .5\,\theta)\mu \tag{4.20}$$
$$\forall\, 0 \leq \theta \leq \theta_6 := \frac{\sqrt{\beta}}{4(1+2\kappa)\sqrt{n}} \min\left\{1\,,\ \left(11.2\sqrt{\beta}\right)^{\frac{-1}{m_p}}\right\}.$$

Due to the obvious fact that

$$(1-2\theta) \leq (1-\theta)^2 \leq (1-\theta) \quad \forall\, \theta \in [0,1],$$

we have

$$(1-2\theta)\mu + \sum_{i=m+1}^{2m} \theta^i (e^T h^i/n) \leq \mu(\theta) \leq (1-\theta)\mu + \sum_{i=m+1}^{2m} \theta^i (e^T h^i/n).$$

Using (4.8), with $\alpha = 2$ and $\tau = 4/\sqrt{\beta}$, we obtain

$$\left| \sum_{i=m+1}^{2m} \theta^i \left(\frac{e^T h^i}{n} \right) \right| \leq \frac{1}{\sqrt{n}} \sum_{i=m+1}^{2m} \theta^i \left\| h^i \right\|_2 < .5\,\theta\mu$$

$\forall\, \theta \in [0,\, \theta_6]$. Using Lemma 4.4 and the sum of a geometric series with ratio .1, we deduce that for any $\theta \in [0,\, \frac{\sqrt{\beta}}{40(1+2\kappa)\sqrt{n}}]$ it holds that

$$\mu'(\theta) = -(1+\epsilon-2\epsilon\theta)\mu + \sum_{i=m+1}^{2m} i\theta^{i-1} \left(\frac{e^T h^i}{n} \right) \leq -\mu + \frac{1}{\sqrt{n}} \sum_{i=m+1}^{2m} i\theta^{i-1} \left\| h^i \right\|_2$$
$$\leq -\mu + 8\mu\sqrt{\beta} \sum_{i=m}^{2m-1} \left(\frac{4(1+2\kappa)\theta\sqrt{n}}{\sqrt{\beta}} \right)^i < -\mu + 8\mu\sqrt{\beta} \frac{.1^m}{1-.1}$$
$$< -\mu + 8\mu \frac{.1^m}{.9} < 0.$$

Since $\theta_6 > \frac{\sqrt{\beta}}{44.8(1+2\kappa)\sqrt{n}} > \frac{\sqrt{\beta}}{50(1+2\kappa)\sqrt{n}}$, we conclude that

$$(1-2.5\,\theta)\mu \le \mu(\theta) \le (1-.5\,\theta)\mu \quad \text{and} \quad \mu'(\theta) < 0 \tag{4.21}$$

$$\forall \theta \in \left[0,\ \frac{\sqrt{\beta}}{50(1+2\kappa)\sqrt{n}}\right]. \tag{4.22}$$

Next, we claim that the quantity θ_2 from (3.11) used in the computation of the predictor step-length satisfies

$$\begin{aligned}\theta_2 \ge \theta_7 &:= \frac{\sqrt{\beta}}{4(1+2\kappa)\sqrt{n}} \min\left\{1\,,\left(11.2\sqrt{\beta}\right)^{\frac{-1}{m_p}}\,,\left(\frac{(1+2\kappa)\delta}{2\beta}\right)^{\frac{1}{m_p+1}}\right\}\\ &\ge \frac{\chi_7}{(1+2\kappa)n^{\frac{1}{2}+\frac{m_c+1}{2m_c(m_p+1)}}}\,, \qquad (4.23)\end{aligned}$$

$$\chi_7 := \frac{1}{4}\min\left\{\frac{1}{11.2},\left(\frac{(1-\beta)\underline{\gamma}\chi_5}{4}\right)^{\frac{1}{2}}\right\}. \tag{4.24}$$

Using (4.20) with $n \ge 8$ we obtain

$$\mu(\theta) \ge (1-2.5\theta_6)\mu \ge \left(1-\frac{2.5}{8\sqrt{2}}\right)\mu \ge .7\mu \quad \forall\,\theta \in [0,\,\theta_6].$$

By taking $\gamma = 0$, and $\beta+\delta$ instead of β, in (4.17), using (4.7) with $\alpha = 1/(.7\delta)$, we deduce that

$$f(\theta) \ge \beta+\delta - \frac{\sum_{i=m+1}^{2m}\theta^i \left\| h^i \right\|_2}{\mu(\theta)} \ge \beta+\delta - \frac{1}{.7\mu}\sum_{i=m+1}^{2m}\theta^i \left\| h^i \right\|_2 \ge \beta \quad \forall\,\theta \in [0,\,\theta_7],$$

which proves that $\theta_2 \ge \theta_7$. From the definition of θ_7 it follows that

$$\begin{aligned}\theta_7 &\ge \frac{\sqrt{\beta}}{4(1+2\kappa)\sqrt{n}}\min\left\{\frac{1}{11.2\sqrt{\beta}},\frac{1}{\sqrt{\beta}}\left(\frac{(1+2\kappa)\delta}{2}\right)^{\frac{1}{m_p+1}}\right\}\\ &\ge \frac{1}{4(1+2\kappa)\sqrt{n}}\min\left\{\frac{1}{11.2},\left(\frac{(1-\beta)\underline{\gamma}\theta_5}{4}\right)^{\frac{1}{m_p+1}}\right\}\\ &\ge \frac{\sqrt{\beta}}{4(1+2\kappa)\sqrt{n}}\min\left\{\frac{1}{11.2},\left(\frac{(1-\beta)\underline{\gamma}\chi_5}{4n^{\frac{m_c+1}{2m_c}}}\right)^{\frac{1}{m_p+1}}\right\}\\ &\ge \frac{\chi_7}{(1+2\kappa)n^{\frac{1}{2}+\frac{m_c+1}{2m_c(m_p+1)}}}.\end{aligned}$$

Bounding the decrease of the duality gap. Due to the fact that the duality gap decreases both in the predictor step and the corrector step, a complete analysis of the decreases of the duality gap has to be done by studying a succession of corrector-predictor-corrector steps. Assume that we are at iteration k and have a point $z^k \in \mathcal{D}(\beta)$ with normalized duality gap μ_k. We follow the notations of Algorithm 1. The corrector step produces a point $\overline{z}^{\,k} \in \mathcal{D}(\beta+\delta)$, with δ given by (4.18). The corresponding normalized duality gap clearly satisfies $\overline{\mu}_k \le \mu_k$, but a bound on the decrease of the duality gap cannot be given at this stage. The corrector is followed by a predictor that produces a point $z^{k+1} = z(\theta_p) \in \mathcal{D}(\beta)$ with duality gap

$\mu_{k+1} = \mu(\theta_p) = \min_{0\le\theta\le\theta_2} \mu(\theta)$. We have $\theta_7 \le \theta_2$ and $\theta_7 \le \theta_6$, so that according to (4.20)

$$\mu_{k+1} \le \mu(\theta_7) \le (1 - .5\,\theta_7)\,\overline{\mu}_k \le \left(1 - \frac{\chi_7}{2\,n^{\frac{1}{2}+\frac{m_c+1}{2m_c(m_p+1)}}}\right)\overline{\mu}_k, \quad \overline{\mu}_k \le \mu_k. \tag{4.25}$$

The above relation is sufficient for proving polynomial complexity, but it does not take into account the contribution of the corrector step. A finer analysis is needed in order to account for that. We distinguish two cases:

(a) $\theta_2 \ge \frac{\sqrt{\beta}}{50(1+2\kappa)\sqrt{n}}$. According to (4.21), in this case we have

$$\mu_{k+1} = \min_{0\le\theta\le\theta_2} \mu(\theta) \le \left(1 - \frac{\sqrt{\beta}}{100(1+2\kappa)\sqrt{n}}\right)\overline{\mu}_k\,;$$

(b) $\theta_2 < \frac{\sqrt{\beta}}{50(1+2\kappa)\sqrt{n}}$. In this case $\mu(\theta)$ is decreasing on the interval $[0,\,\theta_2]$, by virtue of (4.21), and by using (4.23) we deduce that $\theta_p = \theta_2, f(\theta_p) = \beta$. The latter equality must be true, since if $f(\theta_p) > \beta$, then, by a continuity argument, it follows that $\theta_2 > \theta_p$, which is a contradiction (see the definition of θ_2 (3.11)). But if $f(\theta_p) = \beta$, then, according to (3.8), in the next corrector step we have $\theta_c > .2\,\theta_5$, so that

$$\overline{\mu}_{k+1} < (1 - .2\,\underline{\sigma}(1-\overline{\gamma})\theta_5)\,\mu_{k+1} \le \left(1 - \frac{\underline{\sigma}(1-\overline{\gamma})\chi_5}{5(1+2\kappa)\,n^{\frac{1}{2}+\frac{1}{2m_c}}}\right)\mu_{k+1}.$$

In conclusion, for any $k \ge 0$ we have

$$\overline{\mu}_{k+1} \le \mu_{k+1} \le \left(1 - \frac{\sqrt{\beta}}{100(1+2\kappa)\sqrt{n}}\right)\overline{\mu}_k$$

or

$$\overline{\mu}_{k+1} < \left(1 - \frac{\underline{\sigma}(1-\overline{\gamma})\chi_5}{5(1+2\kappa)\,n^{\frac{1}{2}+\frac{1}{2m_c}}}\right)\left(1 - \frac{\chi_7}{2(1+2\kappa)\,n^{\frac{1}{2}+\frac{m_c+1}{2m_c(m_p+1)}}}\right)\overline{\mu}_k.$$

By taking

$$\overline{\chi} := \min\left\{\frac{\sqrt{\beta}}{100},\ \frac{\underline{\sigma}(1-\overline{\gamma})\chi_5}{5}\,,\ \frac{\chi_7}{2}\right\},$$

we deduce that

$$\overline{\mu}_{k+1} \le \left(1 - \frac{\overline{\chi}}{(1+2\kappa)n^{\frac{1}{2}+\upsilon}}\right)\overline{\mu}_k, \quad k = 0, 1, \ldots,$$

where υ is given by (4.11). The proof is complete. □

As an immediate consequence of the above theorem we obtain the following complexity result.

COROLLARY 4.7. *Algorithm* 1 *produces a point* $z = \lceil\, x,\, s \,\rfloor \in \mathcal{D}(\beta)$ *with* $x^T s \le \varepsilon$, *in at most* $O((1+\kappa)n^{1/2+\upsilon}\log(x^{0\,T}s^0/\varepsilon))$ *iterations, where* υ *is given by* (4.11).

It follows that if the order of either the corrector or the predictor is larger than a multiple of $\log n$, then Algorithm 1 has $O\left((1+\kappa)\sqrt{n}L\right)$-iteration complexity.

TABLE 4.1

n	10^4	10^5	10^6	10^7	10^8	10^9	10^{10}
$\lceil n^{.1} \rceil$	3	4	4	6	7	8	11

COROLLARY 4.8. *If* $\max\{m_c, m_p\} = \Omega(\log n)$, *then Algorithm* 1 *produces a point* $z = \lceil x, s \rfloor \in \mathcal{D}(\beta)$ *with* $x^T s \leq \varepsilon$, *in at most* $O((1+\kappa)\sqrt{n}\log(x^{0\,T}s^0/\varepsilon))$.

Proof. Under the hypothesis of the corollary there is a constant ϑ, such that $\upsilon \leq \vartheta/\log n$. Hence $n^{1/2+\upsilon} \leq n^{\frac{\vartheta}{\log n}}\sqrt{n} = e^{\vartheta}\sqrt{n}$. □

Due to the fact that $\lim_{n\to\infty} n^{1/n^{\omega}} = 1$ for any $\omega \in (0,1)$, in applications we can choose $m_p = \lceil n^{\omega} \rceil$ for some value of $\omega \in (0,1)$. This choice was initially suggested by Roos (private communication) and subsequently used in [24] and [16]. A correspondence between n and $\lceil n^{\omega} \rceil$ with $\omega = 0.1$ is shown in Table 4.1.

5. Superlinear convergence. In this section we show that the duality gap of the sequence produced by Algorithm 1 is superlinearly convergent. The result is based on the following lemma, which is a consequence of the results about the analyticity of the central path from [19].

LEMMA 5.1. *If HLCP* (2.1) *is sufficient, then the solution of* (3.2) *with* $\gamma = 0$ *satisfies*

$$u^i = O(\mu^i), \quad v^i = O(\mu^i), \quad i = 1, \ldots, m, \text{ if } \text{HLCP } (2.1) \text{ is nondegenerate}$$

and

$$u^i = O(\mu^{i/2}), \quad v^i = O(\mu^{i/2}), \quad i = 1, \ldots, m, \text{ if } \text{HLCP } (2.1) \text{ is degenerate.}$$

By using the above lemma we obtain the following superlinear convergence result, which is a trivial extension of the corresponding result in [15] to sufficient linear complementarity problems.

THEOREM 5.2. *The sequence* μ_k *produced by Algorithm* 1 *satisfies*

$$\mu_{k+1} = O(\mu_k^{m_p+1}) \text{ if } \text{HLCP } (2.1) \text{ is nondegenerate}$$

and

$$\mu_{k+1} = O(\mu_k^{(m_p+1)/2}) \text{ if } \text{HLCP } (2.1) \text{ is degenerate.}$$

6. Conclusions. We have presented a corrector-predictor interior-point algorithm for sufficient HLCP acting in a wide neighborhood of the central path.

The corrector of order m_c is used to improve both the centrality and the complementarity gap. The predictor of order m_p follows each corrector step to further decrease the complementarity gap. If $\max\{m_c, m_p\} = \Omega(\log n)$, then the iteration complexity of the algorithms is $O((1+\kappa)\sqrt{n}L)$. Although the complexity of our algorithm depends on κ, the algorithm itself does not, so that the algorithm works for the class of sufficient HLCP. Our algorithm has the best known iteration complexity for sufficient linear complementarity problems and is superlinearly convergent even for degenerate problems. The cost of implementing one iteration of our algorithm is $O(n^3)$ arithmetic operations.

REFERENCES

[1] M. Anitescu, G. Lesaja, and F. A. Potra, *Equivalence between different formulations of the linear complementarity problem*, Optim. Methods Softw., 7 (1997), pp. 265–290.

[2] K. M. Anstreicher and R. A. Bosch, *A new infinity-norm path following algorithm for linear programming*, SIAM J. Optim., 5 (1995), pp. 236–246.

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

[4] C. C. Gonzaga, *Complexity of predictor-corrector algorithms for LCP based on a large neighborhood of the central path*, SIAM J. Optim., 10 (1999), pp. 183–194.

[5] P.-F. Hung and Y. Ye, *An asymptotical ($O\sqrt{n}L$)-iteration path-following linear programming algorithm that uses wide neighborhoods*, SIAM J. Optim., 6 (1996), pp. 570–586.

[6] J. Ji, F. A. Potra, and S. Huang, *Predictor-corrector method for linear complementarity problems with polynomial complexity and superlinear convergence*, J. Optim. Theory Appl., 85 (1995), pp. 187–199.

[7] M. Kojima, N. Megiddo, T. Noma, and A. Yoshise, *A Unified Approach to Interior Point Algorithms for Linear Complementarity Problems*, Lecture Notes in Comput. Sci. 538, Springer-Verlag, New York, 1991.

[8] J. Miao, *A quadratically convergent $O((1+k)\sqrt{n}L)$-iteration algorithm for the $P_*(k)$-matrix linear complementarity problem*, Math. Programming, 69 (1995), pp. 355–368.

[9] 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.

[10] R. D. C. Monteiro and S. J. Wright, *Local convergence of interior-point algorithms for degenerate monotone LCP*, Comput. Optim. Appl., 3 (1994), pp. 131–155.

[11] R. C. Monteiro, I. Adler, and M. G. Resende, *A polynomial-time primal-dual affine scaling algorithm for linear and convex quadratic programming and its power series extension*, Math. Oper. Res., 15 (1990), pp. 191–214.

[12] J. Peng, T. Terlaky, and Y. Zhao, *A predictor-corrector algorithm for linear optimization based on a specific self-regular proximity function*, SIAM J. Optim., 15 (2005), pp. 1105–1127.

[13] F. A. Potra, *An $O(nL)$ infeasible interior–point algorithm for LCP with quadratic convergence*, Ann. Oper. Res., 62 (1996), pp. 81–102.

[14] F. A. Potra, *A superlinearly convergent predictor-corrector method for degenerate LCP in a wide neighborhood of the central path with $O(\sqrt{n}l)$-iteration complexity*, Math. Program., 100 (2004), pp. 317–337.

[15] F. A. Potra, *Corrector-predictor methods for monotone linear complementarity problems in a wide neighborhood of the central path*, Math. Program., to appear.

[16] F. A. Potra and X. Liu, *Predictor-corrector methods for sufficient linear complementarity problems in a wide neighborhood of the central path*, Optim. Methods Softw., 20 (2005), pp. 145–168.

[17] F. A. Potra and R. Sheng, *A large-step infeasible-interior-point method for the P_*-matrix LCP*, SIAM J. Optim., 7 (1997), pp. 318–335.

[18] J. Stoer and M. Wechs, *Infeasible-interior-point paths for sufficient linear complementarity problems and their analyticity*, Math. Programming, 83 (1998), pp. 407–423.

[19] J. Stoer, M. Wechs, and S. Mizuno, *High order infeasible-interior-point methods for solving sufficient linear complementarity problems*, Math. Oper. Res., 23 (1998), pp. 832–862.

[20] J. Stoer, *High order long-step methods for solving linear complementarity problems*, Ann. Oper. Res., 103 (2001), pp. 149–159.

[21] J. F. Sturm, *Superlinear convergence of an algorithm for monotone linear complementarity problems, when no strictly complementary solution exists*, Math. Oper. Res., 24 (1999), pp. 72–94.

[22] H. Väliaho, *P_*-matrices are just sufficient*, Linear Algebra Appl., 239 (1996), pp. 103–108.

[23] Y. Ye and K. Anstreicher, *On quadratic and $O(\sqrt{n}L)$ convergence of predictor-corrector algorithm for LCP*, Math. Programming, 62 (1993), pp. 537–551.

[24] G. Zhao, *Interior point algorithms for linear complementarity problems based on large neighborhoods of the central path*, SIAM J. Optim., 8 (1998), pp. 397–413.

SIAM J. OPTIM.
Vol. 17, No. 3, pp. 891–919
© 2006 Society for Industrial and Applied Mathematics

A REGULARIZED SAMPLE AVERAGE APPROXIMATION METHOD FOR STOCHASTIC MATHEMATICAL PROGRAMS WITH NONSMOOTH EQUALITY CONSTRAINTS*

FANWEN MENG† AND HUIFU XU‡

Abstract. We investigate a class of two stage stochastic programs where the second stage problem is subject to nonsmooth equality constraints parameterized by the first stage variant and a random vector. We consider the case when the parametric equality constraints have more than one solution. A regularization method is proposed to deal with the multiple solution problem, and a sample average approximation method is proposed to solve the regularized problem. We then investigate the convergence of stationary points of the regularized sample average approximation programs as the sample size increases. The established results are applied to stochastic mathematical programs with P_0-variational inequality constraints. Preliminary numerical results are reported.

Key words. sample average approximation, Karush–Kuhn–Tucker conditions, regularization methods, P_0-variational inequality, convergence of stationary points

AMS subject classifications. 90C15, 90C30, 90C31, 90C33

DOI. 10.1137/050638242

1. Introduction.

In this paper, we study the following stochastic mathematical program:

$$\begin{array}{ll} \min & \mathbb{E}\left[f(x, y(x, \xi(\omega)), \xi(\omega))\right] \\ \text{s.t.} & x \in \mathcal{X}, \end{array} \tag{1}$$

where $\mathcal{X}$ is a nonempty compact subset of $\mathbb{R}^m$, $f : \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}$ is continuously differentiable, $\xi : \Omega \to \Xi \subset \mathbb{R}^k$ is a vector of random variables defined on probability space $(\Omega, \mathcal{F}, P)$, $\mathbb{E}$ denotes the mathematical expectation, and $y(x, \xi(\omega))$ is *some measurable selection* (which will be reviewed in section 2.2) from the set of solutions of the following system of equations:

$$H(x, y, \xi(\omega)) = 0, \tag{2}$$

where $H : \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^n$ is a piecewise smooth vector-valued function. Piecewise smooth function is a large class of locally Lipschitz continuous functions which cover most practical problems [29]. For the simplicity of notation, we will write $\xi(\omega)$ as ξ, and this should be distinguished from where ξ is a deterministic vector of Ξ in a context. Throughout this paper, we assume that the probability measure P of our considered space $(\Omega, \mathcal{F}, P)$ is nonatomic.

The model is slightly different from the standard two stage stochastic programming model where the second stage decision variate y is chosen to either minimize or maximize $f(x, y, \xi)$ for given x and every realization of $\xi(\omega)$. See an excellent survey

*Received by the editors August 16, 2005; accepted for publication (in revised form) May 15, 2006; published electronically October 24, 2006. This research was supported by United Kingdom Engineering and Physical Sciences Research Council grant GR/S90850/01.

http://www.siam.org/journals/siopt/17-3/63824.html

†School of Mathematics, University of Southampton, Southampton SO17 1BJ, United Kingdom. Current address: The Logistics Institute - Asia Pacific, National University of Singapore, 1 Law Link, 119260 Singapore (tlimf@nus.edu.sg).

‡School of Mathematics, University of Southampton, Southampton SO17 1BJ, United Kingdom (h.xu@soton.ac.uk).

by Ruszczyński and Shapiro [27, Chapters 1 and 2] for the latter. In some practical instances, finding an optimal solution for the second stage problem may be very difficult and/or expensive. For example, in a two stage stochastic Stackelberg–Nash–Cournot model [32], x is the leader's decision variable and y is the followers' Nash–Cournot equilibrium vector with each component representing a follower's decision variable. The followers' equilibrium problem can be reformulated as a system of nonsmooth equations like (2) which depends on the leader's decision variable x and realization of uncertainty ξ in market demand. In the case when the followers have multiple equilibria, the "selection" of an optimal $y(x,\xi)$ at the second stage can be interpreted as the leader's attitude towards the followers' multiple equilibria: an optimistic attitude leads to a selection in favor of his utility, whereas a pessimistic attitude goes to an opposite selection. See [32, section 2] for details. Alternatively such an optimal selection can be interpreted as that the leader puts in some resources so that the followers reach an equilibrium in his favor. In either interpretation, finding such an optimal $y(x,\xi)$ implies additional cost to the leader.

Our argument here is that the leader may not necessarily select an extreme equilibrium (which minimizes/maximizes $f(x,y,\xi)$); instead, he may select one of the feasible equilibria $y(x,\xi)$ under the following circumstances: (a) minimizing/maximizing $f(x,y,\xi)$ with respect to y may be difficult or even impossible numerically; for instance, one can obtain only a local optimal solution or a stationary point; (b) in the case when an optimal solution is obtainable, the cost for obtaining such a solution in the second stage overweighs the overall benefit; for instance, the leader is a dominant player, while the followers are small players and the range of possible followers' equilibria is very narrow; (c) the chance of followers reaching one equilibrium or another is equal, and the leader is unaware of which particular equilibrium may be actually reached in the future and has no intention of putting in any additional resources to influence it.

Of course, such a selection must be consistent for all x and ξ; in other words, $y(x,\xi)$ must be a single-valued function with some measurability or even continuity. Note that there may exist many such selections, and the leader considers only one of them. This means that the leader takes a neutral attitude towards the follower's every possible equilibrium. Note also that in this paper, the selection is not arbitrary and is guided by a regularization method to be discussed shortly.

The argument can be extended to two stage stochastic mathematical programs with equilibrium constraints (SMPECs)

$$\begin{array}{ll} \min & \mathbb{E}[f(x,y(x,\xi),\xi)] \\ \text{s.t.} & x \in \mathcal{X}, \end{array} \tag{3}$$

where $y(x,\xi)$ solves

$$\begin{array}{ll} \min\limits_{y} & f(x,y,\xi) \\ \text{s.t.} & F(x,y,\xi)^T(v-y) \geq 0 \quad \forall v \in \mathcal{C}(x,\xi), \end{array} \tag{4}$$

f and F are continuously differentiable function, ξ is a random variable, and $\mathcal{C}(x,\xi)$ is a random convex set. SMPECs were initially studied by Patriksson and Wynter [18]. Like deterministic MPECs, SMPECs have various potential applications in engineering and economics, etc. [9, 34]. Over the past few years, SMPECs have received increasing investigation from perspectives of both stochastic programming and MPEC; see [17, 30, 35, 36] and the references therein. Observe that the second stage problem

(4) is a deterministic parametric MPEC that is intrinsically nonconvex. Finding an optimal solution for MPECs is often difficult if not impossible. Consequently it may be a realistic approach to take *some* feasible measurable solution at the second stage (which is a solution of the variational inequality problem in the constraint) rather than trying to find an optimal one. Note that MPECs can be easily reformulated as a nonsmooth system of equations, and this is the very reason why we consider general nonsmooth equality constraints (2). We will discuss these in detail in section 5.

Having motivated our model, we next explain how to find the unspecified $y(x,\xi)$ in (1). Our idea can be outlined as follows. We approximate function H with some function R parameterized by a small positive number μ and then solve the following equation:

$$R(x,y,\xi,\mu)=0. \tag{5}$$

Of course, R cannot be any function, and it must be constructed according to the structure of H. First, it must coincide with H when $\mu=0$; second, it must have some nice topological properties such as Lipschitz continuity and directional differentiability. Finally and perhaps most importantly, (5) must have a unique solution for every $x\in\mathcal{X}$, $\xi\in\Xi$, and nonzero μ. We specify these needed properties in a definition of R (Definition 2.1) and regard R as a *regularization* in consistency with the terminology in the literature [19, 11]. Using the regularization method, we expect that an implicit function $\tilde{y}(x,\xi,\mu)$ defined by (5) approximates a measurable feasible solution $y(x,\xi)$ of (2), and consequently we can utilize the program

$$\begin{array}{ll}\min & \mathbb{E}\left[f(x,\tilde{y}(x,\xi,\mu),\xi)\right]\\ \text{s.t.} & x\in\mathcal{X}\end{array} \tag{6}$$

to approximate the true problem (1), where $y(x,\xi)$ is the limit of $\tilde{y}(x,\xi,\mu)$ as $\mu\to 0$.

We then propose a sample average approximation (SAA) method to solve (6). The SAA method and its variants, known under various names such as "stochastic counterpart method," "sample-path method," "simulated likelihood method," etc., were discussed in the stochastic programming and statistics literature over the years. See, for instance, [22, 25, 3, 31] for general stochastic problems and [17, 4, 30, 32, 36] for SMPECs.

We investigate the convergence of the SAA problem of (6) as $\mu\to 0$ and sample size tends to infinity. Since the underlying functions are piecewise smooth and nonconvex in general, our analysis focuses on stationary points rather than local or global optimal solutions. For this purpose, we study the optimality conditions for both the true and the regularized problems. We introduce a kind of generalized Karush–Kuhn–Tucker (KKT) condition for characterizing both true and regularized problems in terms of Clarke generalized Jacobians (subdifferentials). Rockafellar and Wets [23] investigated KKT conditions for a class of two stage convex stochastic programs and derived some "basic Kuhn–Tucker conditions" in terms of convex subdifferentials. More recent discussions on the optimality conditions can also be found in books by Birge and Louveaux [5] and Ruszczyński and Shapiro [27]. These conditions rely on the convexity of underlying functions and hence cannot be applied to our problems, which are nonconvex.

The main contributions of this paper can be summarized as follows: we show that under some conditions, the solution of (5) approximates a measurable solution of (2); we then show that with probability 1 (w.p.1 for short) an accumulation point of a sequence of generalized stationary points of the regularized problem is a generalized

stationary point of the true problem. We propose an SAA method to solve the regularized problem (6) and show that w.p.1 an accumulation point of the sequence of the stationary points of the regularized SAA problem is a generalized stationary point of the true problem as sample size tends to infinity and parameter μ tends to zero. Finally, we apply the established results to a class of SMPECs where the underlying function is a P_0-function.

The rest of the paper is organized as follows. In section 2, we discuss the regularization scheme. In section 3, we investigate the generalized stationary points of both the regularized problem and the true problem. In section 4, we study the convergence of the SAA program of the regularized problem. We then apply the established results to a class of stochastic MPEC problems in section 5. Some preliminary numerical results are reported in section 6.

2. Preliminaries and a regularization scheme. In this section, we characterize the function R in (5) and investigate the approximation of (6) to (1) as $\mu \to 0$.

Throughout this paper, we use the following notation. We use $\|\cdot\|$ to denote the Euclidean norm of a vector, a matrix, and a compact set of matrices. Specifically, if $\mathcal{M}$ is a compact set of matrices, then $\|\mathcal{M}\| := \max_{M\in\mathcal{M}} \|M\|$. We use $\operatorname{dist}(x, \mathcal{D}) := \inf_{x'\in\mathcal{D}} \|x - x'\|$ to denote the distance between point x and set $\mathcal{D}$. Here $\mathcal{D}$ may be a subset of $\mathbb{R}^n$ or a subset of matrix space $\mathbb{R}^{n\times n}$. Given two compact sets $\mathcal{C}$ and $\mathcal{D}$, we use $\mathbb{D}(\mathcal{C}, \mathcal{D}) := \sup_{x\in\mathcal{C}} \operatorname{dist}(x, \mathcal{D})$ to denote the distance from set $\mathcal{C}$ to set $\mathcal{D}$. For two sets $\mathcal{C}$ and $\mathcal{D}$ in a metric space, $\mathcal{C} + \mathcal{D}$ denotes the usual Minkowski addition, and $\mathcal{C}\mathcal{D} := \{CD \mid \text{ for all } C \in \mathcal{C}, \text{ for all } D \in \mathcal{D}\}$ represents the multiplication. We use $\mathcal{B}(x, \delta)$ to denote the closed ball in $\mathbb{R}^n$ with radius δ and center x, that is, $\mathcal{B}(x, \delta) := \{x' \in \mathbb{R}^m : \|x' - x\| \le \delta\}$. For a vector-valued function $g : \mathbb{R}^m \to \mathbb{R}^l$, we use $\nabla g(x)$ to denote the classical Jacobian of g when it exists. In the case when $l = 1$, that is, g is a real-valued function, $\nabla g(x)$ denotes the gradient of g which is a row vector. We use $\overline{\lim}$ to denote the outer limit of a sequence of vectors and a set-valued mapping. We let $\mathbb{R}_{++} := \{x \mid x > 0, x \in \mathbb{R}\}$ and $\mathbb{R}^2_{++} := \{(x, y) \mid x > 0, y > 0, x, y \in \mathbb{R}\}$. For a set-valued mapping $\mathcal{A}(u, v) : \mathbb{R}^n \times \mathbb{R}^m \to 2^{\mathbb{R}^{n\times m}}$, $\pi_u\mathcal{A}(u, v)$ denotes the set of all $n \times n$ matrices M such that, for some $n \times m$ matrix N, the $n \times (n+m)$ matrix $[M\ N]$ belongs to $\mathcal{A}(u, v)$.

2.1. Preliminaries. We first present some preliminaries about the Clarke generalized Jacobian of random functions which will be used throughout the paper.

Let $G : \mathbb{R}^j \to \mathbb{R}^l$ be a locally Lipschitz continuous vector-valued function. Recall that the Clarke generalized Jacobian [10] of G at $x \in \mathbb{R}^j$ is defined as

$$\partial G(x) := \operatorname{conv}\left\{\lim_{y\to x, y\in D_G} \nabla G(y)\right\},$$

where D_G denotes the set of points at which G is Fréchet differentiable, $\nabla G(y)$ denotes the usual Jacobian of G, and "conv" denotes the convex hull of a set. It is well known that the Clarke generalized Jacobian $\partial G(x)$ is a convex compact set [10]. In the case that $j = l$, $\partial G(x)$ consists of square matrices. We say $\partial G(x)$ is nonsingular if every matrix in set $\partial G(x)$ is nonsingular, and in this case we use $\partial G(x)^{-1}$ to denote the set of all inverse matrices of $\partial G(x)$.

In later discussions, particularly sections 2–4, we will have to deal with mathematical expectation of the Clarke generalized Jacobians of locally Lipschitz random functions. For this purpose, we recall some basics about measurability of a random set-valued mapping.

Let $V \subset \mathbb{R}^n$ be a compact set of $\mathbb{R}^n$ and $\xi : \Omega \to \Xi \subset \mathbb{R}^k$ be a random vector defined on the probability space $(\Omega, \mathcal{F}, P)$ (note that we use the same notation ξ and Ξ as in (1), although we do not have to in this general discussion). A random set-valued mapping $\mathcal{A}(\cdot, \xi) : V \to 2^{\mathbb{R}^{n\times m}}$ is said to be *closed-valued* if for every $v \in V$ and $\xi \in \Xi$ (a realization of $\xi(\omega)$), $\mathcal{A}(v, \xi)$ is a closed set. Let $\mathfrak{B}$ denote the space of nonempty, closed subsets of $\mathbb{R}^{n\times m}$ equipped with Hausdorff distance. Then $\mathcal{A}(v, \xi(\cdot))$ can also be viewed as a single-valued mapping from Ω to $\mathfrak{B}$. For a fixed $v \in V$, $\mathcal{A}(v, \xi(\cdot)) : \Omega \to 2^{\mathbb{R}^{n\times m}}$ is said to be *measurable* if for every closed set $B \subset \mathbb{R}^{n\times m}$, $\{\omega : \mathcal{A}(v, \xi(\omega)) \bigcap B \neq \emptyset\}$ belongs to the σ-algebra $\mathcal{F}$. Alternatively, by viewing $\mathcal{A}(v, \xi(\cdot))$ as a single-valued mapping, we can say that $\mathcal{A}(v, \xi(\cdot))$ is measurable if and only if for every $B \in \mathfrak{B}$, $\mathcal{A}(v, \xi(\cdot))^{-1}B$ is $\mathcal{F}$-measurable. See Theorem 14.4 of [24].

We now define the expectation of $\mathcal{A}(v, \xi(\omega))$. A *selection* of a random set $\mathcal{A}(v, \xi(\omega))$ is a random matrix $A(v, \xi) \in \mathcal{A}(v, \xi)$ (which means $A(v, \xi(\omega))$ is measurable). Selections exist. The *expectation of* $\mathcal{A}(v, \xi(\omega))$, denoted by $\mathbb{E}[\mathcal{A}(v, \xi(\omega))]$, is defined as the collection of $\mathbb{E}[A(v, \xi(\omega))]$, where $A(v, \xi(\omega))$ is a selection. For a detailed discussion in this regard, see [1, 2] and the references therein.

Finally, we need the following definitions concerning matrices and functions. A matrix $M \in \mathbb{R}^{l\times l}$ is called a P_0-matrix if for any $x \neq 0$, there exists $i \in \{1, \ldots, l\}$ such that $x_i(Mx)_i \geq 0$ and $x_i \neq 0$. It is evident that a positive semidefinite matrix is a P_0-matrix. A function $G : \mathcal{D} \subset \mathbb{R}^l \to \mathbb{R}^l$ is said to be (over set $\mathcal{D}$) a P_0-function if for all $u, v \in \delta$ with $u \neq v$,

$$\max_{\substack{i\in\{1,\ldots,l\}\\ u_i\neq v_i}} (u_i - v_i)[G_i(u) - G_i(v)] \geq 0.$$

For a continuously differentiable function G, if $\nabla G(x)$ is a P_0-matrix for all $x \in \mathcal{D}$, then $G(x)$ is a P_0-function on $\mathcal{D}$. For a comprehensive discussion of the properties of the above matrices and functions, we refer readers to the book [11].

2.2. A regularization scheme. We specify the regularized approximation outlined in section 1 and investigate the limiting behavior of the implicit function defined by the regularized approximation problem (6) as $\mu \to 0$.

Throughout this paper $\partial H(x, y, \xi)$ denotes the Clarke generalized Jacobian of H at (x, y, ξ), and $\partial R(x, y, \xi, \mu)$ denotes the Clarke generalized Jacobian of H at (x, y, ξ, μ).

DEFINITION 2.1. *Let $\mu \in [0, \mu_0]$, where μ_0 is a positive number. A continuous function $R : \mathcal{X} \times \mathbb{R}^n \times \Xi \times [0, \mu_0] \to \mathbb{R}^n$ is said to be a* regularization *of H if the following hold:*

(i) *for every $x \in \mathcal{X}$, $y \in \mathbb{R}^n$, $\xi \in \Xi$, $R(x, y, \xi, 0) = H(x, y, \xi)$;*

(ii) *$R(x, y, \xi, \mu)$ is locally Lipschitz continuous and piecewise smooth on $\mathcal{X} \times \mathbb{R}^n \times \Xi \times [0, \mu_0]$;*

(iii) *for every $x \in \mathcal{X}$, $y \in \mathbb{R}^n$, and $\xi \in \Xi$,*

$$\overline{\lim_{\mu\downarrow 0}}\, \pi_x \partial R(x, y, \xi, \mu) \subset \pi_x \partial H(x, y, \xi), \quad \overline{\lim_{\mu\downarrow 0}}\, \pi_y \partial R(x, y, \xi, \mu) \subset \pi_y \partial H(x, y, \xi);$$

(iv) *equation $R(x, y, \xi, \mu) = 0$ defines a unique locally Lipschitz continuous function $\tilde{y} : \mathcal{X} \times \Xi \times (0, \mu_0) \to \mathbb{R}^n$ such that $R(x, \tilde{y}(x, \xi, \mu), \xi, \mu) = 0$ for every $x \in \mathcal{X}, \mu \in (0, \mu_0), \ \xi \in \Xi$.*

We call μ a regularization parameter (or variable).

The definition contains three elements. First, a regularization is a parameterized continuous approximation and is locally Lipschitz continuous with respect to the

regularization parameter when it is viewed as an additional variable. Second, the regularization (part (iii)) satisfies some kind of Jacobian consistency [8] that was widely used in smoothing methods when R is a smoothing of H; see [21] and the references therein. A sufficient condition for this is that R is strictly differentiable at $\mu = 0$ (when μ is treated as a variable). Third, the regularization scheme defines a unique function $\tilde{y}$ that approximates a measurable solution $y(x, \xi)$ of (2). We shall investigate the existence of such $\tilde{y}$ in Proposition 2.3.

Remark 2.2. Part (iv) of Definition 2.1 is implied by the following *uniform nonsingularity condition*: for every $(x, \xi, \mu) \in \mathcal{X} \times \Xi \times (0, \mu_0)$, there exists y such that $R(x, y, \xi, \mu) = 0$; $\pi_y \partial R(x, y, \xi, \mu)$ is uniformly nonsingular; i.e., there exists a positive constant $C > 0$ such that for every $x \in \mathcal{X}, y \in \mathbb{R}^n, \xi \in \Xi, \mu \in (0, \mu_0)$, $\|[\pi_y \partial R(x, y, \xi, \mu)]^{-1}\| \leq C$. The uniform nonsingularity implies that the outer limit of $\pi_y \partial R(x, y, \xi, \mu)$ as $\mu \to 0$ is a strict subset of $\pi_y \partial H(x, y, \xi)$, which does not include singular matrices.

PROPOSITION 2.3. *Let R be a function satisfying conditions* (i)–(iii) *of Definition* 2.1 *and the uniform nonsingularity condition hold. Then R is a regularization of H.*

Proof. It suffices to verify (iv) in Definition 2.1; that is, (5) defines a unique locally Lipschitz continuous implicit function $\tilde{y} : \mathcal{X} \times \Xi \times (0, \mu_0) \to \mathbb{R}^n$ such that $R(x, \tilde{y}(x, \xi, \mu), \xi, \mu) = 0$ for all $x \in \mathcal{X}, \xi \in \Xi, \mu \in (0, \mu_0)$. With the uniform nonsingularity of $\pi_y \partial R$, the existence of such an implicit function on $\mathcal{X} \times \mathbb{R}^n \times (0, \mu_0)$ comes straightforwardly from [36, Lemma 2.3]. ∎

In the analysis of sections 3 and 4, we will not assume the uniform nonsingularity. Instead we will assume a regularization R with nonsingularity of $\pi_y \partial R$ and other conditions which are weaker than the uniform nonsingularity. Note that not every function has a regularized approximation. Our definition here is motivated by the functions reformulated from equilibrium constraints. See section 5, particularly Example 5.5, for a detailed explanation. In what follows we investigate the properties of $\tilde{y}(x, \xi, \mu)$, in particular, its limit as the regularization parameter μ tends to zero.

THEOREM 2.4. *Let R be a regularization of H and $\tilde{y}(x, \xi, \mu)$ be the implicit function defined as in Definition* 2.1. *Assume that $\lim_{\mu \downarrow 0} \tilde{y}(x, \xi, \mu)$ exists for every $x \in \mathcal{X}$ and $\xi \in \Xi$, that is,*

$$y(x, \xi) := \lim_{\mu \downarrow 0} \tilde{y}(x, \xi, \mu), \quad x \in \mathcal{X}, \xi \in \Xi. \tag{7}$$

Suppose that there exists a positive measurable function $\kappa_1(\xi) > 0$ such that $\|\tilde{y}(x, \xi, \mu)\| \leq \kappa_1(\xi)$ for all $(x, \mu) \in \mathcal{X} \times (0, \mu_0)$ and that $\mathbb{E}[\kappa_1(\xi)] < \infty$. Then the following statements hold:

(i) *$y(x, \xi)$ is a solution function of* (2) *on $\mathcal{X} \times \Xi$, and $y(x, \xi(\cdot)) : \Omega \to \mathbb{R}^n$ is measurable for every $x \in \mathcal{X}$;*

(ii) *if, in addition, there exists a measurable positive function $L(\xi) > 0$ such that*

$$\|\tilde{y}(x'', \xi, \mu) - \tilde{y}(x', \xi, \mu)\| \leq L(\xi)\|x'' - x'\| \quad \forall x', x'' \in \mathcal{X}, \tag{8}$$

then $y(\cdot, \xi)$ is Lipschitz continuous on $\mathcal{X}$ for every $\xi \in \Xi$.

Proof. Part (i). By Definition 2.1, we have $R(x, \tilde{y}(x, \xi, \mu), \xi, \mu) = 0$ for $x \in \mathcal{X}, \xi \in \Xi, \mu \in (0, \mu_0)$. By Definition 2.1 and the continuity of R in y, it follows that

$$\lim_{\mu \downarrow 0} R(x, \tilde{y}(x, \xi, \mu), \xi, \mu) = R(x, y(x, \xi), \xi, 0) = H(x, y(x, \xi), \xi),$$

which indicates that $y(x, \xi)$ is a solution of (2) for $(x, \xi) \in \mathcal{X} \times \Xi$.

To show the measurability of $y(x,\xi(\omega))$, observe that since $\tilde{y}(x,\xi,\mu)$ is continuous in ξ, then $\tilde{y}(x,\xi(\cdot),\mu):\Omega\to\mathbb{R}^n$ is measurable. Moreover, since $\tilde{y}(x,\xi,\mu)$ is bounded by $\kappa_1(\xi)$ and $\mathbb{E}[\kappa_1(\xi)]<\infty$, by the Lebesgue dominated convergence theorem, $y(x,\xi(\cdot))$ is measurable.

Part (ii). For $x',x''\in\mathcal{X}$, by (8), $\|\tilde{y}(x'',\xi,\mu)-\tilde{y}(x',\xi,\mu)\|$ is dominated by $L(\xi)\|x''-x'\|$. The latter is integrable. By the Lebesgue dominated convergence theorem, we have from (8)

$$\begin{aligned}\|y(x'',\xi)-y(x',\xi)\| &= \|\lim_{\mu\downarrow 0}(\tilde{y}(x'',\xi,\mu)-\tilde{y}(x',\xi,\mu))\|\\ &= \lim_{\mu\downarrow 0}\|\tilde{y}(x'',\xi,\mu)-\tilde{y}(x',\xi,\mu)\|\le L(\xi)\|x''-x'\| \quad \forall x',x''\in\mathcal{X}.\end{aligned}$$

This completes the proof. □

The theorem above shows that we may obtain a measurable solution of (2) through the process of regularization. Note that our assumption on the existence of limit (7) may be relaxed. Indeed if the sequence of functions $\tilde{y}(\cdot,\cdot,\mu)$ has multiple accumulation points, each of which is Lipschitz continuous, then $y(x,\xi)$ can be taken from any of them. Our assumption is to simplify the consequent discussion, and also we expect this to be satisfied in practical instances; see Example 5.5. The boundedness condition for $\tilde{y}(x,\xi,\mu)$ holds under the uniform nonsingularity condition (Remark 2.2). Throughout the rest of this paper, the $y(x,\xi)$ in the true problem (1) refers to the limit of $\tilde{y}(x,\xi,\mu)$ as $\mu\to 0$.

3. Generalized Karush–Kuhn–Tucker (GKKT) conditions. In this section, we investigate the KKT conditions of both true problem (1) and regularized program (6). Our purpose is to show that w.p.1 the stationary points of the regularized problem converge to a stationary point of the true problem (1) as the regularization parameter is driven to zero; therefore the regularized problem (6) is a reasonable approximation of the true problem.

3.1. GKKT conditions of the true problem. Let R be a regularization of H and $\tilde{y}(x,\xi,\mu)$ the solution of (5), let

$$y(x,\xi)=\lim_{\mu\downarrow 0}\tilde{y}(x,\xi,\mu) \quad \text{for } x\in\mathcal{X},\xi\in\Xi,$$

and let $y(\cdot,\xi)$ be Lipschitz continuous. In this subsection, we investigate the true problem (1) associated with $y(x,\xi)$. We first define a set which resembles the set of Lagrange multipliers in nonlinear programming.

DEFINITION 3.1. *For $(x,\xi)\in\mathcal{X}\times\Xi$, let*

$$\begin{aligned}\Lambda(x,\xi) :=& \operatorname{conv}\{\lambda(x,\xi)\in\mathbb{R}^n \mid 0\in\nabla_y f(x,y(x,\xi),\xi)\\ &+\lambda(x,\xi)\pi_y\partial H(x,y(x,\xi),\xi)\}.\end{aligned} \tag{9}$$

Note that $\lambda(x,\xi)$ is a row vector. Note also that when $y(x,\xi)$ is an optimal solution of

$$\begin{aligned}\min_y \quad & f(x,y,\xi)\\ \text{s.t.} \quad & H(x,y,\xi)=0,\end{aligned} \tag{10}$$

$\Lambda(x,\xi)$ contains the Lagrange multipliers of (10) in that the first-order necessary condition of (10) can be written as $0\in\nabla_y f(x,y(x,\xi),\xi)+\lambda(x,\xi)\partial_y H(x,y(x,\xi),\xi)$, and

by [10, Proposition 2.3.16], $\partial_y H(x, y(x,\xi),\xi) \subset \pi_y \partial H(x, y(x,\xi),\xi)$. Note further that $\Lambda(x,\xi)$ is nonempty if and only if the set $\{\lambda(x,\xi) \in \mathbb{R}^n \mid 0 \in \nabla_y f(x, y(x,\xi),\xi) + \lambda(x,\xi)\ \pi_y \partial H(x, y(x,\xi),\xi)\}$ is nonempty. For every $\lambda(x,\xi)$ of the latter set, there exists a matrix $M \in \pi_y \partial H(x, y(x,\xi),\xi)$ such that $0 = \nabla_y f(x, y(x,\xi),\xi) + \lambda(x,\xi)M$. A necessary and sufficient condition for $\Lambda(x,\xi)$ to be nonempty is that there exists $M \in \pi_y \partial H(x, y(x,\xi),\xi)$ such that $\operatorname{rank}([\nabla_y f^T, M^T]) = \operatorname{rank}(M^T)$. Moreover, the set Λ is bounded if and only if M is of full row rank. The following remark discusses the particular case when $\pi_y \partial H$ is nonsingular.

Remark 3.2. If $\pi_y \partial H(x, y(x,\xi),\xi)$ is nonsingular for $x \in \mathbb{R}^m$ and $\xi \in \Xi$, then we have

$$\Lambda(x,\xi) := -\nabla_y f(x, y(x,\xi),\xi)\operatorname{conv}([\pi_y \partial H(x, y(x,\xi),\xi)]^{-1}).$$

The set contains the Lagrange multipliers of the standard second stage minimization problem (10), since the nonsingularity of the Jacobian guarantees $y(x,\xi)$ to be the only feasible solution and hence trivially the optimal solution! If H is continuously differentiable in x, y, and ξ, then $\pi_y \partial H(x,y,\xi)$ reduces to $\nabla_y H(x,y,\xi)$, and $\Lambda(x,\xi)$ reduces to a singleton,

$$\Lambda(x,\xi) = -\nabla_y f(x, y(x,\xi),\xi)\nabla_y H(x, y(x,\xi),\xi)^{-1}.$$

This corresponds to the classical Lagrange multiplier of a standard second stage minimization problem (10).

In this paper, we consider the case when (2) has multiple solutions; therefore $\pi_y \partial H(x,y,\xi)$ cannot be nonsingular. For the simplicity of discussion, we make a blanket assumption that $\Lambda(x,\xi)$ is nonempty for $x \in \mathcal{X}$ and $\xi \in \Xi$, which implies that there exists at least one matrix $M \in \pi_y \partial H(x, y(x,\xi),\xi)$ such that $\operatorname{rank}([\nabla_y f^T, M^T]) = \operatorname{rank}(M^T)$. Using the notion of Λ, we can define the following optimality conditions for the true problem associated with $y(x,\xi)$.

Definition 3.3. *Let $\Lambda(x,\xi)$ be defined as in Definition* 3.1. *A point $x \in \mathbb{R}^m$ is called a* generalized stationary point *of the true problem* (1) *if*

$$0 \in \mathbb{E}[\nabla_x f(x, y(x,\xi),\xi) + \Lambda(x,\xi)\pi_x \partial H(x, y(x,\xi),\xi)] + \mathcal{N}_{\mathcal{X}}(x), \tag{11}$$

where the expectation is taken over the integrable elements of the set-valued integrand, and $\mathcal{N}_{\mathcal{X}}(x)$ denotes the normal cone of $\mathcal{X}$ at $x \in \mathbb{R}^m$ [6]; *that is,*

$$\mathcal{N}_{\mathcal{X}}(x) := [T_{\mathcal{X}}(x)]^- = \{\varsigma \in \mathbb{R}^m \mid \langle \varsigma, d\rangle \le 0 \quad \forall d \in T_{\mathcal{X}}(x)\},$$

where $T_{\mathcal{X}}(x) := \limsup_{t \downarrow 0}(\mathcal{X} - x)/t$. We call (11) *a GKKT condition of the true problem* (1).

For the set of generalized stationary points to be well defined, it must contain all local minimizers of the true problem. In what follows we discuss this and the relationship between the GKKT conditions with other possible KKT conditions. For this purpose, we need to state the following implicit function theorem for piecewise smooth functions.

Lemma 3.4. *Consider an underdetermined system of nonsmooth equations $P(y,z) = 0$, where $P : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^m$ is piecewise smooth. Let $(\bar{y},\bar{z}) \in \mathbb{R}^m \times \mathbb{R}^n$ be such that $P(\bar{y},\bar{z}) = 0$. Suppose that $\pi_y \partial P(\bar{y},\bar{z})$ is nonsingular. Then*

(i) *there exist neighborhoods Z of $\bar{z}$ and Y of $\bar{y}$ and a piecewise smooth function $y : Z \to Y$ such that $y(\bar{z}) = \bar{y}$ and, for every $z \in Z$, $y = y(z)$ is the unique solution of the problem $P(y,z) = 0$, $y \in Y$;*

(ii) *for* $z \in Z$,

$$\partial y(z) \subset \operatorname{conv}\{-V^{-1}U : [V,U] \in \partial P(y(z),z),\ V \in \mathbb{R}^{m\times m},\ U \in \mathbb{R}^{m\times n}\}. \tag{12}$$

Proof. The existence of an implicit function comes essentially from the Clarke implicit function theorem [10]. The piecewise smoothness and the differential inclusion (12) follow from [21, Proposition 4.8] straightforwardly. ∎

Note that the purpose of (12) is to give an estimate of the Clarke generalized Jacobian of the implicit function using the Clarke generalized Jacobian of P, which is relatively easier to obtain. The estimate may be improved under some index consistency conditions; see [21] for details. With Lemma 3.4, we are ready to discuss our GKKT condition. The proposition below establishes a relation between (local) minimizers of the true problem and the generalized stationary points defined in Definition 3.3 under some circumstances.

PROPOSITION 3.5. *Let x^* be a local minimizer of the true problem (associated with the limit function $y(x,\xi)$) and $[\pi_y \partial H(x^*, y(x^*,\xi),\xi)]^{-1}$ be nonsingular. Let $\nabla f(x, y(x,\xi),\xi)$, $[\pi_y \partial H(x, y(x,\xi),\xi)]^{-1}$, and $\pi_x \partial H(x, y(x,\xi),\xi)$ be bounded by a positive integrable function for all x in a neighborhood of x^*. Then x^* is a generalized stationary point of the true problem.*

Proof. By Lemma 3.4,

$$\partial y(x,\xi) \subset \operatorname{conv}([-\pi_y \partial H(x, y(x,\xi),\xi)]^{-1})\pi_x \partial H(x, y(x,\xi),\xi)$$

for x close to x^*. Let $v(x,\xi) := f(x, y(x,\xi),\xi)$. Under the boundedness conditions of $\nabla f(x, y(x,\xi),\xi)$, $[\pi_y \partial H(x, y(x,\xi),\xi)]^{-1}$, and $\pi_x \partial H(x, y(x,\xi),\xi)$, we have

$$\begin{aligned}
0 &\in \partial \mathbb{E}[v(x^*,\xi)] + \mathcal{N}_{\mathcal{X}}(x^*) \subset \mathbb{E}[\partial_x v(x^*,\xi)] + \mathcal{N}_{\mathcal{X}}(x^*) \\
&= \mathbb{E}[\nabla_x f(x^*, y(x^*,\xi),\xi) + \nabla_y f(x^*, y(x^*,\xi),\xi)\partial y(x^*,\xi)] + \mathcal{N}_{\mathcal{X}}(x^*) \\
&\subset \mathbb{E}[\nabla_x f(x^*, y(x^*,\xi),\xi) - \nabla_y f(x^*, y(x^*,\xi),\xi)\operatorname{conv}([\pi_y \partial H(x^*, y(x^*,\xi),\xi)]^{-1}) \\
&\quad \times \pi_x \partial H(x^*, y(x^*,\xi),\xi)] + \mathcal{N}_{\mathcal{X}}(x^*) \\
&= \mathbb{E}[\nabla_x f(x^*, y(x^*,\xi),\xi) + \Lambda(x^*,\xi)\pi_x \partial H(x^*, y(x^*,\xi),\xi)] + \mathcal{N}_{\mathcal{X}}(x^*). \qquad (13)
\end{aligned}$$

The inclusion $\partial \mathbb{E}[v(x^*,\xi)] \subset \mathbb{E}[\partial_x v(x^*,\xi)]$ is deduced from the fact that the Clarke generalized directional derivative of $\mathbb{E}[v(x,\xi)]$ is bounded by the expected value of the Clarke generalized directional derivative of $v(x,\xi)$. See [33, Proposition 2.12]. The conclusion follows. ∎

Remark 3.6. In the case when ∂H is singular, if

$$\nabla_y f(x^*, y(x^*,\xi),\xi)\partial y(x^*,\xi) \subset \Lambda(x^*,\xi)\pi_x \partial H(x^*, y(x^*,\xi),\xi), \tag{14}$$

then we can draw a similar conclusion.

Note that if $v(x,\xi)$ is regular at x^* in the sense of Clarke [10, Definition 2.3.4] and $\|\partial_x v(x,\xi)\|$ is bounded by some integrable function $\eta(\xi)$, then by [15, Proposition 5.1], $\partial \mathbb{E}[v(x^*,\xi)] = \mathbb{E}[\partial_x v(x^*,\xi)]$; consequently, equality holds in the first inclusion of (13). This implies that the set of stationary points satisfying $0 \in \partial \mathbb{E}[v(x^*,\xi)] + \mathcal{N}_{\mathcal{X}}(x^*)$ coincides with the set of stationary points satisfying $0 \in \mathbb{E}[\partial_x v(x^*,\xi)] + \mathcal{N}_{\mathcal{X}}(x^*)$. Our discussions above show that all these stationary points are contained in the set of the generalized stationary points satisfying (11) under some appropriate conditions, which means the latter gives a bound or an estimate of the former. To see how precise the estimate is, we need to look at the second inclusion in (13) or the inclusion in (14). The former relies on the index consistency of the piecewise smooth function H

in y at the considered point [21]. The latter depends on the structure of Λ and $\pi_x\partial H$. In general the inclusions are strict but perhaps not very loose. See [21, section 5] for the comparisons of various GKKT conditions for deterministic nonsmooth equality constrained minimization problems.

3.2. GKKT conditions of the regularized problem. We now consider the GKKT conditions of the regularized program (6). Throughout this subsection and section 4, we make the following assumption.

Assumption 3.7. Let R be a regularization of H. $\pi_y\partial R(x,\tilde{y}(x,\xi,\mu),\xi,\mu)$ is nonsingular for $x\in\mathcal{X}$, $\xi\in\Xi$, $\mu\in(0,\mu_0)$.

This assumption is rather moderate and is satisfied by many regularizations. See section 5 for a detailed discussion. Let us define the mapping of multipliers of the regularized problem.

DEFINITION 3.8. *For $(x,\xi,\mu)\in\mathcal{X}\times\Xi\times(0,\mu_0)$, let*

$$\begin{aligned}\Lambda^{\mathrm{reg}}(x,\xi,\mu) :=&\mathrm{conv}\{\lambda(x,\xi)\in\mathbb{R}^n \mid 0\in\nabla_y f(x,\tilde{y}(x,\xi,\mu),\xi)\\ &+\lambda(x,\xi)\pi_y\partial R(x,\tilde{y}(x,\xi,\mu),\xi,\mu)\}.\end{aligned}\tag{15}$$

Since $\pi_y\partial R(x,\tilde{y}(x,\xi,\mu),\xi,\mu)$ is nonsingular, then Λ^{reg} can be rewritten as

$$\Lambda^{\mathrm{reg}}(x,\xi,\mu) = -\nabla_y f(x,\tilde{y}(x,\xi,\mu),\xi)\mathrm{conv}([\pi_y\partial R(x,\tilde{y}(x,\xi,\mu),\xi,\mu)]^{-1}).\tag{16}$$

Obviously, Λ^{reg} contains the set of Lagrange multipliers of the trivial second stage regularized problem:

$$\min_y f(x,y,\xi)\ \text{ s.t. }\ R(x,y,\xi,\mu)=0,$$

since $\tilde{y}(x,\xi,\mu)$ is the unique feasible solution. Using the notion of Λ^{reg}, we define the stationary point of the regularized problem.

DEFINITION 3.9. *Let $\Lambda^{\mathrm{reg}}(x,\xi,\mu)$ be defined as in Definition* 3.8. *A point $x\in\mathbb{R}^m$ is called a* generalized stationary point *of the regularized problem* (6) *if*

$$0\in\mathbb{E}[\nabla_x f(x,\tilde{y}(x,\xi,\mu),\xi)+\Lambda^{\mathrm{reg}}(x,\xi,\mu)\pi_x\partial R(x,\tilde{y}(x,\xi,\mu),\xi,\mu)]+\mathcal{N}_{\mathcal{X}}(x).\tag{17}$$

We call condition (17) *a GKKT condition for the regularized problem* (6). *Note that this definition depends on the function $\tilde{y}(x,\xi,\mu)$.*

Let $\tilde{v}(x,\xi,\mu) := f(x,\tilde{y}(x,\xi,\mu),\xi)$. Obviously, $\tilde{v}(\cdot,\xi,\mu)$ is locally Lipschitz continuous, since $\tilde{y}(\cdot,\xi,\mu)$ is locally Lipschitz continuous by assumption. Note that by Lemma 3.4

$$\partial\tilde{y}(x,\xi,\mu)\subset -\mathrm{conv}([\pi_y\partial R(x,\tilde{y}(x,\xi,\mu),\xi,\mu)]^{-1})\pi_x\partial R(x,\tilde{y}(x,\xi,\mu),\xi,\mu).$$

If $x^*\in\mathcal{X}$ be a local minimizer of the regularized problem, then under some appropriate measurable conditions (of $\partial_x\tilde{v}$, etc.) we have

$$\begin{aligned}0&\in\partial\mathbb{E}[\tilde{v}(x^*,\xi,\mu)]+\mathcal{N}_{\mathcal{X}}(x^*)\subset\mathbb{E}[\partial_x\tilde{v}(x^*,\xi,\mu)]+\mathcal{N}_{\mathcal{X}}(x^*)\\ &=\mathbb{E}[\nabla_x f(x^*,\tilde{y}(x^*,\xi,\mu),\xi)+\nabla_y f(x^*,\tilde{y}(x^*,\xi,\mu),\xi)\partial_x\tilde{y}(x^*,\xi,\mu)]+\mathcal{N}_{\mathcal{X}}(x^*)\\ &\subset\mathbb{E}[\nabla_x f(x^*,\tilde{y}(x^*,\xi,\mu),\xi)-\nabla_y f(x^*,\tilde{y}(x^*,\xi,\mu),\xi)\\ &\quad\times\mathrm{conv}([\pi_y\partial R(x^*,\tilde{y}(x^*,\xi,\mu),\xi,\mu)]^{-1})\pi_x\partial R(x^*,\tilde{y}(x^*,\xi,\mu),\xi,\mu)]+\mathcal{N}_{\mathcal{X}}(x^*)\\ &=\mathbb{E}[\nabla_x f(x^*,\tilde{y}(x^*,\xi,\mu),\xi)+\Lambda^{\mathrm{reg}}(x^*,\xi)\pi_x\partial R(x^*,\tilde{y}(x^*,\xi,\mu),\xi,\mu)]\\ &\quad+\mathcal{N}_{\mathcal{X}}(x^*),\end{aligned}\tag{18}$$

which implies that an optimal solution of the regularized problem (6) is a generalized stationary point.

3.3. Convergence analysis of the regularized problem. In this subsection, we investigate the convergence of the stationary points of regularized problem as $\mu \to 0$. We first state the following intermediate result.

LEMMA 3.10. *Let R be a regularization of H and $\Lambda^{\rm reg}$ be defined as in Definition* 3.8. *Suppose that there exists a function $\nu_1(\xi) > 0$ such that*

$$\|\Lambda^{\rm reg}(x,\xi,\mu)\| \le \nu_1(\xi) \quad \forall (x,\mu) \in \mathcal{X} \times (0,\mu_0) \tag{19}$$

and that $\mathbb{E}[\nu_1(\xi)] < \infty$. Then $\Lambda^{\rm reg}(\cdot,\xi,\cdot)$ is upper semicontinuous on $\mathcal{X} \times (0,\mu_0)$, and

$$\overline{\lim_{\mu \downarrow 0}} \Lambda^{\rm reg}(x,\xi,\mu) \subset \Lambda(x,\xi), \quad x \in \mathcal{X}. \tag{20}$$

Proof. The upper semicontinuity of $\Lambda^{\rm reg}$ on $\mathcal{X} \times (0,\mu_0)$ follows from (16), the upper semicontinuity of $\pi_y \partial R(x,y,\xi,\mu)$, and $\nabla_y f(x,y,\xi)$ with respect to x,y,μ. In what follows we show (20). By the definition of $\Lambda^{\rm reg}$,

$$-\nabla_y f(x,\tilde{y}(x,\xi,\mu),\xi) \in \Lambda^{\rm reg}(x,\xi,\mu)\pi_y \partial R(x,\tilde{y}(x,\xi,\mu),\xi,\mu). \tag{21}$$

By Definition 2.1(iii),

$$\overline{\lim_{\mu \downarrow 0}} \pi_y \partial R(x,\tilde{y}(x,\xi,\mu),\xi,\mu) \subset \pi_y \partial H(x,y(x,\xi),\xi).$$

Therefore $\pi_y \partial R(x,\tilde{y}(x,\xi,\mu),\xi,\mu)$ is bounded for μ close to 0. This and condition (19) allow us to take an outer limit on both sides of (21)

$$\begin{aligned} -\nabla_y f(x,y(x,\xi),\xi) &\in \overline{\lim_{\mu \downarrow 0}} [\Lambda^{\rm reg}(x,\xi,\mu)\pi_y \partial R(x,\tilde{y}(x,\xi,\mu),\xi,\mu)] \\ &\subset \overline{\lim_{\mu \downarrow 0}} \Lambda^{\rm reg}(x,\xi,\mu) \overline{\lim_{\mu \downarrow 0}} \pi_y \partial R(x,\tilde{y}(x,\xi,\mu),\xi,\mu). \end{aligned}$$

Then we arrive at

$$-\nabla_y f(x,y(x,\xi),\xi) \in \left[\overline{\lim_{\mu \downarrow 0}} \Lambda^{\rm reg}(x,\xi,\mu)\right] \pi_y \partial H(x,y(x,\xi),\xi)).$$

The conclusion follows immediately from the definition of Λ. □

Note that the boundedness condition (19) is satisfied if $[\pi_y \partial R(x,\tilde{y}(x,\xi,\mu),\xi,\mu)]^{-1}$ is uniformly bounded (see Remark 2.2) and f is uniformly Lipschitz continuous in y. We are now ready to present the main result of this section concerning the convergence of the stationary points of the regularized problem.

THEOREM 3.11. *Suppose that assumptions in Theorem* 2.4 *are satisfied. Suppose also that there exists a function $\kappa_2(\xi)$, where $\mathbb{E}[\kappa_2(\xi)] < \infty$, such that for all $(x,\xi,\mu) \in \mathcal{X} \times \Xi \times (0,\mu_0)$,*

$$\max\{\|\nabla_x f(x,\tilde{y}(x,\xi,\mu),\xi)\|, \|\pi_x \partial R(x,\tilde{y}(x,\xi,\mu),\xi,\mu)\|\} \le \kappa_2(\xi). \tag{22}$$

Let $\{x(\mu)\}$ be a sequence of generalized stationary points of the regularized problem (6). *Assume that x^* is an accumulation point of the sequence as $\mu \to 0$. Suppose that condition* (19) *holds and that $\mathbb{E}[\kappa_2(\xi)(1+\nu_1(\xi))] < \infty$. Then w.p.*1 *$x^*$ is a generalized stationary point of the true problem* (1), *that is,*

$$0 \in \mathbb{E}[\nabla_x f(x^*,y(x^*,\xi),\xi) + \Lambda(x^*,\xi)\pi_x \partial H(x^*,y(x^*,\xi),\xi)] + \mathcal{N}_{\mathcal{X}}(x^*).$$

Proof. We use the Lebesgue dominated convergence theorem to prove the result. Let

$$\mathcal{K}(x,\xi,\mu) := \nabla_x f(x,\tilde{y}(x,\xi,\mu),\xi) + \Lambda^{\rm reg}(x,\xi,\mu)\pi_x \partial R(x,\tilde{y}(x,\xi,\mu),\xi,\mu).$$

First note that $x(\mu)$ is a generalized stationary point of (6), that is,

$$0 \in \mathbb{E}[\mathcal{K}(x(\mu),\xi,\mu)] + \mathcal{N}_{\mathcal{X}}(x(\mu)). \tag{23}$$

Note that, by Lemma 3.10, $\overline{\lim}_{\mu\downarrow 0}\Lambda^{\rm reg}(x,\xi,\mu) \subset \Lambda(x,\xi)$. By (19), $\Lambda^{\rm reg}(x,\xi,\mu)$ is uniformly dominated by $\nu_1(\xi)$ for μ sufficiently small and x close to x^*. On the other hand, by (22), $\nabla_x f(x,y,\xi)$ and $\pi_x\partial R(x,\tilde{y}(x,\xi,\mu),\xi,\mu)$ are uniformly dominated by $\kappa_2(\xi)$ for μ sufficiently small. Hence, $\mathcal{K}(x,\xi,\mu)$ is uniformly dominated by $\kappa_2(\xi)(1+\nu_1(\xi))$ for μ small enough and x sufficiently close to x^*. Note that $\mathbb{E}[\kappa_2(\xi)(1+\nu_1(\xi))] < \infty$; by the Lebesgue dominated convergence theorem, we then have

$$\begin{aligned}
&\overline{\lim_{\mu\downarrow 0}}\,\mathbb{E}[\mathcal{K}(x(\mu),\xi,\mu)] = \mathbb{E}\left[\overline{\lim_{\mu\downarrow 0}}\,\mathcal{K}(x(\mu),\xi,\mu)\right]\\
&= \mathbb{E}\left[\overline{\lim_{\mu\downarrow 0}}\,[\nabla_x f(x(\mu),\tilde{y}(x(\mu),\xi,\mu),\xi) + \Lambda^{\rm reg}(x(\mu),\xi,\mu)\pi_x\partial R(x(\mu),\tilde{y}(x(\mu),\xi,\mu),\xi,\mu)]\right].
\end{aligned}$$

By Theorem 2.4 and Definition 2.1, we have by taking a subsequence if necessary on $\{x(\mu)\}$

$$\overline{\lim_{\mu\downarrow 0}}\,\nabla_x f(x(\mu),\tilde{y}(x(\mu),\xi,\mu),\xi) = \nabla_x f(x^*,y(x^*,\xi),\xi)$$

and

$$\overline{\lim_{\mu\downarrow 0}}\,\pi_x\partial R(x(\mu),\tilde{y}(x(\mu),\xi,\mu),\xi,\mu) \subset \pi_x\partial H(x^*,y(x^*,\xi),\xi).$$

In addition, notice that $\overline{\lim}_{\mu\downarrow 0}\Lambda^{\rm reg}(x(\mu),\xi,\mu) \subset \Lambda(x^*,\xi)$. Thus, with (23), it yields that

$$0 \in \mathbb{E}[\nabla_x f(x^*,y(x^*,\xi),\xi) + \Lambda(x^*,\xi)\pi_x\partial H(x^*,y(x^*,\xi),\xi)] + \mathcal{N}_{\mathcal{X}}(x^*).$$

This completes the proof. ∎

Note that when $f(x,y,\xi)$ and $H(x,y,\xi)$ are uniformly Lipschitz in x, condition (22) is satisfied, since $\pi_x\partial R(x,\tilde{y}(x,\xi,\mu),\xi,\mu)$ approximates $\pi_x\partial H(x,y(x,\xi),\xi)$ following Definition 2.1(iii).

4. Sample average approximations. In this section, we propose a sample average approximation (SAA) method for solving the regularized program (6). SAA methods have been extensively investigated in SMPECs recently. See, for instance, [17, 30, 32, 35, 36]. Our convergence analysis is similar to that in [36]. However, there are two main differences: (a) $\tilde{y}(x,\xi,\mu)$ is a solution of a regularized equation (2), which may be nonsmooth, while in [36] $\tilde{y}(x,\xi,\mu)$ is an implicit smoothing of $y(x,\xi)$ and is smooth in x; (b) $y(x,\xi)$ is the limit of $\{\tilde{y}(x,\xi,\mu)\}_{\mu\to 0}$ which satisfies (2) but it not necessarily a unique implicit function of (2).

Let $\xi^1,\ldots,\xi^N$ be an independent, identically distributed sample of ξ. We consider the following SAA program:

$$\begin{aligned}
\min_{x\in\mathcal{X},y^1,\ldots,y^N} \quad & f_N(x,y^1,\ldots,y^N) := \frac{1}{N}\sum_{i=1}^N f(x,y^i,\xi^i)\\
\text{s.t.} \quad & R(x,y^i,\xi^i,\mu) = 0, \quad i=1,\ldots,N.
\end{aligned} \tag{24}$$

Here $\mu > 0$ is a small positive number which may depend on sample size N in practical computation. Problem (24) is essentially a deterministic continuous minimization problem with variables x and $y^1, \ldots, y^N$. It can also be regarded as a two stage stochastic program with finite discrete distribution. Choosing which numerical method for solving (24) depends on the structure and size of the problem. If the problem is of relatively small size, and R is smooth, then many existing nonlinear programming methods may be readily applied to solving the problem. When R is not continuously differentiable, we need to employ those which can deal with nonsmoothness. Bundle methods and aggregate subgradient methods are effective ones.

In the case when the problem size is large, decomposition methods which are popular in dealing with large scale stochastic programs seem to be the choice. Of course, choosing which particular decomposition method also depends on the structure of the problem such as linearity, convexity, separability, and sparsity of the underlying functions. Higle and Sen [13] and Ruszczyński [26] presented a comprehensive discussion and review of various decomposition methods for solving two stage stochastic programs. We refer readers to them and the references therein for the methods.

Note that our model (1) is motivated by SMPECs; hence it might be helpful to explain how (24) is possibly solved when applied to SMPECs. For many practical SMPEC problems such as the stochastic leader-followers problem and capital expansion problem, f is convex in y, whereas $f(x, y(x, \xi), \xi)$ is usually nonconvex in x. Moreover, the feasible set of variable y is governed by a complementarity constraint and is often nonconvex. This means that the feasible set of variable y^i defined by an equality constraint in (24) is nonconvex when $\mu = 0$. However, since we assume $\pi_y R$ is nonsingular for $\mu > 0$, the equation has a unique solution y^i for given x and ξ^i; that is, the feasible set of y^i is a singleton. This implies the minimization with respect to variable y^i is trivial theoretically, albeit not numerically, and this can be achieved by solving an N system of equations simultaneously. Based on these observations, if we can solve (24), then we are likely to obtain a point $(x_N(\mu), y_N^1(\mu), \ldots, y_N^N(\mu))$ with $x_N(\mu)$ being a stationary point, while $y_N^i(\mu)$ is the unique global minimizer which depends on $x_N(\mu)$. Alternatively, we can say that $x_N(\mu)$ is a stationary point of (6).

In what follows, we focus on the Clarke stationary points of (24) given the nonsmooth and nonconvex nature of the problem. Following Hiriart-Urruty [14], we can write down the GKKT conditions of (24) as follows:

$$
\left\{
\begin{array}{l}
0 \in \dfrac{1}{N} \displaystyle\sum_{i=1}^{N} \nabla_x f(x, y^i, \xi^i) + \sum_{i=1}^{N} \lambda^i \partial_x R(x, y^i, \xi^i, \mu) + \mathcal{N}_{\mathcal{X}}(x), \\[2ex]
0 \in \dfrac{1}{N} \begin{pmatrix} \nabla_y f(x, y^1, \xi^1) \\ \vdots \\ \nabla_y f(x, y^N, \xi^N) \end{pmatrix} + \begin{pmatrix} \lambda^1 \partial_y R(x, y^1, \xi^1, \mu) \\ \vdots \\ \lambda^N \partial_y R(x, y^N, \xi^N, \mu) \end{pmatrix}, \\[2ex]
0 = R(x, y^i, \xi^i, \mu), \quad i = 1, \ldots, N.
\end{array}
\right. \tag{25}
$$

Since by assumption for every $(x, \xi) \in \mathcal{X} \times \Xi$, equation $R(x, y, \xi, \mu) = 0$ has a unique solution $\tilde{y}(x, \xi, \mu)$, then y^i in (25) can be expressed as $\tilde{y}(x, \xi^i, \mu)$, $i = 1, \ldots, N$.

Consequently, the above GKKT conditions can be rewritten as

(26)
$$\begin{cases} 0 \in \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \nabla_x f(x, \tilde{y}(x,\xi^i,\mu),\xi^i) + \sum_{i=1}^{N} \lambda^i \partial_x R(x,\tilde{y}(x,\xi^i,\mu),\xi^i,\mu) + \mathcal{N}_{\mathcal{X}}(x), \\ 0 \in \dfrac{1}{N}\begin{pmatrix} \nabla_y f(x,\tilde{y}(x,\xi^1,\mu),\xi^1) \\ \vdots \\ \nabla_y f(x,\tilde{y}(x,\xi^N,\mu),\xi^N) \end{pmatrix} + \begin{pmatrix} \lambda^1 \partial_y R(x,\tilde{y}(x,\xi^1,\mu),\xi^1,\mu) \\ \vdots \\ \lambda^N \partial_y R(x,\tilde{y}(x,\xi^N,\mu),\xi^N,\mu) \end{pmatrix}. \end{cases}$$

Note that by [10, Proposition 2.3.16],

$$\partial_x R(x,y,\xi,\mu) \subset \pi_x \partial R(x,y,\xi,\mu) \quad \text{and} \quad \partial_y R(x,y,\xi,\mu) \subset \pi_y \partial R(x,y,\xi,\mu).$$

In addition, under Assumption 3.7, $\pi_y \partial R(x,y^i,\xi^i,\mu)$ is nonsingular; then we replace λ^i, $i=1,\dots,N$, in (25) with

$$-\frac{1}{N}\nabla_y f(x,y^i,\xi^i)\mathrm{conv}([\pi_y \partial R(x,y^i,\xi^i,\mu)]^{-1}), \quad i=1,\dots,N.$$

By writing y^i as $\tilde{y}(x,\xi^i,\mu)$, we may consider a weaker GKKT condition than (26) as

$$\begin{aligned} 0 \in \frac{1}{N}\sum_{i=1}^{N} & [\nabla_x f(x,\tilde{y}(x,\xi^i,\mu),\xi^i) \\ & - \nabla_y f(x,\tilde{y}(x,\xi^i,\mu),\xi^i)\mathrm{conv}([\pi_y \partial R(x,\tilde{y}(x,\xi^i,\mu),\xi^i,\mu)]^{-1})\pi_x \partial R(x,\tilde{y}(x,\xi^i,\mu),\xi^i,\mu)] \\ & + \mathcal{N}_{\mathcal{X}}(x). \end{aligned}$$

The "weaker" is in the sense that a point x satisfying (26) must satisfy the above equation but not vice versa. Let $\Lambda^{\mathrm{reg}}(x,\xi,\mu)$ be defined as in (16). Then the above equation can be written as

$$\begin{aligned} 0 \in \frac{1}{N}\sum_{i=1}^{N} & \left[\nabla_x f(x,\tilde{y}(x,\xi^i,\mu),\xi^i) + \Lambda^{\mathrm{reg}}(x,\xi^i,\mu)\pi_x \partial R(x,\tilde{y}(x,\xi^i,\mu),\xi^i,\mu)\right] \\ & + \mathcal{N}_{\mathcal{X}}(x). \end{aligned} \tag{27}$$

We say that a point $x \in \mathcal{X}$ is a *generalized stationary point* of the reduced regularized SAA problem (24) if it satisfies (27). In what follows, we investigate the convergence of the generalized stationary points as the sample size tends to infinity. We consider two cases: (a) μ is set small and fixed, and the sample size N tends to infinity; (b) μ depends on the sample size and is reduced to zero as N increases to infinity.

We establish the following theorem, which states the convergence results of generalized stationary points.

THEOREM 4.1. *Let assumptions in Theorem* 3.11 *hold,* $\kappa_3(\xi) := \max(\nu_1(\xi), \kappa_2(\xi))$, *and* $\mathbb{E}[\kappa_3(\xi)(1+\kappa_3(\xi))] < \infty$. *Then the following statements hold:*

(i) *Let* $\mu > 0$ *be fixed. If* $\{x_N(\mu)\}$ *is a sequence of generalized stationary points which satisfy* (27), *then w.p.*1 *an accumulation point of the sequence is a generalized stationary point of the regularized problem* (6); *that is,*

$$0 \in \mathbb{E}\left[\mathcal{G}(x,\xi,\mu)\right] + \mathcal{N}_{\mathcal{X}}(x),$$

where $\mathcal{G}(x,\xi,\mu) := \nabla_x f(x,\tilde{y}(x,\xi,\mu),\xi) + \Lambda^{\mathrm{reg}}(x,\xi,\mu)\pi_x \partial R(x,\tilde{y}(x,\xi,\mu),\xi,\mu)$.

(ii) *Let $\mu = \mu_N$, where $\mu_N \to 0$ as $N \to \infty$, and $\{x(\mu_N)\}$ be a sequence of generalized stationary points which satisfy* (27). *Suppose that $\|\pi_x \partial H(x, y(x,\xi),\xi)\|$ is also bounded by $\kappa_2(\xi)$ in* (22). *If x^* is an accumulation point of $\{x(\mu_N)\}$, then w.p.*1 *x^* is a generalized stationary point of the true problem* (1); *that is, x^* satisfies*

$$0 \in \mathbb{E}[\mathcal{L}(x,\xi)] + \mathcal{N}_{\mathcal{X}}(x), \tag{28}$$

where $\mathcal{L}(x,\xi) := \nabla_x f(x, y(x,\xi),\xi) + \Lambda(x,\xi)\pi_x \partial H(x, y(x,\xi),\xi)$, Λ, and $y(x,\xi)$ are as given in section 3.

Proof. Part (i). By assumption, there exists a unique $\tilde{y}(x,\xi,\mu)$ such that

$$R(x, \tilde{y}(x,\xi,\mu),\xi,\mu) = 0$$

for every $(x,\xi,\mu) \in \mathcal{X} \times \Xi \times (0,\mu_0)$. Since $\partial R(\cdot, \tilde{y}(\cdot,\xi,\mu),\xi,\mu)$ is an upper semicontinuous, compact set-valued mapping, then $\mathcal{G}(\cdot,\cdot,\mu)$ is also an upper semicontinuous and compact set-valued mapping on $\mathcal{X}$ for every $\xi \in \Xi$. It follows from (19) and (22) that $\mathcal{G}(x,\xi,\mu)$ is uniformly dominated by $\kappa_3(\xi)(1+\kappa_3(\xi))$, which is integrable by assumption. Assume without loss of generality that $\{x_N(\mu)\} \to \{x^*\}$. Since $x_N(\mu)$ is a generalized stationary point of problem (24), we have by definition

$$0 \in \frac{1}{N}\sum_{i=1}^{N} \mathcal{G}(x_N(\mu),\xi^i,\mu) + \mathcal{N}_{\mathcal{X}}(x_N(\mu)). \tag{29}$$

For any sufficiently small $\delta > 0, \gamma > 0$, we estimate

$$\mathbb{D}\left(\frac{1}{N}\sum_{i=1}^{N} \mathcal{G}(x_N(\mu),\xi^i,\mu), \quad \mathbb{E}[\mathcal{G}_\delta(x^*,\xi,\mu)] + \gamma\mathcal{B}\right),$$

where $\mathcal{G}_\delta(x^*,\xi,\mu) := \bigcup_{x\in\mathcal{B}(x^*,\delta)} \mathcal{G}(x,\xi,\mu)$ and $\mathbb{E}[\mathcal{G}_\delta(x^*,\xi,\mu)] = \bigcup_{G\in\mathcal{G}_\delta(x^*,\xi,\mu)} \mathbb{E}[G]$. Note that

$$\begin{aligned}
&\mathbb{D}\left(\frac{1}{N}\sum_{i=1}^{N} \mathcal{G}(x_N(\mu),\xi^i,\mu), \quad \mathbb{E}[\mathcal{G}_\delta(x^*,\xi,\mu)] + \gamma\mathcal{B}\right) \\
&\le \mathbb{D}\left(\frac{1}{N}\sum_{i=1}^{N} \mathcal{G}(x_N(\mu),\xi^i,\mu), \quad \frac{1}{N}\sum_{i=1}^{N} \mathcal{G}_{\delta/2}(x^*,\xi^i,\mu) + \gamma/2\mathcal{B}\right) \\
&+ \mathbb{D}\left(\frac{1}{N}\sum_{i=1}^{N} \mathcal{G}_{\delta/2}(x^*,\xi^i,\mu) + \gamma/2\mathcal{B}, \quad \mathbb{E}[\mathcal{G}_\delta(x^*,\xi,\mu)] + \gamma\mathcal{B}\right).
\end{aligned}$$

By Lemma 3.2 of [36], the second term on the right-hand side of the equation tends to zero w.p.1 as $N \to \infty$. On the other hand, for N large enough such that $x_N(\mu) \in \mathcal{B}(x^*,\delta)$, by definition $\mathcal{G}(x_N(\mu),\xi^i,\mu) \subset \mathcal{G}_\delta(x^*,\xi^i,\mu)$, which leads to

$$\mathbb{D}\left(\frac{1}{N}\sum_{i=1}^{N} \mathcal{G}(x_N(\mu),\xi^i,\mu), \quad \frac{1}{N}\sum_{i=1}^{N} \mathcal{G}_\delta(x^*,\xi^i,\mu) + \gamma\mathcal{B}\right) = 0$$

for N sufficiently large. Hence, by (29), it follows that

$$0 \in \mathbb{E}[\mathcal{G}_\delta(x^*,\xi,\mu)] + \mathcal{N}_{\mathcal{X}}(x^*) + \gamma\mathcal{B} \text{ w.p.1.}$$

By the Lebesgue dominated convergence theorem and noticing the arbitrariness of δ and γ, we get the desired conclusion.

Part (ii). We now treat μ in $\mathcal{G}(x,\xi,\mu)$ as a variable and define

$$\hat{\mathcal{G}}(x,\xi,\mu) := \begin{cases} \mathcal{G}(x,\xi,\mu), & \mu > 0, \\ \mathcal{A}(x,\xi), & \mu = 0, \end{cases}$$

where $\mathcal{A}(x,\xi) := \nabla_x f(x,y(x,\xi),\xi) + \overline{\lim}_{\mu\downarrow 0} \Lambda^{\rm reg}(x,\xi,\mu)\pi_x \partial H(x,y(x,\xi),\xi)$. By assumption, it follows that $\hat{\mathcal{G}}(\cdot,\xi,\cdot) : \mathcal{X}\times[0,\mu_0) \to 2^{\mathbb{R}^n}$ is an upper semicontinuous and compact set-valued mapping for every $\xi\in\Xi$. Conditions (19) and (22) and the bound $\kappa_2(\xi)$ on $\pi_x\partial H(x,y(x,\xi),\xi)$ imply that $\hat{\mathcal{G}}(x,\xi,\mu)$ is bounded by $\kappa_3(\xi)(1+\kappa_3(\xi))$, which is integrable by assumption. Since $x(\mu_N)$ is a generalized stationary point of problem (24) with $\mu=\mu_N$, it follows that

$$0 \in \frac{1}{N}\sum_{i=1}^N \hat{\mathcal{G}}(x(\mu_N),\xi^i,\mu_N) + \mathcal{N}_{\mathcal{X}}(x(\mu_N)). \tag{30}$$

Assume without loss of generality that $x(\mu_N)\to x^*$ as $N\to\infty$. For any small $\delta>0$ and $\gamma>0$, we will show that

$$\mathbb{D}\left(\frac{1}{N}\sum_{i=1}^N \hat{\mathcal{G}}(x(\mu_N),\xi^i,\mu_N), \quad \mathbb{E}\left[\mathcal{L}_\delta(x^*,\xi)\right] + \gamma\mathcal{B}\right) \to 0, \quad \text{w.p.1 as } N\to\infty,$$

where $\mathcal{L}_\delta(x^*,\xi) = \hat{\mathcal{G}}_\delta(x^*,\xi,0), \hat{\mathcal{G}}_\delta(x,\xi,\mu) = \bigcup_{(x',\mu')\in\mathcal{B}(x,\delta)\times[0,\delta]} \mathcal{G}(x',\xi,\mu')$. Note that

$$\begin{aligned}
&\mathbb{D}\left(\frac{1}{N}\sum_{i=1}^N \hat{\mathcal{G}}(x(\mu_N),\xi^i,\mu_N), \quad \mathbb{E}\left[\mathcal{L}_\delta(x^*,\xi)\right] + \gamma\mathcal{B}\right) \\
&\le \mathbb{D}\left(\frac{1}{N}\sum_{i=1}^N \hat{\mathcal{G}}(x(\mu_N),\xi^i,\mu_N), \quad \frac{1}{N}\sum_{i=1}^N \hat{\mathcal{G}}_{\delta/2}(x^*,\xi^i,0) + \gamma/2\mathcal{B}\right) \\
&+ \mathbb{D}\left(\frac{1}{N}\sum_{i=1}^N \hat{\mathcal{G}}_{\delta/2}(x^*,\xi^i,0) + \gamma/2\mathcal{B}, \quad \mathbb{E}\left[\mathcal{L}_\delta(x^*,\xi)\right] + \gamma\mathcal{B}\right).
\end{aligned}$$

By Lemma 3.2 of [36], the second term on the right-hand side of the equation above tends to zero as $N\to\infty$. On the other hand, since $\hat{\mathcal{G}}(x(\mu_N),\xi^i,\mu_N) \subset \hat{\mathcal{G}}_\delta(x^*,\xi^i,0)$ for any $(x(\mu_N),\mu_N)\in\mathcal{B}(x^*,\delta)\times[0,\delta]$, hence the first term on the right-hand side equals zero for N sufficiently large. Since γ and δ are arbitrarily small, thereby, the conclusion follows immediately by virtue of the Lebesgue dominated convergence theorem and (30). The proof is completed. □

Theorem 4.1 states that if μ is fixed, then w.p.1. an accumulation point of a sequence of the generalized stationary points of the regularized SAA problem (24) is a generalized stationary point of the regularized problem (6). In the case that μ depends on sample size N and is reduced to zero as $N\to\infty$, then w.p.1. an accumulation point of a sequence of the generalized stationary points of the regularized SAA problem (24) is a generalized stationary point of the true problem.

5. Applications to SMPECs.

5.1. Stochastic program with variational constraints. In this section, we apply the results established in the preceding sections to the following stochastic mathematical programs with boxed constrained variational inequality (BVI) constraints:

$$\min_{x\in\mathcal{X}} \quad \mathbb{E}\left[f(x, y(x,\xi),\xi)\right], \tag{31}$$

where $y(x,\xi)$ is a measurable solution to the VI problem

$$F(x,y,\xi)^T(z-y)\geq 0 \quad \forall z\in\Upsilon, \tag{32}$$

where $\mathcal{X}$ is a nonempty compact subset of $\mathbb{R}^m$, $f:\mathbb{R}^m\times\mathbb{R}^n\times\mathbb{R}^k\to\mathbb{R}$ is continuously differentiable, $\xi:\Omega\to\Xi\subset\mathbb{R}^k$ is a vector of random variables defined on probability space $(\Omega,\mathcal{F},P)$ with nonatomic P, $F:\mathbb{R}^m\times\mathbb{R}^n\times\mathbb{R}^k\to\mathbb{R}^n$ is continuously differentiable, $F(x,\cdot,\xi)$ is a P_0-function for every $(x,\xi)\in\mathcal{X}\times\Xi$, $\Upsilon:=\{y\in\mathbb{R}^n \mid a\leq y\leq b\}$, $a\in\{\mathbb{R}\cup\{-\infty\}\}^n$, $b\in\{\mathbb{R}\cup\{\infty\}\}^n$, and $a<b$ (componentwise). Here we assume that Ξ is a compact set. Notice that if we set $a=0$ and $b=\infty$, then problem (31) is reduced to the stochastic mathematical programs with complementarity constraints. For simplicity in analysis, we assume all components in a or b are finite or infinite simultaneously. In other words, we will focus on the following cases: (i) $a\in\mathbb{R}^n$ and $b\in\mathbb{R}^n$; (ii) $a\in\mathbb{R}^n$, $b=\infty$; (iii) $a=-\infty$, $b\in\mathbb{R}^n$; (iv) $a=-\infty$, $b=\infty$.

For every $(x,\xi)\in\mathcal{X}\times\Xi$, the constraint of the second stage problem (32) is actually a parametric BVI problem. Throughout this section, we assume that the BVI has at least one solution for every $(x,\xi)\in\mathcal{X}\times\Xi$.

Let $\Pi_\Upsilon(y)$ be the Euclidean projection of y onto Υ. Then the parametric BVI can be reformulated as a parameterized normal equation

$$H(x,y,\xi):=F(x,\Pi_\Upsilon(y),\xi)+y-\Pi_\Upsilon(y)=0, \quad (x,\xi)\in\mathcal{X}\times\Xi, \tag{33}$$

in the sense that if $y(x,\xi)$ is a solution of (33), then $\bar{y}(x,\xi):=\Pi_\Upsilon(y(x,\xi))$ is a solution of the BVI problem, and conversely, if $\bar{y}(x,\xi)$ is a solution of the BVI, then $y(x,\xi):=\bar{y}(x,\xi)-F(x,\bar{y}(x,\xi),\xi)$ is a solution of (33). Consequently, we can reformulate (31) as

$$\min_{x\in\mathcal{X}} \quad \mathbb{E}\left[f(x,\Pi_\Upsilon(y(x,\xi)),\xi)\right], \tag{34}$$

where $y(x,\xi)$ is a measurable solution to the following equation:

$$H(x,y,\xi)=0. \tag{35}$$

Obviously, H defined in (33) is locally Lipschitz continuous and piecewise smooth with respect to x,y,ξ as Υ is a box. The nonsmoothness of Π_Υ may result in the ill-posedness of (33). Note also that since F is a P_0-function, the parametric BVI may have multiple solutions, and consequently (33) may have multiple solutions for every x and ξ. In what follows, we use a smoothed regularization method to deal with (33). First, we will use the Gabriel–Moré smoothing function to smooth Π_Υ [12], and consequently we get a smooth approximation of (33)

$$\hat{R}(x,v,y,\xi):=F(x,p(v,y),\xi)+y-p(v,y)=0, \quad (x,\xi)\in\mathcal{X}\times\Xi. \tag{36}$$

Here $p:\mathbb{R}^n\times\mathbb{R}^n\to\mathbb{R}^n$ is continuously differentiable, except at the point $(v,y)\in\mathbb{R}^n\times\mathbb{R}^n$, such that $v_i=0$ for some $i\in\{1,2,\ldots,n\}$, and for any $(v,y)\in\mathbb{R}^n\times\mathbb{R}^n$, $p(v,y)\in\Upsilon$. We set $v_i=\mu$ and apply the well-known *Tikhonov* regularization to $\hat{R}$

$$R(x,y,\xi,\mu):=\hat{R}(x,\mu e,y,\xi)+\mu y=F(x,p(\mu e,y),\xi)+y-p(\mu e,y)+\mu y, \tag{37}$$

where $e = (1, 1, \ldots, 1) \in \mathbb{R}^n$ and μ is a positive parameter.

In the following analysis, we use the well-known Chen–Harker–Kanzow–Smale (CHKS) smoothing function [7, 16] to smooth the components of $\Pi_\Upsilon(y)$:

$$\phi(\alpha, c, d, \beta) = (c + \sqrt{(c-\beta)^2 + 4\alpha^2})/2 + (d - \sqrt{(d-\beta)^2 + 4\alpha^2})/2,$$

where $(\alpha, \beta) \in \mathbb{R}_{++} \times \mathbb{R}$ and $(c, d) \in \mathbb{R} \times \mathbb{R}$. Then

$$\begin{aligned} p_i(\mu e, y) &:= \phi(\mu, a_i, b_i, y_i) \\ &= (a_i + \sqrt{(a_i - y_i)^2 + 4\mu^2})/2 + (b_i - \sqrt{(b_i - y_i)^2 + 4\mu^2})/2, \quad i = 1, \ldots, n. \end{aligned}$$

For any $\mu \in \mathbb{R}_{++}$ and $(x, \xi) \in \mathcal{X} \times \Xi$, since $p(\mu e, y)$ is continuously differentiable with respect to y, it follows that $R(x, y, \xi, \mu)$ is continuously differentiable with respect to x, y for almost every ξ. In what follows, we will verify that R defined in (37) satisfies Definition 2.1 for regularization functions. We first state the following result.

LEMMA 5.1. *Given $(x, \xi) \in \mathcal{X} \times \Xi$ and $\mu > 0$, let R be defined as in* (37). *Then the Jacobian $\nabla_y R(x, y, \xi, \mu)$ is nonsingular.*

Proof. First, we have

$$\nabla_y R(x, y, \xi, \mu) = \nabla_y F(x, p(\mu e, y), \xi) D(\mu, y) + \mu I + I - D(\mu, y),$$

where $D(\mu, y) = \operatorname{diag}(d_1(\mu, y), \ldots, d_n(\mu, y))$ and $d_i(\mu, y) = \partial p_i(\mu e, y)/\partial y_i \in [0, 1]$ for every $i \in \{1, \ldots, n\}$. Since $F(x, y, \xi)$ is a P_0-function with respect to y and $p(\mu e, y) \in \Upsilon$, then $\nabla_y F(x, p(\mu e, y), \xi)$ is a P_0-matrix. For $u \in \mathbb{R}^n$, let

$$[\nabla_y F(x, p(\mu e, y), \xi) D(\mu, y) + \mu I + I - D(\mu, y)]u = 0.$$

We claim that $D(\mu, y)u = 0$. Assume that $[D(\mu, y)u]_i \neq 0$ for any i; we then have

$$\begin{aligned} [D(\mu, y)u]_i [\nabla_y F(x, p(\mu e, y), \xi) D(\mu, y)u]_i &= -[D(\mu, y)u]_i [(\mu + 1)u - D(\mu, y)u]_i \\ &= -(\mu + 1) d_i(\mu, y) u_i^2 + d_i^2(\mu, y) u_i^2 < 0, \end{aligned}$$

which contradicts the definition of P_0-matrix of $\nabla_y F(x, p(\mu e, y), \xi)$. So, $D(\mu, y)u = 0$. Hence, $(\mu + 1)u = 0$, which derives $u = 0$. Thus, $\nabla_y R(x, y, \xi, \mu)$ is nonsingular. ∎

LEMMA 5.2. *Let $\{x^k\}$, $\{y^k\}$, $\{\xi^k\}$ be sequences in $\mathcal{X}$, $\mathbb{R}^n$, Ξ and let $\{\mu^k\}$ be a sequence in any closed subset of $(0, \mu_0)$ with $\{\|y^k\|\} \to \infty$ as $k \to \infty$, where μ_0 is a small positive number. Then $\|R(x^k, y^k, \xi^k, \mu^k)\| \to \infty$ as $k \to \infty$.*

See a detailed proof in the appendix.

By Lemmas 5.1 and 5.2, we are ready to show that function R constructed in (37) is a regularization of H.

PROPOSITION 5.3. *Let μ_0 be a small positive number. Function R defined in* (37) *is a regularization of H as defined in Definition* 2.1. *Moreover, $\tilde{y}$ is continuously differentiable on $\mathcal{X} \times \Xi \times (0, \mu_0)$.*

The proof is long. We move it to the appendix.

Based on the above discussions, we can convert the true problem (31) (or equivalently, (34)) to the following regularized program:

$$\min_{x \in \mathcal{X}} \quad \mathbb{E}\left[f(x, \Pi_\Upsilon(\tilde{y}(x, \xi, \mu)), \xi)\right], \tag{38}$$

where $\tilde{y}(x, \xi, \mu)$ uniquely solves $R(x, y, \xi, \mu) = 0$, $(x, \xi, \mu) \in \mathcal{X} \times \Xi \times (0, \mu_0)$.

In the next subsection, we will investigate a numerical method for solving (38) by using its SAA.

5.2. SAA program. In this subsection, we consider the SAA program of (38)

$$\begin{aligned} \min_{x\in\mathcal{X},\ y^1,\dots,y^N} \quad & \frac{1}{N}\sum_{i=1}^{N} f(x,\Pi_\Upsilon(y^i),\xi^i) \\ \text{s.t.} \quad & R(x,y^i,\xi^i,\mu)=0, \quad i=1,\dots,N, \end{aligned} \tag{39}$$

where μ is a small positive number. Analogous to the discussion in section 4, we can write down the GKKT conditions of (39) as follows:

$$\begin{cases} 0 \in \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \nabla_x f(x,\Pi_\Upsilon(y^i),\xi^i) + \sum_{i=1}^{N} \lambda^i \nabla_x R(x,y^i,\xi^i,\mu) + \mathcal{N}_{\mathcal{X}}(x), \\ 0 \in \dfrac{1}{N}\begin{pmatrix} \nabla_y f(x,\Pi_\Upsilon(y^1),\xi^1)\partial\Pi_\Upsilon(y^1) \\ \vdots \\ \nabla_y f(x,\Pi_\Upsilon(y^N),\xi^N)\partial\Pi_\Upsilon(y^N) \end{pmatrix} + \begin{pmatrix} \lambda^1\nabla_y R(x,y^1,\xi^1,\mu) \\ \vdots \\ \lambda^N\nabla_y R(x,y^N,\xi^N,\mu) \end{pmatrix}, \\ 0 = R(x,y^i,\xi^i,\mu), \quad i=1,\dots,N. \end{cases} \tag{40}$$

Following similar arguments as in section 4, we derive

$$\begin{aligned} 0 \in & \frac{1}{N}\sum_{i=1}^{N} \left[\nabla_x f(x,\Pi_\Upsilon(\tilde{y}(x,\xi^i,\mu)),\xi^i) + \Lambda^{\mathrm{reg}}(x,\xi^i,\mu)\nabla_x R(x,\tilde{y}(x,\xi^i,\mu),\xi^i,\mu)\right] \\ & + \mathcal{N}_{\mathcal{X}}(x), \end{aligned} \tag{41}$$

where

$$\Lambda^{\mathrm{reg}}(x,\xi^i,\mu) = -\nabla_y f(x,\Pi_\Upsilon(\tilde{y}(x,\xi^i,\mu)),\xi^i)\partial\Pi_\Upsilon(\tilde{y}(x,\xi^i,\mu))\nabla_y R(x,\tilde{y}(x,\xi^i,\mu),\xi^i,\mu)^{-1}.$$

Note that for any $(x,y,\xi)\in\mathcal{X}\times\mathbb{R}^n\times\Xi$ and $\mu>0$,

$$\nabla_x H(x,y,\xi) = \nabla_x R(x,y,\xi,\mu) = \nabla_x F(x,\Pi_\Upsilon(y),\xi)$$

and

$$\partial\Pi_\Upsilon(y) \subset \{M\in\mathbb{R}^{n\times n} \mid M=\mathrm{diag}(d_1,\dots,d_n),\ \ d_i\in[0,1]\}.$$

Obviously, $\partial\Pi_\Upsilon(y)$ is bounded for any $y\in\mathbb{R}^n$. Assume that $\lim_{\mu\downarrow 0}\tilde{y}(x,\xi,\mu)$ exists. Let $y(x,\xi)=\lim_{\mu\downarrow 0}\tilde{y}(x,\xi,\mu)$ on $\mathcal{X}\times\Xi$, and for $(x,\xi)\in\mathcal{X}\times\Xi$

$$\begin{aligned} \Lambda(x,\xi) := \mathrm{conv}\{\lambda(x,\xi)\in\mathbb{R}^n \mid & 0\in\nabla_y f(x,\Pi_\Upsilon(y(x,\xi)),\xi)\partial\Pi_\Upsilon(y(x,\xi)) \\ & + \lambda(x,\xi)\pi_y\partial H(x,y(x,\xi),\xi)\}. \end{aligned}$$

Following a similar argument as in Theorem 4.1, we derive convergence results for (39) below.

THEOREM 5.4. *Suppose that there exist a function $\kappa_4(\xi)$ and a constant $\mu_0>0$ such that for all $(x,\xi,\mu)\in\mathcal{X}\times\Xi\times(0,\mu_0)$*

$$\begin{aligned} & \max\{\|\nabla_x f(x,\Pi_\Upsilon(\tilde{y}(x,\xi,\mu),\xi))\|, \|\nabla_x F(x,\Pi_\Upsilon(\tilde{y}(x,\xi,\mu)),\xi)\|, \|\Lambda^{\mathrm{reg}}(x,\xi,\mu)\|\} \\ & \le \kappa_4(\xi) \end{aligned} \tag{42}$$

with $\mathbb{E}[\kappa_4(\xi)]<\infty$. Suppose that $\mathbb{E}[\kappa_4(\xi)(1+\kappa_4(\xi))]<\infty$. Then

(i) *for fixed $\mu > 0$, w.p.*1 *an accumulation point of the sequence of the generalized stationary points $\{x_N(\mu)\}$ of* (39) *satisfies*

$$0 \in \mathbb{E}\left[\hbar(x,\xi,\mu)\right] + \mathcal{N}_{\mathcal{X}}(x),$$

where

$$\hbar(x,\xi,\mu) := \nabla_x f(x, \Pi_{\Upsilon}(\tilde{y}(x,\xi,\mu)),\xi) + \Lambda^{\mathrm{reg}}(x,\xi,\mu)\nabla_x F(x,\Pi_{\Upsilon}(\tilde{y}(x,\xi,\mu)),\xi);$$

(ii) *if $\mu = \mu_N$, where $\mu_N \to 0$ as $N \to \infty$, and $\{x(\mu_N)\}$ is a sequence of generalized stationary points of* (39), *then w.p.*1 *an accumulation point of $\{x(\mu_N)\}$ satisfies*

$$0 \in \mathbb{E}[\mathcal{M}(x,\xi)] + \mathcal{N}_{\mathcal{X}}(x), \tag{43}$$

where

$$\mathcal{M}(x,\xi) := \nabla_x f(x, \Pi_{\Upsilon}(y(x,\xi)),\xi) + \Lambda(x,\xi)\nabla_x F(x,\Pi_{\Upsilon}(y(x,\xi)),\xi).$$

Note that the boundedness condition in (42) on $\|\nabla_x f(x,\Pi_{\Upsilon}(\tilde{y}(x,\xi,\mu)),\xi)\|$ and $\|\nabla_x F(x,\Pi_{\Upsilon}(\tilde{y}(x,\xi,\mu)),\xi)\|$ is satisfied if f and F are uniformly globally Lipschitz with respect to x. The boundedness condition on $\Lambda^{\mathrm{reg}}(x,\xi,\mu)$ is satisfied if $f(x,y,\xi)$ is uniformly globally Lipschitz with respect to y and $\pi_y \partial H(x,y(x,\xi),\xi)$ is uniformly nonsingular. In particular, if $H(x,y,\xi)$ is regular in the sense of [20] in y at $y(x,\xi)$, then $\pi_y \partial H(x,y(x,\xi),\xi)$ is nonsingular. See [20] for a detailed discussion in this regard.

Example 5.5. Consider the following stochastic mathematical program:

$$\min_{x\in\mathcal{X}} \quad \mathbb{E}\left[f(x,y(x,\xi),\xi)\right]. \tag{44}$$

Here $f : \mathbb{R}\times\mathbb{R}^2\times\mathbb{R}\to\mathbb{R}$ is given as

$$f(x,y,\xi) = 2(y_1 - \arctan y_1) + 4y_1^4 y_2/(1+y_1^2)^2 + 1 + x + \xi,$$

and $y(x,\xi)$ is any measurable solution of the following BVI problem:

$$F(x,y(x,\xi),\xi)^T(z - y(x,\xi)) \geq 0 \quad \forall z \in \Upsilon, \tag{45}$$

where $\Upsilon = \mathbb{R}^2_+$, $\mathcal{X} = [0,1]$, ξ can be any random variable that can take values on the interval $\Xi := [-1,-1/4]$, and $F(x,y,\xi) = (0, y_1 + y_2 + x + \xi - 1)^T$. Evidently, F is continuously differentiable and $F(x,\cdot,\xi)$ is a P_0-function for every $(x,\xi)\in\mathcal{X}\times\Xi$ and

$$\nabla_x F(x,y,\xi) = (0,1)^T. \tag{46}$$

Note that f is continuously differentiable on $\mathbb{R}\times\mathbb{R}^2\times\mathbb{R}$, $f'_x(x,y,\xi) = f'_\xi(x,y,\xi) = 1$, and

$$\nabla_y f(x,y,\xi) = \left(2y_1^2/(1+y_1^2) + 16y_1^3y_2/(1+y_1^2)^3,\quad 4y_1^4/(1+y_1^2)^2\right). \tag{47}$$

In this example, we have $H(x,y,\xi) = (y_1 - \max\{0,y_1\},\ \max\{0,y_1\} + y_2 + x + \xi - 1)$. It is not hard to obtain the solution set of the VI problem (45) as follows: for $(x,\xi)\in\mathcal{X}\times\Xi$,

$$\mathcal{Y}(x,\xi) := \{y\in\mathbb{R}^2_+ : y_1 \geq 1 - x - \xi,\ y_2 = 0\} \cup \{y\in\mathbb{R}^2_+ : y_1 + y_2 - 1 + x + \xi = 0\}.$$

Obviously, $\mathcal{Y}$ is a set-valued mapping on $\mathcal{X} \times \Xi$.

We now consider the regularization of the VI problem (45), in which given a regularization parameter, we expect to derive a unique solution function on $\mathcal{X} \times \Xi$. Note here that $\Upsilon = \mathbb{R}^2_+$, we have $a_i = 0$, and $b_i = \infty$, $i = 1, 2$. Then we get the ith component of the smoothing function $p_i(\mu e, y) = \phi(\mu, 0, \infty, y_i)$, $i = 1, 2$, $\mu > 0$, where ϕ is the reduced CHKS smoothing NCP function: $\phi(\alpha, 0, \infty, \beta) = (\sqrt{\beta^2 + 4\alpha^2} + \beta)/2, (\alpha, \beta) \in \mathbb{R}_{++} \times \mathbb{R}$. By definition, we have $p_i(\mu e, y) = (\sqrt{y_i^2 + 4\mu^2} + y_i)/2$, $i = 1, 2, \mu > 0$. And

$$R(x, y, \xi, \mu) = \begin{pmatrix} -(\sqrt{y_1^2 + 4\mu^2} + y_1)/2 + (1 + \mu)y_1 \\ (\sqrt{y_1^2 + 4\mu^2} + y_1)/2 + (1 + \mu)y_2 + x + \xi - 1 \end{pmatrix}.$$

Evidently, R is continuously differentiable on $\mathcal{X} \times \mathbb{R}^2 \times \Xi \times (0, \infty)$.

After some basic manipulations, we derive the unique solution of $R(x, y, \xi, \mu) = 0$ for any $\mu > 0$, $(x, \xi) \in \mathcal{X} \times \Xi$ as follows:

$$\tilde{y}(x, \xi, \mu) = \left(\sqrt{\mu/(\mu + 1)}, \quad (1 - x - \xi)/(1 + \mu) - \sqrt{\mu/(\mu + 1)} \right)^T.$$

Moreover, for any $x \in \mathcal{X}$ and $\mu > 0$, $\|\tilde{y}(x, \xi, \mu)\| \le \kappa_1(\xi)$ with $\mathbb{E}[\kappa_1(\xi)] < \infty$, where $\kappa_1(\xi) = 7 - \xi$, and

$$\|\tilde{y}(x'', \xi, \mu) - \tilde{y}(x', \xi, \mu)\| \le L(\xi)\|x'' - x'\| \qquad \text{for any } x'', x' \in \mathcal{X},$$

where $L(\xi)$ can be taken as any measurable positive function satisfying $1 \le \mathbb{E}[L(\xi)] < \infty$, say, $L(\xi) = 1 - \xi$. Also, $\lim_{\mu \downarrow 0} \tilde{y}(x, \xi, \mu) = y(x, \xi) = (0, 1 - x - \xi) \in \mathcal{Y}(x, \xi)$, $(x, \xi) \in \mathcal{X} \times \Xi$. Thereby, all conditions in Theorem 2.4 are satisfied. Obviously, $y(x, \xi)$ is measurable for every $x \in \mathcal{X}$ and Lipschitz continuous in x.

In addition, by Proposition 5.3, the regularization R defined above satisfies parts (i)–(iv) of Definition 2.1, where μ_0 can be chosen any small positive number.

Next, we investigate the boundedness condition (42). After some simple calculations, we can see that $\tilde{y}(x, \xi, \mu) \in \mathbb{R}^2_{++}$ for any $(x, \xi, \mu) \in \mathcal{X} \times \Xi \times (0, \hat{\mu})$, where $\hat{\mu} = (\sqrt{5} - 2)/4$. So, $\Pi_\Upsilon(\tilde{y}(x, \xi, \mu)) = \tilde{y}(x, \xi, \mu)$ and

$$\partial \Pi_\Upsilon(\tilde{y}(x, \xi, \mu)) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

On the other hand, we have

$$\nabla_y R(x, y, \xi, \mu)^{-1} = \varpi(y, \mu)^{-1} \begin{pmatrix} 1 + \mu & 0 \\ -\frac{1}{2}\left(1 + \frac{y_1}{\sqrt{y_1^2 + 4\mu^2}}\right) & -\frac{1}{2}\left(1 + \frac{y_1}{\sqrt{y_1^2 + 4\mu^2}}\right) + \mu \end{pmatrix},$$

where $\varpi(y, \mu) = (1 + \mu)[\frac{1}{2}(1 - y_1/\sqrt{y_1^2 + 4\mu^2}) + \mu]$. Then it follows that

$$\nabla_y R(x, \tilde{y}(x, \xi, \mu), \xi, \mu)^{-1} = \frac{1 + 2\mu}{2\mu(1 + \mu)^2} \begin{pmatrix} 1 + \mu & 0 \\ -\dfrac{1 + \mu}{1 + 2\mu} & \dfrac{2\mu^2 - 1}{1 + 2\mu} \end{pmatrix}.$$

In addition, we have

$$\begin{aligned}&\nabla_y f(x,\tilde{y}(x,\xi,\mu),\xi)\\&=\frac{2\mu}{2\mu+1}\left((1+8\sqrt{\mu(\mu+1)})\left(\frac{1-x-\xi-\sqrt{\mu(\mu+1)}}{(2\mu+1)^2}\right),\quad \frac{2\mu}{2\mu+1}\right).\end{aligned}$$

Then, for $\mu\in(0,\hat{\mu})$ and $(x,\xi)\in\mathcal{X}\times\Xi$, it follows that

$$\begin{aligned}\Lambda^{\rm reg}(x,\xi,\mu)&=-\nabla_y f(x,\Pi_\Upsilon(\tilde{y}(x,\xi,\mu)),\xi)\partial\Pi_\Upsilon(\tilde{y}(x,\xi,\mu))\nabla_y R(x,\tilde{y}(x,\xi,\mu),\xi,\mu)^{-1}\\&=-\left(\frac{8\sqrt{\mu}(1-x-\xi)}{\sqrt{1+\mu}(1+2\mu)^2}+\frac{1-6\mu-4\mu^2}{(1+\mu)(1+2\mu)^2},\quad \frac{2\mu(2\mu^2-1)}{(1+\mu)^2(1+2\mu)^2}\right).\end{aligned}$$

By applying some basic operations, we have $\|\Lambda^{\rm reg}(x,\xi,\mu)\|<\varrho(\xi)$ for any $(x,\mu)\in\mathcal{X}\times(0,\mu_0)$, where $\varrho(\xi):=\sqrt{2}+8\sqrt{2\hat{\mu}}(1-\xi)$. Assume that ξ follows a uniform distribution with parameters -1 and $-1/4$, i.e., $\xi\sim U(-1,-1/4)$. Then $\mathbb{E}[\varrho(\xi)]=\sqrt{2}+13\sqrt{2\hat{\mu}}<\infty$ and $\mathbb{E}[\varrho(\xi)(1+\varrho(\xi))]<\infty$. Hence, we can choose $\kappa_4(\xi)=\varrho(\xi)$ in Theorem 5.4. This, together with (46) and (47), shows that the boundedness condition (42) holds in this example. □

Note that in this example, one may ask why the objective function f is not chosen in a simpler form, say, a linear function of y rather than in a complex form as it stands. The answer is that we use this example not only to illustrate how regularization works for this particular SMPEC problem but also to demonstrate how the boundedness conditions (19) in Lemma 3.10 and (42) of Theorem 5.4 can be satisfied. In this particular example, if f is made linear in y, then we are not able to guarantee the boundedness of the set $\Lambda^{\rm reg}$, although this does not mean the method will not work.

6. Preliminary numerical results. We have carried out numerical tests on the regularized SAA scheme for stochastic problems with VI constraints. In this section, we report some preliminary numerical results. Such stochastic problems are artificially made by ourselves, since there are few test problems on SMPECs in the literature. The tests are carried out by implementing mathematical programming codes in MATLAB 6.5 installed in a PC with Windows XP operating system. We use the MATLAB built-in solver *fmincon* for solving the regularized SAA problems.

6.1. Estimating the optimal value of the regularized problem. The following methodology of constructing statistical lower and upper bounds was suggested in [28]. Given $\mu>0$, let $v(\mu)$ denote the optimal value of the regularized problem (38) and $\tilde{v}_N(\mu)$ the optimal value of (39). It is known [28] that $\mathbb{E}[\tilde{v}_N(\mu)]\le v(\mu)$. To estimate the expected value $\mathbb{E}[\tilde{v}_N(\mu)]$, we generate M independent samples of ξ, $\{\xi_j^1,\dots,\xi_j^N\}$, $j=1,\dots,M$, each of size N. For each sample j, solve the corresponding SAA problem (39), which can be written as

$$\begin{aligned}\min_{x\in\mathcal{X}\ y^1,\dots,y^N}\quad&\frac{1}{N}\sum_{i=1}^N f(x,\Pi_\Upsilon(y^i),\xi_j^i)\\\text{s.t.}\quad&R(x,y^i,\xi_j^i,\mu)=0,\quad i=1,\dots,N.\end{aligned}\tag{48}$$

Let $\tilde{v}_N^j(\mu)$, $j=1,\dots,M$, denote the corresponding optimal value of problem (48). Compute

$$L_{N,M}(\mu):=\frac{1}{M}\sum_{j=1}^M\tilde{v}_N^j(\mu),$$

which is an unbiased estimate of $\mathbb{E}[\tilde{v}_N(\mu)]$. Then $L_{N,M}(\mu)$ provides a statistical lower bound for $v(\mu)$. An estimate of variance of the estimator $L_{N,M}(\mu)$ can be computed as

$$s_L^2(M;\mu) := \frac{1}{M(M-1)} \sum_{j=1}^{M} \left(\tilde{v}_N^j(\mu) - L_{N,M}(\mu) \right)^2.$$

Let $v(x,\xi,\mu) = f(x, \Pi_\Upsilon(\tilde{y}(x,\xi,\mu)),\xi)$ and $\tilde{\vartheta}(x,\mu) = \mathbb{E}[v(x,\xi,\mu)]$. Then an upper bound for the optimal value $v(\mu)$ can be obtained by the fact that $\tilde{\vartheta}(\bar{x},\mu) \geq v(\mu)$ for any $\bar{x} \in \mathcal{X}$. Hence, by choosing $\bar{x}$ to be a near-optimal solution, for example, by solving one SAA problem and using an unbiased estimator of $\tilde{\vartheta}(\bar{x},\mu)$, we can obtain an estimate of an upper bound for $v(\mu)$. To do so, generate M' independent batches of samples: $\{\xi_j^1,\ldots,\xi_j^{N'}\}$, $j = 1,\ldots,M'$, each of size N'. For $x \in \mathcal{X}$, let $\tilde{v}_{N'}^j(x,\mu) := \frac{1}{N'} \sum_{i=1}^{N'} v(x,\xi_j^i,\mu)$. Then $\mathbb{E}[\tilde{v}_{N'}^j(x,\mu)] = \tilde{\vartheta}(x,\mu)$. Compute

$$U_{N',M'}(\bar{x};\mu) := \frac{1}{M'} \sum_{j=1}^{M'} \tilde{v}_{N'}^j(\bar{x},\mu),$$

which is an unbiased estimate of $\tilde{\vartheta}(\bar{x},\mu)$. So, $U_{N',M'}(\bar{x};\mu)$ is an estimate of an upper bound on $v(\mu)$. An estimate of variance of the estimator $U_{N',M'}(\bar{x},\mu)$ can be computed as

$$s_U^2(\bar{x},M';\mu) := \frac{1}{M'(M'-1)} \sum_{j=1}^{M'} \left(\tilde{v}_{N'}^j(\bar{x},\mu) - U_{N',M'}(\bar{x},\mu) \right)^2.$$

Note that in this part, for each $j = 1,\ldots,M'$ and $i = 1,\ldots,N'$, we need to solve the following repeated subproblems:

$$\begin{array}{ll} \min & f(\bar{x}, \Pi_\Upsilon(y), \xi_j^i) \\ \text{s.t.} & R(\bar{x}, y, \xi_j^i, \mu) = 0; \end{array}$$

then the corresponding optimal value is $v(\bar{x},\xi_j^i,\mu)$. Hence, we can obtain $\tilde{v}_{N'}^j(\bar{x},\mu)$, $U_{N',M'}(\bar{x};\mu)$, and $s_U^2(\bar{x},M';\mu)$. Note that, in practice, we may choose $\bar{x}$ to be any of the solutions of the M regularized SAA problems (48) by generating independent samples $\{\xi_j^1,\ldots,\xi_j^N\}$, $j = 1,\ldots,M$. In fact, we will use $\bar{x}_N^j$, the *best* optimal solution which estimates the smallest optimal value $v(\mu)$, to compute the upper bound estimates, and the optimality gap.

Using the lower bound estimate and the objective function value estimate of the optimal value, $v(\mu)$, of the first stage regularized problem as discussed above, we compute an estimate of the optimality gap of the solution $\bar{x}$ and the corresponding estimated variance as follows:

$$Gap_{N,M,N',M'}(\bar{x}) := U_{N',M'}(\bar{x};\mu) - L_{N,M}(\mu), \quad S_{\text{Gap}}^2 := s_L^2(M;\mu) + s_U^2(\bar{x},M';\mu).$$

6.2. Preliminary computational results. In the following test problem, we choose different values for the regularization parameter μ and sample sizes N, M, N', and M'. We report the lower and upper bounds, $L_{N,M}$ and $U_{N',M'}$, of $v(\mu)$, the sample variances, s_L, s_U, and the estimate of the optimality gap, *Gap*, of the solution candidate $\bar{x}_N^j$, the variance of the gap estimator S_{Gap}.

TABLE 1
Summary of lower and upper bounds on $v(\mu)$, the optimality gap.

μ	N	M	N'	M'	$L_{N,M}$	s_L	$\bar{x}_N^j$	$U_{N',M'}$	s_U	Gap	S_{Gap}
10^{-3}	200	10	200	10	.7345	.0118	.4928	.7632	.0138	.0287	.0181
10^{-4}	200	10	200	10	.7657	.0142	.5056	.7719	.0150	.0062	.0207
10^{-5}	200	10	200	10	.7749	.0138	.4948	.7841	.0127	.0092	.0188
10^{-3}	300	10	300	10	.7295	.0104	.4837	.7406	.0096	.0111	.0141
10^{-4}	300	10	300	10	.7506	.0018	.4988	.7574	.0118	.0069	.0167
10^{-5}	300	10	300	10	.7668	.0120	.5071	.7727	.0149	.0059	.0191

Example 6.1. Consider the following problem:

$$\begin{array}{ll} \min & \mathbb{E}[x^2 + y_2(x,\xi)^2] \\ \text{s.t.} & 0 \le x \le 1, \end{array} \tag{49}$$

where $y(x,\xi)$ is a solution of the following complementarity problem, which is a special case of VI problems:

$$0 \le F(x,y,\xi) \perp y \ge 0, \qquad F(x,y,\xi) = (0, y_1 + y_2 + x + \xi - 1)^T,$$

where ξ is a random variable with truncated standard normal distribution on $[-1, 1]$. Using the regularization scheme, we can convert the above problem into the following problem:

$$\begin{array}{ll} \min\limits_{x,y} & \mathbb{E}[x^2 + (\max(0, y_2))^2] \\ \text{s.t.} & R(x,y,\xi,\mu) = 0, \quad 0 \le x \le 1, \end{array}$$

where μ is a small positive parameter tending to 0 and $R(x,y,\xi,\mu)$ is given in Example 5.5. Note that the limit of the corresponding unique solution function $\tilde{y}(x,\xi,\mu)$ of $R(x,y,\xi,\mu) = 0$ equals $y(x,\xi) := (0, 1 - x - \xi)$. After basic operations, we can derive the optimal solution of problem (49) associated with $y(x,\xi)$ as $x^* = 0.5$, and the optimal value is $f^* = 0.77454$ (obtained from Maple). The test results are displayed in Table 1.

The results show that both optimal solutions and values of the regularized SAA problems approximate those of the true problem very well as sample size increases and the regularization parameter is driven to zero. More numerical tests are needed to evaluate the performance of the proposed method, but this is beyond the scope of this paper.

Appendix. *Proof of Lemma* 5.2. We first define an index set $\mathcal{I}_0^\infty := \{i \mid \{y_i^k\}$ is unbounded, $i = 1, \ldots, n\}$. By assumption, $\mathcal{I}_0^\infty$ is nonempty, and for all $i \in \mathcal{I}_0^\infty$, $|y_i^k| \to \infty$ as $k \to \infty$. In the following analysis, we will consider the following cases: (i) $a = -\infty$, $b = \infty$; (ii) $a \in \mathbb{R}^n$, $b = \infty$; (iii) $a = \infty$, $b \in \mathbb{R}^n$; and (iv) $a, b \in \mathbb{R}^n$.

Case (i). Since $a = -\infty$, $b = \infty$, we have $p(\mu e, y) = y$. Then $R(x,y,\xi,\mu) = F(x,y,\xi) + \mu y$. We now construct a bounded sequence $\{w^k\}$ by letting $w_i^k = 0$ if $i \in \mathcal{I}_0^\infty$ and $w_i^k = y_i^k$ otherwise. Since F is a P_0-function in y, hence for any k,

$$\begin{aligned} 0 &\le \max_{1 \le i \le n} (y_i^k - w_i^k)[F_i(x^k, y^k, \xi^k) - F_i(x^k, w^k, \xi^k)] \\ &= \max_{i \in \mathcal{I}_0^\infty} y_i^k [F_i(x^k, y^k, \xi^k) - F_i(x^k, w^k, \xi^k)] \\ &= y_{i_0}^k [F_{i_0}(x^k, y^k, \xi^k) - F_{i_0}(x^k, w^k, \xi^k)]. \end{aligned} \tag{50}$$

Here i_0 denotes an index in $\mathcal{I}_0^\infty$ at which the maximum value is attained. Without loss of generality, we may assume that the above index i_0 is independent of k. Since $\mathcal{X}$ and Ξ are compact, and $\{w^k\}$ is bounded, hence $\{F_{i_0}(x^k, w^k, \xi^k)\}$ is bounded by virtue of the continuity of F_{i_0}. We now consider two cases: $y_{i_0}^k \to \infty$; $y_{i_0}^k \to -\infty$. In the former case, it follows from (50) that $\{F_{i_0}(x^k, y^k, \xi^k)\}$ does not tend to $-\infty$. Since $\{\mu^k\}$ is contained in a closed interval of $(0, \mu_0)$, hence $F_{i_0}(x^k, y^k, \xi^k) + \mu^k y_{i_0}^k \to \infty$, which implies that $\|F(x^k, y^k, \xi^k) + \mu^k y^k\| \to \infty$. Similarly, in the latter case, we have that $\{F_{i_0}(x^k, y^k, \xi^k)\}$ does not tend to ∞ by (50). Thereby, $F_{i_0}(x^k, y^k, \xi^k) + \mu^k y_{i_0}^k \to -\infty$. Thus, in both cases, we have $\|R(x^k, y^k, \xi^k, \mu^k)\| \to \infty$.

Case (ii). Note that in this case

$$p_i(\mu e, y) = (a_i + \sqrt{(a_i - y_i)^2 + 4\mu^2} + y_i)/2, \quad i = 1, \ldots, n.$$

Then it is not hard to show that for each i, and any sequences $\{y_i^l\}$ and $\{\mu^l\}$ satisfying $y_i^l \to \infty$ and μ^l being in a closed subset of $(0, \mu_0)$ for all l, we have

$$\lim_{l\to\infty} [y_i^l - p_i(\mu^l e, y^l)] = 0. \tag{51}$$

Let

$$\mathcal{I}_+^\infty := \{i \in \mathcal{I}_0^\infty \mid \{y_i^k\} \to \infty\} \quad \text{and} \quad \mathcal{I}_-^\infty := \{i \in \mathcal{I}_0^\infty \mid \{y_i^k\} \to -\infty\}.$$

We now consider two cases: $\mathcal{I}_+^\infty = \emptyset$; $\mathcal{I}_+^\infty \neq \emptyset$. In Case (i), we have $\{y_i^k\} \to -\infty$ for all $i \in \mathcal{I}_0^\infty$. Then it is easy to show that $\lim_{k\to\infty} p_i(\mu^k e, y^k) = a_i$ for all $i \in \mathcal{I}_0^\infty$. Thus, $\{p(\mu^k e, y^k)\}$ is bounded. Noticing the boundedness of $\{x^k\}$ and $\{\xi^k\}$ and by virtue of the continuity of F_i, $i \in \mathcal{I}_0^\infty$, it follows that

$$\left|\big(R(x^k, y^k, \xi^k, \mu^k)\big)_i\right| = \left|F_i(x^k, p(\mu^k e, y^k), \xi^k) - p_i(\mu^k e, y^k) + y_i^k + \mu^k y_i^k\right| \to \infty.$$

In Case (ii), evidently, $\lim_{k\to\infty} p_i(\mu^k e, y^k) = \infty$ or a_i for $i \in \mathcal{I}_+^\infty$ or $\mathcal{I}_-^\infty$. We now define a sequence $\{v^k\}$ with $v_i^k := 0$ if $i \in \mathcal{I}_+^\infty$; $v_i^k := p_i(\mu^k e, y^k)$ if $i \in \mathcal{I}_-^\infty$; $v_i^k := p_i(\mu^k e, y^k)$ if $i \notin \mathcal{I}_0^\infty$. Based on the above arguments, evidently, $\{v^k\}$ is bounded. By the notion of P_0-function, we have

$$\begin{aligned} 0 &\le \max_{1\le i\le n} (p_i(\mu^k e, y^k) - v_i^k)[F_i(x^k, p(\mu^k e, y^k), \xi^k) - F_i(x^k, v^k, \xi^k)] \\ &= \max_{i\in\mathcal{I}_+^\infty} p_i(\mu^k e, y^k)[F_i(x^k, p(\mu^k e, y^k), \xi^k) - F_i(x^k, v^k, \xi^k)] \\ &= p_j(\mu^k e, y^k)[F_j(x^k, p(\mu^k e, y^k), \xi^k) - F_j(x^k, v^k, \xi^k)], \end{aligned} \tag{52}$$

where $j \in \mathcal{I}_+^\infty$ such that the maximum value is attained at j, without loss of generality, which is assumed to be independent of k. By assumption, F_j is continuous, and note that $\{x^k\}$, $\{\xi^k\}$, $\{v^k\}$ are bounded; hence $\{F_j(x^k, v^k, \xi^k)\}$ is bounded as well. In addition, since $p_j(\mu^k e, y^k) \to \infty$, thus by (52), $\{F_j(x^k, p(\mu^k e, y^k), \xi^k)\}$ does not tend to $-\infty$. Thereby, $F_j(x^k, p(\mu^k e, y^k), \xi^k) + \mu^k y_j^k \to \infty$. On the other hand, note that

$$\begin{aligned} \big(R(x^k, y^k, \xi^k, \mu^k)\big)_j &= F_j(x^k, p(\mu^k e, y^k), \xi^k) + y_j^k - p_j(\mu^k e, y^k) + \mu^k y_j^k \\ &= F_j(x^k, p(\mu^k e, y^k), \xi^k) + \mu^k y_j^k + y_j^k - p_j(\mu^k e, y^k). \end{aligned}$$

By (51), $y_j^k - p_j(\mu^k e, y^k) \to 0$. So, $\big(R(x^k, y^k, \xi^k, \mu^k)\big)_j \to \infty$. Therefore,

$$\|R(x^k, y^k, \xi^k, \mu^k)\| \to \infty.$$

Case (iii). In this case, the arguments are similar to Case (ii). Here we omit them for brevity.

Case (iv). Note that for any $i \in \mathcal{I}_0^\infty$, $\lim_{k\to\infty} p_i(\mu^k e, y^k) = \lim_{k\to\infty} \phi(\mu^k, a_i, b_i, y_i^k)$ equals b_i if $y_i^k \to \infty$ or a_i if $y_i^k \to -\infty$. Then $\{p(\mu^k e, y^k)\}$ is bounded; thereby, $\{F_i(x^k, p(\mu^k e, y^k), \xi^k)\}$ is bounded as well for $i \in \mathcal{I}_0^\infty$. Hence,

$$\left|\left(R(x^k, y^k, \xi^k, \mu^k)\right)_i\right| = \left|F_i(x, p(\mu^k e, y^k), \xi^k) + y_i^k - p_i(\mu^k e, y^k) + \mu^k y_i^k\right| \to \infty.$$

Thereby, $\|R(x^k, y^k, \xi^k, \mu^k)\| \to \infty$. □

Proof of Proposition 5.3. Note that R is continuous on $\mathcal{X}\times\mathbb{R}^n\times\Xi\times[0,\mu_0]$. We now check parts (i)–(iv) in Definition 2.1. Obviously, part (i) holds, since $p(0, y) = \Pi_\Upsilon(y)$ for any $y \in \mathbb{R}^n$. By [20, Theorem 3.1], $p(\mu e, y)$ is continuously differentiable at any $(\mu, y) \in \mathbb{R}_{++} \times \mathbb{R}^n$. Then R is continuously differentiable on $\mathcal{X} \times \mathbb{R}^n \times \Xi \times (0, \mu_0]$. Note also that $H(x, y, \xi)$ is piecewise smooth; hence R is piecewise smooth on $\mathcal{X} \times \mathbb{R}^n \times \Xi \times [0, \mu_0]$. Thereby, part (ii) holds.

We now consider part (iii). Note that

$$\pi_x\partial R(x, y, \xi, \mu) = \partial_x R(x, y, \xi, \mu) = \nabla_x F(x, p(\mu e, y), \xi),$$

$$\pi_x\partial H(x, y, \xi) = \partial_x H(x, y, \xi) = \nabla_x F(x, \Pi_\Upsilon(y), \xi)$$

and $\lim_{\mu\downarrow 0} p(\mu e, y) = \Pi_\Upsilon(y)$. Then we have

$$\lim_{\mu\downarrow 0} \partial_x R(x, y, \xi, \mu) = \lim_{\mu\downarrow 0} \nabla_x F(x, p(\mu e, y), \xi) = \nabla_x F(x, \Pi_\Upsilon(y), \xi) = \partial_x H(x, y, \xi)$$

for any $(x, y, \xi) \in \mathcal{X} \times \mathbb{R}^n \times \Xi$. On the other hand, noticing that $\pi_y\partial H(x, y, \xi) = \partial_y H(x, y, \xi) = (\nabla_y F(x, \Pi_\Upsilon(y), \xi) - I)\partial\Pi_\Upsilon(y) + I$, and by Lemma 5.1, $\pi_y\partial R(x, y, \xi, \mu) = \partial_y R(x, y, \xi, \mu) = \nabla_y R(x, y, \xi, \mu) = (\nabla_y F(x, p(\mu e, y), \xi) - I)D(\mu, y) + \mu I + I$. Hence, to show $\lim_{\mu\downarrow 0} \pi_y\partial R(x, y, \xi, \mu) \subset \pi_y\partial H(x, y, \xi)$, it suffices to prove $\lim_{\mu\downarrow 0} D(\mu, y) \subset \partial\Pi_\Upsilon(y)$. Note that for any $y \in \mathbb{R}^n$,

$$\partial\Pi_\Upsilon(y) = \begin{bmatrix} \partial\Pi_{[a_1,b_1]}(y_1) & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & \partial\Pi_{[a_n,b_n]}(y_n) \end{bmatrix},$$

where $\partial\Pi_{[a_i,b_i]}(y_i)$ equals 0 if $y_i \in (-\infty, a_i) \cup (b_i, \infty)$; 1 if $y_i \in (a_i, b_i)$; $[0, 1]$ if $y_i = a_i$ or b_i. Then, after some basic manipulations, $\lim_{\mu\downarrow 0} d_i(\mu, y) = \lim_{\mu\downarrow 0} \partial p_i(\mu e, y)/\partial y_i$ equals 0 if $y_i \in (-\infty, a_i) \cup (b_i, \infty)$; 1 if $y_i \in (a_i, b_i)$; 1/2 if $y_i = a_i$ or b_i. Hence, $\lim_{\mu\downarrow 0} d_i(\mu, y) \subset \partial\Pi_{[a_i,b_i]}(y_i)$ for each i; thereby, $\lim_{\mu\downarrow 0} D(\mu, y) \subset \partial\Pi_{[a,b]}(y) (= \partial\Pi_\Upsilon(y))$. Thus, $\lim_{\mu\downarrow 0} \nabla_y R(x, y, \xi, \mu) \subset \partial_y H(x, y, \xi)$ for any $(x, y, \xi) \in \mathcal{X} \times \mathbb{R}^n \times \Xi$. So, part (iii) holds.

Finally, we prove part (iv). Define a mapping $G : \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^k \times \mathbb{R} \to \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^k \times \mathbb{R}$ by $G(x, y, \xi, \mu) := (x, R(x, y, \xi, \mu), \xi, \mu)$. Then G is continuously differentiable on $\mathcal{X}\times\mathbb{R}^n\times\Xi\times(0, \mu_0)$. We first show that G is a diffeomorphism on $\mathcal{X}\times\mathbb{R}^n\times\Xi\times(0, \mu_0)$; that is, G has a differentiable inverse function on $\mathcal{X} \times \mathbb{R}^n \times \Xi \times (0, \mu_0)$. It is well known that a necessary and sufficient condition for function G to be a diffeomorphism is the nonsingularity of the Jacobian ∇G at every point (x, y, ξ, μ) and the closedness of G; that is, the image $G(S)$ of any closed set $S \subset \mathcal{X} \times \mathbb{R}^n \times \Xi \times (0, \mu_0)$ is closed.

We now prove the closedness of G. Let S be a closed subset of $\mathcal{X}\times\mathbb{R}^n\times\Xi\times(0,\mu_0)$. Assume for the sake of a contradiction that $G(S)$ is not closed. Then there exists a convergent sequence $\{w^k\}\subset G(S)$ such that $\lim_{k\to\infty} w^k = w^0$, but $w^0\notin G(S)$. By definition, there exists a sequence $\{z^k\}$ with $z^k=(x^k,y^k,\xi^k,\mu^k)\in S$ such that $w^k = G(z^k)$. We consider two cases: (i) $\{z^k\}$ is bounded; (ii) $\{z^k\}$ is unbounded. In case (i), obviously, there exists a convergent subsequence, $\{z^{k_l}\}$, of $\{z^k\}$ with $\{z^{k_l}\}$ tending to z^0. Hence, $z^0\in S$ given the closeness of S. Thus, $\lim_{l\to\infty} w^{k_l} = \lim_{l\to\infty} G(z^{k_l}) = G(z^0)$. Clearly, $G(z^0)\in G(S)$. Since $w^{k_l}\to w^0$, then $w^0 = G(z^0)\in G(S)$, which leads to a contradiction as desired. In case (ii), without loss of generality, we assume that $\|z^k\|\to\infty$ as $k\to\infty$. With the help of the compactness of $\mathcal{X}$ and Ξ, there exists a subsequence $\{z^{k_l}\}$ of $\{z^k\}$ such that $\{x^{k_l}\}$, $\{\xi^{k_l}\}$, and $\{\mu^{k_l}\}$ are bounded, while $\|y^{k_l}\|\to\infty$ as $l\to\infty$. Then, by Lemma 5.2, $\lim_{l\to\infty}\|R(x^{k_l},y^{k_l},\xi^{k_l},\mu^{k_l})\| = \infty$. Thus,

$$\lim_{l\to\infty}\|w^{k_l}\| = \lim_{l\to\infty}\|G(z^{k_l})\| = \infty$$

by noticing $w^{k_l} = G(z^{k_l}) = \left(x^{k_l}, R(x^{k_l},y^{k_l},\xi,\mu^{k_l}),\xi^{k_l},\mu^{k_l}\right)$. This contradicts the fact that $\lim_{l\to\infty} w^{k_l} = w^0$. Therefore, G is closed on $\mathcal{X}\times\mathbb{R}^n\times\Xi\times(0,\mu_0)$.

Next, we prove the nonsingularity of ∇G. By Lemma 5.1, we can easily see that $\nabla G(x,y,\mu,\xi)$ is nonsingular at any point $(x,y,\xi,\mu)\in\mathcal{X}\times\mathbb{R}^n\times\Xi\times(0,\mu_0)$. Hence, G is a diffeomorphism on $\mathcal{X}\times\mathbb{R}^n\times\Xi\times(0,\mu_0)$. Let G^{-1} denote its inverse function. For any $(x,y,\xi,\mu)\in\mathcal{X}\times\mathbb{R}^n\times\Xi\times(0,\mu_0)$, we then have $(G^{-1}(x,y,\xi,\mu))_x = x$, $(G^{-1}(x,y,\xi,\mu))_\xi = \xi$, and $(G^{-1}(x,y,\xi,\mu))_\mu = \mu$. Furthermore, for any $(p,t,q)\in\mathcal{X}\times\Xi\times(0,\mu_0)$, equation $G(x,y,\xi,\mu) = (p,0,t,q)$ has a unique solution $(x,y,\xi,\mu) = G^{-1}(p,0,t,q)$. Clearly, $x=p$, $\xi=t$, and $\mu=q$. Let $y=\tilde{y}(p,t,q) := \left(G^{-1}(p\ 0\ t\ q)\right)_y$. By virtue of the arbitrariness of p, t, and q, we obtain the unique solution of $\tilde{y}$ defined on $\mathcal{X}\times\Xi\times(0,\mu_0)$, which satisfies $R(x,\tilde{y}(x,\xi,\mu),\xi,\mu)=0$ for $(x,\xi,\mu)\in\mathcal{X}\times\Xi\times(0,\mu_0)$. Thereby, part (iv) is satisfied. In addition, note that G is continuously differentiable on $\mathcal{X}\times\mathbb{R}^n\times\Xi\times(0,\mu_0)$ by assumption. This leads to the continuous differentiability of $\tilde{y}$ immediately.

In conclusion, based on the above arguments, function R defined in (37) satisfies Definition 2.1. This completes the proof. ∎

Acknowledgments. The authors would like to thank Alexander Shapiro for his constructive comments on an earlier version of this paper. They would also like to thank two anonymous referees for insightful comments which led to a significant improvement of the paper and the associate editor, Andrzej Ruszczyński, for organizing a quick and quality review.

REFERENCES

[1] Z. Artstein and R. A. Vitale, *A strong law of large numbers for random compact sets*, Ann. Probability, 3 (1975), pp. 879–882.

[2] Z. Artstein and S. Hart, *Law of large numbers for random sets and allocation processes*, Math. Oper. Res., 6 (1981), pp. 485–492.

[3] F. Bastin, C. Cirillo, and P. Toint, *Convergence theory for nonconvex stochastic programming with an application to mixed logit*, Math. Program., 108 (2006), pp. 207–234.

[4] S. I. Birbil, G. Gürkan, and O. Listes, *Simulation-Based Solution of Stochastic Mathematical Programs with Complementarity Constraints: Sample-Path Analysis*, working paper, Center for Economic Research, Tilburg University, Tilburg, The Netherlands, 2004.

[5] J. Birge and F. Louveaux, *Introduction to Stochastic Programming*, Springer-Verlag, New York, 1997.

[6] J. F. Bonnans and A. Shapiro, *Perturbation Analysis of Optimization Problems*, Springer-Verlag, New York, 2000.
[7] B. Chen and P. T. Harker, *A non-interior-point continuation method for linear complementarity problems*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 1168–1190.
[8] X. Chen, L. Qi, and D. Sun, *Global and superlinear convergence of the smoothing Newton's method and its application to general box constrained variational inequalities*, Math. Comp., 67 (1998), pp. 519–540.
[9] S. Christiansen, M. Patriksson, and L. Wynter, *Stochastic bilevel programming in structural optimization*, Struct. Multidiscip. Optim., 21 (2001), pp. 361–371.
[10] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.
[11] F. Facchinei and J.-S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Springer-Verlag, New York, 2003.
[12] S. A. Gabriel and J. J. Moré, *Smoothing of mixed complementarity problems*, in Complemenarity and Variational Problems: State of the Art, M. C. Ferris and J. S. Pang, eds., SIAM, Philadelphia, 1997, pp. 105–116.
[13] J. L. Higle and S. Sen, *Stochastic Decomposition: A Statistical Method for Large Scale Stochastic Linear Programming*, Kluwer Academic Publishers, Boston, 1996.
[14] J.-B. Hiriart-Urruty, *Refinements of necessary optimality conditions in nondifferentiable programming* I, Appl. Math. Optim., 5 (1979), pp. 63–82.
[15] T. Homem-de-Mello, *Estimation of derivatives of nonsmooth performance measures in regenerative systems*, Math. Oper. Res., 26 (2001), pp. 741–768.
[16] C. Kanzow, *Some noninterior continuation methods for linear complementarity problems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 851–868.
[17] G.-H. Lin, X. Chen, and M. Fukushima, *Smoothing Implicit Programming Approaches for Stochastic Mathematical Programs with Linear Complementarity Constraints*, http://www. amp.i.kyoto-u.ac.jp/tecrep/index-e.html (2003).
[18] M. Patriksson and L. Wynter, *Stochastic mathematical programs with equilibrium constraints*, Oper. Res. Lett., 25 (1999), pp. 159–167.
[19] 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.
[20] L. Qi, D. Sun, and G. Zhou, *A new look at smoothing Newton methods for nonlinear complementarity problems and box constrained variational inequalities*, Math. Program., 87 (2000), pp. 1–35.
[21] D. Ralph and H. Xu, *Implicit smoothing and its application to optimization with piecewise smooth equality constraints*, J. Optim. Theory Appl., 124 (2005), pp. 673–699.
[22] S. M. Robinson, *Analysis of sample-path optimization*, Math. Oper. Res., 21 (1996), pp. 513–528.
[23] R. T. Rockafellar and R.J.-B. Wets, *Stochastic convex programming: Kuhn-Tucker conditions*, J. Math. Econom., 2 (1975), pp. 349–370.
[24] R. T. Rockafellar and R.J.-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
[25] R. Y. Rubinstein and A. Shapiro, *Discrete Events Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Methods*, John Wiley and Sons, New York, 1993.
[26] A. Ruszczyński, *Decomposition methods*, in Stochastic Programming, Handbooks Oper. Res. Management Sci. 10, A. Ruszczyński and A. Shapiro, eds., North–Holland, Amsterdam, 2003, pp. 141–211.
[27] A. Ruszczyński and A. Shapiro, *Stochastic Programming*, Handbooks Oper. Res. Management Sci. 10, North–Holland, Amsterdam, 2003.
[28] T. Santoso, S. Ahmed, M. Goetschalckx, and A. Shapiro, *A stochastic programming approach for supply chain network design under uncertainty*, European J. Oper. Res., 167 (2005), pp. 96–115.
[29] S. Scholtes, *Introduction to Piecewise Smooth Equations*, Habilitation, University of Karlsruhe, Karlsruhe, Germany, 1994.
[30] A. Shapiro, *Stochastic mathematical programs with equilibrium constraints*, J. Optim. Theory Appl., 128 (2006), pp. 223–243.
[31] A. Shapiro and T. Homem-de-Mello, *On the rate of convergence of optimal solutions of Monte Carlo approximations of stochastic programs*, SIAM J. Optim., 11 (2000), pp. 70–86.
[32] A. Shapiro and H. Xu, *Stochastic Mathematical Programs with Equilibrium Constraints, Modeling and Sample Average Approximation*, http://www.optimization-online.org/DB_HTML/2005/01/1046.html (2005).
[33] R.J.-B. Wets, *Stochastic programming*, in Optimization, Handbooks Oper. Res. Management Sci. 1, G. L. Nemhauser, A. Rinnooy, and M. Todd, eds., North–Holland, Amsterdam,

1989, pp. 573–629.
[34] H. Xu, *An MPCC approach for stochastic Stackelberg-Nash-Cournot equilibrium*, Optimization, 54 (2005), pp. 27–57.
[35] H. Xu, *An implicit programming approach for a class of stochastic mathematical programs with complementarity constraints*, SIAM J. Optim., 16 (2006), pp. 670–696.
[36] H. Xu and F. Meng, *Convergence analysis of sample average approximation methods for a class of stochastic mathematical programs with equality constraints*, Math. Oper. Res., to appear.

SIAM J. OPTIM.
Vol. 17, No. 3, pp. 920–942
© 2006 Society for Industrial and Applied Mathematics

GLOBAL OPTIMIZATION OF POLYNOMIALS USING GRADIENT TENTACLES AND SUMS OF SQUARES*

MARKUS SCHWEIGHOFER†

Abstract. We consider the problem of computing the global infimum of a real polynomial f on $\mathbb{R}^n$. Every global minimizer of f lies on its gradient variety, i.e., the algebraic subset of $\mathbb{R}^n$ where the gradient of f vanishes. If f attains a minimum on $\mathbb{R}^n$, it is therefore equivalent to look for the greatest lower bound of f on its gradient variety. Nie, Demmel, and Sturmfels proved recently a theorem about the existence of sums of squares certificates for such lower bounds. Based on these certificates, they find arbitrarily tight relaxations of the original problem that can be formulated as semidefinite programs and thus be solved efficiently. We deal here with the more general case when f is bounded from below but does not necessarily attain a minimum. In this case, the method of Nie, Demmel, and Sturmfels might yield completely wrong results. In order to overcome this problem, we replace the gradient variety by larger semialgebraic subsets of $\mathbb{R}^n$ which we call gradient tentacles. It now gets substantially harder to prove the existence of the necessary sums of squares certificates.

Key words. global optimization, polynomial, preorder, sum of squares, semidefinite programming

AMS subject classifications. Primary, 13J30, 90C26; Secondary, 12Y05, 13P99, 14P10, 90C22

DOI. 10.1137/050647098

1. Introduction.

Throughout this article, $\mathbb{N} := \{1, 2, \dots\}$, $\mathbb{R}$, and $\mathbb{C}$ denote the sets of natural, real, and complex numbers, respectively. We fix $n \in \mathbb{N}$ and consider real polynomials in n variables $\bar{X} := (X_1, \dots, X_n)$. These polynomials form a commutative ring

$$\mathbb{R}[\bar{X}] := \mathbb{R}[X_1, \dots, X_n].$$

1.1. The problem.

We consider the problem of computing good approximations for the global infimum

$$f^* := \inf\{f(x) \mid x \in \mathbb{R}^n\} \in \mathbb{R} \cup \{-\infty\}$$

of a polynomial $f \in \mathbb{R}[\bar{X}]$. Since f^* is the greatest lower bound of f, it is equivalent to compute

$$f^* = \sup\{a \in \mathbb{R} \mid f - a \geq 0 \text{ on } \mathbb{R}^n\} \in \mathbb{R} \cup \{-\infty\}. \tag{1}$$

To solve this hard problem, it has become a standard approach to approximate f^* by exchanging in (1) the nonnegativity constraint

$$f - a \geq 0 \quad \text{on } \mathbb{R}^n \tag{2}$$

by a computationally more feasible condition and analyze the error caused by this substitution. Typically, the choice of this replacement is related to the interplay between (globally) nonnegative polynomials, sums of squares of polynomials, and semidefinite optimization (also called semidefinite programming).

*Received by the editors December 8, 2005; accepted for publication (in revised form) May 17, 2006; published electronically October 24, 2006. This work was supported by Deutsche Forschungsgemeinschaft, grant "Barrieren."

http://www.siam.org/journals/siopt/17-3/64709.html

†Universität Konstanz, Fachbereich Mathematik und Statistik, 78457 Konstanz, Germany (Markus.Schweighofer@uni-konstanz.de).

1.2. Method based on the fact that every sum of squares of polynomials is nonnegative (Shor [Sho], Shor and Stetsyuk [SS], Parrilo and Sturmfels [PS], et al.). We start with the most basic ideas concerning these connections which can be found in greater detail in the references just cited. A first try is to replace condition (2) by the constraint

$$f - a \text{ is a sum of squares in the polynomial ring } \mathbb{R}[\bar{X}], \tag{3}$$

since every sum of squares in $\mathbb{R}[\bar{X}]$ is obviously nonnegative on $\mathbb{R}^n$.

The advantage of (3) over (2) is that sums of squares of polynomials can be nicely parametrized. Fix a column vector v whose entries are a basis of the vector space $\mathbb{R}[\bar{X}]_d$ of all real polynomials of degree $\leq d$ in n variables ($d \in \mathbb{N}_0 := \{0\} \cup \mathbb{N}$). This vector has a certain length $k = \dim \mathbb{R}[\bar{X}]_d$. It is easy to see that the map from the vector space $S\mathbb{R}^{k\times k}$ of symmetric $k \times k$-matrices to $\mathbb{R}[\bar{X}]_{2d}$ defined by $M \mapsto v^T M v$ is surjective. Using the spectral theorem for symmetric matrices, it is not hard to prove that a polynomial $f \in \mathbb{R}[\bar{X}]_{2d}$ is a sum of squares in $\mathbb{R}[\bar{X}]$ if and only if $f = v^T M v$ for some positive semidefinite matrix $M \in S\mathbb{R}^{k\times k}$. Use the following remark, which is an easy exercise (write the polynomials as sums of their homogeneous parts).

Remark 1. In any representation $f = \sum_i g_i^2$ of a polynomial $f \in \mathbb{R}[\bar{X}]_{2d}$ as a sum of squares $g_i \in \mathbb{R}[\bar{X}]$, we have necessarily $\deg g_i \leq d$.

The described parametrization shows that the modified problem (where we exchange (2) by (3)), i.e., the problem to compute

$$f^{\text{sos}} := \sup\{a \in \mathbb{R} \mid f - a \text{ is a sum of squares in } \mathbb{R}[\bar{X}]\} \in \mathbb{R} \cup \{-\infty\}, \tag{4}$$

can be written as a *semidefinite optimization problem* (also called semidefinite program or SDP for short), i.e., as the problem of minimizing (or maximizing) an affine linear function on the intersection of the cone of positive semidefinite matrices with an affine subspace in $S\mathbb{R}^{k\times k}$. For solving SDPs, there exist very good numerical algorithms, perhaps almost as good as for linear optimization problems. Linear optimization can be seen as the restriction of semidefinite optimization to diagonal matrices, i.e., a method to minimize an affine linear function on the intersection of the cone $\mathbb{R}^k_{\geq 0}$ with an affine subspace of $\mathbb{R}^k$. Speaking very vaguely, most concepts from linear optimization carry over to semidefinite optimization because every symmetric matrix can be diagonalized. We refer readers, for example, to [Tod] for an introduction to semidefinite programming.

Whereas computing f^* as defined in (1) is a very hard problem, it is relatively easy to compute (numerically to a given precision) f^{sos} defined in (4). Of course, the question of how f^* and f^{sos} are related arises. Since (3) implies (2), it is clear that $f^{\text{sos}} \leq f^*$. The converse implication (and thus $f^{\text{sos}} = f^*$) holds in some cases: A globally nonnegative polynomial

- in one variable or
- of degree at most two or
- in two variables of degree at most four

is a sum of squares of polynomials. We refer readers to [Rez] for an overview of these and related old facts. However, recently Blekherman has shown in [Ble] that for fixed degree $d \geq 4$ and high number of variables n only a very small portion (in some reasonable sense) of the globally nonnegative polynomials of degree at most d in n variables are sums of squares. In particular, f^{sos} will often differ from f^*. For example, the *Motzkin polynomial*

$$M := X^2Y^2(X^2 + Y^2 - 3Z^2) + Z^6 \in \mathbb{R}[X, Y, Z] \tag{5}$$

is nonnegative but not a sum of squares (see [Rez, PS]). We have $M^* = 0$ but $M^{\mathrm{sos}} = -\infty$. The latter follows from the fact that M is homogeneous and not a sum of squares by the following remark applied to $f := M - a$ for $a \in \mathbb{R}$ (which can again be proved easily by considering homogeneous parts).

Remark 2. If f is a sum of squares in $\mathbb{R}[\bar{X}]$, then so is the highest homogeneous part (the leading form) of f.

We see that the basic problem with this method (computing f^{sos} by solving an SDP and hoping that f^{sos} is close to f^*) is that polynomials positive on $\mathbb{R}^n$ in general do not have a representation as a sum of squares, a fact that Hilbert already knew.

1.3. The Positivstellensatz. In the 17th of his famous of 23 problems, Hilbert asked whether every (globally) nonnegative (real) polynomial (in several variables) was a sum of squares of *rational functions*. Artin answered this question affirmatively in 1926, and today there exist numerous refinements of his solution. One of them is the *Positivstellensatz* (in analogy to Hilbert's Nullstellensatz). It is often attributed to Stengle [Ste], who clearly deserves credit for finding it independently and making it widely known. However, Prestel [PD, section 4.7] recently discovered that Krivine [Kri] knew the result about 10 years earlier in 1964. Here we state only the following special case of the Positivstellensatz.

THEOREM 3 (Krivine). *For every $f \in \mathbb{R}[\bar{X}]$, the following are equivalent:*

(i) *$f > 0$ on $\mathbb{R}^n$.*

(ii) *There are sums of squares s and t in $\mathbb{R}[\bar{X}]$ such that $sf = 1 + t$.*

By this theorem, we have, of course, that f^* is the supremum over all $a \in \mathbb{R}$ such that there are sums of squares $s, t \in \mathbb{R}[\bar{X}]$ with $s(f - a) = 1 + t$. When one tries to write this as an SDP, there are two obstacles.

First, each SDP involves matrices of a fixed (finite) size. But with matrices of a fixed size, we can parametrize sums of squares only up to a certain degree. We need therefore to impose a degree restriction on s and t. There are no (at least up to now) *practically relevant* degree bounds that could guarantee that such a restriction would not affect the result. We refer readers to the tremendous work [Scd] of Schmid on degree bounds. This first obstacle, namely the question of degrees of the sums of squares, will accompany us throughout the article. The answer will always be to model the problem not as a single SDP but as a whole *sequence* of SDPs, each SDP corresponding to a certain degree restriction. As one solves one SDP after the other, the degree restriction gets less restrictive, and one hopes for fast convergence of the optimal values of the SDPs to f^*. For newcomers in the field, it seems at first glance unsatisfactory having to deal with a whole sequence of SDPs rather than a single SDP. But, after all, it is only natural that a very hard problem cannot be modeled by an SDP of a reasonable size so that one has to look for good *relaxations* of the problem which can be dealt with more easily and to which the techniques of mathematical optimization can be applied.

The second obstacle is much more severe. It is the fact that the unknown polynomial $s \in \mathbb{R}[\bar{X}]$ is multiplied with the unknown $a \in \mathbb{R}$ on the left-hand side of the constraint $s(f - a) = 1 + t$. This makes the formulation as an SDP (even after having imposed a restriction on the degree of s and t) impossible (or at least highly nonobvious). Of course, if one *fixes* $a \in \mathbb{R}$ and a degree bound $2d$ for s and t, then the question of whether there exist sums of squares s and t of degree at most $2d$ such that $s(f - a) = 1 + t$ is equivalent to the feasibility of an SDP. But this plays (at least currently) only a role as a criterion that *might help to decide* whether a certain fixed (or guessed) $a \in \mathbb{R}$ is a strict lower bound of f. We refer readers to [PS] for

more details. What one needs are representation theorems for positive polynomials that are better suited for optimization than the Positivstellensatz (even if they are sometimes less aesthetic).

1.4. "Big ball" method proposed by Lasserre [L1]. In the last 15 years, a lot of progress has been made in proving existence of sums of squares certificates which can be exploited for optimization (although most of the new results were obtained without having in mind the application in optimization which has been established more recently). The first breakthrough was perhaps Schmüdgen's theorem [Sch, Corollary 3], all of whose proofs use the Positivstellensatz. In this article, we will prove a generalization of Schmüdgen's theorem, namely Theorem 9. In [L1], Lasserre uses the following special case of Schmüdgen's theorem which has already been proved by Cassier [Cas, Théorème 4] and which can even be derived easily from [Kri, Théorème 12].

THEOREM 4 (Cassier). *For $f \in \mathbb{R}[\bar{X}]$ and $R \geq 0$, the following are equivalent:*

(i) *$f \geq 0$ on the closed ball centered at the origin of radius R.*

(ii) *For all $\varepsilon > 0$, there are sums of squares s and t in $\mathbb{R}[\bar{X}]$ such that*

$$f + \varepsilon = s + t(R^2 - \|\bar{X}\|^2).$$

Here and in the following, we use the notation

$$\|\bar{X}\|^2 := X_1^2 + \cdots + X_n^2 \in \mathbb{R}[\bar{X}].$$

Similar to section 1.2, it can be seen that for any fixed $d \in \mathbb{N}_0$, computing the supremum over all $a \in \mathbb{R}$ such that $f - a = s + t(R^2 - \|\bar{X}\|^2)$ for some sums of squares $s, t \in \mathbb{R}[\bar{X}]$ of degree at most $2d$ amounts to solving an SDP. Therefore one gets a sequence of SDPs parametrized by $d \in \mathbb{N}_0$. Theorem 4 can now be interpreted as a convergence result; namely, the sequence of optimal values of these SDPs converges to the minimum of f on the closed ball around the origin with radius R. If one has a polynomial $f \in \mathbb{R}[\bar{X}]$ attaining a minimum on $\mathbb{R}^n$ and for which one knows, moreover, a big ball on which this minimum is attained, this method is good for computing f^*. Of course, if one does not know such a big ball in advance, one might choose larger and larger R. But at the same time one might have to choose a bigger and bigger degree restriction $d \in \mathbb{N}_0$, and it is not really clear how to get a sequence of SDPs that converges to f^*.

1.5. Lasserre's high order perturbation method [L2]. Recently, Lasserre used in [L2] a theorem of Nussbaum from operator theory to prove the following result that can be exploited in a similar way for global optimization of polynomials.

THEOREM 5 (Lasserre). *For every $f \in \mathbb{R}[\bar{X}]$, the following are equivalent:*

(i) *$f \geq 0$ on $\mathbb{R}^n$.*

(ii) *For all $\varepsilon > 0$, there is $r \in \mathbb{N}_0$ such that*

$$f + \varepsilon \sum_{i=1}^{n} \sum_{k=0}^{r} \frac{X_i^{2k}}{k!} \text{ is a sum of squares in } \mathbb{R}[\bar{X}].$$

Note that (ii) implies that $f(x) + \varepsilon \sum_{i=1}^{n} \exp(x_i) \geq 0$, for all $x \in \mathbb{R}^n$ and $\varepsilon > 0$, which in turn implies (i). In condition (ii), r depends on ε and f. Using real algebra and model theory, Netzer showed that in fact r depends only on ε, n, the degree of f, and a bound on the size of the coefficients of f [Net, LN].

1.6. "Gradient perturbation" method proposed by Jibetean and Laurent [JL]. The most standard idea for finding the minimum of a function everybody knows from calculus is to compute critical points, i.e., the points where the gradient vanishes. It is a natural question whether the power of classical differential calculus can be combined with the relatively new ideas using sums of squares. Fortunately, it can and the rest of the article will be about how to merge both concepts, sums of squares and differential calculus.

If a polynomial $f \in \mathbb{R}[\bar{X}]$ attains a minimum in $x \in \mathbb{R}^n$, i.e., $f(x) \leq f(y)$ for all $y \in \mathbb{R}^n$, then the gradient ∇f of f vanishes at x, i.e., $\nabla f(x) = 0$. However, there are polynomials that are bounded from below on $\mathbb{R}^n$ and yet do not attain a minimum on $\mathbb{R}^n$. The simplest example is perhaps

$$f := (1 - XY)^2 + Y^2 \in \mathbb{R}[X, Y] \tag{6}$$

for which we have $f > 0$ on $\mathbb{R}^n$ but $f^* = 0$, since $\lim_{x\to\infty} f(x, \frac{1}{x}) = 0$. In the following,

$$(\nabla f) := \left(\frac{\partial f}{\partial X_1}, \ldots, \frac{\partial f}{\partial X_n}\right) \subseteq \mathbb{R}[\bar{X}]$$

denotes the ideal generated by the partial derivatives of f in $\mathbb{R}[\bar{X}]$. We call this ideal the *gradient ideal* of f.

Without going into details, the basic idea of Jibetean and Laurent in [JL] is again to apply a perturbation to f. Instead of adding a truncated exponential like Lasserre, they just add $\varepsilon \sum_{i=1}^n X_i^{2(d+1)}$ for small $\varepsilon > 0$ when $\deg f = 2d$. If $f > 0$ on $\mathbb{R}^n$, then the perturbed polynomial $f_\varepsilon := f + \varepsilon \|\bar{X}\|^{2(d+1)}$ is again a sum of squares but this time only modulo its gradient ideal (∇f_ε). In this case, this is quite easy to prove, since it turns out that this ideal will be zero-dimensional; i.e., $\mathbb{R}[\bar{X}]/(\nabla f_\varepsilon)$ is a finite-dimensional real algebra. We will later see in Theorems 6 and 46 that this finite dimensionality is not needed for the sums of squares representation. But the work of Jibetean and Laurent exploits the finite dimensionality in many ways. We refer readers to [JL] for details.

1.7. "Gradient variety" method by Nie, Demmel, and Sturmfels [NDS]. The two perturbation methods just sketched rely on introducing very small coefficients in a polynomial. These small coefficients might lead to SDPs which are hard to solve because of numerical instability. It is therefore natural to think of another method which avoids perturbation entirely. Nie, Demmels, and Sturmfels considered, for a polynomial $f \in \mathbb{R}[\bar{X}]$, its *gradient variety*

$$V(\nabla f) := \{x \in \mathbb{C}^n \mid \nabla f(x) = 0\}.$$

This is the algebraic variety corresponding to the radical of the gradient ideal (∇f). It can be shown that a polynomial $f \in \mathbb{R}[\bar{X}]$ is constant on each irreducible component of the gradient variety (see [NDS] or use an unpublished algebraic argument of Scheiderer based on Kähler differentials). This is the key to show that a polynomial $f \in \mathbb{R}[\bar{X}]$ nonnegative on its gradient variety is a sum of squares modulo its gradient ideal in the case where the ideal is radical. In the general case where the gradient ideal is not necessarily radical, the same thing still holds for polynomials *positive* on their gradient variety. The following is essentially [NDS, Theorem 9] (confer also the recent work [M2]). We will later prove a generalization of this theorem as a by-product. See Corollary 47.

THEOREM 6 (Nie, Demmel, and Sturmfels). *For every $f \in \mathbb{R}[\bar{X}]$ attaining a minimum on $\mathbb{R}^n$, the following are equivalent:*

(i) $f \geq 0$ *on* $\mathbb{R}^n$.
(ii) $f \geq 0$ *on* $V(\nabla f) \cap \mathbb{R}^n$.
(iii) *For all* $\varepsilon > 0$, *there exists a sum of squares* s *in* $\mathbb{R}[\bar{X}]$ *such that*

$$f + \varepsilon \in s + (\nabla f).$$

Moreover, (ii) *and* (iii) *are equivalent for all* $f \in \mathbb{R}[\bar{X}]$.

For each degree restriction $d \in \mathbb{N}_0$, the problem of computing the supremum over all $a \in \mathbb{R}$ such that

$$f - a = s + p_1 \frac{\partial f}{\partial X_1} + \cdots + p_n \frac{\partial f}{\partial X_n}$$

for some sum of squares s in $\mathbb{R}[\bar{X}]$ and polynomials $p_1, \ldots, p_n$ of degree at most d can be expressed as an SDP. Theorem 6 shows that the optimal values of the corresponding sequence of SDPs (indexed by d) tend to f^*, provided that f attains a minimum on $\mathbb{R}^n$. However, if f does not attain a minimum on $\mathbb{R}^n$, the computed sequence still tends to the infimum of f on its gradient variety, which might, however, now be very different from f^*. Take, for example, the polynomial f from (6). It is easy to see that $V(\nabla f) = \{0\}$, and therefore the method computes $f(0) = 1$ instead of $f^* = 0$. In [NDS, section 7], the authors write:

> "This paper proposes a method for minimizing a multivariate polynomial $f(x)$ over its gradient variety. We assume that the infimum f^* is attained. This assumption is nontrivial, and we do not address the (important and difficult) question of how to verify that a given polynomial $f(x)$ has this property."

1.8. Our "gradient tentacle" method. The reason why the method just described might fail is that the global infimum of a polynomial $f \in \mathbb{R}[\bar{X}]$ is not always a *critical value* of f, i.e., a value that f takes on at least on one of its critical points in $\mathbb{R}^n$. Now there is a well-established notion of *generalized critical values* which includes also the *asymptotic critical values* (a kind of critical value at infinity we will introduce in Definition 12).

In this article, we will replace the real part $V(\nabla f) \cap \mathbb{R}^n$ of the gradient variety by several larger semialgebraic sets on which the partial derivatives do not necessarily vanish but get very small far away from the origin. These semialgebraic sets often look like tentacles, and that is what we will call them. All tentacles we will consider are defined by a single polynomial inequality that depends only on the polynomial

$$\|\nabla f\|^2 := \left(\frac{\partial f}{\partial X_1}\right)^2 + \cdots + \left(\frac{\partial f}{\partial X_n}\right)^2$$

and expresses that this polynomial gets very small. Given a polynomial f for which one wants to compute f^*, the game will consist in finding a tentacle such that two things will hold at the same time:

- There exist suitable sums of squares certificates for nonnegativity on the tentacle.
- The infimum of f on $\mathbb{R}^n$ and on the tentacle coincide.

One can imagine that these two properties are hardly compatible. Taking $\mathbb{R}^n$ as a tentacle would, of course, ensure the second condition, but we have discussed in section 1.2 that the first one would be badly violated. The other extreme would be to take the empty set as a tentacle. Then the first condition would trivially be

satisfied, whereas the second would fail badly. How we will roughly be able to find the balancing act between the two requirements is as follows: The second condition will be satisfied by known nontrivial theorems about asymptotic behavior of polynomials at infinity. The existence of suitable sums of squares certificates will be based on the author's (real) algebraic work [Sr1] on iterated rings of bounded elements (also called real holomorphy rings).

1.9. Contents of the article. The article is organized as follows. In section 2, we prove a general sums of squares representation theorem which generalizes Schmüdgen's theorem, mentioned in section 1.4. This representation theorem is interesting in itself and will be used in the subsequent sections. In section 3, we introduce a gradient tentacle (see Definition 17) which is defined by the polynomial inequality

$$\|\nabla f\|^2 \|\bar{X}\|^2 \leq 1.$$

We call this gradient tentacle *principal*, since we can prove that it does the job in a large number of cases (see Theorem 25), and there is hope that it works in fact for all polynomials $f \in \mathbb{R}[\bar{X}]$ bounded from below. Indeed, we have not found any counterexamples (see Open Problem 33). In case this hope were disappointed, we present in section 4 a collection of other gradient tentacles (see Definition 41) defined by the polynomial inequalities

$$\|\nabla f\|^{2N}(1+\|\bar{X}\|^2)^{N+1} \leq 1 \qquad (N \in \mathbb{N}).$$

Their advantage is that if $f \in \mathbb{R}[\bar{X}]$ is bounded from below and N is large enough for this particular f, then we can prove that the corresponding tentacle does the job (see Theorems 46 and 50). We call these tentacles higher gradient tentacles, since the degree of the defining inequality gets unfortunately high when N gets big, which certainly has negative consequences for the complexity of solving the SDPs arising from these tentacles. However, if f attains a minimum on $\mathbb{R}^n$, then any choice of $N \in \mathbb{N}$ will be good. Conclusions are drawn in section 5.

2. The sums of squares representation. In this section, we prove the important sums of squares representation theorem we will need in the following sections. It is a generalization of Schmüdgen's Positivstellensatz (see [PD, Sch]), which is also of independent interest. Schmüdgen's result is not to be confused with the (classical) Positivstellensatz we described in the introduction. The connection between the two is that all known proofs of Schmüdgen's result use the classical Positivstellensatz. Our result, Theorem 9, is much harder to prove than Schmüdgen's result. Its proof relies on the theory of iterated *rings of bounded elements* (also called real holomorphy rings) described in [Sr1].

DEFINITION 7. *For any polynomial $f \in \mathbb{R}[\bar{X}]$ and subset $S \subseteq \mathbb{R}^n$, the set $R_\infty(f,S)$ of* asymptotic values *of f on S consists of all $y \in \mathbb{R}$ for which there exists a sequence $(x_k)_{k\in\mathbb{N}}$ of points $x_k \in S$ such that*

$$\lim_{k\to\infty} \|x_k\| = \infty \qquad \text{and} \qquad \lim_{k\to\infty} f(x_k) = y. \tag{7}$$

We now recall the important notion of a preordering of a commutative ring. Except in the proof of Theorem 9, we need this concept only for the ring $\mathbb{R}[\bar{X}]$.

DEFINITION 8. *Let A be a commutative ring (with 1). A subset $T \subseteq A$ is called a* preordering *if it contains all squares f^2 of elements $f \in A$ and is closed under*

addition and multiplication. The preordering generated *by* $g_1, \dots, g_m \in A$

$$T(g_1, \dots, g_m) = \left\{ \sum_{\delta \in \{0,1\}^m} s_\delta g_1^{\delta_1} \dots g_m^{\delta_m} \mid s_\delta \text{ is a sum of squares in } A \right\} \tag{8}$$

is by definition the smallest preordering containing $g_1, \dots, g_m$.

If $g_1, \dots, g_m \in \mathbb{R}[\bar{X}]$ are polynomials, then the elements of $T(g_1, \dots, g_m)$ have obviously the geometric property that they are nonnegative on the (basic closed semi-algebraic) set S they define by (9). The next theorem is a partial converse. Namely, if a polynomial satisfies on S some stronger geometric condition, then it lies necessarily in $T(g_1, \dots, g_m)$. In case that S is compact, the conditions (a) and (b) below are empty and the theorem is Schmüdgen's Positivstellensatz (see [PD, Sch]). The more general version we need here is quite hard to prove.

THEOREM 9. *Let* $f, g_1, \dots, g_m \in \mathbb{R}[\bar{X}]$ *and set*

$$S := \{x \in \mathbb{R}^n \mid g_1(x) \geq 0, \dots, g_m(x) \geq 0\}. \tag{9}$$

Suppose that

(a) f *is bounded on* S,

(b) f *has only finitely many asymptotic values on* S *and all of these are positive, i.e.,* $R_\infty(f, S)$ *is a finite subset of* $\mathbb{R}_{>0}$, *and*

(c) $f > 0$ *on* S.

Then $f \in T(g_1, \dots, g_m)$.

Proof. Write $R_\infty(f, S) = \{y_1, \dots, y_s\} \subseteq \mathbb{R}_{>0}$ and consider the polynomial

$$h := \prod_{i=1}^{s} (f - y_i).$$

This polynomial is "on S small at infinity" by which we mean that for every $\varepsilon > 0$ there exists $k \in \mathbb{N}$ such that for all $x \in S$ with $\|x\| \geq k$, we have $|h(x)| < \varepsilon$.

To show this, assume the contrary. Then there exists $\varepsilon > 0$ and a sequence $(x_k)_{k \in \mathbb{N}}$ of points $x_k \in S$ with $\lim_{k \to \infty} \|x_k\| = \infty$ and

$$|h(x_k)| \geq \varepsilon \qquad \text{for all } k \in \mathbb{N}. \tag{10}$$

Because the sequence $(f(x_k))_{k \in \mathbb{N}}$ is bounded by hypothesis (a), we find an infinite subset $I \subseteq \mathbb{N}$ such that the subsequence $(f(x_k))_{k \in I}$ converges. The limit must be one of the asymptotic values of f on S, i.e., $\lim_{k \in I, k \to \infty} f(x_k) = y_i$ for some $i \in \{1, \dots, s\}$. Using (a), it follows that $\lim_{k \in I, k \to \infty} h(x_k) = 0$, contradicting (10).

Let $A := (\mathbb{R}[\bar{X}], T)$, where $T := T(g_1, \dots, g_m)$. The set

$$H'(A) := \{p \in \mathbb{R}[\bar{X}] \mid N \pm p \in T \text{ for some } N \in \mathbb{N}\}$$

is a subring of A (see, e.g, [Sr1, Definition 1.2]). We endow $H'(A)$ with the preordering $T' := T \cap H'(A)$ and consider it also as a preordered ring. By [Sr1, Corollary 3.7], the smallness of h at infinity proved above is equivalent to $h \in S_\infty(A)$ in the notation of [Sr1]. By [Sr1, Corollary 4.17], we have $S_\infty(A) \subseteq H'(A)$ and consequently $h \in H'(A)$. The advantage of $H'(A)$ over A is that its preordering is Archimedean, i.e., $T' + \mathbb{Z} = H'(A)$. According to an old criterion, for an element to be contained in an Archimedean preordering (see, for example, [PD, Proposition 5.2.3 and Lemma 5.2.7]

or [Sr1, Theorem 1.3]), our claim $f \in T'$ follows if we can show that $\varphi(f) > 0$ for all ring homomorphisms $\varphi : H'(A) \to \mathbb{R}$ with $\varphi(T') \subseteq \mathbb{R}_{\geq 0}$. For all such homomorphisms possessing an extension $\bar{\varphi} : A \to \mathbb{R}$ with $\bar{\varphi}(T) \subseteq \mathbb{R}_{\geq 0}$, this follows from hypothesis (c) because it is easy to see that such an extension $\bar{\varphi}$ must be evaluation $p \mapsto p(x)$ in the point $x := (\bar{\varphi}(X_1), \ldots, \bar{\varphi}(X_n)) \in S$. Using the theory in [Sr1], we will see that the only possibility for such a φ not to have such an extension $\bar{\varphi}$ is that $\varphi(h) = 0$. Then we will be done, since $\varphi(h) = 0$ implies $\varphi(f) = y_i > 0$ for some i. We have used here that $f \in H'(A)$, which follows from $h \in H'(A)$, since $H'(A)$ is integrally closed in A (see [Sr1, Theorem 5.3]).

So let us now use [Sr1]. By [Sr1, Corollary 3.7 and Theorem 4.18], the smallness of h at infinity means that

$$A_h = H'(A)_h,$$

where we deal on both sides of this equation with the localization of a preordered ring by the element h (see [Sr1, pages 24 and 25]). If $\varphi : H'(A) \to \mathbb{R}$ is a ring homomorphism with $\varphi(T') \subseteq \mathbb{R}_{\geq 0}$ and $\varphi(h) \neq 0$, then φ extends to a ring homomorphism $\tilde{\varphi} : A_h = H'(A)_h \to \mathbb{R}$ with $\tilde{\varphi}(T_h) = \tilde{\varphi}(T'_h) \subseteq \mathbb{R}_{\geq 0}$. Then $\bar{\varphi} := \tilde{\varphi}|_A$ is the desired extension of φ. □

Example 10. Consider the polynomials

$$h_N := 1 - Y^N (1+X)^{N+1} \in \mathbb{R}[X,Y] \qquad (N \in \mathbb{N}) \tag{11}$$

in two variables. We fix $N \in \mathbb{N}$ and apply Theorem 9 with $f = h_{N+1}$, $m = 3$, $g_1 = X$, $g_2 = Y$, and $g_3 = h_N$. The set S defined by the g_i as in (9) is a subset of the first quadrant which is bounded in the Y-direction but unbounded in the X-direction. Of course, we have $0 \leq h_N \leq 1$ and

$$0 \leq Y(1+X) \leq \frac{1}{\sqrt[N]{1+X}} \qquad \text{on } S$$

showing that 0 is the only asymptotic value of

$$1 - h_{N+1} = (1 - h_N) Y (1+X)$$

on S and therefore $R_\infty(h_{N+1}, S) = \{1\}$. It follows also that $0 \leq h_{N+1} \leq 1$ on S. By Theorem 9, we obtain

$$h_{N+1} + \varepsilon \in T(X, Y, h_N) \tag{12}$$

for all $\varepsilon > 0$.

The following lemma shows that (12) holds even for $\varepsilon = 0$, a fact that does not follow from Theorem 9. This lemma will be interesting later to compare the quality of certain SDP relaxations (see Proposition 49). In its proof, we will explicitly construct a representation of h_{N+1} as an element of $T(X, Y, h_N)$. Only part of this explicit representation will be needed in the following, namely an explicit polynomial $g \in T(X,Y)$ such that $h_{N+1} \in T(X,Y) + g h_N \subseteq T(X, Y, h_N)$. This explains the formulation of the statement. Theorem 9 will not be used in the proof but gave us good hope before we had the proof. The role of Theorem 9 in this article is above all to prove Theorems 25 and 46.

LEMMA 11. *For the polynomials h_N defined by* (11)*, we have*

$$h_{N+1} - \left(1 + \frac{1}{N}\right) Y(1+X) h_N \in T(X,Y).$$

Proof. For a new variable Z,

$$\begin{aligned}(Z-1)^2\sum_{k=0}^{N-1}(N-k)Z^k &= (Z-1)^2\left(N\sum_{k=0}^{N-1}Z^k - Z\sum_{k=1}^{N-1}kZ^{k-1}\right)\\ &= (Z-1)^2\left(N\frac{Z^N-1}{Z-1} - Z\frac{\partial}{\partial Z}\left(\frac{Z^N-1}{Z-1}\right)\right)\\ &= N(Z-1)(Z^N-1) - Z((Z-1)NZ^{N-1} - (Z^N-1))\\ &= Z^{N+1} - (N+1)Z + N.\end{aligned}$$

Specializing Z to $z := Y(1+X)$, we have therefore

$$\begin{aligned}Nh_{N+1} - (N+1)zh_N &= N(1 - z^{N+1}(1+X)) - (N+1)z(1 - z^N(1+X))\\ &= z^{N+1}X + (z^{N+1} - (N+1)z + N)\\ &= z^{N+1}X + (z-1)^2\sum_{k=0}^{N-1}(N-k)z^k \in T(X,Y).\end{aligned}$$

Dividing by $N = (\sqrt{N})^2$ yields our claim. □

3. The principal gradient tentacle. In this section, we associate with every polynomial $f \in \mathbb{R}[\bar{X}]$ a gradient tentacle which is a subset of $\mathbb{R}^n$ containing the real part of the gradient variety of f and defined by a single polynomial inequality whose degree is not more than twice the degree of f. The infimum of any polynomial $f \in \mathbb{R}[\bar{X}]$ bounded from below on $\mathbb{R}^n$ will coincide with the infimum on its principal gradient tentacle (see Theorem 19). Under some technical assumption (see Definition 20) which is not known to be necessary (see Open Problem 33), we prove a sums of squares certificate for nonnegativity of f on its principal gradient tentacle which is suitable for optimization purposes. This representation theorem (Theorem 25) is of independent interest, and its proof is mainly based on the nontrivial representation theorem from the previous section and a result of Parusiński on the behavior of polynomials at infinity [P1, Theorem 1.4]. In section 3.2, we outline how to get a sequence of SDPs growing in size whose optimal values tend to f^* for any f satisfying the conditions of Theorem 25 (or perhaps for any f with $f^* > -\infty$ if the answer to Open Problem 33 is yes). In sections 3.3 and 3.4, we give a MATLAB code for the sums of squares optimization toolboxes YALMIP [Löf] and SOSTOOLS [PPS] that produces and solves these SDP relaxations. This short and simple code is meant for readers who have little experience with such toolboxes and want nevertheless to try our proposed method on their own. In section 3.5, we provide simple examples which have been calculated using the YALMIP code from section 3.3.

We start by recalling the concept of asymptotic critical values developed by Rabier in his 1997 milestone paper [Rab]. For simplicity, we stay in the setting of real polynomials right from the beginning (though part of this theory make sense in a much broader context).

DEFINITION 12. *Suppose $f \in \mathbb{R}[\bar{X}]$. The set $K_0(f)$ of* critical values *of f consists of all $y \in \mathbb{R}$ for which there exists $x \in \mathbb{R}^n$ such that $\nabla f(x) = 0$ and $f(x) = y$. The set $K(f)$ of* generalized critical values *of f consists of all $y \in \mathbb{R}$ for which there exists a sequence $(x_k)_{k\in\mathbb{N}}$ in $\mathbb{R}^n$ such that*

$$\lim_{k\to\infty}\|\nabla f(x_k)\|(1+\|x_k\|) = 0 \qquad \text{and} \qquad \lim_{k\to\infty} f(x_k) = y. \tag{13}$$

The set $K_\infty(f)$ *of* asymptotic critical values *consists of all* $y \in \mathbb{R}$ *for which there exists a sequence* $(x_k)_{k\in\mathbb{N}}$ *in* $\mathbb{R}^n$ *such that* $\lim_{k\to\infty} \|x_k\| = \infty$ *and* (13) *hold.*

The following proposition is easy.

PROPOSITION 13. *The set of generalized critical values of a polynomial* $f \in \mathbb{R}[\bar{X}]$ *is the union of its set of critical and asymptotic critical values, i.e.,*

$$K(f) = K_0(f) \cup K_\infty(f).$$

The following notions go back to Thom [Tho].

DEFINITION 14. *Suppose* $f \in \mathbb{R}[\bar{X}]$. *We say that* $y \in \mathbb{R}$ *is a* typical value *of* f *if there is neighborhood* U *of* y *in* $\mathbb{R}$ *and a smooth (i.e.,* C^∞*) manifold* F *such that* $f|_{f^{-1}(U)} : f^{-1}(U) \to U$ *is a (not necessarily surjective) trivial smooth fiber bundle; i.e., there exist a smooth manifold* F *and a* C^∞ *diffeomorphism* $\Phi : f^{-1}(U) \to F \times U$ *such that* $f|_{f^{-1}(U)} = \pi_2 \circ \Phi$, *where* $\pi_2 : F \times U \to U$ *is the canonical projection. We call* $y \in \mathbb{R}$ *an* atypical value *of* f *if it is not a typical value of* f. *The set of all atypical values of* f *is denoted by* $B(f)$ *and called the* bifurcation set *of* f.

Note that a Φ as in the above definition induces a C^∞ diffeomorphism $f^{-1}(y) \to F \times \{y\} \cong F$ for every $y \in U$. In this context, the preimages $f^{-1}(y)$ are called fibers and F is called *the* fiber. We do not require that the fiber bundle $f|_{f^{-1}(U)} : f^{-1}(U) \to U$ is surjective (if it is not, then the image is necessarily empty). Hence the fiber F may be empty, and a *typical value* is not necessarily a value taken on by f. We make use of the following well-known theorem (see, e.g., [KOS, Theorem 3.1]).

THEOREM 15. *Suppose* $f \in \mathbb{R}[\bar{X}]$. *Then* $B(f) \subseteq K(f)$ *and* $K(f)$ *is finite.*

The advantage of $K(f)$ over $K_0(f)$ is that $f^* \in K(f)$ even if f does not attain a minimum on $\mathbb{R}^n$. This is an easy consequence of Theorem 15. See Theorem 19.

Example 16. Consider again the polynomial $f = (1-XY)^2+Y^2 \in R[X,Y]$ from (6) that does not attain its infimum $f^* = 0$ on $\mathbb{R}^2$. Calculating the partial derivatives, it is easy to see that the origin is the only critical point of f. Because f takes the value 1 at the origin, we have $K_0(f) = \{1\}$ and therefore $f^* = 0 \notin K_0(f)$. Clearly, we have $0 \in B(f)$, since $f^{-1}(-y) = \emptyset \neq f^{-1}(y)$ for small $y \in \mathbb{R}_{>0}$. By Theorem 15, we have therefore $0 \in K_\infty(f) \subseteq K(f)$. To show this directly, a first guess would be that $\|\nabla f(x, \frac{1}{x})\|(1 + \|(x, \frac{1}{x})\|)$ tends to zero when $x \to \infty$ because $\lim_{x\to\infty} f(x, \frac{1}{x}) = 0$. But in fact, this expressions tends to 2 when $x \to \infty$. However, a calculation shows that $\lim_{x\to\infty} \|\nabla f(x, \frac{1}{x})\|(1 + \|(x, \frac{1}{x} - \frac{1}{x^3})\|) = 0$.

DEFINITION 17. *For a polynomial* $f \in \mathbb{R}[\bar{X}]$, *we call*

$$S(\nabla f) := \{x \in \mathbb{R}^n \mid \|\nabla f(x)\|\|x\| \leq 1\}$$

the principal gradient tentacle *of* f.

Remark 18. In the definition of $S(\nabla f)$, the inequality $\|\nabla f(x)\|\|x\| \leq 1$ could be exchanged by $\|\nabla f(x)\|\|x\| \leq R$ for some constant $R > 0$. Then all subsequent results will still hold with obvious modifications. Using an R different from 1 might have in certain cases a practical advantage (see section 3.6). However, we decided to stay with this definition in order to not get too technical and to keep the paper readable.

As expressed by the notation $S(\nabla f)$, polynomials f with the same gradient ∇f have the same gradient tentacle; in other words,

$$S(\nabla(f + a)) = S(\nabla f) \qquad \text{for all } a \in \mathbb{R}.$$

The first important property of $S(\nabla f)$ is stated in the following immediate consequence of Theorem 15.

THEOREM 19. *Suppose $f \in \mathbb{R}[\bar{X}]$ is bounded from below. Then $f^* \in K(f)$ and therefore $f^* = \inf\{f(x) \mid x \in S(\nabla f)\}$.*

Proof. By Theorem 15, it suffices to show that $f^* \in B(f)$. Assume that $f^* \notin B(f)$; i.e., f^* is a typical value of f. Then for all y in a neighborhood of f^*, the fibers $f^{-1}(y)$ are smoothly diffeomorphic to each other. But this is absurd, since $f^{-1}(y)$ is empty for $y < f^*$ but certainly not empty in a neighborhood of f^*. □

Let $\mathbb{P}^{n-1}(\mathbb{C})$ denote the $(n-1)$-dimensional complex projective space over $\mathbb{C}$. For a homogeneous polynomial f and a point $z \in \mathbb{P}^{n-1}(\mathbb{C})$, we simply say $f(z) = 0$ to express that f vanishes on (a nonzero point of) the straight line $z \subseteq \mathbb{C}^n$. Following [P1], we give the following definition.

DEFINITION 20. *We say that a polynomial $f \in \mathbb{C}[\bar{X}]$ has only isolated singularities at infinity if $f \in \mathbb{C}$ (i.e., f is constant) or $d := \deg f \geq 1$ and there are only finitely many $z \in \mathbb{P}^{n-1}(\mathbb{C})$ such that*

$$\frac{\partial f_d}{\partial X_1}(z) = \cdots = \frac{\partial f_d}{\partial X_n}(z) = f_{d-1}(z) = 0, \tag{14}$$

where $f = \sum_i f_i$ and each $f_i \in \mathbb{C}[\bar{X}]$ is zero or homogeneous of degree i.

As shown in [P1, section 1.1], the geometric interpretation of the above definition is that the projective closure of a generic fiber of f has only isolated singularities.

Remark 21. A generic complex polynomial has only isolated singularities at infinity. In fact, much more is true: A generic polynomial $f \in \mathbb{C}[\bar{X}]$ of degree $d \geq 1$ has *no* isolated singularities at infinity in the sense that there is no $z \in \mathbb{P}^{n-1}(\mathbb{C})$ such that (14) holds. In more precise words, to every $d \geq 2$, there exists a complex polynomial relation that is valid for all coefficient tuples of polynomials $f \in \mathbb{C}[\bar{X}]$ of degree d for which (14) has an infinite number of solutions. This follows from the fact that for a generic homogeneous polynomial $g \in \mathbb{C}[\bar{X}]$ of degree $d \geq 1$, there are only finitely many points $z \in \mathbb{P}^{n-1}(\mathbb{C})$ such that $\frac{\partial f}{\partial X_i}(z) = 0$ for all i. See [Kus, Théorème II] or [Shu, Proposition 1.1.1].

Remark 22. In the two variable case $n = 2$, every polynomial $f \in \mathbb{C}[\bar{X}]$ has only isolated singularities at infinity. This is clear, since (14) defines an algebraic subvariety of $\mathbb{P}^1(\mathbb{C})$.

The following theorem follows easily from [P1, Theorem 1.4].

THEOREM 23. *Suppose $f \in \mathbb{R}[\bar{X}]$ has only isolated singularities at infinity. Then*

$$R_\infty(f, S(\nabla f)) \subseteq K(f).$$

In particular, $R_\infty(f, S(\nabla f))$ is finite; i.e., f has only finitely many asymptotic values on its principal gradient tentacle.

Proof. Let $(x_k)_{k\in\mathbb{N}}$ be a sequence of points $x_k \in S(\nabla f)$ and $y \in \mathbb{R}$ such that $\lim_{k\to\infty} ||x_k|| = \infty$ and $\lim_{k\to\infty} f(x_k) = y \notin K_0(f)$. We show that $y \in K_\infty(f)$ using implication (i) $\Longrightarrow$ (ii) in [P1, Theorem 1.4]. Because of our sequence $(x_k)_{k\in\mathbb{N}}$, it is impossible that there exist $N \geq 1$ and $\delta > 0$ such that for all $x \in \mathbb{R}^n$ with $\|x\|$ sufficiently large and $f(x)$ sufficiently close to y, we have

$$\|x\|\|\nabla f(x)\| \geq \delta \sqrt[n]{\|x\|}.$$

This means that condition (ii) in [P1, Theorem 1.4] is violated. The implication (i) $\Longrightarrow$ (ii) in [P1, Theorem 1.4] yields that $y \in B(f)$ (here we use that $y \notin K_0(f)$). But $B(f) \subseteq K(f)$ by Theorem 15. This shows $y \in K(f) \setminus K_0(f) \subseteq K_\infty(f)$ by Proposition 13. □

LEMMA 24. *Every $f \in \mathbb{R}[\bar{X}]$ is bounded on $S(\nabla f)$.*

Proof. By the Łojasiewicz inequality at infinity [Spo, Theorem 1], there exist $c_1, c_2 \in \mathbb{N}$ such that for all $x \in \mathbb{C}^n$,

$$|f(x)| \geq c_1 \implies |f(x)| \leq c_2 \|\nabla f(x)\| \|x\|.$$

Then $|f| \leq \max\{c_1, c_2\}$ on $S(\nabla f)$. ∎

3.1. The principal gradient tentacle and sums of squares. Here comes one of the main results of this article which is interesting on its own but can later be read as a convergence result for a sequence of optimal values of SDPs (Theorem 30).

THEOREM 25. *Let $f \in \mathbb{R}[\bar{X}]$ be bounded from below. Furthermore, suppose that f has only isolated singularities at infinity (which is always true in the two variable case $n = 2$) or the principal gradient tentacle $S(\nabla f)$ is compact. Then the following are equivalent:*

(i) *$f \geq 0$ on $\mathbb{R}^n$.*
(ii) *$f \geq 0$ on $S(\nabla f)$.*
(iii) *For every $\varepsilon > 0$, there are sums of squares of polynomials s and t in $\mathbb{R}[\bar{X}]$ such that*

$$f + \varepsilon = s + t(1 - \|\nabla f\|^2 \|\bar{X}\|^2). \tag{15}$$

Proof. First of all, the polynomial $g := 1 - \|\nabla f\|^2 \|\bar{X}\|^2$ is a polynomial describing the principal gradient tentacle

$$S := \{x \in \mathbb{R}^n \mid g(x) \geq 0\} = S(\nabla f).$$

Because sums of squares of polynomials are globally nonnegative on $\mathbb{R}^n$, identity (15) can be viewed as a certificate for $f \geq -\varepsilon$ on S. Hence it is clear that (iii) implies (ii). For the reverse implication, we apply Theorem 9 (with $m = 1$ and $g_1 := g$) to $f + \varepsilon$ instead of f. We have to check only the hypotheses. Condition (a) is clear from Lemma 24. By Theorem 23, we have that $R_\infty(f, S)$ is a finite set if f has only isolated singularities at infinity. If $S(\nabla f)$ is compact, the set $R_\infty(f, S)$ is even empty. Since $f \geq 0$ on S by hypothesis, this set contains clearly only nonnegative numbers. This shows condition (b); i.e., $R_\infty(f + \varepsilon, S) = \varepsilon + R_\infty(f, S)$ is a finite subset of $\mathbb{R}_{>0}$. Finally, the hypothesis $f \geq 0$ on S gives $f + \varepsilon > 0$ on S, which is condition (c). Therefore (ii) and (iii) are proved to be equivalent. The equivalence of (i) and (ii) is an immediate consequence of Theorem 19. ∎

Remark 26. Let $f \in \mathbb{R}[\bar{X}]$ be bounded from below and $S(\nabla f)$ be compact. Then f attains its infimum f^*. To see this, observe that the equivalence of (i) and (ii) in the preceding theorem implies

$$\begin{aligned} f^* &= \sup\{a \in \mathbb{R} \mid f - a \geq 0 \text{ on } \mathbb{R}^n\} \\ &= \sup\{a \in \mathbb{R} \mid f - a \geq 0 \text{ on } S(\nabla f)\} \\ &= \min\{f(x) \mid x \in S(\nabla f)\}. \end{aligned}$$

The following observation is proved in the same way as Remark 2.

Remark 27. If f is a sum of squares in the ring $\mathbb{R}[[\bar{X}]]$ of formal power series, then its lowest (nonvanishing) homogeneous part must be a sum of squares in $\mathbb{R}[\bar{X}]$.

Remark 28. There are polynomials $f \in \mathbb{R}[\bar{X}]$ such that $f \geq 0$ on $\mathbb{R}^n$, but there is no representation (15) for $\varepsilon = 0$. To see this, take a polynomial $f \in \mathbb{R}[\bar{X}]$ such that

$f \geq 0$ on $\mathbb{R}^n$, but f is not a sum of squares in the ring $\mathbb{R}[[\bar{X}]]$ of formal power series (the Motzkin polynomial from (5) is such an example by the preceding remark). Then a representation (15) with $\varepsilon = 0$ is impossible, since the polynomial $1 - \|\nabla f\|^2 \|\bar{X}\|^2$ has a positive constant term and is therefore a square in $\mathbb{R}[[\bar{X}]]$.

3.2. Optimization using the gradient tentacle and sums of squares. Theorem 25 shows that under certain conditions, computation of f^* amounts to computing the supremum over all a such that $f - a = s + t(1 - \|\nabla f\|^2 \|\bar{X}\|^2)$ for some sums of squares s and t in $\mathbb{R}[\bar{X}]$. As sketched in the introduction, sums of squares *of bounded degree* can be nicely parametrized by positive semidefinite matrices. This motivates the following definition.

DEFINITION 29. *For all polynomials $f \in \mathbb{R}[\bar{X}]$ and all $k \in \mathbb{N}_0$, we define $f_k^* \in \mathbb{R} \cup \{\pm\infty\}$ as the supremum over all $a \in \mathbb{R}$ such that $f - a$ can be written as a sum*

$$f - a = s + t(1 - \|\nabla f\|^2 \|\bar{X}\|^2), \tag{16}$$

where s and t are sums of squares of polynomials with $\deg t \leq 2k$.

Here and in the following, we use the convention that the degree of the zero polynomial is $-\infty$ so that $t = 0$ is allowed in the above definition. Note that when the degree of t in (16) is restricted, then automatically also the degree of s.

Therefore the problem of computing f_k^* can be written as an SDP. How to do this is already suggested in our introduction. It goes exactly as in the well-known method of Lasserre for optimization of polynomials on compact basic closed semialgebraic sets. We refer readers to [L1, M1, Sr2] for the details. There are, anyway, several toolboxes for MATLAB (a software for numerical computation) which can be used to create and solve the corresponding SDPs without knowing these details. The toolboxes we know are YALMIP [Löf] (which is very flexible and good for much more than sums of squares things), SOSTOOLS [PPS] (which has a very flexible and nice syntax), GloptiPoly [HL] (very easy to use for simple problems), and SparsePOP [KKW] (specialized for sparse polynomials). Besides MATLAB and such a toolbox one needs also an SDP solver for which the toolbox provides an interface.

A side remark that we want to make here is that to each SDP there is a dual SDP, and it is desirable from the theoretical and practical point of view that *strong duality* holds; i.e., the optimal value of the primal and dual SDP coincide. For the SDPs arising from Definition 29, strong duality holds. This follows from the fact that principal gradient tentacles (unlike gradient varieties) always have nonempty interior (they always contain a small neighborhood of the origin). For a proof confer [L1, Theorem 4.2], [M1, Corollary 3.2], or [Sr2, Corollary 21]. Here we will neither define the dual SDP nor discuss its interpretation in terms of the so-called moment problem.

Recalling the definition of f^{sos} in (4), we have obviously

$$f^{\text{sos}} \leq f_0^* \leq f_1^* \leq f_2^* \leq \dots, \tag{17}$$

and if f is bounded from below, then all f_k^* are lower bounds (perhaps $-\infty$) of f^* by Theorem 19. Note that the technique from Jibetean and Laurent (see section 1.6) gives upper bounds for f^* so that it complements nicely our method. It is easy to see that Theorem 25 can be expressed in terms of the sequence $f_0^*, f_1^*, f_2^*, \dots$ as follows.

THEOREM 30. *Let $f \in \mathbb{R}[\bar{X}]$ be bounded from below. Suppose that f has only isolated singularities at infinity (e.g., $n = 2$) or the principle gradient tentacle $S(\nabla f)$ is compact. Then the sequence $(f_k^*)_{k \in \mathbb{N}}$ converges monotonically increasing to f^*.*

The following example shows that it is, unfortunately, in general not true that $f_k^* = f^*$ for big $k \in \mathbb{N}$.

Example 31. Let f be the Motzkin polynomial from (5). By Theorem 30, we have $\lim_{k\to\infty} f_k = 0$. But it is not true that $f_k = 0$ for some $k \in \mathbb{N}$. By Definition 29, this would imply that for all $\varepsilon > 0$, there is an identity (15) with sums of squares s and t such that $\deg s \leq k$. Because $S(\nabla f)$ has nonempty interior (note that $\nabla f(1,1,1) = 0$, since $f(1,1,1) = 0$), we can use [PS, Proposition 2.6(b)] (see [Sr2, Theorem 4.5] for a more elementary exposition) to see that such an identity would then also have to exist for $\varepsilon = 0$. But this is impossible, as we have seen in Remark 28.

Unfortunately, the assumption that f is bounded from below is necessary in Theorem 30, as shown by the following trivial example.

Example 32. Consider $f := X \in \mathbb{R}[X]$ (i.e., let $n = 1$ and write X instead of X_1). Then $K(f) = \emptyset$, $S(\nabla f) = [-1, 1]$, and $(f_k^*)_{k\in\mathbb{N}}$ converges monotonically increasing to $\inf\{f(x) \mid -1 \leq x \leq 1\} = -1 \neq -\infty = f^*$.

Open Problem 33. *Do Theorems* 25 *and* 30 *hold without the hypothesis that f has only isolated singularities at infinity or $S(\nabla f)$ is compact?*

By the above arguments, it is easy to see that this question could be answered in the affirmative if $R_\infty(f, S(\nabla f))$ were finite for all polynomials $f \in \mathbb{R}[\bar{X}]$ bounded from below on $\mathbb{R}^n$. But this is not true, as the following counterexample shows. We are grateful to Zbigniew Jelonek for pointing out to us this adaption of an example of Parusiński [P2, Example 1.11].

Example 34. Consider the polynomial $h := X + X^2Y + X^4YZ \in \mathbb{R}[X, Y, Z]$, set $f := h^2$, and define for fixed $a > 0$ the curve

$$\gamma : \mathbb{R}_{>0} \to \mathbb{R}^3 : s \mapsto \left(s, \frac{2a}{s^2}, -\frac{\left(1 + \frac{s}{4a}\right)}{2s^2} \right).$$

Observe that

$$h(\gamma(s)) = \frac{3}{4}s + a \qquad \text{and} \qquad \frac{\partial h}{\partial X}(\gamma(s)) = 0,$$

and therefore $f(\gamma(s)) = (\frac{3}{4}s + a)^2$ and

$$\|\nabla f\|^2(\gamma(s)) = 4f\|\nabla h\|^2(\gamma(s)) = 4s^4 \left(\frac{3}{4}s + a\right)^2 \left(\left(\frac{1}{2} - \frac{s}{8a}\right)^2 + (2a)^2 \right).$$

It follows that $\|\nabla f\|^2(\gamma(s))\|\gamma(s)\|^2$ equals

$$\left(4s^6 + 16a^2 + \left(1 + \frac{s}{4a}\right)^2\right) \left(\frac{3}{4}s + a\right)^2 \left(\left(\frac{1}{2} - \frac{s}{8a}\right)^2 + (2a)^2 \right),$$

which tends to $(16a^2 + 1)a^2(1/4 + 4a^2)$ for $s \to 0$. We now see that for $s \to 0$, $\|\gamma(s)\|$ tends to infinity, $f(\gamma(s))$ tends to a^2, and, when a is a sufficiently small positive number, $\|\nabla f\|^2(\gamma(s))\|\gamma(s)\|^2$ tends to a real number smaller than 1. This shows that $a^2 \in R_\infty(f, S(\nabla f))$ for every sufficiently small positive number a. Hence f is an example of a polynomial bounded from below such that $R_\infty(f, S(\nabla f))$ is infinite.

3.3. Implementation in YALMIP. We show here how to encode computation of f_k^* (as well as of $f_{-1}^* := f^{\text{sos}}$) for any $k \in \mathbb{N}$ with YALMIP. First, declare the variables appearing in the polynomial f (here x and y) as well as the variable a to maximize.

```
sdpvar x y a
```

Now specify the polynomial f and the degree bound k (-1 for computing f^{sos}). Here we take the dehomogenization $f := M(X, Y, 1)$, where M is the Motzkin polynomial introduced in (5).

```
f = x^4 * y^2 + x^2 * y^4 - 3 * x^2 * y^2 + 1, k = 0
```

Now compute the partial derivatives with respect to the variables (here x and y) and specify the polynomial g defining the gradient tentacle.

```
df = jacobian(f, [x y]), g = 1 - (df(1)^2 + df(2)^2) * (x^2 + y^2)
```

Define a polynomial variable t of degree $\leq 2k$ and impose the constraints that t and $f - a - tg$ are sums of squares (for some reason the current version of YALMIP does here not accept a degree zero polynomial t so that this has to be modeled as a scalar variable).

```
if k > 0
 v = monolist([x; y], 2*k), coeffVec = sdpvar(length(v), 1)
 t = coeffVec' * v
 constraints = set(sos(f - a - t * g)) + set(sos(t))
elseif k == 0
 coeffVec = sdpvar(1, 1), t = coeffVec
 constraints = set(sos(f - a - t * g)) + set(t > 0)
else
 coeffVec = []
 constraints = set(sos(f - a))
end
```

Now solve the SDP and output the result for a.

```
solvesos(constraints, -a, [], [a; coeffVec]), double(a)
```

3.4. Implementation in SOSTOOLS. Below we give an SOSTOOLS code which is even slightly easier to read but essentially analogous to the YALMIP code. In contrast to the YALMIP code above, the Symbolic Math Toolbox is required to execute the code below.

```
syms x y a t
f = x^4 * y^2 + x^2 * y^4 - 3 * x^2 * y^2 + 1, k = 0
df = jacobian(f, [x y]), g = 1 - (df(1)^2 + df(2)^2) * (x^2 + y^2)
prog = sosprogram([x; y], a)

if k > 0
 v = monomials([x; y], [0 : k]), [prog, t] = sossosvar(prog, v)
 prog = sosineq(prog, f - a - t * g)
elseif k == 0
 prog = sosdecvar(prog, t), prog = sosineq(prog, t)
 prog = sosineq(prog, f - a - t * g)
else
 prog = sosineq(prog, f - a)
end

prog = sossetobj(prog, -a), prog = sossolve(prog)
sosgetsol(prog, a)
```

3.5. Numerical results. The following examples have been computed on an ordinary PC with MATLAB 7, YALMIP 3, and the SDP solver SeDuMi 1.1. Most of the computations took a few seconds, some of them a few minutes. The first example

corresponds exactly to the code in section 3.3. To compute the others, the variables, the polynomial f, and the degree bound k have to be changed in that code.

Example 35. Let $f := M(X,Y,1)$ be the dehomogenization of the Motzkin polynomial M from (5), i.e., $f := M(X,Y,1) = X^4Y^2 + X^2Y^4 - 3X^2Y^2 + 1 \in \mathbb{R}[X,Y]$. We have $f^* = 0$ but $f^{\text{sos}} = -\infty$ (the latter is an easy exercise). If we execute the program from section 3.3 with $k = -1$ instead of $k = 0$, the computer answers that the SDP is infeasible, which means indeed that $f^{\text{sos}} = -\infty$. Executing the same program for $k = 0, 1, 2$ yields $f_0^* \approx -0.0017$, $f_1^* \approx -0.0013$, and $f_2^* \approx 0.000066$, which is already very close to $f^* = 0$. By Theorem 30, the sequence $f_0, f_1, f_2, \ldots$ converges monotonically to $f^* = 0$. But the computed value $f_2^* \approx 0.000066$ is positive so that there are obviously numerical problems. Confer [PS, Example 2].

Example 36. Define $f := M(X,1,Z) \in \mathbb{R}[X,Z]$, where M is the Motzkin polynomial from (5), i.e., $f = X^4 + X^2 + Z^6 - 3X^2Z^2 \in \mathbb{R}[X,Z]$. Computation yields $f^{\text{sos}} \approx -0.1780$, $f_0^* \approx -5.1749 \cdot 10^{-5}$, $f_1^* \approx -1.2520 \cdot 10^{-7}$, and $f_2^* = 8.7662 \cdot 10^{-10}$, which "equals numerically" $f^* = 0$. This is in accordance with Theorem 25, which guarantees convergence to f^*, since we are in the two variable case. Confer [PS, Example 3].

Example 37. Consider the Berg polynomial $f := X^2Y^2(X^2 + Y^2 - 1) \in \mathbb{R}[X,Y]$ with global minimum $f^* = -1/27$ attained in $(\pm 1/\sqrt{3}, \pm 1/\sqrt{3})$. We have $f^{\text{sos}} = -\infty$, and running the corresponding program gives indeed an output saying that the corresponding SDP is infeasible. The computed optimal values of the first principal tentacle relaxations are $f_0^* \approx -0.0564$, $f_1^* \approx -0.0555$, $f_2^* \approx -0.0371$, and $f_3^* \approx -0.0370 \approx -1/27 = f^*$. Confer [L1, Example 3], [NDS, Example 3], and [JL, Example 4].

Example 38. Being a polynomial in two variables of degree at most four, we have that for $f := (X^2+1)^2 + (Y^2+1)^2 - 2(X+Y+1)^2 \in \mathbb{R}[X,Y]$, $f - f^*$ must be a sum of squares (see introduction) whence $f^* = f^{\text{sos}}$. By computation, we obtain for all values $f^{\text{sos}}, f_0^*, f_1^*, f_2^*$ approximately -11.4581. That all these computed values are the same can be expected by $f^* = f^{\text{sos}}$ and the monotonicity (17). Confer [L1, Example 2] and [JL, Example 3].

Example 39. In [LL], it is shown that

$$f := \sum_{i=1}^{5} \prod_{j \neq i} (X_i - X_j) \in \mathbb{R}[X_1, X_2, X_3, X_4, X_5]$$

is nonnegative on $\mathbb{R}^5$ but not a sum of squares of polynomials. Therefore $f^{\text{sos}} = -\infty$ by Remark 2, since f is homogeneous. The SDP solver detects indeed infeasibility of the corresponding SDP. We have computed $f_0^* \approx -0.2367$, $f_1^* \approx -0.0999$, and $f_2^* \approx -0.0224$. Solving the SDP relaxation computing f_2^* already took the time of a coffee break. As in [JL, Example 6], we observe therefore that minimizing f is after the change of variables $X_i \mapsto X_1 - Y_i$ $(i = 2, 3, 4, 5)$ equivalent to minimizing

$$h := Y_2Y_3Y_4Y_5 + \sum_{i=2}^{5} (-Y_i) \prod_{j \neq i} (Y_j - Y_i) \in \mathbb{R}[Y_2, Y_3, Y_4, Y_5].$$

Computing h^{sos} results in infeasibility. The numerical results using the principle gradient tentacle are $h_0^* \approx -0.2380$, $h_1^* \approx -0.0351$, $h_2^* \approx -0.0072$, $h_3^* \approx -0.0019$, and $h_4^* \approx -0.00086285$, which is already very close to $h^* = 0$. The condition in Theorem 30 is satisfied neither for f nor for h, and yet it seems that we have convergence to

h^*. This is a typical observation that might give hope that Open Problem 33 has a positive answer.

Example 40. Consider once more the polynomial $f = (1 - XY)^2 + Y^2$ from (6) and Example 16 that does not attain its infimum $f^* = 0$ on $\mathbb{R}^2$. Since this polynomial is by definition a sum of squares, we have $f^{\text{sos}} = 0 = f^*$ and therefore $f_k^* = 0$ for all $k \in \mathbb{N}$ by (17). By computation, we get $f^{\text{sos}} \approx 1.5142 \cdot 10^{-12}$, which is almost zero but also $f_0^* \approx 0.0016$, $f_1^* \approx 0.0727$, and $f_2^* \approx 0.1317$, which shows that there are big numerical problems. We have verified that the corresponding SDPs have nevertheless been solved quite accurately. The problem is that small numerical errors in the coefficients of a polynomial can perturb its infimum quite a lot whenever the infimum is not attained (or attained very far from the origin). It should be subject to further research how to fight this problem. Anyway, the gradient tentacle method still performs in this example much better than the gradient variety method, which yields the wrong answer 1 (as described in section 1.7). The method of Jibetean and Laurent gives the best results in this case [JL, Example 5].

3.6. Numerical stability. If the coefficients of f and $\|\nabla f\|\|\bar{X}\|$ have an order of magnitude very different from 1, then the defining polynomial $g = 1 - \|\nabla f\|^2\|\bar{X}\|^2$ for the gradient tentacle should be better exchanged by $R - \|\nabla f\|^2\|\bar{X}\|^2$, where R is a real number of that order of magnitude. This is justified by Remark 18.

Example 40 and other experiments that we did with polynomials bounded from below that do not attain a minimum are a bit disappointing and show that for this "hard" class of polynomials (exactly the class we were attacking), a lot of work remains to be done, at least on the numerical side. The corresponding SDPs tend to be numerically unstable.

For polynomials attaining their minimum, the method in [NDS] is often much more efficient, e.g., for Example 39.

4. Higher gradient tentacles. In this section, we associate with every polynomial $f \in \mathbb{R}[\bar{X}]$ a *sequence* of gradient tentacles. Each of these is defined by a polynomial inequality just as the principal tentacle from section 3 was. But the degree of this polynomial inequality for the Nth tentacle in this sequence will be roughly $2N$ times the degree of f. This has the disadvantage that the corresponding SDP relaxations get very big for large N. Also, we have to deal for *each* N with a sequence of SDPs. All in all, we have therefore a double sequence of SDPs. The advantage is, however, that we can prove a sums of squares representation theorem (Theorem 46) applicable for *all* $f \in \mathbb{R}[\bar{X}]$ bounded from below independently of the answer to Open Problem 33. Again, we think that this theorem is also of theoretical interest. Implementation of the higher gradient tentacle method is analogous to sections 3.3 and 3.4. This time we do not give numerical examples because of Open Problem 33, Remark 21, and numerical problems for big N.

Definition 41. *For $f \in \mathbb{R}[\bar{X}]$ and $N \in \mathbb{N}$, we call*

$$S(\nabla f, N) := \{x \in \mathbb{R}^n \mid \|\nabla f(x)\|^{2N}(1 + \|x\|^2)^{N+1} \leq 1\}$$

the Nth gradient tentacle *of f.*

A trivial fact that one should keep in mind is that $\|\nabla f(x)\|^2(1 + \|x\|^2) \leq 1$ and in particular $\|\nabla f(x)\|\|x\| \leq 1$ for all $x \in S(\nabla f, N)$. This shows that

$$V(\nabla f) \cap \mathbb{R}^n \subseteq S(\nabla f, 1) \subseteq S(\nabla f, 2) \subseteq S(\nabla f, 3) \subseteq \cdots \subseteq S(\nabla f).$$

The definition of $S(\nabla f, N)$ is motivated by the following definition, which is taken from [KOS, page 79].

DEFINITION 42. *Suppose $f \in \mathbb{R}[\bar{X}]$ and $N \in \mathbb{N}$. The set $K_\infty^N(f)$ consists of all $y \in \mathbb{R}$ for which there exists a sequence $(x_k)_{k\in\mathbb{N}}$ in $\mathbb{R}^n$ such that*

$$\lim_{k\to\infty} ||x_k|| = \infty, \qquad \lim_{k\to\infty} \|\nabla f(x_k)\|\|x_k\|^{1+\frac{1}{N}} = 0, \quad \text{and} \qquad \lim_{k\to\infty} f(x_k) = y. \tag{18}$$

Clearly, we have

$$K_\infty^1(f) \subseteq K_\infty^2(f) \subseteq K_\infty^3(f) \subseteq \cdots \subseteq K_\infty(f).$$

The next lemma says that this chain actually gets stationary and reaches $K_\infty(f)$. For the proof, we refer readers to [KOS, Lemma 3.1].

LEMMA 43 (Kurdyka, Orro, and Simon). *For all $f \in \mathbb{R}[\bar{X}]$, there exists $N \in \mathbb{N}$ such that*

$$K_\infty(f) = K_\infty^N(f).$$

Now we prove for sufficiently large gradient tentacles what Theorem 19 was for the principal gradient tentacle (which contains all higher gradient tentacles).

THEOREM 44. *Suppose $f \in \mathbb{R}[\bar{X}]$ is bounded from below. Then $f^* \in K(f)$ and there is $N_0 \in \mathbb{N}$ such that for all $N \geq N_0$,*

$$f^* = \inf\{f(x) \mid x \in S(\nabla f, N)\}. \tag{19}$$

Proof. We know already from Theorem 19 that $f^* \in K(f)$. By Proposition 13, at least one of the following two cases therefore must occur. The first case is that $f^* \in K_0(f)$. Then f^* is attained by f on its gradient variety and therefore on the Nth gradient tentacle for actually all $N \in \mathbb{N}$. Hence we can set $N_0 := 1$. In the second case, $f^* \in K_\infty(f)$, we can choose some $N_0 \in \mathbb{N}$ such that $f^* \in K_\infty^N(f)$ by the previous lemma. Then $f^* \in K_\infty^N(f)$ for any $N \geq N_0$. This means that there exists a sequence $(x_k)_{k\in\mathbb{N}}$ satisfying (18). Therefore $\|\nabla f(x)\|\|x_k\|^{1+1/N} \leq \frac{1}{2}$ and consequently

$$\|\nabla f(x_k)\|^{2N}(1 + \|x_k\|^2)^{N+1} \leq \|\nabla f(x_k)\|^{2N}(2\|x_k\|^2)^{N+1} \leq 1$$

for all large k, since $\|x_k\| \geq 1$ and $2^{N+1} \leq 2^{2N}$. This shows that $x_k \in S(\nabla f, N)$ for all large k, which implies our claim. ∎

The great advantage of the higher gradient tentacles over the principal one is that they are *always* small enough to admit only finitely many asymptotic values; i.e., there is no counterpart to Example 34.

THEOREM 45. *For every $f \in \mathbb{R}[\bar{X}]$, $R_\infty(f, S(\nabla f)) \subseteq K_\infty(f)$. In particular, every $f \in \mathbb{R}[\bar{X}]$ has only finitely many asymptotic values on each of its higher gradient tentacles; i.e., the set $R_\infty(f, S(\nabla f, N))$ is finite for all $N \in \mathbb{N}$.*

Proof. Let $y \in \mathbb{R}$ be such that (7) holds for some sequence $(x_k)_{k\in\mathbb{N}}$ of points $x_k \in S(\nabla f, N)$. By Definition 41,

$$\|\nabla f(x_k)\|^N \|x_k\|^N \leq \frac{1}{\|x_k\|} \to 0 \qquad \text{for } k \to \infty,$$

implying (13). This shows that $y \in K_\infty(f)$. ∎

4.1. Higher gradient tentacles and sums of squares. We are now able to prove the third important sums of squares representation theorem of this article besides Theorems 9 and 25.

THEOREM 46. *For all $f \in \mathbb{R}[\bar{X}]$ bounded from below, there is $N_0 \in \mathbb{N}$ such that for all $N \geq N_0$, the following are equivalent:*

(i) *$f \geq 0$ on $\mathbb{R}^n$.*
(ii) *$f \geq 0$ on $S(\nabla f, N)$.*
(iii) *For every $\varepsilon > 0$, there are sums of squares of polynomials s and t in $\mathbb{R}[\bar{X}]$ such that*

$$f + \varepsilon = s + t(1 - \|\nabla f\|^{2N}(1 + \|\bar{X}\|^2)^{N+1}). \tag{20}$$

Moreover, these conditions are equivalent for all f attaining a minimum on $\mathbb{R}^n$ and all $N \in \mathbb{N}$. Finally, (ii) *and* (iii) *are equivalent for all $f \in \mathbb{R}[\bar{X}]$ and $N \in \mathbb{N}$.*

Proof. We first show that (ii) and (iii) are always equivalent. To see this, observe that $g_1 := 1 - \|\nabla f\|^{2N}\|\bar{X}\|^{2N+2}$ is a polynomial that defines the set $S := \{x \in \mathbb{R}^n \mid g_1 \geq 0\} = S(\nabla f, N)$. Because sums of squares of polynomials are globally nonnegative on $\mathbb{R}^n$, identity (20) can be viewed as a certificate for $f \geq -\varepsilon$ on S. Hence it is clear that (iii) implies (ii). For the reverse implication, we apply Theorem 9 to $f + \varepsilon$ instead of f. We have to check only the hypotheses. Condition (a) is clear from Lemma 24. By Theorem 45, we have that $R_\infty(f, S)$ is a finite set. Since $f \geq 0$ on S by hypothesis, this set contains clearly only nonnegative numbers. This shows condition (b); i.e., $R_\infty(f + \varepsilon, S) = \varepsilon + R_\infty(f, S)$ is a finite subset of $\mathbb{R}_{>0}$. Finally, the hypothesis $f \geq 0$ on S gives $f + \varepsilon > 0$ on S, which is condition (c).

Now suppose that $f \in \mathbb{R}[\bar{X}]$ attains a minimum $f(x^*) = f^*$ in a point $x^* \in \mathbb{R}^n$. Then $\nabla f(x^*) = 0$ and therefore $x^* \in S(\nabla f, N)$ for all $N \in \mathbb{N}$. This shows that (i) and (ii) are in this case equivalent for *all* $N \in \mathbb{N}$.

By what has already been proved, it remains only to show that (i) and (ii) are equivalent for large $N \in \mathbb{N}$ when $f \in \mathbb{R}[\bar{X}]$ is bounded from below but does not attain a minimum. But in this case, (19) holds by Theorem 44 yielding the equivalence of the first two conditions. □

Without needing it for our application, we draw the following immediate corollary. Taking $N = 1$ in the second part of this corollary yields Theorem 6 of Nie, Demmel, and Sturmfels.

COROLLARY 47. *Suppose $f \in \mathbb{R}[\bar{X}]$ and $f \geq 0$ on $V(\nabla f) \cap \mathbb{R}^n$. Then $f + \varepsilon$ is for all $\varepsilon > 0$ a sum of squares modulo any principal ideal generated by a power of the polynomial $\|\nabla f\|^2(1 + \|\bar{X}\|^2)$; i.e., for every $\varepsilon > 0$ and $N \in \mathbb{N}$, there is a sum of squares s in $\mathbb{R}[\bar{X}]$ and a polynomial $p \in \mathbb{R}[\bar{X}]$ such that*

$$f = s + p(\|\nabla f\|^2(1 + \|\bar{X}\|^2))^N.$$

In particular, $f + \varepsilon$ is for all $\varepsilon > 0$ a sum of squares modulo each power of its gradient ideal; i.e., for every $\varepsilon > 0$ and $N \in \mathbb{N}$, there is a sum of squares s in $\mathbb{R}[\bar{X}]$ such that

$$f \in s + (\nabla f)^N.$$

Proof. The second claim follows from the first one. The first claim follows immediately from implication (i) $\Longrightarrow$ (iii) in Theorem 46, which always holds for all $N \in \mathbb{N}$. □

4.2. Optimization using higher gradient tentacles and sums of squares. The following definition can be motivated in the same way as Definition 29 in section 3.

DEFINITION 48. *For all polynomials $f \in \mathbb{R}[\bar{X}]$, all $N \in \mathbb{N}$, and all $k \in \mathbb{N}_0$, we define $f^*_{N,k} \in \mathbb{R} \cup \{\pm\infty\}$ as the supremum over all $a \in \mathbb{R}$ such that $f - a$ can be written as a sum*

$$f - a = s + t(1 - \|\nabla f\|^{2N}(1 + \|\bar{X}\|^2)^{N+1}), \tag{21}$$

where s and t are sums of squares of polynomials with $\deg t \leq 2k$.

Again, as outlined in section 3, computation of $f_{N,k}$ amounts to solving an SDP for each fixed $N \in \mathbb{N}$ and $k \in \mathbb{N}_0$. Recalling the definition of f^{sos} in (4), we have for each fixed $N \in \mathbb{N}$,

$$f^{\text{sos}} \leq f^*_{N,0} \leq f^*_{N,1} \leq f^*_{N,2} \leq \cdots,$$

and if f is bounded from below, then all $f^*_{N,k}$ are lower bounds of f^* by Theorem 44. It would be desirable to also have information on how the $f_{N,k}$ are related to each other when not only k but also N varies. All we know about that is the following proposition.

PROPOSITION 49. *For all $f \in \mathbb{R}[\bar{X}]$, $N \in \mathbb{N}$, and $k \in \mathbb{N}_0$,*

$$f^*_{N+1,k} \leq f^*_{N,k+d}.$$

Proof. Let us define the polynomials h_N as in (11) and substitute into the identity proved in Lemma 11 the polynomials $\|\nabla f\|^2$ for Y and $\|\bar{X}\|^2$ for $\bar{X}$. Then we get

$$1 - \|\nabla f\|^{2(N+1)}(1 + \|\bar{X}\|^2)^{N+2} = p + q(1 - \|\nabla f\|^{2N}(1 + \|\bar{X}\|^2)^{N+1}), \tag{22}$$

where p and

$$q := \left(1 + \frac{1}{N}\right)\|\nabla f\|^2(1 + \|\bar{X}\|^2)$$

are sums of squares of polynomials. The degree of q is no higher than $2(d-1)+2 = 2d$. Now if for $a \in \mathbb{R}$ we have an identity

$$f - a = s + t(1 - \|\nabla f\|^{2(N+1)}(1 + \|\bar{X}\|^2)^{N+2})$$

for sums of squares s and t with $\deg t \leq 2k$, then for the same a

$$f - a = (s + tp) + tq(1 - \|\nabla f\|^{2N}(1 + \|\bar{X}\|^2)^{N+1})$$

and $\deg(tq) \leq 2(k + d)$. ∎

We conclude by interpreting Theorem 46 as a convergence result concerning the optimal values $f^*_{N,k}$ of the proposed relaxations. This is the counterpart to Theorem 30 from section 2.

THEOREM 50. *For all $f \in \mathbb{R}[\bar{X}]$ bounded from below, $(f^*_{N,k})_{k\in\mathbb{N}}$ converges monotonically increasing to f^*, provided that $N \in \mathbb{N}$ is sufficiently large (depending on f). If f attains a minimum on $\mathbb{R}^n$, $(f^*_{N,k})_{k\in\mathbb{N}}$ converges monotonically increasing to f^* no matter what $N \in \mathbb{N}$ is.*

5. Conclusions. We have proposed a method for computing numerically the infimum of a real polynomial in n variables which is bounded from below on $\mathbb{R}^n$. As in [JL] and [NDS], the approach is to find semidefinite relaxations relying on sums of squares certificates and critical point theory. As one could expect, polynomials that do not attain a minimum on $\mathbb{R}^n$ (that are either unbounded from below or have a finite infimum that is not attained) are particularly hard to handle. In [JL], this problem (among others) was solved by perturbing the coefficients of the polynomial to guarantee a minimum (in particular, boundedness from below). Though the results in [JL] are quite good, we are convinced that one should also look for other methods that avoid perturbations and the danger of numerical ill-conditioning coming along with them. Proving sums of squares representations for polynomials positive on their gradient variety, it was shown by Nie, Demmel, and Sturmfels [NDS] that an approach without perturbation is possible. The computational performance of their method is extremely good. However, for polynomials that do not attain a minimum, their method yields wrong answers. Combining considerable machinery from differential geometry and real algebraic geometry, we have shown that part of this limitation can be removed. By using our gradient tentacles instead of the gradient variety, polynomials that do not attain a minimum but are bounded from below can also be handled. Our method has three major problems. First, we do not address the important question of how to check efficiently if a polynomial is bounded from below. For such polynomials, our method still gives a wrong answer (see Example 32). Second, it turns out that solving SDPs that arise from a polynomial that does not attain a minimum takes sometimes a surprisingly long time. And third, small numerical inaccuracies might lead to big changes in the infimum of a polynomial if the infimum is not attained. All three problems should be subject to further research. Polynomials not attaining a minimum remain hard to handle in practice. On the theoretical side, we have combined the theory of generalized critical values with the theory of real holomorphy rings and have obtained new interesting characterizations of nonnegative polynomials.

Acknowledgments. We are most grateful to Zbigniew Jelonek for the discussions in Passau, where he showed us Example 34 and Parusinski's theorem, Theorem 23. Our thanks go also to Mohab Safey El Din for shifting our attention to Theorem 19, to Richard Leroy for helping to prove Lemma 11, and to Krzysztof Kurdyka for interesting discussions in Paris.

REFERENCES

[Ble] G. Blekherman, *There are Significantly More Nonnegative Polynomials than Sums of Squares*, preprint, http://arxiv.org/abs/math.AG/0309130.

[Cas] G. Cassier, *Problème des moments sur un compact de $\mathbb{R}^n$ et décomposition de polynômes à plusieurs variables*, J. Funct. Anal., 58 (1984), pp. 254–266.

[HL] D. Henrion and J. Lasserre, *GloptiPoly: Global optimization over polynomials with Matlab and SeDuM*, ACM Trans. Math. Software, 29 (2003), pp. 165–194.

[JL] D. Jibetean and M. Laurent, *Semidefinite approximations for global unconstrained polynomial optimization*, SIAM J. Optim., 16 (2005), pp. 490–514.

[KKW] M. Kojima, S. Kim, and H. Waki, *Sparsity in sums of squares of polynomials*, Math. Program., 103 (2005), pp. 45–62.

[KOS] K. Kurdyka, P. Orro, and S. Simon, *Semialgebraic Sard theorem for generalized critical values*, J. Differential Geom., 56 (2000), pp. 67–92.

[Kri] J. Krivine, *Anneaux préordonnés*, J. Anal. Math., 12 (1964), pp. 307–326.

[Kus] A. Kushnirenko, *Polyèdres de Newton et nombres de Milnor*, Invent. Math., 32 (1976), pp. 1–31.

[L1] J. B. Lasserre, *Global optimization with polynomials and the problem of moments*, SIAM J. Optim., 11 (2001), pp. 796–817.

[L2] J. B. Lasserre, *A sum of squares approximation of nonnegative polynomials*, SIAM J. Optim., 16 (2006), pp. 751–765.

[LL] A. Lax and P. Lax, *On sums of squares*, Linear Algebra Appl., 20 (1978), pp. 71–75.

[LN] J. B. Lasserre and T. Netzer, *SOS approximations of nonnegative polynomials via simple high degree perturbations*, Math. Z., to appear.

[Löf] J. Löfberg, YALMIP: A MATLAB *Toolbox for Rapid Prototyping of Optimization Problems*, http://control.ee.ethz.ch/∼joloef/yalmip.php.

[M1] M. Marshall, *Optimization of polynomial functions*, Canad. Math. Bull., 46 (2003), pp. 575–587.

[M2] M. Marshall, *Representations of Nonnegative Polynomials, Degree Bounds, and Applications to Optimization*, preprint, http://math.usask.ca/∼marshall/.

[NDS] J. Nie, J. Demmel, and B. Sturmfels, *Minimizing polynomials via sum of squares over the gradient ideal*, Math. Program., 106 (2006), pp. 587–606.

[Net] T. Netzer, *High Degree Perturbations of Nonnegative Polynomials*, Diplomarbeit, Universität Konstanz, Konstanz, Germany, 2005, http://www.math.uni-konstanz.de/∼netzer/.

[P1] A. Parusiński, *On the bifurcation set of a complex polynomial with isolated singularities at infinity*, Compositio Math., 97 (1995), pp. 369–384.

[P2] A. Parusiński, *A note on singularities at infinity of complex polynomials*, in Symplectic Singularities and Geometry of Gauge Fields, Banach Center Publ. 39, Polish Academy of Sciences, Warsaw, Poland, 1997, pp. 131–141.

[PD] A. Prestel and C. Delzell, *Positive Polynomials*, Springer Monogr. Math., Springer, Berlin, 2001.

[PPS] S. Prajna, A. Papachristodoulou, P. Seiler, and P. Parrilo, SOSTOOLS *and its control applications*, in Positive Polynomials in Control, Lecture Notes in Control and Inform. Sci. 312, Springer, Berlin, 2005, pp. 273–292.

[PS] P. Parrilo and B. Sturmfels, *Minimizing polynomial functions*, in Algorithmic and Quantitative Real Algebraic Geometry, DIMACS Ser. Discrete Math. Theoret. Comput. Sci. 60, AMS, Providence, RI, 2003, pp. 83–99.

[Rab] P. Rabier, *Ehresmann fibrations and Palais–Smale conditions for morphisms of Finsler manifolds*, Ann. of Math. (2), 146 (1997), pp. 647–691.

[Rez] B. Reznick, *Some concrete aspects of Hilbert's* 17*th problem*, in Real Algebraic Geometry and Ordered Structures, Contemp. Math. 253, AMS, Providence, RI, 2000, pp. 251–272.

[Sch] K. Schmüdgen, *The K-moment problem for compact semi-algebraic sets*, Math. Ann., 289 (1991), pp. 203–206.

[Scd] J. Schmid, *On the degree complexity of Hilbert's* 17*th problem and the Real Nullstellensatz*, Habilitationsschrift, Universität Dortmund, Dortmund, Germany, 1998.

[Sho] N. Shor, *Class of global minimum bounds of polynomial functions*, Cybernetics, 23 (1987), pp. 731–734.

[Shu] E. Shustin, *Critical points of real polynomials, subdivisions of Newton polyhedra, and topology of real algebraic hypersurfaces*, in Topology of Real Algebraic Varieties and Related Topics, Amer. Math. Soc. Transl. Ser. 2, AMS, Providence, RI, pp. 203–223.

[Spo] S. Spodzieja, *Lojasiewicz inequalities at infinity for the gradient of a polynomial*, Bull. Polish Acad. Sci. Math., 50 (2002), pp. 273–281.

[Sr1] M. Schweighofer, *Iterated rings of bounded elements and generalizations of Schmüdgen's Positivstellensatz*, J. Reine Angew. Math., 554 (2003), pp. 19–45. Erratum available at http://arxiv.org/abs/math.AC/0510675.

[Sr2] M. Schweighofer, *Optimization of polynomials on compact semialgebraic sets*, SIAM J. Optim., 15 (2005), pp. 805–825.

[SS] N. Shor and P. Stetsyuk, *Modified r-algorithm to find the global minimum of polynomial functions*, Cybern. Syst. Anal., 33 (1997), pp. 482–497.

[Ste] G. Stengle, *A Nullstellensatz and a Positivstellensatz in semialgebraic geometry*, Math. Ann., 207 (1974), pp. 87–97.

[Tho] R. Thom, *Ensembles et morphismes stratifies*, Bull. Amer. Math. Soc., 75 (1969), pp. 240–284.

[Tod] M. Todd, *Semidefinite optimization*, Acta Numer., 10 (2001), pp. 515–560.

SIAM J. OPTIM.
Vol. 17, No. 4, pp. 943–968
© 2006 Society for Industrial and Applied Mathematics

STATIONARITY RESULTS FOR GENERATING SET SEARCH FOR LINEARLY CONSTRAINED OPTIMIZATION*

TAMARA G. KOLDA†, ROBERT MICHAEL LEWIS‡, AND VIRGINIA TORCZON§

Abstract. We present a new generating set search (GSS) approach for minimizing functions subject to linear constraints. GSS is a class of direct search optimization methods that includes generalized pattern search. One of our main contributions in this paper is a new condition to define the set of conforming search directions that admits several computational advantages. For continuously differentiable functions we also derive a bound relating a measure of stationarity, which is equivalent to the norm of the gradient of the objective in the unconstrained case, and a parameter used by GSS algorithms to control the lengths of the steps. With the additional assumption that the derivative is Lipschitz, we obtain a big-O bound. As a consequence of this relationship, we obtain subsequence convergence to a KKT point, even though GSS algorithms lack explicit gradient information. Numerical results indicate that the bound provides a reasonable estimate of stationarity.

Key words. constrained optimization, linear constraints, global convergence analysis, direct search, generating set search, generalized pattern search, derivative-free methods, stopping criteria

AMS subject classifications. 90C56, 90C30, 65K05

DOI. 10.1137/S1052623403433638

1. Introduction.

We consider a class of direct search methods called *generating set search* (GSS) [15] which encompasses methods such as generalized pattern search [33, 18, 19] and certain classes of derivative-free optimization methods [21, 22, 23, 24]. The problem of interest is the linearly constrained minimization problem:

$$\begin{aligned} &\text{minimize} && f(x) \\ &\text{subject to} && Ax \le b. \end{aligned} \tag{1.1}$$

Here $f : \mathbb{R}^n \to \mathbb{R}$, A is an $m \times n$ matrix, and b is a vector in $\mathbb{R}^m$. Both A and b are assumed to be explicitly available. No assumption of nondegeneracy of the constraints is made. Let Ω denote the feasible region

$$\Omega = \{\, x \;\mid\; Ax \le b \,\}.$$

We assume that the objective f is continuously differentiable on Ω but that the gradient is not computationally available because no procedure exists for computing the gradient and it cannot be approximated accurately.

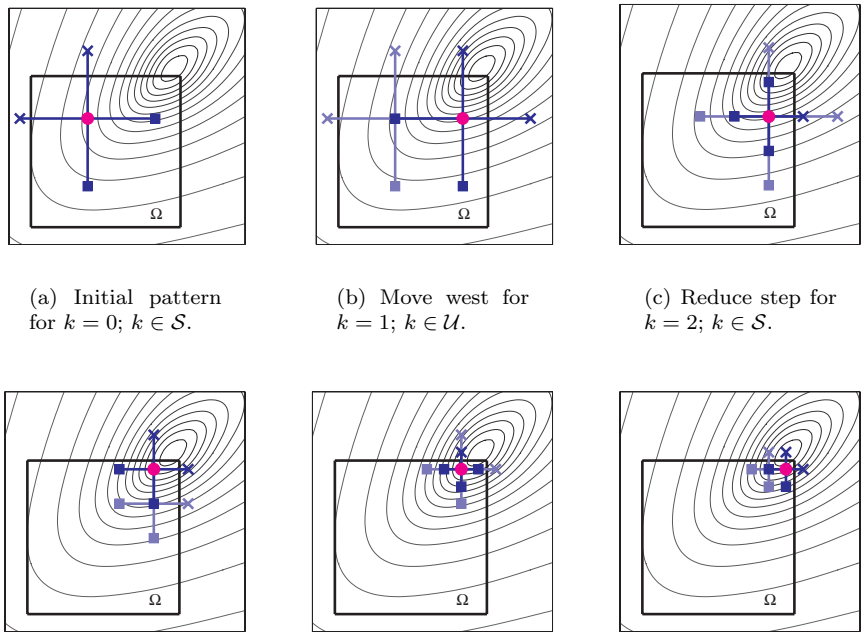
*Received by the editors August 21, 2003; accepted for publication (in revised form) April 24, 2006; published electronically November 22, 2006. The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. Copyright is owned by SIAM to the extent not limited by these rights.

http://www.siam.org/journals/siopt/17-4/43363.html

†Computational Sciences and Mathematics Research Department, Sandia National Laboratories, Livermore, CA 94551-9217 (tgkolda@sandia.gov). This work was supported by the Mathematical, Information, and Computational Sciences Program of the U.S. Department of Energy, under contract DE-AC04-94AL85000 with Sandia Corporation.

‡Department of Mathematics, College of William & Mary, P.O. Box 8795, Williamsburg, VA 23187-8795 (buckaroo@math.wm.edu). This research was supported by the Computer Science Research Institute at Sandia National Laboratories.

§Department of Computer Science, College of William & Mary, P.O. Box 8795, Williamsburg, VA 23187-8795 (va@cs.wm.edu). This research was supported by the Computer Science Research Institute at Sandia National Laboratories.



(a) Initial pattern for $k = 0$; $k \in \mathcal{S}$.

(b) Move west for $k = 1$; $k \in \mathcal{U}$.

(c) Reduce step for $k = 2$; $k \in \mathcal{S}$.

(d) Move north for $k = 3$; $k \in \mathcal{U}$.

(e) Reduce step for $k = 4$; $k \in \mathcal{S}$.

(f) Move west for $k = 5$; $k \in \mathcal{U}$.

FIG. 1.1. *Coordinate search with exact penalization applied to the modified Broyden tridiagonal function with bound constraints.*

1.1. An illustrative example. We illustrate an instance of a GSS method in Figure 1.1. We consider coordinate search applied to the two-dimensional modified Broyden tridiagonal function [4, 26], a standard test problem, with the addition of bounds on the variables. Level curves of the function are shown in the background, and the feasible region is the box labeled Ω. The current iterate x_k is indicated by a circle; this is the point with the lowest value of f found so far, also known as the *best point*. If there are no constraints, a coordinate search method evaluates the function at the $2n$ *trial points* defined by taking a step of a specified length from x_k along the positive and negative coordinate directions, i.e., the *search directions*. The iterates must remain feasible with respect to the bound constraints present in this problem, which means that infeasible trial points are not considered. Terminal crosses show infeasible trial points; solid squares indicate feasible trial points. The lighter versions given in (b)–(f) indicate the search directions and trial points from the previous iteration.

To establish notation and give context for the discussion that follows, we give an outline of a GSS method. Details are developed throughout the paper; complete statements of the algorithms can be found in section 5.

Let $x_0 \in \Omega$ be the initial iterate, and let Δ_0 be the initial choice for the *step-length control parameter* with $\Delta_0 > \Delta_{\text{tol}} > 0$, where Δ_{tol} serves as a measure for termination. The search proceeds for iterations $k = 0, 1, 2, \ldots$ until $\Delta_k < \Delta_{\text{tol}}$.

The first step in each iteration is to select a set of search directions. The number

of search directions is denoted by p_k and the set of search directions by

$$\mathcal{D}_k = \{d_k^{(1)}, \ldots, d_k^{(p_k)}\}.$$

The second step in each iteration is to construct feasible trial points of the form

$$x_k + \tilde{\Delta}_k^{(i)} d_k^{(i)}, \quad i \in \{1, \ldots, p_k\},$$

with $\tilde{\Delta}_k^{(i)} \in [0, \Delta_k]$ chosen to ensure feasibility. These trial points are where the objective function may be evaluated in the search for a new best point to replace x_k.

The third step is to determine whether the iteration is successful or unsuccessful and correspondingly update x and Δ. If one of the trial points reduces the objective function value by an acceptable amount, then that trial point becomes the new iterate x_{k+1}. The step-length control parameter may either be increased or, more usually, left unchanged so that $\Delta_{k+1} = \Delta_k$. In this case the iteration is deemed *successful* and k is assigned to the set of successful iterates denoted by $\mathcal{S}$. Otherwise, none of the trial points improves the value of the objective function, so the step Δ_k is reduced, e.g., $\Delta_{k+1} = \frac{1}{2}\Delta_k$, and the next iterate is unchanged, i.e., $x_{k+1} = x_k$. In this case the iteration is deemed *unsuccessful* and k is assigned to the set of unsuccessful iterates denoted by $\mathcal{U}$.

1.2. Goals of this paper. A primary contribution of this paper is a new condition on the set of search directions $\mathcal{D}_k$ that is flexible but also sufficient to ensure desirable convergence properties of the algorithm. Key to our new results is the way in which the classification of constraints as being nearly binding is tied to Δ_k, the step-length control parameter.

The following measure of stationarity, introduced in [5], is central to our analysis: for $x \in \Omega$,

$$\chi(x) \equiv \max_{\substack{x+w \in \Omega \\ \| w \| \leq 1}} -\nabla f(x)^T w.$$

As discussed in [6], $\chi(x)$ is a continuous function on Ω. Furthermore, $\chi(x) = 0$ for $x \in \Omega$ if and only if x is a Karush–Kuhn–Tucker (KKT) point of the linearly constrained problem.

In Theorem 6.4, under certain assumptions, we show that at unsuccessful iterations there is a big-O relationship between the step-length control parameter and the measure of stationarity:

$$\chi(x_k) = O(\Delta_k) \quad \text{for } k \in \mathcal{U}. \tag{1.2}$$

This means that as Δ_k is reduced, the upper bound on the value of the measure of stationarity is also reduced. Relationship (1.2) is analogous to the unconstrained minimization result (see [8, section 3] or [15, section 3.6]):

$$\| \nabla f(x_k) \| = O(\Delta_k) \quad \text{for } k \in \mathcal{U}. \tag{1.3}$$

Results (1.2) and (1.3) support using the magnitude of Δ_k as a test for termination. In section 7 we give numerical illustrations of relationship (1.2).

Another consequence of (1.2) is that it leads directly to a global convergence result (Theorem 6.5) showing that a subsequence of the iterates converges to a KKT point:

$$\liminf_{k \to \infty} \chi(x_k) = 0. \tag{1.4}$$

The latter follows immediately from (1.2) once the result $\liminf_{k \to \infty} \Delta_k = 0$ from [33] is invoked, thus further simplifying prior global convergence analyses.

1.3. Related work. The GSS methods we propose for solving linearly constrained problems are *feasible-point* methods; i.e., they require all iterates to be feasible. They also share many features with classical feasible directions methods that rely on derivatives [2, 35, 36], especially in the way in which they handle the proximity of the current iterate to the boundary of the feasible region.

Most prior related work has used similar mechanisms for identifying the set of nearly binding linear constraints [25, 34, 19, 1, 30] and [24, Algorithm 2]. Constraints were identified as being nearly binding by considering either the Euclidean distance from the current iterate to the constraint faces [25, 19, 24, 1] or the magnitude of the constraint residual $|\, a_i^T x - b_i \,|$ at the current iterate [34, 30]. A constraint was treated as binding if one of the preceding measures fell below some fixed threshold.

The convergence properties of GSS algorithms rely on the presence at each iteration of a theoretically necessary set of search directions, which we call *core directions*. In the work just cited ([25, 34, 19, 1, 30] and [24, Algorithm 2]), the core directions are all the generators for a *set* of cones. There are situations where the resulting number of search directions is quite large. Since Δ_k can be reduced only at the conclusion of an unsuccessful iteration, and each unsuccessful iteration requires the evaluation of the function at the trial points defined by core directions, there is incentive to try and keep the cardinality of the set of core directions small when the cost of computing f at a feasible point is appreciable.

Algorithm 1 of [24] addresses this concern. Its core directions are the generators of a single cone. However, the *only* allowable search directions are the core directions—the set of search directions cannot be augmented.

The approach we advocate here is a compromise. Our set of core directions is smaller than in [25, 34, 19, 1, 30] and [24, Algorithm 2], but the choice of search directions is more flexible than Algorithm 1 of [24]. The core set need only contain generators for a single cone, but accommodates additional search directions. As reported in [17], the computational advantages of this compromise are appreciable in terms of reducing the number of search directions per iteration, reducing the total number of iterations, and reducing the total number of function evaluations.

Another focus of the work reported here is on establishing (1.2) and a related result regarding the projection of the direction of steepest descent onto the polar of the cone defined by the working set of constraints. Proposition 7.1 in [19] also established a relationship between Δ_k and a different measure of stationarity. The quantity

$$q(x) \equiv P_\Omega\left(x - \nabla f(x)\right) - x, \tag{1.5}$$

where P_Ω denotes the projection onto Ω, is a continuous function of x with the property that $q(x) = 0$ for $x \in \Omega$ if and only if x is a KKT point. In [19, Proposition 7.1] it is shown that

$$\| \, q(x_k) \, \| = O(\sqrt{\Delta_k}) \quad \text{for } k \in \mathcal{U}, \tag{1.6}$$

a result that is neither as satisfying nor as useful as that in (1.2).

Continuing along the lines we began in [15], here we incorporate the sufficient decrease step acceptance criterion from [23, 22, 24], while also preserving a version of the algorithm that requires only simple decrease, as in the work in [19, 1, 30]. The sufficient decrease condition simplifies the analysis. More importantly, the sufficient decrease condition gives us greater flexibility in how we maintain feasibility in the

presence of linear constraints. In particular, using a sufficient decrease acceptance criterion makes steps onto the boundary straightforward.

As mentioned in section 1.2, given (1.2) it is straightforward to prove convergence of a subsequence to a KKT point. The approach to convergence analysis in [1, 30] takes a different tack by focusing on the directional derivatives along the search directions and considering whether limit points of the sequence of iterates are KKT points. This allows a relaxation of the smoothness assumptions on f. If f is not assumed to be continuously differentiable, but is only assumed to be strictly differentiable at limit points of the sequence of iterates, the results in [1, 30] show that those limit points are KKT points. However, subsequence convergence to KKT points in the nonsmooth case is not guaranteed by the results in [1, 30] and, in fact, may not be realized [15].

1.4. Organization. The paper is organized as follows. In section 2, we describe the conditions on the set of core directions for GSS methods applied to problems with linear constraints. As we saw in Figure 1.1, GSS algorithms may generate trial points that are infeasible, so in section 3 we describe how feasibility is maintained. In section 4 we discuss the globalization strategies. Formal statements of GSS algorithms for solving linearly constrained problems are given in section 5. We present two general algorithms. The first (Algorithm 5.1) uses a sufficient decrease condition as in [22, 24]. The second (Algorithm 5.2) uses a simple decrease condition as in [18, 19]. Results showing the stationarity properties of these algorithms are derived in section 6. In section 7 we discuss what the analysis reveals about using Δ_k to test for stationarity and demonstrate its effectiveness on two test problems. In section 8, we summarize the results and their importance. Appendix A contains a discussion of $\chi(x)$ and its use as a measure of stationarity. Appendix B contains geometric results on cones and polyhedra.

2. Search directions. GSS methods for linearly constrained optimization need to choose $\mathcal{D}_k$, the set of search directions, at each iteration. In this section, we describe the conditions we place on $\mathcal{D}_k$ to guarantee (1.2), and thus (1.4). Since GSS methods do not use gradient information, they cannot directly identify descent directions. Instead, the set $\mathcal{D}_k$ must include enough search directions to guarantee that at least one of them is a descent direction and, moreover, allows a sufficiently long step within the feasible region if x_k is not a KKT point. To describe the conditions on the sets of search directions, we start in section 2.1 by reviewing some standard concepts regarding finitely generated cones. Then, in section 2.2, we show how to use the constraints $Ax \leq b$ to define cones that mirror the geometry of the boundary of the polyhedron Ω near the current iterate x_k. Finally, in section 2.3, we detail the conditions placed on the set $\mathcal{D}_k$ to ensure that, for every iteration of any GSS algorithm, there exists at least one direction along which it is possible to take a step of sufficient length while remaining inside Ω.

2.1. Cones and generators. A *cone* K is a set that is closed under nonnegative scalar multiplication, i.e., K is a cone if $x \in K$ implies $\alpha x \in K$ for all $\alpha \geq 0$. The *polar* of a cone K, denoted K°, is defined by

$$K^\circ = \{\ v \ \mid\ w^T v \leq 0 \text{ for all } w \in K\ \}$$

and is itself a cone. Given a convex cone K and any vector v, there is a unique closest point of K to v, the *projection* of v onto K, which we denote by v_K. Given a vector v and a convex cone K, any vector v can be written as $v = v_K + v_{K^\circ}$ and $v_K^T v_{K^\circ} = 0$ [27, 12].

A set of vectors $\mathcal{G}$ *generates* a cone K if K is the set of all nonnegative linear combinations of elements of $\mathcal{G}$. A cone K is *finitely generated* if it can be generated by a finite set of vectors. For any finite set of vectors $\mathcal{G}$, we define

$$\kappa(\mathcal{G}) = \inf_{\substack{v \in \mathbb{R}^n \\ v_K \neq 0}} \max_{d \in \mathcal{G}} \frac{v^T d}{\| v_K \| \, \| d \|}, \quad \text{where } K \text{ is the cone generated by } \mathcal{G}. \tag{2.1}$$

This is a generalization of the quantity given in [15, (3.10)], where $\mathcal{G}$ generates $\mathbb{R}^n$. Note that the value $\kappa(\mathcal{G})$ is a property of the set $\mathcal{G}$—not of the cone K. See Proposition 10.3 in [19] for a proof of the following result.

PROPOSITION 2.1. *If* $\mathcal{G} \neq \{\mathbf{0}\}$, *then* $\kappa(\mathcal{G}) > 0$.

A special case occurs if $\mathcal{G}$ generates $\mathbb{R}^n$. In this case, a set of generators is a *positive spanning set* [7]. Thus a positive spanning set is like a linear spanning set but with the additional requirement that all the coefficients be nonnegative. One particular choice of generating set for $\mathbb{R}^n$ is the set of the positive and negative unit coordinate vectors

$$\{e_1, e_2, \ldots, e_n, -e_1, -e_2, \ldots, -e_n\},$$

which is the set of search directions used for the illustration of coordinate search in Figure 1.1.

2.2. Tangent and normal cones. Let a_i^T be the ith row of the constraint matrix A and let

$$\mathcal{C}_i = \{\ y \ \mid \ a_i^T y = b_i \ \}$$

denote the set where the ith constraint is binding. The set of indices for the binding constraints at x is $I(x) = \{\ i \ \mid \ x \in \mathcal{C}_i \ \}$. The *normal cone* at a point x, denoted by $N(x)$, is the cone generated by the binding constraints, i.e., the cone generated by the set $\{\ a_i \ \mid \ i \in I(x) \ \} \cup \{\mathbf{0}\}$. The presence of $\{\mathbf{0}\}$ means that $N(x) = \{\mathbf{0}\}$ if there are no binding constraints. The *tangent cone*, denoted by $T(x)$, is the polar of the normal cone. Further discussion of the tangent and polar cones in the context of optimization can be found, for instance, in [31, 12, 13, 29].

In our case, we are not only interested in the binding constraints, but also in the nearby constraints. Given $x \in \Omega$, the indices of the *ε-binding constraints* are given by

$$I(x, \varepsilon) = \{\ i \ \mid \ \operatorname{dist}(x, \mathcal{C}_i) \leq \varepsilon \ \}. \tag{2.2}$$

The vectors a_i for $i \in I(x, \varepsilon)$ are the outward-pointing normals to the faces of the boundary of Ω within distance ε of x. The idea of using ε-binding constraints is identical to one sometimes used in gradient-based feasible directions methods, e.g., [2, section 2.5].

Given $x \in \Omega$, we define the *ε-normal cone* $N(x, \varepsilon)$ to be the cone generated by the set $\{\ a_i \ \mid \ i \in I(x, \varepsilon) \ \} \cup \{\mathbf{0}\}$. The presence of $\{\mathbf{0}\}$ means that $N(x, \varepsilon) = \{\mathbf{0}\}$ if $I(x, \varepsilon) = \emptyset$. The corresponding polar cone is the *ε-tangent cone* $T(x, \varepsilon)$. Observe that if $\varepsilon = 0$, then these are just the standard normal and tangent cones; that is, $N(x, 0) = N(x)$ and $T(x, 0) = T(x)$.

Examples of ε-normal and ε-tangent cones are illustrated in Figure 2.1. The set $x + T(x, \varepsilon)$ approximates the feasible region near x, where "near" is with respect to the value of ε. Note that if $I(x, \varepsilon) = \emptyset$, so that $N(x, \varepsilon) = \{\mathbf{0}\}$, then $T(x, \varepsilon) = \mathbb{R}^n$; in other words, if the boundary is more than distance ε away, then the problem looks

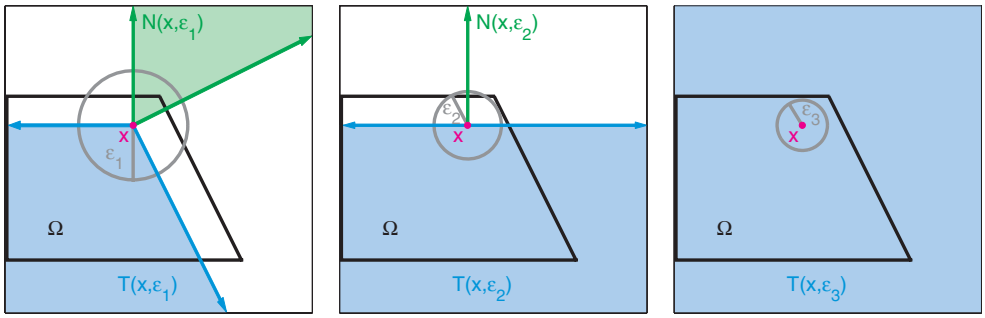


FIG. 2.1. *The cones $N(x,\varepsilon)$ and $T(x,\varepsilon)$ for the values ε_1, ε_2, and ε_3. Note that for this example, as ε varies from ε_1 to 0, there are only the three distinct pairs of cones illustrated ($N(x,\varepsilon_3) = \{\mathbf{0}\}$).*

unconstrained in the ε-neighborhood of x, as can be seen in the third example in Figure 2.1. Observe that one can proceed from x along any direction in $T(x,\varepsilon)$ for a distance of at least ε, and remain inside the feasible region; this is formalized in Proposition 2.2. Overall, the number of *distinct* ε-normal cones (and consequently the number of distinct ε-polar cones) is finite; see Proposition 2.3.

PROPOSITION 2.2. *If $x \in \Omega$, and $v \in T(x,\varepsilon)$ satisfies $\|v\| \leq \varepsilon$, then $x + v \in \Omega$.*

Proof. Let $x \in \Omega$, and $v \in T(x,\varepsilon)$ with $\|v\| \leq \varepsilon$. Since $v \in T(x,\varepsilon) = (N(x,\varepsilon))^\circ$, $a_i^T v \leq 0$ for all $i \in I(x,\varepsilon)$. Thus, $x+v$ satisfies all constraints with $i \in I(x,\varepsilon)$ because

$$a_i^T(x+v) = a_i^T x + a_i^T v \leq b + 0 = b.$$

Meanwhile, if $i \notin I(x,\varepsilon)$, the face $\mathcal{C}_i$ where the ith constraint is binding is more than distance ε away from x. Thus, $x + v \in \Omega$. ∎

PROPOSITION 2.3. *For all $x \in \Omega$ and $\varepsilon > 0$, there are at most 2^m distinct sets $I(x,\varepsilon)$. Consequently, there are at most 2^m distinct cones $N(x,\varepsilon)$ and at most 2^m distinct cones $T(x,\varepsilon)$.*

Proof. Each $I(x_k,\varepsilon_k)$ is a subset of $\{1,\ldots,m\}$, of which there are exactly 2^m possible subsets, including the empty set. The remainder of the proof follows directly from the definitions of $N(x,\varepsilon)$ and $T(x,\varepsilon)$. ∎

2.3. Conditions on the search directions. We now state the conditions on the sets of search directions for GSS for linearly constrained optimization.

At each iteration, a linearly constrained GSS method assembles $\mathcal{D}_k$, the set of search directions. We partition $\mathcal{D}_k$ into two subsets that play different roles in the analysis:

$$\mathcal{D}_k = \mathcal{G}_k \cup \mathcal{H}_k.$$

The set $\mathcal{G}_k$ is required to generate $T(x_k,\varepsilon_k)$ and is called the set of *core directions*. The requirement that the set of search directions contain a set of generators for $T(x_k,\varepsilon_k)$ (which is always $\mathbb{R}^n$ in the unconstrained case) is what led to the name *generating set search* [15].

The (possibly empty) set $\mathcal{H}_k$ accommodates any remaining directions in $\mathcal{D}_k$, the presence of which may prove instrumental in efforts to accelerate the overall progress of the search. For instance, using $\mathcal{H}_k = \{a_i : i \in I(x_k,\varepsilon_k)\}$ can be advantageous computationally [17].

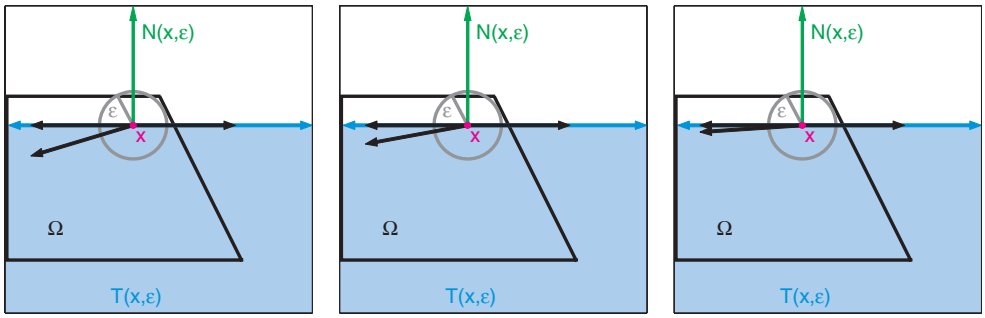


FIG. 2.2. *Condition* 1 *is needed to avoid a sequence of* $\mathcal{G}_k$*'s for which* $\kappa(\mathcal{G}_k) \to 0$.

Our focus here is on the conditions on $\mathcal{G}_k$. The set $\mathcal{H}_k$ accommodates additional directions suggested by heuristics to improve the progress of the search, but has little effect on the analysis. The generating set for $T(x_k, \varepsilon_k)$ contained in $\mathcal{G}_k$ is crucial.

CONDITION 1. There exists a constant $\kappa_{\min} > 0$, independent of k, such that for every k for which $T(x_k, \varepsilon_k) \neq \{\mathbf{0}\}$, the set $\mathcal{G}_k$ generates $T(x_k, \varepsilon_k)$ and satisfies $\kappa(\mathcal{G}_k) \geq \kappa_{\min}$.

Even though there are only finitely many ε-tangent cones $T(x, \varepsilon)$, the set of possible generators for each cone is not necessarily unique, as seen in Figure 2.2. The lower bound $\kappa_{\min}$ from Condition 1 precludes a sequence of $\mathcal{G}_k$'s for which $\kappa(\mathcal{G}_k) \to 0$. Such a situation is depicted in Figure 2.2 for

$$\mathcal{G} = \left\{ \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ -\eta \end{pmatrix} \right\}$$

with three choices of $\eta > 0$. If $-\nabla f(x) = (0, \ -1)^T$, neither of the first two elements of $\mathcal{G}$ are descent directions. Furthermore, since

$$\kappa(\mathcal{G}) \leq \max_{d \in \mathcal{G}} \frac{-\nabla f(x)^T d}{\| \nabla f(x) \| \, \| d \|} = \frac{\eta}{\sqrt{1+\eta^2}} < \eta,$$

the remaining element in $\mathcal{G}$ will be an increasingly poor descent direction if $\eta \to 0$. A nonzero lower bound on $\kappa(\mathcal{G})$, as in Condition 1, will keep the angle between v and at least one generator bounded away from 90°; see [15, sections 2.2 and 3.4.1] for further discussion.

A simple technique to ensure Condition 1 is satisfied is as follows. Let $k_2 > k_1$. If $I(x_{k_2}, \varepsilon_{k_2}) = I(x_{k_1}, \varepsilon_{k_1})$, use the same generators for $T(x_{k_2}, \varepsilon_{k_2})$ as were used for $T(x_{k_1}, \varepsilon_{k_1})$. Recall from Proposition 2.3 that there at most 2^m distinct index sets $I(x, \varepsilon)$ and their corresponding ε-tangent cones $T(x, \varepsilon)$. It then follows that there are at most 2^m distinct sets $\mathcal{G}$ if the same set of generators is always used to generate a particular ε-tangent cone. Since by Proposition 2.1 each $\mathcal{G} \neq \{\mathbf{0}\}$ has a strictly positive value for $\kappa(\mathcal{G})$, and since this technique ensures there are only finitely many $\mathcal{G}_k$'s, we can set $\kappa_{\min} = \min\{\kappa(\mathcal{G}_k) : T(x_k, \varepsilon_k) \neq \{\mathbf{0}\}\}$. Thus, Condition 1 is satisfied.

We have not yet indicated how to compute the generators for a given $T(x_k, \varepsilon_k)$ so as to assemble $\mathcal{G}_k$. If the working set $\{ a_i \mid i \in I(x_k, \varepsilon_k) \}$ is linearly independent, then it is straightforward to calculate the generators of $T(x_k, \varepsilon_k)$ as described in

[25, 19, 17]. If the set $\{ a_i \mid i \in I(x_k, \varepsilon_k) \}$ is linearly dependent (e.g., in the degenerate case), then it is also possible to calculate the generators as described in [17]. In the latter case, the experience reported in [17] suggests that the worst-case computational complexity bounds do not indicate expected performance. For example, for one problem illustrated in [17], the worst-case estimate indicates that more than 4×10^{17} vectors need to be considered when, in fact, only one vector was needed and this one vector was easily identified in less than one-seventh of a second on a conventional workstation using Fukuda's `cddlib` package [9].

Finally, all the core directions must be uniformly bounded; see Condition 2.

CONDITION 2. There exist $\beta_{\max} \geq \beta_{\min} > 0$, independent of k, such that for every k for which $T(x_k, \varepsilon_k) \neq \{\mathbf{0}\}$, the following holds:

$$\beta_{\min} \leq \|d\| \leq \beta_{\max} \qquad \text{for all } d \in \mathcal{G}_k.$$

Condition 2 is easy to satisfy, say, by normalizing all search directions so that $\beta_{\min} = \beta_{\max} = 1$. However, there may be situations where it makes sense to allow the directions in $\mathcal{G}_k$ to accommodate scaling information. This poses no difficulties for the analysis, so long as there are lower and upper bounds, independent of k, on the norm of each $d \in \mathcal{G}_k$.

3. Choosing the step lengths. Given a set of search directions, the length of the step along each direction is dictated by the step-length control parameter Δ_k. In the unconstrained case, the set of trial points at iteration k would be

$$\left\{ x_k + \Delta_k d_k^{(i)} \mid i = 1, \ldots, p_k \right\},$$

where

$$\mathcal{D}_k = \left\{ d_k^{(1)}, d_k^{(2)}, \ldots, d_k^{(p_k)} \right\}.$$

In the constrained case, however, some of those trial points may be infeasible. Thus, the trial points are instead defined by

$$\left\{ x_k + \tilde{\Delta}_k^{(i)} d_k^{(i)} \mid i = 1, \ldots, p_k \right\},$$

where

$$\tilde{\Delta}_k^{(i)} \in [0, \Delta_k]$$

is chosen so that $x_k + \tilde{\Delta}_k^{(i)} d_k^{(i)} \in \Omega$. The main requirement on choosing $\tilde{\Delta}_k^{(i)}$ is that a full step is used if possible, as formally stated in the following condition.

CONDITION 3. If $x_k + \Delta_k d_k^{(i)} \in \Omega$, then $\tilde{\Delta}_k^{(i)} = \Delta_k$.

The simplest formula for choosing $\tilde{\Delta}_k^{(i)} \in [0, \Delta_k]$ that satisfies Condition 3 is

$$\tilde{\Delta}_k^{(i)} = \begin{cases} \Delta_k & \text{if } x_k + \Delta_k d_k^{(i)} \in \Omega, \\ 0 & \text{otherwise.} \end{cases} \tag{3.1}$$

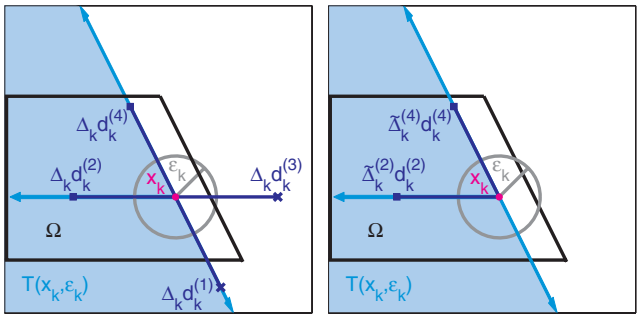


FIG. 3.1. *The step-length control parameter Δ_k may lead to infeasible trial points. The effect of using* (3.1) *is that infeasible points simply are not considered as candidates to replace x_k.*

This corresponds to a form of exact penalization (see [15, section 8.1]) since the effect of (3.1) is to reject (by setting $\tilde{\Delta}_k^{(i)} = 0$) any step $\Delta_k d_k^{(i)}$ that would generate an infeasible trial point. Since the constraints are assumed to be explicit (i.e., A and b are known), verifying the feasibility of a trial point is straightforward. This strategy is illustrated in Figure 3.1.

More sophisticated strategies can be employed for choosing $\tilde{\Delta}_k^{(i)}$ when $x_k + \Delta_k d_k^{(i)}$ is infeasible. Since alternatives for choosing $\tilde{\Delta}_k^{(i)}$ depend on the globalization strategy, we defer the discussion of further examples to section 4.

4. Globalization. Globalization of GSS refers to the conditions that are enforced to ensure that

$$\liminf_{k\to\infty} \Delta_k = 0. \tag{4.1}$$

These conditions affect the decision of whether or not to accept a trial point as the next iterate and how to update Δ_k. Globalization strategies for GSS are discussed in detail in [15, section 3.7]. Here we review those features that are relevant to our analysis of algorithms for the linearly constrained case.

In any GSS algorithm, x_k is always the best feasible point discovered thus far; i.e., $f(x_k) \le f(x_j)$ for all $j \le k$. However, different conditions are imposed on *how much better* a trial point must be to be accepted as the next iterate.

In general, for an iteration to be considered successful we require that

$$\begin{aligned} &x_k + \tilde{\Delta}_k d_k \in \Omega \quad \text{and} \quad f(x_k + \tilde{\Delta}_k d_k) < f(x_k) - \rho(\Delta_k) \\ &\text{for some } d_k \in \mathcal{D}_k \quad \text{and} \quad \tilde{\Delta}_k \in [0, \Delta_k]. \end{aligned} \tag{4.2}$$

The function $\rho(\cdot)$ is called the *forcing function* and must satisfy Condition 4.

CONDITION 4 (general requirements on the forcing function).
1. The function $\rho(\cdot)$ is a nonnegative continuous function on $[0, +\infty)$.
2. The function $\rho(\cdot)$ is $o(t)$ as $t \downarrow 0$; i.e., $\lim_{t\downarrow 0} \rho(t)/t = 0$.
3. The function $\rho(\cdot)$ is nondecreasing; i.e., $\rho(t_1) \le \rho(t_2)$ if $t_1 \le t_2$.

Both $\rho(\Delta) \equiv 0$ and $\rho(\Delta) = \alpha\Delta^p$, where $\alpha > 0$ and $p > 1$, satisfy Condition 4. The first choice also requires globalization via a rational lattice, which is discussed in

section 4.2. The second choice can be used with globalization via a sufficient decrease condition, which is discussed in section 4.1.

In the case of a successful iteration (i.e., one that satisfies (4.2)), the next iterate is defined by

$$x_{k+1} = x_k + \tilde{\Delta}_k d_k \quad \text{for } k \in \mathcal{S}.$$

(Recall from section 1.1 that the set of indices of all successful iterations is denoted by $\mathcal{S}$.) In addition, Δ_k is updated according to

$$\Delta_{k+1} = \phi_k \Delta_k, \ \phi_k \geq 1 \quad \text{for } k \in \mathcal{S}.$$

The parameter ϕ_k is called the *expansion parameter*.

For the kth iteration to be unsuccessful, it must be the case that

$$x_k + \Delta_k d \notin \Omega \ \text{ or } \ f(x_k + \Delta_k d) \geq f(x_k) - \rho(\Delta_k) \text{ for every } d \in \mathcal{G}_k. \tag{4.3}$$

When the iteration is unsuccessful, the best point is unchanged:

$$x_{k+1} = x_k \quad \text{for } k \in \mathcal{U}.$$

(Recall from section 1.1 that the set of indices of all unsuccessful iterations is denoted by $\mathcal{U}$.) In addition, the step-length control parameter is reduced:

$$\Delta_{k+1} = \theta_k \Delta_k, \ \theta_k \in (0, 1) \quad \text{for } k \in \mathcal{U}.$$

The parameter θ_k is called the *contraction parameter*.

There are intimate connections between choosing the ϕ_k or θ_k in the update for Δ_k and guaranteeing that (4.1) holds. Further requirements depend on the particular choice of globalization strategy, and so are given in sections 4.1 and 4.2.

4.1. Globalization via a sufficient decrease condition. In the context of gradient-based nonlinear programming algorithms, the enforcement of a sufficient decrease condition on the step is well established (e.g., [10, 28, 29], or see the discussion in [15, section 2.2]). In the context of gradient-based methods, enforcing a sufficient decrease condition ties the choice of the step-length control parameter to the expected decrease, as estimated by the initial rate of decrease $-\nabla f(x_k)^T d_k$. In the context of GSS methods, the underlying assumption is that the value of $\nabla f(x_k)$ is unavailable—which means that the types of sufficient decrease conditions often used with gradient-based methods cannot be enforced. However, in [11] an alternative that uses the step-length control parameter, rather than $\nabla f(x_k)$, was introduced and analyzed in the context of linesearch methods for unconstrained minimization. In [21, 22, 23, 24], this basic concept was then extended to both unconstrained and constrained versions of what we here refer to as GSS methods. We now review the essential features of this approach.

Within the context of GSS methods for linearly constrained optimization, a sufficient decrease globalization strategy requires the following of the forcing function $\rho(\cdot)$ and the choice of the contraction parameter θ_k.

> CONDITION 5 (the forcing function for sufficient decrease).
> The forcing function $\rho(\cdot)$ is such that $\rho(t) > 0$ for $t > 0$.

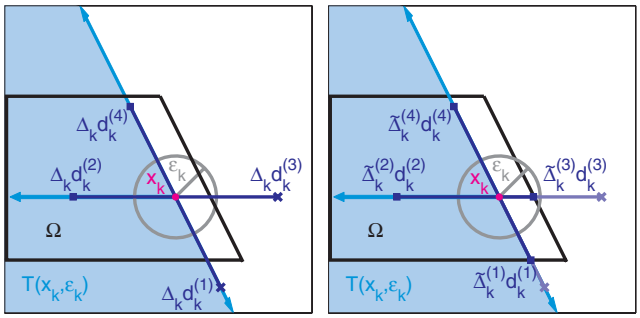


FIG. 4.1. *Observe in the second illustration that globalization via a sufficient decrease condition makes it possible to avoid infeasible trial points by simply stopping at the boundary of* Ω.

CONDITION 6 (contracting Δ_k for sufficient decrease).
A constant $\theta_{\max} < 1$ exists such that $\theta_k \leq \theta_{\max}$ for all k.

Full details are discussed in [15, section 3.7.1], but we include a few salient observations here. The requirements of Condition 5 are easily satisfied by choosing, say, $\rho(\Delta) = 10^{-4}\Delta^2$, while the requirements of Condition 6 are easily satisfied by choosing, say, $\theta_k = \frac{1}{2}$ for all k. The upper bound on the contraction factor θ_k ensures a predictable fraction of reduction on Δ_k at the conclusion of an unsuccessful iteration.

If a sufficient decrease condition is being employed, then we can use an alternative to the exact penalization strategy, given in (3.1), for choosing $\tilde{\Delta}_k^{(i)}$ when $x_k + \Delta_k d_k^{(i)} \notin \Omega$: simply find the step to the nearest constraint from x_k along $d_k^{(i)}$. This is a well-known technique in nonlinear programming (see, for instance, [10, section 5.2] or [28, section 15.4]). In other words, compute $\tilde{\Delta}_k^{(i)}$ as the maximum nonnegative feasible step along $d_k^{(i)}$. This option is illustrated in Figure 4.1.

4.2. Globalization via a rational lattice. Traditionally, direct search methods have relied on simple, as opposed to sufficient, decrease when accepting a step [33]. In other words, it is enough for the step $\tilde{\Delta}_k^{(i)} d_k^{(i)}$ to satisfy $f(x_k + \tilde{\Delta}_k^{(i)} d_k^{(i)}) < f(x_k)$. The trade-off is that when the condition for accepting a step is relaxed to admit simple decrease, further restrictions are required on the types of steps that are allowed. These restrictions are detailed in Conditions 7, 8, and 9.

CONDITION 7 (choosing the directions for the rational lattice).
Let $\mathbf{G} = \cup_{k=0}^{\infty} \mathcal{G}_k$.
1. The set $\mathbf{G}$ is finite and so can be written as $\mathbf{G} = \{g^{(1)}, \ldots, g^{(p)}\}$.
2. Every vector $g \in \mathbf{G}$ is of the form $g \in \mathbb{Z}^n$, where $\mathbb{Z}$ is the set of integers.
3. Every vector $h \in \mathcal{H}_k$ is of the form $h \in \mathbb{Z}^n$.

CONDITION 8 (expanding or contracting Δ_k for the rational lattice).
1. The scalar τ is a fixed rational number strictly greater than 1.
2. For all $k \in \mathcal{S}$, ϕ_k is of the form $\phi_k = \tau^{\ell_k}$, where $\ell_k \in \{0, \dots, L\}$, $L \geq 0$.
3. For all $k \in \mathcal{U}$, θ_k is of the form $\theta_k = \tau^{m_k}$, where $m_k \in \{M, \dots, -1\}$, $M \leq -1$.

CONDITION 9 (choosing the steps for the rational lattice).
$\tilde{\Delta}_k^{(i)}$ satisfies either $\tilde{\Delta}_k^{(i)} = 0$ or $\tilde{\Delta}_k^{(i)} = \tau^{\tilde{m}_k^{(i)}} \Delta_k$, where $\tilde{m}_k^{(i)} \in \{\tilde{M}, \dots, 0\}$, $\tilde{M} \leq 0$.

While the list of requirements in Conditions 7, 8, and 9 looks onerous, they can be satisfied in a straightforward fashion. A discussion of the reasons for these conditions can be found in [19, sections 3.4, 4, and 5]. (A detailed discussion of the rational lattice globalization strategy for the unconstrained case can be found in [15, section 3.7.2].) Here we make only a few pertinent observations.

First, a critical consequence of Conditions 7 and 8 is that when these two conditions are enforced, along with the exact penalization strategy in (3.1), Theorem 5.1 in [19] ensures that all iterates lie on a rational lattice. This fact plays a crucial role in guaranteeing (4.1) when only simple decrease is enforced. Condition 9 is a straightforward extension that preserves the fact that all the iterates lie on a rational lattice while relaxing the exact penalization strategy in (3.1) (an example is shown in Figure 4.2).

Obtaining a finite $\mathbf{G}$ to satisfy part 1 of Condition 7 can be done by following the procedure outlined in section 2.3 (i.e., if $I(x_{k_2}, \varepsilon_{k_2}) = I(x_{k_1}, \varepsilon_{k_1})$ for $k_2 > k_1$, then use the same generators for $T(x_{k_2}, \varepsilon_{k_2})$ as were used for $T(x_{k_1}, \varepsilon_{k_1})$). To satisfy part 2, a standard assumption in the context of simple decrease is that the linear constraints are rational, i.e., $A \in \mathbb{Q}^{m \times n}$, where $\mathbb{Q}$ denotes the set of rational numbers. By clearing denominators, it is then possible—with some care—to obtain a set of integral vectors to generate all possible ε-tangent cones; see [19, section 8] for further discussion. Part 3 is enforced directly.

In Condition 8, the usual choice of τ is 2. The parameter ϕ_k typically is chosen to be 1 so that $\ell_k = 0$ for all k, satisfying the requirement placed on ϕ_k in Condition 8. Usually θ_k is chosen to be $\frac{1}{2}$ so that $m_k = -1$ for all k, satisfying the requirement placed on θ_k in Condition 8. The fact that τ^{-1} is the largest possible choice of θ_k obviates the need to explicitly bound θ_k from above, as was required in Condition 6 for sufficient decrease.

Condition 9 says that it is possible to choose a partial step along a given direction so long as the trial point remains on a rational lattice. One strategy is illustrated in Figure 4.2. Starting with the situation illustrated on the left, along direction $d^{(1)}$, $\tilde{\Delta}_k^{(1)} = 0.5\Delta_k$ yields the feasible trial step $\tilde{\Delta}_k^{(1)} d^{(1)}$ while along direction $d^{(3)}$, $\tilde{\Delta}_k^{(3)} = 0.25\Delta_k$ yields the feasible trial step $\tilde{\Delta}_k^{(3)} d^{(3)}$, as illustrated on the right. These choices for $\tilde{\Delta}_k^{(1)}$ and $\tilde{\Delta}_k^{(3)}$ correspond to choosing $m_k^{(1)} = -1$ and $m_k^{(3)} = -2$, with $\tau = 2$ and $\tilde{M} = -2$.

The general strategy is to find the largest $\tilde{\Delta}_k^{(i)}$ (by finding the largest $m_k^{(i)}$) such that $x_k + \tilde{\Delta}_k^{(i)} d_k^{(i)} \in \Omega$ while satisfying Condition 9. To do so, either reduce Δ_k by a

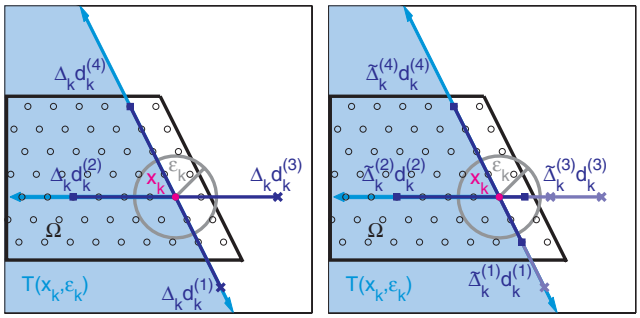


FIG. 4.2. *Globalization via a rational lattice means that the trial points lie on the rational lattice that exists as a consequence of Conditions 7–9. For this example note that while the two reduced steps are near the boundary, the requirement that they remain on the rational lattice means that they may not be on the boundary.*

factor of $1/\tau$ until

$$x_k + \tau^{m_k^{(i)}} \Delta_k d_k^{(i)} \in \Omega \quad \text{with } m_k^{(i)} \geq \tilde{M} \tag{4.4}$$

or set $\tilde{\Delta}_k^{(i)} = 0$ if it is not possible to satisfy (4.4) (for instance, when x_k is on the boundary of the feasible region then *any* step along $d_k^{(i)}$ would be infeasible).

5. GSS algorithms for linearly constrained problems. We now formally state two GSS algorithms for solving linearly constrained optimization problems. The fundamental requirement for both algorithms is that at every iteration k, the set of search directions $\mathcal{D}_k$ must include a set of generators $\mathcal{G}_k$ for the ε-normal cone $T(x_k, \varepsilon_k)$—hence the name *generating set search* methods. The primary requirements on the GSS methods presented here are that they satisfy Conditions 1, 2, 3, and 4. The differences in the two versions given depend on the type of globalization that is used: sufficient decrease in Algorithm 5.1 versus simple decrease in Algorithm 5.2. Sufficient decrease requires Conditions 5 and 6. Simple decrease admits the choice $\rho(\cdot) \equiv 0$, but requires Conditions 7, 8, and 9 in lieu of Conditions 5 and 6.

New in the statements of Algorithms 5.1 and 5.2, and to the analysis that follows, is the way in which ε_k is defined, which has bearing on the construction of the critical set $\mathcal{G}_k \subseteq \mathcal{D}_k$. Here we set $\varepsilon_k = \min\{\varepsilon_{\max}, \beta_{\max}\Delta_k\}$. This selection of ε_k differs from that used in either [19] or [24]. Specifically, in [19] and Algorithm 2 of [24]—as well as earlier in [25], in a slightly restricted form—$\mathcal{G}_k$ is required to contain generators for $T(x_k, \varepsilon)$ for all ε in the interval $[0, \varepsilon_{\max}]$, with $\varepsilon_{\max} > 0$. This means that $\mathcal{G}_k$ may need to contain generators for multiple cones rather than a single cone. Since Δ_k can be reduced only at the conclusion of an unsuccessful iteration, and an unsuccessful iteration requires the verification of (4.3), there is practical incentive to try and keep the cardinality of $\mathcal{G}_k$ manageable when the cost of computing $f(x)$ for $x \in \Omega$ is appreciable. Thus, Algorithm 1 in [24] first introduced the potential for a smaller set of search directions: the set of search directions must exactly generate $T(x_k, \varepsilon_k)$—and *only* $T(x_k, \varepsilon_k)$. Using our notation, this means that $\mathcal{H}_k = \emptyset$ for all k. Furthermore, for Algorithm 1 in [24], ε_k is simply a parameter decreased at unsuccessful iterations as opposed to the particular choice of ε_k given here.

Our requirement that the search directions include generators for $T(x_k, \varepsilon_k)$, with $\varepsilon_k = \min\{\varepsilon_{\max}, \beta_{\max}\Delta_k\}$, is a compromise. On the one hand, it may significantly

ALGORITHM 5.1. LINEARLY CONSTRAINED GSS USING A SUFFICIENT DECREASE GLOBALIZATION STRATEGY

INITIALIZATION.

Let $x_0 \in \Omega$ be the initial guess.

Let $\Delta_{\text{tol}} > 0$ be the tolerance used to test for convergence.

Let $\Delta_0 > \Delta_{\text{tol}}$ be the initial value of the step-length control parameter.

Let $\varepsilon_{\max} > \beta_{\max}\Delta_{\text{tol}}$ be the maximum distance used to identify nearby constraints ($\varepsilon_{\max} = +\infty$ is permissible).

Let $\rho(\cdot)$ be a forcing function satisfying Conditions 4 and 5.

ALGORITHM. For each iteration $k = 0, 1, 2, \ldots$

STEP 1. Let $\varepsilon_k = \min\{\varepsilon_{\max}, \beta_{\max}\Delta_k\}$. Choose a set of search directions $\mathcal{D}_k = \mathcal{G}_k \cup \mathcal{H}_k$ satisfying Conditions 1 and 2.

STEP 2. If there exists $d_k \in \mathcal{D}_k$ and a corresponding $\tilde{\Delta}_k \in [0, \Delta_k]$ satisfying Condition 3 such that $x_k + \tilde{\Delta}_k d_k \in \Omega$ and

$$f(x_k + \tilde{\Delta}_k d_k) < f(x_k) - \rho(\Delta_k),$$

then:

- Set $x_{k+1} = x_k + \tilde{\Delta}_k d_k$.
- Set $\Delta_{k+1} = \phi_k \Delta_k$ for any choice of $\phi_k \geq 1$.

STEP 3. Otherwise, for every $d \in \mathcal{G}_k$, either $x_k + \Delta_k d \notin \Omega$ or

$$f(x_k + \Delta_k d) \geq f(x_k) - \rho(\Delta_k).$$

In this case:

- Set $x_{k+1} = x_k$ (no change).
- Set $\Delta_{k+1} = \theta_k \Delta_k$ for some choice $\theta_k \in (0, 1)$ satisfying Condition 6.

If $\Delta_{k+1} < \Delta_{\text{tol}}$, then terminate.

FIG. 5.1. *Linearly constrained GSS using a sufficient decrease globalization strategy.*

decrease the number of directions in $\mathcal{G}_k$ over that needed when $\mathcal{G}_k$ is required to contain generators for $T(x_k, \varepsilon)$ for all ε in the interval $[0, \varepsilon_{\max}]$. On the other hand, it allows $\mathcal{H}_k \neq \emptyset$—the set of search directions can be augmented in an effort to accelerate the search—without adversely affecting the convergence guarantees for the algorithm.

Yoking the value of ε_k to the value of Δ_k has geometrical motivations. Once Δ_k is small enough, so that $\varepsilon_k = \beta_{\max}\Delta_k$, full steps along directions in $\mathcal{G}_k$ will be feasible, as Figure 2.1 demonstrates.

There is an intuitive practical appeal to allowing—while not requiring—$\mathcal{D}_k$ to

ALGORITHM 5.2. LINEARLY CONSTRAINED GSS USING A RATIONAL LATTICE GLOBALIZATION STRATEGY

INITIALIZATION.

Let $x_0 \in \Omega$ be the initial guess.

Let $\Delta_{\text{tol}} > 0$ be the tolerance used to test for convergence.

Let $\Delta_0 > \Delta_{\text{tol}}$ be the initial value of the step-length control parameter.

Let $\varepsilon_{\max} > \beta_{\max}\Delta_{\text{tol}}$ be the maximum distance used to identify nearby constraints ($\varepsilon_{\max} = +\infty$ is permissible).

Let $\rho(\cdot)$ be a forcing function satisfying Condition 4, e.g., $\rho(\cdot) \equiv 0$ is typical.

ALGORITHM. For each iteration $k = 0, 1, 2, \ldots$

STEP 1. Let $\varepsilon_k = \min\{\varepsilon_{\max}, \beta_{\max}\Delta_k\}$. Choose a set of search directions $\mathcal{D}_k = \mathcal{G}_k \cup \mathcal{H}_k$ satisfying Conditions 1, 2, and 7.

STEP 2. If there exists $d_k \in \mathcal{D}_k$ and a corresponding $\tilde{\Delta}_k \in [0, \Delta_k]$ satisfying Conditions 3 and 9 such that $x_k + \tilde{\Delta}_k d_k \in \Omega$ and

$$f(x_k + \tilde{\Delta}_k d_k) < f(x_k) - \rho(\Delta_k),$$

then:

- Set $x_{k+1} = x_k + \tilde{\Delta}_k d_k$.
- Set $\Delta_{k+1} = \phi_k \Delta_k$ for a choice of $\phi_k \geq 1$ satisfying Condition 8.

STEP 3. Otherwise, for every $d \in \mathcal{G}_k$, either $x_k + \Delta_k d \notin \Omega$ or

$$f(x_k + \Delta_k d) \geq f(x_k) - \rho(\Delta_k).$$

In this case:

- Set $x_{k+1} = x_k$ (no change).
- Set $\Delta_{k+1} = \theta_k \Delta_k$ for some choice $\theta_k \in (0, 1)$ satisfying Condition 8.

If $\Delta_{k+1} < \Delta_{\text{tol}}$, then terminate.

FIG. 5.2. *Linearly constrained GSS using a rational lattice globalization strategy.*

include more search directions. Note that if $T(x_k, \varepsilon_k) \neq \{\mathbf{0}\}$, then the directions in $\mathcal{G}_k$ will move the search along directions that are in some sense "parallel" (the situation is more complicated for $n > 2$) to the faces of the polyhedron that have been identified by the working set. This is best seen in the illustration on the left in Figure 2.1. Intuitively, it makes sense to also allow the search to move *toward* the faces of the polyhedron that have been identified by the working set—particularly

when the solution lies on the boundary of the feasible region. Such intuition is borne out by the numerical results reported in [17].

Before proceeding, we note a technical difference between the presentation of the algorithms in Algorithms 5.1 and 5.2 and what is assumed for the analysis in section 6. In practice, GSS algorithms terminate when the step-length control parameter Δ_k falls below a given threshold $\Delta_{\text{tol}} > 0$. Because this is important to any implementation, we have included it in the statement of the algorithm. In Theorems 6.3, 6.4, and 6.5, however, we assume that the iterations continue ad infinitum (i.e., in the context of the analysis, the reader should assume $\Delta_{\text{tol}} = 0$).

5.1. GSS using a sufficient decrease condition. A linearly constrained GSS algorithm based on a sufficient decrease globalization strategy is presented in Algorithm 5.1. Using a sufficient decrease globalization strategy, as outlined in section 4.1, requires that we enforce two particular conditions. Condition 5 ensures that $\rho(\Delta_k) = 0$ only when $\Delta_k = 0$. Condition 6 ensures that there is sufficient reduction on Δ_k at unsuccessful iterations.

The only assumption on f necessary to show that some subsequence of $\{\Delta_k\}$ converges to zero is that f be bounded below in the feasible region.

THEOREM 5.1 (see Theorem 3.4 of [15]). *Suppose f is bounded below on Ω. Then for a linearly constrained GSS method using a sufficient decrease globalization strategy satisfying Conditions* 4, 5, *and* 6 (*as outlined in Algorithm* 5.1), $\liminf_{k\to\infty} \Delta_k = 0$.

5.2. GSS using a rational lattice. A linearly constrained GSS algorithm based on a rational lattice globalization strategy is presented in Algorithm 5.2. The choice $\rho(\cdot) \equiv 0$ is standard for the rational lattice globalization strategy, which means only simple decrease, i.e., $f(x_k + \tilde{\Delta}_k d_k) < f(x_k)$, is required. We note, however, that a sufficient decrease condition may be employed in conjunction with a rational lattice globalization strategy; see [15, section 3.7.2]. The choice $\rho(\cdot) \equiv 0$ also means that Condition 4 is satisfied automatically. The trade-off for using simple decrease is that additional conditions must be imposed on the choice of admissible $\mathcal{D}_k$ (Condition 7), ϕ_k and θ_k (Condition 8), and $\tilde{\Delta}_k$ (Condition 9).

Using a rational lattice globalization strategy, to show that some subsequence of the step-length control parameters goes to zero, the only assumption placed on f is that the set $\mathcal{F} = \{\, x \in \Omega \mid f(x) \le f(x_0) \,\}$ be bounded. This is a stronger condition on f than is needed when using a sufficient decrease globalization strategy, where all that is required is that f be bounded below. The analysis for the rational lattice globalization strategy requires the sequence $\{x_k\}$ to remain in a bounded set so as to ensure that there is a finite number of lattice points to consider. We could adopt this weaker assumption, though it is not clear how it would be enforced in practice. Instead, assuming that $\mathcal{F}$ is bounded guarantees this requirement.

THEOREM 5.2 (see Theorem 6.5 of [19]). *Assume that $\mathcal{F} = \{x \in \Omega \mid f(x) \le f(x_0)\}$ is bounded and that $A \in \mathbb{Q}^{m\times n}$, where $\mathbb{Q}$ denotes the set of rational numbers. Then for a linearly constrained GSS method using a rational lattice globalization strategy satisfying Conditions* 4, 7, 8, *and* 9 (*as outlined in Algorithm* 5.2), $\liminf_{k\to\infty} \Delta_k = 0$.

6. Stationarity results. At *unsuccessful* iterations of the linearly constrained GSS methods outlined in Algorithms 5.1 and 5.2, we can bound the measure of stationarity $\chi(x_k)$ in terms of Δ_k. To do so, we make the following assumptions.

ASSUMPTION 6.1. *The set* $\mathcal{F} = \{ x \in \Omega \mid f(x) \leq f(x_0) \}$ *is bounded.*

ASSUMPTION 6.2. *The gradient of* f *is Lipschitz continuous with constant* M *on* Ω.

If Assumptions 6.1 and 6.2 hold, then there exists $\gamma > 0$ such that for all $x \in \mathcal{F}$,

$$\| \nabla f(x) \| < \gamma. \tag{6.1}$$

We then have the following results for the algorithms in Algorithms 5.1 and 5.2. Recall from section 2.1 that given a convex cone K and any vector v, we denote the projection of v onto K by v_K.

THEOREM 6.3. *Suppose that Assumption* 6.2 *holds. Consider the linearly constrained GSS algorithms given in Algorithms* 5.1 *and* 5.2*, both of which satisfy Conditions* 1*,* 2*, and* 3*. If* $k \in \mathcal{U}$ *and* ε_k *satisfies* $\varepsilon_k = \beta_{\max}\Delta_k$*, then*

$$\| [-\nabla f(x_k)]_{T(x_k,\varepsilon_k)} \| \leq \frac{1}{\kappa_{\min}} \left(M\Delta_k\beta_{\max} + \frac{\rho(\Delta_k)}{\Delta_k\beta_{\min}} \right). \tag{6.2}$$

Here, $\kappa_{\min}$ *is from Condition* 1*,* M *is from Assumption* 6.2*, and* $\beta_{\max}$ *and* $\beta_{\min}$ *are from Condition* 2*.*

Proof. Clearly, we need only consider the case when $[-\nabla f(x_k)]_{T(x_k,\varepsilon_k)} \neq 0$. Condition 1 guarantees a set $\mathcal{G}_k$ that generates $T(x_k, \varepsilon_k)$. By (2.1) (with $K = T(x_k, \varepsilon_k)$ and $v = -\nabla f(x_k)$) there exists some $\hat{d} \in \mathcal{G}_k$ such that

$$\kappa(\mathcal{G}_k) \, \| [-\nabla f(x_k)]_{T(x_k,\varepsilon_k)} \| \, \| \hat{d} \| \leq -\nabla f(x_k)^T \hat{d}. \tag{6.3}$$

Condition 3 and the fact that iteration k is unsuccessful tell us that

$$f(x_k + \Delta_k d) \geq f(x_k) - \rho(\Delta_k) \quad \text{for all } d \in \mathcal{G}_k \quad \text{for which } x_k + \Delta_k d \in \Omega.$$

Condition 2 ensures that for all $d \in \mathcal{G}_k$, $\| \Delta_k d \| \leq \Delta_k \beta_{\max}$ and, by assumption, $\Delta_k\beta_{\max} = \varepsilon_k$, so we have $\| \Delta_k d \| \leq \varepsilon_k$ for all $d \in \mathcal{G}_k$. Proposition 2.2 then assures us that $x_k + \Delta_k d \in \Omega$ for all $d \in \mathcal{G}_k$. Thus,

$$f(x_k + \Delta_k d) - f(x_k) + \rho(\Delta_k) \geq 0 \quad \text{for all } d \in \mathcal{G}_k. \tag{6.4}$$

Meanwhile, since the gradient of f is assumed to be continuous (Assumption 6.2), we can apply the mean value theorem to obtain, for some $\alpha_k \in (0, 1)$,

$$f(x_k + \Delta_k d) - f(x_k) = \Delta_k \nabla f(x_k + \alpha_k \Delta_k d)^T d \quad \text{for all } d \in \mathcal{G}_k.$$

Putting this together with (6.4),

$$0 \leq \Delta_k \nabla f(x_k + \alpha_k \Delta_k d)^T d + \rho(\Delta_k) \quad \text{for all } d \in \mathcal{G}_k.$$

Dividing through by Δ_k and subtracting $\nabla f(x_k)^T d$ from both sides yields

$$-\nabla f(x_k)^T d \leq \left(\nabla f(x_k + \alpha_k \Delta_k d) - \nabla f(x_k)\right)^T d + \rho(\Delta_k)/\Delta_k \quad \text{for all } d \in \mathcal{G}_k.$$

Since $\nabla f(x)$ is Lipschitz continuous (Assumption 6.2) and $0 < \alpha_k < 1$, we obtain

$$-\nabla f(x_k)^T d \le M\Delta_k \| d \|^2 + \rho(\Delta_k)/\Delta_k \quad \text{for all } d \in \mathcal{G}_k. \tag{6.5}$$

Since (6.5) holds for all $d \in \mathcal{G}_k$, (6.3) tells us that for some $\hat{d} \in \mathcal{G}_k$,

$$\kappa(\mathcal{G}_k) \, \| \, [-\nabla f(x_k)]_{T(x_k,\varepsilon_k)} \, \| \le M\Delta_k \| \, \hat{d} \, \| + \frac{\rho(\Delta_k)}{\Delta_k \, \| \, \hat{d} \, \|}.$$

Using the bounds on $\| \, \hat{d} \, \|$ in Condition 2,

$$\| \, [-\nabla f(x_k)]_{T(x_k,\varepsilon_k)} \, \| \le \frac{1}{\kappa(\mathcal{G}_k)} \left(M\Delta_k \beta_{\max} + \frac{\rho(\Delta_k)}{\Delta_k \beta_{\min}} \right).$$

The theorem then follows from the fact that $\kappa(\mathcal{G}_k) \ge \kappa_{\min}$ (Condition 1). ∎

Theorem 6.4 relates the measure of stationarity $\chi(x_k)$ to the step-length control parameter Δ_k. Before we proceed, we define the following constant (recall that $\kappa(\cdot)$ is defined in (2.1)):

$$\nu_{\min} = \min \left\{ \kappa(\mathcal{A}) : \mathcal{A} = \cup_{i \in I(x,\varepsilon)} \{a_i\}, \; x \in \Omega, \; \varepsilon \ge 0, \; I(x,\varepsilon) \ne \emptyset \right\} > 0. \tag{6.6}$$

We know that $\nu_{\min} > 0$ because there are no more than 2^m possibilities for $\mathcal{A}$.

THEOREM 6.4. *Suppose that Assumptions* 6.1 *and* 6.2 *hold. Consider the linearly constrained GSS algorithms given in Algorithms* 5.1 *and* 5.2*, both of which satisfy Conditions* 1*,* 2*, and* 3*. If $k \in \mathcal{U}$ and $\varepsilon_k = \beta_{\max}\Delta_k$, then*

$$\chi(x_k) \le \left(\frac{M}{\kappa_{\min}} + \frac{\gamma}{\nu_{\min}} \right) \Delta_k \, \beta_{\max} + \frac{1}{\kappa_{\min} \, \beta_{\min}} \frac{\rho(\Delta_k)}{\Delta_k}. \tag{6.7}$$

Here, $\kappa_{\min}$ is from Condition 1*, $\nu_{\min}$ is from* (6.6)*, M is from Assumption* 6.2*, γ is from* (6.1)*, and $\beta_{\max}$ and $\beta_{\min}$ are from Condition* 2*.*

Proof. Since $\varepsilon_k = \Delta_k \beta_{\max}$, Proposition B.2 tells us that

$$\chi(x_k) \le \| \, [-\nabla f(x_k)]_{T(x_k,\varepsilon_k)} \, \| + \frac{\Delta_k \beta_{\max}}{\nu_{\min}} \, \| \, [-\nabla f(x_k)]_{N(x_k,\varepsilon_k)} \, \|.$$

Furthermore, the bound on $\| \, [-\nabla f(x_k)]_{T(x_k,\varepsilon_k)} \, \|$ from Theorem 6.3 holds. The projection onto convex sets is contractive, so $\| \, [-\nabla f(x_k)]_{N(x_k,\varepsilon_k)} \, \| \le \| \, \nabla f(x_k) \, \|$. Under Assumptions 6.1 and 6.2, (6.1) holds, so $\| \, [-\nabla f(x_k)]_{N(x_k,\varepsilon_k)} \, \| \le \gamma$. The result follows. ∎

If we choose either $\rho(\Delta) \equiv 0$ or $\rho(\Delta) = \alpha\Delta^p$ with $\alpha > 0$ and $p \ge 2$, then we obtain an estimate of the form $\chi(x_k) = O(\Delta_k)$.

The constants M, γ, and $\nu_{\min}$ in (6.7) are properties of the linearly constrained optimization problem. The remaining quantities—the bounds on the lengths of the search directions $\beta_{\min}$ and $\beta_{\max}$, as well as $\kappa_{\min}$—are under the control of the algorithm.

Before continuing, we observe that the Lipschitz assumption (Assumption 6.2) can be relaxed. A similar bound can be obtained assuming only continuous differentiability of f. Let ω denote the following modulus of continuity of $\nabla f(x)$: given $x \in \Omega$ and $r > 0$,

$$\omega(x,r) = \max \{ \| \, \nabla f(y) - \nabla f(x) \, \| \;\mid\; y \in \Omega, \; \| \, y - x \, \| \le r \}.$$

Then the proof of Theorem 6.4 yields the bound

$$\chi(x_k) \leq \frac{1}{\kappa_{\min}} \omega(x_k, \Delta_k \beta_{\max}) + \frac{\gamma}{\nu_{\min}} \Delta_k \beta_{\max} + \frac{1}{\kappa_{\min} \beta_{\min}} \frac{\rho(\Delta_k)}{\Delta_k}.$$

Returning to Theorem 6.4, if we recall from Theorems 5.1 and 5.2 that the step-length control parameter Δ_k is manipulated explicitly by GSS methods in a way that ensures $\liminf_{k\to\infty} \Delta_k = 0$, then an immediate corollary is the following first-order convergence result.

THEOREM 6.5. *Suppose that Assumptions* 6.1 *and* 6.2 *hold. Consider either*

(i) *the linearly constrained GSS algorithm in Algorithm* 5.1*, which satisfies Conditions* 1*,* 2*,* 3*,* 4*,* 5*, and* 6*, or*

(ii) *the linearly constrained GSS algorithm in Algorithm* 5.2*, which satisfies Conditions* 1*,* 2*,* 3*,* 4*,* 7*,* 8*, and* 9*, with the additional assumption that A is rational.*

For both algorithms we have $\liminf_{k\to+\infty} \chi(x_k) = 0$.

7. Using Δ_k to terminate GSS methods after unsuccessful iterations. We now present some numerical illustrations of the practical implications of Theorem 6.4. We show that Δ_k can be used as a reasonable measure of stationarity when implementing GSS methods to solve linearly constrained minimization problems. The results in section 6 serve as a justification for terminating the search when $\Delta_k < \Delta_{\text{tol}}$.

To demonstrate that Δ_k is a reasonable measure of stationarity, we show the following results from experiments using an implementation of a GSS method for solving linearly constrained optimization problems (a thorough discussion of the implementation, as well as further numerical results, can be found in [17]).

The first test problem is the following quadratic program (QP) for $n = 8$:

$$\begin{array}{ll} \text{minimize} & f(x) = \sum_{j=1}^{n} j^2 x_j^2 \\ \text{subject to} & 0 \leq x \leq 1, \\ & \sum_{j=1}^{n} x_j \geq 1, \end{array} \tag{7.1}$$

where x_j is the jth component of the vector x. The last constraint is binding at the solution. The second test problem is posed on a pyramid in $\mathbb{R}^3$:

$$\begin{array}{ll} \text{minimize} & f(x) = \sum_{j=1}^{3} [(4-j)^2 (x_j - c_j)^2 - x_j] \\ \text{subject to} & x_3 \geq 0, \\ & x_1 + x_2 + x_3 \leq 1, \\ & x_1 - x_2 + x_3 \leq 1, \\ & -x_1 + x_2 + x_3 \leq 1, \\ & -x_1 - x_2 + x_3 \leq 1, \end{array} \tag{7.2}$$

with $c = (0.01, 0.01, 0.98)^T$. Again, x_j and c_j are the jth components of the vectors x and j, respectively. The solution is at c, which is near the apex of the pyramid. The algorithm actually visits the apex, which is a degenerate vertex insofar as there are four constraints in three variables that meet there.

These two problems were solved using the implementation of Algorithm 5.1 reported in [17]. The forcing function was $\rho(\Delta) = 10^{-4}\Delta^2$. The set of search directions $\mathcal{D}_k$ contained both the set $\mathcal{G}_k$, the generators for the ε-tangent cone $T(x_k, \varepsilon_k)$, as well as the set $\mathcal{H}_k$, which contained the nonzero generators for the ε-normal cone $N(x_k, \varepsilon_k)$.

All search directions were normalized, so $\beta_{\min} = \beta_{\max} = 1$. For these choices, Theorem 6.4 says that $\chi(x_k) = O(\Delta_k)$ at unsuccessful iterations when $\Delta_k \leq \varepsilon_{\max}$.

We used $\theta_k = \frac{1}{2}$ and $\phi_k = 1$ for all k. After any unsuccessful iteration, we recorded the value of Δ_k and computed the value of $\chi(x_k)$. These values are reported in Table 7.1 for unsuccessful iterations with $\varepsilon_k = \Delta_k \beta_{\max}$.

TABLE 7.1
GSS *runs showing decrease in* Δ_k *versus the value of* $\chi(x_k)$ *at unsuccessful iterations.*

Δ_k	$\chi(x_k)$
0.100000000000	0.762038045731
0.050000000000	0.719781449029
0.025000000000	0.683858024464
0.012500000000	0.522963684221
0.006250000000	0.147769116216
0.003125000000	0.009094010555
0.001562500000	0.009042346694
0.000781250000	0.005424114678
0.000390625000	0.002291442563
0.000195312500	0.000803137090
0.000097656250	0.000616656194
0.000048828125	0.000583197890
0.000024414063	0.000134935864
0.000012207031	0.000214535279
0.000006103516	0.000122058457
0.000003051758	0.000033834262
0.000001525879	0.000014798430
0.000000762939	0.000002976275
0.000000381470	0.000003506102
0.000000190735	0.000001047463

(a) The QP in (7.1).

Δ_k	$\chi(x_k)$
0.100000000000	0.009296268053
0.050000000000	0.009296268053
0.025000000000	0.068321041838
0.012500000000	0.001889009252
0.006250000000	0.000193017831
0.003125000000	0.000193017831
0.001562500000	0.003786874320
0.000781250000	0.003080612089
0.000390625000	0.000016499610
0.000195312500	0.000016499610
0.000097656250	0.000004481178
0.000048828125	0.000004481178
0.000024414063	0.000001550420
0.000012207031	0.000007616742
0.000006103516	0.000007616742
0.000003051758	0.000001501552
0.000001525879	0.000000807763
0.000000762939	0.000000008203
0.000000381470	0.000000008203
0.000000190735	0.000000008203

(b) The QP in (7.2).

The point of the results reported in Table 7.1 is not to demand close scrutiny of each entry but rather to demonstrate the trend in the quantities measured. We clearly see the linear relationship between Δ_k and $\chi(x_k)$ that Theorem 6.4 tells us to expect. These results are consistent with findings for the unconstrained case [8] as well as with a long-standing recommendation for using Δ_k as a stopping criterion for direct search methods (see [14, 3, 32]).

One practical benefit of using Δ_k as a measure of stationarity is that it is already present in GSS algorithms; no additional computation is required.

We close with the observation that the effectiveness of Δ_k as a measure of stationarity clearly depends on the value of the constants in the bound in (6.7). For instance, if f is highly nonlinear, so that the Lipschitz constant M is large, then using Δ_k to estimate $\chi(x_k)$ might be misleading. While GSS methods cannot control M, γ, or $\nu_{\min}$, which depend on the linearly constrained optimization problem, a careful implementation of GSS methods for solving linearly constrained optimization problems can control the remaining constants in (6.7). Thus a careful implementation can ensure that Δ_k is a useful measure of stationarity except when f is highly nonlinear (i.e., M

is large with respect to $\| \nabla f \|$) or A is ill-conditioned.

8. Conclusions. The results we have presented are useful in several ways. First, we present a new prescription for how the search directions should conform to the boundary near an iterate x_k. Theorems 6.3 and 6.4 bring out many of the elements common to the approaches described in [18, 19] and [23, 24]. Although the globalization approaches that ensure $\liminf_{k\to\infty} \Delta_k = 0$ differ, the same analysis shows that for both classes of algorithms,

$$\chi(x_k) = O(\Delta_k).$$

This result does not depend on the method of globalization.

Second, the results presented here give theoretical support for terminating GSS methods for linearly constrained optimization when Δ_k falls below some tolerance. Under the assumptions of Theorem 6.4, at the subsequence of unsuccessful iterations ($k \in \mathcal{U}$) we have $\chi(x_k) = O(\Delta_k)$ as $\Delta_k \to 0$. At the same time, Theorem 6.4 also suggests that this stopping criterion may be unsuitable if the objective is highly nonlinear, making clear the need for direct search methods, like all optimization algorithms, to account for scaling.

Theorem 6.3 underlies the use of linearly constrained GSS methods in the augmented Lagrangian framework given in [5]. The latter proceeds by successive approximate minimization of an augmented Lagrangian. The stopping criterion in the subproblems involves the norm of the projection onto $T(x_k, \omega_k)$ of the negative gradient of the augmented Lagrangian, for a parameter $\omega_k \downarrow 0$. In the direct search setting the gradient is unavailable. However, Theorem 6.3 enables us to use Δ_k as an alternative measure of stationarity in the subproblems. Details appear in [16].

Appendix A. Criticality measure for first-order constrained stationarity. Here we discuss $\chi(x)$ and $\| q(x) \|$ in more detail. Because these measures are not novel, we have relegated their discussion to an appendix.

For $x \in \Omega$, progress toward a KKT point of (1.1) is measured by

$$\chi(x) \equiv \max_{\substack{x+w\in\Omega \\ \| w \| \le 1}} -\nabla f(x)^T w. \tag{A.1}$$

This measure was originally proposed in [5] and is discussed at length in section 12.1.4 of [6], where the following properties are noted:

1. $\chi(x)$ is continuous,
2. $\chi(x) \ge 0$, and
3. $\chi(x) = 0$ if and only if x is a KKT point for (1.1).

Showing that $\chi(x_k) \to 0$ as $k \to \infty$ for a subsequence of iterates k constitutes a global first-order stationarity result.

To help better understand this measure, the w's that define $\chi(x)$ in (A.1) are illustrated in Figure A.1 for several choices of $-\nabla f(x)$. Conn, Gould, and Toint [6] observe that $\chi(x)$ can be interpreted as the progress that can be made on a first-order model at x in a ball of radius unity with the constraint of preserving feasibility. They go on to observe that $\chi(x)$ is a direct generalization of $\| \nabla f(x) \|$; in fact, $\chi(x) = \| \nabla f(x) \|$ whenever $\Omega = \mathbb{R}^n$ or $x - \nabla f(x) \in \Omega$.

The work in [19, 20] used the measure $q(x)$ defined in (1.5) (this quantity appears in [6] as equation (12.1.19)), but the resulting stationarity result is unsatisfying in the case of general linear constraints. The quantity $\chi(x)$ turns out to be easier to work with than $q(x)$. The latter involves a projection onto the feasible polyhedron, and if

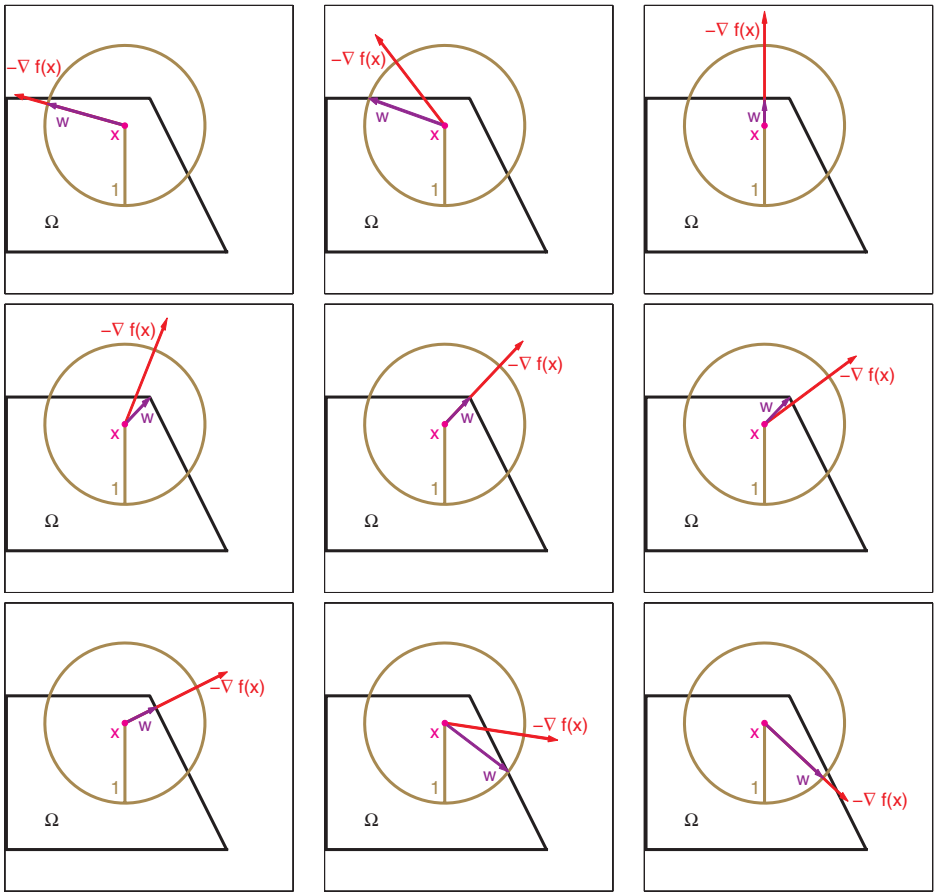


FIG. A.1. *How the* w *in* (A.1) *varies with* $-\nabla f(x)$ *when* $x - \nabla f(x) \notin \Omega$.

the constraints binding at the projection do not correspond to the constraints near x, technical difficulties ensue in relating $q(x)$ to the geometry of the feasible region near x. This is not the case with $\chi(x)$.

Appendix B. Geometric results on cones and polyhedra. Here we present geometrical results having to do with our use of $\chi(\cdot)$ as a measure of stationarity.

The first proposition says that if one can move from x to $x+v$ and remain feasible, then v cannot be too outward-pointing with respect to the constraints near x. Recall from section 2.1 that given a convex cone K and any vector v, there is a unique closest point of K to v, the *projection* of v onto K, which we denote by v_K. Thus $v_{N(x,\varepsilon)}$ is the projection of v onto the ε-normal cone $N(x,\varepsilon)$ while $v_{T(x,\varepsilon)}$ is the projection of v onto the ε-tangent cone $T(x,\varepsilon)$.

PROPOSITION B.1. *If* $x \in \Omega$ *and* $x + v \in \Omega$, *then for any* $\varepsilon \geq 0$, $\| v_{N(x,\varepsilon)} \| \leq \varepsilon/\nu_{\min}$, *where* $\nu_{\min}$ *is the constant from* (6.6).

Proof. Let $N = N(x,\varepsilon)$. The result is immediate if $v_N = 0$, so we need only consider the case when $v_N \neq 0$. Recall that N is generated by the outward-pointing normals to the binding constraints within distance ε of x; thus, the set $\mathcal{A} = \{\, a_i \mid i \in I(x,\varepsilon) \,\}$ generates N. A simple calculation shows that the distance

from x to $\{\ y\ \mid\ a_i^T y = b_i\ \}$ is $(b_i - a_i^T x)/\| a_i \|$, so it follows that

$$\frac{b_i - a_i^T x}{\| a_i \|} \leq \varepsilon \quad \text{for all } i \in I(x, \varepsilon).$$

Meanwhile, since $x + v \in \Omega$, we have

$$a_i^T x + a_i^T v \leq b_i \quad \text{for all } i.$$

The preceding two relations then lead to

$$a_i^T v \leq b_i - a_i^T x \leq \varepsilon \| a_i \| \quad \text{for all } i \in I(x, \varepsilon).$$

Since N is generated by $\mathcal{A} \subseteq \mathbf{A} = \{a_1, \ldots, a_m\}$ and $v_N \neq 0$, by (2.1) and (6.6),

$$\nu_{\min} \| v_N \| \leq \max_{i \in I(x,\varepsilon)} \frac{v^T a_i}{\| a_i \|} \leq \max_{i \in I(x,\varepsilon)} \frac{\varepsilon \| a_i \|}{\| a_i \|} = \varepsilon. \qquad \square$$

For $x \in \Omega$ and $v \in \mathbb{R}^n$, define

$$\hat{\chi}(x; v) = \max_{\substack{x+w \in \Omega \\ \| w \| \leq 1}} w^T v. \tag{B.1}$$

Note from (A.1) that $\chi(x) = \hat{\chi}(x; -\nabla f(x))$. We use v in (B.1) to emphasize that the following results are purely geometric facts about cones and polyhedra.

The following proposition relates $\hat{\chi}(x; v)$ to the projection of v onto the cones $T(x, \varepsilon)$ and $N(x, \varepsilon)$. Roughly speaking, it says that if $\varepsilon > 0$ is small, so that we are only looking at a portion of the boundary very near x, then the projection of v onto $T(x, \varepsilon)$ (i.e., the portion of v pointing into the interior of the feasible region) cannot be small unless $\hat{\chi}(x; v)$ is also small.

PROPOSITION B.2. *If $x \in \Omega$, then for all $\varepsilon \geq 0$,*

$$\hat{\chi}(x; v) \leq \| v_{T(x,\varepsilon)} \| + \frac{\varepsilon}{\nu_{\min}} \| v_{N(x,\varepsilon)} \|,$$

where $\nu_{\min}$ is the constant from (6.6).

Proof. Let $N = N(x, \varepsilon)$ and $T = T(x, \varepsilon)$. Writing v in terms of its polar decomposition, $v = v_N + v_T$, we obtain

$$\hat{\chi}(x; v) = \max_{\substack{x+w \in \Omega \\ \| w \| \leq 1}} w^T v \leq \max_{\substack{x+w \in \Omega \\ \| w \| \leq 1}} w^T v_T + \max_{\substack{x+w \in \Omega \\ \| w \| \leq 1}} w^T v_N.$$

For the first term on the right-hand side we have

$$\max_{\substack{x+w \in \Omega \\ \| w \| \leq 1}} w^T v_T \leq \| v_T \|.$$

Meanwhile, for any w we have

$$w^T v_N = (w_T + w_N)^T v_N \leq w_N^T v_N$$

since $w_T^T v_N \leq 0$. Thus,

$$\max_{\substack{x+w \in \Omega \\ \| w \| \leq 1}} w^T v_N \leq \max_{\substack{x+w \in \Omega \\ \| w \| \leq 1}} \| w_N \| \ \| v_N \|.$$

However, since $x + w \in \Omega$, Proposition B.1 tells us that

$$\| w_N \| \leq \frac{\varepsilon}{\nu_{\min}}.$$

Therefore,

$$\hat{\chi}(x; v) \leq \| v_T \| + \frac{\varepsilon}{\nu_{\min}} \| v_N \|. \qquad \square$$

Acknowledgments. We thank Margaret Wright, the associate editor who handled this paper, along with two anonymous referees, for the many useful comments that led to a significant improvement in this presentation.

REFERENCES

[1] C. Audet and J. E. Dennis, Jr., *Analysis of generalized pattern searches*, SIAM J. Optim., 13 (2003), pp. 889–903.
[2] A. Ben-Israel, A. Ben-Tal, and S. Zlobec, *Optimality in Nonlinear Programming: A Feasible Directions Approach*, John Wiley & Sons, New York, 1981.
[3] M. J. Box, D. Davies, and W. H. Swann, *Non-Linear Optimization Techniques*, ICI Monograph 5, Oliver & Boyd, Edinburgh, UK, 1969.
[4] C. G. Broyden, *A class of methods for solving nonlinear simultaneous equations*, Math. Comp., 19 (1965), pp. 577–593.
[5] A. R. Conn, N. Gould, A. Sartenaer, and P. L. Toint, *Convergence properties of an augmented Lagrangian algorithm for optimization with a combination of general equality and linear constraints*, SIAM J. Optim., 6 (1996), pp. 674–703.
[6] A. R. Conn, N. I. M. Gould, and P. L. Toint, *Trust-Region Methods*, MPS/SIAM Ser. Optim. 1, SIAM, Philadelphia, 2000.
[7] C. Davis, *Theory of positive linear dependence*, Amer. J. Math., 76 (1954), pp. 733–746.
[8] E. D. Dolan, R. M. Lewis, and V. J. Torczon, *On the local convergence properties of pattern search*, SIAM J. Optim., 14 (2003), pp. 567–583.
[9] K. Fukuda, *cdd and cddplus homepage*. From McGill University, Montreal, Canada, http://www.cs.mcgill.ca/∼fukuda/soft/cdd_home/cdd.html, 2005.
[10] P. E. Gill, W. Murray, and M. H. Wright, *Practical Optimization*, Academic Press, London, 1981.
[11] 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.
[12] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms,* I, Springer-Verlag, Berlin, 1993.
[13] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms*, II, Springer-Verlag, Berlin, 1993.
[14] R. Hooke and T. A. Jeeves, *Direct search solution of numerical and statistical problems*, J. Assoc. Comput. Mach., 8 (1961), pp. 212–229.
[15] T. G. Kolda, R. M. Lewis, and V. Torczon, *Optimization by direct search: New perspectives on some classical and modern methods*, SIAM Rev., 45 (2003), pp. 385–482.
[16] T. G. Kolda, R. M. Lewis, and V. Torczon, *A Generating Set Direct Search Augmented Lagrangian Algorithm for Optimization with a Combination of General and Linear Constraints*, Technical report SAND2006-5315, Sandia National Laboratories, Albuquerque, NM and Livermore, CA, 2006.
[17] R. M. Lewis, A. Shepherd, and V. Torczon, *Implementing Generating Set Search Methods for Linearly Constrained Minimization*, Technical report WM–CS–2005–01, Department of Computer Science, College of William & Mary, Williamsburg, VA, 2005.
[18] R. M. Lewis and V. Torczon, *Pattern search algorithms for bound constrained minimization*, SIAM J. Optim., 9 (1999), pp. 1082–1099.
[19] R. M. Lewis and V. Torczon, *Pattern search methods for linearly constrained minimization*, SIAM J. Optim., 10 (2000), pp. 917–941.
[20] 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.

[21] S. Lucidi and M. Sciandrone, *Numerical results for unconstrained optimization without derivatives*, in Nonlinear Optimization and Applications (Proceedings of the International School of Mathematics "G. Stampacchia" 21st Workshop, Erice, Italy, 1995), G. Di Pillo and F. Giannessi, eds., Kluwer Academic/Plenum Publishers, New York, 1996, pp. 261–270.

[22] S. Lucidi and M. Sciandrone, *A derivative-free algorithm for bound constrained optimization*, Comput. Optim. Appl., 21 (2002), pp. 119–142.

[23] S. Lucidi and M. Sciandrone, *On the global convergence of derivative-free methods for unconstrained optimization*, SIAM J. Optim., 13 (2002), pp. 97–116.

[24] S. Lucidi, M. Sciandrone, and P. Tseng, *Objective-derivative-free methods for constrained optimization*, Math. Program., 92 (2002), pp. 37–59.

[25] J. H. May, *Linearly Constrained Nonlinear Programming: A Solution Method That Does Not Require Analytic Derivatives*, Ph.D. thesis, Yale University, New Haven, CT, 1974.

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

[27] J.-J. Moreau, *Décomposition orthgonale d'un espace hilbertien selon deux cônes mutuellement polaires*, C. R. Acad. Sci. Paris, 255 (1962), pp. 238–240.

[28] S. G. Nash and A. Sofer, *Linear and Nonlinear Programming*, McGraw–Hill, New York, 1996.

[29] J. Nocedal and S. Wright, *Numerical Optimization*, Springer Series in Operations Research, Springer, New York, 1999.

[30] C. J. Price and I. D. Coope, *Frames and grids in unconstrained and linearly constrained optimization: A nonsmooth approach*, SIAM J. Optim., 14 (2003), pp. 415–438.

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

[32] W. H. Swann, *Direct search methods*, in Numerical Methods for Unconstrained Optimization, W. Murray, ed., Academic Press, London, New York, 1972, pp. 13–28.

[33] V. Torczon, *On the convergence of pattern search algorithms*, SIAM J. Optim., 7 (1997), pp. 1–25.

[34] W. Yu and Y. Li, *A direct search method by the local positive basis for linearly constrained optimization*, Chinese Ann. Math., 2 (1981), pp. 139–146.

[35] W. I. Zangwill, *Nonlinear Programming; A Unified Approach*, Prentice-Hall, Englewood Cliffs, NJ, 1969.

[36] G. Zoutendijk, *Mathematical Programming Methods*, North–Holland, Amsterdam, 1976.

SIAM J. OPTIM.
Vol. 17, No. 4, pp. 969–996
© 2006 Society for Industrial and Applied Mathematics

CONVEX APPROXIMATIONS OF CHANCE CONSTRAINED PROGRAMS*

ARKADI NEMIROVSKI† AND ALEXANDER SHAPIRO†

Abstract. We consider a chance constrained problem, where one seeks to minimize a convex objective over solutions satisfying, with a given close to one probability, a system of randomly perturbed convex constraints. This problem may happen to be computationally intractable; our goal is to build its computationally tractable approximation, i.e., an efficiently solvable deterministic optimization program with the feasible set contained in the chance constrained problem. We construct a general class of such convex conservative approximations of the corresponding chance constrained problem. Moreover, under the assumptions that the constraints are affine in the perturbations and the entries in the perturbation vector are independent-of-each-other random variables, we build a large deviation-type approximation, referred to as "Bernstein approximation," of the chance constrained problem. This approximation is convex and efficiently solvable. We propose a simulation-based scheme for bounding the optimal value in the chance constrained problem and report numerical experiments aimed at comparing the Bernstein and well-known scenario approximation approaches. Finally, we extend our construction to the case of ambiguous chance constrained problems, where the random perturbations are independent with the collection of distributions known to belong to a given convex compact set rather than to be known exactly, while the chance constraint should be satisfied for every distribution given by this set.

Key words. stochastic programming, chance constraints, convex programming, Monte Carlo sampling, scenario generation, large deviation bounds, ambiguous chance constrained programming

AMS subject classifications. 90C15, 90C25, 90C59

DOI. 10.1137/050622328

1. Introduction.

Let us consider the following optimization problem:

$$\operatorname*{Min}_{x\in X} f(x) \quad \text{subject to} \quad \operatorname{Prob}\big\{F(x,\xi)\le 0\big\}\ge 1-\alpha. \tag{1.1}$$

Here ξ is a random vector with probability distribution P supported on a set $\Xi\subset\mathbb{R}^d$, $X\subset\mathbb{R}^n$ is a nonempty convex set, $\alpha\in(0,1)$, $f:\mathbb{R}^n\to\mathbb{R}$ is a real valued convex function, $F=(f_1,\dots,f_m):\mathbb{R}^n\times\Xi\to\mathbb{R}^m$, and $\operatorname{Prob}(A)$ denotes probability of an event A. Probability constraints of the form appearing in (1.1) arise naturally in various applications and are called *chance* (or probabilistic) constraints. Such constraints can be viewed as a compromise with the requirement of enforcing the constraints $F(x,\xi)\le 0$ for *all* values $\xi\in\Xi$ of the uncertain data vector, which could be too costly or even impossible. Chance constrained optimization problems were introduced in Charnes, Cooper, and Symonds [8], Miller and Wagner [17], and Prékopa [21].

Aside from potential modelling problems with formulation (1.1) (e.g., the necessity to know the probability distribution of the random vector ξ, which in practice is not always easy), there could be serious problems with numerical processing of chance constraints. First, it may happen that the only way to check whether or not a given chance constraint is satisfied at a given point x is to use Monte Carlo simulation, and

*Received by the editors January 10, 2005; accepted for publication (in revised form) May 15, 2006; published electronically November 22, 2006.
http://www.siam.org/journals/siopt/17-4/62232.html
†Georgia Institute of Technology, Atlanta, GA 30332 (nemirovs@isye.gatech.edu, ashapiro@isye.gatech.edu). The first author's research was partly supported by the Binational Science Foundation grant 2002038. The second author's research was partly supported by the NSF grant DMS-0510324.

this becomes too costly when α is small. The second potential difficulty is that even with nice, say affine in x and in ξ, functions $F(x,\xi)$, the feasible set of a chance constraint may happen to be nonconvex, which makes optimization under this constraint highly problematic. It should be mentioned that there are generic situations where the latter difficulty does not occur. First, there exists a wide family of *logarithmically concave distributions* extensively studied by Prékopa [22]; he shows, in particular, that whenever the distribution of a random vector ξ is logarithmically concave, the feasible set of a chance constraint $\mathrm{Prob}\{\xi : Ax \geq \xi\} \geq 1-\epsilon$ (A is a deterministic matrix) or, more generally, the feasible set of a chance constraint $\mathrm{Prob}\{\xi : (x,\xi) \in X\} \geq 1-\epsilon$ (X is a deterministic convex set) is convex. There is also a recent result, due to Lagoa, Li, and Sznaier [16], which states that the feasible set of a scalar chance constraint

$$\mathrm{Prob}\left\{a^T x \leq b\right\} \geq 1-\epsilon \tag{1.2}$$

is convex, provided that the vector $(a^T, b)^T$ of the coefficients has symmetric logarithmically concave density and $\epsilon < 1/2$. Note, however, that in order to process a chance constraint efficiently, we need both efficient computability of the probability in question *and* the convexity of the corresponding feasible set. This combination seems to be a "rare commodity."[1] As far as chance constraint (1.2) is concerned, the only case known to us when both these requirements are satisfied is the one where the random vector $(a^T, b)^T$ is the image, under deterministic affine transformation, of a random vector with rotationally invariant distribution; cf. [16]. The simplest case of this situation is the one when $(a^T, b)^T$ is a normally distributed random vector. There are also other cases (see, e.g., [23, 11]) where a chance constraint can be processed efficiently, but in general the problem still persists; there are numerous situations where the chance constrained version of a randomly perturbed constraint $F(x,\xi) \leq 0$, even as simple-looking a one as the bilinear constraint (1.2), is "severely computationally intractable." Whenever this is the case, a natural course of action is to look for *tractable approximations* of the chance constraint, i.e., for efficiently verifiable *sufficient conditions* for its validity. In addition to being sufficient, such a condition should define a convex and "computationally tractable" set in the x-space, e.g., should be represented by a system of convex inequalities $G(x,u) \leq 0$ in x and, perhaps, in additional variables $u \in \mathbb{R}^s$, with efficiently computable $G(x,u)$. Whenever this is the case, the problem

$$\min_{x \in X,\, u \in \mathbb{R}^s} f(x) \quad \text{subject to} \quad G(x,u) \leq 0 \tag{1.3}$$

is a convex programming problem with efficiently computable objective and constraints and as such it is efficiently solvable.[2] This problem provides a *conservative approximation* of the chance constrained problem of interest, meaning that the projection of the feasible set of (1.3) onto the space of x-variables is contained in the feasible set of the chance constrained problem (1.1), so that an optimal solution to (1.3) is *feasible suboptimal* solution to (1.1).

A general way to build computationally tractable approximations (not necessarily conservative) of chance constrained problems is offered by the *scenario approach* based

[1] For example, let b in (1.2) be deterministic and a be uniformly distributed in the unit box. In this case, the feasible set of (1.2) is convex, provided that $\epsilon < 1/2$, but the left-hand side in (1.2) is difficult to compute: it is known (see Khachiyan [15]) that it cannot be computed within accuracy ϵ in time polynomial in $\dim a$ and $\ln(1/\epsilon)$, unless P=NP.

[2] For a detailed description of tractability issues in continuous optimization and their relation to convexity, see, e.g., [4, Chapter 5].

on Monte Carlo sampling techniques. That is, one generates a sample $\xi^1, \ldots, \xi^N$ of N (independent) realizations of the random vector ξ and approximates (1.1) with the problem

$$(P^N) \qquad \min_{x \in X} f(x) \quad \text{subject to} \quad F(x, \xi^\nu) \leq 0,\ \nu = 1, \ldots, N.$$

The main advantage of this approach is its generality: it imposes no restrictions on the distribution of ξ and on how the data enters the constraints. In order to build (P^N) there is no need even to know what the distribution of ξ is; all we need is to be able to sample from this distribution. Last, but not least, is the "tractability status" of the approximation. The approximation (P^N) is efficiently solvable, provided that the function $F(x, \xi)$ is componentwise convex in x and is efficiently computable, and the sample size N is not too large.

An important theoretical question related to the scenario approximation is the following. The approximation itself is random and its solution may not satisfy the chance constraints. The question is, How large should the sample size N be in order to ensure, with probability of at least $1 - \delta$, that the optimal solution to (P^N) is feasible for the problem of interest (1.1)? To some extent this question was resolved in recent papers of Calafiore and Campi [6, 7] and de Farias and Van Roy [10]. Their results were then extended in [14] to a more complicated case of *ambiguous* chance constraints (that is, the case when the "true" distribution of ξ is assumed to belong to a given family of distributions rather than to be known exactly, while the samples are drawn from a specified reference distribution). The answer to the outlined question, as given in [7], is that if $F(x, \xi)$ is componentwise convex in x, then, under mild additional conditions, with the sample size N satisfying

$$N \geq N^* := \text{Ceil}\left[2n\alpha^{-1}\log(12/\alpha) + 2\alpha^{-1}\log(2/\delta) + 2n\right], \tag{1.4}$$

the optimal solution to (P^N) is, with a probability of at least $1 - \delta$, feasible for the chance constrained problem (1.1). A remarkable feature of this result is that, similar to the scenario approximation itself it, is completely distribution-free.

Aside from the conservativeness (which is a common drawback of all approximations), an intrinsic drawback of the scenario approximation based on (1.4) is that, as is easily seen, the sample size N should be at least inverse proportional to the risk α and thus could be impractically large when the risk is small. Moreover, the sample size as given by (1.4) (and by all other known results of this type) grows linearly with n, which makes it difficult to apply the approach already to medium-size problems (with $\alpha = 0.01$ and $n = 200$, $\delta = 0.01$, the estimate (1.4) results in $N^* = 285,063$). Note that for a properly modified scenario approximation, "bad" dependence of N on α given by (1.4) can be replaced with

$$N = O(1)\left[\log(1/\delta) + dm^2\log(d\log(1/\alpha))\right], \tag{1.5}$$

provided that $F(x, \xi)$ is affine in ξ and ξ has a "nice" distribution, e.g., uniform in a box, or on the vertices of a box, or normal [19].

An alternative to the scenario approximation is an approximation based on "analytical" upper bounding of the probability for the randomly perturbed constraint $F(x, \xi) \leq 0$ to be violated. The simplest approximation scheme of this type was proposed in [2] for the case of a single affine in ξ inequality

$$f_0(x) + \sum_j \xi_j f_j(x) \leq 0 \tag{1.6}$$

(cf., (1.2)). Assuming that ξ_j are independent-of-each-other random variables with zero means varying in segments $[-\sigma_i, \sigma_i]$, it is easy to see that if x satisfies the constraint

$$f_0(x) + \Omega\left(\sum_{j=1}^{d} \sigma_j^2 f_j^2(x)\right)^{1/2} \leq 0, \tag{1.7}$$

where $\Omega > 0$ is a "safety" parameter, then x violates the randomly perturbed constraint (1.6) with probability of at most $\exp\{-\kappa\Omega^2\}$, where $\kappa > 0$ is an absolute constant (as we shall see in section 6, one can take $\kappa = 1/2$). It follows that if all components $f_i(x, \xi)$ are of the form

$$f_i(x, \xi) = f_{i0}(x) + \sum_{j=1}^{d} \xi_j f_{ij}(x), \tag{1.8}$$

then the optimization program

$$\min_{x \in X} f(x) \quad \text{subject to} \quad f_{i0}(x) + \Omega\left(\sum_{j=1}^{d} \sigma_j^2 f_{ij}^2(x)\right)^{1/2} \leq 0,\ i = 1, \ldots, m, \tag{1.9}$$

with $\Omega := \sqrt{2\log(m\alpha^{-1})}$, is an approximation of the chance constrained problem (1.1). This approximation is convex, provided that all $f_{ij}(x)$ are convex and every one of the functions $f_{ij}(x)$ with $j \geq 1$ is either affine or nonnegative. Another, slightly more convenient computationally, analytical approximation of randomly perturbed constraint (1.6) was proposed in [5]. Analytical approximations of more complicated chance constraints, notably a randomly perturbed conic quadratic inequality, are presented in [18]. An advantage of the "analytical" approach as compared to the scenario one is that the resulting approximations are deterministic convex problems with sizes independent of the required value of risk (reliability) α, so that these approximations remain practical also in the case of very small values of α. On the negative side, building an analytical approximation requires structural assumptions on $F(x, \xi)$ and on the stochastic nature of ξ (in all known constructions of this type, ξ_j should be independent of each other and possess "nice" distributions).

In this paper, we develop a new class of analytical approximations of chance constraints, referred to as *Bernstein* approximations.[3] Our major assumptions are that the components of $F(x, \xi)$ are of the form (1.8) with convex $f_{ij}(x)$, and ξ_j are independent of each other and possess distributions with efficiently computable moment generating functions. Besides this, we assume that for every $j \geq 1$ for which not all of the functions $f_{ij}(x)$, $i = 1, \ldots, m$, are affine, the corresponding random variable ξ_j is nonnegative. Under these assumptions, the approximation we propose is an explicit convex program.

After the initial version of this paper was released, we became aware of the paper of Pinter [20] proposing (although not in full generality) Bernstein approximation, even in its advanced "ambiguous" form (see section 6 below). The only (but, we believe, quite important) step ahead in what follows as compared to Pinter's paper is

[3] The construction is based on the ideas used by S. N. Bernstein when deriving his famous inequalities for probabilities of large deviations of sums of independent random variables.

that with our approach the natural scale parameter of Bernstein approximation ("h" in Pinter's paper) becomes a variable rather than an ad hoc chosen constant (as is the case in [20]). Specifically, we manage to represent Bernstein bound in a form which is jointly convex in the original decision variables *and* the scale parameter, which allows one to deal, staying all the time within the convex programming framework, with the bound which is pointwise optimized in the scale parameter.

The rest of the paper is organized as follows. In section 2 we introduce a class of convex conservative approximations of (1.1). Bernstein approximation of (1.1) is derived and discussed in section 3. In section 4, we propose a simple simulation-based scheme for bounding the true optimal value in (1.1), which allows one to evaluate numerically the quality (that is, the conservatism) of various approximations. In section 5, we report some preliminary numerical experiments with Bernstein approximation. Our numerical results demonstrate that this approximation compares favorably with the scenario one. In concluding section 6, we extend Bernstein approximation to the case of *ambiguous uncertainty model*, where the tuple of distributions of (mutually independent) components ξ_j of ξ is assumed to belong to a given convex compact set rather than to be known exactly (cf., [14], where similar extensions of the scenario approach are considered).

2. Convex approximations of chance constrained problems. In this section we discuss convex approximations of chance constrained problems of the form (1.1). As was mentioned in the introduction, chance constrained problems, even simple-looking ones, are often computationally intractable. A natural way to overcome, to some extent, this difficulty is to replace chance constraint problem (1.1) with a *tractable approximation*. That is, with an efficiently solvable problem of the form (1.3). To this end we require the function $G(x,u)$ to be *convex* in (x,u) and efficiently computable. We also would like the constraints $G(x,u) \le 0$ to be *conservative*, in the sense that if for $x \in X$ and u it holds that $G(x,u) \le 0$, then $\mathrm{Prob}\big\{F(x,\xi) \le 0\big\} \ge 1-\alpha$. Thus, feasible solutions to (1.3) induce feasible solutions to (1.1), so that the optimal solution of the approximation is a feasible suboptimal solution of the problem of interest. If these two conditions hold, we refer to (1.3) as a *convex conservative* approximation of the true problem (1.1). Our goal in this section is to construct a special class of convex conservative approximations.

Let us consider first the scalar case of $m = 1$, i.e., $F : \mathbb{R}^n \times \Xi \to \mathbb{R}$. Then the probabilistic (chance) constraint of problem (1.1) is equivalent to the constraint

$$p(x) := \mathrm{Prob}\big\{F(x,\xi) > 0\big\} \le \alpha. \tag{2.1}$$

By $\mathbb{1}_A$ we denote the indicator function of a set A, i.e., $\mathbb{1}_A(z) = 1$ if $z \in A$ and $\mathbb{1}_A(z) = 0$ if $z \notin A$.

Let $\psi : \mathbb{R} \to \mathbb{R}$ be a *nonnegative valued, nondecreasing, convex* function satisfying the following property:

$(*)$ $\psi(z) > \psi(0) = 1$ for any $z > 0$.

We refer to function $\psi(z)$ satisfying the above properties as a (one-dimensional) *generating function*. It follows from $(*)$ that for $t > 0$ and random variable Z,

$$\mathbb{E}[\psi(tZ)] \ge \mathbb{E}\left[\mathbb{1}_{[0,+\infty)}(tZ)\right] = \mathrm{Prob}\{tZ \ge 0\} = \mathrm{Prob}\{Z \ge 0\}.$$

By taking $Z = F(x,\xi)$ and changing t to t^{-1}, we obtain that

$$p(x) \le \mathbb{E}\left[\psi\left(t^{-1}F(x,\xi)\right)\right] \tag{2.2}$$

holds for all x and $t > 0$. Denote that

$$\Psi(x,t) := t\,\mathbb{E}\left[\psi\left(t^{-1}F(x,\xi)\right)\right]. \tag{2.3}$$

We obtain that if there exists $t > 0$ such that $\Psi(x,t) \le t\alpha$, then $p(x) \le \alpha$. In fact this observation can be strengthened to

$$\inf_{t>0}\left[\Psi(x,t) - t\alpha\right] \le 0 \quad \text{implies} \quad p(x) \le \alpha. \tag{2.4}$$

Indeed, let us fix x and set $\phi(t) := \Psi(x,t) - t\alpha$, $Z := F(x,\xi)$. It may happen (case (A)) that $\mathrm{Prob}\{Z > 0\} > 0$. Then there exist $a, b > 0$ such that $\mathrm{Prob}\{Z \ge a\} \ge b$, whence

$$\Psi(x,t) = t\,\mathbb{E}\left[\psi(t^{-1}F(x,\xi))\right] \ge tb\psi(t^{-1}a) \ge tb\left[\psi(0) + (\psi(a) - \psi(0))/t\right]$$

provided that $0 < t < 1$ (we have taken into account that $\psi(\cdot)$ is convex). Since $\psi(a) > \psi(0)$, we conclude that

$$\Psi(x,t) \ge \gamma := b(\psi(a) - \psi(0)) > 0 \quad \text{for } 0 < t < 1,$$

and hence $\liminf_{t\to+0}\phi(t) > 0$. Further, we have

$$\liminf_{t\to\infty}\mathbb{E}\left[\psi(t^{-1}Z)\right] \ge \psi(0) \ge 1,$$

and hence $\liminf_{t\to\infty}\phi(t) = \infty$ due to $\alpha \in (0,1)$. Finally, $\phi(t)$ is clearly lower semicontinuous in $t > 0$. We conclude that if (A) is the case, then $\inf_{t>0}\phi(t) \le 0$ iff there exists $t > 0$ such that $\phi(t) \le 0$, and in this case, as we already know, $p(x)$ indeed is $\le \alpha$. And if (A) is not the case, then the conclusion in (2.4) is trivially true, so that (2.4) is true.

We see that the inequality

$$\inf_{t>0}\left[\Psi(x,t) - t\alpha\right] \le 0 \tag{2.5}$$

is a conservative approximation of (2.1)—whenever (2.5) is true, so is (2.1). Moreover, assume that for every $\xi \in \Xi$ the function $F(\cdot,\xi)$ is convex. Then $G(x,t) := \Psi(x,t) - t\alpha$ is convex. Indeed, since $\psi(\cdot)$ is nondecreasing and convex and $F(\cdot,\xi)$ is convex, it follows that $(x,t) \mapsto t\psi(t^{-1}F(x,t))$ is convex[4]. This, in turn, implies convexity of the expected value function $\Psi(x,t)$, and hence convexity of $G(x,t)$.

We obtain, under the assumption that X, $f(\cdot)$ and $F(\cdot,\xi)$ are convex, that

$$\min_{x\in X,\, t>0} f(x) \quad \text{subject to} \quad \inf_{t>0}\left[\Psi(x,t) - t\alpha\right] \le 0 \tag{2.6}$$

gives a *convex* conservative approximation of the chance constrained problem (1.1).

Clearly the above construction depends on a choice of the generating function $\psi(z)$. This raises the question of what would be a "best" choice of $\psi(z)$. If we consider this question from the point of view of a better (tighter) approximation of the corresponding chance constraints, then the smaller is $\psi(\cdot)$, the better is bound

[4]We have used the well-known fact that if $f(x)$ is convex, so is the function $g(x,t) = tf(t^{-1}x)$, $t > 0$. Indeed, given x', x'', $\lambda \in (0,1)$, and $t', t'' > 0$, and setting $t = \lambda t' + (1-\lambda)t''$, $x = \lambda x' + (1-\lambda)x''$, we have $\lambda t' f(x'/t') + (1-\lambda)t'' f(x''/t'') = t\left[\lambda t' t^{-1} f(x'/t') + (1-\lambda)t'' t^{-1} f(x''/t'')\right] \ge tf\left(t'\lambda t^{-1}(x'/t') + (1-\lambda)t''t^{-1}(x''/t'')\right) = tf(x/t)$.

(2.2). If the right derivative $\psi'_+(0)$ is zero, then $\psi(z) \geq \psi(0) = 1$ for all $z \in \mathbb{R}$, and the above construction produces trivial bounds. Therefore we may assume that $a := \psi'_+(0) > 0$. Since $\psi(0) = 1$ and $\psi(\cdot)$ is convex and nonnegative, we conclude that $\psi(z) \geq \max\{1 + az, 0\}$ for all z, so that the upper bounds (2.2) can be only improved when replacing $\psi(z)$ with the function $\hat{\psi}(z) := \max\{1 + az, 0\}$, which also is a generating function. But the bounds produced by the latter function are, up to scaling $z \leftarrow z/a$, the same as those produced by the function

$$\psi^*(z) := [1 + z]_+, \tag{2.7}$$

where $[a]_+ := \max\{a, 0\}$. That is, from the point of view of the most accurate approximation, the best choice of the generating function ψ is the piecewise linear function ψ^* defined in (2.7).

For the generating function ψ^* defined in (2.7) the approximate constraint (2.5) takes the form

$$\inf_{t>0} \left[\mathbb{E}\big[\big[F(x,\xi) + t\big]_+\big] - t\alpha\right] \leq 0. \tag{2.8}$$

Replacing in the left-hand side $\inf_{t>0}$ with $\inf_t$, we clearly do not affect the validity of the relation; thus, we can rewrite (2.8) equivalently as

$$\inf_{t\in\mathbb{R}} \big[-t\alpha + \mathbb{E}\big[[F(x,\xi) + t]_+\big]\big] \leq 0. \tag{2.9}$$

In that form the constraint is related to the concept of conditional value at risk (CVaR) going back to [13, 21]. Recall that CVaR of a random variable Z is

$$\mathrm{CVaR}_{1-\alpha}(Z) := \inf_{\tau\in\mathbb{R}} \left[\tau + \frac{1}{\alpha}\mathbb{E}[Z - \tau]_+\right]. \tag{2.10}$$

It is easily seen that $\mathrm{CVaR}_{1-\alpha}(Z)$ is a convex and monotone functional on the space of random variables with finite first moment, and that the $(1-\alpha)$-quantile ("value at risk")

$$\mathrm{VaR}_{1-\alpha}(Z) := \inf\left[t : \mathrm{Prob}(Z \leq t) \geq 1 - \alpha\right]$$

of the distribution of Z is a minimizer of the right-hand side in (2.10), so that it always holds that $\mathrm{CVaR}_{1-\alpha}(Z) \geq \mathrm{VaR}_{1-\alpha}(Z)$. Since the chance constraint in (1.1) is nothing but $\mathrm{VaR}_{1-\alpha}[F(x,\xi)] \leq 0$, the constraint

$$\mathrm{CVaR}_{1-\alpha}[F(x,\xi)] \leq 0 \tag{2.11}$$

defines a convex conservative approximation of the chance constraint. The idea of using CVaR as a convex approximation of VaR is due to Rockafellar and Uryasev [24]. Recalling the definition of CVaR, we see that the constraints (2.9) and (2.11) are equivalent to each other.

One of the possible drawbacks of using the "optimal" generating function ψ^* (as compared with the exponential $\psi(z) := e^z$, which we will discuss in the next section) in the above approximation scheme is that it is unclear how to compute efficiently the corresponding function $\Psi(x,t)$ even in the simple case $F(x,\xi) := g_0(x) + \sum_{j=1}^d \xi_j g_j(x)$ of affine in ξ function $F(x,\xi)$ and independent-of-each-other random variables ξ_j with known and simple distributions.

There are several ways how the above construction can be extended for $m > 1$. One simple way is to replace the constraints $f_i(x,\xi) \le 0$, $i = 1,\dots,m$, with one constraint $f(x,\xi) \le 0$, say by taking $f(x,\xi) := \max\{f_1(x,\xi),\dots,f_m(x,\xi)\}$. Note, however, that this may destroy a simple, e.g., affine in ξ, structure of the constraint mapping $F(x,\xi)$. An alternative approach is the following.

Consider a closed convex cone $K \subseteq \mathbb{R}^m_+$ and the corresponding partial order $\succeq_K$, i.e., $z \succeq_K y$ iff $z - y \in K$. Of course, for the nonnegative orthant cone $K := \mathbb{R}^m_+$ the constraint $F(x,\xi) \le 0$ means that $F(x,\xi) \preceq_K 0$. We can also consider some other convex closed cones and define constraints in that form. The corresponding chance constraint can be written in the form

$$p(x) := \mathrm{Prob}\big\{F(x,\xi) \notin -K\big\} < \alpha. \tag{2.12}$$

Let $\psi : \mathbb{R}^m \to \mathbb{R}$ be a nonnegative valued, convex function such that the following hold:

$(\star)$ ψ is K-monotone; i.e., if $z \succeq_K y$, then $\psi(z) \ge \psi(y)$.

$\binom{\star}{\star}$ $\psi(z) > \psi(0) = 1$ for every $z \in \mathbb{R}^m \setminus (-K)$.

We refer to function $\psi(z)$ satisfying these properties as a K-generating function.

By $\binom{\star}{\star}$ we have that $\mathbb{E}[\psi(F(x,\xi))]$ provides an upper bound for $p(x)$, and the corresponding inequality of the form (2.2) holds. Suppose, further, that for every $\xi \in \Xi$ the mapping $F(\cdot,\xi)$ is K-convex; i.e., for any $t \in [0,1]$ and $x, y \in \mathbb{R}^n$,

$$tF(x,\xi) + (1-t)F(y,\xi) \succeq_K F(tx + (1-t)y, \xi).$$

(Note that for $K = \mathbb{R}^m_+$, K-convexity means that $F(\cdot,\xi)$ is componentwise convex.) Then for $\Psi(x,t) := t\,\mathbb{E}[\psi(t^{-1}F(x,\xi))]$, the problem of the form (2.6) gives a convex conservative approximation of the chance constrained problem (1.1).

In such construction for $m > 1$, there is no "best" choice of the K-generating function $\psi(z)$. A natural choice in the case of $K = \mathbb{R}^m_+$ could be

$$\hat{\psi}(z) := \max_{1\le i\le m} [1 + a_i z_i]_+, \tag{2.13}$$

where $a_i > 0$ are "scale parameters."

Yet there is another possible extension of the above approximation scheme for $m > 1$. Let $\alpha_1,\dots,\alpha_m$ be positive numbers such that $\alpha_1 + \cdots + \alpha_m \le \alpha$. The chance constraint of (1.1) is equivalent to $\mathrm{Prob}\,\{\bigcup_{i=1}^m \{\xi : f_i(x,\xi) > 0\}\} < \alpha$. Since

$$\mathrm{Prob}\left\{\bigcup_{i=1}^m \{f_i(x,\xi) > 0\}\right\} \le \sum_{i=1}^m \mathrm{Prob}\big\{f_i(x,\xi) > 0\},$$

it follows that the system of constraints

$$\mathrm{Prob}\big\{f_i(x,\xi) > 0\} \le \alpha_i, \quad i = 1,\dots,m, \tag{2.14}$$

is more conservative then the original chance constraint. We can now apply the one-dimensional construction to each individual constraint of (2.14) to obtain the following convex conservative approximation of the chance constrained problem (1.1):

$$\min_{x\in X} f(x) \quad \text{subject to} \quad \inf_{t>0} [\Psi_i(x,t) - t\alpha_i] \le 0,\ i = 1,\dots,m, \tag{2.15}$$

where $\Psi_i(x,t) := t\mathbb{E}\left[\psi_i(t^{-1}f_i(x,\xi))\right]$, and each $\psi_i(\cdot)$, $i = 1,\dots,m$, is a one-dimensional generating function.

Remark 2.1. An open question related to the approximation (2.15) is how to choose α_i. It would be very attractive to treat these quantities in (2.15) as design variables (subject to the constraints $\alpha_i > 0$ and $\sum_i \alpha_i \leq \alpha$) rather than as parameters. Unfortunately, such an attempt destroys the convexity of (2.15) and thus makes the approximation seemingly intractable. The simplest way to resolve the issue in question is to set

$$\alpha_i := \alpha/m,\ i = 1, \ldots, m. \tag{2.16}$$

3. Bernstein approximation. One of the drawbacks of using the piecewise linear generating functions of the form (2.7) (or (2.13)) is that the corresponding constraint function may be difficult to compute even for relatively simple functions $F(x, \xi)$. In this section we consider the (one-dimensional) generating function $\psi(z) := e^z$. For such a choice of the generating function, constructions of the previous section are closely related to the classical large deviations theory (cf., [9]).

We assume in this section that the following hold:

A1. The components ξ_j, $j = 1, \ldots, d$, of the random vector ξ are independent of other random variables.

We denote by P_j the probability distribution of ξ_j, supported on $\Xi_j \subset \mathbb{R}$ (so that the support of the distribution P of ξ is $\Xi = \Xi_1 \times \cdots \times \Xi_d$), by

$$M_j(t) := \mathbb{E}\left[e^{t\xi_j}\right] = \int \exp(tz) dP_j(z),$$

the moment generating function, and by $\Lambda_j(t) := \log M_j(t)$, the logarithmic moment generating function of ξ_j.

A2. The moment generating functions $M_j(t)$, $j = 1, \ldots, d$, are finite valued for all $t \in \mathbb{R}$ and are efficiently computable.

In fact, we could allow for the moment generating functions to be finite valued just in a neighborhood of $t = 0$. We make the stronger assumption of requiring the moment generating functions to be finite valued for all t in order to simplify the presentation.

A3. The components $f_i(x, \xi)$ in the constraint mapping $F(x, \xi)$ are affine in ξ:

$$f_i(x, \xi) = f_{i0}(x) + \sum_{j=1}^{d} \xi_j f_{ij}(x), \quad i = 1, \ldots, m, \tag{3.1}$$

and the functions $f_{ij}(x)$, $j = 0, 1, \ldots, d$, are well defined and convex on X. Besides this, for every $j \geq 1$ such that $\Xi_j \not\subset \mathbb{R}_+$, all functions $f_{ij}(x)$, $i = 1, \ldots, m$, are affine. In addition, the objective $f(x)$ in (1.1) is well defined and convex on X.

In what follows, we refer to problem (1.1) satisfying the assumptions A1–A3 as *an affinely perturbed convex* chance constrained problem.

Let $z = (z_0, z_1, \ldots, z_d) \in \mathbb{R}^{d+1}$. By A1 and A2, the function

$$\Phi(z) := \log\left(\mathbb{E}\left[\exp\left\{z_0 + \sum_{j=1}^{d} \xi_j z_j\right\}\right]\right) = z_0 + \sum_{j=1}^{d} \Lambda_j(z_j)$$

is well defined and continuous in z. Besides this, it is convex (since, as is well known, the logarithmic moment generating functions are so). Moreover, $\Phi(z)$ is monotone in

z_0 and in every z_j with $j \in J := \{j \geq 1 : \Xi_j \subset \mathbb{R}_+\}$. Finally, one clearly has for $t > 0$ and $p(z) := \text{Prob}\{z_0 + \sum_{j=1}^d \xi_j z_j > 0\}$ that

$$\Phi(t^{-1}z) \geq \log p(z).$$

Consequently, for every $\beta \in (0,1)$,

$$\exists t > 0 : t\Phi(t^{-1}z) - t\log\beta \leq 0 \quad \text{implies} \quad p(z) \leq \beta.$$

Similar to the reasoning which led us to (2.4), the latter implication can be strengthened to

$$\inf_{t>0}\left[t\Phi(t^{-1}z) - t\log\beta\right] \leq 0 \quad \text{implies} \quad p(z) \leq \beta. \tag{3.2}$$

Now consider an affine chance constrained problem with real-valued constraint mapping

$$F(x,\xi) = g_0(x) + \sum_{j=1}^d \xi_j g_j(x).$$

By (3.2), the problem

$$\min_{x\in X} f(x) \quad \text{subject to} \quad \inf_{t>0}\left[g_0(x) + \sum_{j=1}^d t\Lambda_j(t^{-1}g_j(x)) - t\log\alpha\right] \leq 0 \tag{3.3}$$

is a conservative approximation of the chance constrained problem (1.1). In fact this approximation is convex. Indeed, the function

$$G(z,t) := t\Phi(t^{-1}z) - t\log\beta$$

is convex in $(z, t > 0)$ (since $\Phi(z)$ is convex) and is monotone in z_0 and every z_j with $j \in J$, while, by A3, all $g_j(x)$, $j = 0, 1, \ldots, d$, are convex in $x \in X$, and all $g_j(x)$ with $j \geq 1$ such that $j \notin J$ are affine. It follows that the function $G(g_0(x), \ldots, g_d(x), t)$ is convex in $(x \in X, t > 0)$, whence the constraint in (3.3) is convex; the objective is convex by A3, and X was once forever assumed to be convex when formulating (1.1). Thus, (3.3) is a *convex conservative* approximation of an affinely perturbed chance constrained problem with $m = 1$, as claimed.

We can extend the outlined construction to the case of $m > 1$ in a way similar to the construction of problem (2.15). That is, given an affinely perturbed chance constrained problem (1.1), (3.1), we choose $\alpha_i > 0$, $\sum_i \alpha_i \leq \alpha$, and build the optimization problem

$$\begin{array}{ll} \min\limits_{x\in X} & f(x) \\ \text{subject to} & \inf\limits_{t>0}\left[f_{i0}(x) + \sum\limits_{j=1}^d t\Lambda_j(t^{-1}f_{ij}(x)) - t\log\alpha_i\right] \leq 0,\ i = 1,\ldots,m. \end{array} \tag{3.4}$$

Similar to the case of $m = 1$, this problem is a convex conservative approximation of (1.1). We refer to (3.4) as the *Bernstein* approximation of (1.1).

An advantage of the Bernstein approximation over the one discussed in the previous section is that under assumptions A1–A3, Bernstein approximation is an explicit

convex program with efficiently computable constraints and as such is efficiently solvable.

Remark 3.1. A somehow less accurate version of Bernstein approximation was in fact proposed in [2] for the situation where the random variables ξ_j are independent with zero mean and supported on segments $[-\sigma_i, \sigma_i]$. We have cited this result in the introduction; see (1.9). The justification of (1.9) is based on a straightforward bounding from above (going back to Bernstein) of the associated logarithmic moment generating function and demonstrating that if x satisfies (1.7), then the resulting (conservative) version of the corresponding probability bound, as applied to $z = (f_{i0}(x), f_{i1}(x), \ldots, f_{id}(x))$, implies that

$$\text{Prob}\left\{ f_{i0}(x) + \sum_{j=1}^{d} \xi_j f_{ij}(x) > 0 \right\} \le \exp\{-\kappa\Omega^2\}.$$

Clearly, Bernstein approximation as presented here is less conservative than (1.9), since it is based on the corresponding "true" function rather than on its upper bound given solely by the expected values and the sizes of supports of ξ_j.

4. Upper and lower bounds. In general, the approximation-based approach to processing chance constrained problems requires mechanisms for (i) measuring the actual risk (reliability) associated with the resulting solution, and (ii) bounding from below the true optimal value Opt^* of the chance constraint problem (1.1). Task (i) corresponds to the case when the approximation is not necessarily conservative, as it is the case, e.g., with the scenario approximation. With the latter, even applied with the theoretically justified sample size (1.4), there is still a chance $1 - \delta$ that the resulting solution $\bar{x}$ does not satisfy the chance constraint, and we would like to check whether the solution indeed is feasible for (1.1). Task (ii) is relevant to basically all approximations, since they usually are conservative ("for sure," as Bernstein approximation, or "with probability close to 1," as the scenario approximation with sample size (1.4)), and a lower bound on Opt^* allows one to quantify this conservatism.

A straightforward way to measure the actual risk of a given candidate solution $\bar{x} \in X$ is to use Monte Carlo sampling. That is, a sample $\xi^1, \ldots, \xi^{N'}$ of N' realizations of random vector ξ is generated and the probability $p(\bar{x}) := \text{Prob}\big\{F(\bar{x}, \xi) \not\le 0\big\}$ is estimated as Δ/N', where Δ is the number of times the constraint $F(\bar{x}, \xi^\nu) \le 0$, $\nu = 1, \ldots, N'$, is violated. A more reliable upper bound on $p(\bar{x})$ is the random quantity

$$\widehat{\alpha} := \max_{\gamma \in [0,1]} \left\{ \gamma : \sum_{r=0}^{\Delta} \binom{N'}{r} \gamma^r (1-\gamma)^{N'-r} \ge \delta \right\},$$

where $1 - \delta$ is the required confidence level. The quantity $\widehat{\alpha}$ is, with probability of at least $1 - \delta$, an upper bound on $p(\bar{x})$, so that if our experiment results in $\widehat{\alpha} \le \alpha$, we may be sure, "up to probability of bad sampling $\le \delta$," that $\bar{x}$ is feasible for (1.1) and $f(\bar{x})$ is an upper bound on Opt^*. Since the outlined procedure involves only the calculation of quantities $F(\bar{x}, \xi^\nu)$, it can be performed with a large sample size N', and hence feasibility of $\bar{x}$ can be evaluated with a high reliability, provided that α is not too small (otherwise the procedure would require an unrealistically large sample size).

It is more tricky to bound Opt^* from below. Here we propose a bounding scheme as follows. Let us choose three positive integers M, N, L, with $L \le M$, and let

us generate M independent samples $\xi^{1,\mu},\dots,\xi^{N,\mu}$, $\mu=1,\dots,M$, each of size N, of random vector ξ. For each sample we solve the associated optimization problem

$$\min_{x\in X} f(x) \quad \text{subject to} \quad F(x,\xi^{\nu,\mu})\le 0,\ \nu=1,\dots,N, \tag{4.1}$$

and hence calculate its optimal value Opt_μ.

We compute the quantities Opt_μ, $\mu=1,\dots,M$, by treating the infeasibility and unboundedness according to the standard optimization conventions: the optimal value of an infeasible optimization problem is $+\infty$, while for a feasible and unbounded problem from below it is $-\infty$. We then rearrange the resulting quantities $\{\mathrm{Opt}_\mu\}_{\mu=1,\dots,M}$ in nondescending order: $\mathrm{Opt}_{(1)}\le\cdots\le\mathrm{Opt}_{(M)}$ (in the statistics literature these are called the order statistics of the sample $\{\mathrm{Opt}_\mu\}_{\mu=1,\dots,M}$). By definition, the lower bound on the true optimal value is the random quantity $\mathrm{Opt}_{(L)}$.

Let us analyze the resulting bounding procedure. Let x be a feasible point of the true problem (1.1). Then x is feasible for problem (4.1) with probability of at least $\theta_N=(1-\alpha)^N$. When x is feasible for (4.1), we of course have $\mathrm{Opt}_\mu\le f(x)$. Thus, for every $\mu\in\{1,\dots,M\}$ and for every $\varepsilon>0$ we have

$$\theta := \mathrm{Prob}\{\mathrm{Opt}_\mu\le \mathrm{Opt}^*+\varepsilon\}\ge\theta_N.$$

Now, in the case of $\mathrm{Opt}_{(L)}>\mathrm{Opt}^*+\varepsilon$, the corresponding realization of the random sequence $\mathrm{Opt}_1,\dots,\mathrm{Opt}_M$ contains less than L elements which are less than or equal to $\mathrm{Opt}^*+\varepsilon$. Since the elements of the sequence are independent, the probability $\rho(\theta,M,L)$ of the latter event is

$$\rho(\theta,M,L)=\sum_{r=0}^{L-1}\binom{M}{r}\theta^r(1-\theta)^{M-r}.$$

Since $\theta\ge\theta_N$, we have that $\rho(\theta,M,L)\le\rho(\theta_N,M,L)$.

Thus,

$$\mathrm{Prob}\left\{\mathrm{Opt}_{(L)}>\mathrm{Opt}^*+\varepsilon\right\}\le\rho(\theta_N,M,L).$$

Since the resulting inequality is valid for all $\varepsilon>0$, we arrive at the bound

$$\mathrm{Prob}\left\{\mathrm{Opt}_{(L)}>\mathrm{Opt}^*\right\}\le\sum_{r=0}^{L-1}\binom{M}{r}(1-\alpha)^{Nr}\left[1-(1-\alpha)^N\right]^{M-r}. \tag{4.2}$$

We now arrive at the following simple result.

PROPOSITION 4.1. *Given $\delta\in(0,1)$, let us choose positive integers M,N,L in such a way that*

$$\sum_{r=0}^{L-1}\binom{M}{r}(1-\alpha)^{Nr}\left[1-(1-\alpha)^N\right]^{M-r}\le\delta. \tag{4.3}$$

Then with probability of at least $1-\delta$, the random quantity $\mathrm{Opt}_{(L)}$ gives a lower bound for the true optimal value Opt^*.

The question arising in connection with the outlined bounding scheme is how to choose M, N, L. Given a desired reliability $1-\delta$ of the bound *and* M and N, it is easy to specify L: this should be just the largest $L>0$ satisfying condition (4.3). (If no

$L > 0$ satisfying (4.3) exists, the lower bound, by definition, is $-\infty$.) We end up with a question of how to choose M and N. For N given, the larger M is, the better. For given N, the "ideal" bound yielded by our scheme as M tends to infinity is the lower θ_N-quantile of the true distribution of the random variable Opt_1. The larger M, the better we can estimate this quantile from a sample of M independent realizations of this random variable. In reality, however, M is bounded by the computational effort required to solve M problems (4.1). Note that the larger the effort per problem, the larger the sample size N. We have no definite idea how to choose N. As N grows, the distribution of Opt_1 "goes up" in the sense that $\mathrm{Prob}\{\mathrm{Opt}_1 > a\}$ increases for every a. As a result, every lower θ-quantile of this distribution also increases. If our bound were the lower θ-quantile of the distribution of Opt_1, it would grow (that is, improve) with N. Unfortunately, our bound is the (empirical estimate of) the lower θ_N-quantile of the distribution in question, with θ_N decreasing as N grows, and this decrease shifts the bound down. For the time being, we do not know how to balance these two opposite trends, except for a trivial way to test several values of N and to choose the best (the largest) of the resulting bounds. To keep reliability δ by testing k different values of N, would require one to strengthen reliability of every one of the tests, e.g., in accordance with the Bonferroni inequality, by replacing δ in the right-hand side of (4.3) with δ/k.

5. Numerical illustration. We are about to present the results of an illustrative experiment. While the model below is described in financial terms, we do not pretend this toy model is of actual applied value; our only goal here is to compare Bernstein approximations with the scenario approach (see the introduction).

Test problem: optimizing value at risk. The toy test problem we are about to consider is the following. There are $n+1$ assets $0, 1, \ldots, n$ with random returns. The problem is to distribute \$1 between the assets in order to maximize the upper $(1-\alpha)$th quantile of the total profit (that is, the total return of the resulting portfolio minus the initial capital of \$1). The corresponding model is the chance constrained linear programming problem

$$(P_\alpha) \qquad \max_{x\geq 0,\, t\in\mathbb{R}} t - 1 \quad \text{subject to} \quad \mathrm{Prob}\left\{t > \sum_{j=0}^{n} r_j x_j\right\} \leq \alpha,\ \sum_{j=0}^{n} x_j \leq 1,$$

where x_j is the capital invested in asset j, and r_j is the return of this asset.

The data we used in our experiment are as follows:

- There are $n+1 = 65$ assets; asset #0 ("money") has deterministic return $r_0 \equiv 1$, while the returns r_i of the remaining 64 "true" assets are random variables with expectations $\mathbb{E}[r_i] = 1+\rho_i$, with the nominal profits ρ_i varying in $[0, 0.1]$ and growing with i.
- The random variables r_i, $1 \leq i \leq 64$, are of the form

$$r_i = \eta_i + \sum_{\ell=1}^{8} \gamma_{i\ell}\zeta_\ell, \tag{5.1}$$

where $\eta_i \sim \mathcal{LN}(\mu_i, \sigma_i^2)$ (that is, $\log \eta_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$) is the individual noise in ith return, $\zeta_\ell \sim \mathcal{LN}(\nu_\ell, \theta_\ell^2)$ are "common factors" affecting all returns, and $\gamma_{i\ell} \geq 0$ are deterministic "influence coefficients." All "primitive" random variables (64 of η_i's and 8 of ζ_ℓ's) are independent of each other.

We used $\nu_\ell = 0$, $\theta_\ell = 0.1$, $\mu_i = \sigma_i$ (that is, the more promising an asset at average, the more risky it is). The influence coefficients $\gamma_{i\ell}$ and the parameters μ_i were chosen in such a way that $\mathbb{E}\big[\sum_{\ell=1}^{8} \gamma_{i\ell}\zeta_\ell\big] = \rho_i/2$ and $\mathbb{E}[\eta_i] = 1 + \rho_i/2$ for all i.

Processing log-normal distributions. The random returns r_i are linear combinations of independent random variables $\eta_1, \ldots, \eta_{64}$, $\zeta_1, \ldots, \zeta_8$, so that the structure of (P_α) allows for applying Bernstein approximation. The difficulty, however, is that the random variables in question are log-normal and thus the corresponding moment-generating functions are $+\infty$ outside of the origin. This difficulty can be easily circumvented, specifically, as follows. Given a log-normal random variable $\xi \sim \mathcal{LN}(\mu, \sigma^2)$, and positive "threshold probability" $\epsilon > 0$ and "resolution" $\Delta > 0$, we associate with these data a discrete random variable $\widehat{\xi}$ as follows. Let $\pi(s)$ be the $\mathcal{N}(0,1)$-Gaussian density and R be such that $\int_R^\infty \pi(s)ds = \epsilon/2$; we split the segment $[-R, R]$ into bins $[a_k, a_{k+1}]$, $1 \le k < n$, of length $\sigma^{-1}\Delta$ (the last bin can be shorter) and assign the points $b_0 = 0$, $b_k = \exp\{\sigma a_k + \mu\}$, $k = 1, \ldots, n$, probability masses $\nu_k = \int_{a_k}^{a_{k+1}} \pi(s)ds$, where $a_0 = -\infty$ and $a_{n+1} = \infty$. The variable $\widehat{\xi}$ takes the values b_k, $k = 0, \ldots, n$, with probabilities ν_k. Note that this random variable can be thought of as a "rounding" of $\xi \sim \mathcal{LN}(\mu, \sigma^2)$: given a realization a of ξ, we look to which one of the $n+1$ sets $[0, b_1)$, $[b_1, b_2)$,..., $[b_{n-1}, b_n)$, $[b_n, \infty)$ a belongs, and replace a with the left endpoint of this set, thus obtaining a realization $\widehat{a}$ of $\widehat{\xi}$. Note that with our choice of a_i, we always have $\widehat{a}/a \le 1$, and $\widehat{a}/a \ge \exp\{-\Delta\}$ unless $a < b_1$ or $a > b_n$; the latter can happen with probability of at most ϵ. Thus, $\widehat{\xi}$ can be thought of as a lower bound on ξ which with probability of $\ge 1 - \epsilon$ is tight within factor $\exp\{\Delta\}$. Now let us replace in (P_α) underlying log-normal random variables $\eta_1, \ldots, \zeta_8$ with their roundings $\widehat{\eta}_1, \ldots, \widehat{\zeta}_8$. Since we "round down" and all $\gamma_{i\ell}$ are nonnegative, every feasible solution to the resulting chance constrained problem will be feasible for (P_α) as well. At the same time, the new problem is an affinely perturbed chance constrained problem with *discrete* random variables, and building its Bernstein approximation causes no problems at all. This is the scheme we used in our experiments, the parameters being $\epsilon = 10^{-6}$ and $\Delta = 0.0025$. Even with that high (in fact, redundant) quality of discretization, there was no difficulty with handling the resulting discrete random variables—the average, over all 71 discrete random variables in question, number of different values taken by a variable was just ≈ 138, which made computing Bernstein bound a pretty easy task.

Tuning the approximations. Both approximations we are dealing with in our experiments—the scenario and Bernstein one—are conservative in the sense that a solution yielded by an approximation violates the randomly perturbed constraint in question with probability $\alpha_{\rm f}$, which is less than the required risk α (this claim is completely true for Bernstein approximation and is "true with high probability" for the scenario one). Experiments show that the ratio $\alpha/\alpha_{\rm f}$ could be pretty large (see Table 1), which makes it natural to look for ways to reduce the resulting conservatism. To some extent, this can indeed be done via a simple tuning, provided that α is not too small, so that the probabilities of order of α can be measured reliably by Monte Carlo simulations with samples of reasonable size. When tuning Bernstein approximation, we replace the required risk α by a larger quantity α_+, solve the approximation as if the required risk were α_+, and then run Monte Carlo simulation in order to check with a desired reliability whether the actual risk $\alpha_{\rm f}$ of the resulting solution is $\le \alpha$. We then choose the (nearly) largest possible α_+ which meets the outlined requirement and treat the associated solution as the result of our tuning. Of course, tuning can

TABLE 1
Results of experiments with the value-at-risk model.

Quantity	Value	Empirical risk[a]	Inferred risk[a]
Nominal optimal value[b]	0.0950	—	—
Upper bound[c]	0.0799	—	—
Bernstein optimal value (tuned)[d_b]	0.0689	0.043	0.050
Bernstein optimal value[d_a]	0.0586	0.002	0.004
Scenario optimal value (tuned)[e_b]	0.0674	0.040	0.047
Scenario optimal value[e_a] ($N = 14,684$)	0.0557	0.001	0.003
Robust optimal value[f]	0.0000	—	—

be used in the case of scenario approximation as well, with the number of scenarios in the role of tuning parameter.

The experiments. The experiments were conducted for the value of risk $\alpha = 0.05$. The reliability $1 - \delta$ for the scenario approximation (see (1.4)) was set to 0.999. Similarly, the reliability of all other simulation-based inferences (like those on actual risks of various solutions, bound on the true optimal value in the chance constrained problem, etc.) was set to 0.999. The results are presented in Table 1; the reader should be aware that we work with a maximization problem, so that the larger the value of the objective yielded by a method, the better. Therefore, what was before a lower bound on the optimal value in the chance constrained problem becomes an upper bound, etc.

Explanations to Table 1. [a]Empirical risk makes sense only with respect to the optimal values yielded by various methods and is the empirical frequency estimate, taken over 10,000 simulations, of the probability p of violating the randomly perturbed constraint in $(P_{0.05})$ at the solution yielded by the method. Inferred risk is the 0.999-reliable upper bound on p, as inferred from the same 10,000 simulations.

[b]Optimal value in the nominal problem—the one where all randomly perturbed coefficients are set to their expected values.

[c]See section 4. Since $(P_{0.05})$ is a maximization problem, the corresponding construction yields an upper bound on the optimal value in $(P_{0.05})$. The reliability of the bound is 0.999.

[d_a]Optimal value in Bernstein approximation (3.4) of $(P_{0.05})$.

[d_b]Optimal value in tuned Bernstein approximation. In our experiment, the best tuning corresponded to replacing the true value 0.05 of risk with the value 0.3.

[e_a]Optimal value in the scenario approximation (P^N) of $(P_{0.05})$, the sample size N being chosen according to (1.4) (where $n = 66$, $\alpha = 0.05$, and $\delta = 0.001$).

[e_b]Optimal value in tuned scenario approximation. In our experiment, the best tuning corresponded to reducing the number of scenarios with its theoretical value 14,684 to 550.

[f]Optimal value given by robust optimization; under mild regularity assumptions, which hold true in the case of (P), this is the same as the optimal value in (P_α) in the case of $\alpha = 0$. In our case, the robust optimal value is 0, meaning that there is no way to make guaranteed profit, so that the best, in the worst-case setting, policy is to not to invest into "nonmoney" assets at all.

Discussion. A. As far as the objective value is concerned, Bernstein approximation outperforms the (nontuned) scenario approximation; the same is true for the tuned versions of the procedures (this is consistent with all other numerical exper-

iments we have run, including those for test problems of different structure). The differences, although not large, are not negligible (2.2% for tuned approximations).

B. Additional good news about Bernstein approximation is that even with tuning, this still is an implementable routine: the solution and the optimal value in (3.4), (2.16) are well-defined functions of α, and the resulting value of the objective is better, the larger α is. Consequently, tuning becomes an easy-to-implement routine, a kind of bisection: we solve (3.4), (2.16) for a certain value of α and check the actual risk of the resulting solution; if it is worse then necessary, we decrease α in (3.4), otherwise increase it. In contrast to this, the optimal value and the optimal solution of scenario approximation with a given sample size are random. For not too large sample sizes, the variability of these random entities is high, which makes tuning difficult.

C. It should be added that Bernstein approximation in its nontuned form remains practical in the case of very small risks α and/or high design dimension, that is, in situations where the scenario approximation requires samples of unrealistic sizes. To get an impression of the numbers, assume that we want α as small as 0.5% or even 0.1%, while the reliability $1-\delta$ of our conclusions (which in previous experiments was set to 0.999) is now increased to 0.9999. In this case the scenario approximation becomes completely impractical. Indeed, the theoretically valid sample size given by (1.4) becomes 209,571 for $\alpha = 0.5\%$ and 1,259,771 for $\alpha = 0.1\%$, which is a bit too much. Using smaller sample sizes plus tuning also is problematic, since it becomes too complicated to test the risk of candidate solutions by simulation. For example, with $\alpha = 0.005$ and $\alpha = 0.001$, it takes over 100,000 simulations to conclude, with reliability 0.9999, that a given candidate solution which in fact is feasible for $(P_{0.9\alpha})$ is feasible for (P_α).

• At the same time, Bernstein approximation with no tuning is 100% reliable, remains of the same complexity independently of how small is α, and at the uncertainty level 0.5 results in the profits 0.0500 for $\alpha = 0.5\%$ and 0.0445 for $\alpha = 0.1\%$. This is not that bad, given that the robust optimal value in our situation is 0.

The bottom line, as suggested by the experiments (and as such, not conclusive yet) is as follows: The scenario approximation has no advantages whatsoever as compared to the Bernstein one, *provided the latter is applicable* (that is, that we are in the case of a affinely perturbed convex chance constrained problem with known and simple enough distributions of ξ_j).

6. The case of ambiguous chance constraints. As was mentioned in the introduction, one of the basic problems with the formulation of chance constrained problem (1.1) is that it assumes an exact knowledge of the underlying probability distribution P of ξ. Therefore it appears natural to consider "robust" or minimax versions of the chance constrained problems; for results in this direction, see [12, 27, 25, 26, 14] and references therein. When applying the minimax approach to chance constrained problems, one assumes that the distribution P of random vector ξ in (1.1) belongs to a given in advance family $\mathfrak{P}$ of probability distributions supported on a (closed) set $\Xi \subset \mathbb{R}^d$ and replaces the chance constraint in (1.1) with its worst-case, over $P \in \mathfrak{P}$, version, thus arriving at the *ambiguous chance constrained* problem

$$\min_{x\in X} f(x) \quad \text{subject to} \quad \mathrm{Prob}_P\big\{F(x,\xi)\le 0\big\} \ge 1-\alpha \ \ \forall P \in \mathfrak{P}, \tag{6.1}$$

where Prob_P is the P-probability of the corresponding event.

Of course, we can replace the probability constraints in (6.1) with one constraint by taking the minimum of $\mathrm{Prob}_P\big\{F(x,\xi)\le 0\big\}$ with respect to $P \in \mathfrak{P}$. That is,

problem (6.1) is constrained with respect to a "worst" distribution of the considered family $\mathfrak{P}$. We can also write the probability constraints of (6.1) in the following form:

$$\sup_{P\in\mathfrak{P}} \mathbb{E}_P\left[\mathbb{1}_{A_x}\right] \leq \alpha, \tag{6.2}$$

where $A_x := \{\xi \in \Xi : F(x,\xi) \not\leq 0\}$. The "worst-case-distribution" (or minimax) stochastic programming problems were considered in a number of publications (e.g., [12, 27]). When applied to chance constraints, such worst-case-distribution problems are called *ambiguous* chance constrained problems (see [14] and references therein).

For some families of distributions the maximum in the left-hand side of (6.2) can be calculated explicitly. With every family $\mathfrak{P}$ of probability distributions is associated the function

$$\rho(Z) := \sup_{P\in\mathfrak{P}} \mathbb{E}_P\left[Z\right] \tag{6.3}$$

defined on a space of real-valued random variables Z. Formula (6.3) describes a dual representation of so-called *coherent risk measures* introduced by Artzner et al [1]. Consider now the following family:

$$\mathfrak{P} := \big\{P : \gamma_1 P^* \preceq P \preceq \gamma_2 P^*,\ P(\Xi) = 1\big\}. \tag{6.4}$$

Here γ_1 and γ_2 are constants such that $0 \leq \gamma_1 \leq 1 \leq \gamma_2$, P^* is a (reference) probability distribution on Ξ and the notation $P_1 \preceq P_2$ means that for two (not necessarily probability) Borel measures P_1 and P_2 on Ξ it holds that $P_1(A) \leq P_2(A)$ for any Borel set $A \subset \Xi$. The constraint $P(\Xi) = 1$ in (6.3) is written to ensure that P is a probability measure. This family $\mathfrak{P}$ defines a coherent risk measure, which can be written in the following equivalent form:

$$\rho(Z) = \mathbb{E}[Z] + \inf_{\tau\in\mathbb{R}} \mathbb{E}\left[(1-\gamma_1)[\tau - Z]_+ + (\gamma_2 - 1)[Z-\tau]_+\right], \tag{6.5}$$

where all expectations are taken with respect to the reference distribution P^*. In particular, for $\gamma_1 = 0$ and $\kappa := (\gamma_2 - 1)/\gamma_2$,

$$\rho(Z) = \mathrm{CVaR}_\kappa[Z]$$

(cf., [25, 26]).

By the definition (6.4) of $\mathfrak{P}$ we have that $\mathbb{E}_P\left[\mathbb{1}_{A_x}\right] \leq \gamma_2 P^*(A_x)$ for any $P \in \mathfrak{P}$, with the equality holding if $P(A_x) = \gamma_2 P^*(A_x)$. Since $P(\Xi) = 1$, this can be achieved iff $\gamma_2 P^*(A_x) + \gamma_1(1 - P^*(A_x)) \leq 1$, i.e., iff $P^*(A_x) \leq \frac{1-\gamma_1}{\gamma_2-\gamma_1}$. We obtain the following.

> If $\alpha \leq (1-\gamma_1)/(\gamma_2-\gamma_1)$, then the ambiguous chance constrained problem (6.1) with $\mathfrak{P}$ given by (6.4) is equivalent to the chance constrained problem (1.1) with respect to the reference distribution P^* and with rescaled risk $\alpha \leftarrow \alpha^* := \alpha/\gamma_2$.

Another popular example of a coherent risk measure is the mean-upper-absolute semideviation

$$\rho(Z) := \mathbb{E}[Z] + c\mathbb{E}\left(\big[Z - \mathbb{E}[Z]\big]_+\right), \tag{6.6}$$

where $c \in [0,1]$ is a constant and the expectations are taken with respect to a reference distribution P^*. It has the dual representation (6.3) with the corresponding family

$$\mathfrak{P} = \big\{\zeta' : \zeta' = 1 + \zeta - \mathbb{E}[\zeta],\ \|\zeta\|_\infty \leq c\big\}, \tag{6.7}$$

where $\zeta' = dP/dP^*$ denotes the density of P with respect to P^* (cf., [26]). By using the definition (6.6) it is straightforward to calculate that

$$\rho(\mathbb{1}_{A_x}) = P^*(A_x) + 2cP^*(A_x)(1 - P^*(A_x)). \tag{6.8}$$

By solving the quadratic inequality $t + 2ct(1-t) \le \alpha$ for $t = P^*(A_x)$, we obtain that $P^*(A_x) \le \varphi(\alpha)$, where

$$\varphi(\alpha) := \frac{1 + 2c - \sqrt{1 + 4c(1-2\alpha) + 4c^2}}{4c}$$

for $c \in (0,1]$, and $\varphi(\alpha) = \alpha$ if $c = 0$. (Note that for $\alpha \in (0,1)$ and $c \in (0,1]$, it always holds that $\varphi(\alpha) \in (0,\alpha)$.) We obtain the following.

> The ambiguous chance constrained problem (6.1) with $\mathfrak{P}$ given by (6.7) is equivalent to the chance constrained problem (1.1) with respect to the reference distribution P^* and with rescaled reliability parameter $\alpha \leftarrow \alpha^* := \varphi(\alpha)$.

Of course, such explicit reduction of the ambiguous chance constrained problem (6.1) to the regular chance constrained problem (1.1) is possible only for some specific families $\mathfrak{P}$. Our current goal is to develop Bernstein-type approximation of the constraint in (6.1). As before, we restrict ourselves with problems where the "bodies" of the constraints are affine in ξ:

$$\begin{array}{l} \min\limits_{x\in X} f(x) \quad \text{subject to} \\ \inf\limits_{P\in\mathfrak{P}} \mathrm{Prob}_P\left\{\xi : f_{i0}(x) + \sum\limits_{j=1}^{d} \xi_j f_{ij}(x) \le 0,\ i = 1,\dots,m\right\} \ge 1-\alpha. \end{array} \tag{6.9}$$

6.1. Assumptions and construction.

Assumptions. From now on, we make the following assumptions about the "data" of (6.9):

B1. The family $\mathfrak{P}$ of possible distributions of ξ is as follows. Let D_j, $j = 1,\dots,d$, be nonempty compact subsets of the axis, and $\mathcal{M}$ be a nonempty set of tuples $\{P_j\}_{j=1}^d$, where P_j are Borel probability measures on D_j. We assume that

- whenever $\{P_j\}_{j=1}^d$, $\{P'_j\}_{j=1}^d$ are two elements from $\mathcal{M}$, so is $\{\lambda P_j + (1-\lambda)P'_j\}_{j=1}^d$, $\lambda \in [0,1]$ (convexity), and
- whenever a sequence $\{P_j^t\}_{j=1}^d$, $t = 1,2,\dots$, of elements of $\mathcal{M}$ weakly converges to $\{P_j\}_{j=1}^d$ (meaning that $\int f(s)dP_j^t(s) \to \int f(s)dP_j(s)$ as $t \to \infty$ for every j and every continuous and bounded on the axis function f), then $\{P_j\}_{j=1}^d \in \mathcal{M}$ (weak closedness).

We assume that $\mathfrak{P}$ is comprised of all product distributions $P = P_1 \times \cdots \times P_d$ on $\mathbb{R}^d$ with the tuple of marginals $\{P_j\}_{j=1}^d$ running through a given set $\mathcal{M}$ with the outlined properties.

From now on, we equip the set $\mathcal{M}$ underlying, via the outlined construction, the set $\mathfrak{P}$ in question with the weak topology. It is well known that under the above assumptions this topology is yielded by an appropriate metric on $\mathcal{M}$, and that with this metric $\mathcal{M}$ is a compact metric space.

> The simplest example of a set $\mathfrak{P}$ of the outlined structure is as follows. Let D_j be finite subsets of $\mathbb{R}$, let $\Delta := \bigcup_{j=1}^d D_j = \{s_1,\dots,s_K\}$, and let $\mathcal{M}$ be a closed and convex set of matrices $P = [p_{kj}]_{\substack{1\le k\le K\\ 1\le j\le d}}$

with nonnegative entries such that $\sum_k p_{kj} = 1$ for all j and $p_{kj} = 0$ whenever $s_k \notin D_j$. For every $P \in \mathcal{M}$, the jth column P_j of P can be naturally identified with a probability distribution on D_j; the set $\mathfrak{P}$ generated by $\mathcal{M}$ is comprised of all product distributions $P_1 \times \cdots \times P_d$ coming from matrices $P \in \mathcal{M}$.

From now on, we denote a generic element of $\mathcal{M}$ by $Q = \{Q_j\}_{j=1}^d$.

B2. The objective $f(x)$ and all functions $f_{ij}(x)$, $i = 1, \ldots, m$, $j = 0, 1, \ldots, d$, are convex and well defined on X. Moreover, let

$$J := \{j : 1 \le j \le d, \text{ not all functions } f_{ij},\ i = 1, \ldots, m, \text{ are affine}\}.$$

We assume that whenever $j \in J$, the quantities ξ_j and η_j "are always nonnegative," that is, for every $j \in J$

• jth marginal distribution of every $P \in \mathfrak{P}$ is supported on the nonnegative ray, and

• all points $\eta \in \mathcal{U}$ satisfy $\eta_j \ge 0$

(compare with assumption A3 in section 3).

Building Bernstein approximation. For $P = P_1 \times \cdots \times P_d$, let $\widehat{P}$ be the tuple $\{P_j\}_{j=1}^d$, so that when P runs trough $\mathcal{P}$, $\widehat{P}$ runs through $\mathcal{M}$.

Let

$$\begin{aligned} \Phi(z, Q) &:= \log\left(\mathbb{E}_{Q_1 \times \cdots \times Q_d}\left[\exp\left\{z_0 + \sum_{j=1}^d \xi_j z_j\right\}\right]\right) \\ &= z_0 + \sum_{j=1}^d \log\left(\int \exp\{z_j s\} dQ_j(s)\right), \ Q = \{Q_j\}_{j=1}^d \in \mathcal{M}, \\ \widehat{\Phi}(z) &:= \max_{Q \in \mathcal{M}} \Phi(z, Q). \end{aligned} \tag{6.10}$$

By B1, $\Phi(z, Q)$ is a well-defined and continuous function of $(z, Q) \in \mathbb{R}^{d+1} \times \mathcal{M}$ (recall that $\mathcal{M}$ is equipped with w^*-topology). From (6.10) it is also evident that $\Phi(z, Q)$ is convex in $z \in \mathbb{R}^{d+1}$ and concave in $Q \in \mathcal{M}$. From these observations and the compactness of $\mathcal{M}$ it follows that $\widehat{\Phi}(z)$ is well defined everywhere and is convex. Finally, from B2 it follows that $\Phi(z, Q)$ (and therefore $\widehat{\Phi}(z)$) is nondecreasing in z_0 and in every z_j with $j \in J$.

Now let

$$\Theta_Q(z, t) := t\Phi_Q(t^{-1}z), \ \widehat{\Theta}(z, t) := t\widehat{\Phi}(t^{-1}z),$$

so that $\Theta_Q(z, t)$ and $\widehat{\Theta}(z, t)$ are well-defined convex functions in the domain $t > 0$. Same as in section 3, for every $\beta \in (0, 1)$ and every $z \in \mathbb{R}^{d+1}$ we have

$$\inf_{t>0}\left[\Theta_{\widehat{P}}(z, t) - t\log\beta\right] \le 0 \quad \text{implies} \quad \text{Prob}_P\left\{z_0 + \sum_{j=1}^d \xi_j z_j > 0\right\} \le \beta,$$

and we arrive at the following implication:

$$
\begin{aligned}
&P(\beta): && \left\{\forall Q\in\mathcal{M}: \inf_{t>0}\left[\Theta_Q(z,t)-t\log\beta\right]\le 0\right\}\\
&&&\text{implies that}\\
&Q(\beta): && \sup_{P\in\mathfrak{P}}\mathrm{Prob}_P\left\{z_0+\sum_{j=1}^d\xi_j z_j>0\right\}\le\beta.
\end{aligned}
\tag{6.11}
$$

We are about to replace (6.11) with an equivalent and more convenient computationally implication:

$$
\begin{aligned}
&\widehat{P}(\beta): && \left\{\inf_{t>0}\left[\widehat{\Theta}(z,t)-t\log\beta\right]\le 0\right\}\\
&&&\text{implies that}\\
&Q(\beta): && \sup_{P\in\mathfrak{P}}\mathrm{Prob}_P\left\{z_0+\sum_{j=1}^d\xi_j z_j>0\right\}\le\beta.
\end{aligned}
\tag{6.12}
$$

The advantage of (6.12) as compared to (6.11) is that the premise in the latter implication is semi-infinite: to verify its validity, we should check certain conditions for every $Q\in\mathcal{M}$. In contrast to this, the premise in (6.12) requires checking validity of a univariate convex inequality, which can be done by bisection, provided that the function $\widehat{\Theta}$ is efficiently computable. The latter condition is equivalent to efficient computability of the function $\widehat{\Phi}(z)$, which indeed is the case when $\mathcal{M}$ is not too complicated (e.g., is finite-dimensional and computationally tractable).

The validity of (6.12) and the equivalence of (6.11) and (6.12) are given by the following lemma.

LEMMA 6.1. *Let $0<\beta<1$. Then the following holds:*

$$
\widehat{P}(\beta)\quad\text{iff}\quad P(\beta).
\tag{6.13}
$$

Proof. Implication $\Rightarrow$ in (6.13) is evident, since $\widehat{\Theta}(z,t)=\max_{Q\in\mathcal{M}}\Theta_Q(z,t)$. Note that this implication combines with (6.11) to imply the validity of (6.12).

Now let us prove the implication $\Leftarrow$ in (6.13). This is a straightforward consequence of the fact that $\Theta_Q(z,t)$ is concave in Q and convex in $t>0$; for the sake of completeness, we present the corresponding standard reasoning.

As we remember, $\Phi(z,Q)$ is continuous and concave in $Q\in\mathcal{M}$; since $\Theta_Q(z,t)=t\Phi(t^{-1}z,Q)$, the function $\Theta_Q(z,t)$ is continuous in $(t>0,Q\in\mathcal{M})$ and concave in Q; the fact that this function is convex in $t>0$ is already known to us. Now let $P(\beta)$ be valid, and let us prove the validity of $\widehat{P}(\beta)$. Let us fix z and set $\theta(t,Q)=\Theta_Q(z,t)-t\log\beta$, and let $\gamma>0$. By $P(\beta)$, for every $Q\in\mathcal{M}$ there exists $t_Q>0$ such that $\theta(t,Q)<\gamma$. Since $\theta(t,Q)$ is continuous in $Q\in\mathcal{M}$, there exists a neighborhood (in $\mathcal{M}$) V_Q of the point Q such that $\theta(t_Q,Q')\le\gamma$ for all $Q'\in V_Q$. Since $\mathcal{M}$ is a compact set, there exist finitely many points $Q^i\in\mathcal{M}$ such that the corresponding neighborhoods V_{Q^i} cover the entire $\mathcal{M}$. In other words, there exist finitely many positive reals $t_1,\dots,t_N$ such that

$$
\min_{1\le i\le N}\theta(t_i,Q)\le\gamma\ \ \forall Q\in\mathcal{M}.
\tag{6.14}
$$

Since θ is concave and continuous in $Q \in \mathcal{M}$ and $\mathcal{M}$ is convex, (6.14) implies that

$$\exists \lambda^* \in \Delta_N := \left\{ \lambda \in \mathbb{R}^N_+ : \sum_i \lambda_i = 1 \right\} : \sum_i \lambda_i^* \theta(t_i, Q) \le \gamma \ \ \forall Q \in \mathcal{M}. \tag{6.15}$$

The latter conclusion is a standard fact of convex analysis. For the sake of a reader uncomfortable with possible infinite dimension of $\mathcal{M}$, here is a derivation of this fact from the standard von Neumann lemma. For $Q \in \mathcal{M}$, let Λ_Q be the set of those $\lambda \in \Delta_N$ for which $\sum_i \lambda_i \theta(t_i, Q) \le \gamma$; the set Λ_Q clearly is a closed subset of the finite-dimensional compact Δ_N. All we need is to prove that all these sets have a point in common (such a point can be taken as λ^*), and to this end it suffices to prove that all sets Λ_Q from a finite family $\Lambda_{Q_1}, \dots, \Lambda_{Q_M}$, $Q_j \in \mathcal{M}$, have a point in common. But the latter is readily given by the von Neumann lemma as applied to the convex hull Q_N of the points Q_j, $j = 1, \dots, M$ (which is a finite-dimensional convex compact set):

$$\gamma \ge \max_{Q \in Q_N} \min_{\lambda \in \Delta_N} \sum_{i=1}^N \lambda_i \theta(t_i, Q) = \min_{\lambda \in \Delta_N} \max_{Q \in Q_N} \sum_{i=1}^N \lambda_i \theta(t_i, Q)$$

(the inequality is given by (6.14), the equality by the von Neumann lemma; the required point in $\bigcap_i \Lambda_{Q_i}$ is $\operatorname{argmin}_{\lambda \in \Delta_N} \max_{Q \in Q_N} \sum_{i=1}^N \lambda_i \theta(t_i, Q)$).

Since θ is convex in $t > 0$, setting $t_\gamma = \sum_i \lambda_i^* t_i$ we get from (6.15) that $\Theta_Q(t_\gamma, z) - t_\gamma \log \beta \equiv \theta(t_\gamma, Q) \le \sum_i \lambda_i^* \theta(t_i, Q) \le \gamma$ for all $Q \in \mathcal{M}$, whence $\widehat{\Theta}(t_\gamma, z) - t_\gamma \log \beta \equiv \max_{Q \in \mathcal{M}} \Theta_Q(t_\gamma, z) - t_\gamma \log \beta \le \gamma$. Since t_γ is positive by construction and $\gamma > 0$ is arbitrary, we conclude that $\inf_{t>0} \left[\widehat{\Theta}(t_\gamma, z) - t_\gamma \log \beta \right] \le 0$, so that $\widehat{P}(\beta)$ is valid. ∎

Putting things together, we arrive at the following result.

THEOREM 6.2. *Assume that the ambiguous chance constrained problem* (6.9) *satisfies Assumptions* B1 and B2*, and let* α_i, $i = 1, \dots, m$, *be positive reals such that* $\sum_i \alpha_i \le \alpha$. *Then the program*

$$\min_{x \in X} f(x) \ \text{ subject to } \ \underbrace{\inf_{t>0} \left[f_{i0}(x) + t \widehat{\Psi}(t^{-1} z^i[x]) - t \log \alpha_i \right]}_{g_i(x,t)} \le 0, \ i = 1, \dots, m, \tag{6.16}$$

$$z^i[x] \ = \ (f_{i1}(x), \dots, f_{id}(x)), \ \ \widehat{\Psi}(z) = \max_{\{Q_j\}_{j=1}^d \in \mathcal{M}} \sum_{j=1}^d \log \left(\int \exp\{z_j s\} dQ_j(s) \right)$$

is a conservative approximation of problem (6.9)*: every feasible solution to the approximation is feasible for the chance constrained problem. This approximation is a convex program and is efficiently solvable, provided that all* f_{ij} *and* $\widehat{\Psi}$ *are efficiently computable, and* X *is computationally tractable.*

Proof. Function $g_i(x,t)$ is obtained from the function $\theta_i(z,t) := \widehat{\Theta}(z,t) - t \log \alpha_i$ by the substitution

$$(z,t) \leftarrow ((f_{i0}(x), f_{i1}(x), \dots, f_{id}(x)), t).$$

The outer function $\theta_i(z,t)$ is convex and nondecreasing in z_0 and every z_j with $j \in J$ (see the remarks following (6.10)). The inner functions $f_{i0}(x)$, $f_{ij}(x)$, $j \ge 1$, are

convex on X, and functions $f_{ij}(x)$ with $0 < j \notin J$ are affine. It follows that $g_i(x,t)$ is convex in $(t > 0, x \in X)$, so that (6.16) is indeed a convex program. Further, if x is feasible for (6.16), then $x \in X$, and for every i the predicate $\widehat{P}(\alpha_i)$ corresponding to $z = (f_{i0}(x), f_{i1}(x), \ldots, f_{id}(x))$ is valid, which, by (6.12), implies that

$$\sup_{P\in\mathfrak{P}} \mathrm{Prob}_P \left\{ f_{i0}(x) + \sum_{j=1}^{d} \xi_j f_{ij}(x) > 0 \right\} \le \alpha_i.$$

Since $\sum_i \alpha_i \le \alpha$, x is feasible for (6.9). □

Remark 6.1. Assumption B1 requires, among other things, from all distributions $P \in \mathfrak{P}$ to be supported on a common *compact* set $D_1 \times \cdots \times D_d$. This requirement can be straightforwardly relaxed to the requirement for all $P \in \mathfrak{P}$ to have "uniformly light tails": there exists a function $\gamma(t)$, $t > 0$, such that $\exp\{\alpha t\}\gamma(t) \to 0$ as $t \to \infty$ for all α, and for every $Q = \{Q_j\} \in \mathcal{M}$, every j and every $t > 0$ one has $Q_j(\{s : |s| \ge t\}) \le \gamma(t)$.

Examples. In order not to care for nonnegativity of ξ_j's associated with nonaffine $f_{ij}(\cdot)$, we assume from now on that all functions f_{ij}, $j = 1, \ldots, d$, in (6.9) are affine.

Example 1 (*range information on* ξ_j). Assume that all we know about the distributions of ξ is that ξ_j take values in given finite segments (and, as always, that $\xi_1, \ldots, \xi_d$ are independent). By shifting and scaling $f_{ij}(x)$, we may assume w.l.o.g. that ξ_j are independent and take values in $[-1, 1]$. This corresponds to the case where $\mathcal{M}$ is the set of all d-element tuples of Borel probability distributions supported on $[-1, 1]$. Denoting by Π the set of all Borel probability measures on $[-1, 1]$, we have

$$\begin{aligned} \widehat{\Phi}(z) &= z_0 + \sum_{j=1}^{d} \max_{P_j\in\Pi} \log\left(\int \exp\{z_j s\} dP_j(s) \right) = z_0 + \sum_{j=1}^{d} |z_j|, \\ \widehat{\Theta}(z,t) &= t\widehat{\Phi}(t^{-1}z) = z_0 + \sum_{j=1}^{d} |z_j|; \end{aligned}$$

consequently, approximation (6.16) becomes

$$\min_{x\in X} f(x) \quad \text{subject to} \quad \inf_{t>0} \left[f_{i0}(x) + \sum_{j=1}^{d} |f_{ij}(x)| - t \log \alpha_i \right] \le 0,\ i = 1, \ldots, m,$$

or, which is the same due to $\alpha_i \le 1$,

$$\min_{x\in X} f(x) \quad \text{subject to} \quad f_{i0}(x) + \sum_{j=1}^{d} |f_{ij}(x)| \le 0,\ i = 1, \ldots, m. \tag{6.17}$$

As it could be expected, in the situation in question, Bernstein approximation recovers the *robust counterpart* (RC) of the original uncertain problem [3], which in our case is the semi-infinite optimization program:

$$\min_{x\in X} f(x) \quad \text{subject to} \quad f_{i0}(x) + \sum_{j=1}^{d} \xi_j f_{ij}(x) \le 0\ \forall i,\ \forall \xi \in \bigcup_{P\in\mathfrak{P}} \mathrm{supp}(P). \tag{RC}$$

It is clear that in the extreme case we are considering the approximation is *exactly equivalent* to the chance constrained problem (6.9). A relatively good fact of Bernstein

approximation (6.17) is that in our example it is no more conservative than (RC). It is immediately seen that this is a general fact: whenever Bernstein approximation (6.16) is well defined, its feasible set contains the feasible set of (RC).

We see that when all our knowledge on uncertainty is the ranges of ξ_j, both the chance constrained problem (6.9) itself and its Bernstein approximation become the completely worst-case oriented (RC). The situation changes dramatically when we add something to the knowledge of ranges, for example, assume that we know the expected values of ξ_j.

Example 2 (*ranges and expectations of ξ_j are known*). Assume that we know that ξ_j are independent, take values in known finite segments, and have known expectations. As in Example 1, we may further assume w.l.o.g. that ξ_j vary in $[-1,1]$ and have known expectations μ_j, $|\mu_j| \le 1$. We are in the situation where $\mathcal{M}$ is the set of all tuples $\{Q_j\}_{j=1}^d$ with Q_j belonging to the family Π_{μ_j} of all Borel probability distributions on $[-1,1]$ with expectation μ_j, $j = 1,\dots,d$, and $\mathfrak{P}$ is the set of all product distributions on $\mathbb{R}^d$ with the collection of marginal distributions belonging to $\mathcal{M}$. It is easy to see that when $|\mu| \le 1$, then

$$\Lambda_\mu(t) := \max_{Q\in\Pi_\mu} \log\left(\int \exp\{ts\}dQ(s)\right) = \log(\cosh(t) + \mu\sinh(t))^5$$

and that $\Lambda_\mu(0) = 0$, $\Lambda'_\mu(0) = \mu$, and $\Lambda''_\mu(t) \le 1$ for all t, whence

$$\Lambda_\mu(s) \le \mu s + s^2/2 \ \forall s.$$

We therefore have

$$\begin{aligned}
\widehat{\Phi}(z) &:= \max_{P\in\mathfrak{P}} \log\left(\mathbb{E}_P\left\{\exp\{z_0 + \sum_{j=1}^d \xi_j z_j\}\right\}\right)\\
&= z_0 + \sum_{j=1}^d \log\left(\cosh(z_j) + \mu_j\sinh(z_j)\right)\\
&\le \widetilde{\Phi}(z) := z_0 + \sum_{j=1}^d \left[\mu_j z_j + z_j^2/2\right], \qquad (6.18)\\
\widehat{\Theta}(z,t) &:= t\widehat{\Phi}(t^{-1}z) = z_0 + \sum_{j=1}^d t\log\left(\cosh(t^{-1}z_j) + \mu_j\sinh(t^{-1}z_j)\right)\\
&\le \widetilde{\Theta}(z,t) := z_0 + \sum_{j=1}^d \mu_j z_j + (2t)^{-1}\sum_{j=1}^d z_j^2.
\end{aligned}$$

To proceed, we were supposed to compute the functions

$$G(z,\beta) := \inf_{t>0}\left[\widehat{\Theta}(z,t) - t\log\beta\right]$$

[5]Here is the verification: let $\lambda = \sinh(t)$ and $g(s) = \exp\{ts\} - \lambda s$. This function is convex and therefore takes its maximum on $[-1,1]$ at an endpoint; it is immediately seen that this maximum is $g(1) = g(-1) = \cosh(t)$. It follows that when $Q \in \Pi_\mu$, one has $\int \exp\{ts\}dQ(s) = \int g(s)dQ(s) + \lambda\mu = \cosh(t) + \mu\sinh(t)$. The resulting upper bound on $\int \exp\{ts\}dQ(s)$ is achieved when Q is a two-point distribution with mass $(1+\mu)/2$ at 1 and mass $(1-\mu)/2$ at -1.

and write down Bernstein approximation (6.16) of the ambiguous chance constrained problem in question as the convex program

$$\begin{array}{l}\min\limits_{x\in X}\left\{f(x) : G(z^i[x],\alpha_i)\le 0,\ i=1,\dots,m\right\},\\ \qquad z^i[x]=(f_{i0}(x),f_{i1}(x),\dots,f_{id}(x))^T,\end{array}\tag{6.19}$$

where $\alpha_i > 0$ are chosen to satisfy $\sum_i \alpha_i \le \alpha$. While computing $G(z,\beta)$ and its derivatives in z_j numerically (which is all we need in order to solve convex program (6.19) numerically) is easy, a closed form analytic expression for this function seems to be impossible. What we can do analytically is to bound G from above,[6] exploiting the simple upper bound on $\widehat{\Theta}$ presented in (6.18). From the concluding inequality in (6.18) it follows that

$$\begin{array}{rcl} G(z,\beta) & := & \inf\limits_{t>0}\left[\widehat{\Theta}(z,t)-t\log\beta\right]\\ & \le & G_*(z,\beta):=\inf\limits_{t>0}\left[z_0+\sum\limits_{j=1}^d\mu_j z_j+(2t)^{-1}\sum\limits_{j=1}^d z_j^2-t\log\beta\right]\\ & = & z_0+\sum\limits_{j=1}^d\mu_j z_j+\sqrt{2\log(1/\beta)}\left(\sum\limits_{j=1}^d z_j^2\right)^{1/2}.\end{array}\tag{6.20}$$

It follows that the convex optimization program

$$\min_{x\in X}\left\{f(x):\begin{array}{c} f_{i0}(x)+\sum_{j=1}^d\mu_j f_{ij}(x)\\ +\sqrt{2\log(1/\alpha_i)}\left(\sum_{j=1}^d f_{ij}^2(x)\right)^{1/2}\le 0,\ i=1,\dots,m\end{array}\right\}\quad[\textstyle\sum_i\alpha_i\le\alpha]$$

is an approximation (more conservative than Bernstein) of the ambiguous chance constrained problem (6.9), where the independent-of-each-other random perturbations ξ_j are known to vary in $[-1,1]$ and possess expected values μ_j. As could be expected, we have recovered (a slightly refined version of) the results of [2] mentioned in the introduction (see (1.9) and Remark 3.1.

Comparing (6.17) and (6.19)–(6.20), we clearly see how valuable the information on expectations of ξ_j could be, provided that ξ_j are independent (this is the only case we are considering). First of all, from the origin of $G(z,\beta)$ it follows that the left-hand sides of constraints in (6.17) are pointwise and $\ge$ their counterparts in (6.19), so that (6.19) is always less conservative than (6.17). To see how large the corresponding "gap" could be, consider the case when all ξ_j have zero means ($\mu_j = 0$ for all j). In this case, the ith constraint in (6.17) requires from the vector $h_i(x) := (f_{i1}(x),\dots,f_{id}(x))^T$ to belong to the centered at the origin $\|\cdot\|_1$-ball of radius $\rho(x) = -f_{i0}(x)$, let this ball be called $V_1(x)$. The ith constraint in (6.19), first, allows for $h_i(x)$ to belong to $V_1(x)$ (recall that (6.19) is less conservative than (6.17)) and, second, allows for this vector to belong to the centered at the origin $\|\cdot\|_2$-ball $V_2(x)$ of the radius $\kappa^{-1}\rho(x)$, where $\kappa=\sqrt{2\log(1/\alpha_i)}$ (see (6.20) and take into account that $\mu_j\equiv 0$);

[6] It should be stressed that this bounding is completely irrelevant as far as the numerical processing of (6.19) is concerned; the only purpose of the exercise to follow is to link our approach with some previously known constructions.

by convexity, it follows that the ith constraint in (6.19) allows for $h_i(x)$ to belong to the set $V_{1,2}(x) = \mathrm{Conv}\{V_1(x) \cup V_2(x)\} \supset V_1(x)$. When d is not small, the set $V_{1,2}(x)$ is not merely larger, it is "much larger" than $V_1(x)$, and, consequently, the ith constraint in (6.19) is "much less restricting" than its counterpart in (6.17). To get an impression of what "much larger" means, note that the distance from the origin to the boundary of $V_2(x)$ along every direction is $\kappa^{-1}\rho(x)$; the distance to the boundary of $V_{1,2}(x)$ can only be larger. At the same time, the distance from the origin to the boundary of $V_1(x)$ along a randomly chosen direction is, with probability approaching 1 as $d \to \infty$, at most $\sqrt{\pi/2}(1+\delta)d^{-1/2}$ for every fixed $\delta > 0$. Thus, the ratio of the distances, taken along a randomly chosen direction, from the origin to the boundaries of $V_{1,2}(x)$ and of $V_1(x)$ is always ≥ 1, and with probability approaching 1 as $d \to \infty$, is at least $(1-\delta)\kappa^{-1}\sqrt{2d/\pi}$ for every $\delta > 0$; in this sense $V_{1,2}$ is "at average" nearly $\kappa^{-1}\sqrt{2d/\pi}$ times larger in linear sizes than $V_1(x)$. Now, for all practical purposes κ is a moderate constant;[7] thus, we can say that as d grows, approximation (6.19) becomes progressively ("by factor $\sqrt{d}$") less conservative than (6.17).

Coming back to our examples, observe that if $\mathcal{M} = \Pi_1 \times \cdots \times \Pi_d$, where Π_j is a given set in the space of Borel probability distributions on the axis, we have

$$\widehat{\Phi}(z) = z_0 + \sum_{j=1}^{d} \max_{Q \in \Pi_j} \log\left(\int \exp\{z_j s\} dQ(s)\right),$$

and therefore computation of $\widehat{\Phi}(z)$ (which is all we need in order to build Bernstein approximation) reduces to computing the functions $\Lambda^{\Pi}(t) \equiv \max_{Q\in\Pi} \log\left(\int \exp\{ts\}dQ(s)\right)$ for $\Pi = \Pi_1, \ldots, \Pi_d$. In Table 2, we present explicit expressions for $\Lambda^{\Pi}(\cdot)$ for a number of interesting sets Π comprised of distributions with support in a given finite segment (which we w.l.o.g. can assume to be $[-1, 1]$). In the table, Mean$[Q]$, Var$[Q]$ stand for the mean $\int s dQ(s)$ and the second moment $\int s^2 dQ(s)$ of distribution Q; to save notation, we present the expressions for $\exp\{\Lambda^{\Pi}(t)\}$ rather than for Λ^{Π} itself.

Example 3 (*"light tail" families*). In previous examples, all distributions from Π were supported on a fixed finite segment. Now consider the case when Π is comprised of Borel probability distributions P on the axis such that $\mathbb{E}_P[\exp\{|x^r|/r\}] \leq \exp\{\sigma^r/r\}$, where $r \in (1, \infty)$ and $\sigma \in (0, \infty)$ are given parameters. In this case, precise computations of $\Lambda^{\Pi}(t)$ seems to be difficult, but we can point out a tight convex upper bound on $\Lambda^{\Pi}(\cdot)$, specifically,

$$\Lambda^{\Pi}(t) \leq \begin{cases} \sigma|t|, & |t| \leq \sigma^{r-1} \\ \sigma^r/r + |t|^{r_*}/r_*, & |t| \geq \sigma^{r-1}, \end{cases} \quad r_* = r/(r-1). \tag{6.21}$$

This bound coincides with $\lambda_{\Pi}(t)$ when $|t| \leq \sigma^{r-1}$ and coincides with $\Lambda^{\Pi}(t)$ within additive constant $-\log(1 - \exp\{-\sigma^r/r\})$ when $|t| \geq \sigma^{r-1}$.

Here is a justification. It suffices to verify (6.21) when $t \geq 0$. Let $P \in \Pi$. We have $|x|^r/r + t^{r_*}/r_* - tx \geq 0$ for all x, whence $\int \exp\{tx\}dP(x) \leq \int \exp\{|x|^r/r + t^{r_*}/r_*\}dP(x) \leq \exp\{\sigma^r/r + t^{r_*}/r_*\}$; thus, (6.21) holds true when $t \geq \sigma^{r-1}$. Now let us prove that (6.21) is true when $0 \leq t \leq \sigma^{r-1}$. In this range, the bound in (6.21) is true when $t = 0$ and is linear in t, while $\Lambda^{\Pi}(t)$ is convex in t, so that it suffices to verify

[7] With $\alpha_i = \alpha/m$, even risk as small as α =1.e-12 and the number of constraints as large as $m = 10,000,000$ result in $\kappa \leq 9.4$.

TABLE 2

$\exp\{\Lambda^\Pi(\cdot)\}$ *for several families* Π *of univariate distributions. The parameters* μ, σ^2 *are subject to natural restrictions* $|\mu| \le 1$, $\sigma^2 \le 1$, $\mu^2 \le \sigma^2$.

Π	$\exp\{\Lambda^\Pi(t)\}$
$\{Q : \mathrm{supp}(Q) \subset [-1,1],\}$	$\exp\{\lvert t\rvert\}$
$\left\{Q : \begin{array}{c}\mathrm{supp}(Q) \subset [-1,1],\\ Q \text{ is symmetric}\end{array}\right\}$	$\cosh(t)$
$\left\{Q : \begin{array}{c}\mathrm{supp}(Q) \subset [-1,1], Q \text{ is}\\ \text{unimodal w.r.t. } 0^a\end{array}\right\}$	$\dfrac{\exp\{\lvert t\rvert\}-1}{\lvert t\rvert}$
$\left\{Q : \begin{array}{c}\mathrm{supp}(Q) \subset [-1,1], Q\\ \text{is unimodal w.r.t.}\\ 0 \text{ and symmetric}\end{array}\right\}$	$\dfrac{\sinh(t)}{t}$
$\left\{Q : \begin{array}{c}\mathrm{supp}(Q) \subset [-1,1],\\ \mathrm{Mean}[Q] = \mu\end{array}\right\}$	$\cosh(t) + \mu\sinh(t)$
$\left\{Q : \begin{array}{c}\mathrm{supp}(Q) \subset [-1,1],\\ \mu_- \le \mathrm{Mean}[Q] \le \mu_+\end{array}\right\}$	$\cosh(t) + \max[\mu_-\sinh(t), \mu_+\sinh(t)]$
$\left\{Q : \begin{array}{c}\mathrm{supp}(Q) \subset [-1,1]\\ \mathrm{Mean}[Q] = 0, \mathrm{Var}[Q] \le \sigma^2\end{array}\right\}$	$\dfrac{\exp\{-\lvert t\rvert\sigma^2\} + \sigma^2\exp\{\lvert t\rvert\}}{1+\sigma^2}$
$\left\{Q : \begin{array}{c}\mathrm{supp}(Q) \subset [-1,1], Q \text{ is}\\ \text{symmetric, } \mathrm{Var}[Q] \le \sigma^2\end{array}\right\}$	$\sigma^2\cosh(t) + (1-\sigma^2)$
$\left\{Q : \begin{array}{c}\mathrm{supp}(Q) \subset [-1,1],\\ \mathrm{Mean}[Q] = \mu, \mathrm{Var}[Q] \le \sigma^2\end{array}\right\}$	$\begin{cases}\frac{(1-\mu)^2\exp\{t\frac{\mu-\sigma^2}{1-\mu}\}+(\sigma^2-\mu^2)\exp\{t\}}{1-2\mu+\sigma^2}, & t \ge 0\\ \frac{(1+\mu)^2\exp\{t\frac{\mu+\sigma^2}{1+\mu}\}+(\sigma^2-\mu^2)\exp\{-t\}}{1+2\mu+\sigma^2}, & t \le 0\end{cases}$

[a] Q is unimodal w.r.t. 0 if Q is the sum of two measures: a mass at 0 and a measure with density $p(s)$ which is nondecreasing when $t \le 0$ and nonincreasing when $t \ge 0$.

the bound's validity when $t = \sigma^{r-1}$. This we already know, since with $t = \sigma^{r-1}$ we have $\sigma^r/r + t^{r_*}/r_* = t\sigma$. Further, when $0 \le t \le \sigma^{r-1}$, our upper bound coincides with $\Lambda^\Pi(t)$—look what happens when P assigns mass 1 to the point $x = \sigma$. Finally, let $t > \sigma^{r-1}$, and let P assign the mass $\mu = \lambda\exp\{(\sigma^r - t^{r_*})/r\}$ to the point t^{r_*-1} and the mass $1-\mu$ to the point 0; here $\lambda = (1-\exp\{-\sigma^r/r\})\,/\,(1-\exp\{-t^{r_*}/r\})$. Since $t \ge \sigma^{r-1}$, we have $t^{r_*} \ge \sigma^r$, so that $\lambda \le 1$ and $\mu \in [0,1]$; thus, P indeed is a probability distribution. An immediate computation shows that $\int \exp\{|x|^r/r\}dP(x) = \exp\{\sigma^r/r\}$, so that $P \in \Pi$. We now have $\int \exp\{tx\}dP(x) \ge \mu\exp\{t^{r_*}\} = \lambda\exp\{\sigma^r/r + t^{r_*}/r_*\}$, so that $\Lambda^\Pi(t) \ge \sigma^r/r + t^{r_*}/r_* - \log\lambda \ge \sigma^r/r + t^{r_*}/r_* - \log(1-\exp\{-\sigma^r/r\})$.

We could proceed in the same fashion, adding more a priori information on the distribution of ξ; until this information becomes too complicated for numerical processing, it can be "digested" by Bernstein approximation. Instead of moving in this direction, we prefer to present an example of another sort, where the assumptions underlying Theorem 6.2 are severely violated, but the Bernstein approximation scheme still works.

Example 4 (*parametric uncertainty*). Assume that we know a priori that some of ξ_j are normal, and the remaining ones are Poisson; however, we do not know exactly the parameters of the distributions. Specifically, let us parameterize a normal distribution by its mean and variance (note: variance, not standard deviation!), and a Poisson distribution by its natural parameter λ (so that the probability for the corresponding random variable to attain value $i = 0, 1, \ldots$ is $\frac{\lambda^i}{i!}\exp\{-\lambda\}$). Let us arrange parameters of the d distributions in question in a vector ω, and assume that our a priori knowledge is that ω belongs to a known-in-advance convex compact set

Ω. We assume also that the latter set is "realizable" in the sense that every point $\omega \in \Omega$ indeed represents a collection of distributions of the outlined type; specifically, the coordinates of $\omega \in \Omega$ which represent variances of normal distributions and the parameters of the Poisson distributions are positive. Note that our a priori knowledge is incompatible with assumption B1: convexity in the space of parameters has little in common with convexity in the space of distributions. For example, when the mean of a normal distribution with unit variance runs through a given segment, the distribution itself moves along a complicated curve. We can, however, try to use the same approach which led us to Theorem 6.2. Observe that when P_j is the Poisson distribution with parameter λ, we have

$$\begin{aligned}\log\left(\int \exp\{rs\}dP_j(s)\right) &= \log\left(\sum_{i=0}^{\infty}\frac{(\lambda e^r)^i}{i!}\exp\{-\lambda\}\right) = \log(\exp\{\lambda e^r - \lambda\}) \\ &= \lambda\exp\{r\} - \lambda;\end{aligned}$$

the resulting function is continuous, convex in r, as is always the case for the logarithmic moment generating function, and is concave in λ, which is pure luck. We are equally lucky with the normal distribution P_j with mean μ and variance ν:

$$\log\left(\int \exp\{rs\}dP_j(s)\right) = \log\left(\frac{1}{\sqrt{2\pi\nu}}\int \exp\left\{rs - \frac{(s-\mu)^2}{2\nu}\right\}ds\right) = r\mu + \frac{r^2\nu}{2},$$

and the result again is continuous, convex in r and concave in (μ, ν). It follows that if P^ω is the joint distribution of the sequence of d normal/Poisson independent random variables ξ_j, the vector of parameters of the marginal distributions being ω, then, for every vector $z \in \mathbb{R}^{d+1}$, the function

$$\Phi_\omega(z) = \log\left(\mathbb{E}_{P^\omega}\left[\exp\left\{z_0 + \sum_{j=1}^{d}\xi_j z_j\right\}\right]\right)$$

is given by a simple explicit expression, is continuous in $(z \in \mathbb{R}^{d+1}, \omega \in \Omega)$, and is convex in z and concave (in fact even affine) in ω. We now can use the reasoning which led us to Theorem 6.2 and (6.16) to conclude that the optimization problem

$$\begin{gathered}\min_{x\in X} f(x) \quad \text{subject to} \quad \inf_{t>0}\left[t\widehat{\Phi}(t^{-1}z^i[x]) - t\log\alpha_i\right] \le 0,\ i = 1,\dots,m, \\ \widehat{\Phi}(z) = \max_{\omega\in\Omega}\Phi_\omega(z),\ z^i[x] = (f_{i0}(x), f_{i1}(x), \dots, f_{id}(x))\end{gathered}$$

is an approximation of the ambiguous chance constrained problem under consideration, provided that $\alpha_i \in (0,1)$ are such that $\sum_i \alpha_i \le \alpha$. This approximation is convex, provided that all functions f_{ij} are convex and well defined on X and the functions f_{ij} with j's corresponding to normally distributed components in ξ are affine. Finally, our approximation is computationally tractable, provided that $\widehat{\Phi}(\cdot)$ is efficiently computable (which indeed is the case when Ω is computationally tractable).

Acknowledgment. We express our gratitude to Yuri Kan who brought to our attention the paper of Pinter [20].

REFERENCES

[1] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath, *Coherent measures of risk*, Math. Finance, 9 (1999), pp. 203–228.

[2] A. Ben-Tal and A. Nemirovski, *Robust solutions of Linear Programming Problems Contaminated with Uncertain Data*, Math. Program., 88 (2000), pp. 411–424.

[3] A. Ben-Tal and A. Nemirovski, *Robust Optimization—Methodology and Applications*, Math. Program., 92 (2002), pp. 453–480.

[4] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization*, SIAM, Philadelphia, 2001.

[5] D. Bertsimas and M. Sim, *The price of robustness*, Oper. Res., 52 (2004), pp. 35–53.

[6] G. Calafiore and M. C. Campi, *Uncertain convex programs: Randomized solutions and confidence levels*, Math. Program., 102 (2005), pp. 25–46.

[7] G. Calafiore and M. C. Campi, *The scenario approach to robust control design*, IEEE Trans. Automat. Control, 51 (2006), pp. 742–753.

[8] A. Charnes, W. W. Cooper, and G. H. Symonds, *Cost horizons and certainty equivalents: An approach to stochastic programming of heating oil*, Management Science, 4 (1958), pp. 235–263.

[9] A. Dembo and O. Zeitouni, *Large Deviations, Techniques, and Applications*, Springer-Verlag, New York, 1998.

[10] D. P. de Farias and B. Van Roy, *On constraint sampling in the linear programming approach to approximate dynamic programming*, Math. Oper. Res., 29 (2004), pp. 462–478.

[11] D. Dentcheva, A. Prekopa, and A. Ruszczyński, *Concavity and efficient points of discrete distributions in probabilistic programming*, Math. Program., 89 (2000), pp. 55–77.

[12] J. Dupačová, *The minimax approach to stochastic programming and an illustrative application*, Stochastics, 20 (1987), pp. 73–88.

[13] W. K. Klein Haneveld, *Duality in stochastic linear and dynamic programming*, Lecture Notes in Economics and Mathematical Systems 274, Springer-Verlag, Berlin, 1986.

[14] E. Erdoğan and G. Iyengar, *Ambiguous chance constrained problems and robust optimization*, Math. Program., 107 (2006), pp. 37–61.

[15] L. G. Khachiyan, *The problem of calculating the volume of a polyhedron is enumerably hard*, Russian Math. Surveys, 44 (1989), pp. 199–200.

[16] C. M. Lagoa, X. Li, and M. Sznaier, *Probabilistically constrained linear programs and risk-adjusted controller design*, SIAM J. Optim., 15 (2005), pp. 938–951.

[17] L. B. Miller and H. Wagner, *Chance-constrained programming with joint constraints*, Oper. Res., 13 (1965), pp. 930–945.

[18] A. Nemirovski, *On tractable approximations of randomly perturbed convex constraints*, in Proceedings of the 42nd IEEE Conference on Decision and Control, Maui, HI, 2003, pp. 2419–2422.

[19] A. Nemirovski and A. Shapiro, *Scenario approximations of chance constraints*, in Probabilistic and Randomized Methods for Design under Uncertainty, G. Calafiore and F. Dabbene, eds., Springer-Verlag, London, 2005.

[20] J. Pinter, *Deterministic approximations of probability inequalities*, Oper. Res., 33 (1989), pp. 219–239.

[21] A. Prékopa, *On probabilistic constrained programming*, in Proceedings of the Princeton Symposium on Mathematical Programming, Princeton University Press, Princeton, NJ, 1970, pp. 113–138.

[22] A. Prékopa, *Stochastic Programming*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1995.

[23] A. Prékopa, B. Vizvári, and T. Badics, *Programming under probabilistic constraint with discrete random variable*, in New Trends in Mathematical Programming, L. Grandinetti et al., eds., Kluwer Academic Publishers, Boston, 1998, pp. 235–255.

[24] R. T. Rockafellar and S. P. Uryasev, *Optimization of conditional value-at-risk*, J. Risk, 2 (2000), pp. 21–41.

[25] R. T. Rockafellar, S. Uryasev, and M. Zabarankin, *Generalized deviations in risk analysis*, Finance Stoch., 10 (2006), pp. 51–74.

[26] A. Ruszczyński and A. Shapiro, *Optimization of risk measures*, in Probabilistic and Randomized Methods for Design under Uncertainty, G. Calafiore and F. Dabbene, eds., Springer-Verlag, London, 2005, pp. 117–158.

[27] J. Žáčková, *On minimax solutions of stochastic linear programming problems*, Čas. Pěst. Math., 91 (1966), pp. 423–430.

SIAM J. OPTIM.
Vol. 17, No. 4, pp. 997–1014
© 2006 Society for Industrial and Applied Mathematics

DETERMINANT MAXIMIZATION OF A NONSYMMETRIC MATRIX WITH QUADRATIC CONSTRAINTS*

SERGE DÉGERINE† AND ABDELHAMID ZAÏDI‡

Abstract. This paper presents the problem of maximizing the determinant of a real $K \times K$-matrix B, subject to the constraint that each row b_k of B satisfies $b_k^t \Gamma_k b_k \leq 1$, where $\Gamma_1, \ldots, \Gamma_K$ are K given real symmetric positive definite matrices. This problem comes from a specific blind signal separation approach, but the criterion differs from approximate diagonalization criteria usually encountered in this area. Furthermore our criterion corresponds to the following nice geometrical problem: given K ellipsoids in $\mathbf{R}^K, \varepsilon_k = \{x : x^t \Gamma_k x \leq 1\}, k = 1, \ldots, K$, find K vectors, $b_1 \in \varepsilon_1, \ldots, b_K \in \varepsilon_K$, such that the volume of the parallelepiped defined by these vectors is maximum. Existence and uniqueness of the solution are discussed. An iterative algorithm, based on a relaxation technique, is proposed in order to solve this problem, and its convergence is proved under a simple sufficient condition. Some numerical experiments are performed showing the behavior of the algorithm and its comparison with Newton's methods for nonlinear optimization.

Key words. determinant maximization, relaxation technique, nonconvex optimization

AMS subject classifications. 15A15, 49M20, 90C26

DOI. 10.1137/050622821

1. Introduction.

We consider the optimization problem

$$\begin{aligned} &\text{maximize} \quad \det B, \quad B = [b_1, \ldots, b_K]^t, \\ &\text{subject to} \quad b_k^t \Gamma_k b_k \leq 1, \quad k = 1, \ldots, K, \end{aligned} \tag{1.1}$$

where $\mathcal{G} = \{\Gamma_1, \ldots, \Gamma_K\}$ is a set of K given real symmetric positive definite matrices.

From a geometrical point of view, this problem is equivalent to the following: Given K ellipsoids in $\mathbf{R}^K, \varepsilon_k = \{x : x^t \Gamma_k x \leq 1\}, k = 1, \ldots, K$, find K vectors, $b_1 \in \varepsilon_1, \ldots, b_K \in \varepsilon_K$, such that the volume of the parallelepiped defined by these vectors is maximum. Note that the dimension K of the problem is equal to the number of ellipsoids. As we will see in section 3, this nice geometrical problem has an explicit solution for $K = 2$ and for some particular classes of jointly diagonalizable matrices. Otherwise, the existence of a solution is easily proved, but the uniqueness is a tricky problem. This last point will be illustrated by several examples with diagonal matrices when $K = 3$.

We will see in section 3.4 below that this max-det problem, with $|\det B|$ instead of $\det B$, is equivalent to maximizing, with respect to B, the following criterion coming from a blind source separation problem:

$$l(B; \mathcal{G}) = \log |\det B| - \frac{1}{2} \sum_{k=1}^{K} b_k^t \Gamma_k b_k. \tag{1.2}$$

This max-det problem becomes a convex optimization problem, as those considered in [14], only when B is restricted to the cone of symmetric positive definite matrices. We propose an algorithm, using a relaxation technique, for this max-det

*Received by the editors January 17, 2005; accepted for publication (in revised form) July 20, 2006; published electronically November 22, 2006.
http://www.siam.org/journals/siopt/17-4/62282.html

†Laboratory LMC/IMAG, University Joseph Fourier, B.P. 53, 38041 Grenoble cedex 9, France (Serge.Degerine@imag.fr).

‡INSAT, Département Mathématique et Informatique, zone urbaine la Charguia II, 1002 Tunis, Tunisie (abdelhamidzaidi@yahoo.com).

problem in the set of all square matrices. Numerical experiments in the last section show that this algorithm is very fast in comparison with Newton's methods for large values of K. Each step of our algorithm consists of maximizing $l(B;\mathcal{G})$ with respect to one row b_k when the other rows $b_j, j \neq k$, are fixed; $l(B;\mathcal{G})$ is strictly concave with respect to b_k and the solution $\hat{b}_k$ is explicit. Denoting by $\langle\cdot,\cdot\rangle_{\Gamma_k}$ the inner product in $\mathbf{R}^K$, defined by $\langle x, y\rangle_{\Gamma_k} = x^t\Gamma_k y$, and by $\|\cdot\|_{\Gamma_k}$ the corresponding norm, this solution satisfies $\|\hat{b}_k\|_{\Gamma_k} = 1$ and $\langle\hat{b}_k, b_j\rangle_{\Gamma_k} = 0$ for $j \neq k$. So, $\hat{b}_k$ is the normalized projection error of b_k with respect to $\langle\cdot,\cdot\rangle_{\Gamma_k}$ on the subspace spanned by the rows $b_j, j \neq k$. Notice that the existence of a solution to our max-det problem (Proposition 3.9) proves the following relevant fact: Given K inner products $\langle\cdot,\cdot\rangle_{\Gamma_k}, k = 1,\ldots,K$, in $\mathbf{R}^K$, there exist K vectors $b_1,\ldots,b_K$ such that, for each k, b_k is orthogonal to $b_j, j \neq k$, with respect to $\langle\cdot,\cdot\rangle_{\Gamma_k}$.

The criterion $l(B;\mathcal{G})$ comes from a blind source separation method based on the maximum likelihood principle [8]. In this area, other methods lead to approximate diagonalization criteria of a set of N matrices $C_1,\ldots,C_N$ (see, for example, [1], [3], [11], [15]). Here, the required matrix B is such that, for each k, only the off-diagonal elements of the kth row (and column) of $B\Gamma_k B^t$ are set to zero and $N = K$. Thus $l(B;\mathcal{G})$ is not an approximate diagonalization criterion, except for $K = 2$.

The blind source separation problem is presented in section 2, and the max-det problem is studied in section 3. The last section is devoted to the algorithm.

2. The blind source separation problem. Let $X(t) = AS(t), t = 1,\ldots,T$, be the observations of an instantaneous linear mixture of sources $S(\cdot)$. The mixing matrix A is an unknown $K \times K$ nonsingular matrix. The goal is to extract the sources from the observations, with a minimum knowledge about the sources except that they are statistically independent. Independent components analysis exploits only the space independence between the sources [7]. In this case, each component $S_k(\cdot), k = 1,\ldots,K$, of the source process $S(\cdot)$ can be modeled as a sequence of independent identically distributed variables (white sources), provided that there is not more than one Gaussian source in the mixture [3]. Second-order methods exploit the decorrelation hypothesis but require sources having distinct normalized spectra (colored sources) [1]. The method proposed in [8] is in this last category, but, using the maximum entropy principle, $S_k(\cdot)$ is modeled as an $AR(p_k)$ stationary zero-mean Gaussian process. This model is parametrized by the set $\beta_k = \{\beta_k(j), j = 1,\ldots,p_k\}$ of its partial autocorrelation coefficients, since the innovation variance is fixed equal to 1. The innovation variance is a scale parameter which is integrated in the norm of the columns of A. So, $S_k = [S_k(1),\ldots,S_k(T)]^t$ is a zero-mean Gaussian random vector with a $T \times T$ covariance matrix, denoted by $R_k^T(\beta_k)$, which depends only on β_k. Using the independence of the sources, the probability density function of the bloc $S_T = [S(1),\ldots,S(T)]$ equals

$$[2\pi]^{-KT/2}\prod_{k=1}^{K}[\det R_k^T(\beta_k)]^{-1/2}\exp -\frac{1}{2}\sum_{k=1}^{K}S_k^t[R_k^T(\beta_k)]^{-1}S_k.$$

Thus, the log-likelihood function of the observations $X_T = [X(1),\ldots,X(T)] = AS_T$ is

$$l_T(X_T;B,\beta) = T\log|\det B| - \frac{1}{2}\sum_{k=1}^{K}b_k^t X_T[R_k^T(\beta_k)]^{-1}X_T^t b_k$$
$$-\frac{1}{2}\sum_{k=1}^{K}\log[\det R_k^T(\beta_k)] - \frac{KT}{2}\log(2\pi),$$

where $B = A^{-1}$ is the separating matrix and $\beta = \{\beta_1, \ldots, \beta_K\}$. Notice that B and β are two independent parameters. The maximization of $l_T(X_T; B, \beta)$ is obtained through a relaxation technique. When B is fixed, the maximization with respect to β leads to K independent problems of maximum likelihood estimation of AR models. When β is fixed, we recognize criterion (1.2) with

$$\Gamma_k = \frac{1}{T} X_T [R_k^T(\beta_k)]^{-1} X_T, \quad k = 1, \ldots, K.$$

Related approaches, in this blind source separation area, lead to various problems of approximate diagonalization of some sets of matrices [5], [10], [15]. For instance, the problem in [10] is to maximize, with respect to the separating matrix B, the following criterion:

$$\log|\det B| - \frac{1}{2}\sum_{n=1}^{N} \log\det \operatorname{diag}(BC_nB^t),$$

where $C_1, \ldots, C_N$ is a set of $K \times K$ symmetric positive definite matrices obtained from an estimate of the spectral density of $X(\cdot)$ [11]. This criterion measures the global deviation of the matrices BC_nB^t from diagonality and is invariant with respect to the scale of the rows of B.

3. The max-det problem. We consider the first and second derivatives of the criterion $l(B; \mathcal{G})$. Then, we describe some special cases in which an explicit solution is obtained. Finally, we discuss the existence and the uniqueness of the solution.

3.1. Derivatives of the criterion. Let $\{e_k, k = 1, \ldots, K\}$ be the canonical base of $\mathbf{R}^K$. We also use the notations b_{ij} and a_{ij} to designate the entries of B and $A = B^{-1}$, and δ_{ij} is the Kronecker symbol ($\delta_{ij} = 1$ if $i = j$, else 0).

PROPOSITION 3.1. *The first and second derivatives of the criterion $l(B; \mathcal{G})$ are given by*

$$\frac{\partial l(B; \mathcal{G})}{\partial b_{ij}} = a_{ji} - e_j^t \Gamma_i B^t e_i, \tag{3.1}$$

$$\frac{\partial^2 l(B; \mathcal{G})}{\partial b_{kl} \partial b_{ij}} = -a_{li}a_{jk} - \delta_{ki}\Gamma_i(j, l). \tag{3.2}$$

Proof. Let B_{ij} be the matrix formed by deleting row i and column j from B; then

$$\det B = \sum_k b_{ik}(-1)^{i+k} \det B_{ik}, \qquad a_{ji} = (-1)^{i+j} \det B_{ij} (\det B)^{-1}.$$

Using the derivative $(\log|x|)' = x^{-1}$ for $x \neq 0$, we obtain the gradient of $\log\det|B|$,

$$\frac{\partial \log|\det B|}{\partial b_{ij}} = a_{ji}, \qquad \nabla \log\det|B| = \frac{\partial \log|\det B|}{\partial B} = B^{-t}.$$

The differentiation (see [2])

$$0 = dI = d(B^tB^{-t}) = dB^tB^{-t} + B^t d(B^{-t})$$

gives $d(B^{-t}) = -B^{-t}dB^tB^{-t}$. So the quadratic form associated with the Hessian of $\log|\det B|$ is

$$\langle dB, \nabla^2 \log|\det B|(dB)\rangle = -\operatorname{trace}(dB^tB^{-t}dB^tB^{-t}).$$

The identification of the coefficient of $db_{kl}db_{ij}$ gives

$$\frac{\partial^2 \log\det|B|}{\partial b_{kl}\partial b_{ij}} = -a_{li}a_{jk}.$$

The derivatives of the quadratic part in the criterion are obvious. □

As we will see below (Proposition 3.10), the solution of our problem satisfies the optimality conditions $\nabla l(B;\mathcal{G}) = 0$. Using the first derivatives (3.1) and the notation $A = [a_1, \ldots, a_K]$, we have

$$\frac{\partial l(B;\mathcal{G})}{\partial b_k} = a_k - \Gamma_k B^t e_k = 0 \quad \Longleftrightarrow \quad a_k = \Gamma_k B^t e_k.$$

By applying the matrix B to this equality, we obtain the following result.

COROLLARY 3.2. *The optimality conditions $\nabla l(B;\mathcal{G}) = 0$ are equivalent to the following set of equations:*

$$B\Gamma_k B^t e_k = e_k, \quad k = 1, \ldots, K. \tag{3.3}$$

COROLLARY 3.3. *The Hessian matrix is negative definite, $\nabla^2 l(B;\mathcal{G}) \prec 0$, if and only if*

$$\operatorname{trace}(X^2) + \sum_{k=1}^{K} x_k^t B\Gamma_k B^t x_k > 0 \tag{3.4}$$

for all $K \times K$-matrices $X = [x_1, \ldots, x_K]^t \neq 0$.

Proof. Using the second derivatives (3.2), the Hessian is negative definite if and only if

$$\langle dB, \nabla^2 l(B;\mathcal{G})\rangle = -\operatorname{trace}(dB^t B^{-t} dB^t B^{-t}) - \sum_{k=1}^{K} \operatorname{trace}(e_k e_k^t dB\Gamma_k dB^t) < 0$$

for all $dB \neq 0$. Thus the result is given by taking $dB = BX^t$. □

Notice that $\operatorname{trace}(X^2)$ in (3.4) shows that the Hessian is not always negative definite. This point is the main difficulty in the implementation of the Newton methods. On the other hand, the following property is very useful.

LEMMA 3.4. *For any nonsingular square matrix M, we have*

$$l(BM^{-1};\mathcal{G}_M) = l(B;\mathcal{G}) - \log|\det M|,$$

where $\mathcal{G}_M = \{M\Gamma_1 M^t, \ldots, M\Gamma_K M^t\}$.

In order to maximize the criterion in a neighborhood of a nonsingular matrix B, it is apparently easier to maximize $l(X;\mathcal{G}_B)$ with respect to X, since we have

$$\left.\frac{\partial l(X;\mathcal{G}_B)}{\partial x_k}\right|_{X=I} = e_k - B\Gamma_k b_k, \qquad \left.\frac{\partial^2 l(X;\mathcal{G}_B)}{\partial x_l \partial x_k}\right|_{X=I} = -e_k e_l^t - \delta_{lk} B\Gamma_k B^t. \tag{3.5}$$

Let us denote by ∇_B and ∇_B^2 the gradient vector and the Hessian matrix associated with (3.5); i.e., $\nabla_B = [e_1^t - b_1^t\Gamma_1 B^t, \ldots, e_K^t - b_K^t\Gamma_K B^t]^t$ and $\nabla_B^2 = -[E + \Gamma_B]$ with

$$E = \begin{bmatrix} e_1 e_1^t & \cdots & e_K e_1^t \\ \vdots & \vdots & \vdots \\ e_1 e_K^t & \cdots & e_K e_K^t \end{bmatrix}, \qquad \Gamma_B = \begin{bmatrix} B\Gamma_1 B^t & & \\ & \ddots & \\ & & B\Gamma_K B^t \end{bmatrix}.$$

Notice that ∇_B and ∇_B^2 are gradient and Hessian of $l(X;\mathcal{G}_B)$ as a function of $(x_1^t,\dots,x_K^t)^t$, where $X=[x_1,\dots,x_K]^t$ at $X=I$, and are simply obtained from B without computing $A=B^{-1}$. In the algorithm, the new iterate is given by

$$B(n+1)=[I+\lambda_n dB]B(n), \tag{3.6}$$

where λ_n is a scale factor to be chosen and dB corresponds to the Newton step

$$[db_1^t,\dots,db_K^t]^t=dB,\qquad \nabla^2_{B(n)}d_B=-\nabla_{B(n)}.$$

Here, maximizing $l(X;\mathcal{G}_B)$ instead of $l(B;\mathcal{G})$ leads to the introduction of the relative gradient ∇_B. Generally, this approach simplifies the computation of the Hessian, but its more relevant interest is in the statistical properties of the corresponding methods in blind source separation problems [4].

Let $\mathcal{E}$ be a diagonal matrix whose diagonal elements are equal to ±1, called a sign matrix. Then, if B is a solution of (3.3), so is $\mathcal{E}B$ and $l(\mathcal{E}B;\mathcal{G})=l(B;\mathcal{G})$. In what follows, we will say that a matrix M is essentially unique when it is uniquely defined, up to a left multiplication by a sign matrix. Note that a row-permutation of B, which is a classical invariance property in blind source separation, corresponds here to the same permutation on the order of the matrices $\Gamma_1,\dots,\Gamma_K$ in $\mathcal{G}$.

3.2. The case $K=2$. In this situation, an explicit formula for the solution is obtained. A matrix B satisfying (3.3) realizes a joint diagonalization of Γ_1 and Γ_2. This is related to the eigendecomposition problem [9, p. 467].

LEMMA 3.5. *Let Γ_1 and Γ_2 be two symmetric matrices. Assume that Γ_1 is positive definite. Then there exists a matrix W satisfying*

$$W\Gamma_1W^t=I,\quad W\Gamma_2W^t=\Delta,$$

where $\Delta=\operatorname{diag}(\delta_1,\dots,\delta_K)$ is a diagonal matrix. Moreover, the matrix W is essentially unique if and only if the elements of Δ are distinct and arranged in decreasing (or increasing) order.

The rows of W are the eigenvectors of Γ_2, with respect to the inner product $\langle\cdot,\cdot\rangle_{\Gamma_1}$, associated with the eigenvalues given by Δ, i.e., the roots of the polynomial $\det(\Gamma_2-\delta\Gamma_1)$, since we have $\Gamma_2W^t=\Gamma_1W^t\Delta$.

PROPOSITION 3.6. *Let Γ_1 and Γ_2 be two nonproportional symmetric positive definite matrices of order 2. Then the solution of the max-det problem is essentially unique and given by*

$$B=\begin{pmatrix}1&0\\0&\frac{1}{\sqrt{\delta_2}}\end{pmatrix}W,\qquad \det B=\left(\frac{\delta_1}{\delta_2}\right)^{1/4}(\det\Gamma_1\times\det\Gamma_2)^{-1/4},$$

where δ_1,δ_2, and W are elements of the joint diagonalization

$$W\Gamma_1W^t=\begin{pmatrix}1&0\\0&1\end{pmatrix},\quad W\Gamma_2W^t=\begin{pmatrix}\delta_1&0\\0&\delta_2\end{pmatrix},\quad \delta_1>\delta_2>0.$$

Proof. Using Lemma 3.4 with $M=W$, we have $B=CW$, where C is the solution of the max-det problem in which $\Gamma_1=I$ and $\Gamma_2=\Delta$. Now, the solutions of (3.3) are essentially equal to

$$C=\begin{pmatrix}1&0\\0&\frac{1}{\sqrt{\delta_2}}\end{pmatrix},\qquad \tilde{C}=\begin{pmatrix}0&1\\\frac{1}{\sqrt{\delta_1}}&0\end{pmatrix}.$$

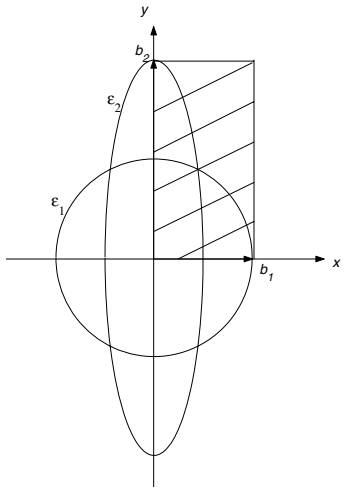


FIG. 3.1. *The max-det problem with $\Gamma_1 = I$ (ellipsoid ϵ_1) and $\Gamma_2 = \Delta$ (ellipsoid ϵ_2): The solution $B = [b_1, b_2]^t$ defines the parallelepiped (hatched) with maximum area when $b_1 \in \epsilon_1$ and $b_2 \in \epsilon_2$.*

Let $\delta = \det C$. The characteristic polynomials of the Hessian matrices are

$$P_C(\lambda) = \delta_2(\delta_2\lambda + 2)(\lambda + 2)(\delta_2^2\lambda^2 + (\delta_1 + 1)\delta_2\lambda + \delta_1 - \delta_2),$$
$$P_{\tilde{C}}(\lambda) = \delta_1(\delta_1\lambda + 2)(\lambda + 2)(\delta_1^2\lambda^2 + (\delta_2 + 1)\delta_1\lambda + \delta_2 - \delta_1).$$

It follows that $P_C(\lambda)$ has four negative roots, whereas $P_{\tilde{C}}(\lambda)$ has one positive and three negative roots. Hence $B = CW$ is the unique matrix, up to a sign factor, that realizes the maximum of (1.2). The equalities

$$\det B = \delta_2^{-1/2} \det W = \delta_2^{-1/2} (\det \Gamma_1)^{-1/2} = \delta_1^{1/2} (\det \Gamma_2)^{-1/2}$$

give the expression of $\det B$. □

Notice that the other solution of (3.3), $\tilde{B} = \tilde{C}W$, corresponds to a saddle point. When the matrices are proportional, $\Gamma_2 = \delta\Gamma_1$, we have $\delta_1 = \delta_2 = \delta$, and the solution is

$$B = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{\sqrt{\delta}} \end{pmatrix} \Gamma_1^{-1/2},$$

where $\Gamma_1^{-1/2}$ is the inverse matrix of any square root of Γ_1.

The solution, for $\Gamma_1 = I$ and $\Gamma_2 = \Delta$, is illustrated by Figure 3.1.

3.3. The case of jointly diagonalizable matrices. We suppose here that the set of matrices $\mathcal{G}$ satisfies $\Gamma_k = A\Delta_k A^t, k = 1, \ldots, K$, with $\Delta_k = \mathrm{diag}(\delta_k(1), \ldots, \delta_k(K))$ a (positive definite) diagonal matrix. This situation corresponds to the asymptotic conditions in the blind source separation problem. We have $X_T = AS_T$ and

$$\Gamma_k = A\frac{1}{T} S_T [R_k^T(\beta_k)]^{-1} S_T^t A^t, \quad k = 1, \ldots, K.$$

Thus, as T goes to infinity, the matrix $\frac{1}{T} S_T [R_k^T(\beta_k)]^{-1} S_T^t$ goes to a diagonal matrix Δ_k because the sources are independent. Moreover, Δ_k satisfies $1 = \delta_k(k) < \delta_k(j), j \neq k$ [8]. This leads to the following result.

PROPOSITION 3.7. *Let $\Gamma_k = A\Delta_k A^t, k = 1, \ldots, K$, be a set of $K \times K$ symmetric positive definite matrices which are jointly diagonalizable. Suppose that each diagonal matrix $\Delta_k = \operatorname{diag}(\delta_k(1), \ldots, \delta_k(K))$ satisfies $1 = \delta_k(k) < \delta_k(j), j \neq k$. Then $B = A^{-1}$ is the essentially unique solution of the max-det problem.*

Proof. Using Lemma 3.4 with $M = A^{-1}$, the solution is $B = UA^{-1}$, where $U = [u_1, \ldots, u_K]^t$ is the solution of the max-det problem associated with the diagonal matrices $\Delta_k, k = 1, \ldots, K$. We have

$$|\det U| \leq \prod_{k=1}^{K} \|u_k\|,$$

where $\|u\|^2 = u^t u$ is the usual Euclidean norm, with equality if and only if $u_k^t u_j = 0$ for $j \neq k$. Now, both constraint $u_k^t \Delta_k u_k \leq 1$ and hypothesis $1 = \delta_k(k) < \delta_k(j), j \neq k$, lead to $\|u_k\| \leq 1$ with equality if and only if $u_k = \pm e_k$. Because U is a sign matrix, the solution $B = A^{-1}$ is essentially unique. ∎

This proposition can be extended as follows.

PROPOSITION 3.8. *Let $\Gamma_k = A\Delta_k A^t, k = 1, \ldots, K$, be a set of $K \times K$ symmetric positive definite matrices which are jointly diagonalizable. The solution B of the max-det problem satisfies*

$$|\det B| \leq |\det A|^{-1} \prod_{k=1}^{K} \delta_k^{-1/2},$$

where $\Delta_k = \operatorname{diag}(\delta_k(1), \ldots, \delta_k(K))$ and $\delta_k = \min_j \delta_k(j), k = 1, \ldots, K$. This upper bound is achieved if and only if there exists a permutation $\{j_1, \ldots, j_K\}$ of $\{1, \ldots, K\}$ such that $\delta_k(j_k) = \delta_k, k = 1, \ldots, K$, and the solution is essentially unique if and only if such a permutation is unique.

Proof. By Lemma 3.4 we can restrict ourself to the case of diagonal matrices. Using the notation and the arguments of the proof above, we have

$$|\det U| \leq \prod_{k=1}^{K} \|u_k\| \leq \prod_{k=1}^{K} \delta_k^{-1/2},$$

and the upper bound is achieved if and only if $U = \Delta^{-1/2} V$, with $\Delta = \operatorname{diag}(\delta_1, \ldots, \delta_K)$ and $V = [v_1, \ldots, v_K]^t$, where v_k is an eigenvector of Δ_k associated with the eigenvalue δ_k and $VV^t = I$. If there exists a permutation $\{j_1, \ldots, j_K\}$ of $\{1, \ldots, K\}$ such that $j_k \in J_k = \{j : \delta_k(j) = \delta_k\}, k = 1, \ldots, K$, then the upper bound is realized by $v_k = \pm e_{j_k}, k = 1, \ldots, K$, and V is essentially a permutation matrix. Suppose now that the upper bound is achieved. Then $V^{\cdot 2} = (v_{ij}^2)$, where $v_{ij}^2 = v_{ij}^2$, is doubly stochastic. So there is a permutation matrix $V^* = (v_{ij}^*)$ such that $v_{ij}^* = 0$ whenever $v_{ij}^2 = 0$ [6, Theorem 20.3, p. 329], and the entries of V^* define the desired permutation. Finally, since any doubly stochastic matrix is a convex combination of permutation matrices [6, Theorem 20.4, p. 330], the solution is essentially unique if and only if such a permutation is unique. ∎

In all other cases, no general result has been obtained. We now give some comments and examples, for diagonal matrices, in order to show the difficulties of the problem.

For any permutation $\{j_1, \ldots, j_K\}$ of $\{1, \ldots, K\}$, the matrix

$$U^t = \left[\frac{e_{j_1}}{\sqrt{\delta_1(j_1)}}, \ldots, \frac{e_{j_K}}{\sqrt{\delta_K(j_K)}} \right], \qquad \det U = \left[\prod_{k=1}^{K} \delta_k(j_k) \right]^{-1/2} \tag{3.7}$$

is a solution of the following system coming from (3.3):

$$U\Delta_k U^t e_k = e_k, \quad k = 1, \dots, K. \tag{3.8}$$

Example 4 below shows that other solutions can exist. From Proposition 3.3, a solution of (3.8) corresponds to a maximum if and only if

$$\frac{\delta_k(j_l)}{\delta_l(j_l)} > \frac{\delta_k(j_k)}{\delta_l(j_k)}, \quad k < l = 1, \dots, K,$$

but it is not proved that the solution of the max-det problem is in this set. Notice that the system (3.8) has a continuum of solutions if, for some $k \neq l$, there exists $i \neq j$ such that $\delta_k(i)\delta_l(j) = \delta_l(i)\delta_k(j)$.

Let us consider now some examples illustrating different situations.

1. For $K = 3, \Delta_1 = \mathrm{diag}(1, 2, 1), \Delta_2 = \mathrm{diag}(1, 1, 2)$, and $\Delta_3 = \mathrm{diag}(2, 1, 1)$, the maximum is equal to 1 and given by the matrices essentially equal to I or to
$$\Pi = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$
2. For $K = 3, \Delta_1 = \mathrm{diag}(1, 1, 1), \Delta_2 = \mathrm{diag}(1, 1, 2)$, and $\Delta_3 = \mathrm{diag}(2, 3, 1)$, the maximum is also equal to 1, but is realized by a continuum of solutions given by the matrices essentially equal to
$$U = \begin{pmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad -\frac{\pi}{2} < \theta \leq \frac{\pi}{2}.$$
3. For $K = 3, \Delta_1 = \mathrm{diag}(1, 1, 1), \Delta_2 = \mathrm{diag}(1, 3, 2)$, and $\Delta_3 = \mathrm{diag}(1, 3, 4)$, the maximum is equal to $1/\sqrt{2}$ and given by the matrices essentially equal to
$$U = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1/\sqrt{2} \\ 1 & 0 & 0 \end{pmatrix}.$$
However, saddle points of the criterion are given by the following matrices:
$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/\sqrt{3} & 0 \\ 0 & 0 & 1/2 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 & 1 \\ \sin\theta & \cos\theta/\sqrt{3} & 0 \\ \cos\theta & -\sin\theta/\sqrt{3} & 0 \end{pmatrix}, \quad -\frac{\pi}{2} < \theta \leq \frac{\pi}{2},$$
and we have observed that our algorithm can converge to such points.
4. For $K = 3, \Delta_1 = \mathrm{diag}(1, 1, 1), \Delta_2 = \mathrm{diag}(3, 9, 11)/12$, and $\Delta_3 = \mathrm{diag}(7, 5, 15)/20$, the maximum is equal to 4 and essentially uniquely realized by
$$U = \begin{pmatrix} 0 & 0 & 1 \\ 2 & 0 & 0 \\ 0 & 2 & 0 \end{pmatrix}.$$
Five solutions of (3.8) are given by
$$U^t = [e_j/\sqrt{\delta_1(j)}, e_k/\sqrt{\delta_2(k)}, e_l/\sqrt{\delta_3(l)}],$$

where $(j,k,l) \neq (3,1,2)$ are the other permutations of $(1,2,3)$. However, in this case, the matrix

$$\begin{pmatrix} 1 & 2 & 1 \\ -1 & 1 & -3 \\ 5 & -1 & -1 \end{pmatrix} \begin{pmatrix} \sqrt{0.1} & 0 & 0 \\ 0 & \sqrt{0.2} & 0 \\ 0 & 0 & \sqrt{0.1} \end{pmatrix}$$

is also a solution of (3.8).

5. For $K = 3, \Delta_1 = \mathrm{diag}(1,2,5), \Delta_2 = \mathrm{diag}(5,1,2)$, and $\Delta_3 = \mathrm{diag}(2,5,1)$, the maximum is equal to 1 and given by the matrices essentially equal to I. The other solutions of (3.8), associated with the permutations of $(1,2,3)$, lead to saddle points except for $(2,3,1)$ for which the matrix

$$\begin{pmatrix} 0 & 1/\sqrt{2} & 0 \\ 0 & 0 & 1/\sqrt{2} \\ 1/\sqrt{2} & 0 & 0 \end{pmatrix}$$

provides a local maximum of the criterion.

3.4. Existence of the solution. The existence of a solution of our max-det problem is obtained by considering the criterion $l(B;\mathcal{G})$ given in (1.2).

PROPOSITION 3.9. *The maximum of $l(B;\mathcal{G})$ is attained by a nonsingular matrix B satisfying the set of equations* (3.3).

Proof. Let λ_k be the smallest eigenvalue of Γ_k. The inequality

$$l(B;\mathcal{G}) \leq \sum_{k=1}^{K} \left[\log \|b_k\| - \frac{\lambda_k}{2} \|b_k\|^2 \right]$$

shows the coercivity of $l(B;\mathcal{G})$. Thus we can restrict ourself to a set of bounded matrices B. Now, in such a set, $l(B;\mathcal{G})$ goes to $-\infty$ when B goes to a singular matrix because $l(B;\mathcal{G}) < \log\det|B|$. Then the maximum is attained by a nonsingular matrix B, and this matrix satisfies (3.3). ∎

Notice that $b_k^t \Gamma_k b_k \leq 1$ leads to $\|b_k\| \leq \lambda_k^{-1}$. So $\prod_{k=1}^{K} \lambda_k^{-1}$ is an upper bound for $\det B$. This upper bound is attained if and only if $B_{ev} = [\lambda_1^{-1/2} v_1, \ldots, \lambda_K^{-1/2} v_K]^t$, where v_k is an eigenvector associated with λ_k, satisfies (3.3). For example, this will be the case for jointly diagonalizable matrices, $\Gamma_k = V \Lambda_k V^t, k = 1, \ldots, K$, with $V = [v_1, \ldots, v_k]^t, V^t V = I$, and λ_k being the smallest element of Λ_k.

From a geometrical point of view, it is natural to maximize $\det B$ in the formulation of our max-det problem. On the other hand, the maximum likelihood principle in the blind source separation problem leads to the presence of $|\det B|$ instead of $\det B$ in the formulation of the criterion $l(B;\mathcal{G})$. In fact, the two optimization problems are equivalent (have the same set of solutions) if we consider either $\det B$ or $|\det B|$ in the two formulations. We have already observed that $l(\mathcal{E}B;\mathcal{G}) = l(B;\mathcal{G})$ for any sign matrix $\mathcal{E}$. Furthermore, $\det \mathcal{E}B = \det \mathcal{E} \times \det B$ with $\det \mathcal{E} = \pm 1$. So the set of matrices that maximize the criterion $l(B;\mathcal{G})$, with $\det B$ instead of $|\det B|$, is the set of matrices that maximize $l(B;\mathcal{G})$ and satisfy $\det B > 0$. This set is also the set of solutions of our max-det problem. More precisely, we have the following result.

PROPOSITION 3.10. *The maximization of $\det B$ subject to the constraint* (1.1) *is equivalent to minimizing the criterion*

$$l^*(B;\mathcal{G}) = -\log\det B + \frac{1}{2} \sum_{k=1}^{K} (b_k^t \Gamma_k b_k - 1).$$

Proof. If B satisfies (1.1), $\mathcal{E}B$ also satisfies (1.1), for any sign matrix $\mathcal{E}$, so we can restrict ourself to the set of matrices B with $\det B > 0$. Then, our max-det problem is equivalent to the minimization of $-\log\det B$ subject to (1.1). Let us consider the Lagrange function associated with

$$L(B,\mu) = -\log\det B + \sum_{k=1}^{K} \mu_k (b_k^t \Gamma_k b_k - 1),$$

where $\mu = (\mu_1, \ldots, \mu_K)^t$ is the vector of Lagrange multipliers. Since $\det B$ is bounded under the constraint (1.1) and $-\log\det B$ goes to $+\infty$ when B tends to a singular matrix, the minimum of $-\log\det B$, subject to (1.1), exists and is attained by a nonsingular matrix B. For such a matrix, we have

$$\frac{\partial b_k^t \Gamma_k b_k}{\partial B} = 2\,[0, \ldots, 0, \Gamma_k b_k, 0, \ldots, 0]^t \neq 0, \quad k = 1, \ldots, K,$$

because $b_k \neq 0$ and Γ_k is positive definite. So, the linear independence constraint qualification holds. Then, there exists μ such that the solution B satisfies the KKT necessary conditions

$$\mu_k \geq 0, \quad \mu_k (b_k^t \Gamma_k b_k - 1) = 0, \quad b_k^t \Gamma_k b_k - 1 \leq 0, \quad k = 1, \ldots, K, \quad \frac{\partial L(B,\mu)}{\partial B} = 0.$$

We have

$$\frac{\partial L(B,\mu)}{\partial B} = 0 \quad \Longleftrightarrow \quad 2\mu_k B \Gamma_k b_k = e_k, \quad k = 1, \ldots, K. \tag{3.9}$$

Setting $\mathcal{S}_B = \{B : \det B > 0, b_k^t \Gamma_k b_k \leq 1, k = 1, \ldots, K\}$ and $\mathcal{S}_\mu = \{\mu : \mu_k \geq 0, k = 1, \ldots, K\}$, the min-max theorem gives

$$\min_{B \in \mathcal{S}_B} \max_{\mu \in \mathcal{S}_\mu} L(B,\mu) \geq \max_{\mu \in \mathcal{S}_\mu} \min_{B \in \mathcal{S}_B} L(B,\mu).$$

From

$$\min_{B \in \mathcal{S}_B} \max_{\mu \in \mathcal{S}_\mu} L(B,\mu) = \min_{B \in \mathcal{S}_B} -\log\det B, \qquad \max_{\mu \in \mathcal{S}_\mu} \min_{B \in \mathcal{S}_B} L(B,\mu) \geq \min_{B \in \mathcal{S}_B} L(B,\mu^*)$$

given by $\mu = 0$ and $\mu^* = (1/2, \ldots, 1/2)^t$, we obtain

$$\min_{B \in \mathcal{S}_B} -\log\det B \geq \min_{B \in \mathcal{S}_B} L(B,\mu^*) = \min_{B \in \mathcal{S}_B} l^*(B;\mathcal{G}) \geq \min_B l^*(B;\mathcal{G}). \tag{3.10}$$

Now, if $\hat{\mathcal{S}}_B$ is the set of matrices B satisfying $\det B > 0$ and (3.3), Proposition 3.9 leads to

$$\min_B l^*(B;\mathcal{G}) = \min_{B \in \hat{\mathcal{S}}_B} l^*(B;\mathcal{G}) = \min_{B \in \hat{\mathcal{S}}_B} -\log\det B \geq \min_{B \in \mathcal{S}_B} -\log\det B, \tag{3.11}$$

because $\hat{\mathcal{S}}_B \subset \mathcal{S}_B$. Relations (3.10) and (3.11) give the equality

$$\min_{B \in \mathcal{S}_B} -\log\det B = \min_B l^*(B;\mathcal{G}) \tag{3.12}$$

and show that

$$\left\{ \hat{B} : l^*(\hat{B};\mathcal{G}) = \min_B l^*(B;\mathcal{G}) \right\} \subseteq \left\{ \hat{B} : -\log\det \hat{B} = \min_{B \in \mathcal{S}_B} -\log\det B \right\}.$$

Conversely, if $\hat{B} \in \mathcal{S}_B$ minimizes $-\log\det B$, it is also a minimum for $l^*(B;\mathcal{G})$ since

$$-\log\det\hat{B} \geq l^*(\hat{B};\mathcal{G}) \geq \min_B l^*(B;\mathcal{G}) = -\log\det\hat{B},$$

and the proof is achieved. Notice that (3.3) corresponds to (3.9) with $\mu = \mu^*$. □

In the proof above, equality (3.12) shows that there is no duality gap and μ^* is a strong duality solution for our max-det problem although this problem is nonconvex.

3.5. The uniqueness problem. We have shown that the solution of our max-det problem is essentially unique for two nonproportional matrices (see Proposition 3.6) and for particular jointly diagonalizable matrices (see Propositions 3.7 and 3.8). Neither a general result for this problem nor a sufficient condition for the solutions of (3.3) to be a set of isolated points has been obtained. Indeed, this last condition guarantees the convergence of our algorithm. In section 3.3 we have seen, for jointly diagonalizable matrices, that the maximum can be reached by two distinct matrices (example 1) or by a continuum set of matrices (example 2). This comes from the particular structure of these matrices, and such nuisance disappears in the blind source separation problem, because $\mathcal{G} = \{\Gamma_k, k = 1,\ldots,K\}$ is a set of estimated matrices. However, the existence of saddle points (example 3) and local maxima (example 5) for the criterion will be probably true more often than not. This comes from the continuity of $l(B,\mathcal{G})$ with respect to $\mathcal{G}$. We suggest computing the upper bound $\prod_{k=1}^K \lambda_k^{-1}$ of $\det B$ and starting the algorithm with the matrix $B_{ev} = [\lambda_1^{-1/2}v_1,\ldots,\lambda_K^{-1/2}v_K]^t$ presented above. These points will be illustrated at the end of the next section.

4. The algorithm. The algorithm and its convergence properties are presented, and some numerical experiments are performed showing its behavior.

We maximize $l(B,\mathcal{G})$ by a relaxation technique on the rows $b_k, k = 1,\ldots,K$, of B. Such an iterative method guarantees that the criterion increases at each step. Then, the convergence of the algorithm can be obtained under reasonable conditions.

4.1. The relaxation technique. The gradient of $l(B,\mathcal{G})$ with respect to b_k vanishes if and only if $e_k = B\Gamma_k b_k$, i.e., $\|b_k\|_{\Gamma_k} = 1$ and $\langle b_k, b_j\rangle_{\Gamma_k} = 0$ for $j \neq k$. The Hessian

$$\nabla^2_{b_k} l(B;\mathcal{G}) = -a_k a_k^t - \Gamma_k, \qquad [a_1,\ldots,a_K] = A = B^{-1}$$

is negative definite. So the maximum of $l(B,\mathcal{G})$, with respect to b_k and when the other rows of B are fixed, has the following explicit solution. Let B be a nonsingular matrix of order K, and let $B_{(k)}$ denote the $(K-1)\times K$-matrix obtained by removing the kth row of B. Hence the orthogonal projector in $\mathbf{R}^K$, according to $\langle\cdot,\cdot\rangle_{\Gamma_k}$, on the subspace spanned by the rows of $B_{(k)}$ is

$$P_k = B_{(k)}^t \left[B_{(k)}\Gamma_k B_{(k)}^t\right]^{-1} B_{(k)}\Gamma_k.$$

Thus the maximum of $l(B,\mathcal{G})$ is realized by $\pm\hat{b}_k$ with

$$\hat{b}_k = \frac{(I-P_k)b_k}{\|(I-P_k)b_k\|_{\Gamma_k}}.$$

Our algorithm proceeds as follows. A nonsingular initial matrix $B(0)$ is chosen. Then the algorithm constructs a sequence $\{B(n), n \geq 0\}$ by modifying successively the rows of the current matrix using the above process. $B(n+1)$ differs from $B(n)$

only by one row; i.e., $b_k(n+1) = \pm\hat{b}_k(n)$ for some k and the sign is chosen in such a way that the first nonzero component of $b_k(n+1)$ is positive. By construction, each matrix $B(n)$ is nonsingular.

4.2. Convergence of the algorithm. We have seen that the criterion $l(B;\mathcal{G})$ is bounded and is strictly concave with respect to b_k. Hence $l(B(n+1);\mathcal{G}) \geq l(B(n);\mathcal{G})$, with equality if and only if $B(n+1) = B(n)$. However, this equality is not sufficient to prove that $B(n)$ satisfies (3.3); this will be true if and only if $B(n+K) = B(n)$. However, we have the following result.

PROPOSITION 4.1. *The sequence* $\{l(B(n);\mathcal{G}), n \geq 0\}$ *is convergent.*

This does not prove the convergence of the sequence $\{B(n), n \geq 0\}$, which requires many intermediate results.

LEMMA 4.2. *The projection* $P_k(n)b_k(n)$*, computed at each step of the algorithm, converges to* 0 *when* n *goes to* $+\infty$*.*

Proof. Assume, without any loss of generality, that the $(n+1)$th step concerns the modification of the first row of $B(n)$. On the one hand, we have, for $n \geq K$,

$$l(B(n+1);\mathcal{G}) - l(B(n);\mathcal{G}) = \log|\det B(n+1)A(n)|,$$

where $A(n) = [a_1(n), \ldots, a_K(n)] = B(n)^{-1}$. On the other hand, we have

$$B(n+1)A(n) = \begin{pmatrix} \langle b_1(n+1), a_1(n)\rangle & \ldots & \ldots & \langle b_1(n+1), a_K(n)\rangle \\ 0 & 1 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & \ldots & 1 \end{pmatrix}.$$

Proposition 4.1 shows that the difference $l(B(n+1);\mathcal{G}) - l(B(n);\mathcal{G})$ goes to 0; therefore the inner product $\langle b_1(n+1), a_1(n)\rangle$ goes to ± 1. By construction we have

$$b_1(n+1) = \pm\frac{b_1(n) - P_1(n)b_1(n)}{\|b_1(n) - P_1(n)b_1(n)\|_{\Gamma_1}},$$

and then

$$\begin{aligned}\langle b_1(n+1), a_1(n)\rangle &= \pm\frac{\langle b_1(n), a_1(n)\rangle - \langle P_1(n)b_1(n), a_1(n)\rangle}{\|b_1(n) - P_1(n)b_1(n)\|_{\Gamma_1}} \\ &= \pm\frac{1}{\|b_1(n) - P_1(n)b_1(n)\|_{\Gamma_1}}.\end{aligned}$$

Therefore, $\|b_1(n) - P_1(n)b_1(n)\|_{\Gamma_1}$ converges to 1 when n goes to $+\infty$. Since

$$\begin{aligned}\|P_1(n)b_1(n)\|^2_{\Gamma_1} &= \|b_1(n)\|^2_{\Gamma_1} - \|b_1(n) - P_1(n)b_1(n)\|^2_{\Gamma_1} \\ &= 1 - \|b_1(n) - P_1(n)b_1(n)\|^2_{\Gamma_1},\end{aligned}$$

one deduces that $\|P_1(n)b_1(n)\|^2_{\Gamma_1}$ converges to 0 when n goes to $+\infty$. □

Now we show that, asymptotically, $B(n)$ satisfies the optimality condition (3.3).

LEMMA 4.3. *The sequence* $\{B(n), n \geq 0\}$ *satisfies*

$$\lim_{n\to+\infty} B(n)\Gamma_k B(n)^t e_k = e_k, \quad k = 1, \ldots, K.$$

Proof. We have to show that, for all $k, j = 1, \ldots, K$,

$$\lim_{n\to+\infty} \langle b_k(n), b_j(n)\rangle_{\Gamma_k} = \delta_{kj}.$$

Let $b_i(n)$ be the row modified at the $(n+1)$th step,

$$b_i(n+1) = \frac{b_i(n) - P_i(n)b_i(n)}{\|b_i(n) - P_i(n)b_i(n)\|_{\Gamma_i}}.$$

For $k = i$, we have $\langle b_i(n+1), b_j(n+1)\rangle_{\Gamma_i} = \delta_{ij}$ by construction. For the same reason, we have $\langle b_j(n), b_i(n)\rangle_{\Gamma_j} = 0$ for $j \neq i$ because of the $(K-1)$ preceding steps (provided that $n \geq K-1$). Hence one obtains

$$\begin{aligned}\langle b_j(n+1), b_i(n+1)\rangle_{\Gamma_j} &= \langle b_j(n), b_i(n+1)\rangle_{\Gamma_j} \\ &= \frac{\langle b_j(n), b_i(n) - P_i(n)b_i(n)\rangle_{\Gamma_j}}{\|b_i(n) - P_i(n)b_i(n)\|_{\Gamma_j}} = -\frac{\langle b_j(n), P_i(n)b_i(n)\rangle_{\Gamma_j}}{\|b_i(n) - P_i(n)b_i(n)\|_{\Gamma_j}}.\end{aligned}$$

Since $\|b_j(n)\|_{\Gamma_j} = 1$ for $j = 1, \ldots, K$ and $P_i(n)b_i(n)$ goes to 0 (according to the Lemma 4.2), one deduces that $\langle b_j(n+1), b_i(n+1)\rangle_{\Gamma_j}$ converges to 0 when n goes to $+\infty$. When both j and k are not equal to i, $\langle b_j(n+1), b_k(n+1)\rangle_{\Gamma_j} = 0$ if the jth row was modified after the kth one, which corresponds to the following situations: $i < k < j$, $k < j < i$, and $j < i < k$. Otherwise, we use the same argument as above from the following equalities. We have

$$\langle b_j(n+1), b_k(n+1)\rangle_{\Gamma_j} = -\frac{\langle b_j(l), P_k(l)b_k(l)\rangle_{\Gamma_j}}{\|b_k(l) - P_k(l)b_k(l)\|_{\Gamma_k}}, \quad l = n - i + k + 1,$$

for $j < k < i$ or $k < i < j$, and

$$\langle b_j(n+1), b_k(n+1)\rangle_{\Gamma_j} = -\frac{\langle b_j(l), P_k(l)b_k(l)\rangle_{\Gamma_j}}{\|b_k(l) - P_k(l)b_k(l)\|_{\Gamma_k}}, \quad l = n - i + k - K + 1,$$

for $i < j < k$. □

LEMMA 4.4. *The sequence* $\{B^{-1}(n), n \geq 0\}$ *is bounded.*

Proof. We recall the notation $A(n) = [a_1(n), \ldots, a_K(n)] = B^{-1}(n)$ and take the usual norm, $\|A\|^2 = \text{trace}(AA^t)$. Let λ_k (resp., μ_k) be the smallest (resp., the largest) eigenvalue of the matrix Γ_k. Since $n \geq K-1$, we have

$$\|A(n)\|^2 = \sum_{k=1}^{K} \|a_k(n)\|^2 \leq \sum_{k=1}^{K} \frac{\mu_k^2}{\lambda_k}.$$

Indeed, assume that $b_k(n)$ is the row defined at the nth step. By the construction process, $b_k(n)$ is characterized by

$$B(n)\Gamma_k b_k(n) = e_k \quad \Longleftrightarrow \quad a_k(n) = \Gamma_k b_k(n).$$

The second expression yields to the inequality $\|a_k(n)\| \leq \mu_k \|b_k(n)\|$. Moreover, $\|b_k(n)\|_{\Gamma_k}^2 = 1$ implies that $\|b_k(n)\|^2 \leq \frac{1}{\lambda_k}$. □

LEMMA 4.5. *The gradient of* $l(B; \mathcal{G})$*, evaluated at* $B(n)$*, converges to* 0 *when* n *goes to* $+\infty$*, and any accumulation point of the sequence* $\{B(n), n \geq 0\}$ *satisfies the optimality condition* (3.3).

Proof. From Proposition 3.1, we have

$$\nabla_{b_k} l(B(n); \mathcal{G}) = a_k(n) - \Gamma_k B^t(n) e_k = A(n)(e_k - B(n)\Gamma_k B^t(n) e_k), \quad k = 1, \ldots, K.$$

The matrix $A(n)$ is bounded and, according to the Lemma 4.3,

$$\lim_{n \to +\infty} B(n)\Gamma_k B^t(n) e_k = e_k \quad \forall\, k = 1, \ldots, K;$$

TABLE 4.1

Behavior of the algorithm in three situations: random (6 cases), saddle points, and local maxima. Variables are computed, in each row, with $N = 1000$ choices of matrices Γ_k and B_{ev} or $R = 100$ random matrices as starting value. See the text for further explanation.

Matrices Γ_k	η	div ev	div r	swe ev	swe r	nonun	$ev \neq r$
Random		0	0.4	11.3	12.2	3	1
		1	0.9	10.5	11.5	1	0
		2	2.3	10.2	11.1	0	0
		3	2.0	11.1	12.3	1	0
		4	2.1	10.1	11.1	2	0
		5	4.1	10.8	11.8	4	1
Saddle points	10^{-3}	0	133.0	44.5	50.2	999	0
	10^{-2}	0	52.6	45.0	55.4	376	0
	10^{-1}	32	38.7	58.2	63.9	111	23
	1	7	6.7	31.3	33.6	29	6
	10	2	2.4	15.0	16.2	11	1
	100	1	1.0	10.6	11.4	4	1
Local maxima	10^{-3}	0	0.0	6.6	11.4	523	0
	10^{-2}	0	0.4	7.7	11.4	137	0
	10^{-1}	0	0.1	10.2	13.4	11	0
	1	2	3.7	25.6	27.7	23	8
	10	2	2.6	15.5	16.6	6	0
	100	0	0	10.8	11.7	5	1

then the first part of the lemma is achieved. Let B^* be an accumulation point of the sequence $\{B(n), n \geq 0\}$. Then there exists a increasing function φ such that the sequence $\{B(\varphi(n)), n \geq 0\}$ converges to B^*. This subsequence satisfies the limit property of Lemma 4.3, and this, using the continuity of the inner product, achieves the second part of the lemma. □

LEMMA 4.6. $\lim_{n\to+\infty} \|B(n+1) - B(n)\| = 0$.

Proof. Assume that $B(n+1)$ differs from $B(n)$ by its kth row; then we have

$$\|B(n+1) - B(n)\| = \|b_k(n+1) - b_k(n)\|,$$
$$b_k(n+1) - b_k(n) = \frac{b_k(n)(1 - \|b_k(n) - P_k(n)b_k(n)\|_{\Gamma_k}) - P_k(n)b_k(n)}{\|b_k(n) - P_k(n)b_k(n)\|_{\Gamma_k}}.$$

Thus $b_k(n+1) - b_k(n)$ converges to 0 when n goes to $+\infty$ since $P_k(n)b_k(n)$ converges to 0 and $\|b_k(n)\|_{\Gamma_k} = 1$. □

Both the lemma and a theorem in Ostrowski [13, p. 173] lead to the following result.

THEOREM 4.7. *The sequence $\{B(n), n \geq 0\}$ given by the algorithm converges; otherwise, the set of solutions of the optimality condition* (3.3) *is a continuum.*

4.3. Numerical experiments. Two sets of experiments are conducted. The former illustrates the behavior of our algorithm, and the latter compares our approach with Newton's methods for nonlinear optimization.

Table 4.1 summarizes results on the behavior of the algorithm. For $K = 3$, three kinds of situations are considered: random, saddle points and local maxima. In the first, the set $\mathcal{G}$ is random; i.e., the matrices $\Gamma_k, k = 1, 2, 3$, are mutually independent with $\Gamma_k = R_k R_k^t$, where R_k is a random matrix whose elements are independent zero-mean Gaussian variables with variance one. In the two other situations, we use

diagonal matrices with random perturbations

$$\Gamma_k = \Delta_k + \eta R_k R_k^t, \quad k = 1, 2, 3,$$

where R_k is as above and η is a scalar factor. The diagonal matrices Δ_k are those introduced in section 3.3 illustrating saddle points (example 3) and local maxima (example 5). Each row of the table corresponds to $N = 1000$ experiments associated with N independent choices of $\mathcal{G}$. For each experiment, our algorithm is used starting with $B(0)$ equal to the matrix B_{ev} using the eigenvectors (see section 3.5) and also to $R = 100$ random matrices $B_r, r = 1, \ldots, R$ (B_r is distributed like R_k above). The algorithm is stopped when the two following conditions are satisfied:

- $abs(l(B(mK + K); \mathcal{G}) - l(B(mK); \mathcal{G})) < 10^{-8}$,
- $\max_{i,j}\{|B_{ij}(mK + K) - B_{ij}(mK)|\} < 10^{-8}$,

or when the number of sweeps reaches 200 (a sweep corresponds to the change of $B(mK)$ into $B(mK + K)$). We use 10^{-8}, instead of 10^{-4} as below, in order to well separate, in each experiment, the distinct values of $l(\hat{B}; \mathcal{G})$ given by distinct values of $B(0)$, for the computation of "nonun" and "$ev \neq r$." Noting $\hat{B}_{ev}$ and $\hat{B}_r$ as the solutions of the algorithm associated with $B(0) = B_{ev}$ and $B(0) = B_r$, variables "nonun" and "$ev \neq r$" in Table 4.1 give, respectively, for the N experiments, the number of cases where $\max_r(l(\hat{B}_r; \mathcal{G})) - \min_r(l(\hat{B}_r; \mathcal{G})) > 10^{-4}$ and $|\max_r(l(\hat{B}_r; \mathcal{G})) - l(\hat{B}_{ev}; \mathcal{G})| > 10^{-4}$. So "nonun" represents the number of experiments for which the solution given by the algorithm is not unique, since it depends on the starting value $B(0)$; "$ev \neq r$" measures the ability of the algorithm to reach the maximum, when the starting value is B_{ev}. Variable "div ev" is the number of divergences of the algorithm (when the number of sweeps reaches 200) when $B(0) = B_{ev}$, and "div r" is the corresponding mean number using R starting values $B(0) = B_r$. Variables "swe ev" and "swe r" are the mean number of sweeps for convergence of the algorithm starting with $B(0) = B_{ev}$ (N trials) and $B(0) = B_r$ ($N \times R$ trials), respectively, computed only on the convergence cases.

Looking at $ev \neq r$ and div ev, it is clear that the algorithm must be initialized with $B(0) = B_{ev}$. In that way, saddle points or local maxima seem to be avoided. The random cases reported here have been obtained with numerous tries in order to select cases with increasing values of div ev. The corresponding values of div r and the relatively stability of swe ev and swe r lead to the following conclusions. Generally, the algorithm converges quickly, except for some "ill-conditioned" sets $\mathcal{G}$ for which the convergence is very slow, whatever the starting value. "nonun" shows that the uniqueness problem is not so crucial in standard situations. "div r" illustrates Theorem 4.7, since the set of solutions of the optimality condition (3.3) is a continuum in the saddle points situation, while these solutions are isolated points, including saddle points, in the local maxima situation.

We compare now our relaxation method with Newton's methods for nonlinear optimization. First, we use the Newton approach described in (3.6). The scale factor λ_n is provided by the quadratic and cubic line search procedure given in section 9.7 of [12]. Furthermore, when the Hessian $H = \nabla^2_{B(n)}$ is not negative definite, we replace it by $H - (\lambda_{max} + 10^{-4})I$, where λ_{max} is the largest eigenvalue of H, in order to guarantee that the Newton direction dB is an ascent direction. Table 4.2 gives results of a set of variables for $K = 3, 5, 8, 10, 12, 15, 20$, and 25. Each row of the table corresponds to $N = 1000$ experiments associated with N independent choices of $\mathcal{G}$ in the random situation. For each experiment, our algorithm, like the Newton method, is used starting with $B(0)$ equal to the matrix B_{ev} using the eigenvectors.

TABLE 4.2

Comparison between the relaxation method (r) and Newton's method (n), using the line search procedure with modification of the Hessian, for eight dimensions K. Variables are computed, in each row, with $N = 1000$ choices of matrices Γ_k and B_{ev} as starting value.

Dim	div	$r-n$	swe r	swe n	cpu r	cpu n	cpu hes	cpu ls	hes	ls
$K=3$	2 - 0	0 - 1	6.1	5.0	0.0088	0.0095	0.0018	0.0023	241	265
$K=5$	1 - 0	1 - 6	8.7	6.6	0.0169	0.0166	0.0060	0.0030	457	489
$K=8$	2 - 0	5 - 10	10.9	7.9	0.0271	0.0710	0.0488	0.0050	661	696
$K=10$	0 - 0	9 - 13	13.3	8.8	0.0386	0.1900	0.1543	0.0063	777	808
$K=12$	1 - 0	14 - 15	14.1	9.3	0.0501	0.4848	0.4225	0.0080	836	860
$K=15$	3 - 0	16 - 22	16.9	10.2	0.0847	1.8966	1.7376	0.0114	914	933
$K=20$	1 - 0	25 - 32	20.4	11.0	0.1498	11.1208	10.6141	0.0150	973	979
$K=25$	3 - 0	31 - 29	22.2	11.8	0.2721	53.3996	51.9327	0.0260	986	987

The stopping rule is the same as above, but with 10^{-4} instead of 10^{-8} and $20K$ instead of 200 for the maximum number of sweeps. Variable "div" gives, for the N experiments, the number of divergences of the two methods, in the order $r-n$. Noting by $\hat{B}_r$ and $\hat{B}_n$ the solutions given by our relaxation method and by Newton's method, variable "$r-n$" represents, for the N experiments, the number of cases where $l(\hat{B}_r;\mathcal{G}) > l(\hat{B}_n;\mathcal{G}) + 2\times 10^{-4}$ and $l(\hat{B}_n;\mathcal{G}) > l(\hat{B}_r;\mathcal{G}) + 2\times 10^{-4}$. Variables "swe r" and "swe n" are the means of sweeps of our algorithm and of Newton's method, computed only on the convergence cases. Variables "cpu r," "cpu n," "cpu hes," and "cpu ls" are the means of cpu times in seconds, computed only on the experiments for which the algorithms converge, of the relaxation algorithm, the Newton's method, the modification of the Hessian, and the line search procedure, respectively. Finally, variables "hes" and "ls" give, for Newton's method, the number of cases where the Hessian has positive eigenvalues and those where the line search procedure is necessary (at least one time during the sweeps of an experiment).

Table 4.2 shows that the performances of the two methods are similar (see variable "$r-n$"). However, our relaxation method is faster (cpu time) than Newton's method, although the mean number of sweeps of Newton's method is lower than that of the relaxation method. Notice that the gain in cpu time increases very quickly with K: from 1 at $K = 5$ to 196 for $K = 25$. As it can be seen from variable "hes," the Hessian frequently has positive eigenvalues, but Newton's method always converges, and we have observed than the Hessian is always negative definite at convergence. An important difference between the two methods is that, in the relaxation technique, the successive values $B(n)$ satisfy the constraint $b_k^t \Gamma_k b_k = 1, k = 1,\ldots,K$, while this constraint is approximatively satisfied only by the solution $\hat{B}_n$ given by the Newton's method. This can explain that "swe r" is greater than "swe n," since the search for the solution is confined on this constrained set in the relaxation technique. Variable "cpu hes" shows that the difference in cpu time between the two methods comes essentially from the computation of the eigenvalues of the Hessian. So, we propose to use a negative definite approximation of $\nabla^2_{B(n)} = E + \Gamma_{B(n)}$ by canceling the off-diagonal blocs of E. The corresponding results are reported in Table 4.3. We observe that the performances of the two methods are still similar (variable "$r-n$"). But this decreases the convergence properties of Newton's method (variables "div" and "swe n"), and the relaxation method remains faster (cpu time) than Newton's method. As suggested by one referee, we also compare our method with Newton's method using the trust region approach. For that, we use the FMINUNC procedure of Matlab 7.0.1, with gradient and Hessian coming from Proposition 3.1. Our stopping rule is

TABLE 4.3

Comparison between the relaxation method (r) and Newton's method (n), using the line search procedure with approximation of the Hessian, for eight dimensions K. Variables are computed, in each row, with $N = 1000$ choices of matrices Γ_k and B_{ev} as starting value.

Dim	div	$r - n$	swe r	swe n	cpu r	cpu n	cpu ls	ls
$K = 3$	2 - 15	3 - 0	5.8	9.7	0.0097	0.0131	0.0034	88
$K = 5$	1 - 13	1 - 2	8.5	16.3	0.0157	0.0196	0.0047	213
$K = 8$	1 - 10	6 - 3	11.3	24.3	0.0268	0.0527	0.0073	347
$K = 10$	0 - 10	3 - 4	12.8	27.7	0.0347	0.0916	0.0076	376
$K = 12$	0 - 10	8 - 5	14.7	33.5	0.0494	0.1835	0.0129	479
$K = 15$	1 - 10	11 - 6	16.9	39.6	0.0758	0.5262	0.0197	508
$K = 20$	2 - 7	16 - 13	20.5	50.6	0.1620	2.2173	0.0320	612
$K = 25$	3 - 8	14 - 19	23.8	55.1	0.2902	6.4221	0.0548	696

TABLE 4.4

Comparison between the relaxation method (r) and Newton's method (n), using the trust region approach, for eight dimensions K. Variables are computed, in each row, with $N = 1000$ choices of matrices Γ_k and B_{ev} as starting value.

Dim	div	$r - n$	swe r	swe n	cpu r	cpu n
$K = 3$	0 - 0	33 - 3	3.6	4.3	0.0051	0.0320
$K = 5$	0 - 0	27 - 5	4.7	5.5	0.0111	0.0461
$K = 8$	0 - 0	16 - 13	5.7	6.7	0.0194	0.0916
$K = 10$	0 - 0	19 - 17	6.4	7.4	0.0243	0.1401
$K = 12$	0 - 0	22 - 23	7.3	8.0	0.0338	0.2106
$K = 15$	0 - 0	20 - 24	8.1	8.7	0.0512	0.4572
$K = 20$	0 - 0	28 - 35	9.4	9.6	0.1033	1.0445
$K = 25$	0 - 0	46 - 53	10.3	10.2	0.1790	3.6607

TABLE 4.5

Behavior of the relaxation method for large dimensions K. Variables are computed, in each row, with $N = 100$ choices of matrices Γ_k and B_{ev} as starting value.

Dim	div r	swe r	cpu r
$K = 25$	0	24.2	0.2912
$K = 30$	0	28.3	0.6337
$K = 50$	0	35.6	5.5955
$K = 75$	2	47.9	23.8969
$K = 100$	3	85.7	109.1694

modified in order to agree with that of this procedure. Thus, our algorithm is stopped when either $abs(l(B(mK+K);\mathcal{G}) - l(B(mK);\mathcal{G})) < 10^{-4}$ or $\max_{i,j}\{|B_{ij}(mK+K) - B_{ij}(mK)|\} < 10^{-4}$, or when the number of sweeps reaches $20K$. This corresponds to $TolFun = 10^{-4}, TolX = 10^{-4}$, and $MaxIter = 20K$ in the options of the FMINUNC procedure. As can be seen in Table 4.4, the performances of the two methods remain similar, and the relaxation method is still faster than Newton's method. The gain in cpu time is less important for large values of K: it increases from 4 for $K = 5$ to 20 for $K = 25$ but is equal to 6 for $K = 3$.

In fact, these methods are not adapted for very large values of K, because the convergence begins slow. This can be seen in Table 4.5 for the relaxation method. In blind sources separation area, small values of K are frequently encountered, and some applications are restricted to $K = 2$.

Acknowledgments. The authors would like to thank the two reviewers for very useful comments and suggestions for clarifying the presentation of the paper.

REFERENCES

[1] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, *A blind source separation technique using second-order statistics*, IEEE Trans. Signal Process., 45 (1997), pp. 434–444.

[2] M. Brookes, *The Matrix Reference Manual*, online at http://www.ee.ic.ac.uk/hp/staff/dmb/matrix/intro.html, 2005.

[3] J.-F. Cardoso and A. Souloumiac, *Blind beamforming for non-Gaussian signals*, in IEE-Proceedings-F, 140 (1993), pp. 362–370.

[4] J.-F. Cardoso and B. H. Laheld, *Equivariant adaptive source separation*, IEEE Trans. Signal Process., 44 (1996), pp. 3017–3030.

[5] J.-F. Cardoso and A. Souloumiac, *Jacobi angles for simultaneous diagonalization*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 161–164.

[6] V. Chvátal, *Linear Programming*, W. H. Freeman, New York/San Francisco, 1983.

[7] P. Comon, *Independent component analysis. A new concept?*, Signal Process., 36 (1994), pp. 287–314.

[8] S. Dégerine and A. Zaïdi, *Separation of an instantaneous mixture of Gaussian autoregressive sources by the exact maximum likelihood approach*, IEEE Trans. Signal Process. 52 (2004), pp. 1499–1512.

[9] G.-H. Golub and F.-V. Loan, *Matrix Computation*, The Johns Hopkins University Press, Baltimore, MD, 1989.

[10] D. T. Pham, *Joint approximate diagonalization of positive definite Hermitian matrices*, SIAM J. Matrix Anal. Appl., 22 (2001), pp. 1136–1152.

[11] D. T. Pham, *Blind separation of instantaneous mixtures of sources via the Gaussian mutual information criterion*, Signal Process., 81 (2001), pp. 850–870.

[12] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing*, 2nd ed., Cambridge University Press, New York, 1992.

[13] A.-M. Ostrowski, *Solution of Equations in Euclidean and Banach Spaces*, 3rd ed., Academic Press, New York, 1973.

[14] 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.

[15] A. Yeredor, *Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation*, IEEE Trans. Signal Process., 50 (2002), pp. 1545–1553.

SIAM J. OPTIM.
Vol. 17, No. 4, pp. 1015–1034
© 2006 Society for Industrial and Applied Mathematics

A PROXIMAL-PROJECTION BUNDLE METHOD FOR LAGRANGIAN RELAXATION, INCLUDING SEMIDEFINITE PROGRAMMING*

KRZYSZTOF C. KIWIEL†

Abstract. We give a proximal bundle method for minimizing a convex function f over a convex set C. It requires evaluating f and its subgradients with a fixed but possibly unknown accuracy $\epsilon > 0$. Each iteration involves solving an unconstrained proximal subproblem and projecting a certain point onto C. The method asymptotically finds points that are ϵ-optimal. In Lagrangian relaxation of convex programs, it allows for ϵ-accurate solutions of Lagrangian subproblems and finds ϵ-optimal primal solutions. For semidefinite programming problems, it extends the highly successful spectral bundle method to the case of inexact eigenvalue computations.

Key words. nondifferentiable optimization, convex programming, proximal bundle methods, Lagrangian relaxation, semidefinite programming

AMS subject classifications. 65K05, 90C25

DOI. 10.1137/050639284

1. Introduction.

We consider the convex constrained minimization problem

$$f_* := \inf\{\, f(u) : u \in C \,\}, \tag{1.1}$$

where C is a nonempty closed convex set in the Euclidean space $\mathbb{R}^n$ with inner product $\langle \cdot, \cdot \rangle$ and norm $|\cdot|$, and $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function. We assume that for a fixed *accuracy tolerance* $\epsilon_f \geq 0$, for each $u \in C$ we can find an *approximate value* f_u and an *approximate subgradient* g_u of f that produce the *approximate linearization* of f:

$$\bar{f}_u(\cdot) := f_u + \langle g_u, \cdot - u \rangle \leq f(\cdot) \quad \text{with} \quad \bar{f}_u(u) = f_u \geq f(u) - \epsilon_f. \tag{1.2}$$

Thus $f_u \in [f(u) - \epsilon_f, f(u)]$ estimates $f(u)$, while $g_u \in \partial_{\epsilon_f} f(u)$; i.e., g_u is a member of the ϵ_f-subdifferential $\partial_{\epsilon_f} f(u) := \{g : f(\cdot) \geq f(u) - \epsilon_f + \langle g, \cdot - u \rangle\}$ of f at u.

Our assumption is realistic in many applications. For instance, if f is a max-type function of the form

$$f(u) := \sup\{\, F_z(u) : z \in Z \,\}, \tag{1.3}$$

where each $F_z : \mathbb{R}^n \to \mathbb{R}$ is convex and Z is an infinite set, then it may be impossible to compute $f(u)$. However, if for some fixed (and possibly *unknown*) tolerance ϵ_f we can find an ϵ_f-maximizer of (1.3), i.e., an element $z_u \in Z$ satisfying $F_{z_u}(u) \geq f(u) - \epsilon_f$, then we may set $f_u := F_{z_u}(u)$ and take g_u as any subgradient of F_{z_u} at u to satisfy (1.2). An important special case arises in *Lagrangian relaxation* [HUL93, Chap. XII], [Lem01], where problem (1.1) with $C := \mathbb{R}^n_+$ is the Lagrangian dual of the primal problem

$$\sup\ \psi_0(z) \quad \text{s.t.} \quad \psi_i(z) \geq 0,\ i = 1\colon n,\ z \in Z, \tag{1.4}$$

*Received by the editors August 30, 2005; accepted for publication (in revised form) July 20, 2006; published electronically December 1, 2006.
http://www.siam.org/journals/siopt/17-4/63928.html
†Systems Research Institute, Polish Academy of Sciences, Newelska 6, 01–447 Warsaw, Poland (kiwiel@ibspan.waw.pl).

with $F_z(u) := \psi_0(z) + \langle u, \psi(z)\rangle$ for $\psi := (\psi_1, \ldots, \psi_n)$. Then, for each multiplier $u \geq 0$, we need only find $z_u \in Z$ such that $f_u := F_{z_u}(y) \geq f(u) - \epsilon_f$ in (1.3) to use $g_u := \psi(z_u)$. For instance, if (1.4) is a *semidefinite program* (SDP) with each ψ_i affine and Z the set of symmetric positive semidefinite matrices of order m with a bounded trace, then $f(u)$ is the maximum eigenvalue of a symmetric matrix $M(u)$ depending affinely on u [Tod01, sect. 6.3], and z_u can be found by computing an approximate eigenvector corresponding to the maximum eigenvalue of $M(u)$ via the Lanczos method [HeK02, HeR00, Nay06].

The recent paper [Kiw06b] extended the proximal bundle methods of [Kiw90] and [HUL93, sect. XV.3] to the inexact setting of (1.2) (see [Hin01, Kiw85, Kiw95, Mil01, Sol03] for earlier related developments, and [Kiw05] for numerical tests). Such methods at each iteration find a trial point that minimizes over C a polyhedral model of f built from accumulated linearizations, stabilized by a quadratic *prox* term centered at a point which is usually the best iterate found so far. Solving this subproblem can require much work for large n even when the set C is polyhedral, including the simplest case of $C = \mathbb{R}^n_+$ used in Lagrangian relaxation.

This paper extends the projection-proximal method of [Kiw99] to the case of inexact linearizations. For this method, we may regard (1.1) as an unconstrained problem $f_* = \inf f_C$ with the *essential objective*

$$f_C := f + i_C, \tag{1.5}$$

where i_C is the *indicator* function of C ($i_C(u) = 0$ if $u \in C$, ∞ otherwise). In its simplest form, the method generates the trial point in two steps. The first *proximal* step minimizes a polyhedral model $\check{f}$ of f, augmented with a quadratic proximal term and a linearization of i_C obtained at the previous iteration, to produce a linearization of $\check{f}$. The second *projection* step minimizes over C this linearization augmented with the proximal term; this amounts to projecting a certain point onto C to produce the trial point and the next linearization of i_C. Thus the standard bundle subproblem is replaced by two subproblems, where the first "unconstrained" subproblem is much easier to solve, and the projection is straightforward if the set C is "simple." Our development is related to the *alternating linearization* approach of [KRR99], in which the prox subproblem for the sum of two functions, such as (1.5), is approximated by two subproblems in which the functions are alternately represented by linear models.

Our extension of [Kiw99] is natural and simple: the original method is run as if the objective linearizations were exact until a test on predicted descent discovers their inaccuracy; then the proximity weight is decreased to produce descent or confirm that the current prox center is ϵ_f-optimal. We show that our method asymptotically estimates the optimal value f_* of (1.1) with accuracy ϵ_f and finds ϵ_f-optimal points. In Lagrangian relaxation, under standard convexity and compactness assumptions on problem (1.4) (see section 5), it finds ϵ_f-optimal primal solutions by combining partial Lagrangian solutions, even when Lagrange multipliers don't exist. These features are essentially "inherited" from the inexact framework of [Kiw06b] (although some technical developments are nontrivial). On the other hand, this paper reorganizes and simplifies the convergence framework of [Kiw06b] and sheds light on several important issues not discussed in there (such as the "true" impact of inexact evaluations, the possible use of "more inexact" null steps, primal recovery for Lagrangian relaxation with subgradient aggregation, and Lagrangian relaxation of equality constraints).

For the important special case where the functions ψ_i of the primal problem (1.4) are affine, we show how to employ *nonpolyhedral* models of f. Each model has the

form $\check{f}(\cdot) := \sup_{z\in\check{Z}} F_z(\cdot)$ stemming from (1.3), where $\check{Z}$ is a closed convex subset of Z. Then the proximal step can be implemented by solving a dual subproblem of minimizing a convex quadratic function over $\check{Z}$ (e.g., via interior-point methods when $\check{Z}$ is simple enough), and the projection on $C := \mathbb{R}^n_+$ is trivial. Further, the dual subproblem solutions estimate ϵ_f-optimal primal solutions asymptotically as above. In particular, our framework extends the highly successful methods of [FGRS06, sect. 3.2] and [ReS06, sect. 3] (see Remark 5.6).

Finally, for SDP (see below (1.4)) our general framework yields extensions of several variants of the spectral bundle method [Hel03, Hel04, HeK02, HeR00, Nay99]. This method employs the nonpolyhedral models discussed above, with $\check{Z}$ constructed from accumulated eigenvectors of the dual objective matrix $M(u)$. The original version of [HeR00] could handle only equality-constrained SDPs. Its extension [HeK02] to inequality-constrained SDPs can be seen as a specialization of the method of [Kiw99]; this helps in distinguishing its "driving force" from "implementation details" (although the latter are, of course, crucial for its performance in practice). Hence the primal recovery result of [Hel04, Thm. 3.6] also follows from our more general results (see Theorems 3.7 and 5.2); in fact, we don't need the assumption of [Hel04, Thm. 3.6] that the dual problem has a solution (see Remark 5.7(i)). Our extension to the case of approximate eigenvectors (see below (1.4)) is relevant for both theory and practice. Namely, while the existing version [HeK02] already employs approximate eigenvectors at so-called null steps (and this saves much work in practice [Hel03, HeK02, Nay99, Nay06]), it requires exact eigenvalues at the remaining descent steps. Our theoretical results show what to expect if approximate eigenvectors are used at descent steps as well, thus opening room for more efficient implementations.

The paper is organized as follows. In section 2 we present our method for general objective models. Its convergence is analyzed in section 3. Various modifications and model choices are given in section 4. Applications to Lagrangian relaxation are studied in section 5.

Our notation is fairly standard. $P_C(u) := \arg\min_C |\cdot - u|$ is the *projector* onto C.

2. The proximal-projection bundle method. Our method generates a sequence of *trial points* $\{u^k\}_{k=1}^\infty \subset C$ for evaluating the approximate values $f_u^k := f_{u^k}$, subgradients $g^k := g_{u^k}$, and linearizations $f_k := \bar{f}_{u^k}$ such that

$$f_k(\cdot) = f_u^k + \langle g^k, \cdot - u^k\rangle \le f(\cdot) \quad \text{with} \quad f_k(u^k) = f_u^k \ge f(u^k) - \epsilon_f, \tag{2.1}$$

as stipulated in (1.2). At iteration k, the current *prox* (or *stability*) *center* $\hat{u}^k := u^{k(l)} \in C$ for some $k(l) \le k$ has the value $f_{\hat{u}}^k := f_u^{k(l)}$ (usually $f_{\hat{u}}^k = \min_{j=1}^k f_u^j$); note that, by (2.1),

$$f_{\hat{u}}^k \in [f(\hat{u}^k) - \epsilon_f, f(\hat{u}^k)]. \tag{2.2}$$

For a model $\check{f}_k \le f$, the next point u^{k+1} approximately solves the prox subproblem

$$\min\ \check{f}_k(\cdot) + i_C(\cdot) + \frac{1}{2t_k}|\cdot - \hat{u}^k|^2, \tag{2.3}$$

where $t_k > 0$ is a *stepsize* that controls the size of $|u^{k+1} - \hat{u}^k|$. To this end, two partial linearizations of (2.3) are employed. First, replacing i_C by its past linearization $\bar{i}_C^{k-1} \le i_C$ in (2.3), we find its solution $\check{u}^{k+1}$ and a linearization $\bar{f}_k \le \check{f}_k$ such that

$\check{u}^{k+1}$ solves (2.3) with $\check{f}_k$, i_C replaced by $\bar{f}_k$, $\bar{\imath}_C^{k-1}$. Next, replacing $\check{f}_k$ by $\bar{f}_k$ in (2.3), we find its solution u^{k+1} and a linearization $\bar{\imath}_C^k \le i_C$ such that u^{k+1} solves (2.3) with $\check{f}_k$, i_C replaced by $\bar{f}_k$, $\bar{\imath}_C^k$. Due to evaluation errors, we may have $f_{\hat{u}}^k < \check{f}_k(\hat{u}^k)$, in which case the *predicted descent* $v_k := f_{\hat{u}}^k - \bar{f}_k(u^{k+1})$ may be nonpositive; then t_k is increased and u^{k+1} is recomputed to decrease $\bar{f}_k(u^{k+1})$ until $v_k > 0$. A *descent* step to $\hat{u}^{k+1} := u^{k+1}$ is taken if $f_u^{k+1} \le f_{\hat{u}}^k - \kappa v_k$ for a fixed $\kappa \in (0,1)$. Otherwise, a *null* step $\hat{u}^{k+1} := \hat{u}^k$ occurs; then $\bar{f}_k$ and the new linearization f_{k+1} are used to produce a better model $\check{f}_{k+1} \ge \max\{\bar{f}_k, f_{k+1}\}$ (e.g., $\check{f}_{k+1} = \max\{\bar{f}_k, f_{k+1}\}$).

Specific rules of our method will be discussed after its formal statement below.

ALGORITHM 2.1.

Step 0 (initialization). Select $u^1 \in C$, a *descent parameter* $\kappa \in (0,1)$, a *stepsize bound* $t_{\min} > 0$, and a *stepsize* $t_1 \ge t_{\min}$. Set $\bar{f}_0 := f_1$ (cf. (2.1)), $\bar{\imath}_C^0 := \langle p_C^0, \cdot - u^1\rangle$ with $p_C^0 := 0$, $\hat{u}^1 := u^1$, $f_{\hat{u}}^1 := f_u^1 := f_{u^1}$, $g^1 := g_{u^1}$ (cf. (2.1)), $i_t^1 := 0$, $k := k(0) := 1$, $l := 0$ ($k(l) - 1$ will denote the iteration of the lth descent step).

Step 1 (model selection). Choose $\check{f}_k : \mathbb{R}^n \to \mathbb{R}$ closed convex and such that

$$\max\{\bar{f}_{k-1}, f_k\} \le \check{f}_k \le f_C. \tag{2.4}$$

Step 2 (proximal point finding). Set

$$\check{u}^{k+1} := \arg\min\left\{ \phi_f^k(\cdot) := \check{f}_k(\cdot) + \bar{\imath}_C^{k-1}(\cdot) + \frac{1}{2t_k}|\cdot - \hat{u}^k|^2 \right\}, \tag{2.5}$$

$$\bar{f}_k(\cdot) := \check{f}_k(\check{u}^{k+1}) + \langle p_f^k, \cdot - \check{u}^{k+1}\rangle \quad \text{with} \quad p_f^k := \frac{1}{t_k}(\hat{u}^k - \check{u}^{k+1}) - p_C^{k-1}. \tag{2.6}$$

Step 3 (projection). Set

$$u^{k+1} := \arg\min\left\{ \phi_C^k(\cdot) := \bar{f}_k(\cdot) + i_C(\cdot) + \frac{1}{2t_k}|\cdot - \hat{u}^k|^2 \right\} = P_C(\hat{u}^k - t_k p_f^k), \tag{2.7}$$

$$\bar{\imath}_C^k(\cdot) := \langle p_C^k, \cdot - u^{k+1}\rangle \quad \text{with} \quad p_C^k := \frac{1}{t_k}(\hat{u}^k - u^{k+1}) - p_f^k, \tag{2.8}$$

$$v_k := f_{\hat{u}}^k - \bar{f}_k(u^{k+1}), \quad p^k := \frac{1}{t_k}(\hat{u}^k - u^{k+1}), \quad \text{and} \quad \epsilon_k := v_k - t_k|p^k|^2. \tag{2.9}$$

Step 4 (stopping criterion). If $\max\{|p^k|, \epsilon_k\} = 0$, stop ($f_{\hat{u}}^k \le f_*$).

Step 5 (stepsize correction). If $v_k < -\epsilon_k$, set $t_k := 10t_k$, $i_t^k := k$, and go back to Step 2.

Step 6 (descent test). Evaluate f_u^{k+1} and g^{k+1} (cf. (2.1)). If the *descent test* holds,

$$f_u^{k+1} \le f_{\hat{u}}^k - \kappa v_k, \tag{2.10}$$

set $\hat{u}^{k+1} := u^{k+1}$, $f_{\hat{u}}^{k+1} := f_u^{k+1}$, $i_t^{k+1} := 0$, $k(l+1) := k+1$, and increase l by 1 (*descent step*); otherwise, set $\hat{u}^{k+1} := \hat{u}^k$, $f_{\hat{u}}^{k+1} := f_{\hat{u}}^k$, and $i_t^{k+1} := i_t^k$ (*null step*).

Step 7 (stepsize updating). If $k(l) = k+1$ (i.e., after a descent step), select $t_{k+1} \ge t_k$; otherwise, either set $t_{k+1} := t_k$ or choose $t_{k+1} \in [t_{\min}, t_k]$ if $i_t^{k+1} = 0$.

Step 8 (loop). Increase k by 1 and go to Step 1.

Several comments on the method are in order. Step 1 may choose the simplest model $\check{f}_k = \max\{\bar{f}_{k-1}, f_k\}$; more efficient choices are given in section 4.4. For a polyhedral model $\check{f}_k$, subproblem (2.5) can be handled via simple quadratic program-

ming (QP) solvers [Kiw86]; in contrast, the more difficult subproblem (2.3) employed in [Kiw06b] requires more sophisticated solvers even for a polyhedral set C [Kiw94]. The projection of (2.7) is easily found if the set C is "simple" (e.g., the Cartesian product of boxes, simplices, and ellipsoids).

We now use the relations of Steps 2 and 3 to derive an optimality estimate, which involves the *aggregate linearization* $\bar{f}_C^k := \bar{f}_k + \bar{\imath}_C^k$ and the *optimality measure*

$$V_k := \max\bigl\{|p^k|, \epsilon_k + \langle p^k, \hat{u}^k\rangle\bigr\}. \tag{2.11}$$

LEMMA 2.2. (i) *The vectors* p_f^k *and* p_C^k *defined in* (2.6) *and* (2.8) *are in fact subgradients,*

$$p_f^k \in \partial \check{f}_k(\check{u}^{k+1}) \quad \text{and} \quad p_C^k \in \partial i_C(u^{k+1}), \tag{2.12}$$

and the linearizations $\bar{f}_k$ *and* $\bar{\imath}_C^k$ *defined in* (2.6) *and* (2.8) *provide the minorizations*

$$\bar{f}_k \le \check{f}_k, \quad \bar{\imath}_C^k \le i_C, \quad \text{and} \quad \bar{f}_C^k := \bar{f}_k + \bar{\imath}_C^k \le f_C. \tag{2.13}$$

(ii) *The* aggregate subgradient p^k *defined in* (2.9) *and the linearization* $\bar{f}_C^k$ *above satisfy*

$$p^k = p_f^k + p_C^k = \frac{\hat{u}^k - u^{k+1}}{t_k}, \tag{2.14}$$

$$\bar{f}_C^k(\cdot) = \bar{f}_k(u^{k+1}) + \langle p^k, \cdot - u^{k+1}\rangle. \tag{2.15}$$

(iii) *The predicted descent* v_k *and the* aggregate linearization error ϵ_k *of* (2.9) *satisfy*

$$v_k = t_k|p^k|^2 + \epsilon_k \quad \text{and} \quad \epsilon_k = f_{\hat{u}}^k - \bar{f}_C^k(\hat{u}^k). \tag{2.16}$$

(iv) *The aggregate linearization* $\bar{f}_C^k$ *is expressed in terms of* p^k *and* ϵ_k *as follows:*

$$f_{\hat{u}}^k - \epsilon_k + \langle p^k, \cdot - \hat{u}^k\rangle = \bar{f}_C^k(\cdot) \le f_C(\cdot). \tag{2.17}$$

(v) *The optimality measure* V_k *of* (2.11) *satisfies* $V_k \le \max\{|p^k|, \epsilon_k\}(1 + |\hat{u}^k|)$ *and*

$$f_{\hat{u}}^k \le f_C(u) + V_k\bigl(1 + |u|\bigr) \quad \text{for all } u. \tag{2.18}$$

(vi) *We have* $v_k \ge -\epsilon_k \Leftrightarrow t_k|p^k|^2/2 \ge -\epsilon_k \Leftrightarrow v_k \ge t_k|p^k|^2/2$. *Moreover,* $v_k \ge \epsilon_k$, $-\epsilon_k \le \epsilon_f$, *and*

$$v_k \ge \max\left\{\frac{t_k|p^k|^2}{2}, |\epsilon_k|\right\} \qquad \text{if} \quad v_k \ge -\epsilon_k, \tag{2.19}$$

$$V_k \le \max\left\{\left(\frac{2v_k}{t_k}\right)^{1/2}, v_k\right\}\bigl(1 + |\hat{u}^k|\bigr) \quad \text{if} \quad v_k \ge -\epsilon_k, \tag{2.20}$$

$$V_k < \left(\frac{2\epsilon_f}{t_k}\right)^{1/2}\bigl(1 + |\hat{u}^k|\bigr) \qquad \text{if} \quad v_k < -\epsilon_k. \tag{2.21}$$

Proof. (i) By (2.5)–(2.6), the optimality condition (using $\nabla\bar{\imath}_C^{k-1} = p_C^{k-1}$; cf. (2.8))

$$0 \in \partial\phi_f^k(\check{u}^{k+1}) = \partial\check{f}_k(\check{u}^{k+1}) + p_C^{k-1} + \frac{\check{u}^{k+1} - \hat{u}^k}{t_k} = \partial\check{f}_k(\check{u}^{k+1}) - p_f^k$$

and the equality $\bar{f}_k(\check{u}^{k+1}) = \check{f}_k(\check{u}^{k+1})$ yield $p_f^k \in \partial\check{f}_k(\check{u}^{k+1})$ and $\bar{f}_k \le \check{f}_k$. By (2.7)–(2.8),

$$0 \in \partial\phi_C^k(u^{k+1}) = p_f^k + \partial i_C(u^{k+1}) + \frac{u^{k+1} - \hat{u}^k}{t_k} = \partial i_C(u^{k+1}) - p_C^k$$

(using $\nabla\bar{f}_k = p_f^k$) and $\bar{\imath}_C^k(u^{k+1}) = i_C(u^{k+1}) = 0$ give $p_C^k \in \partial i_C(u^{k+1})$ and $\bar{\imath}_C^k \le i_C$. Combining both minorizations, we obtain that $\bar{f}_k + \bar{\imath}_C^k \le \check{f}_k + i_C \le f_C$ by (2.4) and (1.5).

(ii) Use the linearity of $\bar{f}_C^k := \bar{f}_k + \bar{\imath}_C^k$, (2.6), (2.8) with $\bar{\imath}_C^k(u^{k+1}) = 0$, and (2.9).

(iii) Rewrite (2.9), using the fact that $\bar{f}_C^k(\hat{u}^k) = \bar{f}_k(u^{k+1}) + t_k|p^k|^2$, by (ii).

(iv) We have $f_{\hat{u}}^k - \epsilon_k = \bar{f}_C^k(\hat{u}^k)$ by (iii), and $\bar{f}_C^k$ is affine by (ii) and minorizes f_C by (i).

(v) Use the Cauchy–Schwarz inequality in the definition (2.11) and in (iv).

(vi) The equivalences follow from the expression of $v_k = t_k|p^k|^2 + \epsilon_k$ in (iii); in particular, $v_k \ge \epsilon_k$. Next, by (2.16), (2.13), and (2.2) with $f_C(\hat{u}^k) = f(\hat{u}^k)$ ($\hat{u}^k \in C$), we have

$$-\epsilon_k = \bar{f}_C^k(\hat{u}^k) - f_{\hat{u}}^k \le f_C(\hat{u}^k) - f_{\hat{u}}^k = f(\hat{u}^k) - f_{\hat{u}}^k \le \epsilon_f.$$

Finally, to obtain the bounds (2.19)–(2.21), use the equivalences together with the facts that $v_k \ge \epsilon_k$, $-\epsilon_k \le \epsilon_f$ and the bound on V_k from assertion (v). ∎

The optimality estimate (2.18) justifies the stopping criterion of Step 4: $V_k = 0$ yields $f_{\hat{u}}^k \le \inf f_C = f_*$; thus, the point $\hat{u}^k$ is ϵ_f-*optimal*; i.e., $f(\hat{u}^k) \le f_* + \epsilon_f$ by (2.2). In the case of exact evaluations ($\epsilon_f = 0$), we have $v_k \ge \epsilon_k \ge 0$ by Lemma 2.2(vi), Step 5 is redundant, and Algorithm 2.1 becomes essentially that of [Kiw99, Alg. 3.1]. When inexactness is discovered via $v_k < -\epsilon_k$, the stepsize t_k is increased to produce descent or confirm that $\hat{u}^k$ is ϵ_f-optimal. Namely, when $\hat{u}^k$ is bounded in (2.21), increasing t_k drives V_k to 0, so that $f_{\hat{u}}^k \le f_*$ asymptotically. Whenever t_k is increased at Step 5, the *stepsize indicator* $i_t^k \ne 0$ prevents Step 7 from decreasing t_k after null steps until the next descent step occurs (cf. Step 6). Otherwise, decreasing t_k at Step 7 aims at collecting more local information about f at null steps.

We now show that an infinite cycle between Steps 2 and 5 means that $\hat{u}^k$ is ϵ_f-optimal.

LEMMA 2.3. *If an infinite cycle between Steps* 2 *and* 5 *occurs, then* $f_{\hat{u}}^k \le f_*$ *and* $V_k \to 0$.

Proof. At Step 5 during the cycle the facts that $V_k < (2\epsilon_f/t_k)^{1/2}(1 + |\hat{u}^k|)$ by (2.21) and $t_k \uparrow \infty$ as the cycle continues give $V_k \to 0$, so that $f_{\hat{u}}^k \le \inf f_C = f_*$ by (2.18). ∎

3. Convergence. In view of Lemma 2.3, we may suppose that the algorithm neither terminates nor cycles infinitely between Steps 2 and 5 (otherwise $\hat{u}^k$ is ϵ_f-optimal). At Step 6, we have $u^{k+1} \in C$ and $v_k > 0$ (by (2.19), since $\max\{|p^k|, \epsilon_k\} > 0$ at Step 4), so that $\hat{u}^{k+1} \in C$ and $f_{\hat{u}}^{k+1} \le f_{\hat{u}}^k$ for all k. We shall show that the asymptotic value $f_{\hat{u}}^\infty := \lim_k f_{\hat{u}}^k$ satisfies $f_{\hat{u}}^\infty \le f_*$. As in [Kiw99, sect. 4], we assume that the model subgradients $p_f^k \in \partial\check{f}_k(\check{u}^{k+1})$ in (2.12) satisfy

$$\{p_f^k\} \text{ is bounded if } \{u^k\} \text{ is bounded.} \tag{3.1}$$

It will be seen in Remark 4.4 that typical models $\check{f}_k$ satisfy this condition automatically.

We first consider the case where only finitely many descent steps occur. After the last descent step, only null steps occur, and the sequence $\{t_k\}$ eventually becomes monotone, since once Step 5 increases t_k, Step 7 can't decrease t_k; thus the limit $t_\infty := \lim_k t_k$ exists. We deal with the cases of $t_\infty = \infty$ in Lemma 3.1 and $t_\infty < \infty$ in Lemma 3.2 below.

LEMMA 3.1. *Suppose there exists $\bar{k}$ such that only null steps occur for all $k \geq \bar{k}$, and $t_\infty := \lim_k t_k = \infty$. Let $K := \{k \geq \bar{k} : t_{k+1} > t_k\}$. Then $V_k \xrightarrow{K} 0$ at Step 5.*

Proof. At iteration $k \in K$, before Step 5 increases t_k for the last time, we have $V_k < (2\epsilon_f/t_k)^{1/2}(1 + |\hat{u}^{\bar{k}}|)$ by (2.21); consequently, $t_k \to \infty$ gives $V_k \xrightarrow{K} 0$. ∎

LEMMA 3.2. *Suppose there exists $\bar{k}$ such that, for all $k \geq \bar{k}$, only null steps occur and Step 5 doesn't increase t_k. Then $V_k \to 0$.*

Proof. First, using partial linearizations of subproblems (2.5) and (2.7), we show that their optimal values $\phi_f^k(\check{u}^{k+1}) \leq \phi_C^k(u^{k+1})$ are nondecreasing and bounded above.

Fix $k \geq \bar{k}$. By the definitions in (2.5)–(2.6), we have $\bar{f}_k(\check{u}^{k+1}) = \check{f}_k(\check{u}^{k+1})$ and

$$\check{u}^{k+1} = \arg\min \left\{ \bar{\phi}_f^k(\cdot) := \bar{f}_k(\cdot) + \bar{\imath}_C^{k-1}(\cdot) + \tfrac{1}{2t_k}|\cdot - \hat{u}^k|^2 \right\} \tag{3.2}$$

from $\nabla\bar{\phi}_f^k(\check{u}^{k+1}) = 0$. Since $\bar{\phi}_f^k$ is quadratic and $\bar{\phi}_f^k(\check{u}^{k+1}) = \phi_f^k(\check{u}^{k+1})$, by Taylor's expansion

$$\bar{\phi}_f^k(\cdot) = \phi_f^k(\check{u}^{k+1}) + \tfrac{1}{2t_k}|\cdot - \check{u}^{k+1}|^2. \tag{3.3}$$

Similarly, by the definitions in (2.7)–(2.8), we have $\bar{\imath}_C^k(u^{k+1}) = i_C(u^{k+1}) = 0$,

$$u^{k+1} = \arg\min \left\{ \bar{\phi}_C^k(\cdot) := \bar{f}_k(\cdot) + \bar{\imath}_C^k(\cdot) + \tfrac{1}{2t_k}|\cdot - \hat{u}^k|^2 \right\}, \tag{3.4}$$

$$\bar{\phi}_C^k(\cdot) = \phi_C^k(u^{k+1}) + \tfrac{1}{2t_k}|\cdot - u^{k+1}|^2. \tag{3.5}$$

Next, to bound the objective values of the linearized subproblems (3.2) and (3.4) from above, we use the minorizations $\bar{f}_k \leq f_C$ and $\bar{\imath}_C^{k-1}, \bar{\imath}_C^k \leq i_C$ of (2.13) with $\hat{u}^k \in C$:

$$\phi_f^k(\check{u}^{k+1}) + \tfrac{1}{2t_k}|\check{u}^{k+1} - \hat{u}^k|^2 = \bar{\phi}_f^k(\hat{u}^k) \leq f(\hat{u}^k), \tag{3.6a}$$

$$\phi_C^k(u^{k+1}) + \tfrac{1}{2t_k}|u^{k+1} - \hat{u}^k|^2 = \bar{\phi}_C^k(\hat{u}^k) \leq f(\hat{u}^k), \tag{3.6b}$$

where the equalities stem from (3.3) and (3.5). Due to the minorization $\bar{\imath}_C^{k-1} \leq i_C$, the objectives of subproblems (3.2) and (2.7) satisfy $\bar{\phi}_f^k \leq \phi_C^k$. On the other hand, since $\hat{u}^{k+1} = \hat{u}^k$, $t_{k+1} \leq t_k$ (cf. Step 7), and $\bar{f}_k \leq \check{f}_{k+1}$ by (2.4), the objectives of (3.4) and the next subproblem (2.5) satisfy $\bar{\phi}_C^k \leq \phi_f^{k+1}$. Altogether, by (3.3) and (3.5), we see that

$$\phi_f^k(\check{u}^{k+1}) + \tfrac{1}{2t_k}|u^{k+1} - \check{u}^{k+1}|^2 = \bar{\phi}_f^k(u^{k+1}) \leq \phi_C^k(u^{k+1}), \tag{3.7a}$$

$$\phi_C^k(u^{k+1}) + \tfrac{1}{2t_k}|\check{u}^{k+2} - u^{k+1}|^2 = \bar{\phi}_C^k(\check{u}^{k+2}) \leq \phi_f^{k+1}(\check{u}^{k+2}). \tag{3.7b}$$

In particular, the inequalities $\phi_f^k(\check{u}^{k+1}) \leq \phi_C^k(u^{k+1}) \leq \phi_f^{k+1}(\check{u}^{k+2})$ imply that the nondecreasing sequences $\{\phi_f^k(\check{u}^{k+1})\}_{k\geq\bar{k}}$ and $\{\phi_C^k(u^{k+1})\}_{k\geq\bar{k}}$, which are bounded above

by (3.6) with $\hat{u}^k = \hat{u}^{\bar{k}}$ for all $k \geq \bar{k}$, must have a common limit, say $\phi_\infty \leq f(\hat{u}^{\bar{k}})$. Moreover, since the stepsizes satisfy $t_k \leq t_{\bar{k}}$ for all $k \geq \bar{k}$, we deduce from the bounds (3.6)–(3.7) that

$$\phi_f^k(\check{u}^{k+1}),\ \phi_C^k(u^{k+1}) \uparrow \phi_\infty, \qquad \check{u}^{k+2} - u^{k+1} \to 0, \tag{3.8}$$

and the sequences $\{\check{u}^{k+1}\}$ and $\{u^{k+1}\}$ are bounded. Then the sequence $\{p_f^k\}$ is bounded by (3.1), and the sequence $\{g^k\}$ is bounded as well, since $g^k \in \partial_{\epsilon_f} f(u^k)$ by (2.1), whereas the mapping $\partial_{\epsilon_f} f$ is locally bounded [HUL93, sect. XI.4.1].

We now show that the *approximation error* $\check{\epsilon}_k := f_u^{k+1} - \bar{f}_k(u^{k+1})$ vanishes. Using the form (2.1) of f_{k+1}, the minorization $f_{k+1} \leq \check{f}_{k+1}$ of (2.4), the Cauchy–Schwarz inequality, and the optimal values of subproblems (2.5) and (2.7) with $\hat{u}^k = \hat{u}^{\bar{k}}$ for $k \geq \bar{k}$, we estimate

$$\begin{aligned}
\check{\epsilon}_k &:= f_u^{k+1} - \bar{f}_k(u^{k+1}) = f_{k+1}(\check{u}^{k+2}) - \bar{f}_k(u^{k+1}) + \langle g^{k+1}, u^{k+1} - \check{u}^{k+2}\rangle \\
&\leq \check{f}_{k+1}(\check{u}^{k+2}) - \bar{f}_k(u^{k+1}) + |g^{k+1}||u^{k+1} - \check{u}^{k+2}| \\
&= \phi_f^{k+1}(\check{u}^{k+2}) - \phi_C^k(u^{k+1}) + \Delta_k - \bar{\imath}_C^k(\check{u}^{k+2}) + |g^{k+1}||u^{k+1} - \check{u}^{k+2}|, \qquad (3.9)
\end{aligned}$$

where $\Delta_k := |u^{k+1} - \hat{u}^{\bar{k}}|^2/2t_k - |\check{u}^{k+2} - \hat{u}^{\bar{k}}|^2/2t_{k+1}$. To see that $\Delta_k \to 0$, note that

$$|\check{u}^{k+2} - \hat{u}^{\bar{k}}|^2 = |u^{k+1} - \hat{u}^{\bar{k}}|^2 + 2\langle \check{u}^{k+2} - u^{k+1}, u^{k+1} - \hat{u}^{\bar{k}}\rangle + |\check{u}^{k+2} - u^{k+1}|^2,$$

$|u^{k+1} - \hat{u}^{\bar{k}}|^2$ is bounded, $\check{u}^{k+2} - u^{k+1} \to 0$ by (3.8), and $t_{\min} \leq t_{k+1} \leq t_k$ for $k \geq \bar{k}$ by Step 7. These properties also give $\bar{\imath}_C^k(\check{u}^{k+2}) \to 0$, since by (2.8) and the Cauchy–Schwarz inequality, we have

$$|\bar{\imath}_C^k(\check{u}^{k+2})| \leq |p_C^k||\check{u}^{k+2} - u^{k+1}| \quad \text{with} \quad |p_C^k| \leq |u^{k+1} - \hat{u}^{\bar{k}}|/t_k + |p_f^k|,$$

where $\{p_f^k\}$ is bounded. Hence, using (3.8) and the boundedness of $\{g^{k+1}\}$ in (3.9) yields $\overline{\lim}_k \check{\epsilon}_k \leq 0$. On the other hand, for $k \geq \bar{k}$ the null step condition $f_u^{k+1} > f_{\hat{u}}^k - \kappa v_k$ gives

$$\check{\epsilon}_k = \big[f_u^{k+1} - f_{\hat{u}}^k\big] + \big[f_{\hat{u}}^k - \bar{f}_k(u^{k+1})\big] > -\kappa v_k + v_k = (1-\kappa)v_k \geq 0,$$

where $\kappa < 1$ by Step 0; we conclude that $\check{\epsilon}_k \to 0$ and $v_k \to 0$. Finally, since $v_k \to 0$, $t_k \geq t_{\min}$ (cf. Step 7), and $\hat{u}^k = \hat{u}^{\bar{k}}$ for $k \geq \bar{k}$, we have $V_k \to 0$ by (2.20). ∎

We may now finish the case of infinitely many consecutive null steps.

LEMMA 3.3. *Suppose that there exists $\bar{k}$ such that only null steps occur for all $k \geq \bar{k}$. Let $K := \{k \geq \bar{k} : t_{k+1} > t_k\}$ if $t_k \to \infty$, $K := \{k : k \geq \bar{k}\}$ otherwise. Then $V_k \xrightarrow{K} 0$.*

Proof. Steps 5–7 ensure that the sequence $\{t_k\}$ is monotone for large k. We have $V_k \xrightarrow{K} 0$ from either Lemma 3.1 if $t_\infty = \infty$, or Lemma 3.2 if $t_\infty < \infty$. ∎

It remains to analyze the case of infinitely many descent steps.

LEMMA 3.4. *Suppose that infinitely many descent steps occur and $f_{\hat{u}}^\infty := \lim_k f_{\hat{u}}^k > -\infty$. Let $K := \{k : f_{\hat{u}}^{k+1} < f_{\hat{u}}^k\}$. Then $\underline{\lim}_{k\in K} V_k = 0$. Moreover, if $\{\hat{u}^k\}$ is bounded, then $V_k \xrightarrow{K} 0$.*

Proof. We have $0 < \kappa v_k \leq f_{\hat{u}}^k - f_{\hat{u}}^{k+1}$ if $k \in K$, $f_{\hat{u}}^{k+1} = f_{\hat{u}}^k$ otherwise (see Step 6). Thus $\sum_{k\in K} \kappa v_k \leq f_{\hat{u}}^1 - f_{\hat{u}}^\infty < \infty$ gives $v_k \xrightarrow{K} 0$ and hence $\epsilon_k, t_k|p^k|^2 \xrightarrow{K} 0$ by (2.19)

and $|p^k| \xrightarrow{K} 0$, using $t_k \geq t_{\min}$ (cf. Step 7). For $k \in K$, $\hat{u}^{k+1} - \hat{u}^k = -t_k p^k$ by (2.9), so

$$|\hat{u}^{k+1}|^2 - |\hat{u}^k|^2 = t_k\big\{t_k|p^k|^2 - 2\langle p^k, \hat{u}^k\rangle\big\}.$$

Sum up and use the facts that $\hat{u}^{k+1} = \hat{u}^k$ if $k \notin K$, $\sum_{k\in K} t_k \geq \sum_{k\in K} t_{\min} = \infty$ to get

$$\overline{\lim_{k\in K}}\big\{t_k|p^k|^2 - 2\langle p^k, \hat{u}^k\rangle\big\} \geq 0$$

(since otherwise $|\hat{u}^k|^2 \to -\infty$, which is impossible). Combining this with $t_k|p^k|^2 \xrightarrow{K} 0$ gives $\underline{\lim}_{k\in K}\langle p^k, \hat{u}^k\rangle \leq 0$. Since also $\epsilon_k, |p^k| \xrightarrow{K} 0$, we have $\underline{\lim}_{k\in K} V_k = 0$ by (2.11).

If $\{\hat{u}^k\}$ is bounded, using $\epsilon_k, |p^k| \xrightarrow{K} 0$ in Lemma 2.2(v) gives $V_k \xrightarrow{K} 0$. ∎

We may now state and prove our principal result.

THEOREM 3.5. (i) *We have* $f_{\hat{u}}^k \downarrow f_{\hat{u}}^\infty \leq f_*$, *and additionally* $\underline{\lim}_k V_k = 0$ *if* $f_* > -\infty$.

(ii) $f_* \leq \underline{\lim}_k f(\hat{u}^k) \leq \overline{\lim}_k f(\hat{u}^k) \leq f_{\hat{u}}^\infty + \epsilon_f$.

Proof. The inequalities in (ii) stem from the facts that $f_* = \inf_C f$, $\{\hat{u}^k\} \subset C$, and $f(\hat{u}^k) \leq f_{\hat{u}}^k + \epsilon_f$ for all k by (2.2). By (ii), if $f_{\hat{u}}^\infty = -\infty$, then $f_* = -\infty$ in (i). Hence, suppose $f_* > -\infty$. Then $f_{\hat{u}}^\infty \geq f_* - \epsilon_f > -\infty$ by (ii). We have $\underline{\lim}_k V_k = 0$ by Lemma 3.3 in the case of finitely many descent steps, or by Lemma 3.4 otherwise. Finally, using $\underline{\lim}_k V_k = 0$ in the estimate (2.18) gives $f_{\hat{u}}^\infty \leq \inf f_C = f_*$. ∎

It is instructive to examine the assumptions of the preceding results.

Remark 3.6. (i) Inspection of the preceding proofs reveals that Theorem 3.5 requires only convexity and finiteness of f on C, and *local boundedness* of the approximate subgradient mapping $u \mapsto g_u$ of f on C (see below (3.8)). In particular, it suffices to assume that f is finite convex on a neighborhood of C.

(ii) The requirement $\max\{\bar{f}_{k-1}, f_k\} \leq \check{f}_k$ of (2.4) is needed only after null steps in the proof of Lemma 3.2. After a descent step (when $k = k(l)$), Step 1 may take any $\check{f}_k \leq f_C$.

We now show that for exact evaluations ($\epsilon_f = 0$), our algorithm has the usual strong convergence properties of typical bundle methods. Instead of requiring that $\inf_k t_k \geq t_{\min} > 0$, as before, we give more general stepsize conditions in the theorem below.

THEOREM 3.7. *Suppose that* $\epsilon_f = 0$. *Let* $U_* := \operatorname{Arg\,min}_C f$ *denote the* (*possibly empty*) *solution set of problem* (1.1). *Then we have the following statements:*

(i) *If only* $l < \infty$ *descent steps occur and* $t_k \downarrow t_\infty > 0$, *then* $\hat{u}^{k(l)} \in U_*$ *and* $V_k \to 0$.

(ii) *Assuming that infinitely many descent steps occur, suppose that* $\sum_{k\in K} t_k = \infty$ *for* $K := \{k : f(\hat{u}^{k+1}) < f(\hat{u}^k)\}$. *Then* $f(\hat{u}^k) \downarrow f_*$. *Moreover, we have the following.*

(a) Let $\check{\epsilon}_k := f(\hat{u}^{k+1}) - \bar{f}_k(\hat{u}^{k+1})$ for $k \in K$. If $U_* \neq \emptyset$ and $\sum_{k\in K} t_k\check{\epsilon}_k < \infty$ (e.g., $\sup_{k\in K} t_k < \infty$), then $\hat{u}^k \to \hat{u}^\infty \in U_*$, and $V_k \xrightarrow{K} 0$ if $\inf_{k\in K} t_k > 0$.

(b) If $U_* = \emptyset$, then $|\hat{u}^k| \to \infty$.

Proof. Since $\epsilon_f = 0$, Step 5 is inactive, and Algorithm 2.1 fits the framework of [Kiw99, Alg. 3.1]. For $l \not\to \infty$, the conclusion follows from Lemma 3.2 and Theorem 3.5. For $l \to \infty$, combine [Kiw99, Thm. 4.4] and the proof of Lemma 3.4. ∎

4. Modifications.

4.1. Looping between subproblems. To obtain a more accurate solution to the prox subproblem (2.3), we may cycle between subproblems (2.5) and (2.7), updating their data as if null steps occur without changing the model $\check{f}_k$. Specifically, for a given *subproblem accuracy threshold* $\check{\kappa} \in (0,1)$, suppose that the following step is inserted after Step 5.

Step 5′ (subproblem accuracy test). If

$$\check{f}_k(u^{k+1}) > f_{\hat{u}}^k - \check{\kappa} v_k, \tag{4.1}$$

set $\bar{\imath}_C^{k-1}(\cdot) := \bar{\imath}_C^k(\cdot)$, $p_C^{k-1} := p_C^k$ and go back to Step 2.

We now give two motivations for the test (4.1) written as (cf. (2.9))

$$\bar{\epsilon}_k := \check{f}_k(u^{k+1}) - \bar{f}_k(u^{k+1}) > (1-\check{\kappa}) v_k.$$

First, when $\bar{\epsilon}_k$ is small relative to v_k, $\check{f}_k$ is correctly approximated by $\bar{f}_k$, so the loop can be broken. Second, since $\bar{f}_k \le \check{f}_k$ (Lemma 2.2(i)) in (2.7), by standard arguments [Kiw99, p. 145], the distance from u^{k+1} to the prox solution of (2.3) is at most $\sqrt{2t_k\bar{\epsilon}_k}$.

The analysis of this modification is given in the following remarks.

Remark 4.1. (i) For any k, each execution of Steps 2 through 5′ is called a loop. First, suppose that finitely many loops occur for each k. By its proof, Lemma 2.2 holds at Step 4 for the current quantities. This suffices for the proofs of Lemmas 2.3, 3.1, and 3.4, whereas the proofs of Lemma 3.3 and Theorem 3.5 will go through once Lemma 3.2 is established. The proof of Lemma 3.2 is modified as follows. For each $k \ge \bar{k}$, (3.6) and (3.7a) hold at each loop, and (3.7b) holds for the final loop. For any preceding loop, letting $\check{u}_{\rm next}^{k+1}$ and $\phi_{f,{\rm next}}^k$ stand for $\check{u}^{k+1}$ and ϕ_f^k produced by Step 2 on the next loop, use the minorization $\bar{f}_k \le \check{f}_k$ of (2.13) in subproblems (3.4) and (2.7) to get $\bar{\phi}_C^k \le \phi_{f,{\rm next}}^k$ and, by (3.5),

$$\phi_C^k(u^{k+1}) + \tfrac{1}{2t_k}|\check{u}_{\rm next}^{k+1} - u^{k+1}|^2 = \bar{\phi}_C^k(\check{u}_{\rm next}^{k+1}) \le \phi_{f,{\rm next}}^k(\check{u}_{\rm next}^{k+1}). \tag{4.2}$$

Then, replacing (3.7b) by (4.2) for all nonfinal loops, we deduce that the optimal values $\phi_f^k(\check{u}^{k+1}) \le \phi_C^k(u^{k+1})$ can't decrease during the loops or when k grows; hence (3.8) and the boundedness of $\{\check{u}^{k+1}\}$ and $\{u^{k+1}\}$ follow as before. For the rest of the proof, let $\check{u}^{k+2}$ in (3.9) stand for the point produced by Step 2 on the first loop at iteration $k+1$, and argue as before.

(ii) Next, suppose that infinitely many loops occur at iteration $k = \check{k}$, for some $\check{k}$. If Step 5 drives t_k to ∞, $f_{\hat{u}}^{\check{k}} \le f_*$ and $V_k \to 0$ by the proof of Lemma 2.3. Hence we may assume that Step 5 doesn't increase t_k at all. To show that $V_k \to 0$ (in which case $f_{\hat{u}}^{\check{k}} \le f_*$ by (2.18)), we suppose that the subdifferential $\partial\check{f}_k$ is locally bounded, and we use a subgradient mapping $C \ni u \mapsto \check{g}_u \in \partial\check{f}_k(u)$. Consider the following modification of Algorithm 2.1. Starting from the first loop at iteration $k = \check{k}$, omit Step 5′; at Step 6 set $f_u^{k+1} := \check{f}_k(u^{k+1})$, $g^{k+1} := \check{g}_{u^{k+1}}$, and $\kappa := \check{\kappa}$; at Step 7, set $t_{k+1} := t_k$; finally, when Step 1 is reached, set $\check{f}_k := \check{f}_{k-1}$. This modification only translates loops into additional iterations with a constant model $\check{f}_k = \check{f}_{\check{k}}$; in particular, only null steps occur, because the descent test (2.10) can't hold with $f_u^{k+1} := \check{f}_k(u^{k+1})$ and $\kappa := \check{\kappa}$ due to the model test (4.1). Further, the "new" linearization $f_{k+1}(\cdot) := f_u^{k+1} + \langle g^{k+1}, \cdot - u^{k+1}\rangle$ satisfies $f_{k+1} \le \check{f}_{k+1}$. Hence, to get $V_k \to 0$, we may use the proof of Lemma 3.2, obtaining boundedness of $\{p_f^k\}$, $\{g^{k+1}\}$ from the boundedness of $\{\check{u}^{k+1}\}$, $\{u^{k+1}\}$ and the local boundedness of $\partial\check{f}_k$.

Note that having $\bar{\imath}_C^{k-1}$ as a model of i_C in subproblem (2.5) is essential only after null steps or loops due to Step 5′. Otherwise, a better model may be constructed as follows. After Step 5 increases t_k, we can set $\bar{\imath}_C^{k-1}(\cdot) := \bar{\imath}_C^k(\cdot)$, $p_C^{k-1} := p_C^k$, or use the more efficient update $u^k := P_C(\hat{u}^k - t_k p_f^k)$, $p_C^{k-1} := (\hat{u}^k - u^k)/t_k - p_f^k$, and $\bar{\imath}_C^{k-1}(\cdot) := \langle p_C^{k-1}, \cdot - u^k\rangle$, which corresponds to resolving subproblem (2.7) before going back to Step 2. Similarly, if $\hat{u}^{k+1} \neq \hat{u}^k$ after Step 7, we may use $\tilde{u} := P_C(\hat{u}^{k+1} - t_{k+1}p_f^k)$, $p_C^k := (\hat{u}^{k+1} - \tilde{u})/t_{k+1} - p_f^k$, and $\bar{\imath}_C^k(\cdot) := \langle p_C^k, \cdot - \tilde{u}\rangle$, where $\tilde{u}$ plays the rôle of u^{k+1}.

4.2. Evaluation errors and relaxed null-step requirements. We now inspect the impact of inexact evaluations on our preceding results, in order to obtain weaker convergence conditions and to provide some practical recommendations.

Our assumption (1.2) on the error tolerance ϵ_f means $\epsilon_f := \sup_{u\in C}[f(u) - f_u] < \infty$. In fact, we need only the weaker condition that $\epsilon_f := \sup_k \epsilon_f^k < \infty$ for the *evaluation errors* $\epsilon_f^k := f(u^k) - f_u^k$ (cf. (2.1)). Thus, for $\epsilon_f := \sup_k \epsilon_f^k$, Theorem 3.5 says that our method produces solutions that are as good as the supplied linearizations.

In fact, the asymptotic accuracy depends *only* on the errors that occur at descent steps. Indeed, at Step 1 we have $\hat{u}^k = u^{k(l)}$ and $f(\hat{u}^k) = f_{\hat{u}}^k + \epsilon_f^{k(l)}$, where $k(l) - 1$ is the iteration number of the lth (i.e., latest) descent step (see Steps 0 and 6). Hence the tolerance ϵ_f in Theorem 3.5(ii) may be replaced by the *asymptotic error*

$$\epsilon_f^\infty := \begin{cases} \epsilon_f^{k(l)} & \text{if only } l < \infty \text{ descent steps occur,} \\ \overline{\lim}_l\, \epsilon_f^{k(l)} & \text{otherwise.} \end{cases} \tag{4.3}$$

In particular, $\epsilon_f^\infty = 0$ if all descent steps happen to be exact. On the other hand, whenever an inexact descent step occurs, then $\epsilon_f^{k+1} := f(u^{k+1}) - f_u^{k+1}$ may potentially determine ϵ_f^∞ (only if $f_u^{k+1} \le f_*$, since $f_{\hat{u}}^\infty \le f_*$ by Theorem 3.5).

Since the asymptotic error is not influenced by the errors occurring at null steps, let us now discuss the case where infinitely many successive null steps occur. Then, by the proof of Lemma 3.2, instead of the requirement $\sup_k \epsilon_f^k < \infty$ (which may be difficult to check for some oracles), it suffices if the following *relaxed null-step requirements* are met:

(a) the sequence $\{g^k\}$ is bounded whenever the sequence $\{u^k\}$ is bounded;

(b) a null step implies that $f_u^{k+1} > f_{\hat{u}}^k - \bar{\kappa} v_k$ for some fixed parameter $\bar{\kappa} \in [\kappa, 1)$.

Condition (a) holds if the mapping $u \mapsto g_u$ is locally bounded on C (cf. Remark 3.6(i)). Condition (b) means that the new linearization f_{k+1} may have any accuracy, as long as it improves the next model sufficiently at u^{k+1}. For $\bar{\kappa} > \kappa$, the oracle may set an indicator $i_{\bar{\kappa}} := 1$ when $\bar{\kappa}$ should replace κ in the descent test (2.10) to accept a shallower null step; $i_{\bar{\kappa}} := 0$ otherwise (i.e., when (2.10) is not modified). Of course, shallow cuts may slow down convergence, but this may be offset by saving the oracle's work per call. To illustrate these requirements, consider the following generalization of the setting of [HeK02].

Example 4.2. Suppose that the objective f has the form $f(\cdot) := \sup_{z\in Z} F_z(\cdot)$ of (1.3) with $F_z(\cdot)$ convex and $\partial F_z(\cdot)$ locally bounded on C, uniformly w.r.t. $z \in Z$. Suppose for each k that the oracle used for approximate evaluation of $f(u^{k+1})$ generates points $z^{(i)} \in Z$, $i = 1, 2, \ldots$, stopping for some i to deliver $f_u^{k+1} := F_{z^{(i)}}(u^{k+1})$ and some $g^{k+1} \in \partial F_{z^{(i)}}(u^{k+1})$. To meet the relaxed null-step requirements, the oracle may stop when $F_{z^{(i)}}(u^{k+1}) > f_{\hat{u}}^k - \bar{\kappa} v_k$ holds, possibly together with other conditions, setting $i_{\bar{\kappa}} := 1$ to force a null step.

Remark 4.3. For an SDP (cf. section 5.6), Example 4.2 accommodates the "inexact null steps" of [HeK02], which can save much work in eigenvalue computations [Hel03, Nay99, Nay06]. In general, when the relaxed null-step requirements are met and the descent steps are exact, then $\epsilon_f^\infty = 0$ in (4.3) and Theorem 3.7 holds (by its proof). In particular, Theorem 3.7 holds for the method of [HeK02].

Insisting that all descent steps be exact may be unrealistic (e.g., as in [Hel03, HeK02, Nay06], where this issue is ignored) or too expensive (cf. [Kiw05]).

For the oracle of Example 4.2, additional stopping criteria may be employed to make a "too inexact" descent step less likely. The general idea is to make the oracle work harder before a descent step is accepted. We distinguish the following two cases.

Case 1. Suppose that the oracle's underestimates $F_{z^{(i)}}(u^{k+1})$ of $f(u^{k+1})$ improve when i grows. Then for a given iteration limit $i_{\max}$ the oracle may stop when either $F_{z^{(i)}}(u^{k+1}) > f_{\hat{u}}^k - \bar{\kappa} v_k$ and $i \le i_{\max}$ (setting $i_{\bar{\kappa}} := 1$ to force a null step), or $F_{z^{(i)}}(u^{k+1}) \le f_{\hat{u}}^k - \kappa v_k$ and $i = i_{\max}$ (setting $i_{\bar{\kappa}} := 0$ for a descent step).

Case 2. In addition to the assumptions of Case 1, suppose that the oracle generates upper bounds $f_{\rm up}^{(i)} \ge f(u^{k+1})$ such that $\bar{f}_{\rm up}^{(i)} - F_{z^{(i)}}(u^{k+1}) \to 0$ if $i \to \infty$. Then the oracle may also stop as soon as for some $i \le i_{\max}$, $\bar{f}_{\rm up}^{(i)} < f_{\hat{u}}^k$, or $f_{\rm up}^{(i)} - F_{z^{(i)}}(u^{k+1}) \le \epsilon_r |F_{z^{(i)}}(u^{k+1})|$ for a given *relative accuracy tolerance* $\epsilon_r > 0$, setting $i_{\bar{\kappa}} := 0$ to promote a descent step.

We add that Case 2 covers oracles employing branch and bound in Lagrangian relaxation of integer programming problems. Then, for difficult Lagrangian subproblems, it pays to use rather loose accuracy requirements, because tighter criteria (e.g., small ϵ_r) may force the oracle to work too long on some calls (see, e.g., [Kiw05]). Fortunately, a typical branch-and-bound oracle generates a good lower bound $F_{z^{(i)}}(u^{k+1})$ quickly (although improving the upper bound $f_{\rm up}^{(i)}$ may need much time). Then the stopping criterion of Case 2 with a moderate tolerance ϵ_r (or another heuristic criterion) may still ensure that the actual error $\epsilon_f^{k+1} := f(u^{k+1}) - f_u^{k+1}$ is small enough. Thus our framework is especially suitable for applications with oracles that deliver reasonably accurate linearizations most of the time, although explicit control of their accuracy might be too costly. (We add that the preceding remarks apply also to the method of [Kiw06b], and they partly explain the good numerical results of [Kiw05].)

4.3. A weaker descent test. As in [Kiw06b, sect. 4.3], at Steps 5 and 6 we may replace the predicted decrease $v_k = t_k|p^k|^2 + \epsilon_k$ (cf. (2.16)) by the smaller quantity $w_k := t_k|p^k|^2/2 + \epsilon_k$. Then the equivalences in Lemma 2.2(vi) are replaced by the fact that

$$w_k \ge -\epsilon_k \quad \Longleftrightarrow \quad \frac{t_k|p^k|^2}{4} \ge -\epsilon_k \quad \Longleftrightarrow \quad w_k \ge \frac{t_k|p^k|^2}{4}.$$

Hence, $w_k \ge -\epsilon_k$ at Step 6 implies $w_k \le v_k \le 3w_k$ and $v_k \ge -\epsilon_k$ for the bounds (2.19)–(2.20), whereas for Step 5, the bound (2.21) is replaced by the fact that

$$V_k < \left(\frac{4\epsilon_{\max}}{t_k}\right)^{1/2} (1 + |\hat{u}^k|) \quad \text{if} \quad w_k < -\epsilon_k.$$

The preceding results extend easily. (In the proof of Lemma 3.2, $f_u^{k+1} > f_{\hat{u}}^k - \kappa w_k$ implies $f_u^{k+1} > f_{\hat{u}}^k - \kappa v_k$, whereas in the proof of Lemma 3.4, $\sum_{k\in K} v_k \le 3 \sum_{k\in K} w_k < \infty$.)

4.4. Linearization accumulation, selection, and aggregation. There are three basic choices of polyhedral models satisfying relation (2.4) rewritten as

$$\max\{\bar{f}_k, f_{k+1}\} \le \check{f}_{k+1} \le f_C. \tag{4.4}$$

First, *accumulation* takes $\check{f}_{k+1} := \max\{\check{f}_k, f_{k+1}\}$, $\check{f}_1 := f_1$; then we may replace f_C by f in (4.4), using the minorizations $\bar{f}_k \le \check{f}_k$ of (2.13) and $f_{k+1} \le f$ of (2.1). In other words, here $\check{f}_k = \max_{j=1}^k f_j$ is the richest model stemming from all the past linearizations, but its storage requirements and QP work per iteration grow with k, so the other choices discussed below are more attractive in practice.

Second, *selection* retains only selected linearizations for its kth model,

$$\check{f}_k(\cdot) := \max_{j\in J_k} f_j(\cdot) \quad \text{with} \quad k \in J_k \subset \{1,\dots,k\}. \tag{4.5}$$

Then $\check{f}_k \le f$ by (2.1), so, in view of (4.4), we need only show how to choose the set J_{k+1} so that $\bar{f}_k \le \check{f}_{k+1}$. Since $p_f^k \in \partial \check{f}_k(\check{u}^{k+1})$ by (2.12) and each f_j is affine in (4.5), there exist multipliers ν_j^k, $j \in J_k$, also known as *convex weights*, such that (cf. [HUL93, Ex. VI.3.4])

$$(p_f^k, 1) = \sum_{j\in J_k} \nu_j^k (\nabla f_j, 1), \quad \nu_j^k \ge 0, \quad \nu_j^k \big[\check{f}_k(\check{u}^{k+1}) - f_j(\check{u}^{k+1})\big] = 0, \quad j \in J_k. \tag{4.6}$$

Then, using relations (2.6) and (4.6), it is easy to obtain the following expansion:

$$(\bar{f}_k, 1) = \sum_{j\in \hat{J}_k} \nu_j^k (f_j, 1) \quad \text{with} \quad \hat{J}_k := \{j \in J_k : \nu_j^k > 0\}. \tag{4.7}$$

In other words, the aggregate linearization $\bar{f}_k$ is a convex combination of the "ordinary" linearizations f_j selected by the *active* set $\hat{J}_k$. Since $\bar{f}_k \le \max_{j\in\hat{J}_k} f_j$, it suffices to choose

$$J_{k+1} \supset \hat{J}_k \cup \{k+1\}. \tag{4.8}$$

Active-set methods for solving subproblem (2.5) [Kiw86, Kiw94] find multipliers ν_j^k such that $|\hat{J}_k| \le n+1$. Hence we can keep $|J_{k+1}| \le \bar{n}$ for any given upper bound $\bar{n} \ge n+2$.

Third, *aggregation* treats the past aggregate linearizations $\bar{f}_j$ like the "ordinary" linearizations f_j, defining $f_{-j} := \bar{f}_j$ for $j = 0\colon k-1$ to replace (4.5) by the aggregate model

$$\check{f}_k(\cdot) := \max_{j\in J_k} f_j(\cdot) \quad \text{with} \quad k \in J_k \subset \{1-k\colon k\}, \quad f_j := \bar{f}_{-j} \text{ for } j \le 0. \tag{4.9}$$

The weights ν_j^k of (4.6) produce $f_{-k} := \bar{f}_k$ via (4.7), and relation (4.8) is replaced by

$$J_{k+1} \supset \{-k, k+1\}, \tag{4.10}$$

so that only $\bar{n} \ge 2$ linearizations may be kept. Formally, if $f_j \le f$ for all $j \in J_k$, then $f_{-k} := \bar{f}_k \le f$ by (4.7); hence, by induction, (4.9)–(4.10) yield (4.4) for all k. Of course, the selection requirement (4.8) may replace (4.10) whenever $|\hat{J}_k| \le \bar{n} - 1$. After a descent step, we can replace (4.8) and (4.10) by $J_{k+1} \ni k+1$ (cf. Remark 3.6(ii)).

Remark 4.4. In the proof of Lemma 3.2, condition (3.1) holds *automatically* for the models discussed above. Indeed, by (4.6) (and induction for aggregation), we have $p_f^k \in \operatorname{co}\{g^j\}_{j=1}^k$ and hence $|p_f^k| \le \max_{j=1}^k |g^j|$, whereas the sequence $\{g^k\}$ is bounded. Similarly, each model $\check{f}_k$ has a bounded subdifferential, as required in Remark 4.1(ii).

5. Lagrangian relaxation.

5.1. The primal problem. Let $\mathcal{Z}$ be a real inner-product space with a finite dimension $\bar{m}$. (We could, of course, always identify $\mathcal{Z}$ with $\mathbb{R}^{\bar{m}}$, but a less concrete approach helps our future development.) In this section we consider the special case where problem (1.1) with $C := \mathbb{R}^n_+$ is the Lagrangian dual problem of the following *primal* convex optimization problem in $\mathcal{Z}$:

$$\psi_0^{\max} := \max\ \psi_0(z) \quad \text{s.t.} \quad \psi_i(z) \geq 0,\ i = 1\colon n,\ z \in Z, \tag{5.1}$$

where $\emptyset \neq Z \subset \mathcal{Z}$ is compact and convex, and each ψ_i is concave and closed (upper semicontinuous) with $\operatorname{dom}\psi_i \supset Z$. The Lagrangian of (5.1) has the form $\psi_0(z) + \langle u, \psi(z)\rangle$, where $\psi := (\psi_1, \ldots, \psi_n)$ and u is a multiplier. Suppose that, at each $u \in C$, the *dual function*

$$f(u) := \max\{\, \psi_0(z) + \langle u, \psi(z)\rangle : z \in Z \,\} \tag{5.2}$$

can be evaluated with *accuracy* $\epsilon_f \geq 0$ by finding a *partial Lagrangian* ϵ_f*-solution*

$$z(u) \in Z \quad \text{such that} \quad f_u := \psi_0(z(u)) + \langle u, \psi(z(u))\rangle \geq f(u) - \epsilon_f. \tag{5.3}$$

Thus f is finite convex and has an ϵ_f-subgradient mapping $g_u := \psi(z(u))$ for $u \in C$. In view of Remark 3.6(i), we suppose that $\psi(z(\cdot))$ is locally bounded on C. (Note that the whole set $\psi(z(C))$ is bounded if $\inf_Z \min_{i=1}^n \psi_i > -\infty$, or the function ψ is continuous on Z.)

5.2. Primal recovery with selection. We first consider our method with linearization selection (cf. section 4.4).

The partial Lagrangian solutions $z^k := z(u^k)$ (cf. (5.3)) and their constraint values $g^k := \psi(z^k)$ determine the linearizations (2.1) as Lagrangian pieces of f in (5.2):

$$f_k(\cdot) = \psi_0(z^k) + \langle \cdot, \psi(z^k)\rangle. \tag{5.4}$$

Using their weights $\{\nu_j^k\}_{j\in J_k}$ (cf. (4.6)), we may estimate a solution to (5.1) via the *aggregate primal solution*

$$\hat{z}^k := \sum_{j\in J_k} \nu_j^k z^j. \tag{5.5}$$

By (4.7), this convex combination is associated with the aggregate linearization $\bar{f}_k$ via

$$(\bar{f}_k, \hat{z}^k, 1) = \sum_{j\in \hat{J}_k} \nu_j^k (f_j, z^j, 1) \quad \text{with} \quad \hat{J}_k := \{j \in J_k : \nu_j^k > 0\}. \tag{5.6}$$

We now derive useful bounds on $\psi_0(\hat{z}^k)$ and $\psi(\hat{z}^k)$, generalizing [Kiw06b, Lem. 5.1].

LEMMA 5.1. *$\hat{z}^k \in Z$, $\psi_0(\hat{z}^k) \geq f_{\hat{u}}^k - \epsilon_k - \langle p^k, \hat{u}^k\rangle$, and $\psi(\hat{z}^k) \geq p_f^k \geq p^k$.*

Proof. By (5.6), $\hat{z}^k \in \operatorname{co}\{z^j\}_{j\in\hat{J}_k} \subset Z$, $\psi_0(\hat{z}^k) \geq \sum_j \nu_j^k \psi_0(z^j)$, and $\psi(\hat{z}^k) \geq \sum_j \nu_j^k \psi(z^j)$ by convexity of Z and concavity of ψ_0, ψ. Since $p_C^k \in \partial i_{\mathbb{R}^n_+}(u^{k+1})$ by (2.12), we have $p_C^k \leq 0$ and $\langle p_C^k, u^{k+1}\rangle = 0$ [HUL93, Ex. III.5.2.6(b)], so $p_f^k = p^k - p_C^k \geq p^k$ by (2.14). Next, using (5.6) with $p_f^k = \nabla \bar{f}_k$ by (2.6) and $\nabla f_j = \psi(z^j)$ by (5.4), we get $\bar{f}_k(0) = \sum_j \nu_j^k \psi_0(z^j)$ and $p_f^k = \sum_j \nu_j^k \psi(z^j)$. Since $\bar{f}_k(0) = \bar{f}_C^k(0) - \bar{\imath}_C^k(0)$ with

$\bar{\imath}_C^k(0) = -\langle p_C^k, u^{k+1}\rangle = 0$ from (2.8), we have $\bar{f}_k(0) = \bar{f}_C^k(0) = f_{\hat{u}}^k - \epsilon_k - \langle p^k, \hat{u}^k\rangle$ by (2.17). Combining the preceding relations yields the conclusion. ∎

In terms of the optimality measure V_k of (2.11), the bounds of Lemma 5.1 imply

$$\hat{z}^k \in Z \quad \text{with} \quad \psi_0(\hat{z}^k) \geq f_{\hat{u}}^k - V_k, \quad \psi_i(\hat{z}^k) \geq -V_k, \quad i = 1\colon n. \tag{5.7}$$

We now show that $\{\hat{z}^k\}$ has cluster points in the set of *ϵ_f-optimal primal solutions* of (5.1),

$$Z_{\epsilon_f} := \{\, z \in Z : \psi_0(z) \geq \psi_0^{\max} - \epsilon_f, \psi(z) \geq 0 \,\}, \tag{5.8}$$

unless this set is empty, i.e., the primal problem is infeasible.

THEOREM 5.2. *Either $f_* = -\infty$ and $f_{\hat{u}}^k \downarrow -\infty$, in which case the primal problem* (5.1) *is infeasible, or $f_* > -\infty$, $f_{\hat{u}}^k \downarrow f_{\hat{u}}^\infty \in [f_* - \epsilon_f, f_*]$, $\overline{\lim}_k f(\hat{u}^k) \leq f_{\hat{u}}^\infty + \epsilon_f$, and $\underline{\lim}_k V_k = 0$. In the latter case, let $K' \subset \mathbb{N}$ be a subsequence such that $V_k \xrightarrow{K'} 0$. Then we have the following:*

(i) *The sequence $\{\hat{z}^k\}_{k\in K'}$ is bounded, and all its cluster points lie in the set Z.*

(ii) *Let $\hat{z}^\infty$ be a cluster point of the sequence $\{\hat{z}^k\}_{k\in K'}$. Then $\hat{z}^\infty \in Z_{\epsilon_f}$.*

(iii) $d_{Z_{\epsilon_f}}(\hat{z}^k) := \inf_{z\in Z_{\epsilon_f}} |\hat{z}^k - z| \xrightarrow{K'} 0$.

Proof. The first assertion follows from Theorem 3.5 (since $f_* = -\infty$ implies primal infeasibility by weak duality). In the second case, using $f_{\hat{u}}^k \downarrow f_{\hat{u}}^\infty \geq f_* - \epsilon_f$ and $V_k \xrightarrow{K'} 0$ in the bounds of (5.7) yields $\underline{\lim}_{k\in K'} \psi_0(\hat{z}^k) \geq f_* - \epsilon_f$ and $\underline{\lim}_{k\in K'} \min_{i=1}^n \psi_i(\hat{z}^k) \geq 0$.

(i) By (5.7), $\{\hat{z}^k\}$ lies in the set Z, which is compact by our assumption.

(ii) We have $\hat{z}^\infty \in Z$, $\psi_0(\hat{z}^\infty) \geq f_* - \epsilon_f$, and $\psi(\hat{z}^\infty) \geq 0$ by the closedness of ψ_0 and ψ. Since $f_* \geq \psi_0^{\max}$ by weak duality (cf. (1.1), (5.1), (5.2)), we get $\psi_0(\hat{z}^\infty) \geq \psi_0^{\max} - \epsilon_f$. Thus $\hat{z}^\infty \in Z_{\epsilon_f}$ by the definition (5.8).

(iii) This follows from (i), (ii), and the continuity of the distance function $d_{Z_{\epsilon_f}}$. ∎

Remark 5.3. (i) For Theorem 5.2, we can replace ϵ_f in (5.8) by ϵ_f^∞ (cf. (4.3)).

(ii) By the proofs of Lemma 2.3 and Theorem 5.2, if an infinite cycle between Steps 2 and 5 occurs, then $V_k \to 0$ yields $d_{Z_{\epsilon_f}}(\hat{z}^k) \to 0$. Similarly, if Step 4 terminates with $V_k = 0$, then $\hat{z}^k \in Z_{\epsilon_f}$. In both cases, we can replace ϵ_f with ϵ_f^∞ (cf. (4.3)).

(iii) Given a tolerance $\epsilon_{\text{tol}} > 0$, the method may stop if

$$\psi_0(\hat{z}^k) \geq f_{\hat{u}}^k - \epsilon_{\text{tol}} \quad \text{and} \quad \psi_i(\hat{z}^k) \geq -\epsilon_{\text{tol}}, \quad i = 1\colon n.$$

Then $\psi_0(\hat{z}^k) \geq \psi_0^{\max} - \epsilon_f - \epsilon_{\text{tol}}$ from $f_{\hat{u}}^k \geq f_* - \epsilon_f$ (cf. (2.2)) and $f_* \geq \psi_0^{\max}$ (weak duality), so that the point $\hat{z}^k \in Z$ is an approximate primal solution of (5.1). This stopping criterion will be satisfied for some k if $f_* > -\infty$ (cf. (5.7) and Theorem 5.2).

5.3. Primal recovery with aggregation. Let us now consider the variant with aggregation based on (4.9), where each linearization f_j has an associated primal point z^j, with $f_j := \bar{f}_{-j}$ and $z^j := \hat{z}^{-j}$ for $j < 0$. Letting $z^0 := z^1$, suppose for induction that $(f_j, z^j) \in \text{co}\{(f_i, z^i)\}_{i=0}^{|j|}$ for $j \in J_k$. For the convex weights ν_j^k satisfying (4.7), let $z^{-k} := \hat{z}^k$ for the aggregate primal solution $\hat{z}^k$ given by (5.6). Since a convex combination of convex combinations of given points is a convex combination of those points, we deduce the existence of convex weights $\bar{\nu}_j^k$ such that

$$(f_{-k}, z^{-k}, 1) := (\bar{f}_k, \hat{z}^k, 1) = \sum_{0\leq j\leq k} \bar{\nu}_j^k (f_j, z^j, 1) \quad \text{with} \quad \bar{\nu}_j^k \geq 0,\ j = 0\colon k. \tag{5.9}$$

In other words, $(f_{-k}, z^{-k}) \in \mathrm{co}\{(f_i, z^i)\}_{i=0}^k$, as required for induction. Replacing (5.6) by (5.9) for Lemma 5.1, we conclude that the preceding convergence results remain valid.

5.4. Handling primal equality constraints. Consider the primal problem (5.1) with additional equality constraints of the form

$$\psi_0^{\max} := \max\ \psi_0(z) \quad \text{s.t.} \quad \psi_{\mathcal{I}}(z) \ge 0,\ \psi_{\mathcal{E}}(z) = 0,\ z \in Z, \tag{5.10}$$

where $\mathcal{I} \cup \mathcal{E} = \{1\colon n\}$, $\mathcal{I} \cap \mathcal{E} = \emptyset$, and $\psi_{\mathcal{E}}$ is affine. For $C := \mathbb{R}_+^{|\mathcal{I}|} \times \mathbb{R}^{|\mathcal{E}|}$, the final bound in Lemma 5.1 becomes $\psi_{\mathcal{I}}(\hat{z}^k) \ge p_{f,\mathcal{I}}^k \ge p_{\mathcal{I}}^k$, $\psi_{\mathcal{E}}(\hat{z}^k) = p_{f,\mathcal{E}}^k = p_{\mathcal{E}}^k$ (using $p_{C,\mathcal{I}}^k \le 0$, $p_{C,\mathcal{E}}^k = 0$, $\langle p_C^k, u^{k+1}\rangle = 0$ as before); the final inequalities in (5.7) are replaced by $\min_{i\in\mathcal{I}} \psi_i(\hat{z}^k) \ge -V_k$, $\max_{i\in\mathcal{E}} |\psi_i(\hat{z}^k)| \le V_k$, and $\psi(z) \ge 0$ in (5.8) by $\psi_{\mathcal{I}}(z) \ge 0$, $\psi_{\mathcal{E}}(z) = 0$. With these replacements, the proof of Theorem 5.2 extends easily (since $\underline{\lim}_{k\in K'} \max_{i\in\mathcal{E}} |\psi_i(\hat{z}^k)| = 0$ yields $\psi_{\mathcal{E}}(\hat{z}^\infty) = 0$ in (ii)).

Remark 5.4. We add that the ideas of sections 4.2, 5.3, and 5.4 can be translated into additional properties of the method of [Kiw06b]. Further, a simplified variant of the latter method is obtained by modifying relations (2.5)–(2.8) as follows. Letting u^{k+1} solve the prox subproblem (2.3), for the subgradients $p_f^k \in \partial \check{f}_k(u^{k+1})$ and $p_C^k \in \partial i_C(u^{k+1})$ such that $p_f^k + p_C^k = (\hat{u}^k - u^{k+1})/t_k$, define $\bar{f}_k$ by (2.6) with $\check{u}^{k+1} := u^{k+1}$ and $\bar{\imath}_C^k$ by (2.8). Then Lemma 2.2 holds by construction, and the proof of Lemma 3.2 simplifies to that of [Kiw06b, Lem. 3.3]. In effect, except for section 4.1, all the preceding results hold for this variant as well.

5.5. Nonpolyhedral objective models. In addition to the assumptions of section 5.1, suppose ψ is affine: $\psi(z) := b - Az$ for some given $b \in \mathbb{R}^n$ and a linear mapping $A : \mathcal{Z} \to \mathbb{R}^n$. Then the Lagrangian of (5.1) has the form

$$L(z,u) := \psi_0(z) + \langle u, \psi(z)\rangle = \psi_0(z) + \langle u, b - Az\rangle \tag{5.11}$$

and $f(\cdot) := \max_{z\in Z} L(z,\cdot)$. Suppose Step 1 selects the (possibly) *nonpolyhedral* model

$$\check{f}_k(\cdot) := \max_{z\in Z_k} L(z,\cdot) \quad \text{with} \quad z^k \in Z_k \subset Z, \tag{5.12}$$

where the set Z_k is closed convex. Since $f_k(\cdot) = L(z^k,\cdot)$ by (5.4), we have $f_k \le \check{f}_k \le f$. Thus, to meet the requirement of (4.4), we need only show how to choose a set $Z_{k+1} \ni z^{k+1}$ so that $\bar{f}_k \le \check{f}_{k+1}$. First, for solving subproblem (2.5) with the model $\check{f}_k$ given by (5.12), we employ the *Lagrangian* $\bar{L} : \mathbb{R}^n \times Z_k \to \mathbb{R}$ of subproblem (2.5) defined by

$$\bar{L}(u,z) := L(z;u) + \langle p_C^{k-1}, u - u^k\rangle + \frac{1}{2t_k}|u - \hat{u}^k|^2, \tag{5.13}$$

so that

$$\phi_f^k(\cdot) = \max\big\{\bar{L}(\cdot,z) : z \in Z_k\big\}. \tag{5.14}$$

For each primal point $z \in Z_k$, the (unique) *Lagrangian solution*

$$u_z := \arg\min \bar{L}(\cdot,z) = \hat{u}^k - t_k\big[\psi(z) + p_C^{k-1}\big] \tag{5.15}$$

substituted for u in (5.13) gives the value of the *dual function* $q : Z_k \to \mathbb{R}$ defined by

(5.16)
$$q(z) := \min \bar{L}(\cdot, z) = \psi_0(z) + \langle \psi(z), \hat{u}^k \rangle + \langle p_C^{k-1}, \hat{u}^k - u^k \rangle - \frac{t_k}{2} |\psi(z) + p_C^{k-1}|^2.$$

Since q is closed and Z_k is compact, the *dual problem* $\max_{Z_k} q$ has at least one solution:

$$\hat{z}^k \in \operatorname{Arg\,max} \{ q(z) : z \in Z_k \}. \tag{5.17}$$

LEMMA 5.5. *Given a dual solution $\hat{z} := \hat{z}^k$ of (5.17), define the Lagrangian solution $\check{u} := u_{\hat{z}}$ by (5.15). Then we have the following statements:*

(i) *The pair $(\check{u}, \hat{z})$ is a saddle-point of the Lagrangian $\bar{L}$ defined by (5.13):*

$$\bar{L}(\check{u}, z) \le \bar{L}(\check{u}, \hat{z}) \le \bar{L}(u, \hat{z}) \quad \forall u \in \mathbb{R}^n, z \in Z_k. \tag{5.18}$$

(ii) *For $\check{u}^{k+1}$, $\bar{f}_k$, and p_f^k defined by (2.5)–(2.6), we have $\check{u}^{k+1} = \check{u}$, $p_f^k = \psi(\hat{z}^k)$,*

$$\check{u}^{k+1} = \hat{u}^k - t_k \big[\psi(\hat{z}^k) + p_C^{k-1} \big], \tag{5.19}$$

$$\bar{f}_k(\cdot) = \psi_0(\hat{z}^k) + \langle \cdot, \psi(\hat{z}^k) \rangle. \tag{5.20}$$

Proof. (i) $\bar{L}$ is convex-concave on $\mathbb{R}^n \times Z_k$, Z_k is compact, and for each $z \in Z_k$, $\bar{L}(u, z) \to \infty$ when $|u| \to \infty$. Hence $\bar{L}$ has a saddle-point $(\bar{u}, \bar{z})$ [HUL93, Thm. VII.4.3.1]. Since $\hat{z} \in \operatorname{Arg\,max}_{Z_k} \min_u \bar{L}(u, \cdot)$ by (5.16)–(5.17), $(\bar{u}, \hat{z})$ is a saddle-point as well [HUL93, Thm. VII.4.2.5]. Then $\bar{L}(\bar{u}, \hat{z}) \le \bar{L}(u, \hat{z})$ $\forall u$ yields $\bar{u} = u_{\hat{z}} = \check{u}$ by (5.15), so that (5.18) holds.

(ii) By (2.5) and (5.14), (5.18) implies $\check{u}^{k+1} = \check{u}$ [HUL93, Thm. VII.4.2.5]. Then (2.6) and (5.15) with $z = \hat{z}$ yield $p_f^k = \psi(\hat{z}^k)$. The left inequality in (5.18) combined with (5.11)–(5.13) gives $\check{f}_k(\check{u}^{k+1}) = \psi_0(\hat{z}^k) + \langle \check{u}^{k+1}, \psi(\hat{z}^k) \rangle$, and then (2.6) yields (5.20). ∎

In view of (5.12) and (5.20), the requirement of (4.4) is met if the set Z_{k+1} satisfies

$$Z_{k+1} \supset \{ \hat{z}^k, z^{k+1} \}, \tag{5.21}$$

in addition to being a closed convex subset of Z. Further, condition (3.1) holds (with $p_f^k = \psi(\hat{z}^k)$, $\hat{z}^k \in Z_k$, Z_k compact, ψ continuous), and the aggregate representation (5.20) can be seen as a special case of (5.6) (with $\hat{J}_k := \{k\}$ and z^k replaced by $\hat{z}^k$ in (5.4)). In effect, the results of section 5.2 hold for this variant as well.

Remark 5.6. (i) We add that for $p_f^k = \psi(\hat{z}^k)$ (and $C := \mathbb{R}^n_+$), (2.7)–(2.8) simplify to

$$u^{k+1} = \max \{ \hat{u}^k - t_k(b - A\hat{z}^k), 0 \} \quad \text{and} \quad p_C^k = \min \left\{ \frac{1}{t_k} \hat{u}^k - b + A\hat{z}^k, 0 \right\}. \tag{5.22}$$

In general, $\langle p_C^{k-1}, u^k \rangle = 0$ from $p_C^{k-1} \in \partial i_C(u^k)$, so we can omit u^k in (5.13) and (5.16). A dual interpretation of (5.22) follows. Since $i_C(\cdot) = \sup\{-\langle \eta, \cdot \rangle : \eta \in \mathbb{R}^n_+\}$, using a dual variable $\eta \in \mathbb{R}^n_+$ for subproblem (2.3), its Lagrangian $\bar{L}(u, z, \eta)$, relaxed solution $u_{z,\eta}$, and dual function $q(z, \eta)$ are given by (5.13), (5.15), and (5.16) with p_C^{k-1} replaced by $-\eta$. Let $\eta^k := -p_C^{k-1}$. The dual problem $\max_{Z_k \times \mathbb{R}^n_+} q$ is treated in a Gauss–Seidel fashion by finding $\hat{z}^k \in \operatorname{Arg\,max}_{Z_k} q(\cdot, \eta^k)$ (cf. (5.17)) and then $\eta^{k+1} := \arg\max_{\mathbb{R}^n_+} q(\hat{z}^k, \cdot)$, for which $u^{k+1} = u_{\hat{z}^k, \eta^{k+1}}$ and $\eta^{k+1} = -p_C^k$ by (5.22).

Thus alternating linearizations of subproblem (2.3) correspond to coordinatewise maximizations of its dual function.

(ii) Suppose that ψ_0 is linear and $Z_k := \mathrm{co}\{z^j\}_{j=1}^k$. Then $z \in Z_k$ iff $z = \sum_j \nu_j z^j$ for a weight vector ν in $N := \{\nu \in \mathbb{R}^k_+ : \sum_j \nu_j = 1\}$. For $F := [\psi_0(z^1), \ldots, \psi_0(z^k)]$ and $G := [g^1, \ldots, g^k]$, we have $\psi_0(z) = F\nu$ and $\psi(z) = G\nu$. Using these representations in (5.16)–(5.17), we may take $\hat{z}^k = \sum_j \nu^k_j z^j$ for any solution ν^k to the dual QP subproblem

$$\nu^k \in \mathrm{Arg\,max}\left\{F\nu + \nu^T G^T \hat{u}^k - \frac{t_k}{2}|G\nu - p_C^{k-1}|^2 : \nu \in N\right\}. \tag{5.23}$$

In effect, our framework comprises the method of [FGRS06, sect. 3.2], which requires exact evaluations. Note that the similarity of $\hat{z}^k$ above to (5.5) is not accidental: the model (5.12) with $Z_k := \mathrm{co}\{z^j\}_{j=1}^k$ is *equivalent* to the polyhedral model (4.5) with $J_k := \{1\colon k\}$ (cf. (5.11) and (5.4)). Other choices of J_k from section 4.4 correspond to $Z_k := \mathrm{co}\{z^j\}_{j\in J_k}$.

(iii) For problem (5.10) with mixed constraints, formula (5.22) is valid for components indexed by $\mathcal{I}$, whereas $u^{k+1}_{\mathcal{E}} = \hat{u}^k_{\mathcal{E}} - t_k(b - A\hat{z}^k)_{\mathcal{E}}$ and $p^k_{C,\mathcal{E}} = 0$. Then the setting of (ii) above comprises the method of [ReS06, sect. 3] (for exact evaluations).

(iv) By Remark 4.1, the results of section 5.2 hold when Step 5′ is used as well, since each $\check{f}_k$ has bounded subgradients (by (5.11)–(5.12) and the compactness of $Z_k \subset Z$).

5.6. SDP via eigenvalue optimization. To discuss applications in SDP, we need the following notation.

We consider the Euclidean space S^m of $m \times m$ real symmetric matrices with the Frobenius inner product $\langle x, y\rangle = \mathrm{tr}\, xy$ (we use lowercase notation for the elements of S^m for consistency with the rest of the text). S^m_+ is the cone of positive semidefinite matrices. The maximum eigenvalue $\lambda_{\max}(y)$ of a matrix $y \in S^m$ and its positive part $\lambda^+_{\max}(y) := \max\{\lambda_{\max}(y), 0\}$ satisfy (see, e.g., [LeO96, Tod01])

$$\lambda_{\max}(y) = \max\big\{\langle y, x\rangle : x \in \Sigma^m\big\} \quad \text{with} \quad \Sigma^m := \big\{x \in S^m_+ : \mathrm{tr}\, x = 1\big\}, \tag{5.24a}$$

$$\lambda^+_{\max}(y) = \max\big\{\langle y, x\rangle : x \in \Sigma^m_{\leq}\big\} \quad \text{with} \quad \Sigma^m_{\leq} := \big\{x \in S^m_+ : \mathrm{tr}\, x \leq 1\big\}. \tag{5.24b}$$

Let $a > 0$, $b \in \mathbb{R}^n$, $c \in S^m$, and $A : S^m \to \mathbb{R}^n$ be linear. Consider the *SDPs*

$$(P_=): \quad \max \langle c, x\rangle \quad \text{s.t.} \quad Ax \leq b, \quad x \in S^m_+, \quad \mathrm{tr}\, x = a, \tag{5.25}$$

$$(P_{\leq}): \quad \max \langle c, x\rangle \quad \text{s.t.} \quad Ax \leq b, \quad x \in S^m_+, \quad \mathrm{tr}\, x \leq a. \tag{5.26}$$

Any SDP can be formulated as $(P_{\leq})$ without the final trace condition. If we know or simply guess an upper bound a on the trace of some optimal solution, we may use $(P_{\leq})$. (For a wrong guess, our method will produce dual values going to $-\infty$, thus indicating primal infeasibility.) Of course, $(P_{\leq})$ can be formulated as $(P_=)$ by adding a slack variable, but this is not really necessary, since our method can handle both. $(P_=)$ is natural in many combinatorial applications, where the trace of all feasible solutions is known [HeR00]; $(P_{\leq})$ is employed in [Nay06] for equality-constrained SDPs.

We can regard $(P_=)$ as an instance of (5.1) with $\mathcal{Z} := S^m$, $\psi_0(z) := \langle c, z\rangle$, $\psi(z) := b - Az$, and $Z := a\Sigma^m$. Then, by (5.2) and (5.24a), the dual function f satisfies

$$f(u) = a\lambda_{\max}(c - A^*u) + \langle b, u\rangle \quad \forall\, u, \tag{5.27}$$

where A^* is the adjoint of A (defined by $\langle z, A^*u\rangle = \langle Az, u\rangle \ \forall\ z \in S^m,\ u \in \mathbb{R}^n$). For each u, the approximate evaluation condition (5.3) is met by $z(u) := ar(u)r(u)^T$, where $r(u) \in \mathbb{R}^m$ is an (ϵ_f/a)-*eigenvector* of the matrix $s(u) := c - A^*u \in S^m$ satisfying

$$r(u)^T s(u) r(u) \geq \lambda_{\max}(s(u)) - \frac{\epsilon_f}{a}, \qquad r(u)^T r(u) = 1. \tag{5.28}$$

Then the ϵ_f-subgradient mapping $u \to g_u := \psi(z(u)) = b - Az(u)$ is bounded on $\mathbb{R}^n$.

Thus we can use the setting of section 5.5 with models $\check{f}_k$ given by (5.12) for sets Z_k satisfying (5.21). In effect, the results of section 5.2 and Remark 5.6 hold for this variant as well.

Remark 5.7. (i) Our dual problem $f_* := \inf_C f$ is equivalent to the standard dual of $(P_=)$, which is strictly feasible. Hence (cf. [Tod01, Thm. 4.1]) if $(P_=)$ is feasible, then its optimal value is finite and equals f_*, although the dual problem need not have solutions. Thus, even for exact evaluations, Theorem 5.2 improves upon [Hel04, Thm. 3.6], which assumes that $\operatorname{Arg\,min}_C f \neq \emptyset$. We show elsewhere [Kiw06a] how to extend a related result of [Hel04, Thm. 4.8], without assuming that $\operatorname{Arg\,min}_C f$ is nonempty and bounded.

(ii) Condition (5.28) is particularly useful when approximate eigenvectors are found by iterative methods (such as the Lanczos method [Hel03, Nay06]) that employ only matrix-vector multiplications to exploit the structure of the matrix $s(u) := c - A^*u$. This condition has the following meaning in the setting of Example 4.2 with $u = u^{k+1}$, $s^{k+1} := s(u^{k+1})$. Suppose that an iterative method generates approximate eigenvectors $r^{(i)} \in \mathbb{R}^m$, $|r^{(i)}| = 1$, $i = 1, 2, \ldots$, stopping for some i to deliver $z^{k+1} := ar^{(i)}r^{(i)^T}$. To meet the relaxed null-step requirements, the method may stop when $ar^{(i)^T}s^{k+1}r^{(i)} + \langle b, u^{k+1}\rangle > f_{\hat{u}}^k - \bar{\kappa}v_k$. If a descent step occurs, then $\epsilon_f^{k+1} = a\lambda_{\max}(s^{k+1}) - ar^{(i)^T}s^{k+1}r^{(i)}$ may potentially determine the asymptotic error ϵ_f^∞ of (4.3). To ensure that ϵ_f^{k+1} is not "too large," we can employ additional stopping criteria based on upper estimates of $\lambda_{\max}(s^{k+1})$ generated as in [Nay06].

(iii) We may employ the following choice of the set Z_k due to [Nay99, Nay06]:

$$Z_k := \left\{ \sum_{j=1}^{\check{j}} \nu_j \check{z}^j + pvp^T : \nu \in \mathbb{R}^{\check{j}}_+, v \in S^r_+, \sum_{j=1}^{\check{j}} \nu_j + \operatorname{tr} v = a \right\}, \tag{5.29}$$

where each $\check{z}^j \in \Sigma^m$ and p is an $m \times r$ orthonormal matrix. The resulting model

$$\check{f}_k(u) = a \max\Big\{ \max_{j=1:\ \check{j}} \langle c - A^*u, \check{z}^j\rangle, \lambda_{\max}\big(p^T(c - A^*u)p\big) \Big\} + \langle b, u\rangle \tag{5.30}$$

attempts to strike a balance between being easy to handle (the polyhedral part) and accurate enough for fast convergence (the semidefinite part). Then the dual subproblem (5.17) can be cast as a conic optimization problem and handled by specialized solvers. Two efficient updates of Z_k satisfying (5.21) are given in [Nay99, sect. 4.4.2] (although they update AZ_k, they can update Z_k as well). For $\check{j} = 1$, (5.29) reduces to the original choice of [HeR00]; again, (5.17) can be solved efficiently as a quadratic SDP [HeK02], and efficient updates of Z_k are given in [Hel03, HeK02].

(iv) For problem $(P_\leq)$ of (5.26), we can take $Z := a\Sigma^m_\leq$. Then (cf. (5.24)), $\lambda^+_{\max}$ replaces $\lambda_{\max}$ in (5.27), and we can take $r(u) := 0$ if $\lambda_{\max}(s(u)) < 0$, using (5.28) otherwise. We can thus stop an iterative eigenvalue computation whenever an upper bound indicates that $\lambda_{\max}(s(u)) < 0$. Of course, the final "=" in (5.29) is replaced by "$\leq$".

Acknowledgments. I would like to thank the Associate Editor, the two anonymous referees, and Claude Lemaréchal for helpful comments.

REFERENCES

[FGRS06] I. Fischer, G. Gruber, F. Rendl, and R. Sotirov, *Computational experience with a bundle approach for semidefinite cutting plane relaxations of Max-Cut and equipartition*, Math. Program., 105 (2006), pp. 451–469.

[HeK02] C. Helmberg and K. C. Kiwiel, *A spectral bundle method with bounds*, Math. Program., 93 (2002), pp. 173–194.

[Hel03] C. Helmberg, *Numerical evaluation of SBmethod*, Math. Program., 95 (2003), pp. 381–406.

[Hel04] C. Helmberg, *A cutting plane algorithm for large scale semidefinite relaxations*, The Sharpest Cut, The Impact of Manfred Padberg and His Work, M. Grötschel, ed., MPS-SIAM Ser. Optim. 4, SIAM, Philadelphia, 2004, pp. 233–256.

[HeR00] C. Helmberg and F. Rendl, *A spectral bundle method for semidefinite programming*, SIAM J. Optim., 10 (2000), pp. 673–696.

[Hin01] M. Hintermüller, *A proximal bundle method based on approximate subgradients*, Comput. Optim. Appl., 20 (2001), pp. 245–266.

[HUL93] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms*, Springer, Berlin, 1993.

[Kiw85] K. C. Kiwiel, *An algorithm for nonsmooth convex minimization with errors*, Math. Comp., 45 (1985), pp. 173–180.

[Kiw86] K. C. Kiwiel, *A method for solving certain quadratic programming problems arising in nonsmooth optimization*, IMA J. Numer. Anal., 6 (1986), pp. 137–152.

[Kiw90] K. C. Kiwiel, *Proximity control in bundle methods for convex nondifferentiable minimization*, Math. Programming, 46 (1990), pp. 105–122.

[Kiw94] K. C. Kiwiel, *A Cholesky dual method for proximal piecewise linear programming*, Numer. Math., 68 (1994), pp. 325–340.

[Kiw95] K. C. Kiwiel, *Approximations in proximal bundle methods and decomposition of convex programs*, J. Optim. Theory Appl., 84 (1995), pp. 529–548.

[Kiw99] K. C. Kiwiel, *A projection-proximal bundle method for convex nondifferentiable minimization*, in Ill-posed Variational Problems and Regularization Techniques, M. Théra and R. Tichatschke, eds., Lecture Notes in Econom. Math. Systems 477, Springer-Verlag, Berlin, 1999, pp. 137–150.

[Kiw05] K. C. Kiwiel, *An Inexact Bundle Approach to Cutting-Stock Problems*, Technical report, Systems Research Institute, Warsaw, 2005.

[Kiw06a] K. C. Kiwiel, *Inexact Dynamic Bundle Methods*, Technical report, Systems Research Institute, Warsaw, 2006.

[Kiw06b] K. C. Kiwiel, *A proximal bundle method with approximate subgradient linearizations*, SIAM J. Optim., 16 (2006), pp. 1007–1023.

[KRR99] K. C. Kiwiel, C. H. Rosa, and A. Ruszczyński, *Proximal decomposition via alternating linearization*, SIAM J. Optim., 9 (1999), pp. 668–689.

[Lem01] C. Lemaréchal, *Lagrangian relaxation*, in Computational Combinatorial Optimization, M. Jünger and D. Naddef, eds., Lecture Notes in Comput. Sci. 2241, Springer-Verlag, Berlin, 2001, pp. 112–156.

[LeO96] A. S. Lewis and M. L. Overton, *Eigenvalue optimization*, Acta Numer., 5 (1996), pp. 149–190.

[Mil01] S. A. Miller, *An Inexact Bundle Method for Solving Large Structured Linear Matrix Inequalities*, Ph.D. thesis, Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA, 2001.

[Nay99] M. V. Nayakkankuppam, *Optimization Over Symmetric Cones*, Ph.D. thesis, Department of Computer Science, New York University, New York, NJ, 1999.

[Nay06] M. V. Nayakkankuppam, *Solving Large-scale Semidefinite Programs in Parallel*, Math. Program., (2006), to appear.

[ReS06] F. Rendl and R. Sotirov, *Bounds for the quadratic assignment problem using the bundle method*, Math. Program., (2006), to appear.

[Sol03] M. V. Solodov, *On approximations with finite precision in bundle methods for nonsmooth optimization*, J. Optim. Theory Appl., 119 (2003), pp. 151–165.

[Tod01] M. J. Todd, *Semidefinite optimization*, Acta Numer., 10 (2001), pp. 515–560.

SIAM J. OPTIM.
Vol. 17, No. 4, pp. 1035–1055
© 2006 Society for Industrial and Applied Mathematics

TRANSPOSITION THEOREMS AND QUALIFICATION-FREE OPTIMALITY CONDITIONS*

HERMANN SCHICHL† AND ARNOLD NEUMAIER†

Abstract. New theorems of the alternative for polynomial constraints (based on the Positivstellensatz from real algebraic geometry) and for linear constraints (generalizing the transposition theorems of Motzkin and Tucker) are proved. Based on these, two Karush–John optimality conditions—holding without any constraint qualification—are proved for single- or multiobjective constrained optimization problems. The first condition applies to polynomial optimization problems only, and gives for the first time necessary and sufficient global optimality conditions for polynomial problems. The second condition applies to smooth local optimization problems and strengthens known local conditions. If some linear or concave constraints are present, the new version reduces the number of constraints for which a constraint qualification is needed to get the Kuhn–Tucker conditions.

Key words. certificate of global optimality, first order optimality conditions, Fritz John conditions, Karush–John conditions, global optimality condition, global optimization, Kuhn–Tucker conditions, Mangasarian–Fromovitz constraint qualification, necessary and sufficient conditions, Positivstellensatz, second order optimality conditions, theorem of the alternative, transposition theorem

AMS subject classification. 90C30

DOI. 10.1137/05063129X

1. Introduction.

In this paper, we present a number of theorems that are useful for the global analysis of optimization problems, i.e., the assessment of their feasibility, and the construction and verification of a global solution. Several of the results are, however, also relevant for local optimization.

In constrained optimization, first and second order optimality conditions play a central role, as they give necessary and/or sufficient conditions for a point to attain a local or global minimum of the problem considered, and thus define the goals that numerical methods should try to satisfy.

The various conditions currently available usually depend on qualitative conditions (concerning smoothness, linearity, convexity, etc.) that delineate the problem class, and on technical conditions, so-called constraint qualifications, that allow one to avoid certain difficulties in proofs or certain known counterexamples.

The proof of the optimality conditions depends crucially on the availability of certain theorems of the alternative, which state that among two alternative existence statements, exactly one can be satisfied. Thus a theorem of the alternative may serve to define certificates whose presence implies the solvability of one alternative and the unsolvability of the other alternative.

Recent advances in global optimization [27, 30] make it possible in many cases to find and verify the global optimality of a solution, or to verify that no feasible point exists. Certificates acquire in this case a special importance, particularly in the context of computer-assisted proofs.

However, in order to apply a necessary optimality condition to rule out candidate solutions, or a sufficient optimality condition to verify the existence of a solution,

*Received by the editors May 11, 2005; accepted for publication (in revised form) May 15, 2006; published electronically December 5, 2006.
http://www.siam.org/journals/siopt/17-4/63129.html
†Fakultät für Mathematik, Universität Wien, Nordbergstr. 15, A-1090 Wien, Austria (Hermann.Schichl@esi.ac.at, Arnold.Neumaier@univie.ac.at; http://www.mat.univie.ac.at/∼neum/).

it is important that these conditions are valid under conditions that can be checked explicitly. Therefore the optimality conditions should not depend on any constraint qualification.

Optimality conditions characterizing the solutions of smooth nonlinear programming problems by first order necessary conditions are often called Fritz John conditions if they apply without constraint qualification, after Fritz John [15], who rediscovered unpublished earlier results of Karush [16]; see Kuhn [18, section 6] for a history. Therefore, we shall refer to such conditions as *Karush–John optimality conditions*.

The importance of the Karush–John conditions stems from the fact that they apply without any hypothesis on the optimization problem (apart from smoothness). For the known (second order) sufficient conditions, a similar result was not known before, sufficiency requiring very strong nondegeneracy conditions. It is therefore remarkable that, *for polynomial optimization problems*, it is possible to formulate necessary and sufficient conditions for (global) optimality, valid without any restriction. These strong results are based on the so-called Positivstellensatz, a polynomial analogue of the transposition theorem for linear systems. The Positivstellensatz is a highly nontrivial tool from real algebraic geometry which has been applied recently also in an algorithmic way for the solution of global polynomial optimization problems. Some of the consequences of the Positivstellensatz are implemented in the packages GloptiPoly (Henrion and Lasserre [12, 13, 14]) and SOSTOOLS (Prajna, Papachristodoulou, and Parrilo [32]).

Related results in this direction are in Lasserre [20]. He proved in his Theorem 4.2 a sufficient condition for global optimality in polynomial optimization problems, which is a special case of our necessary and sufficient conditions. (The unconstrained minimization of the Motzkin polynomial shows that Lasserre's condition is not sufficient.) He shows that his certificates can be interpreted as polynomial multipliers in a fashion analogous to the Kuhn–Tucker optimality conditions. Instead of necessary conditions he obtains under some compactness assumption an infinite sequence of semidefinite relaxations whose optimal values converge to the global optimum.

In this article we derive in section 2 polynomial transposition theorems and deduce from them a global Karush–John condition which is a necessary and sufficient condition for global optimality of *polynomial* programs.

Section 3 then proves a very general transposition theorem for *linear* constraints, establishing a theorem of the alternative from which the transposition theorems of Motzkin [24] and of Tucker [34] (as well as many weaker ones) can be obtained as corollaries. This level of generality is necessary to deduce in section 4 a form of the constraint qualifications for the Kuhn–Tucker optimality conditions for general *smooth nonlinear* programming problems which is stronger (i.e., makes more stringent assertions about the multipliers) than the known Karush–John conditions and also applies for multiple objectives.

Our Karush–John conditions imply derived Kuhn–Tucker conditions with linear independence constraint qualifications for fewer constraints than the conditions found in the literature. In particular, they imply the known result that for concavely (or linearly) constrained problems no constraint qualification is needed.

The new conditions will be incorporated in the COCONUT environment [9] for deterministic global optimization; the local Karush–John conditions from section 4 are already in place.

Notation. In the following, $\mathbb{R}$ is the field of real numbers, and $\mathbb{N}_0$ the set of nonnegative integers. To denote monomials and their degrees, we use the multi-index

notation

$$x^\alpha = x_1^{\alpha_1} \dots x_n^{\alpha_n}, \quad |\alpha| = \alpha_1 + \dots + \alpha_n$$

($x \in \mathbb{R}^n, \alpha \in \mathbb{N}_0^n$). Inequalities ($\leq$, $\geq$) and strict inequalities ($<$, $>$) between vectors and matrices are interpreted componentwise. However, disequality ($\neq$) is the negation of equality ($=$) and hence not interpreted componentwise. The infimum $\inf\{x, y\}$ of two vectors x, y of the same size is taken in the partial order $\leq$, and is equivalent to the componentwise minimum. In particular, the condition $\inf\{x, y\} = 0$ is just the complementarity condition $x \geq 0, y \geq 0, x_i y_i = 0$ for all i. By e we denote a column vector of arbitrary size all of whose entries have the value 1. $[A, B]$ denotes the $m \times (n+p)$-matrix formed by juxtaposition of the $m \times n$-matrix A and the $m \times p$-matrix B. Zero dimensional vectors and matrices (needed to avoid stating many special cases) are handled according to the conventions in de Boor [7]; in addition, any of the relations $=$, $<$, $\leq$ (but not $\neq$) between zero dimensional objects is considered to be valid.

2. Global optimality conditions for polynomials. It is well known that first order (Kuhn–Tucker) optimality conditions for constrained (single-objective) optimization are sufficient for convex problems, but not in general. For nonconvex problems, they must be complemented by second order conditions, which come in two forms—as necessary conditions and as sufficient conditions—and they apply to local optimality only. Moreover, between necessary and sufficient conditions is a theoretical gap, in which various degenerate exceptional situations are possible. It is therefore remarkable that, for polynomial systems, it is possible to bridge this gap and formulate necessary and sufficient conditions for (global) optimality, valid without any restriction.

The following discussion is based on a polynomial analogue of the transposition theorem (Theorem 3.4), the so-called Positivstellensatz, a highly nontrivial result from real algebraic geometry. To present this result, we need some definitions.

$\mathbb{N}_0$ denotes the set of nonnegative integers. $\mathbb{R}[x_{1:n}] := \mathbb{R}[x_1, \dots, x_n]$ denotes the algebra of polynomials in the indeterminates $x_1, \dots, x_n$ with real coefficients. Let $R_i \in \mathbb{R}[x_{1:n}]$ $(i = 1 : k)$ be a finite family of polynomials, combined in the vector $R = (R_1, \dots, R_k)^T$. The *ideal* generated by the R_i is the vector space

$$I\langle R\rangle = I\langle R_1, \dots, R_k\rangle := \left\{ \sum_{i=1}^k a_i R_i \;\middle|\; a_i \in \mathbb{R}[x_{1:n}] \right\}.$$

The *multiplicative monoid* generated by the R_i is the semigroup

$$M\langle R\rangle = M\langle R_1, \dots, R_k\rangle := \left\{ \prod_{i=1}^k R_i^{e_i} \;\middle|\; e_i \in \mathbb{N}_0 \right\}. \tag{1}$$

A *polynomial cone* C is a subset of $\mathbb{R}[x_{1:n}]$ containing all squares a^2 with $a \in \mathbb{R}[x_{1:n}]$, such that $r+s, rs \in C$ whenever $r, s \in C$. The smallest polynomial cone is the set SOS of polynomials which can be represented as sums of squares; we call such polynomials *SOS polynomials*. The *polynomial cone* generated by the R_i is the smallest polynomial cone containing $R_1, \dots, R_k$; it is given by

$$C\langle R\rangle = C\langle R_1, \dots, R_k\rangle = \left\{ y_0 + Y^T R_S \;\middle|\; y_0 \in SOS,\ Y \in SOS^{2^k} \right\}, \tag{2}$$

where R_S denotes the vector containing the 2^k polynomials in the *squarefree part*

$$S\langle R\rangle = S\langle R_1,\ldots,R_k\rangle := \left\{ \prod_{i=1}^{k} R_i^{e_i} \;\middle|\; e_i \in \{0,1\} \right\} \tag{3}$$

of $M\langle R_1,\ldots,R_k\rangle$.

THEOREM 2.1 (polynomial transposition theorem I). *Let P, Q, and R be vectors of polynomials. Then exactly one of the following holds:*

(i) $P(x) \geq 0, \quad Q(x) = 0, \quad R_i(x) \neq 0$ *for* $i = 1,\ldots,k, \quad$ *for some* $x \in \mathbb{R}^n$,

(ii) $f + g + h = 0$ *for some* $f \in C\langle P\rangle$, $g \in I\langle Q\rangle$, *and* $h \in M\langle R_1^2,\ldots,R_k^2\rangle$.

Proof. That conditions (i) and (ii) are mutually inconsistent can easily be seen. Indeed, if (i) holds then for any $f \in C\langle P\rangle$, $g \in I\langle Q\rangle$, and $h \in M\langle R_1^2,\ldots,R_k^2\rangle$, we have $f(x) \geq 0$, $g(x) = 0$, and $h(x) > 0$, whence $f(x) + g(x) + h(x) > 0$, contradicting (ii). That one of the two conditions can always be satisfied is the hard part. It follows from the statement that the inconsistency of (i) implies the solvability of (ii), which is equivalent to the weak Positivstellensatz stated and proved as Theorem 4.4.2 in Bochnak, Coste, and Roy [4]. □

For our application to optimality conditions, we need the following slightly different formulation.

THEOREM 2.2 (polynomial transposition theorem II). *Let P, Q, and R be vectors of polynomials. Then exactly one of the following holds:*

(i) $P(x) \geq 0$, $Q(x) = 0$, *and* $R(x) > 0$ *for some* $x \in \mathbb{R}^n$,

(ii) $f + g + h = 0$ *for some* $f \in C\langle P, R\rangle$, $g \in I\langle Q\rangle$, *and* $h \in M\langle R\rangle$.

Proof. That conditions (i) and (ii) are mutually inconsistent can again easily be seen. Indeed, if (i) holds then for any $f \in C\langle P, R\rangle$, $g \in I\langle Q\rangle$, and $h \in M\langle R\rangle$, we have $f(x) \geq 0$, $g(x) = 0$, and $h(x) > 0$, whence $f(x) + g(x) + h(x) > 0$, contradicting (ii).

If, on the other hand, (i) is inconsistent, this implies that the system $\big(P(x), R(x)\big) \geq 0$, $Q(x) = 0$, and $R(x) \neq 0$ is inconsistent, and by Theorem 2.1, there exist $f \in C\langle P, R\rangle$, $g \in I\langle Q\rangle$, and $h \in M\langle R_1^2,\ldots,R_k^2\rangle$ with $f+g+h = 0$. Since $M\langle R_1^2,\ldots,R_k^2\rangle \subset M\langle R\rangle$, the result follows. □

Both versions have the following common generalization.

THEOREM 2.3 (general polynomial transposition theorem). *Let P, Q, R, and $S_1,\ldots,S_k$ be vectors of polynomials. Then exactly one of the following holds:*

(i) $P(x) \geq 0$, $Q(x) = 0$, $R(x) > 0$, *and* $S_i(x) \neq 0$ *for* $i = 1,\ldots,k$, *for some* $x \in \mathbb{R}^n$,

(ii) $f+g+h = 0$ *for some* $f \in C\langle P, R\rangle$, $g \in I\langle Q\rangle$, *and* $h \in M\langle R, S_1^T S_1,\ldots,S_k^T S_k\rangle$.

Proof. That conditions (i) and (ii) are mutually inconsistent can be proved as before. Given that (i) holds, then for any $f \in C\langle P, R\rangle$, $g \in I\langle Q\rangle$, and $h \in M\langle R, S_1^T S_1,\ldots,$
$S_k^T S_k\rangle$, we have $f(x) \geq 0$, $g(x) = 0$, and $h(x) > 0$, leading to $f(x) + g(x) + h(x) > 0$, contradicting (ii).

The fact that (i) is inconsistent implies that the system of constraints $R(x) \geq 0$, $Q(x) = 0$, and $\big(R(x), S_1(x)^T S_1(x),\ldots,S_k(x)^T S_k(x)\big) > 0$ is inconsistent, and by Theorem 2.2, there exist polynomials $f \in C\langle P, R\rangle$, $g \in I\langle Q\rangle$, and $h \in M\langle R, S_1^T S_1,\ldots,$
$S_k^T S_k\rangle$ with $f + g + h = 0$. □

The equivalence of the three transposition theorems, Theorems 2.1, 2.2, and 2.3, can be seen by taking $R(x) \equiv 1 \in \mathbb{R}$ in Theorem 2.3 and noting that $C\langle P, 1\rangle = C\langle P\rangle$ and $M\langle 1, S_1^T S_1,\ldots,S_k^T S_k\rangle = M\langle S_1^2,\ldots,S_k^2\rangle$ when all S_j are scalars.

The following result gives necessary and sufficient conditions for the global opti-

mality of a feasible point of an optimization problem defined in terms of a polynomial objective function f and polynomial constraints. In most applications, f will be a real-valued function.

However, it is not difficult to state and prove analogous conditions for multiobjective optimization problems, by allowing f to be vector-valued. In this case, optimality is replaced by Pareto optimality, defined as follows. The point $\hat{x}$ is called *weakly Pareto minimal* with respect to the continuous function $f : X \subseteq \mathbb{R}^n \to \mathbb{R}^m$ for f on X if $x \in X$ and there exists a neighborhood of $\hat{x}$ in X which does not contain a point y with $f(y) < f(\hat{x})$.

THEOREM 2.4 (global Karush–John conditions). *Let $\hat{x}$ be a feasible point of the polynomial Pareto optimization problem*

$$\begin{aligned} &\min\ f(x) \\ &s.t.\ \ C(x) \geq 0, \\ &\qquad F(x) = 0, \end{aligned} \tag{4}$$

where $f \in \mathbb{R}[x_{1:n}]^k$ and $C \in \mathbb{R}[x_{1:n}]^m$, $F \in \mathbb{R}[x_{1:n}]^r$ are vectors of polynomials in $x_{1:n}$. Write B for the vector obtained by concatenating C with the vector $G(x) = f(\hat{x}) - f(x)$, so that $B_k = C_k$ for $k \leq m$, and define B_S as indicated by (3). Then the following are equivalent:

(i) *The point $\hat{x}$ is a global weak Pareto minimum of (4).*

(ii) *There are a polynomial $y_0 \in SOS$, polynomial vectors $Y \in SOS^{2^{m+1}}$, and $Z \in \mathbb{R}[x_{1:n}]^r$, and a multi-index $\alpha \in \mathbb{N}_0^k$ with $|\alpha| > 0$ such that*

$$G(x)^\alpha + y_0(x) + Y(x)^T B_S(x) + Z(x)^T F(x) = 0 \tag{5}$$

identically in x.

Moreover, any solution of (5) satisfies

$$y_0(\hat{x}) = 0, \quad \inf\{Y(\hat{x}), B_S(\hat{x})\} = 0, \quad F(\hat{x}) = 0, \tag{6}$$

$$\delta_{|\alpha|\,1} f_i'(\hat{x})^T = B_S{}'(\hat{x})^T Y(\hat{x}) + F'(\hat{x})^T Z(\hat{x}), \tag{7}$$

where $\alpha_i = 1$ and δ_{ik} is the Kronecker symbol.

Proof. $\hat{x}$ is a global weak Pareto minimum of (4) iff the conditions

$$C(x) \geq 0, \quad F(x) = 0, \quad f(x) < f(\hat{x})$$

are inconsistent. Because $f(x) < f(\hat{x})$ iff $G > 0$, the polynomial transposition theorem, Theorem 2.2, applies and shows that this is equivalent to the existence of polynomials $q \in C\langle B\rangle$, $r \in I\langle F\rangle$, and $s \in M\langle G\rangle$ with $q + r + s = 0$. Expressing this more explicitly using (1) and (2) shows this to be equivalent to (ii) without the constraint (6), and α only restricted to being a nonnegative multi-index. The equivalence of (i) and (ii) follows if we show that $|\alpha| \neq 0$.

Since $\hat{x}$ is feasible, we have $B(\hat{x}) \geq 0$, $F(\hat{x}) = 0$, and by construction, $G(\hat{x}) = 0$. Moreover, as a sum of squares, $y_0(\hat{x}) \geq 0$ and $Y(\hat{x}) \geq 0$. Inserting $x = \hat{x}$ into (5) gives, with the Kronecker δ,

$$\delta_{0\,|\alpha|} \leq \delta_{0\,|\alpha|} + y_0(\hat{x}) + Y(\hat{x})^T B_S(\hat{x}) = 0.$$

This indeed forces $|\alpha| > 0$.

We also get $y_0(\hat{x}) = 0$ and $Y(\hat{x})^T B_S(\hat{x}) = 0$. But the latter inner product is a sum of nonnegative terms; hence each product vanishes, giving the complementarity conditions (6). Differentiating the relation (5) and evaluating the result at $\hat{x}$ yield

$$\begin{aligned} 0 = &\sum_{\substack{i=1 \\ \alpha_i>0}}^{k} \alpha_i G^{\alpha-e_i}(\hat{x}) G_i'(\hat{x})^T + y_0'(\hat{x}) + Y'(\hat{x})^T B_S(\hat{x}) + {B_S}'(\hat{x})^T Y(\hat{x}) \\ &+ Z'(\hat{x})^T F(\hat{x}) + F'(\hat{x})^T Z(\hat{x}). \end{aligned} \tag{8}$$

We now note that $y_0'(\hat{x}) = 0$ because $y_0(\hat{x}) = 0$ and y_0 is SOS. Together with the facts that $|\alpha| > 0$, $G(\hat{x}) = 0$, and $F(\hat{x}) = 0$, we can simplify (8) and get

$$0 = \delta_{|\alpha|\,1} G_i'(\hat{x})^T + Y'(\hat{x})^T B_S(\hat{x}) + {B_S}'(\hat{x})^T Y(\hat{x}) + F'(\hat{x})^T Z(\hat{x}) \tag{9}$$

for that i with $\alpha_i = 1$. Finally, whenever $(B_S)_j(\hat{x}) \neq 0$, the complementarity conditions in (6) imply $Y_j(\hat{x}) = 0$ and then $Y_j'(\hat{x})^T = 0$ since Y_j is an SOS. Thus, $Y'(\hat{x})^T B_S(\hat{x}) = 0$, and (9) simplifies further to (7), upon noting that $G_i'(\hat{x}) = -f_i'(\hat{x})$. □

We may interpret the polynomials in (5) as a certificate that $\hat{x}$ is a global optimizer of (4). For applications in practice, one would first try to find $\hat{x}$ by local optimization or a heuristic global search, and then try to prove its globality by solving (5) with the side constraints (6). Note that the conditions are linear, except for the SOS conditions which give semidefinite constraints. Since the degree of the polynomials involved is not known a priori, one would solve a sequence of linear semidefinite feasibility problems on the finite-dimensional spaces of polynomials defined by limiting the total degree of the terms in (5) to $d = 1, 2, 3, \ldots$. Once a certificate is found one can stop.

Our theorem guarantees that this procedure will be finite (though worst case exponential because of the size of Y and B_S) iff $\hat{x}$ is indeed a global minimizer. In contrast, the method of Lasserre [20] yields an infinite sequence of semidefinite relaxations whose optimal values converge (under some compactness assumption) to the global optimum. There is no guarantee that the global optimum is found after finitely many steps. It would be interesting to combine the approaches to a constructive procedure for finding and verifying a global optimizer in finitely many steps.

Of course, an efficient implementation would try to avoid the exponential work in the majority of cases, by using suitable heuristics. For rigorous certification, one would have the additional problem of verifying the existence of an exact certificate close to the computed approximation.

We now relate the global Karush–John conditions to the traditional local conditions.

COROLLARY 2.5 (Kuhn–Tucker conditions). *If $\hat{x}$ is a global optimum of problem* (4) *with $k = 1$ and* (5) *holds with $\alpha = (1)$ then there are vectors $y \geq 0$ and z with*

$$\nabla f(\hat{x}) = C'(\hat{x})^T y + F'(\hat{x})^T z \tag{10}$$

and

$$\inf\{y, C(\hat{x})\} = 0. \tag{11}$$

Proof. We already know by Theorem 2.4 (7) that

$$\nabla f(\hat{x}) = {B_S}'(\hat{x})^T Y(\hat{x}) + F'(\hat{x})^T Z(\hat{x}).$$

We can write that in a slightly expanded way as follows:

$$\nabla f(\hat{x}) = C_S'(\hat{x})^T Y^{(1)}(\hat{x}) + G_1'(\hat{x})^T C_S(\hat{x}) Y^{(2)}(\hat{x}) + F'(\hat{x})^T Z(\hat{x}).$$

Noting that $C_S(\hat{x})Y^{(2)}(\hat{x}) \geq 0$ and $G_1'(\hat{x})^T = -\nabla f(\hat{x})$ and expanding further we see that

$$\gamma \nabla f(\hat{x}) = \sum_{\beta \in \{0,1\}^m} \sum_{\substack{i=1 \\ \beta_i = 1}}^{m} C_i'(\hat{x})^T C^{\beta - e_i}(\hat{x}) Y_\beta^{(1)}(\hat{x}) + F'(\hat{x})^T Z(\hat{x}) = 0,$$

where $\gamma = 1 + C_S(\hat{x})Y^{(2)}(\hat{x}) > 0$. We reorder the sums and get

$$\gamma \nabla f(\hat{x}) = \sum_{i=1}^{m} C_i'(\hat{x})^T \sum_{\substack{\beta \in \{0,1\}^m \\ \beta_i = 1}} C^{\beta - e_i}(\hat{x}) Y_\beta^{(1)}(\hat{x}) + F'(\hat{x})^T Z(\hat{x}) = 0.$$

If we now set

$$y_i = \frac{1}{\gamma} \sum_{\substack{\beta \in \{0,1\}^m \\ \beta_i = 1}} C^{\beta - e_i}(\hat{x}) Y_\beta^{(1)}(\hat{x}), \quad z = \frac{1}{\gamma} Z(\hat{x}), \tag{12}$$

we get the required equality (10). For the complementarity conditions we calculate

$$C_i(\hat{x}) y_i = \frac{1}{\gamma} \sum_{\substack{\beta \in \{0,1\}^m \\ \beta_i = 1}} C^{\beta}(\hat{x}) Y_\beta^{(1)}(\hat{x}) = 0,$$

since by (6) all terms in the sum vanish. ∎

Example 2.6. We consider the simple optimization problem

$$\begin{aligned} &\min \ 2x - x^2 \\ &\text{s.t.} \ \ x \in [-1, 1]. \end{aligned}$$

Clearly, the objective function is concave; hence the global minimum $\hat{f} = -3$ is attained at the bound $\hat{x} = -1$. We can write

$$f(x) - f(\hat{x}) = 2x - x^2 + 3 = (1 + x)(3 - x) = 1(1 + x)(1 - x) + 2(1 + x).$$

Obviously, each term on the right is nonnegative, showing again that $\hat{x}$ is a global minimizer.

From this representation one reads off the certificate ($\alpha = 1$, $y_0 = 0$, $Y^T = (0, 0, 2, 0, 1, 0, 0, 0)$, $Z = (\ )$) satisfying (5), where the components of B_S are arranged in the order $1, 1 + x, 1 - x, x^2 - 2x - 3, 1 - x^2, \ldots$.

While this example is trivial, it shows the essentials. Higher dimensional examples differ only in the complexity of what has to be written. In many other situations, as, e.g., in Example 2.7, the certificate will be very sparse and of low degree, thus simplifying the search for it.

By Theorem 2.4, for every global minimizer *all* terms in (5) have to vanish, i.e., for every global minimizer $\hat{x}$ we have that $G(\hat{x}) = 0$ and (6) and (7) are valid. Using this information frequently allows the identification of all global minimizers of an optimization problem when a certificate is available.

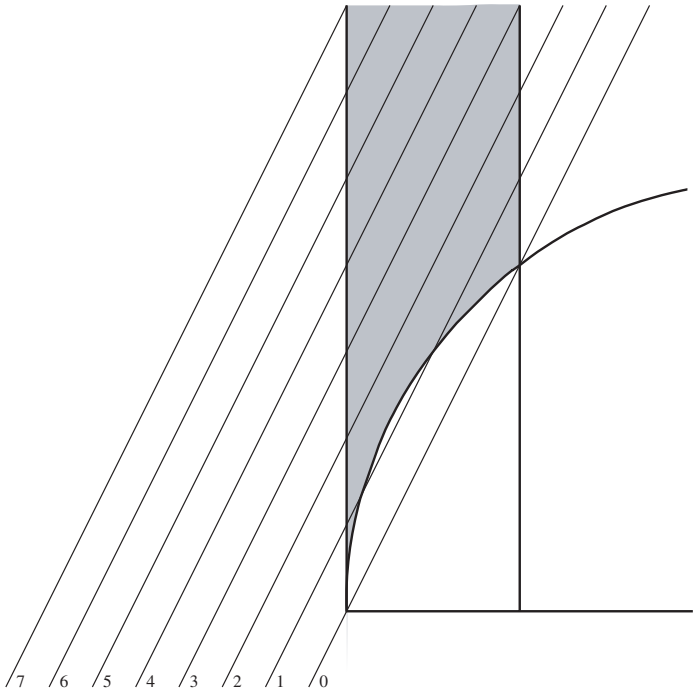


FIG. 1. *Feasible set (grey) and lines of constant value of f for Example* 2.7*, after eliminating z.*

Example 2.7. We consider the optimization problem

$$\begin{aligned} &\min\ z - x \\ &\text{s.t.}\ \ x - y + z = 0, \\ &\qquad x^2 - 10x + y^2 \geq 0, \\ &\qquad x \in [0,2], \quad y \geq 0. \end{aligned} \tag{13}$$

As Figure 1 shows, the point $(0,0,0)$ is a global optimizer with objective function value 0. This can be validated formally by Theorem 2.4, using the following polynomial identity:

$$\begin{aligned} &(x-z) + \tfrac{1}{2}y(2-x) + \tfrac{1}{4}y(x-z) + \tfrac{1}{4}x(2-x) + \tfrac{1}{4}(x^2 - 10x + y^2) \\ &\quad + \tfrac{1}{4}(4+y)(x-y+z) = 0. \end{aligned}$$

By checking when in this identity all terms vanish, we see that for the point $(2,4,2)$ (and no third point), all terms of the certificate vanish as well. Hence (cf. also the figure), this point is another global minimizer of (13), and no other global minimizers exist.

Example 2.8. For the problem

$$\begin{aligned} &\min\ x + y + z \\ &\text{s.t.}\ \ y - y^2 + z \geq 0, \\ &\qquad 2x + y + y^2 + z - 2 \geq 0, \\ &\qquad x \geq 0, \end{aligned}$$

we can find the polynomial identity

$$1-(x+y+z)+\tfrac{1}{2}\cdot(y-y^2+z)+\tfrac{1}{2}\cdot(2x+y+y^2+z-2)=0,$$

which implies that the global minimum value of the objective function is 1. By the complementarity conditions, we find that the two nonlinear inequality constraints must be active at every global minimum, i.e.,

$$\begin{aligned} y-y^2+z&=0,\\ 2x+y+y^2+z-2&=0, \end{aligned}$$

which implies that all

$$\hat{x}\in\left\{\begin{pmatrix}1-s^2\\ s\\ s(s-1)\end{pmatrix}\;\middle|\; s\in[-1,1]\right\}$$

are global optima.

Finally, the following example shows that in (5), the possibility $|\alpha|>1$ may indeed be necessary.

Example 2.9. For every nonnegative integer k the optimization problem

$$\begin{aligned} &\min\ x\\ &\text{s.t.}\ \ x^{2k+1}\geq 0 \end{aligned}$$

admits the unique global optimizer $\hat{x}=0$. The required polynomial identity of smallest degree is

$$(-x)^{\alpha}+1\cdot x^{2k+1}=0,\quad \alpha=2k+1.$$

In the GloptLab package [28], an optimization package currently developed in MATLAB, methods are being implemented and tested, which work along the lines presented in this section. They use the SeDuMi [33] package for the semidefinite programming part and combine the present techniques with branch-and-bound, constraint propagation, interval techniques [25, 26, 27], and linear relaxations. They have produced very promising results. At a later stage, the most successful techniques will be implemented as inference modules for the COCONUT environment [10].

3. Refined linear theorems of the alternative. In the linear case, there is a long tradition of theorems of the alternative, beginning with the lemma of Farkas [11], and culminating in the transposition theorems of Motzkin [24] and Tucker [34]. These transposition theorems are concerned with the solvability of linear constraints of various forms (equations, inequalities, strict inequalities, disequalities); see, e.g., Broyden [6] for some history.

As we shall show, there is a single general transposition theorem, which contains the others as special cases. As for the latter, our starting point is the lemma of Farkas.

LEMMA 3.1 (Farkas). *Let $A\in\mathbb{R}^{m\times n}$ and $g\in\mathbb{R}^n$. Then exactly one of the following conditions can be satisfied.*

(i) $g^Tp<0$, $Ap\geq 0$ *for some* $p\in\mathbb{R}^n$.

(ii) $g=A^Tq$, $q\geq 0$ *for some* $q\in\mathbb{R}^m$.

For the formulation of the transposition theorem and the constraint qualification we define $[\mathbf{1},-u]=:E_u\in\mathbb{R}^{k\times(k+1)}$ with $0<u\in\mathbb{R}^k$ and $\mathbf{1}$ being the identity matrix. We get the following result.

LEMMA 3.2. *For $X \in \mathbb{R}^{n\times k}$ with $\operatorname{rk} X = n$, $0 < u \in \mathbb{R}^k$, $0 < v \in \mathbb{R}^\ell$, and $Y \in \mathbb{R}^{n\times\ell}$ there exists a matrix $0 \le S \in \mathbb{R}^{(k+1)\times(\ell+1)}$ with*

$$XE_uS = YE_v.$$

Proof. Since $\operatorname{rk} X = n$ every $y \in \mathbb{R}^n$ can be written as a linear combination of the columns x_i of X:

$$y = \sum_{i=1}^{k} \lambda_i x_i.$$

Define

$$\mu = \max_{i=1}^{k} \left\{ -\frac{\lambda_i}{u_i}, 0 \right\}.$$

Then

$$y = \sum_{i=1}^{k} (\lambda_i + \mu u_i) x_i + \mu \left(-\sum_{i=1}^{k} u_i x_i \right) = XE_u s,$$

with $0 \ge s_i := \lambda_i + \mu u_i$ and $s_{k+1} := \mu$. Since all columns of $YE_v \in \mathbb{R}^n$, the result follows. ∎

We prove the following general theorem of the alternative, and deduce from it the transposition theorems of Motzkin and Tucker.

THEOREM 3.3 (general linear transposition theorem). *Consider matrices $A \in \mathbb{R}^{m_A\times n}$, $B \in \mathbb{R}^{m_B\times n}$, $C \in \mathbb{R}^{m_C\times n}$, and $D_j \in \mathbb{R}^{m_j\times n}$ with $m_j > 0$ for $j = 1, \dots, N$. Then exactly one of the following holds.*

(i) *$Ax = 0$, $Bx \ge 0$, $Cx > 0$, and $D_j x \ne 0$ for $j = 1, \dots, N$, for some $x \in \mathbb{R}^n$.*

(ii) *We have $m_C > 0$ and there exist $q \in \mathbb{R}^{m_A}$, $r \in \mathbb{R}^{m_B}$, and $s \in \mathbb{R}^{m_C}$ with*

$$A^T q + B^T r + C^T s = 0, \quad r \ge 0, \quad s \ge 0,\ s \ne 0, \tag{14}$$

or for some $j \in \{1, \dots, N\}$ there exist matrices $Q \in \mathbb{R}^{m_A\times(m_j+1)}$ and $R \in \mathbb{R}^{m_B\times(m_j+1)}$ with

$$A^T Q + B^T R = D_j^T E_u, \quad R \ge 0, \tag{15}$$

for some $u > 0$. Moreover, the same alternative holds if in (ii) *a fixed vector $u > 0$ (such as the all-one vector $u = e$) is prescribed.*

Proof. If (i) and (ii) hold simultaneously then multiplying (14) with x^T yields

$$(Bx)^T r + (Cx)^T s = 0.$$

Since $Bx \ge 0$ and $r \ge 0$ we have $(Cx)^T s \le 0$, which is a contradiction to $Cx > 0$ and $s \ge 0$, $s \ne 0$. Multiplying, on the other hand, (15) by x^T we get

$$0 \le (Bx)^T R = [(D_j x)^T, -(D_j x)^T u],$$

whence $D_j x \ge 0$ and $u^T D_j x < 0$ forces $D_j x = 0$, which is a contradiction.

Now assume that (i) cannot be solved. Then, for all $j = 1, \dots, N$ and all $v_j \in \mathbb{R}^{m_j}$, there is no $x \in \mathbb{R}^n$ with

$$Ax = 0, \quad Bx \ge 0, \quad Cx > 0, \quad \text{and} \quad v_j^T D_j x > 0. \tag{16}$$

Writing

$$g := \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad F := \begin{pmatrix} A & 0 \\ -A & 0 \\ B & 0 \\ C & e \\ v_1^T D_1 & 1 \\ \vdots & \vdots \\ v_N^T D_N & 1 \end{pmatrix}, \quad p := \begin{pmatrix} x \\ -\lambda \end{pmatrix},$$

we find that

$$g^T p < 0, \quad Fp \geq 0 \tag{17}$$

is unsolvable for $p \in \mathbb{R}^{n+1}$. By the Lemma of Farkas, Lemma 3.1, we can find $q \in \mathbb{R}^{2m_A + m_B + m_C + N}$ with

$$F^T q = g, \quad q = \begin{pmatrix} \hat{a} \\ \bar{a} \\ b \\ c \\ \mu \end{pmatrix} \geq 0. \tag{18}$$

Writing $a := \hat{a} - \bar{a}$, we find the existence of vectors $a \in \mathbb{R}^{m_A}$, $b \in \mathbb{R}^{m_B}$, $c \in \mathbb{R}^{m_C}$, and $\mu \in \mathbb{R}^N$ (depending on the choice of the v_j) such that

$$A^T a + B^T b + C^T c + \sum_{j=1}^{N} \mu_j D_j^T v_j = 0, \quad e^T \begin{pmatrix} c \\ \mu \end{pmatrix} = 1, \quad b, c, \mu \geq 0. \tag{19}$$

For $M \leq N$, we consider the set S_M consisting of all $(v_1, \ldots, v_{M-1}) \in \mathbb{R}^{m_1} \times \cdots \times \mathbb{R}^{m_{M-1}}$ for which (19) holds with $\mu_j = 0$ for $j \geq M$. Let $S_1 := \emptyset$. Let M be maximal with $S_M \neq \mathbb{R}^{m_1} \times \cdots \times \mathbb{R}^{m_{M-1}}$. If $M = 1$ we get $c \neq 0$, and hence $m_C > 0$, and by setting $q := a$, $r := b$, and $s := c$ we find (14).

Hence we may assume that $M > 1$ and pick $(v_1, \ldots, v_{M-1}) \notin S_M$. Take an arbitrary $v_M \in \mathbb{R}^{m_M}$. We can find vectors a, a', $b \geq 0$, $b' \geq 0$, $c \geq 0$, $c' \geq 0$, $0 \leq \xi, \xi' \in \mathbb{R}^{M-1}$, and numbers $\lambda > 0$ and $\tilde{\lambda} > 0$ with

$$A^T a + B^T b + C^T c + \sum_{j=1}^{M-1} \xi_j D_j^T v_j + \lambda D_M^T v_M = 0, \qquad e^T \begin{pmatrix} c \\ \xi \end{pmatrix} + \lambda = 1, \tag{20}$$

$$A^T a' + B^T b' + C^T c' + \sum_{j=1}^{M-1} \xi'_j D_j^T v_j + \lambda' D_M^T (-v_M) = 0, \quad e^T \begin{pmatrix} c' \\ \xi' \end{pmatrix} + \lambda' = 1. \tag{21}$$

Indeed, assume that we cannot find a, b, c, ξ, and λ with (20). We can get vectors a, b, c, μ satisfying (19). If there are only combinations with $\mu_{M+1:N} \neq 0$, then $(v_1, \ldots, v_M) \notin S_{M+1}$, contradicting the maximality of M. If there is a combination with $\mu_M = 0$, we find $(v_1, \ldots, v_{M-1}) \in S_M$, which is another contradiction. Thus $\mu_M \neq 0$, and we set $\xi := \mu_{1:M-1}$ and $\lambda := \mu_M$. The same argument gives (21).

Combining (20) and (21) leads to

$$A^T\underbrace{\left(\tfrac{1}{\lambda}a+\tfrac{1}{\lambda'}a'\right)}_{:=q}+B^T\underbrace{\left(\tfrac{1}{\lambda}b+\tfrac{1}{\lambda'}b'\right)}_{:=r}+C^T\underbrace{\left(\tfrac{1}{\lambda}c+\tfrac{1}{\lambda'}c'\right)}_{:=s}+\sum_{j=1}^{M-1}\underbrace{\left(\tfrac{\xi_j}{\lambda}+\tfrac{\xi_j'}{\lambda'}\right)}_{:=\nu_j}D_j^Tv_j=0,$$

$$e^T\begin{pmatrix}s\\ \nu\end{pmatrix}=\frac{\lambda+\lambda'}{\lambda\lambda'}-2=:\sigma. \tag{22}$$

If $s\neq 0$ or $\nu\neq 0$ the combination $(\mu_1,\dots,\mu_{M-1}):=\sigma^{-1}\nu\geq 0$, $a:=\sigma^{-1}q$, $b:=\sigma^{-1}r\geq 0$, and $c:=\sigma^{-1}s\geq 0$ proves that $(v_1,\dots,v_{M-1})\in S_M$, which is a contradiction. Thus $s=0$, implying $c=c'=0$, and $\nu=0$, hence $\xi_1=\xi_1'=\cdots=\xi_{N-1}=\xi_{N-1}'=0$, and $\lambda=1$. Since v_M was arbitrary, we have shown that for all $v_M\in\mathbb{R}^{m_M}$ there exist vectors $a\in\mathbb{R}^{m_A}$ and $b\in\mathbb{R}^{m_B}$ with

$$A^Ta+B^Tb+D_M^Tv_M=0. \tag{23}$$

We set $j:=M$ and choose for v_M in turn an arbitrary $u>0$ and the vectors $w_k:=-e_k$ $(k=1,\dots,m_j)$. This gives vectors q' and $r'\geq 0$ with

$$A^Tq'+B^Tr'=-D_j^Tu$$

and vectors q_k and $r_k\geq 0$ $(k=1,\dots,m_j)$ with

$$A^Tq_j+B^Tr_j=-D_j^Tw_j.$$

Forming the matrices $Q:=\left[q_1,\dots,q_{m_j},q'\right]$ and $R:=\left[r_1,\dots,r_{m_j},r'\right]$ finally gives (15). ∎

The well-known theorems of the alternative by Motzkin [24] and Tucker [34] are consequences of this theorem.

THEOREM 3.4 (Motzkin's linear transposition theorem). *Let $B\in\mathbb{R}^{m\times n}$, and let (I,J,K) be a partition of $\{1,\dots,m\}$ with $K\neq\emptyset$. Then exactly one of the following holds:*

(i) $(Bp)_I=0$, $(Bp)_J\geq 0$, $(Bp)_K>0$ *for some* $p\in\mathbb{R}^n$,

(ii) $B^Tq=0$, $q_{J\cup K}\geq 0$, $q_K\neq 0$ *for some* $q\in\mathbb{R}^m$.

Proof. We set $\tilde{A}:=B_{I:}$, $\tilde{B}:=B_{J:}$, $\tilde{C}:=C_{K:}$, $N=0$ and apply Theorem 3.3. ∎

THEOREM 3.5 (Tucker's linear transposition theorem). *Let $B\in\mathbb{R}^{m\times n}$, and let (I,J,K) be a partition of $\{1,\dots,m\}$ with $K\neq\emptyset$. Then exactly one of the following holds:*

(i) $(Bp)_I=0$, $(Bp)_{J\cup K}\geq 0$, $(Bp)_K\neq 0$ *for some* $p\in\mathbb{R}^n$,

(ii) $B^Tq=0$, $q_J\geq 0$, $q_K>0$ *for some* $q\in\mathbb{R}^m$.

Proof. Set $\tilde{A}=-B^T$, define the matrix $\tilde{B}$ whose rows are indexed by $I\cup J\cup K$ and whose columns are indexed by J with $\tilde{B}_{J:}=\mathbf{1}$, $\tilde{B}_{I\cup K,:}=0$, and introduce the matrix $\tilde{C}$ whose rows are indexed by $I\cup J\cup K$ with $\tilde{C}_{K:}=\mathbf{1}$, $\tilde{C}_{I\cup J,:}=0$, and $N=0$. Clearly, case (i) of Theorem 3.3 is equivalent to the solvability of the present (ii). On the other hand, case (ii) of Theorem 3.3 is here equivalent to the existence of vectors q, $r\geq 0$, and $s\geq 0$, $s\neq 0$ with

$$\tilde{A}^Tq+\tilde{B}^Tr+\tilde{C}^Ts=0.$$

Plugging in the definitions of $\tilde{A}$, $\tilde{B}$, and $\tilde{C}$ this becomes

$$-B_{I:}q=0,\quad -B_{J:}q+r=0,\quad -B_{K:}q+s=0,$$

which is clearly equivalent to (i). ∎

For the applications in the next section we need the following corollary of Theorem 3.3.

COROLLARY 3.6. *Let $B \in \mathbb{R}^{m\times n}$, and let (I,J,K) be a partition of $\{1,\dots,m\}$ of $i=|I|$, $j=|J|$, and $k=|K|>0$ elements. Then exactly one of the following holds:*

(i) *If $A = B^T_{K:}E_u$ for any $0 < u \in \mathbb{R}^k$, then* rk $A = k$ *and for some matrix $P \in \mathbb{R}^{n\times(k+1)}$*

$$(B(A+P))_{I:} = 0, \quad (B(A+P))_{J:} \geq 0, \quad \textit{and} \quad (BP)_{K:} = 0.$$

(ii) *$B^Tq = 0$, $q_J \geq 0$, $q_K \neq 0$ for some $q \in \mathbb{R}^m$.*

Proof. We set $\tilde{A} = -B^T$, define $\tilde{B} \in \mathbb{R}^{|I\cup J\cup K|\times|J|}$ with $\tilde{B}_{J:} = \mathbf{1}$, $\tilde{B}_{I\cup K,:} = 0$, construct $\tilde{D} \in \mathbb{R}^{|I\cup J\cup K|\times|K|}$ with $\tilde{D}_{K:} = \mathbf{1}$, $\tilde{D}_{I\cup J,:} = 0$, and set $N = 1$, $m_C = 0$, and $k = |K|$. Clearly, case (i) in Theorem 3.3 is equivalent to the present (ii).

On the other hand, (ii) in Theorem 3.3 is here equivalent to the existence of matrices Q and $R \geq 0$ with

$$B_{I:}Q = 0, \quad B_{J:}Q = R, \quad B_{K:}Q = E_u. \tag{24}$$

This, in turn, is equivalent to (i) by the following argument, for which we introduce the pseudoinverse $B^\dagger_{K:} = B^T_{K:}(B_{K:}B^T_{K:})^{-1}$ of $B_{K:}$.

Let us assume (i). By Lemma 3.2 we can find a matrix $S \geq 0$ with $(B_{K:}B^T_{K:})E_uS = E_u$, and we set $Q := B^\dagger_{K:}E_u + PS$. Then

$$\begin{aligned}
B_{I:}Q &= B_{I:}\big(B^T_{K:}(B_{K:}B^T_{K:})^{-1}E_u + PS\big) = B_{I:}(B^T_{K:}E_u + P)S = 0,\\
B_{J:}Q &= B_{J:}\big(B^T_{K:}(B_{K:}B^T_{K:})^{-1}E_u + PS\big) = B_{J:}(B^T_{K:}E_u + P)S =: R \geq 0,\\
B_{K:}Q &= B_{K:}B^\dagger_{K:}E_u + B_{K:}PS = E_u + 0.
\end{aligned}$$

Now assume (24). The last equation implies $\mathrm{rk}B_{K:} = \mathrm{rk}A = k$, and so the pseudoinverse of $B_{K:}$ exists and Q is of the form $Q = B^\dagger_{K:}E_u + P'$ for some P' with $B_{K:}P' = 0$. By Lemma 3.2 we can find $S \geq 0$ with $(B_{K:}B^T_{K:})^{-1}E_uS = E_u$ and set $P := P'S$. Calculating

$$\begin{aligned}
B_{I:}(A+P) &= B_{I:}(B^T_{K:}E_u + P'S) = B_{I:}(B^\dagger_{K:}E_u + P')S = B_{I:}QS = 0,\\
B_{J:}(A+P) &= B_{J:}(B^T_{K:}E_u + P'S) = B_{J:}(B^\dagger_{K:}E_u + P')S = B_{J:}QS = RS \geq 0,\\
B_{K:}P &= B_{K:}P'S = 0,
\end{aligned}$$

we prove (ii). ∎

4. A refinement of the Karush–John conditions. Karush–John conditions were originally derived—for single-objective optimization with inequality constraints only—by Karush [16], and were rediscovered by John [15]. They were subsequently generalized to mixed equality and inequality constraints, and to multiobjective optimization problems; there is a large literature on the subject, which can be accessed from the references below.

However, the Karush–John conditions in their most general form pose difficulties in applications, because the factor in front of the gradient term may be zero or very small. Therefore, most of the local solvers require a constraint qualification, like that of Mangasarian and Fromovitz [22] (MFCQ), to be able to reduce the Karush–John conditions to the much more convenient Kuhn–Tucker conditions [19]. Thorough discussions of such constraint qualifications can be found for single-objective optimization in Bazaraa, Sherali, and Shetty [3] and Mangasarian [21]. A more recent

account is in Bonnans and Shapiro [5, section 5.2]; there one can also find extensions to conic programming, semi-infinite programming, and infinite-dimensional problems (not considered in the present work). The Karush–John conditions have been investigated in the case of multiobjective optimization in Marusciac [23], though the result implicitly contains a constraint qualification. Further reference can be found in Arrow, Hurwicz, and Uzawa [2], Khanh and Luu [31], and Aghezzaf and Hachimi [1], especially for connections to the constraint qualifications and, e.g., in Cambini [8] for second order conditions.

Deterministic global optimization algorithms cannot take this course, since it is not known beforehand whether the global optimum satisfies an assumed constraint qualification. Therefore, they have to use the Karush–John conditions in their general form (cf., e.g., Kearfott [17]). Unfortunately, the additional constraints needed involve all multipliers and are very inconvenient for the solution process.

In this section we prove a strong version of the Karush–John conditions for nonlinear programming and multiobjective optimization, and a corresponding relaxation of the Mangasarian–Fromovitz constraint qualification (MFCQ). Apart from the inverse function theorem, our main tools are the transposition theorems of the previous section. The treatment is along the lines of the special case of a single objective discussed in our unpublished paper [29].

We consider concave and nonconcave constraints separately, and introduce slack variables to transform all nonconcave constraints into equations. Thus we may write a general nonlinear optimization problem without loss of generality in the form

$$\begin{aligned} &\min\ f(x) \\ &\text{s.t.}\ \ C(x) \geq 0, \quad F(x) = 0. \end{aligned} \tag{25}$$

In many applications, the objective function f will be a real-valued function. However, we allow f to be vector-valued; in this case, optimality is replaced by Pareto optimality.

The form (25), which separates the concave constraints (including bound constraints and general linear constraints) and the remaining nonlinear constraints, is most useful to prove our strong form of the Karush–John conditions. However, in computer implementations, a transformation to this form is not ideal, and the slack variables should be eliminated again from the optimality conditions.

THEOREM 4.1 (general first order optimality conditions). *Let $f : U \to \mathbb{R}^k$, $C : U \to \mathbb{R}^m$, and $F : U \to \mathbb{R}^r$ be functions continuously differentiable on a neighborhood U of $\hat{x} \in \mathbb{R}^n$. If C is convex on U and $\hat{x}$ is a weakly Pareto minimal point of the nonlinear program* (25)*, then there are vectors $\hat{w} \geq 0 \in \mathbb{R}^k$, $\hat{y} \in \mathbb{R}^m$, $\hat{z} \in \mathbb{R}^r$ such that*

$$f'(\hat{x})^T \hat{w} = C'(\hat{x})^T \hat{y} + F'(\hat{x})^T \hat{z}, \tag{26}$$

$$\inf(\hat{y}, C(\hat{x})) = 0, \tag{27}$$

$$F(\hat{x}) = 0, \tag{28}$$

and

$$\hat{w}, \hat{z} \text{ are not both zero.} \tag{29}$$

Proof. We begin by noting that a feasible point $\hat{x}$ of (25) is also a feasible point

for the optimization problem

$$\begin{aligned} &\min\ f(x)\\ &\text{s.t.}\ \ Ax \geq b,\\ &\qquad F(x) = 0, \end{aligned} \tag{30}$$

where J is the set of all components j for which $C(\hat{x})_j = 0$ and

$$A = C'(\hat{x})_{J:}, \quad b = C'(\hat{x})_{J:}\hat{x}.$$

For the indices k corresponding to the set N of inactive constraints, we choose $y_N = 0$ to satisfy condition (27). Since C is convex, we have $C(x) \geq C(\hat{x}) + C'(\hat{x})(x - \hat{x})$. Restricted to the rows J we get $C(x)_J \geq C'(\hat{x})_{J:}(x - \hat{x})$. This fact implies that problem (25) is a relaxation of problem (30) on a neighborhood U of $\hat{x}$. Note that since C is continuous we know that $C(x)_j > 0$ for $k \in N$ in a neighborhood of $\hat{x}$ for all constraints with $C(\hat{x})_j > 0$. Since, by assumption, $\hat{x}$ is weakly Pareto minimal for a relaxation of (30) and a feasible point of (30), it is weakly Pareto minimal for (30) as well. Together with the choice $y_N = 0$ the Karush–John conditions of problem (30) are again conditions (26)–(28). So we have successfully reduced the problem to the case where C is an affine function and all constraints are active at $\hat{x}$.

Thus, in the following, we consider a weakly Pareto minimal point $\hat{x}$ of the optimization problem (30) satisfying

$$A\hat{x} = b. \tag{31}$$

If rk $F'(\hat{x}) < r$ then $z^T F'(\hat{x}) = 0$ has a solution $z \neq 0$, and we can solve (26)–(29) with $\hat{y} = 0$, $\hat{w} = 0$. Hence we may assume that rk $F'(\hat{x}) = r$. This allows us to select a set R of r column indices such that $F'(\hat{x})_{:R}$ is nonsingular. Let B be the $(0,1)$-matrix such that Bs is the vector obtained from $s \in \mathbb{R}^n$ by discarding the entries indexed by R. Then the function $\Phi : C \to \mathbb{R}^n$ defined by

$$\Phi(x) := \begin{pmatrix} F(x) \\ Bx - B\hat{x} \end{pmatrix}$$

has at $x = \hat{x}$ a nonsingular derivative

$$\Phi'(\hat{x}) = \begin{pmatrix} F'(\hat{x}) \\ B \end{pmatrix}.$$

Hence, by the inverse function theorem, Φ defines in a neighborhood of $0 = \Phi(\hat{x})$ a unique continuously differentiable inverse function Φ^{-1} with $\Phi^{-1}(0) = \hat{x}$. Using Φ we can define a curved search path with tangent vector $p \in \mathbb{R}^n$ tangent to the nonlinear constraints satisfying $F'(\hat{x})p = 0$. Indeed, the function defined by

$$s_p(\alpha) := \Phi^{-1}\begin{pmatrix} 0 \\ \alpha Bp \end{pmatrix} - \hat{x}$$

for sufficiently small $\alpha \geq 0$ is continuously differentiable, with

$$s_p(0) = \Phi^{-1}(0) - \hat{x} = 0, \quad \begin{pmatrix} F(\hat{x} + s_p(\alpha)) \\ Bs_p(\alpha) \end{pmatrix} = \Phi\left(\Phi^{-1}\begin{pmatrix} 0 \\ \alpha Bp \end{pmatrix}\right) = \begin{pmatrix} 0 \\ \alpha Bp \end{pmatrix};$$

hence

$$s_p(0) = 0, \quad F(\hat{x} + s_p(\alpha)) = 0, \quad Bs_p(\alpha) = \alpha Bp. \tag{32}$$

Differentiation of (32) at $\alpha = 0$ yields

$$\begin{pmatrix} F'(\hat{x}) \\ B \end{pmatrix} \dot{s}_p(0) = \begin{pmatrix} F'(\hat{x})\dot{s}_p(0) \\ B\dot{s}_p(0) \end{pmatrix} = \begin{pmatrix} 0 \\ Bp \end{pmatrix} = \begin{pmatrix} F'(\hat{x}) \\ B \end{pmatrix} p;$$

hence $\dot{s}_p(0) = p$, i.e., p is indeed a tangent vector to $\hat{x} + s_p(\alpha)$ at $\alpha = 0$.

Since $\hat{x}$ is weakly Pareto minimal, we know that there exists a neighborhood V of $\hat{x}$ in the set of feasible points containing no y with $f(\hat{x}) > f(y)$. Thus, for every $y \in V$ there exists an index j with $f_j(\hat{x}) \le f_j(y)$. Taking an arbitrary curved path γ in the feasible set with $\gamma(0) = \hat{x}$ we conclude that there is an index j with $f_j'(\hat{x})\dot{\gamma}(0) \ge 0$. Hence, there is no direction p along such a curved search path, for which $f'(\hat{x})^T.p < 0$.

Now we consider a direction $p \in \mathbb{R}^n$ such that

$$Ap > 0, \tag{33}$$

$$F'(\hat{x})p = 0. \tag{34}$$

(In contrast to the purely concave case, we need the strict inequality in (33) to take care of curvature terms.) Since $A\hat{x} \ge b$ and (33) imply

$$A(\hat{x} + s_p(\alpha)) = A(\hat{x} + \alpha\dot{s}_p(0) + o(\alpha)) = A\hat{x} + \alpha(Ap + o(1)) \ge b$$

for sufficiently small $\alpha \ge 0$, (32) implies feasibility of the points $\hat{x} + s_p(\alpha)$ for small $\alpha \ge 0$.

Thus, s_p is a curved search path in the feasible set, and we conclude from the discussion above that there is no such p with $f'(\hat{x})^T p < 0$. Thus, (33), (34), and

$$f'(\hat{x})^T p < 0 \tag{35}$$

are inconsistent.

Therefore, the transposition theorem, Theorem 3.4, applies with

$$\begin{pmatrix} -f'(\hat{x}) \\ A \\ F'(\hat{x}) \end{pmatrix}, \quad \begin{pmatrix} w \\ y \\ z \end{pmatrix} \quad \text{in place of } B, q,$$

and shows the solvability of

$$-f'(\hat{x})^T w + A^T y + F'(\hat{x})^T z = 0, \quad w \ge 0, \quad y \ge 0, \quad \begin{pmatrix} w \\ y \end{pmatrix} \ne 0.$$

If we put $\hat{z} = z$, let $\hat{y}$ be the vector with $\hat{y}_J = y$ and zero entries elsewhere, and note that $\hat{x}$ is feasible, we find (26)–(28).

Because of (29), it now suffices to discuss the case where $w = 0$ and $z = 0$, and therefore

$$A^T y = 0, \quad y \ne 0. \tag{36}$$

In this case, $b^T y = (A\hat{x})^T y = \hat{x}^T A^T y = 0$. Therefore any point $x \in U$ satisfies $(Ax - b)^T y = x^T A^T y - b^T y = 0$, and since $y \ge 0$, $Ax - b \ge 0$, we see that the set

$$K := \{i \mid (Ax)_i = b_i \text{ for all } x \in V \text{ with } Ax - b \ge 0\} \tag{37}$$

contains all indices i with $y_i \neq 0$ and hence is nonempty.

Since V is nonempty, the system $A_{K:}x = b_K$ is consistent, and hence equivalent to $A_{L:}x = b_L$, where L is a maximal subset of K such that the rows of A indexed by L are linearly independent. If M denotes the set of indices complementary to K, we can describe the feasible set equivalently by the constraints

$$A_{M:}x \geq b_M, \quad \begin{pmatrix} A_{L:}x - b_L \\ F(x) \end{pmatrix} = 0.$$

This modified description of the feasible set has no equality constraints implicit in the inequality $A_{M:}x \geq b_M$. For a solution $\hat{x}$ of the equivalent optimization problem with these constraints, we find as before vectors $w \geq 0$, $\tilde{y}_M$, and $\binom{\tilde{y}_L}{z}$ such that

$$f'(\hat{x})^T w = A_{M:}^T \tilde{y}_M + \begin{pmatrix} A_{L:} \\ F'(\hat{x}) \end{pmatrix}^T \begin{pmatrix} \tilde{y}_L \\ z \end{pmatrix}, \tag{38}$$

$$\inf(\tilde{y}_M, A_{M:}\hat{x} - b_M) = 0, \tag{39}$$

$$F(x) = 0, \quad A_{K:}\hat{x} - b_K = 0, \tag{40}$$

$$w, \ \begin{pmatrix} \tilde{y}_L \\ z \end{pmatrix} \text{ are not both zero.} \tag{41}$$

Clearly, this yields vectors $\hat{w} = w$, $\hat{y} = \tilde{y}$, and $\hat{z} = z$ satisfying (26) and (27), but now $\tilde{y}_{K \setminus L} = 0$. The exceptional situation $w = 0$, $z = 0$ can no longer occur. Indeed, as before, all indices i with $\tilde{y}_i \neq 0$ lie in K; hence $\tilde{y}_M = 0$ and (38) gives $A_{L:}^T \tilde{y}_L = 0$. Since, by construction, the rows of $A_{L:}$ are linearly independent, this implies $\tilde{y}_L = 0$, contradicting (41). Hence, we have that w and z are not both zero.

It remains to show that we can choose $y \geq 0$ with $A^T y = A_{M:}^T \tilde{y}_M + A_{L:}^T \tilde{y}_L$. From the definition (37) of K we know that the two relations $Ap \geq 0$ and $A_{K:}p \neq 0$ are inconsistent (set $x = \hat{x} + p$). In particular, the relations $Ap \geq 0$ and $\tilde{y}_K^T A_{K:}p < 0$ are inconsistent. By the lemma of Farkas, Lemma 3.1, we conclude the existence of a vector $q \geq 0$ with $A^T q = A_{K:}^T \tilde{y}_K = A_{L:}^T \tilde{y}_L$. Setting $y_M = \tilde{y}_M + q_M$ and $y_K = q_K$ completes the proof. ∎

In contrast to our version of the Karush–John condition, the standard Karush–John condition asserts under our assumptions only that $\hat{w}$, $\hat{y}$, and $\hat{z}$ are not simultaneously zero. Thus the present version gives more information in case that $\hat{w} = 0$. Therefore, weaker constraint qualifications are needed to ensure that $\hat{w} \neq 0$. In that case, the multipliers in (26) can be rescaled so that $\|\hat{w}\| = 1$. However, from a numerical perspective, it may be better to keep the homogeneous formulations, since a tiny w in a well-scaled multiplier vector implies near degeneracy and would give huge multipliers if normalized to $\|\hat{w}\| = 1$.

Note that in view of (27), the condition (29) can be written (after rescaling) in the equivalent form

$$\hat{w} \geq 0, \quad v^T \hat{w} + u^T \hat{y} + \hat{z}^T D \hat{z} = 1, \tag{42}$$

where $v \neq 0$ is an arbitrary nonnegative vector, u is an arbitrary nonnegative vector with $u_J > 0$, $u_N = 0$, and D is an arbitrary diagonal matrix with positive diagonal entries. This form is numerically stable in that all multipliers are bounded and near degeneracies—which would produce huge multipliers in the Kuhn–Tucker conditions—are revealed by a small norm of $\hat{w}$. The lack of a constraint qualification (which

generally cannot be established in finite precision arithmetic anyway) therefore simply appears as the limit $\hat{w} = 0$.

The formulation (42) is particularly useful for the rigorous verification of the existence of a solution of our refined Karush–John conditions in the vicinity of an approximate solution; cf. Kearfott [17, section 5.2.5] for the corresponding use of the standard Karush–John conditions. The advantage of our stronger formulation is that in case there are only a few nonconcave constraints, condition (42) involves only a few variables and hence is a much stronger constraint if constraint propagation techniques [17, 35] are applied to the optimality conditions.

Let $B := F'(\hat{x})^T E_u$ for some $u > 0$. We say that the *constraint qualification (CQ)* is satisfied if $\operatorname{rk} F'(\hat{x}) = r$ and there exists a matrix $Q \in \mathbb{R}^{n\times(r+1)}$ with

$$\begin{aligned} C'(\hat{x})_{J:}(B+Q) &\geq 0, \\ F'(\hat{x})Q &= 0. \end{aligned} \tag{43}$$

COROLLARY 4.2. *If, under the assumptions of Theorem* 4.1*, the constraint qualification* (CQ) *is satisfied then the conclusion of Theorem* 4.1 *holds with* $\hat{w} \neq 0$*.*

Proof. It is obvious that the conclusion of Theorem 4.1 holds with $\hat{w} \neq 0$ if

$$C'(\hat{x})_{J:}^T y_J + F'(\hat{x})^T z = 0,\ y_J \geq 0 \implies z = 0. \tag{44}$$

If (44) is satisfied, we have that $z \neq 0$, $y_J \geq 0$, and $C'(\hat{x})_{J:}^T y_J + F'(\hat{x})^T z = 0$ are inconsistent. By Corollary 3.6 this is equivalent to the constraint qualification. ∎

THEOREM 4.3 (Kuhn–Tucker conditions). *Under the assumption of Theorem* 4.1 *with* f *one-dimensional, if the constraint qualification (CQ) is satisfied, then there are vectors* $\hat{y} \in \mathbb{R}^m, \hat{z} \in \mathbb{R}^r$ *such that*

$$f'(\hat{x})^T = C'(\hat{x})^T \hat{y} + F'(\hat{x})^T \hat{z}, \tag{45}$$

$$\inf(\hat{y}, C(\hat{x})) = 0, \tag{46}$$

$$F(\hat{x}) = 0. \tag{47}$$

Equations (45)–(47) are the *Kuhn–Tucker conditions* for the nonlinear program (25); cf. [19]. The traditional linear independence constraint qualification requires in place of the assumptions in Theorem 4.3 the stronger condition that the rows of $\binom{F'(\hat{x})}{C'(\hat{x})_{J:}}$ are independent. In contrast, our condition allows arbitrary dependence among the rows of $C'(\hat{x})$.

Weaker than the constraint qualification (CQ) is the Mangasarian–Fromowitz constraint qualification (MFCQ), which asserts the existence of a vector q with $C'(\hat{x})_{J:}q > 0$ and $F'(\hat{x})q = 0$. It implies our constraint qualification (CQ), because $Q = q\lambda^T$ satisfies (43) for λ large enough. We now show that (MFCQ) is more restrictive than our new constraint qualification (CQ).

Example 4.4. We reconsider Example 2.7. As we have seen there, the point $\hat{x} = (0,0,0)^T$ is a local (even global) optimum of (13). Figure 1 reveals that there is a degeneracy since the two activities have a common tangent. Thus the constraint qualifications are nontrivial. Clearly, the mapping C defined by transforming the inequality constraints of $C(x) \geq 0$ is convex; hence we can use Theorem 4.1. We have

$$f'(\hat{x}) = \begin{pmatrix} -1 & 0 & 1 \end{pmatrix}, \quad F'(\hat{x}) = \begin{pmatrix} 1 & -1 & 1 \end{pmatrix},$$

$$B = \begin{pmatrix} 1 & -1 \\ -1 & 1 \\ 1 & -1 \end{pmatrix}, \quad C'_{J:}(\hat{x}) = \begin{pmatrix} -10 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

We can use the formulas (12) to calculate the multipliers

$$\hat{y} = \frac{1}{4} \begin{pmatrix} 1 \\ 4 \\ 2 \end{pmatrix}, \quad \hat{z} = 1;$$

then (45) reduces to the identity

$$1 \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} = \frac{1}{4} \begin{pmatrix} -10 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 4 \\ 2 \end{pmatrix} + 1 \cdot \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}.$$

This shows that we can choose $\hat{w} = 1$ in the Karush–John conditions. Indeed, our constraint qualification (CQ) is valid. For

$$Q = \frac{1}{19} \begin{pmatrix} -1 & 1 \\ 1 & -1 \\ 2 & -2 \end{pmatrix}$$

we have

$$C'_{J:}(\hat{x})(B + Q) = 0, \quad F'(\hat{x})Q = 0.$$

However, (MFCQ) is not satisfied since there is no vector q with $C'_{J:}(\hat{x})q > 0$.

5. Conclusions. We presented various theorems of the alternative, and, based on them, derived new optimality conditions that hold without any constraint qualification. These results strengthen known local conditions, but they are also suitable for use in a global optimization context, which was our main motivation for this work.

New and exciting is the fact that, for the first time, it is possible to give necessary and sufficient (global) optimality conditions for polynomial problems. In particular, it is possible to produce (under the idealized assumptions that all semidefinite programs can be solved exactly) certificates for global optimality of a putative solution $\hat{x}$. However, these global results are probably not the best possible.

The failure to find a certificate after all problems up to some maximum degree d have been solved makes it likely that $\hat{x}$ is not a global optimizer of (4). In this case, one would like to have a procedure that guarantees (for sufficiently large but a priori unknown d) to find a feasible point x with a better objective function value than the value at $\hat{x}$. Then a new local optimization could be started from x, resulting in a better candidate for a global optimizer. Work on this is in progress.

Also, at present we have no simple constraint qualification which would guarantee in the single-objective case that the exponent α in the global Karush–John condition of Theorem 2.4 takes the value 1, which is needed to construct from the certificate multipliers satisfying the Kuhn–Tucker conditions. We conjecture that the exponent $e = 1$ is possible iff the Kuhn–Tucker conditions can be satisfied at $\hat{x}$, in particular, under the same (weakened Mangasarian–Fromovitz) constraint qualification as in our Theorem 4.3. This would strengthen the currently weak connections between sections 2 and 3.

REFERENCES

[1] B. Aghezzaf and M. Hachimi, *On a gap between multiobjective optimization and scalar optimization*, J. Optim. Theory Appl., 109 (2001), pp. 431–435.

[2] K. J. Arrow, L. Hurwicz, and H. Uzawa, *Constraint qualifications in maximization problems*, Naval Res. Logist., 8 (1961), pp. 175–191.

[3] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming, Theory and Algorithms*, 2nd ed., Wiley, New York, 1993.

[4] J. Bochnak, M. Coste, and M.-F. Roy, *Real Algebraic Geometry*, Ergeb. Math. Grenzgeb. (3) 36, Springer-Verlag, Berlin, 1998.

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

[6] C. G. Broyden, *A simple algebraic proof of Farkas's lemma and related theorems*, Optim. Methods Softw., 8 (1998), pp. 185–199.

[7] C. de Boor, *An empty exercise*, ACM SIGNUM Newsletter, 25 (1990), pp. 3–7.

[8] R. Cambini, *Second order optimality conditions in multiobjective programming*, Optimization, 44 (1998), pp. 139–160.

[9] COCONUT, *COntinuous CONstraints—Updating the Technology*, http://www.mat.univie.ac.at/∼neum/glopt/coconut/.

[10] COCONUT, *The COCONUT Environment for Global Optimization*, http://www.mat.univie.ac.at/coconut-environment/ (2004).

[11] J. Farkas, *Über die Theorie der einfachen Ungleichungen*, J. Reine Angew. Math., 124 (1902), pp. 1–24.

[12] D. Henrion and J. B. Lasserre, *GloptiPoly: Global optimization over polynomials with MATLAB and SeDuMi*, ACM Trans. Math. Software, 29 (2003), pp. 165–194.

[13] D. Henrion and J. B. Lasserre, *Solving global optimization problems over polynomials with GloptiPoly* 2.1, in Global Optimization and Constraint Satisfaction, Ch. Bliek et al., eds., Springer-Verlag, Berlin, 2003, pp. 43–58.

[14] D. Henrion and J. B. Lasserre, *Detecting global optimality and extracting solutions in GloptiPoly*, in Positive Polynomials in Control, D. Henrion and A. Garuli, eds., Lecture Notes in Control and Inform. Sci. 312, Springer-Verlag, Berlin, 2005, pp. 293–310.

[15] F. John, *Extremum problems with inequalities as subsidiary conditions*, in Fritz John, Collected Papers, Vol. 2, J. Moser, ed., Birkhäuser Boston, Boston, 1985, pp. 543–560.

[16] W. Karush, *Minima of Functions of Several Variables with Inequalities as Side Constraints*, M.Sc. dissertation, Dept. of Mathematics, University of Chicago, Chicago, IL, 1939.

[17] R. B. Kearfott, *Rigorous Global Search: Continuous Problems*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996.

[18] H. W. Kuhn, *Nonlinear programming: A historical note*, in History of Mathematical Programming, J. K. Lenstra et al., eds., North–Holland, Amsterdam, 1991 pp. 82–96.

[19] H. W. Kuhn and A. W. Tucker, *Nonlinear programming*, in Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, J. Neyman, ed., Univeristy of California Press, Berkeley, CA, 1951, pp. 481–492.

[20] J. B. Lasserre, *Global optimization with polynomials and the problem of moments*, SIAM J. Optim., 11 (2001), pp. 796–817.

[21] O. L. Mangasarian, *Nonlinear Programming*, Classics Appl. Math. 10, SIAM, Philadelphia, 1994.

[22] 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.

[23] I. Marusciac, *On Fritz John type optimality criterion in multiobjective optimization*, Anal. Numér. Théor. Approx., 11 (1982), pp. 109–114.

[24] T. S. Motzkin, *Beiträge zur Theorie der Linearen Ungleichungen*, Inaugural Dissertation, Basel, Jerusalem, 1936.

[25] A. Neumaier, *Interval Methods for Systems of Equations*, Cambridge University Press, Cambridge, UK, 1990.

[26] A. Neumaier, *Introduction to Numerical Analysis*, Cambridge University Press, Cambridge, UK, 2001.

[27] A. Neumaier, *Complete search in continuous global optimization and constraint satisfaction*, in Acta Numerica 2004, A. Iserles, ed., Acta Numer. 13, Cambridge University Press, Cambridge, UK, 2004, pp. 271–369.

[28] A. Neumaier and F. Domes, *GloptLab—a Global Optimization Laboratory*, in preparation, 2006.

[29] A. Neumaier and H. Schichl, *Sharpening the Karush-John Optimality Conditions*, manuscript, 2003, http://www.optimization-online.org/DB_HTML/2003/07/691.html.

[30] A. Neumaier, O. Shcherbina, W. Huyer and T. Vinko, *A comparison of complete global optimization solvers*, Math. Program., 103 (2005), pp. 335–336.
[31] P. Q. Khanh and L. M. Luu, *Multifunction optimization problems involving parameters: Necessary optimality conditions*, Optimization, 51 (2002), pp. 577–595.
[32] S. Prajna, A. Papachristodoulou, and A. Parrilo, SOSTOOLS: *Sum of Squares Optimization Toolbox for MATLAB—User's Guide*, manuscript, 2002, http://www.optimization-online.org/DB_HTML/2002/05/483.html.
[33] SeDuMi, http://sedumi.mcmaster.ca (2006).
[34] A. W. Tucker, *Dual Systems of Homogeneous Linear Relations. Linear Inequalities and Related Systems*, H. W. Kuhn and A. W. Tucker, eds., Ann. Math. Stud. 38, Princeton University Press, Princeton, NJ, 1956.
[35] P. Van Hentenryck, L. Michel, and Y. Deville, *Numerica—A Modeling Language for Global Optimization*, MIT Press, Cambridge, MA, 1997.

SIAM J. OPTIM.
Vol. 17, No. 4, pp. 1056-1101
© 2006 Society for Industrial and Applied Mathematics

LYAPUNOV STABILITY OF COMPLEMENTARITY AND EXTENDED SYSTEMS*

M. KANAT CAMLIBEL†, JONG-SHI PANG‡, AND JINGLAI SHEN§

Abstract. A linear complementarity system (LCS) is a piecewise linear dynamical system consisting of a linear time-invariant ordinary differential equation (ODE) parameterized by an algebraic variable that is required to be a solution to a finite-dimensional linear complementarity problem (LCP), whose constant vector is a linear function of the differential variable. Continuing the authors' recent investigation of the LCS from the combined point of view of system theory and mathematical programming, this paper addresses the important system-theoretic properties of exponential and asymptotic stability for an LCS with a C^1 state trajectory. The novelty of our approach lies in our employment of a quadratic Lyapunov function that involves the auxiliary algebraic variable of the LCS; when expressed in the state variable alone, the Lyapunov function is piecewise quadratic, and thus nonsmooth. The nonsmoothness feature invalidates standard stability analysis that is based on smooth Lyapunov functions. In addition to providing sufficient conditions for exponential stability, we establish a generalization of the well-known LaSalle invariance theorem for the asymptotic stability of a smooth dynamical system to the LCS, which is intrinsically a nonsmooth system. Sufficient matrix-theoretic copositivity conditions are introduced to facilitate the verification of the stability properties. Properly specialized, the latter conditions are satisfied by a passive-like LCS and certain hybrid linear systems having common quadratic Lyapunov functions. We provide numerical examples to illustrate the stability results. We also develop an extended local exponential stability theory for nonlinear complementarity systems and differential variational inequalities, based on a new converse theorem for ODEs with B-differentiable right-hand sides. The latter theorem asserts that the existence of a "B-differentiable Lyapunov function" is a necessary and sufficient condition for the exponential stability of an equilibrium of such a differential system.

Key words. complementarity systems, Lyapunov stability, LaSalle's invariance principle, asymptotic and exponential stability

AMS subject classifications. 34A40, 90C33, 93C10, 93D05, 93D20

DOI. 10.1137/050629185

1. Introduction.

Fundamentally linked to a linear hybrid system, a linear complementarity system (LCS) is a piecewise linear dynamical system defined by a linear time-invariant ordinary differential equation (ODE) parameterized by solutions of a finite-dimensional linear complementarity problem (LCP) linearly coupled with the state of the differential equation. LCSs, and also nonlinear complementarity systems (NCSs), belong to the more general class of differential variational inequalities (DVIs) [38]. In the last few years there has been a rapidly growing interest in complementarity systems and DVIs from the mathematical programming community and the systems

*Received by the editors April 14, 2005; accepted for publication (in revised form) June 30, 2006; published electronically December 5, 2006.

http://www.siam.org/journals/siopt/17-4/62918.html

†Department of Mechanical Engineering, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands; and Department of Electronics and Communications Engineering, Dogus University, Istanbul, Turkey (k.camlibel@tue.nl). The work of this author is partially supported by the European Community through the Information Society Technologies thematic program under the project SICONOS (IST-2001-37172).

‡Department of Mathematical Sciences and Department of Decision Science and Engineering Systems, Rensselaer Polytechnic Institute, Troy, NY 12180-3590 (pangj@rpi.edu). The research of this author was partially supported by National Science Foundation Focused Research Group grant DMS 0353216.

§Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY 12180-3590 (shenj2@rpi.edu). The research of this author was partially supported by the National Science Foundation under grant DMS 0508986.

and control community, due to their applications in many areas such as robotics, nonsmooth mechanics, economics, and finance and traffic systems; see the recent review papers [3, 50] and [4, 5, 6, 7, 8, 9, 11, 19, 20, 21, 22, 37, 51, 53, 54] for studies on specific issues pertaining to the LCS.

Stability is a classical issue in dynamical system theory. One of the most widely adopted stability concepts is Lyapunov stability, which plays important roles in system and control theory and in the analysis of engineering systems. In the classical Lyapunov stability theory, we assume that the ODE in consideration has a smooth (at least C^1) right-hand side and the origin is an equilibrium. Furthermore, we assume that there exists a continuously differentiable, positive definite, and coercive function of the system states, which is called a Lyapunov function. If the Lie derivative of such a function along the vector field of the system is nonpositive at all states (in a small neighborhood of the origin), then one can establish stability of the origin in the sense of Lyapunov. On the other hand, if the Lie derivative of such a Lyapunov function along the vector field of the system is negative at all nonzero states (in a small neighborhood of the origin), then the system is asymptotically stable at the origin. In the setting of linear systems, this leads to the well-known Lyapunov equation.

An important extension of the above results is LaSalle's invariance principle [28], which plays a fundamental role in the stability analysis of smooth systems. This theorem says that if the largest invariant set of the zero level of the Lie derivative of the Lyapunov function along the system vector field is a singleton and contains the origin only, then the system is asymptotically stable at the origin. It is known that the singleton condition can be further expressed in terms of certain observability conditions. Thus checking the singleton condition is closely related to the observability analysis of the system.

Extending classical smooth system theory to stability analysis of hybrid and switched systems has received growing attention in recent years. Among the extensive literature on the stability of linear switched systems, we mention a few relevant papers. A multiple-Lyapunov-function approach was proposed in [2]; see also [56] for related discussion. Uniform (asymptotic) stability of switched linear systems is studied in [23] where an extension of LaSalle's invariance principle to certain classes of switched linear systems is addressed. The latter result is further generalized to the stability analysis of switched nonlinear systems [24], where several nonlinear norm-observability notions generalizing classical observability concepts are introduced to obtain sufficient conditions for asymptotic stability using arguments of the LaSalle type. For surveys of recent results, including extensive references, on stability and stabilization of switched linear systems, see [14, 29]. Typically, the mentioned results assume that a Lyapunov-like function exists for each mode's vector field and holds for the entire state space. In many hybrid and switched systems, however, each mode holds only over a subset of the state space, especially for those systems whose switchings are triggered by state evolution, such as the LCS. Hence, the above results are rather restrictive, even for linear switched systems. Due to this concern, the paper [12] had proposed copositive Lyapunov functions for "conewise linear systems" for which the feasible region of each mode is a polyhedral cone. This proposal leads to an interesting study of copositive matrices that satisfy the Lyapunov equation. Similar ideas and relevant results for piecewise linear systems can also be found in [26]. Also employing a copositivity theory, the authors of several recent papers [1, 16, 17, 18] have developed an extensive stability theory for evolutionary variational inequalities (EVIs), including an extension of LaSalle's invariance principle to such systems, nec-

essary conditions for asymptotic stability, application to mechanical systems under frictional contact, and matrix conditions for stability and instability for linear EVIs (LEVIs). The EVIs belong to the class of differential inclusions and are dynamic generalizations of a finite-dimensional variational inequality [15]. In this paper, as an example of a DVI, we briefly discuss the "functional evolutionary variational inequality" (FEVI) as another dynamic generalization of a static finite-dimensional variational inequality (VI); see the system (5.8). In contrast to the EVI, the FEVI always has continuously differentiable solution trajectories, whose stability properties can be established without resorting to the framework of differential inclusions (DIs). Last, we mention [52, section 8.2], which studies the stability of "linear selectionable" DIs. While an LCS is related to such a DI, the two are quite different; consequently, the results from this reference are not applicable to the LCS. See the discussion at the end of subsection 3.3 for details.

It should be emphasized that while complementarity systems, and more generally, differential variational systems via their Karush–Kuhn–Tucker formulations, could be considered as special switched systems, LCSs, NCSs, and DVIs occupy a significant niche in many practical applications and have several distinguished features: inequality constraints on states, state-triggered mode switchings, and an endogenous control variable. These features invalidate much of the known theory of hybrid systems, which often allow arbitrary switchings, and necessitate the employment of the copositivity theory pioneered by such authors as Brogliato, Goeleven, and Schumacher. Another noteworthy point about the switched system theory is that it takes for granted a fundamental "non-Zenoness assumption" (i.e., finite number of switches in finite time) whose satisfaction is the starting point for stability analysis; for complementarity systems, this issue of finite switches is nontrivial and has been rigorously analyzed only very recently [37, 51]; see also [10].

Complementing the aforementioned works, this paper aims at analyzing the asymptotic and exponential stability of classes of nonsmooth differential systems, focusing in particular on the LCSs, NCSs, and DVIs. For an early work on the asymptotic behavior of solutions to the evolutionary nonlinear complementarity problem, see Chapter 3 in the Ph.D. thesis [25]. A key assumption for the class of LCSs treated in our work is that they have C^1 state trajectories for all initial states. Since the right-hand side of such an LCS is a Lipschitz function of state, the results for the LEVIs are not applicable to this class of LCSs; see [16, Remark 10]. Nevertheless, there are LCSs that fall within the framework of the LEVI, and which are therefore amenable to the treatment in the cited reference (see, e.g., Corollary 2 therein) but which cannot be handled by our approach. In contrast to a set-valued approach, our analysis is based to a large extent on the theory of "B-differentiable" functions (see section 2 for a formal definition of such a nonsmooth function). Specifically, unlike many stability results in the literature where the candidate Lyapunov functions are chosen to be continuously differentiable in the state, the nontraditional Lyapunov-like function in our consideration is, in the case of the LCS, quadratic in both the state and the associated algebraic variable; thus it is piecewise quadratic when expressed in the system state only. The nonsmoothness of the resulting Lyapunov function is the novelty of our work, as a result of which mathematical tools that go beyond the scope of the classical Lyapunov stability theory are needed. In this regard, our analysis is in the spirit of [52, Chapter 8]; yet the differential systems considered in our work are of a particular type, whose structure is fully exploited in designing the class of Lyapunov functions. Consequently, we are able to obtain much sharper results than those

derived from the general theory of differential inclusions. In particular, combining LCP theory and stability methods, we obtain asymptotic stability results via an extension of LaSalle's invariance principle; moreover, our stability results for the LCS are expressed in terms of matrix copositivity conditions. Several special cases are highlighted and numerical examples are given. We further extend these results to inhomogeneous LCSs, NCSs, and DVIs, with the latter two classes of systems satisfying the strong regularity condition [43, 15]. The noteworthy point of the latter extension is that it is based on a "converse theorem" of the exponential stability of an equilibrium of an ODE with a "B-differentiable" right-hand side. The latter theorem asserts that the existence of a "B-differentiable Lyapunov function" is a necessary and sufficient condition for the exponential stability of an equilibrium to such a differential system. Incidentally, there is an extensive literature on converse theorems for switched systems, some of which even involve discontinuous Lyapunov functions; see, e.g., [30, 33, 34, 42]. Our main result, Theorem 5.2, differs from the common treatment in switched systems in a major way; namely, our theorem is established for a general ODE with a B-differentiable right-hand side and thus potentially has broader applicability than those restricted to switched systems.

The organization of the rest of the paper is as follows. In the next section, we formally define the LCS, review the notions of stability, asymptotic stability, and exponential stability, and briefly examine some matrix classes related to the LCP [13]. The stability results for the equilibrium $x^e = 0$ of the LCS are presented in section 3, first for the "P-case" which is then extended to a non-P system. Numerical examples illustrating these results and the special case of a single-input-single-output (SISO) system are also given. Sections 4 and 5 address the stability issues of the extended systems; the former section treats the inhomogeneous LCS and the latter the NCS and the DVI, via the above-mentioned converse theorem for a B-differentiable ODE.

2. Linear complementarity systems. An LCS is defined by a tuple of four constant matrices $A \in \Re^{n\times n}$, $B \in \Re^{n\times m}$, $C \in \Re^{m\times n}$, and $D \in \Re^{m\times m}$; it seeks two time-dependent trajectories $x(t) \in \Re^n$ and $u(t) \in \Re^m$ for $t \in [0, T]$ for some $0 < T \leq \infty$ such that

$$\begin{aligned} \dot{x} &= Ax + Bu, \\ 0 \leq u \perp Cx + Du &\geq 0, \\ x(0) &= x^0, \end{aligned} \tag{2.1}$$

where $\dot{x} \equiv dx/dt$ denotes the time derivative of the trajectory $x(t)$, x^0 is the initial condition, and $a \perp b$ means that the two vectors a and b are orthogonal, i.e., $a^T b = 0$. We denote the above LCS by the tuple (A, B, C, D). Obviously, the LCP of finding a vector $u \in \Re^m$ satisfying

$$0 \leq u \perp q + Du \geq 0,$$

which we denote by the pair (q, D) and whose solution set we denote $\mathrm{SOL}(q, D)$, has a lot to do with various properties of the above LCS. We refer the reader to [13] for a comprehensive study of the LCP and also to the two-volume monograph [15] for many advanced solution properties of the LCP that we will freely use throughout this paper. In particular, under the blanket assumption that $B\mathrm{SOL}(Cx, D)$ is a singleton for all $x \in \Re^n$, an assumption which was introduced in [51] and used subsequently in [39], it follows that the LCS (2.1) is equivalent to the ODE

$$\dot{x} = Ax + B\mathrm{SOL}(Cx, D), \quad x(0) = x^0, \tag{2.2}$$

whose right-hand side $Ax + B\mathrm{SOL}(Cx, D)$ is a (single-valued) piecewise linear, and hence Lipschitz continuous and directionally differentiable (i.e., B(ouligand)-differentiable [35]) function of $x \in \Re^n$. (A word about notation: we identify the single vector in $B\mathrm{SOL}(Cx, D)$ with the set itself; thus we talk about the piecewise linear function $x \mapsto B\mathrm{SOL}(Cx, D)$ directly without referring to the element in $B\mathrm{SOL}(Cx, D)$. The same usage applies to other similar contexts.) The class of B-differentiable functions will play a central role throughout this work. Formally, a function $\Phi : \mathcal{D} \subseteq \Re^n \to \Re^m$ is *B-differentiable* at a point x in the open set $\mathcal{D}$ if Φ is Lipschitz continuous in a neighborhood of x contained in $\mathcal{D}$ and directionally differentiable at x; Φ is B-differentiable in $\mathcal{D}$ if it is B-differentiable at every point therein. We refer the reader to [15, Chapter 3] for basic properties of B-differentiable functions.

It follows from the ODE formulation (2.2) that the LCS (2.1) has a unique solution, which we denote $x(t, x^0)$, for all initial conditions $x^0 \in \Re^n$. If the initial condition x^0 is clear from the context, we will simply write $x(t)$ to de-emphasize the dependence of the solution trajectory on the initial condition. Even in this case where the x-trajectory is unique, there is no guarantee that there is a unique u-trajectory, unless D is a P-matrix [13], which implies that $\mathrm{SOL}(q, D)$ is a singleton for all $q \in \Re^m$, or unless the quadruple (A, B, C, D) satisfies the passifiability by pole shifting property and a rank condition [7]. See Proposition 2.2 for a unification of these uniqueness conditions. For our purpose, we are interested in the LCS (2.1) where the x-trajectory is unique and C^1 in time. It turns out that this condition is equivalent to the single-valuedness of $B\mathrm{SOL}(Cx, D)$ as made precise in the following result.

PROPOSITION 2.1. *Let (A, B, C, D) be given. The following two statements are equivalent.*

(a) *For every $x^0 \in \Re^n$, the LCS* (2.1) *has a unique C^1 trajectory $x(t, x^0)$ defined for all $t \geq 0$.*

(b) *For every $x^0 \in \Re^n$, the set $B\mathrm{SOL}(Cx^0, D)$ is a singleton.*

Proof. It remains to show (a) $\Rightarrow$ (b). This is clear because for any $u^0 \in \mathrm{SOL}(Cx^0, D)$, we have $Bu^0 = \dot{x}(0, x^0) - Ax^0$, where $\dot{x}(0, x^0)$ is the time derivative of the unique trajectory $x(t, x^0)$ evaluated at the initial time $t = 0$. ∎

Throughout the discussion of the LCS (2.1), we assume that condition (b) holds. There are simple instances where this condition holds easily. Statement (a) of the following result identifies one such instance; see [51]. The notation $a \circ b$ denotes the Hadamard product of two vectors; i.e., the ith component of $a \circ b$ is equal to $a_i b_i$.

PROPOSITION 2.2. *Suppose that $\mathrm{SOL}(Cx, D) \neq \emptyset$ for all $x \in \Re^n$. The following two statements hold.*

(a) *If $u \circ Du \leq 0 \Rightarrow Bu = 0$, then $B\mathrm{SOL}(Cx, D)$ is a singleton for all $x \in \Re^n$.*

(b) *If $[u \circ Du \leq 0,\ Bu = 0] \Rightarrow u = 0$, then $B\mathrm{SOL}(Cx, D)$ is a singleton for all $x \in \Re^n$ if and only if for every $x^0 \in \Re^n$, there exists a unique pair of trajectories $(x(t, x^0), u(t, x^0))$ defined for all $t \geq 0$ satisfying* (2.1) *such that $x(\cdot, x^0)$ is C^1.*

Proof. For statement (a), it suffices to show that $Bu^1 = Bu^2$ for any two solutions u^1 and u^2 in $\mathrm{SOL}(Cx, D)$. This is easy because any two such solutions must satisfy $u \circ Du \leq 0$ for $u \equiv u^1 - u^2$. For statement (b), it suffices to show the "only if" assertion; in turn it suffices to show the uniqueness of the $u(t, x^0)$ trajectory. But this is also clear in view of the uniqueness of the C^1 trajectory $x(t, x^0)$, which follows from Proposition 2.1. ∎

Remark 2.1. If D is positive semidefinite, then $u \circ Du \leq 0$ implies $(D + D^T)u = 0$. Thus, if the matrix $\begin{bmatrix} D + D^T \\ B \end{bmatrix}$ has full column rank, then the implication $[u \circ Du \leq 0,$

$Bu = 0] \Rightarrow u = 0$ holds. The former rank condition is used in [7] along with the passifiability condition, which implies the positive semidefiniteness of D, to yield the uniqueness of the u-trajectory.

There are many matrix classes in LCP theory; among these, the following are most relevant to this work. A matrix $D \in \Re^{m\times m}$ is a *P-matrix* if $u \circ Du \leq 0 \Rightarrow u = 0$; the matrix D is an *R_0-matrix* if $\mathrm{SOL}(0, D) = \{0\}$; the matrix D is *(strictly) copositive* on a cone $\mathcal{C} \subseteq \Re^m$ if $u^T Du \geq 0$ for all $u \in \mathcal{C}$ ($u^T Du > 0$ for all nonzero $u \in \mathcal{C}$); a copositive matrix D is *copositive plus* on $\mathcal{C}$ if $[u^T Du = 0, u \in \mathcal{C}] \Rightarrow (D + D^T)u = 0$. Properties of these matrices will be used freely in the paper; see [13]. In particular, it is known that a matrix D is P if and only if $\mathrm{SOL}(q, D)$ is a singleton for all $q \in \Re^m$; moreover a constant $c_D > 0$ exists such that $\|u\| \leq c_D\|q\|$ for all $q \in \Re^m$, where u is the unique solution of the LCP (q, D). It is further known that D is an R_0-matrix if and only if $\mathrm{SOL}(q, D)$ is bounded (possibly empty) for all $q \in \Re^m$. Clearly a P-matrix must be R_0. Last, note that if D is copositive on a convex cone $\mathcal{C}$, then

$$[\, u^T Du = 0,\ u \in \mathcal{C}\,] \Rightarrow (D + D^T)u \in \mathcal{C}^*,$$

where $\mathcal{C}^*$ denotes the dual cone of $\mathcal{C}$. Consequently, if D is a symmetric matrix copositive on a convex cone $\mathcal{C}$, then

$$[\, u^T Du = 0,\ u \in \mathcal{C}\,] \Rightarrow [\, \mathcal{C} \ni u \ \perp\ Du \in \mathcal{C}^*\,]. \tag{2.3}$$

We say that $(D, \mathcal{C})$ is an *R_0-pair* if the unique vector satisfying the right-hand complementarity conditions in the above implication is $u = 0$.

The condition that $B\mathrm{SOL}(Cx, D)$ is a singleton is not as restrictive as it seems. Indeed, consider a homogeneous differential affine variational inequality (DAVI)

$$\begin{aligned} \dot{x} &= Ax + Bu, \\ u &\in \mathrm{SOL}(K, Cx, D), \end{aligned} \tag{2.4}$$

where $u \in \mathrm{SOL}(K, Cx, D)$ means that $u \in K$ and

$$(u' - u)^T(Cx + Du) \geq 0 \quad \forall\, u' \in K,$$

with K being the polyhedral cone $\{u \in \Re^m : Eu \leq 0\}$ for some matrix E of appropriate dimension. Introducing a multiplier λ for the constraint in K, we deduce that $u \in \mathrm{SOL}(K, Cx, D)$ if and only if

$$\begin{aligned} 0 &= Cx + Du + E^T\lambda, \\ 0 &\leq -Eu \perp \lambda \geq 0. \end{aligned}$$

If D is positive definite, we can solve for u from the first equation, obtaining $u = -D^{-1}[Cx + E^T\lambda]$, which we can substitute into Eu and Bu. This results in the LCS

$$\begin{aligned} \dot{x} &= [\, A - BD^{-1}C\,]x - BD^{-1}E^T\lambda, \\ 0 &\leq \lambda \perp -ED^{-1}Cx + ED^{-1}E^T\lambda \geq 0. \end{aligned}$$

It is easy to see that the triple of matrices $(B', C', D') \equiv (-BD^{-1}E^T, -ED^{-1}C, ED^{-1}E^T)$ satisfies the property that $B'\mathrm{SOL}(C'x, D')$ is a singleton for all x, due to the positive definiteness of D. More generally, if D is only positive semidefinite (but not necessarily symmetric), it is still possible to convert (2.4) into an LCS (2.1) satisfying the desired singleton property, under suitable conditions; we refer the reader

to [15, Exercise 1.8.10] for a general conversion scheme. In what follows, we illustrate how this conversion can be carried out by assuming that the matrix

$$\begin{bmatrix} D & E^T \\ -E & 0 \end{bmatrix}$$

is nonsingular. Letting $w = -Eu$, we can show that (2.4) is equivalent to

$$\dot{x} = \widehat{A}x + \widehat{B}w,$$
$$0 \leq w \perp \widehat{C}x + \widehat{D}w \geq 0,$$

where

$$\widehat{A} \equiv A - \begin{bmatrix} B & 0 \end{bmatrix} \begin{bmatrix} D & E^T \\ -E & 0 \end{bmatrix}^{-1} \begin{bmatrix} C \\ 0 \end{bmatrix}, \quad \widehat{B} \equiv \begin{bmatrix} B & 0 \end{bmatrix} \begin{bmatrix} D & E^T \\ -E & 0 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ I \end{bmatrix},$$
$$\widehat{C} \equiv -\begin{bmatrix} 0 & I \end{bmatrix} \begin{bmatrix} D & E^T \\ -E & 0 \end{bmatrix}^{-1} \begin{bmatrix} C \\ 0 \end{bmatrix}, \qquad \widehat{D} \equiv \begin{bmatrix} 0 & I \end{bmatrix} \begin{bmatrix} D & E^T \\ -E & 0 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ I \end{bmatrix}.$$

It is not difficult to show that if $\mathrm{SOL}(K, Cx, D) \neq \emptyset$ for all $x \in \Re^n$ and if $(D+D^T)u = 0 \Rightarrow Bu = 0$, then the triple $(\widehat{B}, \widehat{C}, \widehat{D})$ is such that $\widehat{B}\mathrm{SOL}(\widehat{C}x, \widehat{D})$ is a singleton for all $x \in \Re^n$.

2.1. Stability concepts. An important goal of this paper is to derive sufficient conditions for the "equilibrium solution" $x = 0$ of the LCS (2.1) to be "exponentially stable" and "asymptotically stable." While these are well-known concepts in systems theory [28], we offer their formal definitions below for completeness. The setting is a time-invariant system on $\Re^n$,

$$\dot{x} = f(x), \quad x(0) = x^0, \tag{2.5}$$

where $f : \Re^n \to \Re^n$ is Lipschitz continuous. Let $x^e \in \Re^n$ be an *equilibrium* of the system (2.5), i.e., $f(x^e) = 0$, and let $x(t, x^0)$ denote the unique trajectory of (2.5).

DEFINITION 2.3. *The equilibrium x^e of (2.5) is*

(a) *stable in the sense of Lyapunov if, for each $\varepsilon > 0$, there is $\delta_\varepsilon > 0$ such that*

$$\| x^0 - x^e \| < \delta_\varepsilon \Rightarrow \| x(t, x^0) - x^e \| < \varepsilon \quad \forall\, t \geq 0;$$

unstable otherwise;

(b) *asymptotically stable if it is stable and $\delta > 0$ exists such that*

$$\| x^0 - x^e \| < \delta \Rightarrow \lim_{t \to \infty} x(t, x^0) = x^e;$$

(c) *exponentially stable if there exist scalars $\delta > 0$, $c > 0$, and $\mu > 0$ such that*

$$\| x^0 - x^e \| < \delta \Rightarrow \| x(t, x^0) - x^e \| \leq c \| x^0 - x^e \| e^{-\mu t} \quad \forall\, t \geq 0.$$

Clearly, exponential stability implies asymptotic stability, which further implies stability, but not vice versa. For a Lipschitz function $f(x)$ that is positively homogeneous in x, i.e., $f(\tau x) = \tau f(x)$ for all $\tau \geq 0$, we will be interested in the particular equilibrium $x^e = 0$. For the system (2.5) with such an f, we have $x(t, \tau x^0) = \tau x(t, x^0)$

for all $\tau \geq 0$ and all pairs $(t, x^0) \in [0, \infty) \times \Re^n$. For such a function f, stability of $x^e = 0$ is equivalent to *linearly bounded stability*, which means the existence of a constant $\eta > 0$ such that $\|x(t, x^0)\| \leq \eta \|x^0\|$ for all $(t, x^0) \in [0, \infty) \times \Re^n$; asymptotic stability is equivalent to *global asymptotic stability*, which means $\lim_{t \to \infty} x(t, x^0) = 0$ for all $x^0 \in \Re^n$; and exponential stability is equivalent to *global exponential stability*, which means the existence of scalars $c > 0$ and $\mu > 0$ such that $\|x(t, x^0)\| \leq c\|x^0\| e^{-\mu t}$ for all $(t, x^0) \in [0, \infty) \times \Re^n$. Throughout the paper, we will omit the adjective "global" when we deal with the equilibrium $x^e = 0$ for an ODE with a positively homogenous right-hand side.

Returning to the LCS (2.1), we note that, under our blanket assumption, the above definition is applicable to the equivalent system (2.2). Furthermore, since $B\mathrm{SOL}(0, D) = \{0\}$, $x^e = 0$ is indeed an equilibrium of (2.2). Due to its piecewise linearity, the right-hand function $f(x) \equiv Ax + B\mathrm{SOL}(Cx, D)$ is in general not Fréchet differentiable (but is indeed positively homogeneous). Although $f(x)$ is (globally) Lipschitz continuous, the nonsmoothness of $f(x)$ invalidates much of the standard analysis of well-known stability results for smooth dynamical systems; see, e.g., the book [28]. Our goal is to undertake a generalized stability analysis of the system (2.2), taking advantage of the special piecewise linear structure of the function $f(x)$. The resulting theory is a significant advance from the classical linear systems theory and involves matrix-theoretic properties that are based on LCP theory.

Before proceeding to derive sufficient conditions for the asymptotic stability of the equilibrium $x = 0$, we state and prove a necessary condition for the said stability.

PROPOSITION 2.4. *Suppose that $B\mathrm{SOL}(Cx, D)$ is a singleton for all $x \in \Re^n$. A necessary condition for $x^e = 0$ to be an asymptotically stable equilibrium for the LCS* (2.1) *is that for all scalars $\lambda \geq 0$, the following implication holds:*

$$\left.\begin{array}{c} \lambda\, x = Ax + Bu \\ 0 \leq u \perp Cx + Du \geq 0 \end{array}\right\} \Rightarrow x = 0. \tag{2.6}$$

If D is an R_0-matrix, then (2.6) *holds if and only if*

$$\left.\begin{array}{c} \lambda\, x = Ax + Bu \\ 0 \leq u \perp Cx + Du \geq 0 \end{array}\right\} \Rightarrow (\, x, u\,) = 0. \tag{2.7}$$

Proof. Indeed, if (x^*, u^*) is a solution of the system at the left-hand side of (2.6) for some $\lambda^* \geq 0$, then defining the trajectory $(x(t, x^*), u(t, x^*)) = (e^{\lambda^* t} x^*, e^{\lambda^* t} u^*)$ for all $t \geq 0$, we deduce that, $\lim_{t \to \infty} x(t, x^*) = 0$ only if $x^* = 0$. This establishes the implication (2.6). Clearly (2.7) implies (2.6). The converse is also clear, provided that D is an R_0-matrix. □

Remark 2.2. By the implication (2.6), which holds for all $\lambda \geq 0$, and by the homotopy invariance of the degree of a continuous mapping [31], it follows that the index of the map $x \mapsto -Ax - B\mathrm{SOL}(Cx, D)$ at the origin is well defined and equal to 1. (The index of a continuous map at an isolated zero is a well-known topological concept; see the reference.) The latter degree-theoretic necessary condition for asymptotic stability is a special case of a more general result due to Mawhin [32]. The implication (2.7) defines the "mixed R_0"-property of the matrix

$$\begin{bmatrix} A - \lambda I & B \\ C & D \end{bmatrix}.$$

If $A - \lambda I$ is nonsingular, then this property is equivalent to the R_0-property of the Schur complement $D - C(A - \lambda I)^{-1}B$. In this regard, the left-hand system of (2.6) is an instance of a homogeneous "mixed LCP," where there is a mixture of linear equations and standard linear complementarity conditions.

3. Stability results for $x^e = 0$. As in the classical analysis, our approach to the stability analysis of the system (2.1) is based on the existence of a Lyapunov function of a special kind. The novelty of our approach lies in the choice of the Lyapunov function: it is a quadratic function in the pair (x, u), which when expressed in the state variable x alone, is piecewise quadratic, and thus not smooth. At this point, we refer to the habilitation thesis of Scholtes [49] for the precise definition and an extensive study of piecewise differentiable functions; see also [15, Chapter 4]. Results from these references will be used freely in our discussion.

We first consider the case where D is a P-matrix. It follows that $\mathrm{SOL}(Cx, D)$ is a singleton for all $x \in \Re^n$, whose unique element we denote $u(x)$. A constant $c'_D > 0$ exists such that

$$\| u(x) \| \leq c'_D \| x \| \quad \forall\, x \in \Re^m. \tag{3.1}$$

Define three fundamental index sets:

$$\begin{aligned}
\alpha(x) &\equiv \{\, i \,:\, u_i(x) > 0 = (\, Cx + Du(x)\,)_i \,\}, \\
\beta(x) &\equiv \{\, i \,:\, u_i(x) = 0 = (\, Cx + Du(x)\,)_i \,\}, \\
\gamma(x) &\equiv \{\, i \,:\, u_i(x) = 0 < (\, Cx + Du(x)\,)_i \,\}.
\end{aligned}$$

In terms of these index sets, we have

$$u_\alpha(x) = -(\, D_{\alpha\alpha}\,)^{-1} C_{\alpha\bullet} x, \quad u_{\bar\alpha}(x) = 0,$$

where $\alpha = \alpha(x)$ and $\bar\alpha = \beta(x) \cup \gamma(x)$. Let $\mathrm{Gr\ SOL}_{CD}$ denote the graph of the solution function $u(x)$; i.e., $\mathrm{Gr\ SOL}_{CD}$, which is a closed (albeit not necessarily convex) cone, consists of all pairs $(x, u(x))$ for all $x \in \Re^n$. This graph can be described as follows. For each subset α of $\{1, \ldots, m\}$ with complement $\bar\alpha$, define

$$\mathcal{C}_\alpha \equiv \left\{ x \in \Re^n \,:\, \begin{bmatrix} -(\, D_{\alpha\alpha}\,)^{-1} C_{\alpha\bullet} \\ C_{\bar\alpha\bullet} - D_{\bar\alpha\alpha}(\, D_{\alpha\alpha}\,)^{-1} C_{\alpha\bullet} \end{bmatrix} x \geq 0 \right\}$$

and the matrix

$$E_\alpha \equiv \begin{bmatrix} I \\ -(\, D_{\alpha\alpha}\,)^{-1} C_{\alpha\bullet} \\ 0 \end{bmatrix} \in \Re^{(n+m)\times n}.$$

We then have

$$\Re^n = \bigcup_\alpha \mathcal{C}_\alpha \quad \text{and} \quad \mathrm{Gr\ SOL}_{CD} = \bigcup_\alpha \{\, E_\alpha x \,:\, x \in \mathcal{C}_\alpha \,\}. \tag{3.2}$$

The solution function $u(x)$ is piecewise linear in x and thus has directional derivatives given as follows: with

$$u'(x; d) \equiv \lim_{\tau \downarrow 0} \frac{u(x + \tau d) - u(x)}{\tau} \tag{3.3}$$

denoting the directional derivative of u at x along the direction d, $u'(x;d)$ is the unique vector v such that

$$\begin{array}{ll} \text{free } v_i & (Cd+Dv)_i = 0, \quad i \in \alpha(x), \\ 0 \le v_i \perp (Cd+Dv)_i \ge 0, & i \in \beta(x), \\ 0 = v_i, & i \in \gamma(x). \end{array}$$

Thus there exists a subset $\beta_d \subseteq \beta(x)$ such that the directional derivative $u'(x;d)$ is given by

$$u'_{\alpha_d}(x;d) = -(D_{\alpha_d\alpha_d})^{-1}C_{\alpha_d\bullet}\, d, \quad u'_{\bar{\alpha}_d}(x;d) = 0,$$

where $\alpha_d = \alpha(x) \cup \beta_d$ and $\bar{\alpha}_d = \{1,\ldots,m\} \setminus \alpha_d$. Note that we also have

$$u_{\alpha_d}(x) = -(D_{\alpha_d\alpha_d})^{-1}C_{\alpha_d\bullet}\, x, \quad u_{\bar{\alpha}_d}(x) = 0.$$

Since there are only finitely many subsets α_d, a constant $\widehat{c}\,' > 0$ exists such that

$$\| u'(x;d) \| \le \widehat{c}\,' \| d \| \quad \forall\, (x,d) \in \Re^{2n}. \tag{3.4}$$

Based on the LCP functions, we define the LCS map $\mathrm{SOL}'_{\mathrm{LCS}} : x \in \Re^n \to \Re^{2m}$ by

$$\mathrm{SOL}'_{\mathrm{LCS}}(x) \equiv \begin{pmatrix} u(x) \\ u'(x;dx) \end{pmatrix}, \quad \text{where } dx \equiv Ax + Bu(x),$$

and let Gr $\mathrm{SOL}'_{\mathrm{LCS}}$ denote its graph. Unlike Gr SOL_{CD}, which has a fairly simple representation in terms of the index subsets of $\{1,\ldots,m\}$ (cf. (3.2)), Gr $\mathrm{SOL}'_{\mathrm{LCS}}$ is somewhat more complicated to describe using index sets; for one thing, the latter graph is not closed because the function $u'(x;d)$ is in general not continuous in x. We denote the closure of Gr $\mathrm{SOL}'_{\mathrm{LCS}}$ by cl Gr $\mathrm{SOL}'_{\mathrm{LCS}}$. Like Gr SOL_{CD}, Gr $\mathrm{SOL}'_{\mathrm{LCS}}$ is a cone, albeit not necessarily convex.

In terms of $u(x)$, the LCS (2.1) becomes the ODE $\dot{x} = Ax + Bu(x)$ with a piecewise linear right-hand side which vanishes at the origin. In order to analyze the stability properties of the latter equilibrium $x^e = 0$, we postulate the existence of a symmetric matrix

$$M \equiv \begin{bmatrix} P & Q \\ Q^T & R \end{bmatrix} \in \Re^{(n+m)\times(n+m)}$$

that is *strictly copositive* on the cone Gr SOL_{CD}; i.e., $y^T My > 0$ for all nonzero $y \in$ Gr SOL_{CD}. Since the latter is a closed cone, the strict copositivity condition is equivalent to the existence of a scalar $c_M > 0$ such that

$$y^T My \ge c_M\, y^T y \quad \forall\, y \in \text{Gr } \mathrm{SOL}_{CD}. \tag{3.5}$$

In fact, one such choice is $c_M \equiv \min\{y^T My : y \in \text{Gr } \mathrm{SOL}_{CD}, \|y\| = 1\}$, which is well defined and positive. Let

$$V(x,u) \equiv \begin{pmatrix} x \\ u \end{pmatrix}^T \begin{bmatrix} P & Q \\ Q^T & R \end{bmatrix} \begin{pmatrix} x \\ u \end{pmatrix}$$

be the quadratic form associated with the matrix M. The composite function

$$\widehat{V}(x) \equiv V(x, u(x)) = x^T P x + 2x^T Q u(x) + u(x)^T R u(x)$$

is locally Lipschitz continuous and directional differentiable with

$$\widehat{V}\,'(x; v) = 2x^T P v + 2v^T Q u(x) + 2x^T Q u'(x; v) + 2u(x)^T R u'(x; v).$$

Associated with the trajectories $(x(t, x^0), u(t, x^0))$ of the LCS (2.1), where $u(t, x^0) \equiv u(x(t, x^0))$, define

$$\varphi_{x^0}(t) \equiv \widehat{V}(x(t, x^0)) \quad \forall\, t \geq 0.$$

By the chain rule of directional differentiation, the one-sided derivative of $\varphi_{x^0}(t)$ is given by

$$\begin{aligned}\varphi'_{x^0}(t+) &= \lim_{\tau \downarrow 0} \frac{\varphi_{x^0}(t+\tau) - \varphi_{x^0}(t)}{\tau} = \widehat{V}\,'(x(t, x^0); \dot{x}(t, x^0)) \\ &= 2x(t, x^0)^T P \dot{x}(t, x^0) + 2\dot{x}(t, x^0)^T Q u(t, x^0) + 2x^T Q u'(x(t, x^0); \dot{x}(t, x^0)) \\ &\quad + 2u(t, x^0)^T R u'(x(t, x^0); \dot{x}(t, x^0)).\end{aligned}$$

Letting $v(t, x^0) \equiv u'(x(t, x^0); \dot{x}(t, x^0))$ and substituting $\dot{x}(t, x^0) = Ax(t, x^0) + Bu(t, x^0)$, we deduce $\varphi'_{x^0}(t+) = v(t, x^0)^T N(t, x^0)$, where

(3.6)
$$N \equiv \begin{bmatrix} A^T P + PA & PB + A^T Q & Q \\ B^T P + Q^T A & Q^T B + B^T Q & R \\ Q^T & R & 0 \end{bmatrix} \text{ and } z(t, x^0) \equiv \begin{pmatrix} x(t, x^0) \\ u(t, x^0) \\ v(t, x^0) \end{pmatrix} \in \text{Gr SOL}'_{\text{LCS}}.$$

Note that, by (3.4),

$$\| v(t, x^0) \| \leq \widehat{c}\,' \| \dot{x}(t, x^0) \| \leq c_v \| (x(t, x^0), u(t, x^0)) \| \quad \forall\, (t, x^0) \in [0, \infty) \times \Re^n, \tag{3.7}$$

for some constant $c_v > 0$. Employing the notation introduced thus far, the following result provides sufficient conditions for the various kinds of stability to hold for the equilibrium $x^e = 0$ of the LCS (2.1) with a P-matrix D.

THEOREM 3.1. *Let D be a P-matrix. Suppose that matrices P, Q, and R, with P and R symmetric, exist such that M is strictly copositive on Gr SOL_{CD}. The following four statements hold for the equilibrium $x^e = 0$ of (2.1).*

(a) *If $-N$ is copositive on Gr SOL'_{LCS}, then x^e is linearly bounded stable.*

(b) *If $-N$ is strictly copositive on cl Gr SOL'_{LCS}, then x^e is exponentially stable.*

(c) *If $-N$ is copositive on Gr SOL'_{LCS} and*

$$[\, z(t, \xi)^T N z(t, \xi) = 0 \quad \forall\, t \geq 0\,] \Rightarrow \xi = 0, \tag{3.8}$$

then x^e is asymptotically stable.

(d) *If $-N$ is copositive-plus on Gr SOL'_{LCS} and*

$$[\, Nz(t, \xi) = 0 \quad \forall\, t \geq 0\,] \Rightarrow \xi = 0, \tag{3.9}$$

then x^e is asymptotically stable.

Proof. Let $x^0 \in \Re^n$ be arbitrary and let $u^0 \equiv u(x^0)$. Since $\varphi_{x^0}(t) \equiv \widehat{V}(x(t,x^0))$ is locally Lipschitz continuous for $t \geq 0$, it is almost everywhere differentiable on $[0,\infty)$, by Radamacher's theorem [48]. Hence for almost all $t \geq 0$, $\varphi'_{x^0}(t)$ exists and is equal to $\varphi'_{x^0}(t+)$, which is nonpositive, by the copositivity of $-N$ on Gr $\mathrm{SOL}'_{\mathrm{LCS}}$. On the one hand, we have, for some constant $\rho_M > 0$ independent of x^0,

$$\varphi_{x^0}(t) = \varphi_{x^0}(0) + \int_0^t \varphi'_{x^0}(s+)\, ds \leq \varphi_{x^0}(0) = V(x^0,u^0) \leq \rho_M \,\|\,(x^0,u^0)\,\|^2.$$

Hence by (3.1), we deduce that, for some constant $\rho'_M > 0$ independent of x^0,

$$\varphi_{x^0}(t) \leq \rho'_M \,\|\,x^0\,\|^2 \quad \forall\, t \geq 0. \tag{3.10}$$

On the other hand, by (3.5),

$$\varphi_{x^0}(t) = V(x(t,x^0),u(t,x^0)) \geq c_M \,\|\,(\,x(t,x^0),u(t,x^0)\,)\,\|^2 \geq c_M \,\|\,x(t,x^0)\,\|^2.$$

Combining the two inequalities, we obtain $\|x(t,x^0)\| \leq \sqrt{\rho'_M/c_M}\|x^0\|$, establishing the desired linearly bounded stability of $x^e = 0$.

The strictly copositivity of $-N$ on cl Gr $\mathrm{SOL}'_{\mathrm{LCS}}$ implies the existence of a scalar $c_N > 0$ such that $z^T N z \leq -c_N z^T z$ for all $z \in \mathrm{GrSOL}'_{\mathrm{LCS}}$. Hence, for all $x^0 \in \Re^n$ and for all $t \geq 0$, $\varphi'_{x^0}(t+) \leq -c_N\|(x(t,x^0),u(t,x^0),v(t,x^0))\|^2$. By (3.7), we deduce the existence of a constant $c'_M > 0$ such that

$$\varphi_{x^0}(t) \geq c'_M \,\|\,(x(t,x^0),u(t,x^0),v(t,x^0))\,\|^2.$$

Therefore, we obtain, for some constant $c > 0$,

$$\|\,z(t,x^0)\,\|^2 \leq c\left[\,\varphi_{x^0}(0) - \int_0^t \|\,z(s,x^0)\,\|^2\, ds\right] \quad \forall\,(\,t,x^0\,) \in [\,0,\infty\,) \times \Re^n,$$

where $z(t,x^0) \equiv (x(t,x^0),u(t,x^0),v(t,x^0))$. By Gronwall's inequality, we therefore deduce

$$\|\,x(t,x^0)\,\|^2 \leq \|\,z(t,x^0)\,\|^2 \leq c\,\varphi_{x^0}(0)\,e^{-ct} \leq c\,\rho'_M \,\|\,x^0\,\|^2\,e^{-ct},$$

where the last inequality is by (3.10). Consequently, $\|x(t,x^0)\| \leq \sqrt{c\rho'_M}\,\|x^0\|\,e^{-ct/2}$. This establishes part (b) of the theorem. We will postpone the proof of part (c) because it requires an auxiliary result that is of independent interest; see Proposition 3.2 below. Since N is symmetric, it follows that if $-N$ is copositive-plus on Gr $\mathrm{SOL}'_{\mathrm{LCS}}$, then (3.8) and (3.9) are equivalent implications. Hence (d) follows from (c). □

Part (c) of Theorem 3.1 is a generalized LaSalle's theorem for the LCS (2.1). The assumed implication (3.8) resembles a "generalized long-time observability condition" on the zero state of the LCS. Subsequently, we will discuss more about this condition; see subsection 3.1. For now, we note that if $-N$ is copositive on Gr $\mathrm{SOL}'_{\mathrm{LCS}}$ and if $(-N,\mathcal{C})$, where $\mathcal{C}$ is the closure of the convex hull of Gr $\mathrm{SOL}'_{\mathrm{LCS}}$, is an R_0-pair, then (3.8) holds. Indeed, in this case, by (2.3), it follows that $z(t,\xi)^T N z(t,\xi) = 0$ implies $z(t,\xi) = 0$. In particular $\xi = x(0,\xi) = 0$; hence (3.8) holds.

To prove part (c) of Theorem 3.1, we define for each fixed $x^0 \in \Re^n$ the *positive limit set*

$$\Omega(x^0) \equiv \Big\{\, x^\infty \in \Re^n \,:\, \exists\,\{\,t_k\,\} \uparrow \infty \text{ such that } x^\infty = \lim_{k\to\infty} x(t_k,x^0)\,\Big\}.$$

If M is strictly copositive on Gr SOL_{CD} and $-N$ is copositive on cl Gr $\mathrm{SOL}'_{\mathrm{LCS}}$, then $\Omega(x^0)$ is nonempty, by part (a) of Theorem 3.1. Additional properties of this set are summarized below.

PROPOSITION 3.2. *Let D be a P-matrix. If M is strictly copositive on* cl Gr $\mathrm{SOL}_{\mathrm{LCS}}$ *and $-N$ is copositive on* cl Gr $\mathrm{SOL}'_{\mathrm{LCS}}$*, then for every $x^0 \in \Re^n$, the following three statements hold:*

(a) *for every $x^\infty \in \Omega(x^0)$, the trajectory $\{x(t, x^\infty)\}_{t\geq 0} \subset \Omega(x^0)$;*
(b) *a constant σ_{x^0} exists such that $V(x^\infty, \mathrm{SOL}(Cx^\infty, D)) = \sigma_{x^0}$ for all $x^\infty \in \Omega(x^0)$;*
(c) *$\varphi'_{x^\infty}(t) = 0$ for all $x^\infty \in \Omega(x^0)$.*

Proof. Suppose $x^\infty = \lim_{k\to\infty} x(t_k, x^0)$ for some sequence $\{t_k\} \uparrow \infty$. For any $t \geq 0$, we have $x(t + t_k, x^0) = x(t, x(t_k, x^0))$; hence taking limits as $k \uparrow \infty$ and using the continuity of $x(t, \cdot)$ in the second argument, we deduce

$$\lim_{k\to\infty} x(t + t_k, x^0) = x(t, x^\infty),$$

which establishes part (a). To prove part (b), note that since $\varphi'_{x^0}(t+) \leq 0$ for all $t \geq 0$, it follows that $\varphi_{x^0}(t)$ is nonincreasing. Since

$$\varphi_{x^0}(t) = V(x(t, x^0), u(t, x^0)) = \begin{pmatrix} x(t, x^0) \\ u(t, x^0) \end{pmatrix} \begin{bmatrix} P & Q \\ Q^T & R \end{bmatrix} \begin{pmatrix} x(t, x^0) \\ u(t, x^0) \end{pmatrix} \geq 0,$$

by the copositivity of M on Gr $\mathcal{G}_{CD}(x(t, x^0))$, it follows that

$$\lim_{t\to\infty} \varphi_{x^0}(t)$$

exists. With σ_{x^0} denoting the above limit, it follows that $V(x^\infty, u(x^\infty)) = \sigma_{x^0}$ for all $x^\infty \in \Omega(x^0)$. Combining (a) and (b), we deduce that for all $x^\infty \in \Omega(x^0)$, we have

$$\varphi_{x^\infty}(t) = V(x(t, x^\infty), u(t, x^\infty)) = \sigma_{x^0} \quad \forall\, t \geq 0.$$

Thus, $\varphi_{x^\infty}(t)$ is a constant function on $[0, \infty)$. Part (c) is therefore trivial. □

Proof of Theorem 3.1(c). It suffices to show that $\Omega(x^0) = \{0\}$ for all $x^0 \in \Re^n$. Let $x^\infty \in \Omega(x^0)$ be given. By part (c) of Proposition 3.2, we have $0 = \varphi'_{x^\infty}(t) = z(t, x^\infty)^T N z(t, x^\infty)$ for all $t \geq 0$. Hence (3.8) implies $x^\infty = 0$ as desired. □

Admittedly, the conditions in Theorem 3.1 are in general not easy to verify. This is inevitable because most matrix properties in LCP theory are already so. Nevertheless, such difficulties have not prevented the fruitful development of the theory and applications of the LCP and its extensions. Thus we fully expect that Theorem 3.1 is of fundamental importance in the stability theory of the LCS. In what follows, we provide evidence for this optimism by deriving various special results and by giving examples to illustrate the broad applicability of this theorem. We begin by considering the case where both Q and R are taken to be zero. Proposition 3.3 below provides succinct matrix-theoretic conditions that ensure the existence of a "common Lyapunov function" for the LCS. (The study of copositivity has recently received renewed interest in the mathematical programming community; see, e.g., the Ph.D. thesis [41] and the paper [55]. It would be of interest to investigate how these works can be used to help check the conditions obtained herein.)

PROPOSITION 3.3. *Let D be a P-matrix and P be a symmetric positive definite matrix.*

(a) *If, for every* $\alpha \subseteq \{1, \ldots, m\}$,

$$\begin{bmatrix} -(D_{\alpha\alpha})^{-1}C_{\alpha\bullet} \\ C_{\bar{\alpha}\bullet} - D_{\bar{\alpha}\alpha}(D_{\alpha\alpha})^{-1}C_{\alpha\bullet} \end{bmatrix} x \geq 0 \Rightarrow x^T[A - B_{\bullet\alpha}(D_{\alpha\alpha})^{-1}C_{\alpha\bullet}]^T Px \leq 0, \tag{3.11}$$

then $x^e = 0$ *is a linearly bounded stable equilibrium of the* LCS (2.1),

(b) *If, for every* $\alpha \subseteq \{1, \ldots, m\}$,

$$\begin{aligned} &\left\{ \begin{bmatrix} -(D_{\alpha\alpha})^{-1}C_{\alpha\bullet} \\ C_{\bar{\alpha}\bullet} - D_{\bar{\alpha}\alpha}(D_{\alpha\alpha})^{-1}C_{\alpha\bullet} \end{bmatrix} x \geq 0,\ x \neq 0 \right\} \\ &\qquad \Rightarrow x^T[A - B_{\bullet\alpha}(D_{\alpha\alpha})^{-1}C_{\alpha\bullet}]^T Px < 0, \end{aligned} \tag{3.12}$$

then $x^e = 0$ *is an exponentially stable equilibrium of the* LCS (2.1).

(c) *If, for every* $\alpha \subseteq \{1, \ldots, m\}$, (3.11) *holds and*

$$\left. \begin{aligned} &\begin{bmatrix} -(D_{\alpha\alpha})^{-1}C_{\alpha\bullet} \\ C_{\bar{\alpha}\bullet} - D_{\bar{\alpha}\alpha}(D_{\alpha\alpha})^{-1}C_{\alpha\bullet} \end{bmatrix} x \geq 0 \\ &x^T[A - B_{\bullet\alpha}(D_{\alpha\alpha})^{-1}C_{\alpha\bullet}]^T Px = 0 \end{aligned} \right\} \Rightarrow\ x = 0, \tag{3.13}$$

then $x^e = 0$ *is an asymptotically stable equilibrium of the LCS* (2.1).

Proof. With $Q = 0$ and $R = 0$, the matrices M and N become

$$M = \begin{bmatrix} P & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad N = \begin{bmatrix} A^TP + PA & PB & 0 \\ B^TP & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

By (3.1) and the positive definiteness of P, it follows that M is strictly copositive on Gr SOL_{CD}. For any triple $z \equiv (x, u(x), v) \in \mathrm{cl\ Gr\ SOL}'_{\mathrm{LCS}}$ with $x \in \mathcal{C}_\alpha$, we have

$$z^T N z = 2x^T[A - B_{\bullet\alpha}(D_{\alpha\alpha})^{-1}C_{\alpha\bullet}]^T Px.$$

Hence, the proposition follows easily from Theorem 3.1. ∎

Remark 3.1. It should be noted that the resulting matrix M in the above proposition is not positive definite. This illustrates the fact that the strict copositivity of M on Gr SOL_{CD} is not as restrictive as it seems.

A special case of Proposition 3.3 pertains to a "passive-like" LCS for which there exists a *symmetric positive definite* K such that

$$-\begin{bmatrix} A^TK + KA & KB - C^T \\ B^TK - C & -D - D^T \end{bmatrix} \tag{3.14}$$

is positive semidefinite. This class of LCSs is closely related to the class of passive LCSs defined in [4, 7] and to the class of *positive real transfer functions* via the well-known Kalman–Yakubovich–Popov lemma [28]. In essence, we have bypassed the transfer functions and the "minimality" of the tuple (A, B, C, D) and worked directly with the positive semidefinite matrix (3.14). Note that if (3.14) is positive semidefinite, then the matrix D must be positive semidefinite albeit not necessarily

symmetric. It is possible for such a D to be also P without being positive definite; a trivial example is

$$D \equiv \begin{bmatrix} 1 & -2 \\ 0 & 1 \end{bmatrix}.$$

The next result shows how Proposition 3.3 (b) can be applied to such an LCS. This result complements Theorem 11.2 [7] in providing a sufficient condition for a passive-like LCS to be asymptotically stable.

COROLLARY 3.4. *Suppose that D is a P-matrix and there exists a symmetric positive definite matrix K such that* (3.14) *is positive semidefinite. If for every $\alpha \subseteq \{1, \ldots, m\}$,*

$$\left.\begin{array}{c} \begin{bmatrix} -(D_{\alpha\alpha})^{-1}C_{\alpha\bullet} \\ C_{\bar{\alpha}\bullet} - D_{\bar{\alpha}\alpha}(D_{\alpha\alpha})^{-1}C_{\alpha\bullet} \end{bmatrix} x \geq 0 \\ \begin{bmatrix} A^TK + KA & KB_{\bullet\alpha} - (C_{\alpha\bullet})^T \\ (B_{\bullet\alpha})^TK - C_{\alpha\bullet} & -D_{\alpha\alpha} - (D_{\alpha\alpha})^T \end{bmatrix} \begin{bmatrix} I \\ -(D_{\alpha\alpha})^{-1}C_{\alpha\bullet} \end{bmatrix} x = 0 \end{array}\right\} \Rightarrow \ x = 0,$$

then x^e is asymptotically stable.

Proof. It suffices to verify the implication (3.13). Let x satisfy the left-hand condition in the latter implication. Proceeding as before, we deduce

$$\begin{aligned} 0 &= \begin{pmatrix} x \\ u \end{pmatrix}^T \begin{bmatrix} A^TK + KA & KB - C^T \\ B^TK - C & -D - D^T \end{bmatrix} \begin{pmatrix} x \\ u \end{pmatrix} \\ &= x^T \begin{bmatrix} I \\ -(D_{\alpha\alpha})^{-1}C_{\alpha\bullet} \end{bmatrix}^T \begin{bmatrix} A^TK + PA & KB_{\bullet\alpha} - (C_{\alpha\bullet})^T \\ (B_{\bullet\alpha})^TK - C_{\alpha\bullet} & -D_{\alpha\alpha} - (D_{\alpha\alpha})^T \end{bmatrix} \\ &\quad \times \begin{bmatrix} I \\ -(D_{\alpha\alpha})^{-1}C_{\alpha\bullet} \end{bmatrix} x, \end{aligned}$$

which implies, since (3.14) is symmetric positive semidefinite,

$$\begin{bmatrix} A^TK + PA & KB_{\bullet\alpha} - (C_{\alpha\bullet})^T \\ (B_{\bullet\alpha})^TK - C_{\alpha\bullet} & -D_{\alpha\alpha} - (D_{\alpha\alpha})^T \end{bmatrix} \begin{bmatrix} I \\ -(D_{\alpha\alpha})^{-1}C_{\alpha\bullet} \end{bmatrix} x = 0.$$

The desired implication (3.13) follows easily from the assumption of part (b) herein. ∎

The assumption in Proposition 3.3(b) is significantly weaker than the passivity [4, 7] of the LCS tuple (A, B, C, D). The next two examples illustrate this point. The first example has a matrix A that is not negatively stable and the matrix D is not positive semidefinite.

Example 3.1. Consider the tuple with $n = 1$ and $m = 2$:

$$A = 1, \quad B = \begin{bmatrix} 2 & -2 \end{bmatrix}, \quad C = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad \text{and} \quad D = \begin{bmatrix} 1 & 3 \\ 0 & 1 \end{bmatrix}.$$

By an easy calculation, we have

$$A - B_{\bullet\alpha}(D_{\alpha\alpha})^{-1}C_{\alpha\bullet} = \begin{cases} 1 & \text{if } \alpha = \emptyset, \\ -1 & \text{if } \alpha = \{1\}, \\ -1 & \text{if } \alpha = \{2\}, \\ -9 & \text{if } \alpha = \{1, 2\}, \end{cases} \quad \text{and} \quad \mathcal{C}_\alpha = \begin{cases} \{0\} & \text{if } \alpha = \emptyset, \\ (-\infty, 0] & \text{if } \alpha = \{1\}, \\ [0, \infty) & \text{if } \alpha = \{2\}, \\ \{0\} & \text{if } \alpha = \{1, 2\}. \end{cases}$$

Note that with $P = 1$, the matrix $A - B_{\bullet\emptyset}(D_{\emptyset\emptyset})^{-1}C_{\emptyset\bullet}$ is *not negative definite*; nevertheless, the assumption in Proposition 3.3(b) is satisfied.

The next example has the same matrix D but has $A = -1$ so that A is negatively stable. Yet the LCS (A, B, C, D) is still not passive because D is not positive semidefinite. This example shows that passivity is not a necessary condition for exponential stability, even with a negatively stable matrix A.

Example 3.2. Consider the tuple with $n = 1$ and $m = 2$:

$$A = -1, \quad B = \begin{bmatrix} 0 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \text{and} \quad D = \begin{bmatrix} 1 & 3 \\ 0 & 1 \end{bmatrix}.$$

By an easy calculation, we have

$$A - B_{\bullet\alpha}(D_{\alpha\alpha})^{-1}C_{\alpha\bullet} = \begin{cases} -1 & \text{if } \alpha = \emptyset, \\ -1 & \text{if } \alpha = \{1\}, \\ -2 & \text{if } \alpha = \{2\}, \\ -2 & \text{if } \alpha = \{1,2\}, \end{cases} \quad \text{and} \quad \mathcal{C}_\alpha = \begin{cases} [\,0, \infty\,) & \text{if } \alpha = \emptyset, \\ \{0\} & \text{if } \alpha = \{1\}, \\ (\,-\infty, 0\,] & \text{if } \alpha = \{2\}, \\ \{0\} & \text{if } \alpha = \{1,2\}. \end{cases}$$

Again, the assumption in Proposition 3.3(b) is satisfied with $P = 1$.

As noted in the proof of Theorem 3.1(b), the strict copositivity of $-N$ on cl Gr $\mathrm{SOL}'_{\mathrm{LCS}}$ is equivalent to the existence of a constant $\rho_N > 0$ such that

$$-z^T N z \;\geq\; \rho_N \,\|\, z \,\|^2 \quad \forall\, z \;\in\; \mathrm{Gr}\ \mathrm{SOL}'_{\mathrm{LCS}}.$$

Involving only Gr $\mathrm{SOL}'_{\mathrm{LCS}}$, the latter inequality avoids the explicit description of the closure of this graph, which is a nontrivial task. We employ this equivalent condition for the strict copositivity of $-N$ in the example below, for which we establish the asymptotic stability of the equilibrium with the choice of a nonzero pair (Q, R) satisfying part (b) of Theorem 3.1, and to which we cannot apply Proposition 3.3(b). This example combines Example 3.1 and the one in [27, section IV]. As such, the matrix A is not negatively stable.

Example 3.3. Consider the LCS

$$\begin{aligned} \dot{x} \;&=\; \begin{bmatrix} -5 & -4 & 0 \\ -1 & -2 & 0 \\ 0 & 0 & 1 \end{bmatrix} x + \begin{bmatrix} -3 & 0 & 0 \\ -21 & 0 & 0 \\ 0 & 2 & -2 \end{bmatrix} u, \\ 0 \;\leq\; u \;&\perp\; \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & -1 \end{bmatrix} x + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 3 \\ 0 & 0 & 1 \end{bmatrix} u \;\geq\; 0. \end{aligned}$$

We claim that there exists no symmetric positive definite matrix P satisfying the assumptions of Proposition 3.3. Consider the two index sets $\alpha = \emptyset$ and $\alpha = \{1\}$. For these sets, we have

$$\mathcal{C}_\emptyset = \{\, x \in \Re^3 \,:\, x_1 \,\geq\, 0 = x_3 \,\}, \quad \mathcal{C}_{\{1\}} = \{\, x \in \Re^3 \,:\, x_1 \,\leq\, 0 = x_3 \,\}$$

and

$$A - B_{\bullet\emptyset}(D_{\emptyset\emptyset})^{-1}C_{\emptyset\bullet} = \begin{bmatrix} -5 & -4 & 0 \\ -1 & -2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

and

$$A - B_{\bullet\{1\}}(D_{\{1\}\{1\}})^{-1}C_{\{1\}\bullet} = \begin{bmatrix} -2 & -4 & 0 \\ 20 & -2 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

By way of contradiction, suppose that there exists a symmetric and positive definite matrix P such that the assumption in Proposition 3.3(b) is satisfied. This would mean that there exists a symmetric positive definite matrix $\bar{P}$ such that

$$\bar{x}^T(\bar{A}_i^T\bar{P} + \bar{P}\bar{A}_i)\bar{x} < 0 \quad \forall\, \bar{x} \in \bar{\mathcal{C}}_i, \tag{3.15}$$

for $i = 1, 2$, where $\bar{\mathcal{C}}_1 \equiv \{\bar{x} \in \Re^2 \mid \bar{x}_1 \geq 0\}$, $\bar{\mathcal{C}}_2 = \{\bar{x} \in \Re^2 \mid \bar{x}_1 \leq 0\}$, and

$$\bar{A}_1 \equiv \begin{bmatrix} -5 & -4 \\ -1 & -2 \end{bmatrix}, \quad \bar{A}_2 \equiv \begin{bmatrix} -2 & -4 \\ 20 & -2 \end{bmatrix}.$$

Since $\bar{\mathcal{C}}_i$ are both half-spaces, the relations (3.15) hold if and only if $\bar{A}_i^T\bar{P} + \bar{P}\bar{A}_i$ are both negative definite for $i = 1, 2$. As shown in [27, section IV], however, this cannot happen. Next, we claim that $x^e = 0$ is an exponentially stable equilibrium of the LCS by verifying that with

$$P \equiv \begin{bmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad Q \equiv 0, \quad \text{and} \quad R \equiv \begin{bmatrix} 9 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

the assumptions in Theorem 3.1 are satisfied. The strict copositivity of M on Gr SOL_{CD}, is not difficult to verify. We briefly sketch the proof of the strict copositivity of the matrix

$$-N = \left[\begin{array}{ccc|ccc|ccc} 10 & 7 & 0 & 3 & 0 & 0 & 0 & 0 & 0 \\ 7 & 12 & 0 & 63 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -2 & 0 & -2 & 2 & 0 & 0 & 0 \\ - & - & - & - & - & - & - & - & - \\ 3 & 63 & 0 & 0 & 0 & 0 & 9 & 0 & 0 \\ 0 & 0 & -2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ - & - & - & - & - & - & - & - & - \\ 0 & 0 & 0 & 9 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

on the closure of Gr $\mathrm{SOL}'_{\mathrm{LCS}}$. We have $u_1(x) = \max(0, -x_1)$, $u_2(x) = \max(0, -x_3)$, and $u_3(x) = \max(0, x_3)$, With the last two rows and columns of N being identically equal to zero, we need not deal with the directional derivatives of u_2 and u_3. Instead, we focus on

$$u_1'(x_1; dx_1) = \begin{cases} 0 & \text{if } x_1 > 0, \\ -dx_1 & \text{if } x_1 < 0, \\ \max(0, -dx_1) & \text{if } x_1 = 0, \end{cases}$$

where $dx_1 = C_{1\bullet}Ax + C_{1\bullet}Bu(x) = -5x_1 - 4x_2 - 3\max(0, -x_1)$. It suffices to show the existence of a constant $\rho_N > 0$ such that

• $x_1 > 0$ implies

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \max(0,-x_3) \\ \max(0,x_3) \end{pmatrix}^T \left[\begin{array}{ccc|cc} 10 & 7 & 0 & 0 & 0 \\ 7 & 12 & 0 & 0 & 0 \\ 0 & 0 & -2 & -2 & 2 \\ \hline 0 & 0 & -2 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \end{array}\right] \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \max(0,-x_3) \\ \max(0,x_3) \end{pmatrix} \geq \rho_N \| x \|^2,$$

• $x_1 < 0$ implies

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ -x_1 \\ \max(0,-x_3) \\ \max(0,x_3) \\ 2x_1 + 4x_2 \end{pmatrix}^T \left[\begin{array}{ccc|ccc|c} 10 & 7 & 0 & 3 & 0 & 0 & 0 \\ 7 & 12 & 0 & 63 & 0 & 0 & 0 \\ 0 & 0 & -2 & 0 & -2 & 2 & 0 \\ \hline 3 & 63 & 0 & 0 & 0 & 0 & 9 \\ 0 & 0 & -2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 9 & 0 & 0 & 0 \end{array}\right] \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ -x_1 \\ \max(0,-x_3) \\ \max(0,x_3) \\ 2x_1 + 4x_2 \end{pmatrix}$$
$$\geq \rho_N \|x\|^2,$$

• and ($x_1 = 0$ implies)

$$\begin{pmatrix} x_2 \\ x_3 \\ \max(0,-x_3) \\ \max(0,x_3) \\ \max(0,4x_2) \end{pmatrix}^T \left[\begin{array}{cc|cc|c} 12 & 0 & 0 & 0 & 0 \\ 0 & -2 & -2 & 2 & 0 \\ \hline 0 & -2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 \end{array}\right] \begin{pmatrix} x_2 \\ x_3 \\ \max(0,-x_3) \\ \max(0,x_3) \\ \max(0,4x_2) \end{pmatrix}$$
$$\geq \rho_N \left\| \begin{pmatrix} x_2 \\ x_3 \end{pmatrix} \right\|^2.$$

We will leave it to the reader to verify that the desired constant ρ_N indeed exists in view of the positive definiteness of certain appropriate matrices.

3.1. Role of observability. The implication (3.8) can be refined by employing an explicit analytic expansion for the vector $z(t,\xi)$ for $t > 0$ sufficiently small. The expansion enables the application of the following known fact about an analytic function expressed in series form.

LEMMA 3.5. *Consider the univariate real-analytic function*

$$\psi(t) \equiv \sum_{j=0}^{\infty} a_j\, t^j, \quad t \geq 0,$$

where $\{a_j\}_{j\geq 0}$ is a given sequence of scalars. The following three statements are valid:

(a) *in order for $\psi(t) > 0$ for all $t > 0$ sufficiently small, it is necessary and sufficient that the sequence of coefficients $\{a_j\}_{j\geq 0}$ be lexicographically positive; i.e., these coefficients are not all zero and the first nonzero coefficient is positive;*

(b) *in order for* $\psi(t) \geq 0$ *for all* $t > 0$ *sufficiently small, it is necessary and sufficient that the sequence of coefficients* $\{a_j\}_{j\geq 0}$ *be lexicographically nonnegative; i.e., either all coefficients are zero or the sequence is lexicographically positive;*

(c) *in order for* $\psi(t) = 0$ *for all* $t > 0$ *sufficiently small, it is necessary and sufficient that* $a_j = 0$ *for all* $j \geq 0$.

If the coefficients a_j *are given by* $e^T G^j \xi$ *for some* n*-vectors* e *and* ξ *and* $n \times n$ *matrix* G*, the above conditions on the infinite sequence* $\{a_j\}_{j\geq 0}$ *can be replaced by the finite sequence* $\{a_j\}_{j=0}^{n-1}$.

For a given pair of matrices $G \in \Re^{k\times k}$ and $H \in \Re^{\ell\times k}$, the unobservable space of (H, G), denoted $\overline{O}(H, G)$, is the set of vectors $\xi \in \Re^k$ such that $HG^j\xi = 0$ for all $j = 0, 1, \ldots, k-1$. In contrast to this linear subspace, the semiunobservable cone of (H, G), denoted $\overline{SO}(H, G)$, is the set of vectors $\xi \in \Re^n$ such that the family of scalars $\{H_{i\bullet}G^j\xi\}_{j=0}^{k-1}$ is lexicographically nonnegative for all $i = 1, \ldots, \ell$. The two sets $\overline{O}(H, G)$ and $\overline{SO}(H, G)$ have played an important role in the observability analysis of the LCS [37]; they have an equally important role here in the asymptotic stability analysis of the LCS. We also define the open subset $SO(H, G)$ of $\overline{SO}(H, G)$ consisting of vectors $\xi \in \Re^n$ such that the family of scalars $\{H_{i\bullet}G^j\xi\}_{j=0}^{k-1}$ is lexicographically positive for all $i = 1, \ldots, \ell$. Note that $0 \notin SO(H, G)$.

The one-sided directional derivative $u'(x(t, x^0); CAx(t, x^0) + CBu(t, x^0))$ is the unique vector $v(t, x^0)$ satisfying

(3.16)

$$\begin{array}{lcll}
\text{free } v_i(t,x^0) & & (\,CAx(t,x^0)+CBu(t,x^0)+Dv(t,x^0)\,)_i = 0, & i \in \alpha(x(t,x^0)),\\
0 \leq v_i(t,x^0) & \perp & (\,CAx(t,x^0)+CBu(t,x^0)+Dv(t,x^0)\,)_i \geq 0, & i \in \beta(x(t,x^0)),\\
0 = v_i(t,x^0), & & (\,CAx(t,x^0)+CBu(t,x^0)+Dv(t,x^0)\,)_i \text{ free}, & i \in \gamma(x(t,x^0)).
\end{array}$$

By a strong non-Zeno result for an LCS with a P-matrix D [37], we deduce the existence of a time $\tau_0 > 0$ and a triple of index sets $(\alpha_{\rm n}, \beta_{\rm n}, \gamma_{\rm n})$, both dependent on the initial condition x^0, such that $(\alpha(x(t, x^0)), \beta(x(t, x^0)), \gamma(x(t, x^0))) = (\alpha_{\rm n}, \beta_{\rm n}, \gamma_{\rm n})$ for all $t \in (0, \tau_0]$. For all such times t, the system (3.16) becomes

$$\begin{array}{lcll}
 & & (\,CAx(t,x^0)+CBu(t,x^0)+Dv(t,x^0)\,)_i = 0, & i \in \alpha_{\rm n},\\
0 \leq v_i(t,x^0) & \perp & (\,CAx(t,x^0)+CBu(t,x^0)+Dv(t,x^0)\,)_i \geq 0, & i \in \beta_{\rm n},\\
0 = v_i(t,x^0), & & & i \in \gamma_{\rm n}.
\end{array}$$

The latter is a mixed LCP of the P-type. As explained in [37], there exist a scalar $\widehat{\tau} \in (0, \tau_0]$ and a subset $\beta_a \subseteq \beta_{\rm n}$ with complement $\bar{\beta}_a \equiv \beta_{\rm n} \setminus \beta_a$ such that for all $t \in (0, \widehat{\tau}\,]$, the unique solution $v(t, x^0)$ of the above mixed LCP satisfies (where $\mathcal{K} \equiv \alpha_{\rm n} \cup \beta_a$)

$$v_{\mathcal{K}}(t, x^0) = -(\,D_{\mathcal{K}\mathcal{K}}\,)^{-1} C_{\mathcal{K}\bullet} \begin{bmatrix} A & B \end{bmatrix} \begin{pmatrix} x(t, x^0) \\ u(t, x^0) \end{pmatrix}.$$

Note that $v_{\beta_a}(t, x^0) \geq 0$, $v_{\bar{\beta}_a}(t, x^0) = 0$, and

$$\left\{\, C_{\bar{\beta}_a\bullet} - D_{\bar{\beta}_a\mathcal{K}}(\,D_{\mathcal{K}\mathcal{K}}\,)^{-1} C_{\mathcal{K}\bullet} \right\} \begin{bmatrix} A & B \end{bmatrix} \begin{pmatrix} x(t, x^0) \\ u(t, x^0) \end{pmatrix} \geq 0.$$

Provided that τ_0 is sufficiently small, we have

$$\text{supp}(u(x^0)) \subseteq \alpha_{\rm n} \subseteq \mathcal{K} \subseteq \alpha_{\rm n} \cup \beta_{\rm n} \subseteq \{\, i : (\,Cx^0 + Du(x^0)\,)_i = 0 \,\}.$$

In terms of the index set $\mathcal{K}$, we have

$$\begin{array}{rl} 0 < \\ 0 = \end{array} \begin{pmatrix} u_{\alpha_{\rm n}}(t,x^0) \\ u_{\beta_a}(t,x^0) \end{pmatrix} = - \begin{bmatrix} D_{\alpha_{\rm n}\alpha_{\rm n}} & D_{\alpha_{\rm n}\beta_a} \\ D_{\beta_a\alpha_{\rm n}} & D_{\beta_a\beta_a} \end{bmatrix}^{-1} \begin{bmatrix} C_{\alpha_{\rm n}\bullet} \\ C_{\beta_a\bullet} \end{bmatrix} x(t,x^0)$$

and

$$\begin{array}{rl} 0 = \\ 0 < \end{array} \left\{ \begin{bmatrix} C_{\bar\beta_a\bullet} \\ C_{\gamma_{\rm n}\bullet} \end{bmatrix} - \begin{bmatrix} D_{\bar\beta_a\alpha_{\rm n}} & D_{\bar\beta_a\beta_t} \\ D_{\gamma_{\rm n}\alpha_{\rm n}} & D_{\gamma_{\rm n}\beta_t} \end{bmatrix} \begin{bmatrix} D_{\alpha_{\rm n}\alpha_{\rm n}} & D_{\alpha_{\rm n}\beta_a} \\ D_{\beta_a\alpha_{\rm n}} & D_{\beta_a\beta_a} \end{bmatrix}^{-1} \begin{bmatrix} C_{\alpha_{\rm n}\bullet} \\ C_{\beta_a\bullet} \end{bmatrix} \right\} x(t,x^0).$$

Substituting the expression for $u_{\mathcal{K}}(t,x^0)$ into the ODE $\dot x = Ax + Bu$ and noting that $u_i(t,x^0) = 0$ for all $i \notin \mathcal{K}$, we deduce

$$x(t,x^0) = \sum_{j=0}^{\infty} \frac{t^j}{j\,!}\, A(\mathcal{K})^j x^0, \quad \overline{C}(\mathcal{K})x(t,x^0) = \sum_{j=0}^{\infty} \frac{t^j}{j\,!}\, \overline{C}(\mathcal{K})A(\mathcal{K})^j x^0,$$

$$\begin{bmatrix} A & B \end{bmatrix} \begin{pmatrix} x(t,x^0) \\ u(t,x^0) \end{pmatrix} = \sum_{j=0}^{\infty} \frac{t^j}{j\,!} \begin{bmatrix} A & B_{\bullet\mathcal{K}} \end{bmatrix} \begin{bmatrix} I \\ \overline{C}_{\mathcal{K}\bullet}(\mathcal{K}) \end{bmatrix} A(\mathcal{K})^j x^0,$$

where $A(\mathcal{K}) \equiv A - B_{\bullet\mathcal{K}}(D_{\mathcal{K}\mathcal{K}})^{-1}C_{\mathcal{K}\bullet}$, and with $\bar{\mathcal{K}} \equiv \{1,\ldots,m\} \setminus \mathcal{K}$,

$$\overline{C}(\mathcal{K}) \equiv \begin{bmatrix} -(\,D_{\mathcal{K}\mathcal{K}}\,)^{-1}C_{\mathcal{K}\bullet} \\ C_{\bar{\mathcal{K}}\bullet} - D_{\bar{\mathcal{K}}\mathcal{K}}(\,D_{\mathcal{K}\mathcal{K}}\,)^{-1}C_{\mathcal{K}\bullet} \end{bmatrix}$$

and

$$\overline{D}(\mathcal{K}) \equiv \overline{C}(\mathcal{K}) \begin{bmatrix} A & B_{\bullet\mathcal{K}} \end{bmatrix} \begin{bmatrix} I \\ \overline{C}_{\mathcal{K}\bullet}(\mathcal{K}) \end{bmatrix}.$$

By Lemma 3.5, in order for $v_{\beta_a}(t,x^0) \geq 0 = u_{\beta_a}(t,x^0)$ to hold for all $t > 0$ sufficiently small, it is necessary and sufficient that $x^0 \in \overline{SO}(\overline{D}_{\beta_a\bullet}(\mathcal{K}), A(\mathcal{K})) \cap \overline{O}(\overline{C}_{\beta_a\bullet}(\mathcal{K}), A(\mathcal{K}))$. Moreover, if $\alpha_{\rm n} \neq \emptyset$, then since $u_{\alpha_{\rm n}}(t,x^0) > 0$ for all $t > 0$ sufficiently small, we must have $x^0 \in SO(\overline{C}_{\alpha_{\rm n}\bullet}(\mathcal{K}), A(\mathcal{K}))$. Similarly, if $\gamma_{\rm n} \neq \emptyset$, we also have $x^0 \in SO(\overline{C}_{\gamma_{\rm n}\bullet}(\mathcal{K}), A(\mathcal{K}))$.

Turning our attention to the implication (3.8), we note that $Nz(t,x^0)$ is equal to

$$\begin{bmatrix} A^TP + PA & PB_{\bullet\mathcal{K}} + A^TQ_{\bullet\mathcal{K}} & Q_{\bullet\mathcal{K}} \\ (B_{\bullet\mathcal{K}})^TP + (Q_{\bullet\mathcal{K}})^TA & (B_{\bullet\mathcal{K}})^TQ_{\bullet\mathcal{K}} + (Q_{\bullet\mathcal{K}})^TB_{\bullet\mathcal{K}} & R_{\mathcal{K}\mathcal{K}} \\ (Q_{\bullet\mathcal{K}})^T & R_{\mathcal{K}\mathcal{K}} & 0 \end{bmatrix} \begin{pmatrix} x(t,x^0) \\ u_{\mathcal{K}}(t,x^0) \\ v_{\mathcal{K}}(t,x^0) \end{pmatrix}$$

$$= \sum_{j=0}^{\infty} \frac{t^j}{j\,!} \begin{bmatrix} A^TP + PA & PB_{\bullet\mathcal{K}} + A^TQ_{\bullet\mathcal{K}} & Q_{\bullet\mathcal{K}} \\ (B_{\bullet\mathcal{K}})^TP + (Q_{\bullet\mathcal{K}})^TA & (B_{\bullet\mathcal{K}})^TQ_{\bullet\mathcal{K}} + (Q_{\bullet\mathcal{K}})^TB_{\bullet\mathcal{K}} & R_{\mathcal{K}\mathcal{K}} \\ (Q_{\bullet\mathcal{K}})^T & R_{\mathcal{K}\mathcal{K}} & 0 \end{bmatrix}$$

$$\times \begin{bmatrix} I \\ \overline{C}_{\mathcal{K}\bullet}(\mathcal{K}) \\ \overline{D}_{\mathcal{K}\bullet}(\mathcal{K}) \end{bmatrix} A(\mathcal{K})^j x^0.$$

Define

$$\overline{N}(\mathcal{K}) \equiv \begin{bmatrix} A^TP + PA & PB_{\bullet\mathcal{K}} + A^TQ_{\bullet\mathcal{K}} & Q_{\bullet\mathcal{K}} \\ (B_{\bullet\mathcal{K}})^TP + (Q_{\bullet\mathcal{K}})^TA & (B_{\bullet\mathcal{K}})^TQ_{\bullet\mathcal{K}} + (Q_{\bullet\mathcal{K}})^TB_{\bullet\mathcal{K}} & R_{\mathcal{K}\mathcal{K}} \\ (Q_{\bullet\mathcal{K}})^T & R_{\mathcal{K}\mathcal{K}} & 0 \end{bmatrix} \begin{bmatrix} I \\ \overline{C}_{\mathcal{K}\bullet}(\mathcal{K}) \\ \overline{D}_{\mathcal{K}\bullet}(\mathcal{K}) \end{bmatrix}.$$

By Lemma 3.5, in order for $Nz(t, x^0) = 0$ for all $t \geq 0$ sufficiently small, it is necessary and sufficient that $x^0 \in \overline{O}(\overline{N}(\mathcal{K}), A(\mathcal{K}))$.

Based on the above discussion, we state and prove the following result which is derived from a refinement of the implication (3.8).

PROPOSITION 3.6. *Let D be a P-matrix. Suppose there exist symmetric matrices P and R and a matrix Q such that M is strictly copositive on* Gr SOL_{CD} *and $-N$ is copositive-plus on* Gr $\mathrm{SOL}'_{\mathrm{LCS}}$. *Assume further that the following two conditions hold for all triples of index sets (α, β, γ) partitioning $\{1, \ldots, m\}$ and for all subsets β_a of β, with $\mathcal{K} \equiv \alpha \cup \beta_a$:*

(a) *for $\alpha = \gamma = \emptyset$,*

$$\overline{SO}(\overline{D}_{\mathcal{K}\bullet}(\mathcal{K}), A(\mathcal{K})) \cap \overline{O}(\overline{C}_{\mathcal{K}\bullet}(\mathcal{K}), A(\mathcal{K})) \cap \overline{O}(\overline{N}(\mathcal{K}), A(\mathcal{K})) = \{0\}, \tag{3.17}$$

(b) *for $\alpha \cup \gamma \neq \emptyset$,*

$$\begin{aligned} &SO(\overline{C}_{\alpha\cup\gamma\bullet}(\mathcal{K}), A(\mathcal{K})) \cap \overline{SO}(\overline{D}_{\beta_a\bullet}(\mathcal{K}), A(\mathcal{K})) \\ &\quad \cap \overline{O}(\overline{C}_{\beta_a\bullet}(\mathcal{K}), A(\mathcal{K})) \cap \overline{O}(\overline{N}(\mathcal{K}), A(\mathcal{K})) = \emptyset; \end{aligned} \tag{3.18}$$

then $x^e = 0$ is an asymptotically stable equilibrium of the LCS (2.1).

Proof. It suffices to show that the implication (3.8) holds. Let ξ satisfy the left-hand side of (3.8). Thus, in particular, $Nz(t, \xi) = 0$ for all $t > 0$ sufficiently small. Following the above argument, we consider the pair of index sets $(\alpha_{\mathrm{n}}, \mathcal{K})$ associated with the trajectories $u(x(t, \xi))$ and $u'(x(t, \xi); dx(t, \xi))$, where $dx(t, \xi) \equiv CAx(t, \xi) + CBu(x(t, \xi))$. The empty intersection (3.18) implies that $\alpha_{\mathrm{n}} = \gamma_{\mathrm{n}} = \emptyset$. Since ξ belongs to the intersection of the three sets in the left-hand side of (3.17), the latter condition then yields $\xi = 0$ as desired. ∎

3.2. A SISO system. We illustrate Proposition 3.6 for a single-input-single-output (SISO) system, which has $m = 1$ and $D = 1$ (the latter is assumed without loss of generality). We write c^T for C and b for B. Thus the SISO LCS is of the form

$$\dot{x} = Ax + b \max(0, -c^Tx). \tag{3.19}$$

In this case, we have $u(x) = \max(0, -c^Tx)$ and

$$\mathrm{SOL}'_{\mathrm{LCS}}(x) = \begin{cases} \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right\} & \text{if } c^Tx > 0, \\ \left\{ \begin{pmatrix} 0 \\ \max(0, -c^TAx) \end{pmatrix} \right\} & \text{if } c^Tx = 0, \\ \left\{ \begin{pmatrix} -c^Tx \\ -c^T(A - bc^T)x \end{pmatrix} \right\} & \text{if } c^Tx < 0. \end{cases}$$

The reader can easily check that Gr $\mathrm{SOL}'_{\mathrm{LCS}}$ is not closed; nevertheless, one can verify that

$$\text{cl Gr SOL}'_{\mathrm{LCS}} = \begin{cases} \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right\} & \text{if } c^T x > 0, \\ \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ -c^T Ax \end{pmatrix} \right\} & \text{if } c^T x = 0, \\ \left\{ \begin{pmatrix} -c^T x \\ -c^T(A - bc^T)x \end{pmatrix} \right\} & \text{if } c^T x < 0. \end{cases}$$

The matrix $M \equiv \left[\begin{smallmatrix} P & q \\ q^T & r \end{smallmatrix} \right]$ is strictly copositive on Gr SOL_{CD} if and only if

$$\left\{ [c^T x \geq 0,\, x \neq 0] \Rightarrow x^T P x > 0 \right\} \text{ and } \left\{ [c^T x < 0] \Rightarrow x^T [P - qc^T - cq^T + rcc^T] x > 0 \right\}.$$

In turn, this holds if and only if P and $P - qc^T - cq^T + rcc^T$ are both positive definite. To see this, suppose that the above two implications hold. If $c^T x < 0$, then $c^T(-x) > 0$; thus $0 < (-x)^T P(-x) = x^T P x$. Hence P must be positive definite. This together with the second implication establishes the positive definiteness of $P - qc^T - cq^T + rcc^T$. The converse is obvious.

The matrix

$$-N \equiv - \begin{bmatrix} A^T P + PA & Pb + A^T q & q \\ b^T P + q^T A & q^T b + b^T q & r \\ q^T & r & 0 \end{bmatrix}$$

is copositive on Gr $\mathrm{SOL}'_{\mathrm{LCS}}$ if and only if

$$c^T x \geq 0 \Rightarrow x^T (A^T P + PA) x \leq 0,$$

$$c^T x \leq 0 \Rightarrow \begin{pmatrix} x \\ -c^T x \\ -c^T(A - bc^T)x \end{pmatrix}^T \begin{bmatrix} A^T P + PA & Pb + A^T q & q \\ b^T P + q^T A & q^T b + b^T q & r \\ q^T & r & 0 \end{bmatrix} \begin{pmatrix} x \\ -c^T x \\ -c^T(A - bc^T)x \end{pmatrix} \leq 0.$$

In turn the above implications hold if and only if $-(A^T P + PA)$ and

(3.20)

$$- \begin{bmatrix} I & -c & -(A^T - cb^T)c \end{bmatrix} \begin{bmatrix} A^T P + PA & Pb + A^T q & q \\ b^T P + q^T A & q^T b + b^T q & r \\ q^T & r & 0 \end{bmatrix} \begin{bmatrix} I \\ -c^T \\ -c^T(A - bc^T) \end{bmatrix}$$

are both positive semidefinite and thus copositive-plus. We examine the two conditions (3.17) and (3.18) in Proposition 3.6. For (3.17) where $\alpha = \gamma = \emptyset$, there are two cases: $\mathcal{K} = \emptyset$ or $\{1\}$. For $\mathcal{K} = \emptyset$, (3.17) stipulates that $\overline{O}(A^T P + PA, A) = \{0\}$. For $\mathcal{K} = \{1\}$, we have

$$\overline{N}(1) = N \begin{bmatrix} I \\ -c^T \\ -c^T(A - bc^T) \end{bmatrix},$$

and the condition (3.17) stipulates that

$$\{0\} = \overline{SO}(-c^T(A-bc^T), A-bc^T) \cap \overline{O}(-c^T, A-bc^T) \cap \overline{O}(\overline{N}(1), A-bc^T)$$
$$= \overline{O}(c^T, A) \cap \overline{O}(\overline{N}(1), A-bc^T) = \overline{O}(c^T, A) \cap \overline{O}(A^TP+PA, A),$$

which is implied by the former case. For (3.18), there are 2 subcases: $\alpha = \{1\}$ or $\gamma = \{1\}$. For $\alpha = \{1\}$, the condition (3.18) stipulates that $SO(-c^T, A-bc^T) \cap \overline{O}(\overline{N}(1), A-bc^T) = \emptyset$. For $\gamma = \{1\}$, the condition (3.18) stipulates that $SO(c^T, A-bc^T) \cap \overline{O}(A^TP+PA, A) = \emptyset$, which is implied by $\overline{O}(A^TP+PA, A) = \{0\}$ because $0 \notin SO(c^T, A-bc^T)$.

Summarizing the above analysis, we present a sufficient condition for $x^e = 0$ to be an asymptotically stable equilibrium of the SISO LCS (3.19).

PROPOSITION 3.7. *If there exist a symmetric positive definite matrix P, a vector q, and a scalar r such that*

(a) $P - qc^T - cq^T + rcc^T$ *is positive definite,*
(b) $-(A^TP+PA)$ *and* (3.20) *are both positive semidefinite,*
(c) $\overline{O}(A^TP+PA, A) = \{0\}$,
(d) $SO(-c^T, A-bc^T) \cap \overline{O}(\overline{N}(1), A-bc^T) = \emptyset$,

then $x^e = 0$ is an asymptotically stable equilibrium of the SISO LCS (3.19). *If the two matrices in* (b) *are positive definite, then $x^e = 0$ is exponentially stable.*

3.3. Extension to non-P systems. In this subsection, we extend Theorem 3.1 to the case where D is not a P-matrix; but we assume the blanket condition that $B\mathrm{SOL}(Cx, D)$ is a singleton for all $x \in \Re^n$. The extension turns out to be technically nontrivial; for one thing, Gr $\mathrm{SOL}'_{\mathrm{LCS}}$ ceases to exist because $\mathrm{SOL}(Cx, D)$ is no longer a single-valued function, and thus we cannot employ its directional derivatives as defined by (3.3). In addition to the main result, Theorem 3.12, we also obtain a stability result for a passive LCS without assuming the P-property of D; see Corollary 3.13.

To carry out the extended analysis, we assume that the matrices Q and R are such that $Q\mathrm{SOL}(Cx, D)$ and $R\mathrm{SOL}(Cx, D)$ are both singletons for all $x \in \Re^n$. Among other things, the single-valuedness of $R\mathrm{SOL}(Cx, D)$ yields the following important property of the quadratic term $\mathrm{SOL}(Cx, D)^T R\mathrm{SOL}(Cx, D)$.

PROPOSITION 3.8. *Let R be a symmetric matrix. Suppose that $R\mathrm{SOL}(Cx, D)$ is a singleton for all $x \in \Re^n$. The function $x \mapsto \mathrm{SOL}(Cx, D)^T R\mathrm{SOL}(Cx, D)$ is a single-valued piecewise quadratic function on $\Re^n$. In other words, for any four vectors $u^i \in \mathrm{SOL}(Cx, D)$, $i = 1, 2, 3, 4$, it holds that $(u^1)^T Ru^2 = (u^3)^T Ru^4$; moreover, this function is continuous in x and there exist finitely many matrices $\{E^j\}_{j=1}^K \subset \Re^{n\times n}$ for some integer $K > 0$ such that $\mathrm{SOL}(Cx, D)^T R\mathrm{SOL}(Cx, D) \in \{x^TE^jx\}_{j=1}^K$ for every $x \in \Re^n$.*

Proof. For any $u^i \in \mathrm{SOL}(Cx, D)$, $i = 1, 2, 3, 4$, we have $Ru^1 = Ru^2 = Ru^3 = Ru^4$. Hence by the symmetry of R, we have

$$(u^1)^T Ru^2 = (u^3)^T Ru^2 = (u^3)^T Ru^4.$$

Next, we show that the function $x \mapsto \mathrm{SOL}(Cx, D)^T R\mathrm{SOL}(Cx, D)$ is continuous. This follows easily from the single-valuedness of this map and the fact that the LCP solution map $q \mapsto \mathrm{SOL}(q, D)$ is *pointwise upper Lipschitz continuous* [13, 15, 44] on $\Re^m$; i.e., for every $q \in \Re^m$, there exist positive scalars c and ε such that

$$\| q' - q \| < \varepsilon \Rightarrow \mathrm{SOL}(q', D) \subseteq \mathrm{SOL}(q, D) + c \| q' - q \| \mathcal{B},$$

where $\mathcal{B}$ is the unit ball in $\Re^m$. Indeed, let $\{x^k\} \subset \Re^n$ be any sequence of vectors converging to some vector $x^\infty \in \Re^n$. Let $\{u^k\} \subset \Re^m$ be such that $u^k \in \mathrm{SOL}(Cx^k, D)$ for every k. By the above continuity property of the LCP solution map, it follows that there exists a corresponding sequence $\{\widehat{u}^k\}$ such that $\widehat{u}^k \in \mathrm{SOL}(Cx^\infty, D)$ for every k and $\lim_{k\to\infty} \| u^k - \widehat{u}^k \| = 0$. By the single-valuedness of $\mathrm{SOL}(Cx^\infty, D)^T R\mathrm{SOL}(Cx^\infty, D)$ and $R\mathrm{SOL}(Cx^\infty, D)$, we can write

$$\begin{aligned}(u^k)^T R u^k &= (\widehat{u}^k)^T R \widehat{u}^k + 2(u^k - \widehat{u}^k)^T R \widehat{u}^k + (u^k - \widehat{u}^k)^T R (u^k - \widehat{u}^k) \\ &= \mathrm{SOL}(Cx^\infty, D)^T R\mathrm{SOL}(Cx^\infty, D) + 2(u^k - \widehat{u}^k)^T R\mathrm{SOL}(Cx^\infty, D) \\ &\quad + (u^k - \widehat{u}^k)^T R (u^k - \widehat{u}^k).\end{aligned}$$

Passing to the limit $k \to \infty$ easily establishes $\lim_{k\to\infty} (u^k)^T R u^k = \mathrm{SOL}(Cx^\infty, D)^T R\mathrm{SOL}\,(Cx^\infty, D)$. Finally, we postpone the identification of the matrices E^j after our description of the structure of $\mathrm{SOL}(Cx, D)$ that immediately follows this proof. □

It is well known that the graph of the set-valued LCP solution map $\mathcal{S}_D : q \mapsto \mathrm{SOL}(q, D)$ is the union of finitely many polyhedra in $\Re^m$; this property is the basis for proving the upper Lipschitz continuity of this map used in the above proof. For the purpose of introducing a closed graph that plays the role of $\mathrm{Gr\ SOL}'_{\mathrm{LCS}}$, which is not available in the non-P case, we first define certain subsets of the polyhedra that compose the graph $\mathrm{Gr\ SOL}_{CD}$. The derivation below is closely related to the development in [39, section 5.1] where we have identified a "linear Newton approximation" for the single-valued map $B\mathrm{SOL}(Cx, D)$.

For every vector $x \in \Re^n$, let $\mathcal{L}(x)$ be the (necessarily nonempty) family of pairs of index subsets α and $\mathcal{J}$ of $\{1, \ldots, m\}$ such that (a) $\alpha \subseteq \mathcal{J}$, (b) the columns of $D_{\mathcal{J}\alpha}$ are linearly independent, and (c) there exists $u \in \mathrm{SOL}(Cx, D)$ such that $\mathrm{supp}(u) \subseteq \alpha$ and $\mathcal{J} \subseteq \{i : (Cx + Du)_i = 0\}$, where $\mathrm{supp}(u) \equiv \{i : u_i > 0\}$ is the *support* of the vector u. Here, we adopt the convention that an empty set of vectors is linearly independent; under this convention, if $0 \in \mathrm{SOL}(Cx, D)$, then $\mathcal{L}(x)$ includes all pairs $(\emptyset, \mathcal{J})$ for all subsets $\mathcal{J} \subseteq \{i : (Cx)_i = 0\}$. For a given pair $(\alpha, \mathcal{J})$ in $\mathcal{L}(x)$, by (b), the solution u in (c) is unique and given by

$$u_\alpha = -\left[\, (D_{\mathcal{J}\alpha})^T D_{\mathcal{J}\alpha} \,\right]^{-1} (D_{\mathcal{J}\alpha})^T C_{\mathcal{J}\bullet} x, \quad u_{\bar{\alpha}} = 0, \tag{3.21}$$

where $\bar{\alpha}$ is the complement of α in $\{1, \ldots, m\}$. Notice that the converse is not true; namely, for a given solution $u \in \mathrm{SOL}(Cx, D)$, it is possible for multiple pairs $(\alpha, \mathcal{J})$ in $\mathcal{L}(x)$ to give rise to the same u, via (3.21). Define the set-valued map

$$\mathcal{G}_{CD} : x \mapsto \mathcal{G}_{CD}(x) \equiv \left\{ \begin{pmatrix} -\left[(D_{\mathcal{J}\alpha})^T D_{\mathcal{J}\alpha}\right]^{-1} (D_{\mathcal{J}\alpha})^T C_{\mathcal{J}\bullet} x \\ 0 \end{pmatrix} : (\alpha, \mathcal{J}) \in \mathcal{L}(x) \right\}.$$

Clearly, $\mathrm{Gr}\ \mathcal{G}_{CD} \subseteq \mathrm{Gr\ SOL}_{CD}$. It is easily seen that $\mathrm{Gr}\ \mathcal{G}_{CD}$ is a cone in $\Re^{n+m}$; subsequently, we will show that it is closed. Like the LCP solution graph, $\mathrm{Gr}\ \mathcal{G}_{CD}$ is not necessarily convex. In general, $\mathrm{Gr}\ \mathcal{G}_{CD}$ is a *proper* subset of $\mathrm{Gr\ SOL}_{CD}$; for instance, if D is a singular matrix, then any positive vector that is a solution of the LCP (Cx, D) is not an element of the former graph. Moreover, due to the finite number of index sets, a positive constant $\rho_G > 0$ exists such that

$$\sup\{\, \| u \| \,:\, u \in \mathcal{G}_{CD}(x) \,\} \le \rho_G \| x \| \quad \forall\, x \in \Re^n. \tag{3.22}$$

For any matrix $W \in \Re^{p\times m}$, define the family

$$\mathcal{T}_W(x) \equiv \left\{ -W_{\bullet\alpha} \left[(D_{\mathcal{J}\alpha})^T D_{\mathcal{J}\alpha}\right]^{-1} (D_{\mathcal{J}\alpha})^T C_{\mathcal{J}\bullet} : (\alpha, \mathcal{J}) \in \mathcal{L}(x) \right\},$$

where, by convention, we define $W_{\bullet\alpha}[(D_{\mathcal{J}\alpha})^T D_{\mathcal{J}\alpha}]^{-1}(D_{\mathcal{J}\alpha})^T C_{\mathcal{J}\bullet}$ to be the zero matrix if $\alpha = \emptyset$. In general,

$$\{\, Ex \,:\, E \in \mathcal{T}_W(x)\,\} \subseteq W\mathrm{SOL}(Cx, D)$$

with equality holding if $W\mathrm{SOL}(Cx, D)$ is a singleton. Suppose that $W\mathrm{SOL}(Cx, D)$ is a singleton for all $x \in \Re^n$. It then follows that the piecewise linear map $h_W(x) \equiv W\mathrm{SOL}(Cx, D)$ is B-differentiable everywhere on $\Re^n$. Thus the directional derivative $h'_W(x; v)$ of h_W at x along the direction v is well defined and, according to standard theory [49], is an element of the set $\{Ev : E \in \mathcal{A}_W(x)\}$, where $\mathcal{A}_W(x) \equiv \{E : Ex = h_W(x)\}$ is the set of *active pieces* of h_W at x. The following result sharpens this representation of $h'_W(x; v)$ by restricting to the pieces in $\mathcal{T}_W(x)$, which is clearly a subfamily of $\mathcal{A}_W(x)$.

PROPOSITION 3.9. *Let $W \in \Re^{p\times m}$ be such that $W\mathrm{SOL}(Cx, D)$ is a singleton for all $x \in \Re^n$. For the piecewise linear function $h_W(x) \equiv W\mathrm{SOL}(Cx, D)$, it holds that $h'_W(x; v) \in \{Ev : E \in \mathcal{T}_W(x)\}$ for all x and v in $\Re^n$.*

Proof. For each $\tau > 0$, $h_W(x+\tau v) = E^\tau(x+\tau v)$, where, for any pair of index sets $(\alpha_\tau, \mathcal{J}_\tau)$ in $\mathcal{L}(x+\tau v)$, $E^\tau \equiv -W_{\bullet\alpha_\tau}[(D_{\mathcal{J}_\tau\alpha_\tau})^T D_{\mathcal{J}_\tau\alpha_\tau}]^{-1}(D_{\mathcal{J}_\tau\alpha_\tau})^T C_{\mathcal{J}_\tau\bullet}$. Thus we have (a) $\alpha_\tau \subseteq \mathcal{J}_\tau$, (b) the columns of $D_{\mathcal{J}_\tau\alpha_\tau}$ are linearly independent, and (c) there exists $u^\tau \in \mathrm{SOL}(C(x+\tau v), D)$ such that $\mathrm{supp}(u^\tau) \subseteq \alpha_\tau$ and $\mathcal{J}_\tau \subseteq \{i : [C(x+\tau v)+Du^\tau]_i = 0\}$. In fact, u^τ is given by (3.21),

$$u^\tau_{\alpha_\tau} = -\left[(D_{\mathcal{J}_\tau\alpha_\tau})^T D_{\mathcal{J}_\tau\alpha_\tau}\right]^{-1}(D_{\mathcal{J}_\tau\alpha_\tau})^T C_{\mathcal{J}_\tau\bullet}(x+\tau v), \quad u^\tau_{\bar\alpha_\tau} = 0,$$

where $\bar\alpha_\tau$ is the complement of α_τ in $\{1, \ldots, m\}$. Let $\{\tau_k\}$ be an arbitrary sequence of positive scalars converging to zero for which there exists a pair $(\alpha_\infty, \mathcal{J}_\infty)$ such that $(\alpha_{\tau_k}, \mathcal{J}_{\tau_k}) = (\alpha_\infty, \mathcal{J}_\infty)$ for all k (there must be at least one such sequence for every pair (x, v) because there are only finitely many pairs of index sets). The corresponding sequence of solutions $\{u^{\tau_k}\}$ converges to a vector, say, u^∞, which must be a solution of the LCP (Cx, D), by the continuity of the latter solution with respect to Cx. Moreover, for all k sufficiently large, we have

$$\mathrm{supp}(u^\infty) \subseteq \mathrm{supp}(u^{\tau_k}) \subseteq \alpha_\infty \subseteq \mathcal{J}_\infty \subseteq \{i : (Cx + Du^\infty)_i = 0\}$$

by a simple limiting argument. Thus the pair $(\alpha_\infty, \mathcal{J}_\infty)$ belongs to $\mathcal{L}(x)$ and $E^{\tau_k} \in \mathcal{T}_W(x)$ for all k sufficiently large. Writing $E^\infty \equiv E^{\tau_k}$ for all such k, we have

$$h_W(x + \tau_k v) - h_W(x) = E^{\tau_k}(x + \tau_k v) - E^\infty x = \tau_k\, E^\infty v,$$

from which we obtain $h'_W(x; v) = E^\infty v$, where $E^\infty \in \mathcal{T}_W(x)$, as desired. ∎

Dealing with a symmetric matrix, the next result completes the proof of Proposition 3.8. For a symmetric $m \times m$ matrix R, define the finite family of symmetric matrices $\widehat{\mathcal{T}}_R(x) \subset \Re^{n\times n}$:

$$\left\{ -(C_{\mathcal{J}\bullet})^T D_{\mathcal{J}\alpha}\left[(D_{\mathcal{J}\alpha})^T D_{\mathcal{J}\alpha}\right]^{-1} R_{\alpha\alpha}\left[(D_{\mathcal{J}\alpha})^T D_{\mathcal{J}\alpha}\right]^{-1}(D_{\mathcal{J}\alpha})^T C_{\mathcal{J}\bullet} : (\alpha, \mathcal{J}) \in \mathcal{L}(x)\right\}.$$

PROPOSITION 3.10. *Let $R \in \Re^{m\times m}$ be symmetric such that $R\mathrm{SOL}(Cx, D)$ is a singleton for all $x \in \Re^n$. For the piecewise quadratic function $\widehat{h}_R(x) \equiv \mathrm{SOL}(Cx, D)^T R\mathrm{SOL}(Cx, D)$, it holds that*

(a) *$\widehat{h}_R(x) = x^T \widehat{E} x$ for all $\widehat{E} \in \widehat{\mathcal{T}}_R(x)$;*

(b) *$\widehat{h}'_R(x; v) \in \{2x^T \widehat{E} v : \widehat{E} \in \widehat{\mathcal{T}}_R(x)\}$ for all x and v in $\Re^n$.*

Proof. It suffices to prove part (b). As a piecewise quadratic function, the directional derivative $\widehat{h}'_R(x;v)$ exists. For each $\tau > 0$, $\widehat{h}_R(x+\tau v) = (x+\tau v)^T \widehat{E}^\tau (x+\tau v)$, where

$$\widehat{E}^\tau \equiv (C_{\mathcal{J}_\tau \bullet})^T D_{\mathcal{J}_\tau \alpha_\tau} \left[(D_{\mathcal{J}_\tau \alpha_\tau})^T D_{\mathcal{J}_\tau \alpha_\tau} \right]^{-1} R_{\alpha_\tau \alpha_\tau} \left[(D_{\mathcal{J}_\tau \alpha_\tau})^T D_{\mathcal{J}_\tau \alpha_\tau} \right]^{-1} (D_{\mathcal{J}_\tau \alpha_\tau})^T C_{\mathcal{J}_\tau \bullet}$$

for any pair of index sets $(\alpha_\tau, \mathcal{J}_\tau) \in \mathcal{L}(x+\tau v)$. As in the proof of Proposition 3.9, we can take a sequence of positive scalars $\{\tau_k\}$ converging to zero and a fixed pair $(\alpha_\infty, \mathcal{J}_\infty)$ such that $(\alpha_{\tau_k}, \mathcal{J}_{\tau_k}) = (\alpha_\infty, \mathcal{J}_\infty)$ for all k. It is now easy to complete the proof. □

We apply the above results to the singled-valued function:

$$\widehat{V}(x) \equiv V(x, \mathrm{SOL}(Cx, D)) = x^T P x + 2x^T Q \mathrm{SOL}(Cx, D) + \mathrm{SOL}(Cx, D)^T R \mathrm{SOL}(Cx, D),$$

assuming that $Q\mathrm{SOL}(Cx, D)$ and $R\mathrm{SOL}(Cx, D)$ are both singletons for all $x \in \Re^n$. Under this assumption, $\widehat{V}(x) = V(x, \mathcal{G}_{CD}(x))$ is piecewise quadratic and

$$\widehat{V}\,'(x;v) = 2x^T P v + 2v^T Q \mathrm{SOL}(Cx, D) + 2x^T E^Q v + 2x^T \widehat{E}^R v,$$

where $E^Q \equiv -Q_{\bullet\alpha} \left[(D_{\mathcal{J}\alpha})^T D_{\mathcal{J}\alpha} \right]^{-1} (D_{\mathcal{J}\alpha})^T C_{\mathcal{J}\bullet} \in \mathcal{T}_Q(x)$ and

$$\widehat{E}^R \equiv (\, C_{\mathcal{J}\bullet})^T D_{\mathcal{J}\alpha} \left[(\, D_{\mathcal{J}\alpha})^T D_{\mathcal{J}\alpha} \right]^{-1} R_{\alpha\alpha} \left[(\, D_{\mathcal{J}\alpha})^T D_{\mathcal{J}\alpha} \right]^{-1} (\, D_{\mathcal{J}\alpha})^T C_{\mathcal{J}\bullet} \in \widehat{\mathcal{T}}_R(x)$$

for some pair $(\alpha, \mathcal{J}) \in \mathcal{L}(x)$; note that we can choose the same pair $(\alpha, \mathcal{J})$ for the directional derivatives of $Q\mathrm{SOL}(Cx, D)$ and $\mathrm{SOL}(Cx, D)^T R\mathrm{SOL}(Cx, D)$ because (cf. the proofs of Propositions 3.9 and 3.10) both derivatives were derived from $\mathrm{SOL}(C(x+\tau v), D)$ corresponding to the same v. Since $Q\mathrm{SOL}(Cx, D) = E^Q x$, we have $\widehat{V}\,'(x;v) = 2x^T[P + E^Q + (E^Q)^T + \widehat{E}^R]v$. Note that the matrix $P + E^Q + (E^Q)^T + \widehat{E}^R$ is symmetric. With $\varphi_{x^0}(t) \equiv \widehat{V}(x(t, x^0))$, we have

$$\begin{aligned} \varphi'_{x^0}(t+) &= \widehat{V}\,'(x(t,x^0); \dot{x}(t,x^0)) \\ &= 2\, x(t,x^0)^T \left[P + E^Q_t + (E^Q_t)^T + \widehat{E}^R_t \right] \left(Ax(t,x^0) + B\mathrm{SOL}(Cx(t,x^0), D) \right), \end{aligned}$$

where the equality in the second line is by a simple substitution, and for each $t > 0$,

$$E^Q_t \equiv -Q_{\bullet\alpha_t} (D_{\alpha_t \mathcal{J}_t} D_{\mathcal{J}_t \alpha_t})^{-1} D_{\alpha_t \mathcal{J}_t} C_{\mathcal{J}_t \bullet} \in \mathcal{T}_Q(x(t, x^0)),$$

and

$$\begin{aligned} \widehat{E}^R_t \equiv (C_{\mathcal{J}_t \bullet})^T D_{\mathcal{J}_t \alpha_t} \left[(D_{\mathcal{J}_t \alpha_t})^T D_{\mathcal{J}_t \alpha_t} \right]^{-1} R_{\alpha_t \alpha_t} \left[(D_{\mathcal{J}_t \alpha_t})^T D_{\mathcal{J}_t \alpha_t} \right]^{-1} \\ \times (D_{\mathcal{J}_t \alpha_t})^T C_{\mathcal{J}_t \bullet} \in \widehat{\mathcal{T}}_R(x(t, x^0)) \end{aligned}$$

for some pair $(\alpha_t, \mathcal{J}_t) \in \mathcal{L}(x(t, x^0))$. Corresponding to any such pair of index sets, letting $z(t, x^0) \equiv (x(t, x^0), u(t, x^0), v(t, x^0))$,

$$\begin{pmatrix} u_{\alpha_t}(t, x^0) \\ u_{\bar{\alpha}_t}(t, x^0) \end{pmatrix} \equiv \begin{pmatrix} -\left[(D_{\mathcal{J}_t \alpha_t})^T D_{\mathcal{J}_t \alpha_t} \right]^{-1} (D_{\mathcal{J}_t \alpha_t})^T C_{\mathcal{J}_t \bullet} x(t, x^0) \\ 0 \end{pmatrix} \in \mathrm{Gr}\mathcal{G}_{CD}(x(t, x^0)),$$

$$\begin{pmatrix} v_{\alpha_t}(t, x^0) \\ v_{\bar{\alpha}_t}(t, x^0) \end{pmatrix} \equiv \begin{pmatrix} -\left[(D_{\mathcal{J}_t \alpha_t})^T D_{\mathcal{J}_t \alpha_t} \right]^{-1} (D_{\mathcal{J}_t \alpha_t})^T C_{\mathcal{J}_t \bullet} (Ax(t, x^0) + B\mathrm{SOL}(Cx(t, x^0), D)) \\ 0 \end{pmatrix},$$

where $\bar{\alpha}_t$ is the complement of α_t in $\{1, \dots, m\}$, we obtain

$$\begin{aligned} Qu(t,x^0) = E_t^Q x(t,x^0), \quad Bu(t,x^0) = B\mathrm{SOL}(Cx(t,x^0),D), \\ x(t,x^0)^T \widehat{E}_t^R (Ax(t,x^0) + B\mathrm{SOL}(Cx(t,x^0),D)) = u(t,x^0)^T Rv(t,x^0), \\ x(t,x^0)^T E_t^Q (Ax(t,x^0) + B\mathrm{SOL}(Cx(t,x^0),D)) = x(t,x^0)^T Qv(t,x^0), \end{aligned}$$

and $\varphi'_{x^0}(t+) = z(t,x^0)^T N z(t,x^0)$, where N is the same matrix defined by (3.6). Note that a constant $\rho_{\widehat{G}} > 0$ exists satisfying

$$\| v(t,x^0) \| \le \rho_{\widehat{G}} \| (x(t,x^0), u(t,x^0)) \| \quad \forall (t,x^0) \in [0,\infty) \times \Re^n. \tag{3.23}$$

Augmenting the map $\mathcal{G}_{CD}$, define

$$\widehat{\mathcal{G}}_{\mathrm{LCS}} : x \mapsto \left\{ \begin{pmatrix} -\left[(D_{\mathcal{J}\alpha})^T D_{\mathcal{J}\alpha} \right]^{-1} (D_{\mathcal{J}\alpha})^T C_{\mathcal{J}\bullet} x \\ 0 \\ -\left[(D_{\mathcal{J}\alpha})^T D_{\mathcal{J}\alpha} \right]^{-1} (D_{\mathcal{J}\alpha})^T C_{\mathcal{J}\bullet} (Ax + B\mathrm{SOL}(Cx,D)) \\ 0 \end{pmatrix} : (\alpha, \mathcal{J}) \in \mathcal{L}(x) \right\}.$$

Note that the pair $(u(t,x^0), v(t,x^0))$ defined above belongs to $\widehat{\mathcal{G}}_{\mathrm{LCS}}(x(t,x^0)) \subset \Re^{2m}$. In what follows, we let $z(t,x^0)$ denote any triple in Gr $\widehat{\mathcal{G}}_{\mathrm{LCS}}$ such that $\varphi'_{x^0}(t+) = z(t,x^0)^T N z(t,x^0)$. We next show that the two graphs Gr $\mathcal{G}_{CD}$ and Gr $\widehat{\mathcal{G}}_{\mathrm{LCS}}$ are closed.

PROPOSITION 3.11. *Both maps $\mathcal{G}_{CD}$ and $\widehat{\mathcal{G}}_{\mathrm{LCS}}$ have closed graphs.*

Proof. We prove the claim only for $\widehat{\mathcal{G}}_{\mathrm{LCS}}$. Let $\{x^k\}$ be a sequence converging to x^∞. For each k, let $(\alpha_k, \mathcal{J}_k) \in \mathcal{L}(x^k)$ be such that

$$\lim_{k\to\infty} \begin{pmatrix} -\left[(D_{\mathcal{J}_k\alpha_k})^T D_{\mathcal{J}_k\alpha_k} \right]^{-1} (D_{\mathcal{J}_k\alpha_k})^T C_{\mathcal{J}_k\bullet} x^k \\ 0 \\ -\left[(D_{\mathcal{J}_k\alpha_k})^T D_{\mathcal{J}_k\alpha_k} \right]^{-1} (D_{\mathcal{J}_k\alpha_k})^T C_{\mathcal{J}_k\bullet} (Ax^k + B\mathrm{SOL}(Cx^k,D)) \\ 0 \end{pmatrix}$$

exists. As in the proof of Proposition 3.9, there exist an infinite subset κ of $\{1,2,\dots\}$ and a pair $(\alpha_\infty, \mathcal{J}_\infty) \in \mathcal{L}(x^\infty)$ such that $(\alpha_k, \mathcal{J}_k) = (\alpha_\infty, \mathcal{J}_\infty)$ for all $k \in \kappa$. Since $B\mathrm{SOL}(Cx,D)$ is continuous in x, the displayed limit is therefore equal to

$$\begin{pmatrix} -\left[(D_{\mathcal{J}_\infty\alpha_\infty})^T D_{\mathcal{J}_\infty\alpha_\infty} \right]^{-1} (D_{\mathcal{J}_\infty\alpha_\infty})^T C_{\mathcal{J}_\infty\bullet} x^\infty \\ 0 \\ -\left[(D_{\mathcal{J}_\infty\alpha_\infty})^T D_{\mathcal{J}_\infty\alpha_\infty} \right]^{-1} (D_{\mathcal{J}_\infty\alpha_\infty})^T C_{\mathcal{J}_\infty\bullet} (Ax^\infty + B\mathrm{SOL}(Cx^\infty,D)) \\ 0 \end{pmatrix}.$$

The closedness of the graph Gr $\widehat{\mathcal{G}}_{\mathrm{LCS}}$ follows. ∎

The above discussion makes it clear that the LCS (2.1) is related to a "linear selectionable DI"; see Smirnov [52, section 8.2]. Nevertheless, there are significant differences between the two kinds of systems; such differences therefore dismiss the applicability of the stability results in the cited reference to the LCS. If $x(t)$ is a

solution of (2.1), then $\dot{x}(t) \in \mathcal{A}(x(t))$, where the set-valued map $\mathcal{A} : \Re^n \to \Re^n$ is given by

$$\mathcal{A}(x) \equiv \{(A+E)x : E \in \mathcal{T}_B(x)\},$$

with the family $\mathcal{T}_B(x)$ being finite and dependent on the state. In contrast, in order for the DI $\dot{x}(t) \in \widehat{\mathcal{A}}(x(t))$ to be *linear selectionable*, there must exist a constant convex compact set $\mathcal{M}$ of real $n \times n$ matrices such that $\widehat{\mathcal{A}}(x) \equiv \{Mx : M \in \mathcal{M}\}$. Clearly, there are noticeable differences between the two sets $\mathcal{A}(x)$ and $\widehat{\mathcal{A}}(x)$; for instance, the latter is always convex, whereas the former consists of only finitely many vectors. In fact, linear selectionable DIs are like hybrid systems with "state independent switchings" [23], and the LCS is a hybrid system with state-triggered switchings.

The following result extends Theorem 3.1 to a non-P matrix D. The same proof applies.

THEOREM 3.12. *Suppose that* $B\mathrm{SOL}(Cx, D)$ *is a singleton for all* $x \in \Re^n$. *Assume further matrices* P, Q, *and* R, *with* P *and* R *symmetric, exist such that*

(A1) $Q\mathrm{SOL}(Cx, D)$ *and* $R\mathrm{SOL}(Cx, D)$ *are singletons for all* $x \in \Re^n$;

(A2) M *is strictly copositive on* $\mathrm{Gr}\ \mathcal{G}_{CD}$.

Let $z(t, x^0)$ *denote any triple in* $\mathrm{Gr}\ \widehat{\mathcal{G}}_{\mathrm{LCS}}$ *such that* $\varphi'_{x^0}(t+) = z(t, x^0)^T N z(t, x^0)$. *The following four statements hold for the equilibrium* $x^e = 0$ *of* (2.1).

(a) *If* $-N$ *is copositive on* $\mathrm{Gr}\ \widehat{\mathcal{G}}_{\mathrm{LCS}}$, *then* x^e *is linearly bounded stable.*

(b) *If* $-N$ *is strictly copositive on* $cl\ \mathrm{Gr}\ \widehat{\mathcal{G}}_{\mathrm{LCS}}$, *then* x^e *is exponentially stable.*

(c) *If* $-N$ *is copositive on* $\mathrm{Gr}\ \widehat{\mathcal{G}}_{\mathrm{LCS}}$ *and* (3.8) *holds, then* x^e *is asymptotically stable.*

(d) *If* $-N$ *is copositive-plus on* $\mathrm{Gr}\ \widehat{\mathcal{G}}_{\mathrm{LCS}}$ *and* (3.9) *holds, then* x^e *is asymptotically stable.*

Complementing Corollary 3.4, the next result is a specialization of the above theorem to a passive LCS.

COROLLARY 3.13. *Assume that* $\mathrm{SOL}(Cx, D) \neq \emptyset$ *for all* $x \in \Re^n$ *and that* $(D + D^T)u = 0 \Rightarrow Bu = 0$. *If the quadruple* (A, B, C, D) *is passive with a passifying matrix* K *such that the only vector* x *for which*

$$\begin{bmatrix} A^T K + KA & KB_{\bullet\alpha} - (C_{\alpha\bullet})^T \\ (B_{\bullet\alpha})^T K - C_{\alpha\bullet} & -D_{\alpha\alpha} - (D_{\alpha\alpha})^T \end{bmatrix} \begin{bmatrix} I \\ -\left[(D_{\mathcal{J}\alpha})^T D_{\mathcal{J}\alpha}\right]^{-1} (D_{\mathcal{J}\alpha})^T C_{\mathcal{J}\bullet} \end{bmatrix} x = 0$$

for some pair $(\alpha, \mathcal{J}) \in \mathcal{L}(x)$ *is the zero vector, then* $x^e = 0$ *is an asymptotically stable equilibrium of the LCS* (2.1).

Proof. Since D is positive semidefinite, the assumption $(D + D^T)u = 0 \Rightarrow Bu = 0$ implies that $B\mathrm{SOL}(Cx, D)$ is a singleton for all $x \in \Re^n$. The remaining proof is similar to that of part (b) of Corollary 3.4 and is not repeated. □

4. An inhomogeneous extension. The stability results in the last section can be extended to a "generalized LCS" [38], which has exactly the same structure as the LCS except that the nonnegative orthant is replaced by an arbitrary polyhedral cone and its dual. Such an extension is significant because the generalized LCS is a much broader class of nonsmooth dynamical system than the LCS; for instance, it includes the case of a mixed LCP to be satisfied by the algebraic variable and also the case of more general linear constraints on the latter variable than nonnegativity. The generalized LCS also arises from the approximation of inhomogeneous (cf. Corollary 4.6) and nonlinear systems (see section 5). All the piecewise linearity properties that we

have employed for the LCP have known extensions to the generalized LCP defined over a polyhedral cone. Based on these extended LCP results, we can easily generalize the Lyapunov stability theory to the generalized LCS without difficulty. The reason we have chosen to focus on the LCS is because this is a fundamental system in its own right with important applications in diverse fields.

Instead of presenting the details of the extended stability results, which will not involve significantly new ideas, we present below a Lyapunov stability theory for an inhomogeneous differential affine system, via a reduction to an equivalent homogeneous system. At the end of the section, we introduce a general reduction approach that paves the way to the treatment of differential nonlinear systems that is the topic of section 5.

Consider the following inhomogeneous LCS with D being a P-matrix:

$$\begin{aligned} \dot{x} &= p + Ax + Bu, \\ 0 \le u \perp q + Cx + Du &\ge 0, \\ x(0) &= x^0, \end{aligned} \tag{4.1}$$

where $p \in \Re^n$ and $q \in \Re^m$ are constant vectors and the other matrices are defined in the same way as before. To avoid triviality, we assume throughout that $(p, q) \neq 0$. By the P-property of the matrix D, we deduce that the unique solution $u(x)$ to the LCP $(q + Cx, D)$ is globally Lipschitz continuous in x; hence, for any $x^0 \in \Re^n$, there exists a unique continuously differentiable solution $x(t, x^0)$ for all $t \ge 0$ satisfying $\dot{x} = p + Ax + Bu(x)$ and $x(0) = x^0$. Since the right-hand side of the latter ODE is not positively homogeneous in x, the solution $x(t, \cdot)$ is no longer positively homogeneous in the initial condition. Therefore, the local asymptotic/exponential stability of an equilibrium of (4.1) does not imply its global asymptotic/exponential stability. Such an equilibrium is a vector $x^e \in \Re^n$ such that $0 = p + Ax^e + Bu(x^e)$. In order to analyze the stability of such a vector x^e, let

$$\begin{aligned} \alpha_e &\equiv \{\, i \,:\, u_i^e > 0 = (\, q + Cx^e + Du^e \,)_i \,\}, \\ \beta_e &\equiv \{\, i \,:\, u_i^e = 0 = (\, q + Cx^e + Du^e \,)_i \,\}, \\ \gamma_e &\equiv \{\, i \,:\, u_i^e = 0 < (\, q + Cx^e + Du^e \,)_i \,\} \end{aligned}$$

be the three fundamental index sets corresponding to the pair (x^e, u^e), where $u^e \equiv u(x^e)$ and define the matrices

$$\begin{aligned} \widehat{A} &\equiv A - B_{\bullet\alpha_e}(D_{\alpha_e\alpha_e})^{-1}C_{\alpha_e\bullet}, & \widehat{B}_{\bullet\beta_e} &\equiv B_{\bullet\beta_e} - B_{\bullet\alpha_e}(D_{\alpha_e\alpha_e})^{-1}D_{\alpha_e\beta_e}, \\ \widehat{C}_{\beta_e\bullet} &\equiv C_{\beta_e\bullet} - D_{\beta_e\alpha_e}(D_{\alpha_e\alpha_e})^{-1}C_{\alpha_e\bullet}, & \widehat{D}_{\beta_e\beta_e} &\equiv D_{\beta_e\beta_e} - D_{\beta_e\alpha_e}(D_{\alpha_e\alpha_e})^{-1}D_{\alpha_e\beta_e}. \end{aligned}$$

We say that x^e is an *isolated* zero of the equation $0 = p + Ax + Bu(x)$ if a neighborhood of x^e exists within which x^e is the only zero of the equation. A similar definition applies to the "isolatedness" of the pair (x^e, u^e) in part (b) of the proposition below.

PROPOSITION 4.1. *Let D be a P-matrix. The following three statements are equivalent.*

(a) *x^e is an isolated zero of the equation $0 = p + Ax + Bu(x)$;*

(b) *the pair (x^e, u^e) is an isolated solution of the mixed LCP in the variables $(x, u) \in \Re^{n+m}$:*

$$\begin{aligned} 0 &= p + Ax + Bu, \\ 0 \le u \perp q + Cx + Du &\ge 0; \end{aligned}$$

(c) *the following homogeneous mixed LCP has a unique solution* $(z, v) = (0, 0)$*:*

$$\begin{aligned} 0 &= \widehat{A}z + \widehat{B}_{\bullet\beta_e} v, \\ 0 \le v \perp \widehat{C}_{\beta_e \bullet} z &+ \widehat{D}_{\beta_e\beta_e} v \ge 0. \end{aligned}$$

Any one of the above three conditions is necessary for x^e *to be an asymptotically stable equilibrium of* (4.1).

Proof. (a) $\Leftrightarrow$ (b). Clearly (a) implies (b). The converse holds by the P-property of D.

(b) $\Leftrightarrow$ (c). This follows from [15, Corollary 3.3.9] and the fact that $D_{\alpha_e\alpha_e}$ is nonsingular.

To see that any one of the three conditions (a)–(c) is necessary for x^e to be an asymptotically stable equilibrium of (4.1), assume for the sake of contradiction that there exists a sequence $\{x^k\}$ of zeros of the equation $0 = p + Ax + Bu(x)$ such that $x^k \neq x^e$ for all k and $\lim_{k\to\infty} x^k = x^e$. Each such zero x^k, for k sufficiently large, defines a stationary trajectory $x^k(t, x^k) = x^k$ for all $t \ge 0$ that violates the asymptotic stability of x^e. □

Next we show that the stability (resp., asymptotic/exponential stability) of the equilibrium x^e of the inhomogeneous LCS (4.1) is equivalent to the linearly bounded stability (resp., global asymptotic/exponential stability) of the equilibrium $z = 0$ of the homogeneous LCS

$$\begin{aligned} \dot{z} &= \widehat{A}z + \widehat{B}_{\bullet\beta_e} v, \\ 0 \le v \perp \widehat{C}_{\beta_e \bullet} z &+ \widehat{D}_{\beta_e\beta_e} v \ge 0, \end{aligned} \tag{4.2}$$

which has a C^1 solution trajectory $z(t, z^0)$ for every initial condition $z^0 = z(0)$. Via this equivalence, the results in the previous sections can then be applied to yield sufficient conditions for the respective stability properties to hold for the inhomogeneous LCS (4.1).

PROPOSITION 4.2. *Let* D *be a P-matrix. The equilibrium* x^e *of the LCS* (4.1) *is stable (resp., asymptotically/exponentially stable) if and only if* $z^e = 0$ *is a linearly boundedly stable (resp., global asymptotically/exponentially stable) equilibrium of the homogeneous LCS* (4.2).

Proof. Sufficiency. Suppose that $z^e = 0$ is a linearly boundedly stable equilibrium of the homogeneous LCS (4.2). Hence there exists a constant $\eta > 0$ such that for all solution trajectory $z(t, z^0)$ of (4.2) satisfying $z(0, z^0) = z^0$, it holds that $\|z(t, z^0)\| \le \eta\|z^0\|$ for all $(t, z^0) \in [0, \infty) \times \Re^n$. We need to show that for every $\varepsilon > 0$, a constant $\delta_\varepsilon > 0$ exists such that for all $\|x^0 - x^e\| < \delta_\varepsilon \Rightarrow \limsup_{t\ge 0} \|x(t, x^0) - x^e\| < \varepsilon$. The proof lies in showing that for x^0 sufficiently close to x^e, the trajectory $\widehat{z}(t, z^0) \equiv x(t, x^0) - x^e$, which has $\widehat{z}(0, z^0) = x^0 - x^e \equiv z^0$, is a solution of the homogeneous LCS (4.2). Once the latter claim is established, the stability of x^e follows; so do the asymptotic and exponential stability. To prove the claim, let x^0 be given and let $(z(t, z^0), v(t, z^0))$ be the unique solution trajectory of (4.2) satisfying $z(0, z^0) = z^0$; it suffices to show that for all x^0 sufficiently close to x^e, $\widehat{z}(t, z^0) = z(t, z^0)$ for all $t \ge 0$. We do this by producing a suitable trajectory $\widehat{u}(t, x^0)$ such that the pair $(z(t, z^0) + x^e, \widehat{u}(t, x^0))$ satisfies (4.1); by the uniqueness of the solution to the latter LCS, we then deduce $\widehat{z}(t, z^0) = z(t, z^0)$ for all $t \ge 0$ as desired. In turn, to produce

the $\widehat{u}(t,x^0)$ trajectory, let $\widehat{u}_{\beta_e}(t,x^0) \equiv v(t,z^0)$, $\widehat{u}_{\gamma_e}(t,x^0) \equiv 0$, and

$$\begin{aligned}\widehat{u}_{\alpha_e}(t,x^0) &\equiv -(D_{\alpha_e\alpha_e})^{-1}\left[q_{\alpha_e} + C_{\alpha_e\bullet}(z(t,z^0)+x^e) + D_{\alpha_e\beta_e}\widehat{u}_{\beta_e}(t,x^0)\right] \\ &= u^e_{\alpha_e} - (D_{\alpha_e\alpha_e})^{-1}\left[C_{\alpha_e\bullet}z(t,z^0) + D_{\alpha_e\beta_e}v(t,x^0)\right].\end{aligned}$$

We have

$$\begin{aligned}&q_{\beta_e} + C_{\beta_e\bullet}(z(t,z^0)+x^e) + D_{\beta_e\alpha_e}\widehat{u}_{\alpha_e}(t,x^0) + D_{\beta_e\beta_e}\widehat{u}_{\beta_e}(t,x^0) \\ &\quad = \widehat{C}_{\beta_e\bullet}z(t,z^0) + \widehat{D}_{\beta_e\beta_e}v(t,z^0)\end{aligned}$$

and

$$\begin{aligned}&q_{\gamma_e} + C_{\gamma_e\bullet}(z(t,z^0)+x^e) + D_{\gamma_e\alpha_e}\widehat{u}_{\alpha_e}(t,x^0) + D_{\gamma_e\beta_e}\widehat{u}_{\beta_e}(t,x^0) \\ &\quad = q_{\gamma_e} + C_{\gamma_e\bullet}x^e + D_{\gamma_e\alpha_e}u^e_{\alpha_e} + \widehat{C}_{\gamma_e\bullet}z(t,z^0) + \widehat{D}_{\gamma_e\beta_e}v(t,x^0),\end{aligned}$$

where $\widehat{C}_{\gamma_e\bullet} \equiv C_{\gamma_e\bullet} - D_{\gamma_e\alpha_e}(D_{\alpha_e\alpha_e})^{-1}C_{\alpha_e\bullet}$ and $\widehat{D}_{\gamma_e\beta_e} \equiv D_{\gamma_e\beta_e} - (D_{\alpha_e\alpha_e})^{-1}D_{\alpha_e\beta_e}$. Note that both $u^e_{\alpha_e}$ and $q_{\gamma_e} + C_{\gamma_e\bullet}x^e + D_{\gamma_e\alpha_e}u^e_{\alpha_e}$ are positive. Being the Schur complement of a P-matrix, $\widehat{D}_{\beta_e\beta_e}$ is itself a P-matrix. Hence there exists a constant $L_v > 0$ such that

$$\|\, v(t,z^0)\| \le L_v \,\|\, z(t,z^0)\,\| \;\le\; L_v\,\eta\,\|\, z^0\,\| \quad \forall\, t \ge 0,$$

where the second inequality is by the linearly bounded stability of the equilibrium $z^e = 0$ for the homogeneous LCS (4.2). Consequently, provided that x^0 is sufficiently close to x^e, or equivalently, that z^0 is sufficiently close to the origin, $\widehat{u}_{\alpha_e}(t,x^0)$ and $q_{\gamma_e} + C_{\gamma_e\bullet}(z(t,z^0)+x^e) + D_{\gamma_e\alpha_e}\widehat{u}_{\alpha_e}(t,x^0) + D_{\gamma_e\beta_e}\widehat{u}_{\beta_e}(t,x^0)$ remain positive for all $t \ge 0$. Hence for all such x^0, $\widehat{u}(t,x^0) \in \mathrm{SOL}(q + C(z(t,z^0)+x^e), D)$ for all $t \ge 0$.

Since $0 = p + Ax^e + Bu^e = p + Ax^e + B_{\bullet\alpha_e}u^e_{\alpha_e} = p + Ax^e - B_{\bullet\alpha_e}(D_{\alpha_e\alpha_e})^{-1}C_{\alpha_e\bullet}x^e$, we have

$$\begin{aligned}\frac{d(z(t,z^0)+x^e)}{dt} &= \widehat{A}z(t,z^0) + \widehat{B}v(t,z^0) \\ &= [\,A - B_{\bullet\alpha_e}(\,D_{\alpha_e\alpha_e}\,)^{-1}C_{\alpha_e\bullet}\,]z(t,z^0) \\ &\quad + [\,B_{\bullet\beta_e} - B_{\bullet\alpha_e}(\,D_{\alpha_e\alpha_e}\,)^{-1}D_{\alpha_e\beta_e}\,]v(t,z^0) \\ &= p + A(\,z(t,z^0)+x^e) + B\widehat{u}(t,x^0).\end{aligned}$$

We have therefore verified all the required conditions for the pair $(z(t,z^0)+x^e, \widehat{u}(t,x^0))$ to be a solution of (4.1). This establishes the sufficiency part of the proposition.

Necessity. Suppose that x^e is a stable equilibrium of the LCS (4.1). We may choose $\varepsilon > 0$ sufficiently small such that for all x satisfying $\|x - x^e\| < \varepsilon$, we have $u_{\alpha_e}(x) > 0$ and $(q + Cx + Du(x))_{\gamma_e} > 0$. Corresponding to such an ε, let $\delta_\varepsilon > 0$ be such that $\|x^0 - x^e\| < \delta_\varepsilon \Rightarrow \|x(t,x^0) - x^e\| < \varepsilon$ for all $t \ge 0$. Consequently, for any such x^0, we have $[q + Cx(t,x^0) + Du(x(t,x^0))]_{\alpha_e} = 0$ and $u_{\gamma_e}(x(t,x^0)) = 0$. Since $(q + Cx^e + Du^e)_{\alpha_e} = 0$, we deduce

$$C_{\alpha_e\bullet}(\,x(t,x^0) - x^e) + D_{\alpha_e\alpha_e}(\,u(x(t,x^0)) - u^e\,)_{\alpha_e} + D_{\alpha_e\beta_e}u_{\beta_e}(t,x^0) = 0,$$

which yields

$$(u(x(t,x^0)) - u^e)_{\alpha_e} = -(D_{\alpha_e\alpha_e})^{-1}\left[C_{\alpha_e\bullet}(x(t,x^0) - x^e) + D_{\alpha_e\beta_e}u_{\beta_e}(t,x^0)\right]. \tag{4.3}$$

Substituting this and using $(q + Cx^e + Du^e)_{\beta_e} = 0$, we deduce

$$[\, q + Cx(t, x^0) + Du(x(t, x^0))\,]_{\beta_e} = \widehat{C}_{\beta_e \bullet}(\, x(t, x^0) - x^e) \, + \widehat{D}_{\beta_e \beta_e} u_{\beta_e}(t, x^0).$$

Hence $u_{\beta_e}(t, x^0)$ satisfies

$$0 \, \leq \, u_{\beta_e}(t, x^0) \, \perp \, \widehat{C}_{\beta_e \bullet}(\, x(t, x^0) - x^e) \, + \widehat{D}_{\beta_e \beta_e} u_{\beta_e}(t, x^0) \, \geq \, 0$$

for all $t \geq 0$. Furthermore,

$$\begin{aligned} \frac{d(x(t, x^0) - x^e)}{dt} &= p + Ax(t, x^0) + Bu(x(t, x^0)) \\ &= A(x(t, x^0) - x^e) + B_{\bullet \alpha_e}(u(t, x^0) - u^e)_{\alpha_e} + B_{\bullet \beta_e} u_{\beta_e}(t, x^0) \\ &= \widehat{A}(x(t, x^0) - x^e) + \widehat{B}_{\bullet \beta_e} u_{\beta_e}(t, x^0). \end{aligned}$$

Therefore, by the uniqueness of the solution trajectory to (4.2), we deduce that $z(t, z^0) \equiv x(t, x^0) - x^e$ is the unique solution trajectory satisfying (4.2) and $z(0, z^0) = z^0 \equiv x^0 - x^e$, along with the auxiliary algebraic trajectory $v(t, z^0) \equiv u_{\beta_e}(x(t, x^0))$. Consequently, the stability, and thus the linearly bounded stability, of the equilibrium $z^e = 0$ for (4.2) follows readily; so do the global asymptotic and global exponential stability, provided that the equilibrium x^e is, respectively, asymptotically and exponentially stable for (4.1). □

An interesting special case occurs when x^e is *nondegenerate*; i.e., when the index set β_e is empty. In this case, for all x sufficiently close to x^e, the LCP $(q + Cx, D)$ is equivalent to a system of linear equations. As such, intuitively speaking, the stability of x^e can be established via classical system-theoretic results. A formal statement of this assertion is presented below whose proof follows easily from Proposition 4.2.

COROLLARY 4.3. *Let D be a P-matrix. Suppose that the equilibrium x^e of the LCS* (4.1) *is nondegenerate. The following statements are equivalent.*

(a) *x^e is asymptotically stable;*
(b) *x^e is exponentially stable;*
(c) *the matrix $\widehat{A}$ is negatively stable, i.e., there exists a symmetric positive definite matrix K such that $\widehat{A}^T K + K\widehat{A}$ is negative definite.*

Proof. If x^e is nondegenerate, then the system (4.2) becomes the ODE: $\dot{z} = \widehat{A}z$, whose unique solution is given by $z(t, z^0) = e^{t\widehat{A}} z^0$ for all $t \geq 0$. The conclusion of the corollary now follows from classical linear systems theory and Proposition 4.2. □

The proof of Proposition 4.2 can be significantly simplified, and in fact, the proposition itself can be extended considerably, by exploiting an approximation property of a piecewise affine function. In spite of the generalization discussed below, the proof given above is of interest for several reasons: one, it helps us to understand the generalized result; two, it expresses the reduced homogeneous system (4.2) in a form that enables a direct application of the results in section 3, and three, this reduction argument can be extended to a nonlinear complementarity system.

The following lemma is the cornerstone of the generalization of Proposition 4.2. It extends an obvious global property of affine functions to a local property of piecewise affine functions. For a proof of the lemma, see section 2.2.2 (particularly expression (2.2)) in [49] and [15, Exercise 4.8.10].

LEMMA 4.4 (Scholtes). () *Let $f : \Re^n \to \Re^m$ be a piecewise affine function. For every $x \in \Re^n$, there exists a neighborhood $\mathcal{N}_x$ of x such that $f(y) = f(x) + f'(x; y - x)$ for all $y \in \mathcal{N}_x$.*

Notice that the directional derivative $f'(x;\cdot)$ is a piecewise linear function of the second argument; in particular, it is positively homogeneous. In general, a piecewise affine function is not differentiable. Thus the ODE $\dot{x} = f(x)$ has a nonsmooth right-hand side. The following result is the promised generalization of Proposition 4.2; the proof is essentially an abstraction of that of the cited proposition.

PROPOSITION 4.5. *Let $f : \Re^n \to \Re^n$ be a piecewise affine function with $f(x^e) = 0$. The equilibrium x^e is stable (resp., asymptotically/exponentially stable) for the ODE $\dot{x} = f(x)$ if and only if $z^e = 0$ is a linearly boundedly stable (resp., asymptotically/exponentially stable) equilibrium of the ODE $\dot{z} = f'(x^e; z)$.*

Proof. Since f is piecewise affine on $\Re^n$, it is globally Lipschitz continuous there. Hence the initial-value ODE

$$\dot{x} = f(x), \quad x(0) = x^0 \tag{4.4}$$

has a unique solution $x(t, x^0)$ for all $x^0 \in \Re^n$. The same is true of the ODE

$$\dot{z} = f'(x^e; z), \quad z(0) = z^0 \tag{4.5}$$

for all $z^0 \in \Re^n$. Suppose that x^e is a locally stable equilibrium of the ODE $\dot{x} = f(x)$. Let $\varepsilon > 0$ be such that $f(x) = f'(x^e; x - x^e)$ for all x satisfying $\|x - x^e\| < \varepsilon$. Corresponding to this ε, let $\delta_\varepsilon > 0$ be such that $\|x^0 - x^e\| < \delta_\varepsilon \Rightarrow \|x(t, x^0) - x^e\| < \varepsilon$ for all $t \geq 0$. It follows that $z(t, z^0) \equiv x(t, x^0) - x^e$ is the unique solution trajectory of (4.5) satisfying $z(0, z^0) = z^0 \equiv x^0 - x^e$. Hence $z^e = 0$ is a linearly bounded stable equilibrium of (4.5), by the positive homogeneity of $f'(x^e; \cdot)$. The other assertions of the proposition can be proved similarly. □

Instead of showing how Proposition 4.2 is a special instance of Proposition 4.5, we consider the more general inhomogeneous DAVI,

$$\begin{array}{l} \dot{x} = p + Ax + Bu, \\ u \in \mathrm{SOL}(K, q + Cx, D), \end{array} \tag{4.6}$$

where K is a polyhedron in $\Re^m$. We assume that the pair (K, D) is "coherently oriented" [46, 15]. This condition is necessary and sufficient for the AVI $(K, \widehat{q}, D)$ to have a unique solution for all vectors $\widehat{q} \in \Re^m$; moreover, under this condition, such a solution function is necessarily a piecewise affine function of $\widehat{q}$. Hence, letting $u(x)$ be the unique element of $\mathrm{SOL}(K, q + Cx, D)$, the DAVI (4.6) is equivalent to the ODE with a piecewise affine right-hand side: $\dot{x} = p + Ax + Bu(x)$. (Incidentally, this equivalence remains valid if the coherent orientation of the pair (K, D) is weakened to the condition that $B\mathrm{SOL}(K, q + Cx, D)$ is a singleton for all $x \in \Re^n$; nevertheless this weakening necessitates a modification of the following discussion about the directional derivatives, which becomes much more involved. For simplicity, we continue to assume the coherent orientation condition.) If (K, D) is coherently oriented, then the directional derivative $u'(x; dx)$ of the solution function $u(x)$ along a direction $dx \in \Re^n$ is the unique solution v to the generalized LCP

$$\mathcal{C}(x) \ni v \perp Cdx + Dv \in \mathcal{C}(x)^*,$$

where $\mathcal{C}(x)$ is the "critical cone" of the AVI $(K, q + Cx, D)$ at the solution $u(x)$, and $\mathcal{C}(x)^*$ is the dual of $\mathcal{C}(x)$; specifically, $\mathcal{C}(x) \equiv \mathcal{T}(K; u(x)) \cap (q + Cx + Du(x))^\perp$, where $\mathcal{T}(K; u(x))$ denotes the tangent cone of K at $u(x) \in K$ (as in convex analysis [47]) and the superscript denotes the orthogonal complement. It should be pointed out

that both $\mathcal{C}(x)$ and its dual are polyhedral cones. For details of these results, we refer the reader to [15, Volume I, section 4.3].

Applying Proposition 4.5 to the DAVI (4.6), we obtain the following result, which requires no further proof.

COROLLARY 4.6. *Let K be a polyhedron in $\Re^m$. Suppose that the pair (K, D) is coherently oriented. Let x^e satisfy $0 = p + Ax^e + Bu(x^e)$. The equilibrium x^e of* (4.6) *is stable (resp., asymptotically/exponentially stable) if and only if $z^e = 0$ is a linearly boundedly stable (resp., asymptotically/exponentially stable) equilibrium of the differential complementarity system*

$$\begin{aligned} \dot{z} &= Az + Bv, \\ \mathcal{C}(x^e) \ni v \perp Cz + Dv &\in \mathcal{C}(x^e)^*, \end{aligned} \tag{4.7}$$

where $\mathcal{C}(x^e) \equiv \mathcal{T}(K; u(x^e)) \cap (q + Cx^e + u(x^e))^\perp$. □

As mentioned in the beginning of this section, it is possible to extend the Lyapunov stability results for the LCS to the generalized LCS (4.7). Instead of repeating the derivation, we proceed to the other major topic of this paper, to be addressed in the next section. There, we establish a partial generalization of Proposition 4.2 and Corollary 4.6 that deals with the exponential stability of nonlinear systems; see Propositions 5.7 and 5.10.

5. Exponential stability of nonlinear systems via a converse theorem. So far our development has been restricted to systems with linear structures. In this section, we extend our treatment to nonlinear systems via the so-called *Lyapunov indirect method* of "first-order approximation." The results in this section are of the exponential stability type. Due to the nonsmoothness of the solution function to the LCP/AVI, it seems difficult to develop an asymptotic stability theory for nonlinear systems without relying on exponential stability.

The cornerstone of the extended treatment of nonlinear systems is a converse theorem for the exponential stability of an equilibrium to an ODE with a B-differentiable right-hand side that is not F(réchet)-differentiable. In general, if the right-hand side of the ODE is not F-differentiable, the solution map of the ODE is not a differentiable function of the initial condition; nevertheless, the latter map remains B-differentiable, provided that the right-hand function of the ODE is so. This is formally stated in the following result whose proof can be found in the recent paper [39, Theorem 7].

LEMMA 5.1. *Suppose that for a given $\xi \in \Re^n$, f is B-differentiable in a neighborhood of a solution trajectory $x(t, \xi)$ of the ODE* (4.4) *for $t \in [0, T]$. For each $t \in [0, T]$, the solution map $x(t, \cdot)$ of the ODE* (4.4) *is B-differentiable at ξ; the directional derivative*

$$x'_\xi(t, \xi; \eta) \equiv \lim_{\tau \downarrow 0} \frac{x(t, \xi + \tau\eta) - x(t, \xi)}{\tau}$$

of $x(t, \cdot)$ at ξ along the direction η is the unique solution $y(t)$ to the variational equation $\dot{y}(t) = f'(x(t, \xi); y(t))$, $y(0) = \eta$.

The following result gives a necessary and sufficient condition for an equilibrium of the ODE (4.4) to be exponentially stable in terms of the existence of a nonsmooth Lyapunov function satisfying certain conditions. Since the latter function is not necessarily differentiable, the result does not follow from standard system theory; see, e.g., [28, Chapter 3]. Moreover, whereas the proof is inspired by that of Theorem 3.12 in the cited reference, some details are different as the Lyapunov function is not

continuously differentiable. In particular, conditions (b) and (c) are normally stated in terms of the F-derivatives of V; here they are expressed in terms of directional derivatives.

THEOREM 5.2. *Suppose that f is Lipschitz continuous in a neighborhood $\mathcal{N}_0$ of the origin and that $f(0) = 0$. The following two statements hold.*

(I) *If there exist positive constants $c_1 < c_2$, and c_3, a neighborhood $\mathcal{N} \subseteq \mathcal{N}_0$ of $x^e = 0$, and a Lipschitz continuous and directionally differentiable function V in $\mathcal{N}$ such that*
  (a) $c_1\|x^0\|^2 \leq V(x^0) \leq c_2\|x^0\|^2$ *for all* $x^0 \in \mathcal{N}$,
  (b) $V'(x^0; f(x^0)) \leq -c_3\|x^0\|^2$ *for all* $t \geq 0$ *and all* $x^0 \in \mathcal{N}$,
  then $x^e = 0$ is an exponentially stable equilibrium of the ODE (4.4).

(II) *Conversely, if $x^e = 0$ is an exponentially stable equilibrium of the* ODE (4.4) *and if f is additionally directionally differentiable in $\mathcal{N}_0$, then there exist positive constants c_1, c_2, c_3, and c_4, a neighborhood $\mathcal{N} \subseteq \mathcal{N}_0$ of $x^e = 0$, and a Lipschitz continuous and directionally differentiable function V in $\mathcal{N}$ such that (a), (b), and (c) hold, where*
  (c) $|V'(x^0; z) - V'(x^0; z')| \leq c_4\|x^0\|\|z - z'\|$ *for all* $x^0 \in \mathcal{N}$ *and all* z, z' *in* $\Re^n$.

Proof. Without loss of generality, we take $\mathcal{N}$ to be an open ball centered at the origin and with radius $r > 0$. We claim that under the assumption in (I), by defining the neighborhood

$$\mathcal{N}' \equiv \{\, z \in \Re^n \,:\, \|\, z \,\| \,\leq \sqrt{c_1/c_2}\, r/2 \,\},$$

a unique solution trajectory $x(t, x^0)$ exists satisfying the ODE (4.4) for all $t \geq 0$ and all $x^0 \in \mathcal{N}'$; moreover, $\|x(t, x^0)\| < r/2$ for all such pairs (t, x^0). Notice that the existence and uniqueness of such a trajectory do not follow directly from basic ODE theory because f is assumed to be Lipschitz continuous only in $\mathcal{N}_0$ and not everywhere. Let $x^0 \in \mathcal{N}'$; clearly $\|x^0\| < r/2$ because $c_1 < c_2$. Hence there is a time $t_0 > 0$ such that the trajectory $x(t, x^0)$ exists and is unique for all $t \in [0, t_0]$. We claim that $\|x(t, x^0)\| < r/2$ for all t in the domain of definition of the trajectory. Assume for the sake of contradiction that there exists $\widetilde{t} \in (0, t_0]$ such that $\|x(\widetilde{t}, x^0)\| = r/2$ and that $\|x(t, x^0)\| < r/2$ for all $t \in [0, \widetilde{t}\,)$. For all $\varepsilon > 0$ sufficiently small, we can write

$$\begin{aligned} V(x(\widetilde{t}, x^0)) - V(x^0) = &\int_0^{\widetilde{t}-\varepsilon} V'(x(s, x^0); f(x(s, x^0)))\, ds \\ &+ \int_{\widetilde{t}-\varepsilon}^{\widetilde{t}} V'(x(s, x^0); f(x(s, x^0)))\, ds < 0, \end{aligned}$$

where the first summand in the right-hand side is nonpositive by (b) and the second summand is negative because $\|x(s, x^0)\|$ is near $r/2 > 0$ for all $s \in [\widetilde{t} - \varepsilon, \widetilde{t}]$. Hence,

$$c_1 \,\|\, x(\widetilde{t}, x^0) \,\|^2 \,\leq\, V(x(\widetilde{t}, x^0)) < V(x^0) \,\leq\, c_2 \,\|\, x^0 \,\|^2,$$

which implies $\|x(\widetilde{t}, x^0)\|^2 < r^2/4$, which is a contradiction. Thus, $\|x(t, x^0)\| < r/2$ for all $t \in [0, t_0]$. Let

$$\begin{aligned} t_* \,\equiv\, \sup\{\, \bar{t} \,\geq\, t_0 \,:\;& \text{the trajectory } x(t, x^0) \text{ exists, is unique,} \\ & \text{and satisfies } \|x(t, x^0)\| < r/2 \text{ for all } t \in [\,0, \bar{t}\,]\,\}. \end{aligned}$$

It follows that there exists $\varepsilon > 0$ such that for all $\widetilde{t} \in [0, t_*)$, the trajectory $x(t, x^0)$ can be continued beyond time $\widetilde{t}$ for at least ε duration. Since ε is independent of $\widetilde{t}$, we must have $t_* = \infty$. Hence, the trajectory $x(t, x^0)$ exists, is unique, and remains in $\mathcal{N}$ for all $t \geq 0$. By condition (b), the trajectory $x(t, x^0)$ must satisfy $V'(x(t, x^0); f(x(t, x^0))) \leq -c_3 \|x(t, x^0)\|^2$ for all $t \geq 0$ and all $x^0 \in \mathcal{N}'$. From this point on, we can follow the same line of proof as in Theorem 3.1(b) to complete the proof of the exponential stability of x^e. This establishes part (I) of the theorem.

Conversely, to show (II), let $\mathcal{N} \subseteq \mathcal{N}_0$ be a subneighborhood of the equilibrium such that for some positive constants ν and κ, $\|x(t, x^0)\| \leq \kappa e^{-\nu t} \|x^0\|$ for all $t \geq 0$ and all $x^0 \in \mathcal{N}$ and that $x(t, x^0) \in \mathcal{N}_0$ for all such pairs (t, x^0). Define

$$V(z) \equiv \int_0^T x(\tau, z)^T x(\tau, z)\, d\tau, \quad z \in \mathcal{N},$$

where the upper limit $T > 0$ will be determined later. It is clear that V is Lipschitz continuous in $\mathcal{N}$. To show that V is directionally differentiable, we need to show that the limit

$$\lim_{\tau \downarrow 0} \frac{V(z + \tau h) - V(z)}{\tau}$$

exists for all $h \in \Re^n$. We have

$$V(z + \tau h) - V(z) = \int_0^T \left[\left(x(s, z + \tau h) - x(s, z) \right)^T \left(x(s, z + \tau h) + x(s, z) \right) \right] ds.$$

By the Lipschitz property of $x(\tau, \cdot)$ and the exponential bound of $x(\tau, z)$, it follows by the Lebesgue convergence theorem that we can interchange the integral with the limit as $\tau \downarrow 0$ and obtain

$$\lim_{\tau \downarrow 0} \frac{V(z + \tau h) - V(z)}{\tau} = 2 \int_0^T x'_\xi(s, z; h)^T x(s, z)\, ds,$$

where we have used Lemma 5.1 to justify the well-definedness of the directional derivative $x'_\xi(\tau, z; h)$ (this is where the directional differentiability of f is needed). In particular, we have

$$V'(x^0; f(x^0)) = 2 \int_0^T x'_\xi(s, x^0; f(x^0))^T x(s, x^0)\, ds.$$

By Lemma 5.1, $x'_\xi(s, x^0; f(x^0))$ is the unique function $y(s)$ satisfying $\dot{y}(s) = f'(x(s, x^0); y(s))$ and $y(0) = f(x^0)$. It is easy to verify that the function $y(s) \equiv f(x(s, x^0))$ satisfies the latter initial-value ODE because $\dot{x}(s, x^0) = f(x(s, x^0))$. Hence $x'_\xi(s, x^0; f(x^0)) = f(x(s, x^0))$; thus

$$\begin{aligned} V'(x^0; f(x^0)) &= 2 \int_0^T f(x(\tau, x^0))^T x(\tau, x^0)\, d\tau \\ &= 2 \int_0^T \dot{x}(\tau, x^0)^T x(\tau, x^0)\, d\tau = \left[\| x(T, x^0) \|^2 - \| x^0 \|^2 \right] \\ &\leq -(1 - \kappa^2 e^{-2\nu T}) \| x^0 \|^2. \end{aligned}$$

Choosing $T \equiv (\ln(2\kappa^2))/(2\nu)$, we deduce $V'(x^0; f(x^0)) \leq -\|x^0\|^2/2$. Hence (b) holds with $c_3 \equiv 1/2$. To prove (a), note that

$$V(z) \leq \int_0^T \kappa^2 e^{-2\nu\tau} \| z \|^2 d\tau \leq \frac{\kappa^2}{2\nu} \left(1 - e^{-2\nu T}\right) \| z \|^2.$$

Moreover, letting $L > 0$ be a Lipschitz constant of f in $\mathcal{N}$, and by shrinking $\mathcal{N}$ if necessary, we have $\|x(t, x^0)\| \geq e^{-Lt}\|x^0\|$ for all $(t, x^0) \in [0, \infty) \times \mathcal{N}$. Consequently, we can deduce

$$V(z) \geq \frac{1 - e^{-2LT}}{2L} \| z \|^2 \quad \forall\, z \in \mathcal{N}.$$

Hence (a) holds with appropriate positive constants c_1 and c_2. To prove (c), note that

$$V'(x; z) - V'(x; z') = \lim_{\tau \downarrow 0} \frac{V(x + \tau z) - V(x + \tau z')}{\tau}.$$

Substituting the definition of the function V and taking absolute values, we deduce

$$\begin{aligned} &| V'(x^0; z) - V'(x^0; z') | \\ &\quad \leq \int_0^T \lim_{\tau \downarrow 0} \frac{\| x(s, x^0 + \tau z) - x(s, x^0 + \tau z') \| \, \| x(s, x^0 + \tau z) + x(s, x^0 + \tau z') \|}{\tau} \, ds \\ &\quad \leq c_4 \| z - z' \| \, \| x^0 \| \end{aligned}$$

for some constant $c_4 > 0$, where we have used the Lipschitz continuity of the solution map $x(t, \cdot)$ and the finiteness of the time T. □

We call a B-differentiable function V satisfying conditions (a), (b), and (c) in Theorem 5.2 a *B-differentiable Lyapunov function* for the nonsmooth ODE (4.4) at its equilibrium. An important consequence of Theorem 5.2 is the next perturbation result pertaining to the persistence of the exponential stability property. Notice that while the nominal function f is required to be B- (and thus directionally) differentiable, the perturbed function g is required to be only locally Lipschitz continuous. This observation is important as we see in the subsequent Corollary 5.5 that not requiring the perturbed function g to be directionally differentiable has its benefit.

COROLLARY 5.3. *Let f be Lipschitz continuous and directionally differentiable in a neighborhood $\mathcal{N}_0$ of an equilibrium x^e of f. Suppose that x^e is exponentially stable for the* ODE (4.4). *For every function g such that $g(x^e) = 0$, g is Lipschitz continuous in $\mathcal{N}_0$, and*

$$\lim_{x \to x^e} \frac{f(x) - g(x)}{\| x - x^e \|} = 0; \tag{5.1}$$

x^e is an exponentially stable equilibrium of the ODE*: $\dot{x} = g(x)$.*

Proof. Without loss of generality, we may take $x^e = 0$. Let V be a B-differentiable Lyapunov function for the ODE (4.4). According to part (I) of Theorem 5.2 applied to the function g, it suffices to show that a neighborhood $\mathcal{N}' \subseteq \mathcal{N}$ and a constant $c_3' > 0$ exist such that $V'(x^0; g(x^0)) \leq -c_3'\|x^0\|^2$ for all $x^0 \in \mathcal{N}'$. By properties (b) and (c) of V, we have

$$\begin{aligned} V'(x^0; g(x^0)) &= V'(x^0; f(x^0)) + [\, V'(x^0; g(x^0)) - V'(x^0; f(x^0)) \,] \\ &\leq -c_3 \| x^0 \|^2 + c_4 \| x^0 \| \, \| f(x^0) - g(x^0) \| \\ &= -c_3 \| x^0 \|^2 \left(1 - \frac{c_4}{c_3} \frac{\| f(x^0) - g(x^0) \|}{\| x^0 \|} \right). \end{aligned}$$

By (5.1), the existence of $\mathcal{N}'$ and c_3' with the desired property is clear. □

Remark 5.1. The limit condition (5.1) postulates that f and g are "first-order approximations" of each other near x^e. This condition, along with the directional differentiability of f at x^e, implies that the perturbed function g is directionally differentiable at x^e also, but not necessarily at other points.

We present a consequence of Corollary 5.3 that pertains to the ODE where the right-hand side is a "composite nonsmooth" function of a particular kind. Specifically, let $f(x) \equiv \Phi(x, u(x))$, where $\Phi(x, y)$ is a B-differentiable function of two arguments $(x, y) \in \Re^{n+m}$ and $u(x)$ is a B-differentiable function of x. We first state a lemma pertaining to the B-differentiability of such a function f.

LEMMA 5.4. *Let $\Phi : \Re^{n+m} \to \Re^{\ell}$ be Lipschitz continuous in a neighborhood of $(x^0, y^0) \in \Re^{n+m}$. Suppose that $\Phi(\cdot, y^0)$ and $\Phi(x^0, \cdot)$ are directionally (and thus B-) differentiable at x^0 and y^0, respectively. If*

$$\lim_{(x^0,y^0)\neq(x,y)\to(x^0,y^0)} \frac{\Phi(x,y) - \Phi(x^0,y) - (\,\Phi(\cdot,y^0)\,)'(x^0; x-x^0)}{\|\,x-x^0\,\|} = 0, \tag{5.2}$$

then Φ is directionally (and thus B-) differentiable at (x^0, y^0) and

$$\Phi'((x^0,y^0);(dx,dy)) = (\,\Phi(\cdot,y^0)\,)'(x^0;dx) + (\,\Phi(x^0,\cdot)\,)'(y^0;dy). \tag{5.3}$$

Thus, if $u : \Re^n \to \Re^m$ is B-differentiable at x^0, then so is $f(x) \equiv \Phi(x, u(x))$ and

$$f'(x^0;z) = (\,\Phi(\cdot,u(x^0))\,)'(x^0;z) + (\,\Phi(x^0,\cdot))'(u(x^0);u'(x^0;z)).$$

Proof. The B-differentiability of Φ at (x^0, y^0) and the directional derivative formula (5.3) follow from [45]; see also [15, Exercise 3.7.4]. The B-differentiability of the composite function f and the formula for its directional derivative $f'(x^0; z)$ follow from the chain rule of B-differentiation; see [15, Proposition 3.1.6]. □

Remark 5.2. The limit (5.2) is essential for (5.3) to hold; without the former, the latter need not hold. See [15].

The next result formally establishes the above-mentioned consequence of Corollary 5.3.

COROLLARY 5.5. *Let $u : \Re^n \to \Re^m$ be B-differentiable at $x^e \in \Re^n$ and let $\Phi : \Re^{n+m} \to \Re^n$ be Lipschitz continuous in a neighborhood of $(x^e, u^e) \in \Re^{n+m}$, where $u^e \equiv u(x^e)$ and (x^e, u^e) satisfies $\Phi(x^e, u^e) = 0$. Suppose that $\Phi(\cdot, u^e)$ and $\Phi(x^e, \cdot)$ are directionally differentiable at x^e and u^e, respectively, and that*

$$\lim_{(x^e,u^e)\neq(x,u)\to(x^e,u^e)} \frac{\Phi(x,u) - \Phi(x^e,u) - (\,\Phi(\cdot,u^e)\,)'(x^e; x-x^e)}{\|\,x-x^e\,\|} = 0.$$

If the equilibrium x^e is exponentially stable for the ODE (4.4)*, where $f(x) \equiv \Phi(x, u(x))$, then $z^e = 0$ is an exponentially stable equilibrium of the homogeneous* ODE $\dot{z} = (\Phi(\cdot, u^e))'(x^e; z) + (\Phi(x^e, \cdot))'(u^e; u'(x^e; z))$*. The converse is valid if additionally the right-hand side of the latter* ODE *is directionally differentiable in z.*

Proof. We have

$$\begin{aligned} f(x) = f'(x^e; x-x^e) + e(x) = (\Phi(\cdot,u^e))'(x^e;x-x^e) \\ +(\Phi(x^e,\cdot))'(u^e;u'(x^e;x-x^e)) + e(x), \end{aligned}$$

where $\lim_{x\to x^e} e(x)/\|x-x^e\| = 0$. Since x^e is locally exponentially stable for the ODE (4.4) if and only if $\widetilde{x}^e \equiv 0$ is locally exponentially stable for the ODE $\dot{\widetilde{x}} = f(\widetilde{x} + x^e)$,

and since $f(x^e) = \Phi(x^e, u^e) = 0$, we have

$$\lim_{\widetilde{x}\to 0} [\, f(\widetilde{x} + x^e) - (\, \Phi(\cdot, u^e)\,)'(x^e; \widetilde{x}) + (\, \Phi(x^e, \cdot))'(u^e; u'(x^e; \widetilde{x}))\,]/\| \widetilde{x}\, \| = 0,$$

and since the function $z \mapsto (\Phi(\cdot, u^e))'(x^e; z) + (\Phi(x^e, \cdot))'(u^e; u'(x^e; z))$ is positively homogeneous and Lipschitz continuous, the first assertion of the corollary follows from Corollary 5.3. So does the second. □

We further specialize Corollary 5.5 to the case where Φ is F-differentiable and u is piecewise smooth. Specifically, we say that a function $\Psi : \Re^n \to \Re^m$ is PC^1 (piecewise C^1) near a point $x^0 \in \Re^n$ if there exist a neighborhood $\mathcal{N}$ of x^0 and finitely many C^1 functions $\{f^1, \ldots, f^k\}$ near x^0 for some positive integer k such that $\Psi(x) \in \{f^1(x), \ldots, f^k(x)\}$ for all $x \in \mathcal{N}$. Basic properties of the family of PC^1 functions can be found in [49, 15]. In particular, it is known that a PC^1 function must be B-differentiable. Based on this remark, the result below does not require further proof.

Corollary 5.6. *Let $u : \Re^n \to \Re^m$ be PC^1 near $x^e \in \Re^n$ and let $\Phi : \Re^{n+m} \to \Re^n$ be F-differentiable in a neighborhood of $(x^e, u^e) \in \Re^{n+m}$, where $u^e \equiv u(x^e)$ and (x^e, u^e) satisfies $\Phi(x^e, u^e) = 0$. Let $f(x) \equiv \Phi(x, u(x))$. The following statements are equivalent.*

(a) *x^e is an exponentially stable equilibrium of the* ODE (4.4).

(b) *The* ODE (4.4) *has a B-differentiable Lyapunov function at x^e.*

(c) *$z^e = 0$ is an exponentially stable equilibrium of the* ODE

$$\dot{z} = J_x\Phi(x^e, u^e)z + J_y\Phi(x^e, u^e)u'(x^e; z). \tag{5.4}$$

(d) *The* ODE (5.4) *has a B-differentiable Lyapunov function at the origin.*

It is interesting to compare Corollary 5.6 with Proposition 4.5. The corollary pertains only to exponential stability, whereas the proposition deals with asymptotic stability as well. The difference between the two results is that the former proposition concerns a piecewise linear ODE, whereas the corollary concerns an ODE with a PC^1 right-hand side, for which there is no such exact approximation result as Lemma 4.4.

5.1. Strongly regular DVIs. We wish to apply Corollary 5.6 to the following differential variational inequality (DVI) [38]:

$$\begin{aligned} &\dot{x} = F(x, u), \\ &u \in \mathrm{SOL}(K, H(x, \cdot)), \end{aligned} \tag{5.5}$$

where K is a closed convex set in $\Re^m$ and $F : \Re^{n+m} \to \Re^n$ and $H : \Re^{n+m} \to \Re^m$ are continuously differentiable functions in a neighborhood of a given pair $(x^e, u^e) \in \Re^{n+m}$ that satisfies $F(x^e, u^e) = 0$ and $u^e \in \mathrm{SOL}(K, H(x^e, \cdot))$, with the latter notation meaning that u^e is a solution of the variational inequality (VI) defined by the pair $(K, H(x^e, \cdot))$; i.e., $u^e \in K$ and

$$(\, u - u^e\,)^T H(x^e, u^e) \;\geq\; 0 \quad \forall\, u \in K.$$

The key assumption to be made here is that u^e is a "strongly regular" solution of the VI $(K, H(x^e, \cdot))$. The latter is a well-known property in the theory of finite-dimensional VIs/CPs; it was introduced by Robinson [43]; see also [15]. We have employed this property in several recent studies of the DVI [39, 37] and will use it here as the main setting to facilitate the application of the previous results in the

stability analysis of the given equilibrium pair (x^e, u^e). Notice that we avoid assuming the strong monotonicity of the function $H(x, \cdot)$, which is unnecessarily restrictive in general; see nevertheless the discussion about the functional evolutionary variational inequality (5.8).

Under the strong regularity assumption, it follows that there exist neighborhoods $\mathcal{U}$ of u^e and $\mathcal{V}$ of x^e, and a Lipschitz continuous function $u : \mathcal{V} \to \mathcal{U}$ such that for every $x \in \mathcal{V}$, $u(x)$ is the unique solution of the VI $(K, H(x, \cdot))$ belonging to $\mathcal{U}$ and $u(x^e) = u^e$. Without further restricting the set K, the VI solution map $u(x)$ is not necessarily directionally differentiable; nevertheless, for a large class of closed convex sets K (such as a polyhedron), $u(x)$ is a PC^1 [36] (or a "semismooth" [40]) function near x^e. For these special sets, the DVI (5.5) is therefore, locally near the pair (x^e, u^e), equivalent to an ODE with a composite nonsmooth right-hand side, $\dot{x} = F(x, u(x))$, to which Corollary 5.5 is applicable. Before detailing this application, we make an important remark regarding this approach. Namely, corresponding to a given $x^e \in \Re^n$, there may be multiple vectors u^e satisfying the above-mentioned properties, each of which leads to a particular ODE that could be quite distinct from another. More interestingly, x^e may be exponentially stable with respect to one resulting ODE but not to another. (This is illustrated in Example 5.1.) In other words, the "stability" of x^e is linked to the particular solution of the VI $(K, H(x^e, \cdot))$. For future research, it may be of interest to extend this individual ODE-based stability theory for the nonlinear DVI (5.5) to a broader theory analogous to that for the LCS (2.1) or its affine generalization, the DAVI (4.6), where we have relied on the key assumption that $B\mathrm{SOL}(K, q + Cx, D)$ is a singleton for all $x \in \Re^n$. In such affine cases, in spite of the possible multiplicity of solutions to the AVIs $(K, q+Cx, D)$, the singleton assumption, or equivalently, the assumption of a unique C^1 trajectory $x(t, x^0)$, leads to a unique ODE with a piecewise linear (thus Lipschitz) right-hand side to which Definition 2.3 applies. Incidentally, there are multiple C^1 x-trajectories in the example below.

Example 5.1. Consider the following nonlinear complementarity system (NCS):

$$\begin{aligned} \dot{x} &= x(-1 + 2\sin u), \\ 0 \le u &\perp (x+1)(1 - \sin u) \ge 0, \end{aligned} \tag{5.6}$$

where $x \in \Re$ and $u \in \Re$. It is clear that $x^e = 0$ is an equilibrium. For any $x > -1$, the solutions of the associated nonlinear complementarity problem (NCP) $0 \le u \perp (x+1)(1 - \sin u) \ge 0$ are $u = 0$ and $u = (2k + 1/2)\pi$ for $k \ge 0$. Each of these solutions is strongly regular. The unique solution trajectory to (5.6) that is near the pair $(x^e, u^e) = (0, 0)$ initially is $(x(t, x^0), u(t, x^0)) = (x^0 e^{-t}, 0)$ for all $t \ge 0$. The equilibrium $x^e = 0$ is clearly exponentially stable for the resulting ODE, which is $\dot{x} = -x$. In contrast, the unique solution trajectory to (5.6) that is near the pair $(x^e, \widehat{u}^e) = (0, \pi/2)$ initially is $(x(t, x^0), u(t, x^0)) = (x^0 e^t, \pi/2)$ for all $t \ge 0$. The same equilibrium $x^e = 0$ is clearly *unstable* for the resulting ODE, which is $\dot{x} = x$.

Returning to the general discussion, we fix an implicit VI solution function $u(x)$ as defined above. For simplicity, we focus on the case where K is a polyhedron. It follows that $u(x)$ is a PC^1 function of $x \in \mathcal{V}$. Let $\mathcal{C}(x^e) \equiv \mathcal{T}(K; u(x^e)) \cap H(x^e, u^e)^\perp$ be the critical cone of the linearly constrained VI $(K, H(x^e, \cdot))$. It is known that the directional derivative $u'(x^e; z)$ of the VI solution map $u(x)$ along the direction z is the unique solution v of the generalized LCP:

$$\mathcal{C}(x^e) \ni v \perp J_x H(x^e, u^e) z + J_u H(x^e, u^e) v \in \mathcal{C}(x^e)^*.$$

Based on this characterization of the directional derivative, define the following gen-

eralized LCS that extends (4.7) to the nonlinear case:

$$\begin{aligned} \dot{z} &= J_x F(x^e, u^e) z + J_u F(x^e, u^e) v, \\ \mathcal{C}(x^e) \ni v \perp J_x H(x^e, u^e) z + J_u H(x^e, u^e) v &\in \mathcal{C}(x^e)^*. \end{aligned} \tag{5.7}$$

PROPOSITION 5.7. *Let K be polyhedral and let F and H be C^1 in a neighborhood of the pair (x^e, u^e), where $F(x^e, u^e) = 0$ and u^e is a strongly regular solution of the VI $(K, H(x^e, \cdot))$. Let $\mathcal{V} \times \mathcal{U}$ and $u : \mathcal{V} \to \mathcal{U}$ be, respectively, the neighborhood of (x^e, u^e) and the solution map associated with the strong regularity of u^e. The two statements below are equivalent.*

(a) *x^e is an exponentially stable equilibrium of the ODE: $\dot{x} = F(x, u(x))$.*

(b) *$z^e = 0$ is an exponentially stable equilibrium of the generalized LCS* (5.7).

Proof. This follows readily from Corollary 5.6. □

We illustrate Proposition 5.7 with a *functional evolutionary variational inequality* (FEVI) of the following kind:

$$\dot{x} = \Pi_K(x - \Phi(x)) - x, \tag{5.8}$$

where Π_K denotes the Euclidean projection onto the polyhedron K and Φ is a C^1 function defined on $\Re^n$. The equilibria of this DVI are precisely the solutions of the finite-dimensional VI (K, Φ). Incidentally, there are other dynamical systems whose equilibria are solutions of the VI. The above FEVI is different from the kind of evolutionary variational problems studied in the literature of differential inclusions which rely on a "generalized equation" formulation of the VI; see, e.g., [16]. A distinct advantage of the FEVI (5.8) over the latter kind is that the solution trajectories of (5.8) are all C^1 because the right-hand side is a Lipschitz function of x, whereas those based on the differential inclusion formulation need not be so. In addition, when started at a vector in K, a trajectory of (5.8) will remain in K. The last assertion is established in the result below.

PROPOSITION 5.8. *Let K be a closed convex set and Φ be Lipschitz continuous on K. Let $x(t, x^0)$ denote the unique solution trajectory of* (5.8) *initiated at $x(0) = x^0$. If $x^0 \in K$, then $x(t, x^0) \in K$ for all $t \geq 0$.*

Proof. Considering (5.8) as an ODE with an inhomogeneous right-hand side, we have

$$\begin{aligned} x(t, x^0) &= e^{-t} x^0 + \int_0^t e^{-(t-\tau)} \, \Pi_K\big(x(\tau, x^0) - \Phi(x(\tau, x^0))\big) d\tau \\ &= e^{-t} \, x^0 + (\, 1 - e^{-t} \,) \, \frac{\int_0^t e^{\tau} \, \Pi_K\big(x(\tau, x^0) - \Phi(x(\tau, x^0))\big) \, d\tau}{e^t - 1} \\ &= e^{-t} \, x^0 + (\, 1 - e^{-t} \,) \, \frac{\int_0^t e^{\tau} \, \Pi_K\big(x(\tau, x^0) - \Phi(x(\tau, x^0))\big) \, d\tau}{\int_0^t e^{\tau} \, d\tau}. \end{aligned} \tag{5.9}$$

Since $x(t, x^0)$ is well defined for all t, $\Pi_K\big(x(\tau, x^0) - \Phi(x(\tau, x^0))\big)$ is continuous in τ. Hence, we can represent the integrals in (5.9) by Riemann sums:

$$\begin{aligned} I &\equiv \frac{\int_0^t e^{\tau} \, \Pi_K\big(x(\tau, x^0) - \Phi(x(\tau, x^0))\big) \, d\tau}{\int_0^t e^{\tau} \, d\tau} \\ &= \lim_{k \to \infty} \frac{\sum_{i=1}^k e^{s_i} \, \Pi_K\big(x(s_i, x^0) - \Phi(x(s_i, x^0))\big) \, \frac{t}{k}}{\sum_{i=1}^k e^{s_i} \, \frac{t}{k}}, \end{aligned}$$

where s_i is any point in the subinterval $[\frac{(i-1)}{k}\,t, \frac{i}{k}\,t]$. Since K is a convex set, the vector

$$\left(\sum_{i=1}^{k} e^{s_i}\,\frac{t}{k}\right)^{-1}\left[\sum_{i=1}^{k} e^{s_i}\,\Pi_K\big(x(s_i,x^0)-\Phi(x(s_i,x^0))\big)\,\frac{t}{k}\right],$$

which is a convex combination of vectors in K, belongs to K for all positive integers k. By the closedness of K, it follows that the vector I, and thus $x(t,x^0)$, also belongs to K. □

The system (5.8) is a special DVI with $F(x,u)\equiv u-x$ and $H(x,u)\equiv u-x+\Phi(x)$. Since $H(x,\cdot)$ is strongly monotone, the strong regularity condition holds trivially. Moreover, by the chain rule of B-differentiation, we have, letting $u(x)\equiv\Pi_K(x-\Phi(x))$,

$$u'(x;z)=\Pi_{\mathcal{C}}(z-J\Phi(x)z),$$

where $\mathcal{C}\equiv\mathcal{T}(K;u(x))\cap(u(x)-x+\Phi(x))^{\perp}$ is the critical cone of the polyhedron K at the projected vector $u(x)$; see [15, Chapter 4]. Consequently, the first-order LCS (5.7), which becomes $\dot z=\Pi_{\mathcal{C}}(z-J\Phi(x)z)-z$, is a functional evolutionary version of the finite-dimensional generalized LCP $\mathcal{C}\ni z\perp J\Phi(x)z\in\mathcal{C}^*$ of the same kind as (5.8). Notice that if $x\in\mathrm{SOL}(K,\Phi)$ so that $u(x)=x$, then $\mathcal{C}=\mathcal{T}(K,x)\cap\Phi(x)^{\perp}$ coincides with the critical cone of the VI (K,Φ) at the solution x.

Summarizing the above discussion and invoking Proposition 5.7, we obtain the following result for the FEVI (5.8).

COROLLARY 5.9. *Let K be a polyhedron and let $\Phi:\Re^n\to\Re^n$ be C^1. A solution $x^e\in\mathrm{SOL}(K,\Phi)$ is exponentially stable for the* FEVI (5.8) *if and only if the origin is an exponentially stable equilibrium for the linearized* FEVI*: $\dot z=\Pi_{\mathcal{C}}(z-J\Phi(x^e)z)-z$, where $\mathcal{C}\equiv\mathcal{T}(K;x^e)\cap\Phi(x^e)^{\perp}$.* □

Next, we specialize Proposition 5.7 to the NCS

$$\begin{aligned}\dot x(t)&=F(x(t),u(t)),\\ 0\;\le\;u(t)\;&\perp\;H(x(t),u(t))\;\ge\;0,\end{aligned}\tag{5.10}$$

which is a special case of (5.5) with $K=\Re^m_+$. Let (x^e,u^e) be as specified above. The strong regularity of u^e can be characterized by introducing the three fundamental index sets $(\alpha_e,\beta_e,\gamma_e)$ associated with the pair (x^e,u^e) (cf. the LCS with a P-matrix in section 3):

$$\begin{aligned}\alpha_e&=\{\,i\,:\,u^e_i>0=H_i(x^e,u^e)\,\},\\ \beta_e&=\{\,i\,:\,u^e_i=0=H_i(x^e,u^e)\,\},\\ \gamma_e&=\{\,i\,:\,u^e_i=0<H_i(x^e,u^e)\,\}.\end{aligned}$$

According to these sets, we can partition the (partial) Jacobian matrix $J_uH(x^e,u^e)$ as follows:

$$J_uH(x^e,u^e)\;\equiv\;\begin{bmatrix} J_{u_{\alpha_e}}H_{\alpha_e}(x^e,u^e) & J_{u_{\beta_e}}H_{\alpha_e}(x^e,u^e) & J_{u_{\gamma_e}}H_{\alpha_e}(x^e,u^e)\\ J_{u_{\alpha_e}}H_{\beta_e}(x^e,u^e) & J_{u_{\beta_e}}H_{\beta_e}(x^e,u^e) & J_{u_{\gamma_e}}H_{\beta_e}(x^e,u^e)\\ J_{u_{\alpha_e}}H_{\gamma_e}(x^e,u^e) & J_{u_{\beta_e}}H_{\gamma_e}(x^e,u^e) & J_{u_{\gamma_e}}H_{\gamma_e}(x^e,u^e)\end{bmatrix},$$

where $J_{u_\alpha}H_\beta$ denotes the matrix of partial derivatives $[\partial H_j/\partial u_i]_{(i,j)\in\alpha\times\beta}$. It is known that u^e is a strongly regular solution of the NCP

$$0\;\le\;u\;\perp\;H(x^e,u)\;\ge\;0\tag{5.11}$$

if and only if (a) the principal submatrix $J_{u_{\alpha_e}} H_{\alpha_e}(x^e, u^e)$ is nonsingular, and (b) the Schur complement

(5.12)
$$\widehat{D}_{\beta_e\beta_e} \equiv J_{u_{\beta_e}} H_{\beta_e}(x^e, u^e) - J_{u_{\alpha_e}} H_{\beta_e}(x^e, u^e) \left[J_{u_{\alpha_e}} H_{\alpha_e}(x^e, u^e) \right]^{-1} J_{u_{\beta_e}} H_{\alpha_e}(x^e, u^e)$$

is a P-matrix. Moreover, for every $z \in \Re^n$, the directional derivative $u'(x^e; z)$ is the unique vector $\widehat{u}$ satisfying $\widehat{u}_{\gamma_e} = 0$ and

$$\begin{aligned} J_x H_{\alpha_e}(x^e, u^e) z + J_{\alpha_e} H_{\alpha_e}(x^e, u^e) \widehat{u}_{\alpha_e} + J_{\alpha_e} H_{\beta_e}(x^e, u^e) \widehat{u}_{\beta_e} = 0, \\ 0 \le \widehat{u}_{\beta_e} \perp J_x H_{\beta_e}(x^e, u^e) z + J_{\beta_e} H_{\alpha_e}(x^e, u^e) \widehat{u}_{\alpha_e} + J_{\beta_e} H_{\beta_e}(x^e, u^e) \widehat{u}_{\beta_e} \ge 0, \end{aligned}$$

which, by the nonsingularity of $J_{\alpha_e} H_{\alpha_e}(x^e, u^e)$, is equivalent to the standard LCP

$$0 \le \widehat{u}_{\beta_e} \perp \widehat{C}_{\beta_e \bullet} z + \widehat{D}_{\beta_e\beta_e} \widehat{u}_{\beta_e} \ge 0,$$

where $\widehat{C}_{\beta_e \bullet} \equiv J_x H_{\beta_e}(x^e, u^e) - J_{u_{\alpha_e}} H_{\beta_e}(x^e, u^e) [J_{u_{\alpha_e}} H_{\alpha_e}(x^e, u^e)]^{-1} J_x H_{\alpha_e}(x^e, u^e)$. Define the matrices

$$\begin{aligned} \widehat{A} &\equiv J_x F(x^e, u^e) - J_{u_{\alpha_e}} F(x^e, u^e) [J_{u_{\alpha_e}} H_{\alpha_e}(x^e, u^e)]^{-1} J_x H_{\alpha_e}(x^e, u^e), \\ \widehat{B}_{\bullet\beta_e} &\equiv J_{u_{\beta_e}} F(x^e, u^e) - J_{u_{\alpha_e}} F(x^e, u^e) [J_{u_{\alpha_e}} H_{\alpha_e}(x^e, u^e)]^{-1} J_{u_{\beta_e}} H_{\alpha_e}(x^e, u^e); \end{aligned}$$

consider the homogeneous LCS where the algebraic variable involves only the β_e-components:

(5.13)
$$\begin{aligned} \dot{z} &= \widehat{A} z + \widehat{B}_{\bullet\beta_e} \widehat{u}_{\beta_e}, \\ 0 \le \widehat{u}_{\beta_e} &\perp \widehat{C}_{\beta_e \bullet} z + \widehat{D}_{\beta_e\beta_e} \widehat{u}_{\beta_e} \ge 0. \end{aligned}$$

The results in section 3 can surely be applied to (5.13) to yield sufficient conditions for statement (b) of the following proposition to hold, whose proof follows readily from Proposition 5.7.

PROPOSITION 5.10. *Let F and H be C^1 in a neighborhood of the pair (x^e, u^e), where $F(x^e, u^e) = 0$ and u^e is a strongly regular solution of the* NCP (5.11). *Let $\mathcal{V} \times \mathcal{U}$ and $u : \mathcal{V} \to \mathcal{U}$ be, respectively, the neighborhood of (x^e, u^e) and the solution map associated with the strong regularity of u^e. The following two statements are equivalent.*

(a) *x^e is an exponentially stable equilibrium of the* ODE $\dot{x} = F(x, u(x))$.

(b) *$z^e = 0$ is an exponentially stable equilibrium of the homogeneous* LCS (5.13).

6. Concluding remarks. Based on the combined tools of contemporary finite dimensional LCPs and VIs/CPs and classical Lyapunov stability theory for smooth dynamical systems, we have obtained many stability results for the LCS and its nonlinear generalizations. Part of the novelty of our analysis is the employment of a nontraditional Lyapunov function in both the system state and the auxiliary algebraic variable, which leads to a nondifferentiable Lyapunov function of the state alone. We speculate that this approach might be useful in other contexts, such as in the convergence analysis of iterative algorithms for solving finite-dimensional variational and optimization problems.

The results in this paper have left open some questions that are worthy of further investigation. Foremost among these is the question of whether asymptotic stability

would imply exponential stability for an LCS satisfying the P-property. In this vein, we recall [52, Lemma 8.2] which establishes such an implication for a linear selectionable DI. Yet, as we have noted a few times, the DI result is not applicable to the LCS. Nevertheless, the same implication may be valid for the LCS. Another interesting question is the persistence of asymptotic stability of a B-differentiable differential system under small perturbations; related to the latter question is whether there are analogues of the results in subsection 5.1 for asymptotic stability. Finally, the authors in [16] have established a very interesting necessary degree-theoretic condition for the asymptotic stability of an evolutionary variational inequality. We feel that a further degree-theoretic exploration of the LCS and the DVI is warranted.

Acknowledgment. We are grateful to two referees who have offered many constructive comments that have significantly improved the presentation of the paper.

REFERENCES

[1] S. Adly and D. Goeleven, *A stability theory for second-order nonsmooth dynamical systems with application to friction problems*, J. Math. Pures Appl. (9), 83 (2004), pp. 17–51.

[2] M. Branicky, *Multiple Lyapunov functions and other analysis tools for switched and hybrid systems*, IEEE Trans. Automat. Control, 43 (1998), pp. 475–482.

[3] B. Brogliato, *Some perspectives on analysis and control of complementarity systems*, IEEE Trans. Automat. Control, 48 (2003), pp. 918–935.

[4] M. K. Çamlibel, *Complementarity Methods in the Analysis of Piecewise Linear Dynamical Systems*, Ph.D. thesis, Center for Economic Research, Tilburg University, Tilburg, The Netherlands, 2001.

[5] M. K. Çamlibel, W. P. M. H. Heemels, A. J. van der Schaft, and J. M. Schumacher, *Switched networks and complementarity*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 50 (2003), pp. 1036–1046.

[6] M. K. Çamlibel, W. P. M. H. Heemels, and J. M. Schumacher, *Well-posedness of a class of linear network with ideal diodes*, in Proceedings of the 14th International Symposium of Mathematical Theory of Networks and Systems, Perpignan, France, 2000.

[7] M. K. Çamlibel, W. P. M. H. Heemels, and J. M. Schumacher, *On linear passive complementarity systems*, European J. Control, 8 (2002), pp. 220–237.

[8] M. K. Çamlibel, W. P. M. H. Heemels, and J. M. Schumacher, *Stability and controllability of planar bimodal linear complementarity systems*, in Proceedings of the 42nd IEEE Conference on Decision and Control, Maui, HI, 2003, pp. 1651–1656.

[9] M. K. Çamlibel, W.P.M.H. Heemels, and J. M. Schumacher, *On the controllability of bimodal piecewise linear systems*, in Hybrid Systems: Computation and Control, R. Alur and G. J. Pappas, eds., Lecture Notes in Comput. Sci. 2993, Springer-Verlag, Berlin, 2004, pp. 250–264.

[10] M. K. Çamlibel, J. S. Pang, and J. L. Shen, *Conewise Linear Systems: NonZenoness and Observability*, Preprint, Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY, 2005.

[11] M. K. Çamlibel and J. M. Schumacher, *On the Zeno behavior of linear complementarity systems*, in Proceedings of the 40th IEEE Conference on Decision and Control, Orlando, FL, 2001, pp. 346–351.

[12] M. K. Çamlibel and J. M. Schumacher, *Copositive Lyapunov functions*, Open Problems in Mathematical Systems and Control Theory, in V. D. Blondel and A. Megretski, eds., Princeton University Press, Princeton, NJ, 2004, pp. 189–193.

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

[14] R. A. Decarlo, M. S. Branicky, S. Pettersson, and B. Lennartson, *Perspectives and results on the stability and stabilization of hybrid systems*, Proceedings of the IEEE, 88 (2000), pp. 1069–1082.

[15] F. Facchinei and J.-S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Springer-Verlag, New York, 2003.

[16] D. Goeleven and B. Brogliato, *Stability and instability matrices for linear evolution variational inequalities*, IEEE Trans. Automat. Control, 49 (2004), pp. 521–534.

[17] D. Goeleven and B. Brogliato, *Necessary conditions of asymptotic stability for unilateral dynamical systems*, Nonlinear Anal., 61 (2005), pp. 961–1004.
[18] D. Goeleven, M. Motreanu, and V. Motreanu, *On the stability of stationary solutions of evolution variational inequalities*, Adv. Nonlinear Var. Inequal., 6 (2003), pp. 1–30.
[19] W. P. H. Heemels, *Linear Complementarity Systems: A Study in Hybrid Dynamics*, Ph.D. thesis, Department of Electrical Engineering, Eindhoven University of Technology, Eindhaven, The Netherlands, 1999.
[20] W. P. M. H. Heemels, J. M. Schumacher, and S. Weiland, *Well-posedness of linear complementarity systems*, in Proceedings of the 38th IEEE Conference on Decision and Control, Phoenix, AZ, 1999, pp. 3037–3042.
[21] W. P. M. H. Heemels, J. M. Schumacher, and S. Weiland, *The rational linear complementarity systems*, Linear Algebra Appl., 294 (1999), pp. 93–135.
[22] W. P. M. H. Heemels, J. M. Schumacher, and S. Weiland, *Linear complementarity systems*, SIAM J. Appl. Math., 60 (2000), pp. 1234–1269.
[23] J. P. Hespanha, *Uniform stability of switched linear systems: Extension of LaSalle's invariance principle*, IEEE Trans. Automat. Control, 49 (2004), pp. 470–482.
[24] J. P. Hespanha, D. Liberzon, D. Angeli, and E. D. Sontag, *Nonlinear norm-observability notions and stability of switched systems*, IEEE Trans. Automat. Control, 50 (2005), pp. 154–168.
[25] D. Hipfel, *The Nonlinear Differential Complementarity Problem*, Ph.D. thesis, Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY, 1993.
[26] M. Johansson, *Piecewise linear control systems*, Lecture Notes in Control and Inform. Sci. 284, Springer-Verlag, Berlin, 2003.
[27] M. Johansson and A. Rantzer, *Computation of piecewise quadratic Lyapunov functions for hybrid systems*, IEEE Trans. Automat. Control, 43 (1998), pp. 555–559.
[28] H. Khalil, *Nonlinear Systems*, 3rd ed., Prentice–Hall, Englewood Cliffs, NJ, 2002.
[29] H. Lin and P. J. Antsaklis, *Stability and stabilizability of switched linear systems: A short survey of recent results*, in Proceedings of the 2005 IEEE International Symposium on Intelligent Control, Limassol, Cyprus, 2005, pp. 24–29.
[30] H. Lin and P. J. Antsaklis, *A converse Lyapunov theorem for switched systems*, in Proceedings of the 44th IEEE Conference on Decision and Control, Seville, Spain, 2005, pp. 3291–3296.
[31] N. G. Lloyd, *Degree Theory*, Cambridge University Press, Cambridge, UK, 1978.
[32] J. Mawhin, *Continuation theorems and periodic solutions of ordinary differential equations*, in Topological Methods in Differential Equations and Inclusions, A. Granas, M. Frigon, and G. Sabidussi, eds., NATO Sci. Ser. C, Math. Phys. Sci. 472, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1995, pp. 291–375.
[33] A. N. Michel, *Recent trends in stability of hybrid dynamical systems*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 46 (1999), pp. 120–134.
[34] A. N. Michel and B. Hu, *Towards a stability theory of general hybrid dynamical systems*, Automatica J. IFAC, 35 (1999), pp. 371–384.
[35] J.-S. Pang, *Newton's method for B-differentiable equations*, Math. Oper. Res., 15 (1990), pp. 311–341.
[36] J.-S. Pang and D. Ralph, *Piecewise smoothness, local invertibility, and parametric analysis of normal maps*, Math. Oper. Res., 21 (1996), pp. 401–426.
[37] J.-S. Pang and J. L. Shen, *Strongly regular differential variational systems*, IEEE Trans. Automat. Control, to appear, 2006.
[38] J.-S. Pang and D. E. Stewart, *Differential variational inequalities*, Math. Program. Ser. A, second revision under review (as of April 2006). Available at http://www.rpi.edu/~pangj/PStewart03.pdf.
[39] J.-S. Pang and D. E. Stewart, *Solution dependence on initial conditions in differential variational inequalities*, Math. Program. Ser. B, to appear. Available at http://www.rpi.edu/~pangj/ODEsol_final.pdf.
[40] J.-S. Pang, D. Sun, and J. Sun, *Semismooth homeomorphisms and strong stability of semidefinite and Lorentz complementarity problems*, Math. Oper. Res., 28 (2003), pp. 39–63.
[41] P. A. Parillo, *Structured Semidefinite Programs and Semialgebraic Geometry Methods in Robustness and Optimization*, Ph.D. dissertation, California Institute of Technology, Pasadena, CA, 2000.
[42] S. Pettersson and B. Lennartson, *An LMI approach for stability analysis of nonlinear systems*, in Proceedings of the 4th European Control Conference, Brussels, Belgium, 1997.
[43] S. M. Robinson, *Strongly regular generalized equations*, Math. Oper. Res., 5 (1980), pp. 43–62.

[44] S. M. Robinson, *Some continuity properties of polyhedral multifunctions*, Math. Programming Stud., 14 (1981), pp. 206–214.
[45] S. M. Robinson, *An implicit-function theorem for a class of nonsmooth functions*, Math. Oper. Res., 16 (1991), pp. 292–309.
[46] S. M. Robinson, *Normal maps induced by linear transformations*, Math. Oper. Res., 17 (1992), pp. 691–714.
[47] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[48] R. T. Rockafellar and R.J.-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
[49] S. Scholtes, *Introduction to piecewise differentiable equations*, Habilitation thesis, Institut für Statistik und Mathematische Wirtschaftstheorie, Universität Karlsruhe, Karlsruhe, Germany, 1994.
[50] J. M. Schumacher, *Complementarity systems in optimization*, Math. Program. Ser. B, 101 (2004), pp. 263–296.
[51] J. Shen and J.-S. Pang, *Linear complementarity systems: Zeno states*, SIAM J. Control Optim., 44 (2005), pp. 1040–1066.
[52] G. V. Smirnov, *Introduction to the Theory of Differential Inclusions*, Grad. Stud. Math. 41, AMS, Providence, RI, 2002.
[53] A. J. van der Schaft and J. M. Schumacher, *Complementarity modeling of hybrid systems*, IEEE Trans. Automat. Control, 43 (1998), pp. 483–490.
[54] A. J. van der Schaft and J. M. Schumacher, *An Introduction to Hybrid Dynamical Systems*, Springer-Verlag, London, 2000.
[55] J. F. Sturm and S. Zhang, *On cones of nonnegative quadratic functions*, Math. Oper. Res., 28 (2003), pp. 246–267.
[56] H. Ye, A. N. Michel, and L. Hou, *Stability theory of hybrid dynamical systems*, IEEE Trans. Automat. Control, 43 (1998), pp. 461–474.

SIAM J. OPTIM.
Vol. 17, No. 4, pp. 1102–1128
© 2006 Society for Industrial and Applied Mathematics

SPASELOC: AN ADAPTIVE SUBPROBLEM ALGORITHM FOR SCALABLE WIRELESS SENSOR NETWORK LOCALIZATION*

MICHAEL W. CARTER†, HOLLY H. JIN‡, MICHAEL A. SAUNDERS§, AND YINYU YE§

Abstract. An adaptive rule-based algorithm, SpaseLoc, is described to solve localization problems for ad hoc wireless sensor networks. A large problem is solved as a sequence of very small subproblems, each of which is solved by semidefinite programming relaxation of a geometric optimization model. The subproblems are generated according to a set of sensor/anchor selection rules. Computational results compared with existing approaches show that the SpaseLoc algorithm scales well and provides excellent localization accuracy.

Key words. sensor localization, semidefinite programming, large-scale optimization

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

DOI. 10.1137/040621600

1. Introduction.

Ad hoc wireless sensor networks may contain hundreds or even tens of thousands of inexpensive devices (sensors) that can communicate with their neighbors within a limited radio range. By relaying information to each other, they can transmit signals to a command post anywhere within the network. They have many practical uses in areas such as military applications [15], environment or industrial control and monitoring [7, 9], wildlife monitoring [24], and security monitoring [15]. For example, Southern California Edison's Nuclear Generating Station in San Onofre, CA, has deployed wireless mesh networked sensors from Dust Networks, Inc. to obtain real-time trend data [9]. These data are used to predict which motors are about to fail, so they could be preemptively rebuilt or replaced during scheduled maintenance periods. The use of a wireless sensor network saves the station money and avoids potential machine shutdown. Implementation of a sensor localization algorithm would provide a service that eliminates the need to record every sensor's location and its associated ID number in the network.

Wireless sensor networks are potentially important enablers for many other advanced applications. A huge variety of applications lie ahead. By 2008, there could be 100 million wireless sensors in use, up from about 200,000 in 2005, according to the market-research company Harbor Research. The worldwide market for wireless sensors, it says, will grow from $100 million in 2005 to more than $1 billion by 2009 [18]. This is motivating great effort in academia and industry to explore effective ways to build sensor networks with feature-rich services [12].

One of the important inputs these services build upon is the exact locations of all sensors in the network. The need for *sensor localization* arises because accurate

*Received by the editors December 27, 2004; accepted for publication (in revised form) February 27, 2006; published electronically December 5, 2006.
http://www.siam.org/journals/siopt/17-4/62160.html

†Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, ON, Canada M5S 3G8 (carter@mie.utoronto.ca).

‡Department of Management Science and Engineering, Stanford University, Stanford, CA 94305-4026, and Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, ON, Canada M5S 3G8 (hollyjin@stanford.edu). The author was partially supported by Robert Bosch Corporation.

§Department of Management Science and Engineering, Stanford University, Stanford, CA 94305-4026 (saunders@stanford.edu, yinyu-ye@stanford.edu).

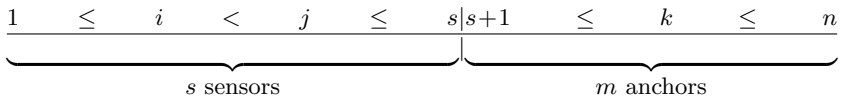


FIG. 1.1. *Indexing of sensors and anchors.*

locations are known for only some of the sensors (which are called *anchors*). If the networks are to achieve their purpose, the locations of the remaining sensors must be determined. One approach to localizing these sensors with unknown locations is to use known anchor locations and distance measurements that neighboring sensors and anchors obtain among themselves. The mathematical problem is to *estimate* all sensors' locations using a sparse data matrix of noisy distance measurements. This leads to a large, nonconvex, constrained optimization problem. Large networks may contain many thousands of sensors, whose locations should be determined accurately and quickly.

1.1. Problem definition. Sensor localization for ad hoc wireless sensor networks aims to find the locations of all sensors in the network, given pairwise distance measurements among some of the sensors and known locations of some of the sensors. The sensors with known locations are called *anchors*. From now on, *sensor* generally means *unlocalized sensor*, excluding anchors. A *node* is any sensor or anchor.

We use a constrained optimization approach to estimate the sensors' locations. The following input, output, and objectives are considered.

Input

Total points: n, the total number of nodes in the network.

Unknown points: s sensors, whose locations $x_i \in \mathcal{R}^2$, $i = 1, \ldots, s$, are to be determined. (We assume the points are on a plane here, but the approach is extended to three dimensions in Jin's thesis [14].)

Known points: m anchors, whose locations $a_k \in \mathcal{R}^2$, $k = s+1, \ldots, n$, are known. (Note that we put anchors at the end of the total points' list without loss of generality, and that $n = s + m$. Index k is specific for indexing anchors. Refer to Figure 1.1 for node indexing.)

Known distance measurements: The nonzero elements of a sparse matrix $\widehat{d}$ containing the readings of certain ranging devices for estimating the distance between two points. $\widehat{d}_{ij}$ is the distance measurement between two sensors x_i and x_j $(i < j \leq s)$, and $\widehat{d}_{ik}$ is the distance measurement between some sensor x_i and anchor a_k $(i \leq s < k)$. The distance measurements are constant data and generally have errors.

Output

Locations: Estimated locations x_i for s sensors.

Objectives

Accuracy: Minimal errors in the estimated sensor locations.

Speed: Fast enough for real-time applications (e.g., networks with moving sensors).

Scalability: Suitable for large-scale deployment (with tens of thousands of nodes).

1.2. Notation. The Euclidean distance between two vectors v and w is defined to be $\|v - w\|$, where $\|\cdot\|$ always means the 2-norm. Nodes are said to be *connected* if the associated measurements $\widehat{d}_{ij}$ or $\widehat{d}_{ik}$ exist. The remaining elements of $\widehat{d}$ are zero.

If a measurement does exist between node i and j but it is zero (i and j are at the same spot), we do not set $\widehat{d}_{ij}$ to zero: we set it to machine precision ϵ instead to distinguish from the case of $\widehat{d}_{ij} = 0$ when two nodes' distance is beyond the sensor device's measuring range.

1.3. Related research work. Sensor localization in ad hoc wireless networks has been a booming research area recently. Hightower and Boriello [12] give an extensive review of the area and available methods. There are many ways to solve the localization problem [6, 8, 10, 13, 17, 19, 20, 21, 22], with two main ones based on triangulation and optimization.

Triangulation methods estimate node locations based on distance measurements between neighboring nodes, and some algorithms use iterative steps to localize all sensors.

Early work using optimization techniques is reported by Doherty, El Ghaoui, and Pister [8]. Ideally the Euclidean distance between neighboring nodes should be fitted in some near-equality sense to the distance measurements:

$$\|x_i - x_j\| \approx \widehat{d}_{ij} \quad \text{and} \quad \|x_i - a_k\| \approx \widehat{d}_{ik}. \tag{1.1}$$

Doherty, El Ghaoui, and Pister formulate a convex optimization model by treating the constraints as $\|x_i - x_j\| \le \widehat{d}_{ij}$ and $\|x_i - a_k\| \le \widehat{d}_{ik}$, and by including certain other convex constraints. This formulation takes advantage of available optimization algorithms, including those for convex optimization. However, the method needs sufficient anchors to be on the boundary of the localization area for it to work effectively.

Biswas and Ye [2] work with the near-equality constraints (1.1), and more importantly introduced a semidefinite programming (SDP) relaxation method in order to retain the benefits of convex optimization. They report that their method yields more accuracy than the approach in [8].

The SDP relaxation approach can solve small problems effectively. The paper [2] reports a few seconds of laptop execution time for a 50-node localization problem. However, the number of constraints in the SDP model is $O(n^2)$, where n is the number of nodes in the network. Even a few-hundred-node problem leads to excessive memory and computation time by available SDP solvers such as DSDP (Benson, Ye, and Zhang [1]) and SeDuMi (Sturm [23]). These solvers are effective for SDP problems with dimension and number of constraints up to a few thousand.

Tseng [25] has presented a second-order cone programming (SOCP) relaxation model that permits solution for problem sizes up to a few thousand using available SOCP solvers. However, the additional relaxation of the original model usually generates larger error rates, and the run-times are high. The author reports CPU times of 330 seconds for 1000 nodes and 3 hours for 2000 nodes using SeDuMi 1.05 [23] and MATLAB 6.1 on a Linux PC.

Biswas and Ye [3] propose a decomposition scheme to overcome the scalability issue with SDP solvers. The anchors in the network are first partitioned into many clusters according to their physical locations, and sensors are assigned to these clusters if they have a direct connection to one of the anchors. Each cluster formulates a subproblem, and the subproblems are solved *independently* on each cluster using the SDP relaxation of [2]. The paper reports results for randomly generated sensor networks of 4000 sensors partitioned into 100 clusters strictly according to their geographic locations. Sensors with distance connections to more than one cluster are included in multiple clusters. The final estimation of their locations is determined

by the cluster that gives the least estimated errors. An execution time of about 4 minutes on a 1.2GHz Pentium laptop is reported for a problem of this size. Thus, the decomposition approach makes large-scale sensor network localization possible on a single processor. The further advantage is that multiple CPUs can be used in a natural way.

1.4. SpaseLoc. A basic tool that we have developed during this research is a rule-based iterative algorithm named SpaseLoc (**s**ub**p**roblem **a**lgorithm for **se**nsor **loc**alization). It is effective for networks involving tens of thousands of sensors and beyond, using a single CPU.

To solve a large localization problem (defined as the *full_problem*), SpaseLoc proceeds iteratively by estimating only a portion of the total sensors' locations at each iteration. Some anchors and sensors are chosen according to a set of rules. They form a sensor localization *subproblem* that can be treated similarly to the basic SDP formulation of Biswas and Ye [2]. The solution from the subproblem is fed back to the *full_problem* and the algorithm iterates again until all sensors are localized.

Computational results show that SpaseLoc can solve small or large problems with excellent accuracy and scalability. It is capable of localizing 4000 nodes with great accuracy in under 20 seconds, and 10000 nodes in under a minute on a 2.4 GHz laptop.

2. The subproblem SDP model. This section reviews the quadratic programming formulation of the sensor localization problem and the SDP relaxation model of Biswas and Ye [2] that the SpaseLoc subproblem is based on. Error analysis is also reviewed here as a reference for later sections.

2.1. Euclidean distance model. Consider a network of sensors and anchors labeled as in Figure 1.1. For any point in the network, there could be three types of distance measurements. Since we generally do not need the distance information between two anchor points, we exclude this type of measurement from now on.

The other types of distance measurements are the two we need for the localization model. First is the distance measurement between two sensors (i and j) with unknown locations; second is the distance measurement between a sensor (i) and an anchor (k) with known location. Corresponding to these two types of distances, we define sets N_1, $\overline{N}_1$, N_2, and $\overline{N}_2$ as follows:

- N_1 includes pairwise sensors (i, j) if $i < j$ and there exists a distance measurement $\widehat{d}_{ij}$:
$$N_1 = \{(i, j) \text{ with known } \widehat{d}_{ij} \text{ and } i < j\}.$$
- $\overline{N}_1$ includes pairwise sensors (i, j) with unknown measurement $\widehat{d}_{ij}$ and $i < j$:
$$\overline{N}_1 = \{(i, j) \text{ with unknown } \widehat{d}_{ij} \text{ and } i < j\}.$$
- N_2 includes pairs of sensor i and anchor k if there exists a measurement $\widehat{d}_{ik}$:
$$N_2 = \{(i, k) \text{ with known } \widehat{d}_{ik}\}.$$
- $\overline{N}_2$ includes pairs of sensor i and anchor k with unknown measurement $\widehat{d}_{ik}$:
$$\overline{N}_2 = \{(i, k) \text{ with unknown } \widehat{d}_{ik}\}.$$

The full set of nodes and pairwise distance measurements form a graph $G = \{V, E\}$, where $V = \{1, 2, \ldots, s, s + 1, \ldots, n\}$ and $E = N_1 \cup N_2$.

Introduce α_{ij} to be the difference between the measured squared distance $(\widehat{d}_{ij})^2$ and the squared Euclidean distance $\|x_i - x_j\|^2$ from sensor i to sensor j. Also, let α_{ik} be the difference between the measured squared distance $(\widehat{d}_{ik})^2$ and the squared Euclidean distance $\|x_i - a_k\|^2$ from sensor i to anchor k. Intuitively, we seek a solution for which the magnitude of these differences is small.

Lower bounds r_{ij} or r_{ik} are imposed if $(i,j) \in \overline{N}_1$ or if $(i,k) \in \overline{N}_2$. Typically each r_{ij} or r_{ik} value is the radio range (also known as *radius*) within which the associated sensors can detect each other.

Biswas and Ye [2] formulate the sensor localization problem as minimizing the ℓ_1-norm of the squared-distance errors α_{ij} and α_{ik} subject to mixed equality and inequality constraints:

$$
\begin{aligned}
\underset{x_i, x_j, \alpha_{ij}, \alpha_{ik}}{\text{minimize}} \quad & \sum_{(i,j)\in N_1} |\alpha_{ij}| \; + \sum_{(i,k)\in N_2} |\alpha_{ik}| \\
\text{subject to} \quad & \|x_i - x_j\|^2 - \alpha_{ij} = (\widehat{d}_{ij})^2 \quad \forall\ (i,j) \in N_1, \\
& \|x_i - a_k\|^2 - \alpha_{ik} = (\widehat{d}_{ik})^2 \quad \forall\ (i,k) \in N_2, \\
& \|x_i - x_j\|^2 \qquad\ \geq r_{ij}^2 \qquad \forall\ (i,j) \in \overline{N}_1, \\
& \|x_i - a_k\|^2 \qquad\ \geq r_{ik}^2 \qquad \forall\ (i,k) \in \overline{N}_2, \\
& x_i,\ x_j \in \mathcal{R}^2, \qquad \alpha_{ij},\ \alpha_{ik} \in \mathcal{R}, \\
& i,j = 1,\ldots,s, \qquad k = s+1,\ldots,n.
\end{aligned}
\tag{2.1}
$$

The above model is a nonconvex constrained optimization problem. As yet there is no effective solution method. In the following subsections, we review Biswas and Ye's [2] relaxation method for solving this problem approximately.

2.2. The Euclidean distance model in matrix form. The distance model (2.1) is reformulated into (2.2) (refer to Biswas and Ye [2]) by introducing matrix variables as follows:

$$
\begin{aligned}
\text{minimize} \quad & \sum_{(i,j)\in N_1} (\alpha_{ij}^+ + \alpha_{ij}^-) \; + \sum_{(i,k)\in N_2} (\alpha_{ik}^+ + \alpha_{ik}^-) \\
\text{subject to} \quad & e_{ij}^T\ Y\ e_{ij} - \alpha_{ij}^+ + \alpha_{ij}^- = (\widehat{d}_{ij})^2 \quad \forall\ (i,j) \in N_1, \\
& \begin{pmatrix} e_i \\ -a_k \end{pmatrix}^T \begin{pmatrix} Y & X^T \\ X & I \end{pmatrix} \begin{pmatrix} e_i \\ -a_k \end{pmatrix} - \alpha_{ik}^+ + \alpha_{ik}^- = (\widehat{d}_{ik})^2 \quad \forall\ (i,k) \in N_2, \\
& e_{ij}^T\ Y\ e_{ij} \geq r_{ij}^2 \qquad \forall\ (i,j) \in \overline{N}_1, \\
& \begin{pmatrix} e_i \\ -a_k \end{pmatrix}^T \begin{pmatrix} Y & X^T \\ X & I \end{pmatrix} \begin{pmatrix} e_i \\ -a_k \end{pmatrix} \geq r_{ik}^2 \qquad \forall\ (i,k) \in \overline{N}_2, \\
& Y = X^T X, \\
& \alpha_{ij}^+,\ \alpha_{ij}^-,\ \alpha_{ik}^+,\ \alpha_{ik}^- \geq 0, \\
& i,j = 1,\ldots,s, \quad k = s+1,\ldots,n,
\end{aligned}
\tag{2.2}
$$

where

- $X = (x_1\ x_2\ \ldots\ x_s)$ is a $2 \times s$ matrix to be determined;
- e_{ij} is a zero column vector except for 1 in location i and -1 in location j, so that

$$\|x_i - x_j\|^2 = e_{ij}^T\ X^T X\ e_{ij};$$

- e_i is a zero column vector except for 1 in position i, so that
$$\|x_i - a_k\|^2 = \begin{pmatrix} e_i \\ -a_k \end{pmatrix}^T \begin{pmatrix} X & I \end{pmatrix}^T \begin{pmatrix} X & I \end{pmatrix} \begin{pmatrix} e_i \\ -a_k \end{pmatrix};$$
- Y is defined to be X^TX;
- The substitutions $\alpha_{ij} = \alpha_{ij}^+ - \alpha_{ij}^-$ and $\alpha_{ik} = \alpha_{ik}^+ - \alpha_{ik}^-$ are made to deal with $|\alpha_{ij}|$ and $|\alpha_{ik}|$ in the normal way.

2.3. The SDP relaxation model. The approach of Biswas and Ye [2] is to relax the constraint $Y = X^TX$ to be $Y \succeq X^TX$, for which an equivalent matrix inequality is (Boyd et al. [5])

$$Z_I \equiv \begin{pmatrix} Y & X^T \\ X & I \end{pmatrix} \succeq 0. \tag{2.3}$$

With the definitions

$$A_I = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}, \qquad b_I = \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix},$$

where $\mathbf{0}$ in A_I is a zero column vector of dimension s, problem (2.2) is relaxed to a linear SDP:

$$\begin{aligned}
\text{minimize} \quad & \sum_{(i,j)\in N_1} (\alpha_{ij}^+ + \alpha_{ij}^-) \ + \sum_{(i,k)\in N_2} (\alpha_{ik}^+ + \alpha_{ik}^-) \\
\text{subject to} \quad & \operatorname{diag}(A_I^T \ Z \ A_I) = b_I, \\
& \begin{pmatrix} e_{ij} \\ \mathbf{0} \end{pmatrix}^T Z \begin{pmatrix} e_{ij} \\ \mathbf{0} \end{pmatrix} - \alpha_{ij}^+ + \alpha_{ij}^- = (\widehat{d}_{ij})^2 \quad \forall\ (i,j) \in N_1, \\
& \begin{pmatrix} e_i \\ -a_k \end{pmatrix}^T Z \begin{pmatrix} e_i \\ -a_k \end{pmatrix} - \alpha_{ik}^+ + \alpha_{ik}^- = (\widehat{d}_{ik})^2 \quad \forall\ (i,k) \in N_2, \\
& \begin{pmatrix} e_{ij} \\ \mathbf{0} \end{pmatrix}^T Z \begin{pmatrix} e_{ij} \\ \mathbf{0} \end{pmatrix} \geq r_{ij}^2 \qquad \forall\ (i,j) \in \overline{N}_1, \\
& \begin{pmatrix} e_i \\ -a_k \end{pmatrix}^T Z \begin{pmatrix} e_i \\ -a_k \end{pmatrix} \geq r_{ik}^2 \qquad \forall\ (i,k) \in \overline{N}_2, \\
& Z \succeq 0, \qquad \alpha_{ij}^+,\ \alpha_{ij}^-,\ \alpha_{ik}^+,\ \alpha_{ik}^- \geq 0, \\
& \qquad\qquad i,j = 1,\dots,s, \quad k = s+1,\dots,n,
\end{aligned} \tag{2.4}$$

where the constraint $\operatorname{diag}(A_I^T Z A_I) = b_I$ ensures that the matrix variable Z's lower right corner is a 2×2 identity matrix I, so that Z takes the form of Z_I in (2.3).

Initially, Biswas and Ye [2, 3] omit the $\geq$ inequalities involving r_{ij} and r_{ik} and solve the resulting problem to obtain an initial solution Z_1. (The inequality constraints increase the problem size dramatically, and Z_1 is likely to satisfy most of them.) They then adopt an "iterative active-constraint generation technique" in which inequalities violated by Z_k are added to the problem and the resulting SDP is solved to give Z_{k+1} ($k = 1, 2, \dots$). The process usually terminates before all constraints are included. Further study of this approach is presented in section 4.1.

2.4. SDP model analysis. Let $\bar{Z} = \begin{pmatrix} \bar{Y} & \bar{X}^T \\ \bar{X} & I \end{pmatrix}$ be a feasible solution of the relaxed SDP (2.4). Assuming the distance measurements are exact (no noise), Biswas and Ye [2] give conditions under which $\bar{X}$ and $\bar{Y}$ solve problem (2.2) exactly as follows:

- $\bar{Z}$ is the unique optimal solution of (2.4), including all inequality constraints.
- In (2.4), there are $2n + n(n+1)/2$ exact pairwise distance measurements.

These conditions ensure that $\bar{Y} = \bar{X}^T\bar{X}$. In practice, distance measurements have noise and we only know that the SDP solution satisfies $\bar{Y} - \bar{X}^T\bar{X} \succeq 0$. This inequality can be used for error analysis of the location estimates provided by the relaxation. For example, $\text{trace}(\bar{Y} - \bar{X}^T\bar{X}) = \sum \tau_i$, where

$$\tau_i \equiv \bar{Y}_{ii} - \|\bar{x}_i\|^2 \geq 0, \tag{2.5}$$

is a measure of deviation of the SDP solution from the desired constraint $Y = X^TX$ (ignoring off-diagonal elements). The individual trace τ_i can be used to evaluate the location estimate $\bar{x}_i$ for sensor i. In particular, we interpret a smaller τ_i to mean higher accuracy in the estimated location x_i. Further explanation is given in [2].

3. SpaseLoc: A scalable localization algorithm. When the number of nodes in (2.4) is large, applying a general SDP solver such as DSDP5.0 [1] or SeDuMi [23] would not scale well. In this section, we present a sequential subproblem approach named SpaseLoc to solve the full localization problem iteratively.

3.1. Adaptive subproblem approach. We call the overall sensor localization problem including all sensors and anchors the *full_problem*. At each iteration, SpaseLoc selects from the *full_problem* a subset of the unlocalized sensors and a subset of the anchors to form a localization *subproblem*. We call the selected sensors in the subproblem *subsensors*, and the selected anchors in the subproblem *subanchors*, These subsensors and subanchors, together with their known distance measurements and known anchors' locations, form a sub-SDP relaxation model to be solved using the same formulation as in (2.4).

In our adaptive approach, the subsensors and subanchors for each subproblem are chosen dynamically according to rule sets. (Rather than using predefined data, every new iteration's subproblem generation is based on the previous iteration's results.) The resulting SDP subproblems are of varying but limited size. Currently they are solved by Benson, Ye, and Zhang's SDP solver DSDP5.0 [1].

SpaseLoc is a *greedy algorithm* in the sense that each subproblem determines the final estimate of the associated sensor locations.

3.2. The SpaseLoc algorithm. The main steps of SpaseLoc are listed below, followed by explanations of the steps and definitions of new terms used therein.

A0. Set *subproblem_size*.
A1. Subproblem creation: Select subsensors and subanchors to be included in the subproblem.
A2. Formulate SDP relaxation model (2.4) based on the chosen subsensors and subanchors, together with the known distances among them and the subanchors' known locations.
A3. Call SDP solver to obtain a solution for the subsensors' locations.
A4. Classify localized subsensors according to their τ_i value.
A5. If all sensors in the network become localized or are determined to be outliers, go to step A6. Otherwise, return to step A1 for the next iteration.
A6. Output all sensor locations and report outliers if any. Stop.

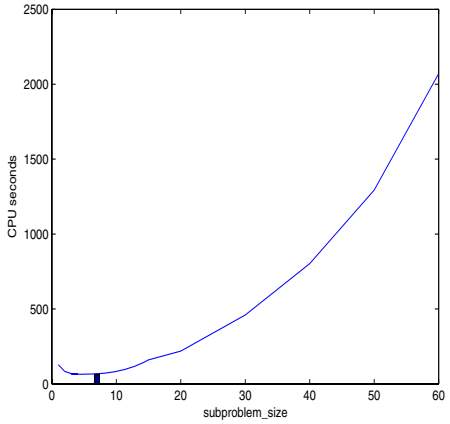


FIG. 3.1. *SpaseLoc execution time as a function of subproblem_size: total nodes* = 10000, *anchors* = 100, *radius* = 0.0226.

In step A0, *subproblem_size* specifies a limit on the number of unlocalized sensors to be included in each subproblem. It can range from 1 to an upper limit value that is potentially solvable by the SDP solver. In our experiments, the upper limit is 150. The most effective *subproblem_size* seems to change with the *full_problem* size, the model parameters such as *radius*, and the SDP solver used. We perform an approximate linesearch to find *subproblem_size* that corresponds to the minimum time because, empirically, the total execution time with all other parameters fixed is essentially a convex function of *subproblem_size*.

For example, when *full_problem* size is 10000 with 100 anchors, *radius* 0.0226, and no noise, *subproblem_size* 7 seems to give the best execution time with the DSDP5.0 solver (refer to Figure 3.1). The search time for *subproblem_size* is not included as part of the SpaseLoc execution time.

Step A1 involves choosing a subset of unlocalized sensors (no more than *subproblem_size*) and an associated subset of nodes with known locations. The latter can include a subset of the original anchors and/or a subset of sensors already localized by a previous subproblem (we define them as *acting anchors*). The rules for choosing subsensors and subanchors in this iteration are discussed in sections 3.4–3.5.

In step A4, the error in sensor i's location is estimated by its individual trace τ_i as discussed in section 2.4. Subsensors whose τ_i value is within a given tolerance τ are labeled as localized and treated as acting anchors for the next iteration, whereas subsensors whose localization error is higher than the tolerance are also labeled as localized but are not used as acting anchors in later iterations. These new acting anchors are labeled with different acting levels as explained in section 3.4. The value of τ has an impact on the localization accuracy. Bigger values allow more localized sensors to be acting anchors, but with possibly greater transitive errors. Smaller values may increase the estimation accuracy for some of the sensors, but could lead to fewer connections to anchors for some unlocalized sensors. A rule of thumb is to use a small τ for networks with high anchor density to achieve potentially more accuracy, and a bigger τ for networks with low anchor density to avoid lacking connections to anchors. In order to avoid the side effect of a bigger τ eliminating too many potential acting anchors, at some later iteration we utilize all localized sensors as acting anchors

(including those whose τ_i value is bigger than the given tolerance τ). This change only starts when the remaining unlocalized sensors are connected to fewer than three anchors. It makes sure that we use acting anchors with higher accuracy first, but if no such acting anchors are available, we use localized points whose locations might be less accurate. In most cases, this is better than using no reference points at all.

In step A5, an unlocalized sensor is called an *outlier* when it does not have any distance information for the algorithm to decide its location. If a sensor has no connection to any anchor, it is classified as an outlier. In addition, if a connected cluster of sensors has no connection to any anchors, then all sensors in the cluster will be outliers.

The next sections explain the subproblem creation procedure used by step A1. Section 3.3 lists steps S1–S8 of the creation procedure itself. Section 3.4 presents rules RS1–RS4 for subsensor selection in step S5. Section 3.5 presents rules RA1–RA3 for subanchor selection in step S7. Section 3.6 illustrates the method for independent subanchor selection used in rules RA2–RA3. Sections 3.7–3.8 discuss the routines used in step S8 to localize sensors that have fewer than 3 connected anchors.

3.3. Subproblem creation procedure. As explained, *subproblem_size* is a predetermined parameter that represents the maximum number of unlocalized sensors that can be selected as subsensors in a subproblem. When there are more than *subproblem_size* unlocalized sensors, we have a choice to make among them.

The subproblem creation procedure makes sure that the choice of subsensors is based first on the number of connected anchors they have, and second on the type of connected anchors such as original anchors and different levels of acting anchors as defined in section 3.4, and that the choice of subanchors is based on a set of rules (section 3.5). The main steps are listed below, followed by explanations of the steps and definitions of new terms used.

S1. Specify *MaxAnchorReq*.

S2. Initialize *AnchorReq* = *MaxAnchorReq*.

S3. Loop through unlocalized sensors, finding all that are connected to at least *AnchorReq* anchors. If *AnchorReq* ≥ 3, determine if there are 3 *independent* subanchors; if not, go to the next sensor.[1] Enter each found sensor into a *candidate subsensor list*, and enter its connected anchors into a corresponding *candidate subanchor list*. Each sensor in the candidate subsensor list has its own candidate subanchor list (so there are as many candidate subanchor lists as the number of sensors in the candidate subsensor list). Let *sub_s_candidate* be the length of the candidate subsensor list.

S4. If $0 <$ *sub_s_candidate* $\leq$ *subproblem_size*, the candidate subsensor list becomes the chosen subsensors list. Go to step S7.

S5. If *sub_s_candidate* $>$ *subproblem_size*, the choice of subsensors is further based on subsensor selection rules RS1–RS4 described in section 3.4. After exactly *subproblem_size* subsensors are selected from the candidate list according these rules, go to step S7.

S6. Now *sub_s_candidate* $= 0$. Reduce *AnchorReq* by 1.
If *AnchorReq* > 0, go to step S3 for another round of subproblem creation.
Otherwise, *AnchorReq* $= 0$ and *sub_s_candidate* $= 0$ indicates that there are still unlocalized sensors left that are not connected to any localized node. We classify them as outliers and exit this procedure to continue at step A6 of

[1] See section 3.6 for dependency definition and independent anchor selection.

section 3.2.

S7. Now that we have a subsensor list and the candidate subanchor lists, choose subanchors using selection rules RA1–RA3 presented in section 3.5.

S8. The subsensors and subanchors are selected and the subproblem creation routine finishes here.
If $AnchorReq \geq 3$, go to step A2 in section 3.2.
If $AnchorReq = 2$, apply the procedure in section 3.7 and go to step A5.
If $AnchorReq = 1$, apply the procedure in section 3.8 and go to step A5.

In step S1, *MaxAnchorReq* determines the initial (maximum) value of *AnchorReq*. It is useful for scalability when connectivity is dense. A smaller *MaxAnchorReq* would generally cause fewer subanchors to be included in the subproblem, thus reducing the number of distance constraints in each SDP subproblem and hence reducing execution time for each iteration. For instance, under ideal conditions (where there is no noise), even if a sensor has 10 distance measurements to 10 anchors, we don't need to include all 10 anchors because we can use 3 to localize that sensor accurately.

In the presence of noise, a bigger *MaxAnchorReq* should reduce the average estimation error. For example, if there is a large distance measurement error from one particular anchor, since *MaxAnchorReq* anchors are all taken into consideration for deciding the sensor's actual location, the large error would be averaged out. Another consideration for setting *MaxAnchorReq* is the trade-off between estimation accuracy and execution speed. If we are in a static environment and would like to have localization as accurate as possible under noise conditions, we might choose a large *MaxAnchorReq*. However, in a real-time environment involving moving sensors, where speed might take priority, we would consider a smaller *MaxAnchorReq*.

In step S2, *AnchorReq* is a dynamic parameter that may decrease in later steps.

In step S4, the subproblem may contain fewer than *subproblem_size* subsensors, which is perfectly acceptable. The alternative is to reduce *AnchorReq* by 1 and find more subsensor candidates that have fewer distance connections. However, this approach might reduce the accuracy of the algorithm, because we do want to localize the subsensors as accurately as possible as the iteration progresses, and the newly localized subsensors could be further used as acting anchors for the next iteration.

In step S6, *AnchorReq* is iteratively reduced by 1 from *MaxAnchorReq* to 0 eventually. This approach allows sensors with at least *AnchorReq* connections to anchors to be localized before sensors with fewer connections to anchors.

As we know, under no-noise conditions, a sensor's location can be uniquely determined by connections to at least 3 independent anchors. If a sensor has connections to only 2 anchors, there are two possible locations; and if there is only 1 connection to an anchor, the sensor can be anywhere on a circle. In step S8, we use heuristic subroutines described in sections 3.7–3.8 to include the sensor's anchors' connected neighboring nodes in the subproblem in order to improve the estimation accuracy.

3.4. Subsensor selection. In step S5, when the number of sensors in the candidate subsensor list is bigger than *subproblem_size*, the choice of subsensors is further based on the types of anchors each sensor is connected to.

First, we introduce the concept of *sensor priority*. We assign a priority to each sensor in the candidate subsensor list. A sensor with a smaller priority value is selected to be localized before one with a bigger priority value. A sensor's priority is based on the types of anchors the sensor is directly connected to. Next, in order to define different types of anchors, we introduce the concept of *anchor acting levels*. All anchors including acting anchors are assigned certain acting levels. Original anchors

TABLE 3.1
An example: priority list when MaxAnchorReq = 3.

Priority value	Level 1 anchor	Level 3 anchor	Level 5 anchor	Level 7 anchor	Level 9 anchor	...	Resulting anchor level
1	≥ 3	any					3
2	$= 2$	≥ 1	any				5
3	$= 1$	≥ 2	any				7
3	$= 2$	$= 0$	≥ 1	any			7
4	$= 1$	$= 1$	≥ 1	any			9
4	$= 2$	$= 0$	$= 0$	≥ 1	any		9
5	$= 1$	$= 1$	$= 0$	≥ 1	any		11
5	$= 2$	$= 0$	$= 0$	$= 0$	≥ 1	any	11
...	$= 0$	total ≥ 3					(11, *bigN*)
bigN	total $= 2$						*bigN*
bigN+1	total $= 1$						*bigN*+1

are always set to acting level 1. Every acting anchor is set to an acting level after it has been localized as a sensor. Essentially, acting anchors are set with acting levels depending on the levels of the anchors that localized them.

The priority rules for selecting subsensors from a candidate subsensor list are as follows:

RS1. When $AnchorReq \geq 3$ and a sensor has at least 3 connected anchors that are independent, the sensor's priority depends on the 3 connected anchors that have the lowest acting levels among all its connected anchors. The sensor's priority value is defined as the summation of these 3 connected anchors' acting levels.

RS2. If the sensor has 3 connected anchors that are dependent, it is ranked with the same priority as when the sensor is connected to only 2 anchors.

RS3. Sensors with 2 anchor connections are ranked with equal priority, independent of the acting levels of the 2 connected anchors. (This can be easily expanded to be more granular according to the connected anchors' acting levels.) Sensors in this category are assigned lower priority than any sensors that have at least 3 independent anchor connections.

RS4. Sensors with 1 anchor connection are ranked with equal priority, independent of the acting level of the connected anchor. (Again, this can be more granular according to the connected anchor's acting level.) Sensors in this category are assigned lower priority than any sensors that have at least 2 anchor connections.

Table 3.1 illustrates the priority list for an example where $MaxAnchorReq = 3$ and the sensor's priority is determined by the 3 anchors that have the lowest acting levels among all the sensor's connected anchors. We can certainly add more granularity by further classifying the acting levels of the sensor's fourth or fifth connected anchors (if any). Although more categorization of the priorities should increase localization accuracy under most noise conditions, more computational effort is required to handle more levels of priorities.

Each item in the table represents the number of anchors with different acting levels that are needed at each priority. The last column represents the resulting acting anchors' acting levels for subsequent iterations. For example, if a sensor has at least three independent connections to anchors, and if 3 of the anchors are original anchors (acting level 1), this sensor belongs to priority 1 as listed in row 1 of the table. Also, when this sensor is localized, it becomes acting anchor level 3 (the summation of the

anchor levels of the three anchors that localized it). Similarly, if a sensor has at least three independent connections to anchors, and if 2 of the anchors are original anchors (acting level 1) and at least 1 of the connected anchors is at acting level 3, then this sensor belongs to priority 2 as listed in row 2 of the table. Also, when this sensor is localized, it becomes acting anchor level 5. The sensors that connect to two anchors belong to the second to last priority, and sensors that connect to only one anchor belong to the last priority. We use a big enough number *bigN* in the implementation to ensure that sensors connected to fewer than 3 anchors are given the lowest priority.

3.5. Subanchor selection. In step S7, for each unlocalized subsensor in the subsensor list, only *AnchorReq* of the connected anchors are allowed to be included in the subproblem. We use the following rules to select subanchors from a candidate subanchor list that contains more than *AnchorReq* anchors.

RA1. Original anchors are selected first, followed by acting anchors with lower acting level.

RA2. The subanchors chosen should be linearly independent.

RA3. Among independent anchors in the candidate subanchor list, we use distance scale-factors to encourage selection of the closest subanchors.

Rules RA2 and RA3 are implemented as in section 3.6. Rule RA3 is based on the assumption that under noise conditions, we trust the shorter distance measurements more than the longer ones.

3.6. Independent subanchors selection. Suppose sensor i is connected to K $(K > 3)$ anchors at locations a_{ik} with corresponding distance measurements $\widehat{d}_{ik}$ $(k = 1, \ldots, K)$. Define the matrices

$$A = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ -a_{i1} & -a_{i2} & \cdots & -a_{iK} \end{pmatrix}, \quad D_1 = \operatorname{diag}(1/\sqrt{1 + \|a_{ik}\|^2}), \quad D_2 = \operatorname{diag}(1/\widehat{d}_{ik}).$$

We select an independent subset by a QR factorization with column interchanges [11]: $B = AD_1D_2$, $BP = QR$, where Q is orthogonal, R is upper-trapezoidal, and P is a permutation chosen to maximize the next diagonal of R at each stage of the factorization. (D_1 normalizes the columns of A, and D_2 biases them in favor of anchors that are closer to sensor i.) If the 3rd diagonal of R is larger than a predefined threshold (10^{-4} is used in our simulation), then the first 3 columns of AP are regarded as independent, and the associated anchors are chosen. Otherwise, all subsets of 3 among the K anchors are regarded as dependent. (In MATLAB, R and P are obtained by a command of the form `[Q,R,P] = qr(B)`.)

3.7. Geometric subroutine (two connected anchors). This section illustrates the heuristic techniques used in step S8 of section 3.3 to localize sensors connected to only two anchors.

When a sensor's connected anchors are also connected to other anchors, this subroutine may improve the accuracy of the sensor's localization, as illustrated by an example in Figure 3.2.

In this example, assume s_1 and s_2 are sensors with unknown locations, and $a_3(1,3)$, $a_4(1,2)$, $a_5(2,2)$, $a_6(4,1)$, $a_7(5,1)$ are anchors with known locations in brackets. Assume that the sensors' radio range is $\sqrt{2}$, and we are also given two distance measurements $\widehat{d}_{13} = 1$ and $\widehat{d}_{14} = \sqrt{2}$ for sensor s_1 and one measurement $\widehat{d}_{27} = 1$ for sensor s_2.

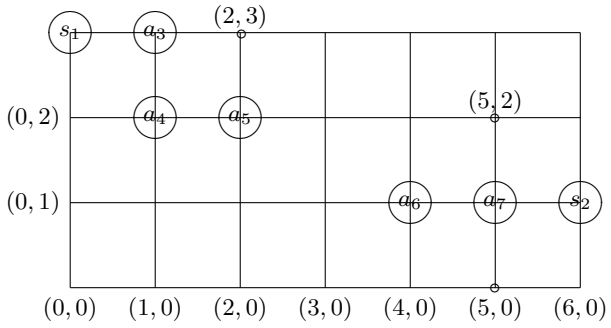


FIG. 3.2. *Sensors with connections to at most two anchors.*

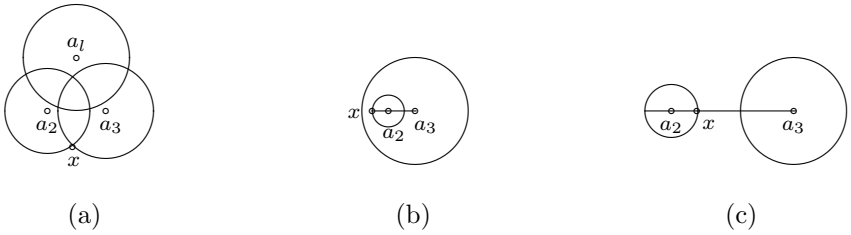


FIG. 3.3. (a) *Sensor with two anchors' circles intersecting.* (b) *Sensor with two anchors, a_2's circle in a_3's.* (c) *Sensor with two anchors' circles disjoint.*

Given two distances $\widehat{d}_{13}$ and $\widehat{d}_{14}$ to two anchors $a_3(1,3)$ and $a_4(1,2)$, we know that s_1 should be at either $(0,3)$ or $(2,3)$. If we only use s_1, $a_3(1,3)$, $a_4(1,2)$ in an SDP subproblem, then SDP relaxation will give a solution near the middle of the two possible points, which would be very close to point $(1,3)$. If there is any anchor (a_5) that is near s_1's connected anchors (a_3 and a_4) with any of the two possible sensor points within their radio range (point $(2,3)$ is within a_5's range), that point $(2,3)$ must not be the real location of s_1, or else s_1 would be connected to this anchor (a_5) as well. Thus we can infer that s_1 must be at the other point $(0,3)$.

Inspired by the above observation, when a sensor has at most 2 connected anchors, we include these anchors' connected anchors in the subproblem (we call them the connected anchors' neighboring anchors) together with the sensor and its directly connected anchors. By including the neighboring anchors, we might hope that the inequality constraints in the SDP relaxation model (2.4) would push the estimation towards the right point. However, because of the relaxation, enforcing inequalities in (2.4) is not equivalent to enforcing them in the distance model (2.2). The added inequality constraints only push the original solution near $(1,3)$ a tiny bit towards s_1's real location $(0,3)$, and the solution essentially stays at around $(1,3)$.

Given the ineffectiveness of the SDP relaxation approach under this condition, we propose instead a geometric approach as illustrated in Figure 3.3. Assume $s_1(x_x, x_y)$ has measurements $\widehat{d}_{12}$ to anchor $a_2(a_{2x}, a_{2y})$ and $\widehat{d}_{13}$ to anchor $a_3(a_{3x}, a_{3y})$. We also assume $\widehat{d}_{12} \le \widehat{d}_{13}$ (we can always swap the two indexes otherwise). Let a_l $(l = 4, \ldots, k)$ be a_2 and/or a_3's neighboring anchors with radio range r_{1l} $(l = 4, \ldots, k)$, and let d_{23} be the known (exact) Euclidean distance between a_2 and a_3.

- If two circles centered at a_2 and a_3 with radii $\widehat{d}_{12}$ and $\widehat{d}_{13}$ intersect each other ($\widehat{d}_{12} + \widehat{d}_{13} \ge d_{23}$ and $\widehat{d}_{13} - \widehat{d}_{12} \le d_{23}$) as in Figure 3.3(a):
  - Two possible locations of s_1 are given by solutions x^* and x^{**} of the

equations

$$\|x - a_2\|^2 = \widehat{d}_{12}^2, \qquad \|x - a_3\|^2 = \widehat{d}_{13}^2.$$

– Sensor s_1's location is selected from x^* and x^{**}, whichever is further away from any neighboring anchor. Thus, for $l = 4$ to k,

if $\|x^* - a_l\|^2 < r_{1l}^2$, then $x = x^{**}$ and stop
else if $\|x^{**} - a_l\|^2 < r_{1l}^2$, then $x = x^*$ and stop.

Otherwise, $x = (x^* + x^{**})/2$ and stop.

- Under noise conditions, the a_2 circle may be inside the a_3 circle ($\widehat{d}_{12} + \widehat{d}_{13} \geq d_{23}$ and $\widehat{d}_{13} - \widehat{d}_{12} > d_{23}$) as in Figure 3.3(b).
  – The solutions x^* and x^{**} of the following equations give two possible points for s_1 on the a_2 circle:

$$(x_x - a_{2x})^2 + (x_y - a_{2y})^2 = \widehat{d}_{12}^2,$$
$$(a_{2x} - a_{3x})(x_y - a_{2y}) = (a_{2y} - a_{3y})(x_x - a_{2x}),$$

  where x is on the line through a_2 and a_3 represented by the second equation.
  – If $\|x^* - a_3\| < \|x^{**} - a_3\|$, then $x = x^{**}$; otherwise $x = x^*$. This guarantees that the point further from a_3 is chosen. Note that we base the sensor's estimation on the closest anchor (a_2 here since $\widehat{d}_{13} \geq \widehat{d}_{12}$), assuming that a shorter measurement is generally more accurate than longer ones, given similar anchor properties.

  The same approach applies when the a_3 circle is inside the a_2 circle ($\widehat{d}_{12} - \widehat{d}_{13} > d_{23}$).
- Under noise conditions, the a_2 and a_3 circles may again have no intersection ($\widehat{d}_{12} + \widehat{d}_{13} < d_{23}$) as in Figure 3.3(c).
  – The solutions x^* and x^{**} of the following equations give two possible points for s_1 on the circle for the anchor with smaller *radius*. Let us assume $\widehat{d}_{12} \leq \widehat{d}_{13}$:

$$(x_x - a_{2x})^2 + (x_y - a_{2y})^2 = \widehat{d}_{12}^2,$$
$$(a_{2x} - a_{3x})(x_y - a_{2y}) = (a_{2y} - a_{3y})(x_x - a_{2x}),$$

  where x is on the line through a_2 and a_3 represented by the second equation.
  – If $\|x^* - a_3\| > \|x^{**} - a_3\|$, then $x = x^{**}$; otherwise $x = x^*$. This guarantees that the point closer to a_3 (in between a_2 and a_3) is chosen.

3.8. Geometric subroutine (one connected anchor). Similar inefficiency occurs in the SDP solution when a sensor connects to only *one* anchor. The SDP solver under this condition gives a solution for the sensor to be in the same location as the sensor's connected anchor. In reality, the sensor could be anywhere on the circle. The SDP gives an average point, at the center of the circle, and that is where the connected anchor is. Even if the anchor's neighboring anchor is included in the SDP subproblem, the inequality constraints are not active most of the time because the SDP solution may not provide optimal solutions all the time.

We propose a heuristic for estimating a sensor's location with only one connecting anchor. The idea is to use one neighboring anchor's radio range information to

FIG. 3.4. (a) *Sensor with one anchor connection a and one neighboring anchor b.* (b) *Sensor with one anchor connection a and two neighboring anchors b, c.*

eliminate the portion of the circle that the sensor would not be on, and then calculate the middle of the other portion of the circle to be the sensor's location. For the example in Figure 3.2, because we know the distance between s_2 and a_7 is 1, we know that s_2 could be anywhere on the circle surrounding a_7 with *radius* 1. Knowing a_7's neighboring anchor node a_6 is not connected to s_2, we know that s_2 would not be in the area surrounding a_6 with *radius* $\sqrt{2}$. Thus, s_2 could be anywhere around the half circle including points $(5,2)$, $(6,1)$, $(5,0)$. The heuristic chooses the middle point between the two circles' intersection points $(5,2)$ and $(5,0)$, which happens to be $(6,1)$ in this example. The heuristic gives better accuracy for the sensor's location than the SDP solution under most conditions. The procedure follows.

- Assume s has one distance measurement $\widehat{d}$ to anchor a, and b is the closest connected neighboring anchor to a with radio range r (refer to Figure 3.4(a)). We assume $a = (a_x, a_y)$, $b = (b_x, b_y)$, $x = (x_x, x_y)$.
- The solutions x^* and x^{**} of the following equations give two possible points s on the circle:

$$(x_x - a_x)^2 + (x_y - a_y)^2 = \widehat{d}^2,$$
$$(a_x - b_x)(x_y - a_y) = (a_y - b_y)(x_x - a_x),$$

  where x is on the line through a and b represented by the second equation.
- If $\|x^* - b\| < r$, then $x = x^{**}$; otherwise $x = x^*$. This guarantees that the point further from b is chosen.

The above heuristic provides a simple way of estimating a sensor's location when the sensor connects to only one anchor. A more complicated approach can be adopted when the connected anchor has more than one neighboring anchor, which can increase the accuracy of the sensor's location. We call it an arc elimination heuristic. The idea is to loop through each of the neighboring anchors and find the portion of the circle that the sensor won't be on, and eliminate that arc as a possible location of the sensor. Eventually, when one or more plausible arcs remain, we choose the middle of the largest arc to be the sensor's location. For example, assume we add one more neighboring anchor c to sensor s's anchor a from the previous example in Figure 3.4(a). The new scenario is shown in Figure 3.4(b). First, we find the intersections (points 1 and 2) of two circles: one at a with *radius* $\widehat{d}$, the other at b with *radius* r. We know that the 1–2 portion of the arc closer to point b won't be the location of s. Second, we find the intersections (points 3 and 4) of two circles: one at a with *radius* $\widehat{d}$, the other at c with *radius* r. We know that the arc 3–4 closer to point c won't be the location of s. Thus we deduce that s must be somewhere on the arc 1–4 further away from b or c. The estimation of s is given in the middle of the arc 1–4. As we see, this method should provide more accuracy than the one-neighboring-anchor approach.

3.9. Subproblem optimality. For the case of one sensor connected to three independent anchors, Biswas and Ye [2] prove when there is no noise that the SDP relaxation (2.4) gives an optimal solution to (2.2). The proof depends on the fact that there are three independent equations and only three variables.

In SpaseLoc, the subproblems are constructed from sensors that have three independent anchors (or acting anchors) where possible. If each of these subsensors (say total s) were included in separate subproblems together with their connected 3 independent anchors, the proof in [2] shows that they would be localized exactly by the SDP approach. If these s subsensors and their connected anchors are treated together in a single subproblem, the larger SDP relaxation contains sets of the same three equations that would arise in the separate SDP relaxations. The equations form a block-diagonal system in the larger SDP. There are $3s$ independent equations and the same number of variables containing only x_i and $y_{ii},\ i = 1, \ldots, s$. The $\widehat{d}_{ik}$ equation in (2.4) reduces under no noise conditions to $y_{ii} = x_i^T x_i$ for all relevant pairs (i, k). The constraint $Y - X^T X \succeq 0$ then guarantees $y_{ij} = x_i^T x_j$ for all $j = 1, \ldots, s$. Hence, the SDP solution for the SpaseLoc subproblem is also rank 2 and gives an exact locations for all subsensors.

4. Computational results. This section explains the simulation method and the setup for experimenting with the SpaseLoc algorithm, then presents results for various parameter settings.

For the simulation, a total number of nodes n (including s sensors and m anchors) is specified in the range 4 to 10000. The locations of these nodes are assigned with a uniform random distribution on a square region of size $r \times r$ where $r = 1$, or put on the grid-points of a regular topology such as a square or an equilateral triangle on the same region. The m anchors are randomly chosen from the given n nodes. We assume all sensors have the same radio range (*radius*) for any given test case. Various radio ranges were tested in the simulation.

Euclidean distances $d_{ij} = \|x_i - x_j\|$ are calculated among all sensor pairs $(i,\ j)$ for $i < j$. We then use $\widehat{d}_{ij}$ to simulate measured distances, where $\widehat{d}_{ij}$ is d_{ij} times a random error simulated by *noise_factor* $\in [0, 1]$. For a given *radius* $\subseteq [0, 1]$ it is defined as follows:

- If $d_{ij} \leq$ *radius*, then $\widehat{d}_{ij} = d_{ij}(1 + \texttt{rn} * \textit{noise\_factor})$, where `rn` is normally distributed with mean zero and variance one. (Any numbers generated outside $(-1, 1)$ are regenerated.)
  In practical networks, depending on the technologies that are being used to obtain the distance measurements, there may be many factors that contribute to the noise level. For example, one way to obtain the distance measurement is to use the received radio signal strength between two sensors. The signal strength could be affected by media or obstacles in between the two sensors. In this study, *noise_factor* is a normally distributed random variable with mean zero and variance one. This model could be replaced by any other noise model in practice.
- If $d_{ij} >$ *radius*, then $\widehat{d}_{ij} = 0$, and the bound $r_{ij} = 1.001 *$ *radius* is used in the SDP model.

In the simulation, we define the average estimation error to be $\frac{1}{s}\sum_{i=1}^{s} \|\bar{x}_i - x_i\|$, where $\bar{x}_i$ is from the SDP solution and x_i is the ith node's true location. In a practical setting, we wouldn't know the node's true location x_i. Instead, we would use the node's trace τ_i (2.5) to gauge the estimation error.

To convey the distribution of estimation errors and trace, we also give the 95%

quartile.

Factors such as noise level, radio range, and anchor densities can directly impact localization accuracy. The sensors' estimated locations are derived directly from the given distance measurements. If the noise level in these measurements is high, the estimation accuracy cannot be high. We also need sufficiently large radio range to achieve accurate localization, because too small a range could cause many sensors to have low connectivity or even be unreachable. Finally, more anchors in the network should help with the estimation accuracy because there are more reference points.

In the following subsections, we present simulation results (most results averaged over 10 runs) to show the accuracy and scalability of the SpaseLoc algorithm. We observe the impact of various radio ranges, anchor densities, and noise levels on the accuracy and performance of the algorithm. Computations were performed on a laptop computer with 2.4 GHz CPU and 1GB system memory, using MATLAB 6.5 [16] for SpaseLoc and a Mex interface to DSDP5.0 (Benson, Ye, and Zhang [1]) for the SDP solutions.

4.1. Effect of inequality constraints in SDP relaxation model. As we discussed in section 3.7, because of the $Y \succeq X^T X$ constraint relaxation, enforcing the r_{ij}^2 and r_{ik}^2 inequality constraints in (2.4) is not equivalent to enforcing them in the distance model (2.2). In order to observe the effectiveness of including these inequality constraints, we conduct simulations with the following three strategies, according to the number of times we check for violated inequality constraints and then include them to obtain a new solution.

- I0. This corresponds to solving the SDP problem with equality constraints only. (No inequality constraints are ever added.) The final solution is optimal for problem (2.4) without the inequality constraints involving r_{ij}^2 and r_{ik}^2.
- I1. This corresponds to solving the SDP problem with all equalities (and no inequalities) first, and then adding violated inequality constraints and resolving it at most once.
- I2. This corresponds to solving the SDP problem with all equalities (and no inequalities) first, and then adding violated inequality constraints and resolving one or more times until all inequalities are satisfied. The final solution is an optimal solution to problem (2.4).

Our experimental results show that the added inequality constraints do not always provide better localization accuracy, but can greatly increase the execution time. In this section, we illustrate the inequality constraints' impact through two simulation examples: one with no noise but low connectivity; the other with full connectivity but with noise.

In our first example, we run simulation results on a network of 100 randomly uniform-distributed sensors with *radius* 0.2275 and 10 randomly selected anchors. One of the sensors happens to be connected to only two other nodes. The sensors are localized with the full SDP and with SpaseLoc, using each of the I0, I1, I2 strategies in turn. In addition, we examine each case with or without our geometric routines for SpaseLoc. The results are shown in Figure 4.1 and Table 4.1.

Figure 4.1 shows there is a sensor connected to only 2 anchors. For full SDP shown in (a), no violated inequalities are ever found, so full SDP with I0, I1, or I2 has only one SDP call and always generates the same results. For SpaseLoc in (b) with I0 and no geometric routine, SDP is called 47 times (with no subsequent check for violated constraints). It produces the similar estimation accuracy as the full SDP approach but with much improved performance. In (c), SpaseLoc with I1 or

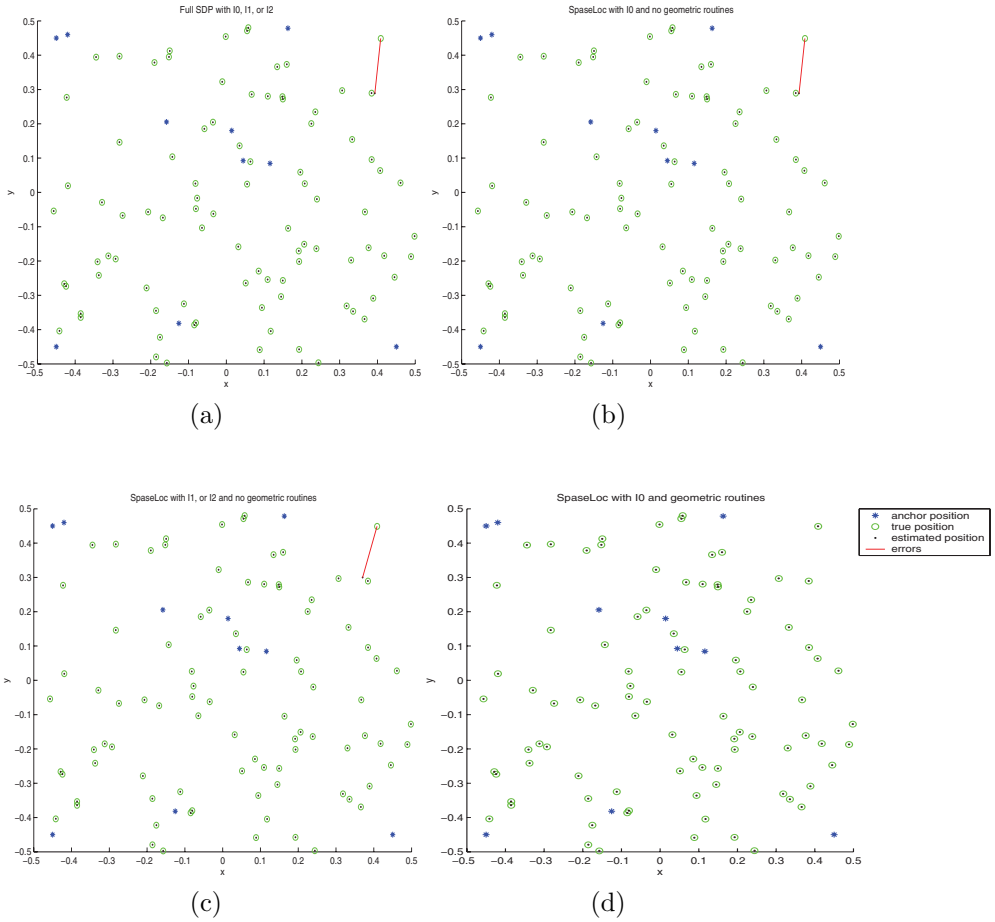


FIG. 4.1. *Inequality impact on accuracy:* 100 *nodes,* 10 *anchors, no noise, radius* 0.2275.

TABLE 4.1
Inequality impact on accuracy and speed: 100 *nodes,* 10 *anchors, no noise, radius* 0.2275.

Methods	Error	95% Error	Time	SDP's
Full SDP with I0 or I1 or I2	1.7877e-3	1.7483e-10	11.97	1
SpaseLoc with I0 and no geometric routines	1.7890e-3	1.1684e-7	0.38	47
SpaseLoc with I1 or I2 and no geometric routines	1.7134e-3	1.1684e-7	0.42	48
SpaseLoc with I0 and geometric routines	1.4679e-7	1.1523e-7	0.35	46

I2 produces the same results, which means violated inequalities are found only once. Comparing (b) and (c), we see that including violated inequalities does improve the estimation accuracy a little. Best of all in (d), SpaseLoc with I0 and our geometric routines localizes all sensors with virtually no error.

Table 4.1 shows that adding violated inequalities increases execution time slightly for SpaseLoc.

In our second example, in order to observe the effectiveness of the inequality constraints under noise conditions, we run simulations for a network of 100 nodes whose true locations are at the vertices of an equilateral triangle grid. Ten anchors are placed at the middle grid-point of each row, and the *radius* is 0.25. A *noise_factor*

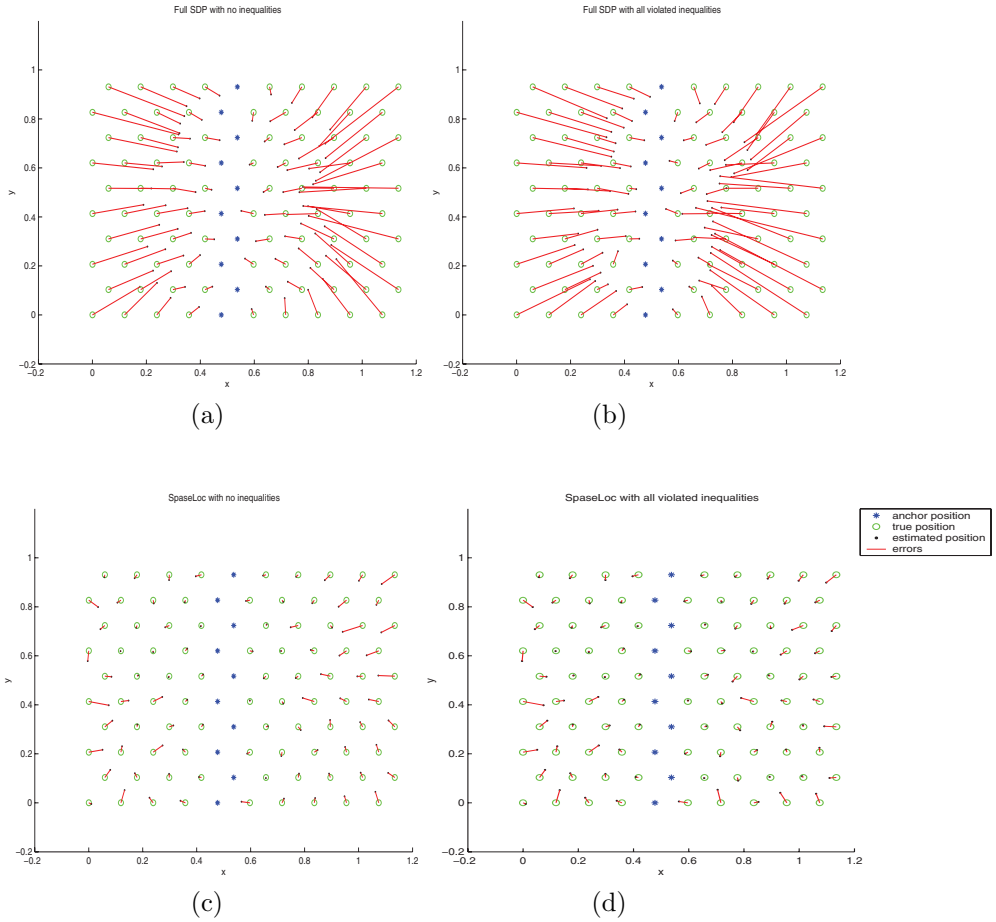


FIG. 4.2. *Inequality impact on accuracy:* 100 *nodes,* 10 *anchors, noise_factor* 0.1*, radius* 0.25.

TABLE 4.2
Inequality impact on accuracy and speed: 100 *nodes,* 10 *anchors, noise_factor* 0.1*, radius* 0.25.

Methods	Error	Time	SDP's
Full SDP with I0	0.1268	13.87	1
Full SDP with I1	0.1292	34.20	2
Full SDP with I2	0.1403	134.50	4
SpaseLoc with I0	0.0231	0.45	54
SpaseLoc with I1 or I2	0.0203	0.51	56

of 0.1 is applied to the distance measurements. The sensors are localized with either full SDP or SpaseLoc using I0, I1, I2 in turn without geometric routines. (Although we do not activate the geometric routines in this experiment, they are not a factor here because the localization error is not caused by low connectivity but by the noisy measurements.) The results are shown in Figure 4.2 and Table 4.2. Figure 4.2(b) and (d) correspond to strategy I1 or I2 for full SDP and SpaseLoc.

As we can see, adding violated inequalities for full SDP not only increases the execution times dramatically, but also increases the localization error. For SpaseLoc, adding violated inequalities improves the estimation accuracy slightly. Note that I1

and I2 produce the same results for SpaseLoc.

In summary, the first experiment shows that when the errors are caused by *low connectivity*, SpaseLoc with geometric routines and no inequality constraints (I0) outperforms SpaseLoc with inequalities (I1 or I2) and all of the full SDP options. Given this observation, from now on we only use SpaseLoc with geometric routines, which means the geometric routines are used instead of SDP to localize sensors connected to less than 3 anchors.

The second experiment indicates that under noise conditions, although adding violated inequalities does not seem to improve the estimation accuracy for full SDP, it does improve accuracy for SpaseLoc.

In the subsequent sections, we continue to examine the inequality constraints' effects on accuracy and speed.

4.2. Accuracy and speed comparison: Full SDP versus SpaseLoc. For very small networks, the SDP approach is both accurate and efficient. (This is vital to SpaseLoc, as many small subproblems must be solved using SDP.) However, the performance of the full SDP approach deteriorates rapidly with network size.

Table 4.3 shows the localization results using full SDP (a) and using SpaseLoc (b) for a range of examples with various numbers of nodes whose true locations in the network are at the vertices of an equilateral triangle grid. Anchors are placed at the middle grid-point of each row. A *noise_factor* of 0.1 is applied to the distance measurements.

Let us first look at the impact of I0, I1, and I2 on estimation accuracy. Table 4.3 (a) shows that for full SDP, 8 errors with I1 are bigger than with I2, and 5 errors with I2 are bigger than with I1. Comparing I0 with I2, we see that for each strategy, 9 errors in I0 are bigger than the errors for the other strategy. It appears that full SDP with added inequalities does not always improve the estimation accuracy. For SpaseLoc, I1 and I2 generate almost equivalent estimation accuracy; I0 has 8 errors that are bigger than with I1, while I1 has 5 errors bigger than with I0. Therefore, the added inequalities provide only marginal accuracy improvement for SpaseLoc.

Now let us compare full SDP with SpaseLoc. Figures 4.3–4.4 plot results for full SDP with I0 and SpaseLoc with I0 for two of these examples: 9 and 49 nodes, including 3 and 7 anchors placed at the grid-point in the middle of each row. As we can see from these two figures and Table 4.3, for localizing 4 and 9 nodes, full SDP and SpaseLoc show comparable performance. Beyond that size, the contrast grows rapidly. For localizing 49 nodes, SpaseLoc is 10 times faster than the full SDP method, with more than four times the accuracy. For 400 nodes, SpaseLoc with strategies I0, I1, and I2 is, respectively, 800, 2500, and 8500 times faster than full SDP with the same strategies, while achieving 10 times greater accuracy. Thus, the full SDP model becomes less effective as problem size increases. In fact, for problem sizes above 49 nodes, the average estimation error using full SDP becomes so large that the computed solution is of little value.

It may seem nonintuitive that SpaseLoc's greedy approach could produce smaller errors than the full SDP method. However, all of the SDP problems and subproblems of the form (2.4) are *relaxations* of Euclidean models of the form (2.2). As we discussed in section 3.9, SpaseLoc always tries to create a subproblem whose subsensors have three independent anchor connections, so that the SDP solution is exact. The same conclusion cannot be drawn under noise conditions, but experimentally the relaxations under noise conditions appear to be *tighter* in SpaseLoc's subproblems than in the single large SDP.

TABLE 4.3
Accuracy and speed comparison between full SDP and SpaseLoc.

(a) Full SDP

Number of nodes	Radio range	Error			Time (sec)			SDP calls		
		I0	I1	I2	I0	I1	I2	I0	I1	I2
4	2.24	0.0317	0.0317	0.0317	0.01	0.01	0.01	1	1	1
9	1.12	0.1267	0.1203	0.1203	0.02	0.05	0.05	1	2	2
16	0.75	0.0837	0.0703	0.0680	0.10	0.21	0.35	1	2	3
25	0.56	0.0938	0.1170	0.1170	0.37	0.80	1.26	1	2	3
36	0.45	0.0719	0.0618	0.0561	0.81	1.88	3.02	1	2	3
49	0.40	0.1190	0.1190	0.1190	2.10	5.33	5.33	1	2	2
64	0.40	0.1218	0.0919	0.0954	3.43	9.21	21.60	1	2	4
81	0.40	0.1380	0.0894	0.0885	7.26	19.66	59.05	1	2	5
100	0.25	0.1268	0.1292	0.1403	13.87	34.20	140.26	1	2	4
121	0.40	0.1157	0.1088	0.1091	23.24	81.62	182.74	1	2	3
144	0.21	0.1480	0.1899	0.1891	37.76	168.43	584.23	1	2	4
169	0.40	0.1283	0.1141	0.1217	71.87	278.72	692.12	1	2	4
196	0.18	0.1404	0.1275	0.1286	151.52	461.97	1081.35	1	2	4
225	0.40	0.1568	0.1589	0.1571	232.31	752.75	2408.67	1	2	5
256	0.15	0.1429	0.1375	0.1370	356.86	1089.52	3260.33	1	2	5
324	0.14	0.1685	0.1685	0.1685	962.66	2620.20	2620.20	1	2	2
361	0.13	0.1734	0.1842	0.1833	1391.04	5051.05	15281.26	1	2	4
400	0.12	0.1819	0.1970	0.1968	1662.22	5950.34	20321.60	1	2	4

(b) SpaseLoc

Number of nodes	Radio range	Error			Time (sec)			SDP calls		
		I0	I1	I2	I0	I1	I2	I0	I1	I2
4	2.24	0.0317	0.0317	0.0317	0.02	0.02	0.02	1	1	1
9	1.12	0.0513	0.0513	0.0513	0.04	0.04	0.04	6	6	6
16	0.75	0.0615	0.0559	0.0559	0.06	0.19	0.09	8	9	9
25	0.56	0.0597	0.0608	0.0608	0.13	0.13	0.13	12	13	13
36	0.45	0.0364	0.0294	0.0294	0.17	0.20	0.20	20	23	23
49	0.40	0.0252	0.0252	0.0252	0.21	0.21	0.21	26	26	26
64	0.40	0.0272	0.0273	0.0273	0.30	0.34	0.34	38	42	42
81	0.40	0.0286	0.0295	0.0295	0.37	0.41	0.41	49	53	53
100	0.25	0.0232	0.0203	0.0203	0.46	0.49	0.49	54	56	56
121	0.40	0.0238	0.0227	0.0227	0.57	0.61	0.61	74	77	77
144	0.21	0.0230	0.0237	0.0237	0.69	0.70	0.70	84	89	89
169	0.40	0.0200	0.0190	0.0190	0.80	0.84	0.84	100	106	106
196	0.18	0.0177	0.0177	0.0177	0.98	1.08	1.08	84	90	90
225	0.40	0.0226	0.0207	0.0208	1.07	1.41	1.47	94	109	110
256	0.15	0.0208	0.0235	0.0249	1.21	1.44	1.50	118	131	132
324	0.14	0.0179	0.0178	0.0178	1.64	1.70	1.70	157	158	158
361	0.13	0.0218	0.0217	0.0217	1.89	2.01	2.01	177	181	181
400	0.12	0.0176	0.0175	0.0175	2.02	2.37	2.37	184	201	201

In the following sections, we run more simulations only with SpaseLoc.

4.3. Scalability. Table 4.4 shows simulation results for 49 to 10000 randomly uniform-distributed sensors being localized using SpaseLoc with strategies I0, I1, and I2. The node numbers 49, 100, 225, ... are squares k^2, and the *radius* is the minimum value that permits localization on a regular $k \times k$ grid. The number of anchors changes with the number of sensors and is chosen to be k. Noise is not included in this simulation. When the items under I1, I2 are empty, it means that they are equal to the values under I0 in the same row.

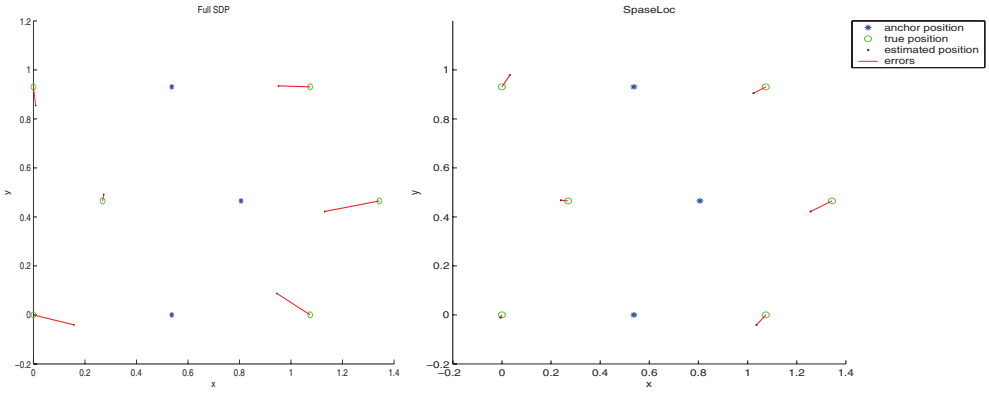


FIG. 4.3. 9 *nodes on equilateral-triangle grids,* 3 *anchors,* 0.1 *noise, radius* 1.12.

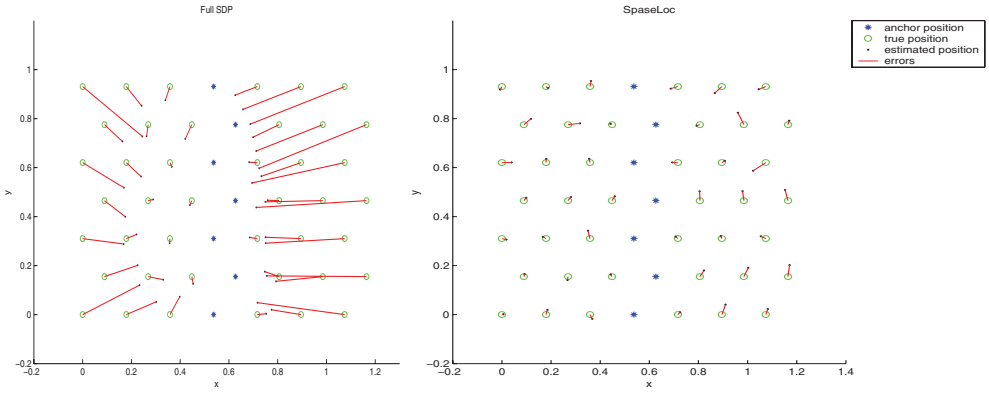


FIG. 4.4. 49 *nodes on equilateral-triangle grids,* 7 *anchors,* 0.1 *noise, radius* 0.40.

TABLE 4.4
SpaseLoc scalability. Strategies I1 and I2 generate same results.

Nodes	An_ chors	Ra_ dius	Sub- size	Error		95% Error	Time		SDP's	
				I0	I1,I2	I0,I1,I2	I0	I1,I2	I0	I1,I2
49	7	0.3412	3	4.5840e-8		3.4449e-8	0.18		18	
100	10	0.2275	3	1.4679e-7		1.1523e-7	0.35		46	
225	15	0.1462	3	4.4940e-7		3.1248e-7	0.82		112	
529	23	0.0931	3	2.1662e-6		8.9873e-7	2.02		278	
1089	33	0.0620	3	1.1969e-4		7.1510e-5	4.48		587	
2025	45	0.0451	4	1.4917e-4	1.4115e-4	9.6639e-5	8.85	9.28	1006	1007
3969	63	0.0334	4	1.2399e-4		7.2414e-5	18.79		1867	
5041	71	0.0319	6	1.5172e-4		1.1918e-4	27.19		2210	
6084	78	0.0290	6	1.7126e-4		1.1475e-4	33.66		2742	
7056	84	0.0269	7	5.2369e-5		4.0388e-5	40.59		3117	
8100	90	0.0251	7	2.7376e-4	2.7353e-4	1.7071e-4	47.87	49.71	3564	3566
9025	95	0.0238	7	2.1141e-4	2.1977e-4	1.6039e-4	54.41	56.03	3957	3958
10000	100	0.0226	7	2.0269e-4		1.5836e-4	59.33		4452	

We find that strategies I1 and I2 produce the same results, and I0 gives essentially the same. This is because the inaccuracy of the estimation is caused purely by low connectivity, not by noisy distance measurements. Empirically we see that the program scales well: almost linearly in the number of nodes in the network. Indeed, the computational complexity of the SpaseLoc algorithm is of order n, the number of

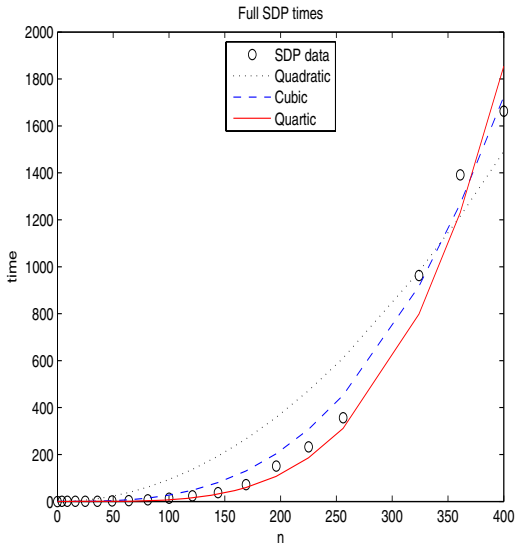


FIG. 4.5. *SDP computational complexity.*

sensors in the network, even though the full SDP approach has much greater complexity, as we now show.

We know that in the full SDP model (2.4), the number of constraints is $O(n^2)$, and in each iteration of its interior-point algorithm the SDP solver needs to solve a *sparse* linear system of equations whose dimension is the number of constraints. Figure 4.5 plots the CPU time for strategy I0 from Table 4.3(a) as well as three curves of the form $time = a_p n^p$ for $p = 2, 3, 4$, where a_p is determined by a least-squares fit. It appears that the SDP complexity with strategy I0 lies somewhere between $O(n^3)$ and $O(n^4)$.

In SpaseLoc, we partition the full problem into p subproblems of size q or less, where $p \times q = n$. We generally set q to be much smaller than n, ranging from 2 to around 10 in most of our simulations. If t represents the execution time taken by the full SDP method for a 10-node network, in the worst case the computation time for SpaseLoc is $t \times O(p)$. Thus, SpaseLoc is really linear in p in theory. Since we can assume q to be a parameter ranging from 2 to 10, with worst case 2, we know that $O(p) = O(n/q) \leq O(n/2) = O(n)$. Now we can see that SpaseLoc's computation time is $O(n)$.

In the remaining subsections we choose the middle network size from Table 4.4 (nodes = 3969) to observe the effect of varying radio range, number of anchors, and noise.

4.4. Radio range impact. With a fixed total number of randomly uniform-distributed nodes (3969, of which 63 are anchors), Table 4.5 shows the direct impact of *radius* in the range 0.0304 to 0.0334 on accuracy and performance.

Strategies I1 and I2 produce essentially the same results, and with slightly better accuracy than I0 for 8 of the 16 *radius* values, while I0 produces slightly better accuracy than I1 or I2 in 4 cases. However, I1 and I2 take more time than I0 because they need more SDP calls.

TABLE 4.5
Radio range impact: nodes = 3969, anchors = 63, no noise, sub_size = 5.

radius	Error			95% Error			Time			SDP's		
	I0	I1	I2	I0	I1	I2	I0	I1	I2	I0	I1	I2
0.0304	2.444e-3	2.359e-3	2.359e-3	4.035e-4	5.757e-4	5.757e-4	18.03	18.60	18.60	1743	1688	1689
0.0306	1.122e-3	1.123e-3	1.123e-3	5.638e-4	5.644e-4	5.644e-4	18.13	18.70	18.70	1747	1749	1749
0.0308	2.460e-3	1.412e-3	1.412e-3	7.952e-4	6.039e-4	6.039e-4	18.64	19.39	19.51	1879	1895	1896
0.0310	1.087e-3	1.083e-3	1.083e-3	4.424e-4	4.397e-4	4.397e-4	18.39	19.05	19.05	1809	1814	1814
0.0312	2.480e-3	2.481e-3	2.481e-3	3.142e-4	3.146e-4	3.146e-4	18.30	18.93	18.93	1715	1717	1717
0.0314	5.464e-4	5.337e-4	5.337e-4	2.612e-4	2.612e-4	2.612e-4	18.90	19.53	19.53	1897	1900	1900
0.0316	4.828e-4	4.827e-4	4.827e-4	2.645e-4	2.645e-4	2.645e-4	18.95	19.54	19.54	1916	1917	1917
0.0318	3.018e-4	3.013e-4	3.013e-4	1.955e-4	1.955e-4	1.955e-4	19.06	19.65	19.65	1911	1913	1913
0.0320	4.214e-4	4.214e-4	4.214e-4	1.781e-4	1.781e-4	1.781e-4	18.91	19.49	19.49	1847	1848	1848
0.0322	2.842e-4	2.842e-4	2.842e-4	1.702e-4	1.702e-4	1.702e-4	18.89	19.45	19.45	1894	1895	1895
0.0324	5.213e-4	5.495e-4	5.495e-4	2.968e-4	3.020e-4	3.020e-4	18.91	19.58	19.71	1859	1865	1866
0.0326	4.091e-4	4.033e-4	4.033e-4	2.323e-4	2.315e-4	2.315e-4	18.96	19.51	19.51	1890	1893	1893
0.0328	2.299e-4	2.289e-4	2.289e-4	1.363e-4	1.363e-4	1.363e-4	18.87	19.46	19.46	1921	1922	1922
0.0330	2.057e-4	2.160e-4	2.161e-4	9.435e-5	9.450e-5	9.450e-5	18.83	19.52	19.64	1873	1875	1876
0.0332	6.192e-4	6.439e-4	6.439e-4	3.557e-4	3.557e-4	3.557e-4	19.37	20.04	20.04	1849	1853	1853
0.0334	1.240e-4	1.240e-4	1.240e-4	7.241e-5	7.241e-5	7.241e-5	18.79	18.79	18.79	1867	1867	1867

TABLE 4.6
Number of anchors impact: nodes = 3969, radius = 0.0334, no noise, sub_size = 5.

Anchors	Error		95% Error		Time		SDP's	
	I0	I1,I2	I0	I1,I2	I0	I1,I2	I0	I1,I2
40	1.052e-3	1.052e-3	8.408e-4	8.409e-4	19.30	19.88	1906	1908
50	1.109e-3	1.128e-3	7.748e-4	7.745e-4	19.38	20.15	1861	1865
100	8.782e-4	7.280e-4	5.337e-4	5.115e-4	19.16	19.96	1870	1872
150	2.716e-4	2.717e-4	1.025e-4	1.025e-4	18.86	19.60	1806	1808
200	4.889e-5	4.872e-5	1.473e-5	1.473e-5	18.77	19.52	1795	1796
250	1.716e-5	1.699e-5	7.760e-6	7.760e-6	18.55	19.32	1748	1749
300	1.538e-5	1.521e-5	4.408e-6	4.408e-6	18.20	18.99	1750	1751
350	7.533e-6	7.365e-6	2.858e-6	2.858e-6	18.13	18.93	1684	1685
400	6.383e-6	6.215e-6	1.841e-6	1.841e-6	18.16	18.96	1560	1591

As we see, increasing *radius* leads to increased accuracy and only slightly more computational time. The simulation could assist sensor network designers in selecting a radio range to achieve a desired estimation accuracy with little concern about algorithm speed.

4.5. Number of anchors impact. With constant *radius* (0.0334) and the same randomly distributed nodes (3969), Table 4.6 shows the impact of the number of anchors, ranging from 1% to 10% of the total number of points. (Noise is not included.)

Strategies I1 and I2 produce identical results. Comparing I0 with I1 or I2, we see that added inequalities slightly improve the average error consistently, although the 95% error remains essentially the same. Increasing the number of anchors in the network improves the estimation accuracy in general, with no obvious impact on algorithm speed. However, we don't see accuracy improvement when the number of anchors reaches more than 10% of the total points. This analysis is beneficial for designers to avoid the cost of deploying unnecessary anchors.

4.6. Noise impact. With constant *radius* (0.0334) and the same randomly distributed nodes (3969), Table 4.7 shows the impact of noise conditions on accuracy and performance.

We see that strategies I1 and I2 do not provide consistent improvement over I0 for both average and 95% error, yet they always increase execution time. Also, more noise in the network has a direct impact on estimation accuracy. Simulations of this kind may help designers determine the measurement noise level that will give an acceptable estimation error.

TABLE 4.7
Noise_factor impact: nodes = 3969, *anchors* = 400, *radius* = 0.0334, *sub_size* = 5.

Noise	Error			95% Error			Time			SDP's		
factor	I0	I1	I2	I0	I1	I2	I0	I1	I2	I0	I1	I2
0.01	9.60e-4	9.59e-4	9.59e-4	1.90e-6	3.16e-4	3.16e-4	20.38	21.18	21.20	1622	1623	1623
0.05	3.15e-3	8.33e-3	8.33e-3	3.16e-4	3.57e-3	3.57e-3	21.56	21.75	21.86	1479	1253	1256
0.10	6.87e-3	7.36e-3	1.03e-2	1.91e-3	5.17e-3	6.95e-3	21.46	24.73	22.94	1447	1592	1230
0.20	1.55e-2	1.57e-2	1.65e-2	4.95e-3	1.16e-2	1.25e-2	21.55	25.67	25.83	1208	1433	1390
0.30	1.51e-2	1.48e-2	1.46e-2	1.17e-2	1.27e-2	1.24e-2	21.09	29.76	31.78	1411	1829	1844
0.40	1.98e-2	1.79e-2	1.79e-2	1.32e-2	1.57e-2	1.57e-2	21.30	32.05	35.15	1523	1985	2073
0.50	3.05e-2	2.35e-2	2.28e-2	2.86e-2	2.16e-2	2.08e-2	22.07	35.27	39.26	1608	2157	2252

5. Summary and extensions. We have shown that SpaseLoc achieves the aims of accuracy, speed, and scalability with a single processor on very large networks. It takes full advantage of the recent SDP approach of Biswas and Ye [2]. The latter has computational complexity $O(n^p)$, where n is the network size and p is between 3 and 4, but we use it on multiple tiny subproblems to obtain an algorithm with essentially linear complexity. On a 2.4GHz laptop with 1GB memory, SpaseLoc maintains efficiency and provides accurate location estimation for networks with 10000 sensors and beyond.

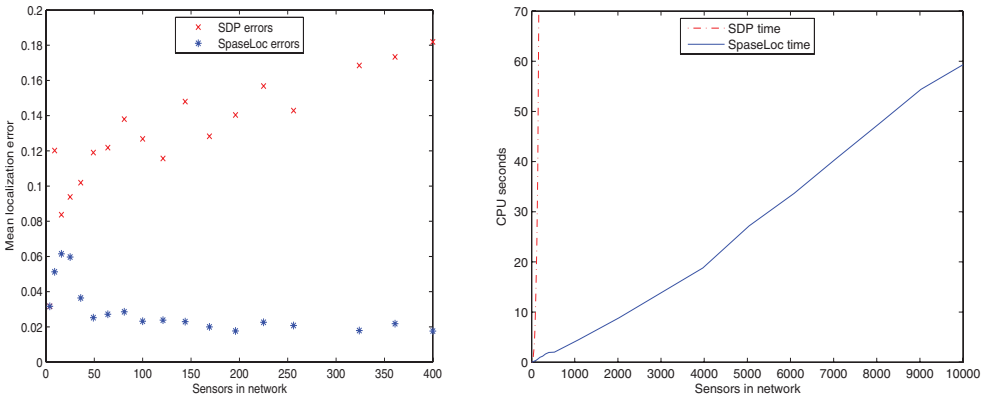


FIG. 5.1. *Accuracy and performance comparison.*

Figure 5.1 compares localization results for our SpaseLoc algorithm and the full SDP approach [2] for various sized networks. The left-hand figure shows a comparison in terms of estimation accuracy for localizing various sizes of networks when sensors are placed at the vertices of an equilateral triangle grid with 0.1 *noise_factor* added to distance measurements (data is taken from Table 4.3). It shows clearly that SpaseLoc provides much improved localization accuracy.

The right-hand graph summarizes results in terms of execution time on various network sizes. Data for the full SDP method is taken from Table 4.3, and data for SpaseLoc is taken from Table 4.4. The figure confirms near-linear complexity for SpaseLoc.

5.1. More general problems. In Jin [14], SpaseLoc is used as a building block for more general localization algorithms. A dynamic version can estimate moving sensors' locations in real time, and a three-dimensional version extends its utility further. For clustered and distributed environments, it is shown how to use SpaseLoc

in parallel (on multiple large subproblems) to obtain essentially linear complexity on clustered networks of unlimited size. Finally, a preprocessor for SpaseLoc has been developed in [14] to localize sensors in anchorless networks.

5.2. A bootstrap procedure. SpaseLoc works effectively when Step A1 (subproblem creation) finds subsensors connected to at least 3 anchors. A difficult situation arises if there are more than 3 anchors in the network but no subsensor is directly connected to 3 anchors. A network with anchors placed at the borders of the region is such an example. SpaseLoc's subproblems will involve sensors connected to only 2 or 1 anchors, leading to a less accurate final solution.

When there is sufficient connection information for sensors to be indirectly connected to at least 3 anchors through other sensors, the full SDP approach can find a solution. We are developing a procedure to choose a subproblem in the above situation. It will include the anchors, certain subsensors, and the sensors on each shortest path from a subsensor to an anchor.

5.3. Alternative subproblem solvers. At present, most of the SpaseLoc subproblems are solved by the SDP approach of Biswas and Ye [2]. This is an approximation method that may produce large errors with noisy data. A recent development by Biswas et al. [4] adds regularization terms to the SDP problem and uses a gradient-descent method to refine the SDP solution. Significant accuracy improvement is reported. An advantage of SpaseLoc is that it can solve each subproblem by any method that is effective on small networks. Our next step is to experiment with such approaches, including that of [4] and various triangulation-based methods.

Acknowledgments. We gratefully acknowledge valuable technical advice from Profs. Henry Wolkowicz, Scott Rogers, and Daniel Frances and two perceptive referees. Thanks also to Dr. Steve Benson for his expert advice on using the DSDP5.0 solver and to Prof. Kenneth Holmström for his help in fine-tuning parts of the MATLAB implementation of SpaseLoc. We are further indebted to Robert Bosch Corporation and the talented Bosch RTC team: Sharmila Ravula, Lakshmi Venkatraman, Bhaskar Srinivasan, Abtin Keshavarzian, Hauke Schmidt, and Karsten Funk; they provided the inspiration and vision for our algorithms to be practical.

REFERENCES

[1] S. J. Benson, Y. Ye, and X. Zhang, *DSDP Website*, http://www-unix.mcs.anl.gov/~benson or http://www.stanford.edu/~yyye/Col.html (1998–2005).

[2] P. Biswas and Y. Ye, *Semidefinite programming for ad hoc wireless sensor network localization*, in Proceedings of the Third International Symposium on Information Processing in Sensor Networks, Berkeley, CA, 2004.

[3] P. Biswas and Y. Ye, *A distributed method for solving semidefinite programs arising from ad hoc wireless sensor network localization*, in Multiscale Optimization Methods and Applications, Nonconvex Optim. Appl. 82, Springer, New York, 2006.

[4] P. Biswas, T. Liang, K. Toh, T. Wang, and Y. Ye, *Semidefinite programming approaches for sensor network localization with noisy distance measurements*, IEEE Trans. Automation Sci. Engrg., to appear.

[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] N. Bulusu, J. Heidemann, and D. Estrin, *GPS-less Low Cost Outdoor Localization for Very Small Devices*, Technical report 00-729, Computer Science Department, University of Southern California, Los Angeles, CA, 2000.

[7] D. Culler and W. Hong, *Wireless sensor networks*, Comm. ACM, 47 (2004), pp. 30–33.

[8] L. Doherty, L. El Ghaoui, and K. Pister, *Convex position estimation in wireless sensor networks*, in Proceedings of the IEEE Infocom 2001, Anchorage, AK, 2001, pp. 1655–1663.

[9] A. Dragoon, *Small wonders*, CIO Magazine, January 15, 2005.

[10] D. Ganesan, B. Krishnamachari, A. Woo, D. Culler, D. Estrin, and S. Wicker, *An Empirical Study of Epidemic Algorithms in Large-Scale Multihop Wireless Networks*, Report UCLA/CSD-TR-02-0013, Computer Science Department, University of California at Los Angeles, Los Angeles, CA, 2002.
[11] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, 1996.
[12] J. Hightower and G. Boriello, *Location systems for ubiquitous computing*, IEEE Computer, 34 (2001), pp. 57–66.
[13] A. Howard, M. J. Mataric, and G. S. Sukhatme, *Relaxation on a mesh: A formalism for generalized localization*, in Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS01), Maui, HI, 2001, pp. 1055–1060.
[14] H. H. Jin, *Scalable Sensor Localization Algorithms for Wireless Sensor Networks*, Ph.D. thesis, University of Toronto, Toronto, Canada, 2005. (Joint research conducted at Stanford University.)
[15] M. Lawlor, *Small systems, big business*, Signal Magazine, January 2005.
[16] MATLAB 6.5, Release 13 with Service Pack 1, The MathWorks, Inc., Natick, MA, 2003.
[17] D. Niculescu and B. Nath, *Ad hoc positioning system*, in Proceedings of the IEEE GlobeCom 2001, San Antonio, TX, 2001, pp. 2926–2931.
[18] A. Ricadela, *Sensors everywhere*, Information Week, January 24, 2005.
[19] C. Savarese, J. Rabaey, and K. Langendoen, *Robust positioning algorithm for distributed ad hoc wireless sensor networks*, in Proceedings of the USENIX Technical Annual Conference, Monterey, CA, 2002.
[20] A. Savvides, C.-C. Han, and M. B. Srivastava, *Dynamic fine-grained localization in ad hoc networks of sensors*, in Proceedings of the ACM/IEEE International Conference on Mobile Computing and Networking (MOBICON), Rome, Italy, 2001, pp. 166–179.
[21] A. Savvides, H. Park, and M. B. Srivastava, *The bits and flops of the n-hop multilateration primitive for node localization problems*, in Proceedings of the 1st ACM International Workshop on Wireless Sensor Networks and Applications (WSNA'02), Atlanta, GA, ACM Press, New York, 2002, pp. 112–121.
[22] Y. Shang, W. Ruml, Y. Zhang, and M. Fromherz, *Localization from mere connectivity*, in Proceedings of the Fourth ACM International Symposium on Mobile Ad Hoc Networking and Computing, Annapolis, MD, ACM Press, New York, 2003, pp. 201–212.
[23] J. F. Sturm, *Let SeDuMi Seduce You*, http://fewcal.kub.nl/sturm/software/sedumi.html (October 2001).
[24] R. Szewczyk, E. Osterweil, J. Polastre, M. Hamilton, A. Mainwaring, and D. Estrin, *Habitat monitoring with sensor networks*, Comm. ACM, 47 (2004), pp. 34–44.
[25] P. Tseng, *SOCP relaxation for nonconvex optimization*, presented at ICCOPT 1, Rensselaer Polytechnic Institute, Troy, NY, 2004.

SIAM J. OPTIM.
Vol. 17, No. 4, pp. 1129–1153
© 2006 Society for Industrial and Applied Mathematics

INTERIOR POINT TRAJECTORIES AND A HOMOGENEOUS MODEL FOR NONLINEAR COMPLEMENTARITY PROBLEMS OVER SYMMETRIC CONES*

AKIKO YOSHISE†

Abstract. We study the continuous trajectories for solving monotone nonlinear mixed complementarity problems over symmetric cones. While the analysis in [L. Faybusovich, *Positivity*, 1 (1997), pp. 331–357] depends on the optimization theory of convex log-barrier functions, our approach is based on the paper of Monteiro and Pang [*Math. Oper. Res.*, 23 (1998), pp. 39–60], where a vast set of conclusions concerning continuous trajectories is shown for monotone complementarity problems over the cone of symmetric positive semidefinite matrices. As an application of the results, we propose a homogeneous model for standard monotone nonlinear complementarity problems over symmetric cones and discuss its theoretical aspects. Consequently, we show the existence of a path having the following properties: (a) The path is bounded and has a trivial starting point without any regularity assumption concerning the existence of feasible or strictly feasible solutions. (b) Any accumulation point of the path is a solution of the homogeneous model. (c) If the original problem is solvable, then every accumulation point of the path gives us a finite solution. (d) If the original problem is strongly infeasible, then, under the assumption of Lipschitz continuity, any accumulation point of the path gives us a finite certificate proving infeasibility.

Key words. complementarity problem, symmetric cone, homogeneous algorithm, existence of trajectory, interior point method, detecting infeasibility

AMS subject classifications. 90C22, 90C25, 90C33, 65K05, 46N10

DOI. 10.1137/04061427X

1. Introduction.

Let $(V, \circ)$ be a Euclidean Jordan algebra with an identity element e. We denote by K the symmetric cone of V which is a self-dual closed convex cone such that for any two elements $x \in \mathrm{int}K$ and $y \in \mathrm{int}K$, there exists an invertible map $\Gamma : V \to V$ satisfying $\Gamma(K) = K$ and $\Gamma(x) = y$. It is known that a cone in V is symmetric if and only if it is the cone of squares of V given by $K = \{x \circ x : \ x \in V\}$.

Faybusovich [6] studied the linear monotone complementarity problem (LCP) over symmetric cones of the form

$$\begin{array}{lll} \text{(LCP)} & \text{find} & (x, y) \in K \times K \\ & \text{s.t.} & (x, y) \in (a, b) + L, \ x \circ y = 0, \end{array} \tag{1.1}$$

where $(a, b) \in V \times V$ and $L \subseteq K \times K$ is a linear subspace with $\dim L = \dim V$ having the monotone property, i.e., $\langle x, y \rangle \geq 0$ if $(x, y) \in L$. The author showed the existence of the central path of the form

$$\{(x, y) \in \mathrm{int}K \times \mathrm{int}K : \ x \circ y = \mu e, \ \mu > 0\}$$

whenever the LCP has an interior feasible solution $(x, y) \in ((a, b)+L) \cap (\mathrm{int}K \times \mathrm{int}K)$, based on primal-dual interior point methods for linear programs [14, 17].

*Received by the editor September 1, 2004; accepted for publication (in revised form) June 7, 2006; published electronically December 5, 2006. This research was supported in part by Grant-in-Aid for Scientific Research ((C)1856052, (C)17560050, (B)18310101), and for Exploratory Research 18651076 of the Ministry of Education, Culture, Sports, Science and Technology of Japan.
http://www.siam.org/journals/siopt/17-4/61427.html
†Graduate School of Systems and Information Engineering, The University of Tsukuba, Tsukuba, Ibaraki 305-8573, Japan (yoshise@sk.tsukuba.ac.jp).

Note that the first extension of primal-dual methods to a more general setting than linear programs was achieved by Nesterov and Todd [21, 22] who developed the powerful theoretical concept of self-scaled barrier functions. It is known that the self-scaled cones associated with the self-scaled barriers are closely related to the symmetric cones [1, 10, 11, 27]. See also [15, 7, 23, 28, 25, 24, 3] for other extensions of primal-dual methods to the positive semidefinite cones, the symmetric cones, the self-scaled cones, or the homogeneous cones.

In most of the papers cited above, the analyses depend on the optimization theory of convex barrier functions. In this paper, apart from the theories of barrier functions, the trajectory of an interior point map is discussed in view of homeomorphisms of continuous maps. There have been studies of various types of central paths using the theory of homeomorphisms for some special cases of symmetric cones, i.e., the non-negative orthant and the cone of symmetric positive semidefinite matrices (see [9, 13, 29, 20]). We extend these results and set out basic properties concerning the existence of central paths for the monotone nonlinear complementarity problems over symmetric cones. Consequently, a first analysis of the trajectory for complementarity problems over the second order cone is provided.

As an application of the results, we give a homogeneous model for solving the problems. The homogeneous model for monotone complementarity problems was first proposed by Andersen and Ye [2] for solving the problems over the n-dimensional positive orthant. Unlike for linear optimization cases, the model is given by a nonlinear system even if the original problems are linear. However, the path associated with the model has some remarkable features as described below:

(a) The path exists, is bounded, and has a trivial starting point without any regularity assumption concerning the existence of feasible or strictly feasible solutions.
(b) Any accumulation point of the path is a solution of the homogeneous model.
(c) If the original problem is solvable, then every accumulation point of the path gives us a finite certificate proving feasibility.
(d) If the original problem is strongly infeasible, then, under the assumption of (scaled) Lipschitz continuity, every accumulation point of the path gives us a finite certificate proving infeasibility.

We show that a path having the above properties also exists for the monotone complementarity problems over symmetric cones.

Consider the following nonlinear and mixed complementarity problem:

$$\begin{array}{lll} \text{(CP)} & \text{Find} & (x, y, z) \in K \times K \times \Re^m \\ & \text{s.t.} & F(x, y, z) = 0,\ x \circ y = 0, \end{array} \tag{1.2}$$

where $F : K \times K \times \Re^m \to V \times \Re^m$ is a continuous map. Many problems can be cast into CPs having the monotone property, e.g., any primal and dual linear optimization problems over symmetric cones, and the robust Nash equilibrium problem introduced by Hayashi, Yamashita, and Fukushima [12].

The map F appearing in our homogeneous model is not necessarily defined on the boundary of the set $K \times K \times \Re^m$. Some *asymptotic* definitions are then introduced as follows:

- The CP is *asymptotically feasible* if and only if there exists a bounded sequence $\{x^{(k)}, y^{(k)}, z^{(k)}\} \subseteq \mathrm{int}K \times \mathrm{int}K \times \Re^m$ such that

$$\lim_{k\to\infty} F(x^{(k)}, y^{(k)}, z^{(k)}) = 0.$$

- The CP is *asymptotically solvable* if and only if there exists a bounded sequence $\{x^{(k)}, y^{(k)}, z^{(k)}\} \subseteq \mathrm{int}K \times \mathrm{int}K \times \Re^m$ such that

$$\lim_{k\to\infty} F(x^{(k)}, y^{(k)}, z^{(k)}) = 0 \text{ and } \lim_{k\to\infty} x^{(k)} \circ y^{(k)} = 0. \tag{1.3}$$

As long as the asymptotic property is discussed, the map F should be defined only on the set $\mathrm{int}K \times \mathrm{int}K \times \Re^m$ rather than on $K \times K \times \Re^m$.

We also introduce the following definitions to discuss the infeasibility of CPs:

- The CP is *infeasible* if and only if there is no feasible point $(x, y, z) \in K \times K \times \Re^m$ satisfying $F(x, y, z) = 0$.
- The CP is *strongly infeasible* if and only if there is no sequence $\{x^{(k)}, y^{(k)}, z^{(k)}\} \subseteq \mathrm{int}K \times \mathrm{int}K \times \Re^m$ such that $\lim_{k\to\infty} F(x^{(k)}, y^{(k)}, z^{(k)}) = 0$.

We impose the following assumption on F.

ASSUMPTION 1.1.

(i) *F is (x, y)-equilevel-monotone on its domain; i.e., if (x, y, z) and (x', y', z') lie in the domain of F and satisfy $F(x, y, z) = F(x', y', z')$, then $\langle x-x', y-y'\rangle \geq 0$ holds.*

(ii) *F is z-bounded on its domain; i.e., for any sequence $\{(x^{(k)}, y^{(k)}, z^{(k)})\}$ in the domain of F, if $\{(x^{(k)}, y^{(k)})\}$ and $\{F(x^{(k)}, y^{(k)}, z^{(k)})\}$ are bounded, then the sequence $\{z^{(k)}\}$ is also bounded.*

(iii) *$F(x, y, z)$ is z-injective on its domain; i.e., if (x, y, z) and (x, y, z') lie in the domain of F and satisfy $F(x, y, z) = F(x, y, z')$, then $z = z'$ holds.*

The above assumption is the same as the one imposed by Monteiro and Pang [19] for the case of the cone of symmetric matrices. Note that, in contrast to the paper [19], the domain of the map F is not given explicitly in the assumption. The domain is set as the set $\mathrm{int}K \times \mathrm{int}K \times \Re^m$ for observing some basic properties required in constructing a homogeneous model (section 3), and as the set $K \times K \times \Re^m$ for discussing the solvability of the CP (section 4).

The paper is organized as follows.

In section 2, some basic results for symmetric cones are summarized.

Section 3 is devoted to deriving a homeomorphism of the map $H : \mathrm{int}K \times \mathrm{int}K \times \Re^m \to V \times V \times \Re^m$ given by

$$H := \begin{pmatrix} x \circ y \\ F(x, y, z) \end{pmatrix}. \tag{1.4}$$

The main result, Theorem 3.10, ensures that if Assumption 1.1 is satisfied with the domain $\mathrm{int}K \times \mathrm{int}K \times \Re^m$, then the system $H(x, y, z) = h$ has a solution $(x, y, z) \in \mathcal{U} \times \Re^m$ for any $h \in \mathrm{int}K \times F(\mathcal{U} \times \Re^m)$, where the set $\mathcal{U}$ is a subset of $\mathrm{int}K \times \mathrm{int}K$ defined by

$$\mathcal{U} := \{(x, y) \in \mathrm{int}K \times \mathrm{int}K : \ x \circ y \in \mathrm{int}K\}.$$

Suppose that there exists a sequence $\{h^{(k)}\} \subseteq \mathrm{int}K \times F(\mathcal{U} \times \Re^m)$ satisfying $h^{(k)} \to 0$. Theorem 3.10 implies that there exists a sequence $\{(x^{(k)}, y^{(k)}, z^{(k)})\} \subseteq \mathcal{U} \times \Re^m$ that is the set of solutions of the system $H(x^{(k)}, y^{(k)}, z^{(k)}) = h^{(k)}$ for any k. It is easy to see that the sequence $\{(x^{(k)}, y^{(k)}, z^{(k)})\} \subseteq \mathcal{U} \times \Re^m$ satisfies (1.3). Thus, if $\{(x^{(k)}, y^{(k)}, z^{(k)})\}$ is bounded, then the CP is asymptotically solvable. In section 4, the asymptotic solvability of the CP is discussed under Assumption 1.1 with the domain $K \times K \times \Re^m$. The obtained results are direct extensions of the ones in [19].

In section 5, as an application of the results in section 3, a homogeneous model is provided for a special class of CPs. As we described above, a remarkable feature of the model is that the associated trajectory gives certifications on the feasibility or the (strong) infeasibility of the original problem. We show that the result can be extended for the case of Euclidean Jordan algebras. To the best of our knowledge, this is a first homogeneous model for CPs over the cone of symmetric matrices and/or the second order cone, which are special cases of the symmetric cones. Note that for LCPs over the n-dimensional positive orthant and for linear conic optimization problems, several homogeneous models including the homogeneous self-dual embedding model have been studied [32, 16, 30, 4].

Some concluding remarks are given in section 6.

2. Some key lemmas for the symmetric cone. We give a summary of the theory of Euclidean Jordan algebra. Most of the results can be found in the book of Faraut and Korányi [5] and the papers [6, 7, 8, 28].

Let $(V, \circ)$ be a Euclidean Jordan algebra with the identity element e, where $(x, y) \mapsto x \circ y : V \times V \to V$ is a bilinear map satisfying

(i) $x \circ y = y \circ x$,
(ii) $x \circ (y \circ x^2) = (x \circ y) \circ x^2$, where $x^2 = x \circ x$,
(iii) $x \circ e = e \circ x = x$

for all $x, y, z \in V$. Note that the Jordan algebra $(V, \circ)$ is called Euclidean if there exists a symmetric, positive definite quadratic form $\mathcal{Q}$ on V which is also associative, i.e., $\mathcal{Q}(x \circ y, z) = \mathcal{Q}(x, y \circ z)$ for all $x, y, z \in V$. Since $(x, y) \mapsto x \circ y$ is a bilinear map, for each $x \in V$, the linear transformation $L(x)$ is defined by $L(x)y = x \circ y$. For $x \in V$, the *degree* of x is the smallest integer d such that the set $\{e, x, x^2, \ldots, x^d\}$ is linearly independent. The *rank* r of V is the maximum of the degree of x over all $x \in V$. For any element x in V of rank r, we can define the *characteristic polynomial* of x of the form

$$p_x(\lambda) := \lambda^r - a_1(x)\lambda^{r-1} + \cdots + (-1)^r a_r(x)$$

(cf. section 2 of [28]). We call the roots $\lambda_1, \ldots, \lambda_r$ of $p_x(\lambda)$ the *eigenvalues* of x and define

$$\mathrm{tr}(x) := \sum_{i=1}^{r} \lambda_i = a_1(x), \quad \det(x) := \prod_{i=1}^{r} \lambda_i = a_r(x). \tag{2.1}$$

Since $(x, y) \mapsto x \circ y$ is bilinear and $\mathrm{tr}(x \circ y)$ is a symmetric positive definite quadratic form which is associative, i.e., $\mathrm{tr}\,(x \circ (y \circ z)) = \mathrm{tr}\,((x \circ y) \circ z)$ for all $x, y, z \in V$, we define below the canonical inner product $\langle x, y \rangle$ of $x, y \in V$ and the canonical norm of $x \in V$, which we use throughout the paper:

$$\langle x, y \rangle := \mathrm{tr}(x \circ y), \quad \|x\| := \sqrt{\mathrm{tr}(x \circ x)}. \tag{2.2}$$

Note that $\|e\| = \sqrt{r}$. The property $\mathrm{tr}\,(x \circ (y \circ z)) = \mathrm{tr}\,((x \circ y) \circ z)$ implies that for each $x \in V$, $L(x)$ is self-adjoint with respect to $\langle \cdot, \cdot \rangle$, i.e., $\langle L(x)y, z \rangle = \langle y, L(x)z \rangle$ holds for all $y, z \in V$. We use the notation $L(u) \succ 0$ to mean that $L(u)$ is positive definite.

The set of squares $K := \{x^2 : \ x \in V\}$ is the symmetric cone of V, which is self-dual (i.e., $K = K^* := \{y : \ \langle x, y \rangle \geq 0 \ \text{ for all } x \in K\}$) and has the following properties.

PROPOSITION 2.1.

(i) $\mathrm{int}K = \{u \in V : \ L(u) \succ 0\} = \{x^2 : \ x \in V, \ \det(x) \neq 0\}$.

(ii) *If $y \in \mathrm{int}K$ and $\eta > 0$, then the set $\{x \in K : \langle x, y\rangle \leq \eta\}$ is compact.*

Proof. See Theorem III.2.1 together with its proof and Proposition II.2.4 of [5] for (i) and Corollary I.1.6 of [5] for (ii). □

An *idempotent* c is an element of V such that $c^2 = c$. An idempotent is *primitive* if it is nonzero and not given by the sum of two nonzero idempotents. A complete system of orthogonal idempotents is a set $\{c_1, c_2, \ldots, c_k\}$, where

$$c_j \circ c_j = c_j, \quad c_i \circ c_j = 0 \ (i \neq j), \quad \sum_{j=1}^{k} c_j = e.$$

A complete system of orthogonal primitive idempotents is called a *Jordan frame*.

THEOREM 2.2. *Let r be the rank of V.*

(i) *If $x \in V$, then there exist real numbers $\lambda_1, \ldots, \lambda_r$ and a Jordan frame $c_1, \ldots, c_r$ such that $x = \sum_{j=1}^{r} \lambda_j c_j$. Here the numbers λ_j (with their multiplicities) are uniquely determined by x and λ_j's are the eigenvalues (multiplicities included) of x.*

(ii) *Let c be an idempotent in a Jordan algebra, $c^2 = c$. The only possible eigenvalues of $L(c)$ are 0, $\frac{1}{2}$, and 1.*

Proof. See Theorem III.1.2 of [5] for (i) and Proposition III.1.3 of [5] for (ii). □

The corollary below follows from Proposition 2.1 and the above theorem.

COROLLARY 2.3. *Let $x \in V$ and let $\sum_{j=1}^{r} \lambda_j c_j$ be a decomposition of x given by Theorem 2.2. Then*

(i) *$x \in K$ if and only if $\lambda_j \geq 0$ $(j = 1, 2, \ldots, r)$,*

(ii) *$x \in \mathrm{int}K$ if and only if $\lambda_j > 0$ $(j = 1, 2, \ldots, r)$.*

For each $x, y \in V$, define $P(x) := 2L(x)^2 - L(x^2)$ and

$$P(x, y) := \frac{1}{2}(P(x + y) - P(x) - P(y)) = L(x)L(y) + L(y)L(x) - L(x \circ y). \tag{2.3}$$

$P(x)$ is the *quadratic representation* of x and used in several characterizations of x.

PROPOSITION 2.4.

(i) *If $x, y \in V$, then $P(x)e = x^2$ and $P(P(y)x) = P(y)P(x)P(y)$.*

(ii) *If $x, y \in V$ are invertible, then $P(x)^{-1} = P(x^{-1})$ and $P(x)\mathrm{int}K = \mathrm{int}K$.*

(iii) *If $x \in \mathrm{int}K$, then $P(x)^{-1/2} = P(x^{-1/2})$.*

Proof. See Proposition II.3.1, II.3.3, and III.2.2 of [5] for (i) and (ii). Using (i) and (ii), we obtain (iii) as follows:

$$\begin{aligned}[P(x)^{-1/2}]^2 &= P(x)^{-1} = P(x^{-1}) = P(P(x^{-1/2})e) \\ &= P(x^{-1/2})P(e)P(x^{-1/2}) = [P(x^{-1/2})]^2. \end{aligned}$$ □

The following is a collection of technical facts which are often used in the succeeding sections. Before proceeding, we give a definition of the *star-shaped* set in a vector space.

DEFINITION 2.5. *A subset C of a vector space is said to be* star-shaped *if there exists $c^0 \in C$ such that the line segment connecting c^0 to any other point in C is contained entirely in C.*

LEMMA 2.6.

(i) $x \in K$ *if and only if* $\langle x, y\rangle \geq 0$ *for all* $y \in K$.

(ii) *If* $x \in K$ *and* $y \in K$, *then* $\langle x, y\rangle = 0$ *if and only if* $x \circ y = 0$.

(iii) *If* $x \in \mathrm{int}K$ *and* $y \in K$, *then* $\langle x, y\rangle = 0$ *if and only if* $y = 0$.

(iv) *Define*

$$\mathcal{U} := \{(x, y) \in \mathrm{int}K \times \mathrm{int}K : \ x \circ y \in \mathrm{int}K\}. \tag{2.4}$$

Then $\mathcal{U} = \{(x, y) \in K \times K : \ x \circ y \in \mathrm{int}K\}$.

(v) *If* $(x, y) \in \mathcal{U}$, *then* $L(x)L(y) + L(y)L(x) \succ 0$.

(vi) *If* $(x, y) \in \mathcal{U}$, $(\Delta x, \Delta y) \in V \times V$, *and*

$$\langle \Delta x, \Delta y\rangle \geq 0, \quad x \circ \Delta y + y \circ \Delta x = 0 \tag{2.5}$$

hold, then $\Delta x = \Delta y = 0$.

(vii) $\mathcal{U}$ *is a nonempty and open subset of* $\mathrm{int}K \times \mathrm{int}K$ *which is star-shaped.*

(viii) *If* $x \in \mathrm{int}K$, *then* $(x, x) \in \mathcal{U}$.

(ix) $\mathrm{int}K \times \{\alpha e : \ \alpha \in \Re_{++}\} \subseteq \mathcal{U}$, $\{\alpha e : \ \alpha \in \Re_{++}\} \times \mathrm{int}K \subseteq \mathcal{U}$,
$K \times \{\alpha e : \ \alpha \in \Re_{+}\} \subseteq \mathrm{cl}(\mathcal{U})$, $\{\alpha e : \ \alpha \in \Re_{+}\} \times K \subseteq \mathrm{cl}(\mathcal{U})$,
where $\Re_+ := \{\alpha \in \Re : \ \alpha \geq 0\}$ *and* $\Re_{++} := \{\alpha \in \Re : \ \alpha > 0\}$.

Proof. (i) Since the set K is the symmetric cone, K is self-dual, and we obtain (i).

(ii) See Lemma 2.2 of [7].

(iii) It follows from the fact that K is a self-dual convex cone with nonempty interior. See, for example, Exercise 6.22 of [26].

(iv) Let $\bar{\mathcal{U}} := \{(x, y) \in K \times K : \ x \circ y \in \mathrm{int}K\}$. It is obvious that $\mathcal{U} \subseteq \bar{\mathcal{U}}$. Suppose that $(x, y) \in \bar{\mathcal{U}}$ and $x \in K \setminus \mathrm{int}K$. Let $x = \sum_{i=1}^{r} \lambda_i e_i$ and $y = \sum_{j=1}^{r} \mu_j f_j$ be decompositions of x and y given by (i) of Theorem 2.2. Since $x \in K \setminus \mathrm{int}K$, by (i) and (ii) of Corollary 2.3, there exists an index $\bar{i}$ such that $\lambda_{\bar{i}} = 0$. Define $z := \sum_{k=1}^{r} \nu_k e_k$, where

$$\nu_k = \begin{cases} 1 & \text{if } \lambda_k = 0, \\ 0 & \text{otherwise.} \end{cases}$$

Then $x \circ z^2 = 0$ and $z \neq 0$. Thus, we see that $0 = \langle x \circ z^2, y\rangle = \langle x \circ y, z^2\rangle = \langle z, L(x \circ y)z\rangle$ for $z \neq 0$, which implies that $L(x \circ y)$ is not positive definite; a contradiction to the assumption $(x, y) \in \bar{\mathcal{U}}$. Similarly, we obtain $y \in \mathrm{int}K$.

(v) We first show that $P(x, y)$ defined by (2.3) is positive definite for any $x, y \in \mathrm{int}K$. Suppose that $x, y \in \mathrm{int}K$. Using the results of Proposition 2.4, we see that

$$\begin{aligned} P(x)^{-1/2}P(x + y)P(x)^{-1/2} &= P(x^{-1/2})P(x + y)P(x^{-1/2}) && \text{(by (iii))} \\ &= P(P(x^{-1/2})(x + y)) && \text{(by (i))} \\ &= P(P(x^{-1/2})P(x^{1/2})e + P(x^{-1/2})y) && \text{(by (i))} \\ &= P(e + P(x^{-1/2})y) && \text{(by (iii)).} \end{aligned}$$

Similarly,

$$\begin{aligned} P(x)^{-1/2}P(y)P(x)^{-1/2} &= P(x^{-1/2})P(y)P(x^{-1/2}) && \text{(by (iii))} \\ &= P(P(x^{-1/2})(y)) && \text{(by (i)).} \end{aligned}$$

Therefore, we have

$$P(x+y) - P(x) - P(y) = P(x)^{1/2}[P(e+P(x^{-1/2})y) - P(e) - P(P(x^{-1/2})y)]P(x)^{1/2}.$$

Let $z := P(x^{-1/2})y$. It follows from (2.3) that

$$P(e+z) - P(e) - P(z) = 2[L(e)L(z) + L(z)L(e) - L(e \circ z)] = 2L(z).$$

Note that $z = P(x^{-1/2})y \in \mathrm{int}K$ since $x^{-1/2}$ is invertible and $y \in \mathrm{int}K$ (see (ii) of Proposition 2.4). Thus, $P(x+y) - P(x) - P(y) = 2L(z) \succ 0$. Since $x \circ y \in \mathrm{int}K$ implies that $L(x \circ y)$ is positive definite, using (2.3) again, we can conclude that

$$\begin{aligned} L(x)L(y) + L(y)L(x) &= P(x,y) + L(x \circ y) \\ &= [P(x+y) - P(x) - P(y)] + L(x \circ y) \succ 0. \end{aligned}$$

(vi) Let $(x,y) \in \mathcal{U}$. Since $x \in \mathrm{int}K$, $L(x)$ is invertible by (i) of Proposition 2.1. Suppose that $(\Delta x, \Delta y)$ satisfy (2.5). The equation in (2.5) implies $\Delta y + L(x)^{-1}L(y)\Delta x = 0$ and $\langle \Delta x, \Delta y \rangle + \langle \Delta x, L(x)^{-1}L(y)\Delta x \rangle = 0$. By the inequality in (2.5), we have $\langle \Delta x, L(x)^{-1}L(y)\Delta x \rangle \leq 0$. Define $\Delta \tilde{x} = L(x)^{-1}\Delta x$. Then

$$\begin{aligned} 0 &\geq \langle \Delta x,\ L(x)^{-1}L(y)\Delta x \rangle \\ &= \langle \Delta \tilde{x},\ L(y)L(x)\Delta \tilde{x} \rangle \\ &= \langle \Delta \tilde{x},\ (L(x)L(y) + L(y)L(x))\Delta \tilde{x} \rangle / 2, \end{aligned}$$

which implies that $\Delta \tilde{x} = 0$ by the facts $(x,y) \in \mathcal{U}$ and (v) above. Finally, we see that

$$\Delta x = L(x)\Delta \tilde{x} = 0 \quad \text{and} \quad \Delta y = -L(x)^{-1}L(y)\Delta x = 0.$$

(vii) By the fact that $(e,e) \in \mathcal{U}$ and by the continuity of the operators $x \circ y$ and $L(x)L(y) + L(y)L(x)$, the set $\mathcal{U}$ is a nonempty open subset of $\mathrm{int}K \times \mathrm{int}K$. Let $(x,y) \in \mathcal{U}$. For $\theta \in [0,1]$, define

$$(x(\theta), y(\theta)) := (\theta e + (1-\theta)x, \theta e + (1-\theta)y) = \theta(e,e) - (1-\theta)(x,y).$$

To see that the set $\mathcal{U}$ is star-shaped, it suffices to show that $(x(\theta), y(\theta)) \in \mathcal{U}$ for any $\theta \in [0,1]$. Since the set $\mathrm{int}K \times \mathrm{int}K$ is convex, we have $(x(\theta), y(\theta)) \in \mathrm{int}K \times \mathrm{int}K$ for any $\theta \in [0,1]$. In addition, $x(\theta) \circ y(\theta)$ turns out to be

$$x(\theta) \circ y(\theta) = \theta^2 e + \theta(1-\theta)(x+y) + (1-\theta)^2 x \circ y,$$

where $e \in \mathrm{int}K$, $x + y \in \mathrm{int}K$, and $x \circ y \in \mathrm{int}K$. By the convexity of the cone $\mathrm{int}K$, we see that $x(\theta) \circ y(\theta) \in \mathrm{int}K$ and $(x(\theta), y(\theta)) \in \mathcal{U}$ for any $\theta \in [0,1]$.

(viii) For any $x \in \mathrm{int}K$, (i) of Proposition 2.1 implies $x \circ x \in \mathrm{int}K$ and $(x,x) \in \mathcal{U}$.

(ix) Since $\mathrm{int}K$ is a convex cone, for any $x \in \mathrm{int}K$ and $\alpha \in \Re_{++}$, it must hold that $\alpha e \in \mathrm{int}K$ and $x \circ (\alpha e) = \alpha x \in \mathrm{int}K$. Thus $\mathrm{int}K \times \{\alpha e :\ \alpha \in \Re_{++}\} \subseteq \mathcal{U}$ and $K \times \{\alpha e :\ \alpha \in \Re_{+}\} \subseteq \mathrm{cl}(\mathcal{U})$. By the symmetricity $x \circ y = y \circ x$, we obtain the assertion. □

A Euclidean Jordan algebra is called *simple* if it cannot be represented as the orthogonal direct sum of two Jordan algebras.

PROPOSITION 2.7.

(i) *Any Euclidean Jordan algebra V is, in a unique way, an orthogonal direct sum of simple Euclidean Jordan algebras $V_1, V_2, \ldots, V_m$. That is, any element $x \in V$ is uniquely represented by $x = \sum_{i=1}^{m} x_i$, where $x_i \in V_i$ $(i = 1,2,\ldots,m)$ and $x_i \circ x_j = 0$ $(i \neq j)$.*

(ii) *Let V be a simple Euclidean Jordan algebra. Then, for any $u \in V$,*

$$\mathrm{Tr}(L(u)) = \frac{n}{r}\mathrm{tr}(u).$$

(iii) *Let V be a simple Euclidean Jordan algebra. Then, for any nonzero idempotent c of V,*

$$0 < \sqrt{\frac{r}{2n}} \leq \|c\| = \sqrt{\langle e, c\rangle} \leq \sqrt{r}.$$

(iv) *Let V be a Euclidean Jordan algebra. Then, there exist $\omega_1 > 0$ and $\omega_2 > 0$ for which $0 < \omega_1 \leq \|c\| = \sqrt{\langle e, c\rangle} \leq \omega_2$ holds for any nonzero idempotent c of V.*

Proof. See Propositions III.4.4 and III.4.2 of [5] for (i) and (ii), respectively.

(iii) Since any nonzero idempotent c is an element of K, by (i) of Corollary 2.3, all eigenvalues of c are nonnegative and $\mathrm{tr}(c)$ is positive. The assertion (iii) follows from (ii) of the proposition, (ii) of Theorem 2.2, and

$$0 < \|c\| = \sqrt{\mathrm{tr}(c^2)} = \sqrt{\mathrm{tr}(c)} = \sqrt{\frac{r}{n}\mathrm{Tr}(L(c))}, \quad \langle e, c\rangle = \mathrm{tr}(e \circ c) = \mathrm{tr}(c) = \|c\|^2.$$

(iv) Let V be a Euclidean Jordan algebra. By the assertion (i) above, V is given by an orthogonal direct sum of simple Jordan algebras $V_1, V_2, \ldots, V_m$. For each $i = 1, 2, \ldots, m$, let us denote the dimension and the rank of V_i by n_i and r_i, respectively.

Suppose that c is a nonzero idempotent of V. Then c is given by $c = \sum_{i=1}^m c_i$ for some $c_i \in V_i$ $(i = 1, 2, \ldots, m)$. The orthogonality of V_is and the fact that $c^2 = c$ ensure that $c = c^2 = \sum_{i=1}^m c_i^2$. Therefore, by the uniqueness of the representation, we see that $c_i^2 = c_i$ $(i = 1, 2, \ldots, m)$; i.e., c is given by the sum of c_is which are idempotents of the simple Jordan algebras V_i $(i = 1, 2, \ldots, m)$.

Note that the orthogonality of V_is also ensures that

$$\|c\|^2 = \sum_{i=1}^m \|c_i\|^2 = \sum_{i=1}^m \langle e, c_i\rangle = \langle e, c\rangle.$$

Since $c \neq 0$, there exists $c_i \neq 0$ for some i, and by the assertion (iii) above, we obtain the following inequalities:

$$0 < \min_i \left\{\sqrt{\frac{r_i}{2n_i}}\right\} \leq \|c\| = \sqrt{\langle e, c\rangle} \leq \sqrt{\sum_{i=1}^m r_i}. \qquad \square$$

3. Homeomorphism of an interior point map. In this section, we extend the results in [19] to the case of symmetric cones and show the homeomorphism of an interior point map using the results in section 2.

The arguments used in the section are quite analogous to the ones in [19] and we omit some details in the proofs. Note that we restrict the domain of the map F to $\mathrm{int}K \times \mathrm{int}K \times \Re^m$ and it causes subtle differences in the results.

Here we introduce some notation and definitions. If M and N are two metric spaces, we denote the set of continuous functions from M to N by $C(M, N)$. For given $G \in C(M, N)$, $D \subseteq M$, and $E \subseteq N$, we define

$$G(D) := \{G(u) : \ u \in D\}, \quad G^{-1}(E) := \{u \in M : \ G(u) \in E\}.$$

We also denote "G restricted to the pair (D, E)" by $G|_{(D,E)}$. $G \in C(M, N)$ is called a *homeomorphism* from M onto N if G is bijective from M onto N and G and G^{-1} are continuous on M and N, respectively. We denote the set of homeomorphisms from M onto N by $\mathrm{Hom}(M, N)$. $G : M \to N$ is called a *local homeomorphism* from M onto N if for each $x \in M$, there exist open neighborhoods M_x of x and N_x of $G(x)$ such that $G|_{(M_x,N_x)} \in \mathrm{Hom}(M_x, N_x)$. In addition, for $a, b \in \Re$, $[a, b]$ denotes the line segment $\{x \in \Re : \ a \le x \le b\}$.

DEFINITION 3.1 (section 2 of [18], section 2.2 of [19]).

(i) *A metric space M is* connected *if there exists no partition (V_1, V_2) of M for which V_1 and V_2 are nonempty and open.*

(ii) *A metric space M is* path-connected *if for any two points $u_0, u_1 \in M$, there exists a path, i.e., a continuous function $p : [0, 1] \to M$ such that $p(0) = u_0$ and $p(1) = u_1$.*

(iii) *A metric space M is* simply connected *if it is path-connected and for any path $p : [0, 1] \to M$ with $p(0) = p(1) = u$, there exists a continuous map $\alpha : [0, 1] \times [0, 1] \to M$ such that $\alpha(s, 0) = p(s)$ and $\alpha(s, 1) = u$ for all $s \in [0, 1]$ and $\alpha(0, t) = \alpha(1, t) = u$ for all $t \in [0, 1]$.*

(vi) *The map $G \in C(M, N)$ is said to be* proper *with respect to the set $E \subseteq N$ if the set $G^{-1}(K) \subseteq M$ is compact for any compact set $K \subseteq E$. If G is proper with respect to N, we simply say that G is proper.*

It is easy to see that any star-shaped set (see Definition 2.5) in a normed vector space is simply connected.

The goal of this section is to show that the map H defined by (1.4) gives a homeomorphism between $\mathcal{U} \times \Re^m$ and $\mathrm{int}K \times F(\mathcal{U} \times \Re^m)$ under Assumption 1.1 (cf. Theorem 3.8 for affine maps and Theorem 3.10 for general maps). The homeomorphism ensures that there exists a unique path $H^{-1}(P) \subset \mathcal{U} \times \Re^m$ for any path $P \subseteq \mathrm{int}K \times F(\mathcal{U} \times \Re^m)$. As we will see in Theorem 3.12, if an additional assumption (Assumption 3.11) holds, then the set $\mathrm{int}K \times F(\mathcal{U} \times \Re^m)$ is open and convex. Therefore, if $0 \in \mathrm{cl}(\mathrm{int}K \times F(\mathcal{U} \times \Re^m))$ then we may choose the path as $P = \{th : \ t \in (0, 1]\} \subseteq \mathrm{int}K \times F(\mathcal{U} \times \Re^m)$ for any $h \in \mathrm{int}K \times F(\mathcal{U} \times \Re^m)$. The set $H^{-1}(P)$ is a so-called *(weighted) interior point trajectory* which is an important element in the development of interior point algorithms (cf. Corollary 4.4 and Theorems 5.4 and 5.5).

PROPOSITION 3.2 (Theorem 1 of [18], Proposition 1 of [19]). *Let M and N be metric spaces such that M is path-connected and N is simply connected. Suppose that $G : M \to N$ is a local homeomorphism. Then G is proper if and only if $G \in \mathrm{Hom}(M, N)$.*

PROPOSITION 3.3 (Corollary 1 of [18], Proposition 2 of [19]). *Let $G \in C(M, N)$, $M_0 \subseteq M$, and $N_0 \subseteq N$.*

(i) *Suppose that G, $M_0 \subseteq M$, and $N_0 \subseteq N$ satisfy the following conditions:*
   (a) *$G|_{(M_0,N)}$ is a local homeomorphism,*
   (b) *$G(M_0) \cap N_0 \neq \emptyset$,*
   (c) *$G(M \setminus M_0) \cap N_0 = \emptyset$, and*
   (d) *G is proper with respect to a subset E such that $N_0 \subseteq E \subseteq N$.*
   Then $G|_{(M_0 \cap G^{-1}(N_0), N_0)}$ is a proper local homeomorphism.

(ii) *Suppose that G, $M_0 \subseteq M$, and $N_0 \subseteq N$ satisfy the conditions* (a)–(d) *in* (i) *above and the additional condition below:*
   (e) *N_0 is connected.*
   Then $G(M_0) \supseteq N_0$ and $G(\mathrm{cl}(M_0)) \supseteq E \cap \mathrm{cl}(N_0)$.

PROPOSITION 3.4 (Corollary 3 of [18], Proposition 3 of [19]). *Let M be a path-connected metric space and let V be an n-dimensional real vector space. Suppose that $G : M \to V$ is a local homeomorphism and that $G^{-1}([y_0, y_1])$ is compact for any pair of points $y_0, y_1 \in G(M)$. Then, $G|_{(M,G(M))} \in \mathrm{Hom}(M, G(M))$ and $G(M)$ is convex.*

The following three lemmas are analogous to Lemmas 2–4 of [19]. See the proofs in [19] for detailed arguments.

LEMMA 3.5 (cf. Lemma 2 of [19]). *Let $F : \mathrm{int}K \times \mathrm{int}K \times \Re^m \to V \times \Re^m$ be a continuous map which satisfies Assumption* 1.1. *Let H be the map defined by* (1.4). *If the map H restricted to $\mathcal{U} \times \Re^m$ is a local homeomorphism, then the map H is proper with respect to $\mathrm{int}K \times F(\mathcal{U} \times \Re^m)$.*

Proof. Let C be a compact subset of $\mathrm{int}K \times F(\mathcal{U} \times \Re^m)$. Then the set $H^{-1}(C)$ is closed since H is continuous. The boundedness of the set $H^{-1}(C)$ can be obtained by using (i) and (ii) of Assumption 1.1 and (ii) of Proposition 2.1, with $\langle x, y\rangle = \mathrm{tr}(x \circ y)$. □

LEMMA 3.6 (cf. Lemma 3 of [19]). *Let $F : V \times V \times \Re^m \to V \times \Re^m$ be an affine map and let F^0 be the linear part of F.*

(i) *F is (x, y)-equilevel-monotone if and only if for any $(\Delta x, \Delta y, \Delta z) \in V \times V \times \Re^m$, $F^0(\Delta x, \Delta y, \Delta z) = 0$ implies that $\langle \Delta x, \Delta y\rangle \geq 0$.*

(ii) *F is z-injective if and only if for any $\Delta z \in \Re^m$, $F^0(0, 0, \Delta z) = 0$ implies that $\Delta z = 0$.*

(iii) *F is z-injective if and only if F is z-bounded.*

Proof. See the proof of Lemma 3 in [19]. □

LEMMA 3.7 (cf. Lemma 4 of [19]). *Let $F : V \times V \times \Re^m$ be an affine map which is (x, y)-equilevel-monotone and z-injective. Then H restricted to $\mathcal{U} \times \Re^m$ is a local homeomorphism.*

Proof. Since $\mathcal{U} \times \Re^m$ is an open set ((vii) of Lemma 2.6), it suffices to show that the derivative map $H'(x, y, z) : V \times V \times \Re^m \to V \times V \times \Re^m$ is an isomorphism for all $(x, y, z) \in \mathcal{U} \times \Re^m$. The isomorphism follows from (i) and (ii) of Lemma 3.6 and (vi) of Lemma 2.6. □

In the following theorem, we consider that F is affine. The theorem leads us to an important technical lemma, Lemma 3.9.

THEOREM 3.8 (cf. Theorem 1 of [19]). *Let $F : V \times V \times \Re^m \to V \times \Re^m$ be an affine map which is (x, y)-equilevel-monotone, z-injective, and z-bounded on $V \times V \times \Re^m$. Then the map H defined by* (1.4) *satisfies*

(i) *H is proper with respect to $\mathrm{int}K \times F(\mathcal{U} \times \Re^m)$,*

(ii) *H maps $\mathcal{U} \times \Re^m$ homeomorphically onto $\mathrm{int}K \times F(\mathcal{U} \times \Re^m)$.*

Proof. Define

$$\begin{aligned} &M := \mathrm{int}K \times \mathrm{int}K \times \Re^m,\ N := V \times V \times \Re^m,\ E := \mathrm{int}K \times F(\mathcal{U} \times \Re^m), \\ &M_0 := \mathcal{U} \times \Re^m,\ N_0 := \mathrm{int}K \times F(\mathcal{U} \times \Re^m),\ G := H|_{(M,N)}. \end{aligned} \tag{3.1}$$

We can easily see that

$$N_0 \subseteq E \subseteq N, \quad M_0 \subseteq H^{-1}(N_0). \tag{3.2}$$

(i) Since F is (x, y)-equilevel-bounded and z-injective, Lemma 3.7 and (iii) of Lemma 3.6 ensure that $H|_{(M_0,N)} = G|_{(M_0,N)}$ is a local homeomorphism and z-bounded. Thus the map F satisfies Assumption 1.1 and by Lemma 3.5, $H|_{(M,E)}$ is proper with respect to $E = \mathrm{int}K \times F(\mathcal{U} \times \Re^m)$.

(ii) Note that (i) above ensures that the requirement (d) of Proposition 3.3 is satisfied. In addition, Lemma 3.7 implies that $G|_{(M_0,N)} = H_{(M_0,N)}$ is a local homeomorphism, i.e., the requirement (a) holds. By similar discussions in the proof of Theorem 1 of [19], we can see that other requirements (b) and (c) are also satisfied.

Since (3.2) ensures the relation $M_0 \subseteq M_0 \cap H^{-1}(N_0) = M_0 \cap G^{-1}(N_0)$, we obtain that the map H restricted to

$$(M_0, N_0) = (\mathcal{U} \times \Re^m, \mathrm{int}K \times F(\mathcal{U} \times \Re^m))$$

is a proper local homeomorphism. Next, we show that $H(\mathcal{U}\times\Re^m) = \mathrm{int}K\times F(\mathcal{U}\times\Re^m)$ by using (ii) of Proposition 3.3. It is clear that

$$G(M_0) = H(\mathcal{U} \times \Re^m) \subseteq \mathrm{int}K \times F(\mathcal{U} \times \Re^m) = N_0.$$

To obtain the inverse inclusion, we should mention that the set N_0 is connected. In fact, by (vii) of Lemma 2.6, $\mathcal{U}\times\Re^m$ is star-shaped and hence path-connected. Since F is continuous, the sets $F(\mathcal{U}\times\Re^m)$ and $N_0 = \mathrm{int}K\times F(\mathcal{U}\times\Re^m)$ are also path-connected, and hence connected. Thus, applying (ii) of Proposition 3.3, we obtain

$$H(\mathcal{U} \times \Re^m) = G(M_0) \supseteq N_0 = \mathrm{int}K \times F(\mathcal{U} \times \Re^m).$$

Let us show that $G \in H(M_0, N_0)$. In (vii) of Lemma 2.6, we have seen that the set $\mathcal{U}$ is star-shaped. Since F is affine, both of the sets M_0 and N_0 are star-shaped and hence simply connected. By the local homeomorphism of G, the assertion $G \in H(M_0, N_0)$ follows from Proposition 3.2. □

LEMMA 3.9 (cf. Lemma 5 of [19]). *For any* $(x_0, y_0), (x_1, y_1) \in \mathcal{U}$, *if* $\langle x_0 - x_1, y_0 - y_1\rangle \geq 0$ *and* $x_0 \circ y_0 = x_1 \circ y_1$, *then* $(x_0, y_0) = (x_1, y_1)$.

Proof. See the proof of Lemma 5 of [19]. □

Since the set $\mathcal{U}$ is defined regardless of F, the above lemma is applicable to the case where the map F is nonlinear. The following theorem is our main result.

THEOREM 3.10 (cf. Theorem 2 of [19]). *Suppose that a continuous map* $F : \mathrm{int}K \times \mathrm{int}K \times \Re^m \to V \times \Re^m$ *satisfies Assumption* 1.1. *Then the map* H *defined by* (1.4) *satisfies the following properties:*

(i) H *is proper with respect to* $\mathrm{int}K \times F(\mathcal{U} \times \Re^m)$.

(ii) H *maps* $\mathcal{U} \times \Re^m$ *homeomorphically onto* $\mathrm{int}K \times F(\mathcal{U} \times \Re^m)$.

Proof. Define the sets M, N, E, M_0, and N_0 and the map G as in (3.1). Then (3.2) holds even in this case.

(i) To show the local homeomorphism of $H|_{(M_0,N)} = G|_{(M_0,N)}$, we use Lemma 3.9 instead of Lemma 3.7. Since $H|_{(M_0,N)}$ is a continuous map from an open subset of $V \times V \times \Re^m$ into the same space, by the domain invariance theorem, it suffices to show that $H|_{(M_0,N)}$ is one-to-one.

Suppose that $(\hat{x}, \hat{y}, \hat{z}), (\tilde{x}, \tilde{y}, \tilde{z}) \in \mathcal{U} \times \Re^m$ satisfy $H(\hat{x}, \hat{y}, \hat{z}) = H(\tilde{x}, \tilde{y}, \tilde{z})$, i.e.,

$$F(\hat{x}, \hat{y}, \hat{z}) = F(\tilde{x}, \tilde{y}, \tilde{z}), \quad \hat{x} \circ \hat{y} = \tilde{x} \circ \tilde{y}.$$

Since F is (x, y)-equilevel-monotone, we see that

$$\langle \hat{x} - \tilde{x}, \hat{y} - \tilde{y}\rangle \geq 0, \quad \hat{x} \circ \hat{y} = \tilde{x} \circ \tilde{y}.$$

By Lemma 3.9, the above relations imply that $(\hat{x}, \hat{y}) = (\tilde{x}, \tilde{y})$ and by the z-injectivity of F, we have $(\hat{x}, \hat{y}, \hat{z}) = (\tilde{x}, \tilde{y}, \tilde{z})$. Thus, $H|_{(M_0,N)}$ is one-to-one and maps M_0 homeomorphically onto $H(M_0)$. By the local homeomorphism of $H|_{(M_0,N)}$ together with

the equilevel-monotonicity and the z-boundedness of the map F, Lemma 3.5 ensures that H is proper with respect to the set $E = \mathrm{int}K \times F(\mathcal{U} \times \Re^m)$.

(ii) In the proof of (i) above, we have already seen that (a) and (d) of Proposition 3.3 hold. By the same discussions as in the proof of (ii) of Theorem 3.8, we can see that the assumptions (b) and (c) of Proposition 3.3 are also satisfied. Since F is continuous, $F(\mathcal{U} \times \Re^m)$ and $N_0 = \mathrm{int}K \times F(\mathcal{U} \times \Re^m)$ are also path-connected. Thus, as in the proof of (ii) of Theorem 3.8 again, we obtain that

$$H(\mathcal{U} \times \Re^m) = G(M_0) = N_0 = \mathrm{int}K \times F(\mathcal{U} \times \Re^m). \qquad \square$$

Next, we discuss the convexity of the set $F(\mathcal{U} \times \Re^m)$, which is a key property in section 5. We impose the additional assumption below on the map F(cf. [19]).

ASSUMPTION 3.11. *F is (x, y)-everywhere-monotone on the domain with respect to the set $V \times \Re^m$, i.e., there exist continuous functions ϕ from the domain of F to the set $V \times \Re^m$ and $c : (V \times \Re^m) \times (V \times \Re^m) \to \Re$ such that $c(r, r) = 0$ for any $r \in V \times \Re^m$ and*

$$\langle x - x', y - y' \rangle \geq \langle r - r', \phi(x, y, z) - \phi(x', y', z') \rangle_{V \times \Re^m} + c(r, r')$$

holds for any (x, y, z) and (x', y', z') in the domain of F satisfying $F(x, y, z) = r$ and $F(x', y', z') = r'$. Here we define

$$\langle (a, b), (a', b') \rangle_{V \times \Re^m} = \langle a, a' \rangle + b^T b'$$

for any $(a, b), (a', b') \in V \times \Re^m$.

It can be easily seen that if F is (x, y)-everywhere-monotone then $r = r'$ implies that $\langle x - x', y - y' \rangle \geq 0$ and F is (x, y)-equilevel-monotone.

THEOREM 3.12 (cf. Theorem 3 of [19]). *Suppose that a continuous map $F : \mathrm{int}K \times \mathrm{int}K \times \Re^m \to V \times \Re^m$ satisfies Assumptions* 1.1 *and* 3.11. *Then the set $F(\mathcal{U} \times \Re^m)$ is an open convex set.*

Proof. It suffices to show that $H(\mathcal{U} \times \Re^m)$ is open and convex since $H(\mathcal{U} \times \Re^m) = \mathrm{int}K \times F(\mathcal{U} \times \Re^m)$ holds by (ii) of Theorem 3.10. Since $\mathcal{U} \times \Re^m$ is open ((vii) of Lemma 2.6), (ii) of Theorem 3.10 implies that the set $H(\mathcal{U} \times \Re^m)$ is also open. Define

$$M := \mathcal{U} \times \Re^m, \quad N := V \times V \times \Re^m, \quad G := H|_{(M,N)}.$$

Then the set M is path-connected by (vii) of Lemma 2.6 and G is a local homeomorphism by (ii) of Theorem 3.10. Using (ii) of Proposition 2.1 and (iv) of Lemma 2.6, we can obtain the compactness of the set $G^{-1}([w_0.w_1])$ by similar arguments as in the proof of Theorem 3 of [19]. Therefore, the convexity of the set $H(\mathcal{U} \times \Re^m)$ follows from Proposition 3.4. $\square$

4. Solvability of the CP. In this section, we discuss the solvability of the CP assuming that the map F is defined and continuous on the set $K \times K \times \Re^m$ instead of $\mathrm{int}K \times \mathrm{int}K \times \Re^m$. The following results are direct extensions of the ones in [19] to the case of symmetric cones and obtained similarly as in the previous section. In the proofs below, we give only the differences arising from the expansion of the domain of F.

LEMMA 4.1 (cf. Lemma 2 of [19] and Lemma 3.5). *Let $F : K \times K \times \Re^m \to V \times \Re^m$ be a continuous map which is (x, y)-equilevel-monotone and z-bounded on $K \times K \times \Re^m$. Let H be the map defined by* (1.4). *If the map H restricted to $\mathcal{U} \times \Re^m$ is a local homeomorphism, then the map H is proper with respect to $V \times F(\mathcal{U} \times \Re^m)$.*

Proof. Let C be a compact subset of $V \times F(\mathcal{U} \times \Re^m)$. Since the domain $K \times K \times \Re^m$ of F is closed, by the continuity of H, the set $H^{-1}(C)$ is always closed. The argument to obtain the boundedness of $H^{-1}(C)$ is the same as in the proof of Lemma 3.5. □

THEOREM 4.2 (cf. Theorem 2 of [19] and Theorem 3.10). *Suppose that a continuous map $F : K \times K \times \Re^m \to V \times \Re^m$ satisfies Assumption* 1.1. *Then the map H defined by* (1.4) *satisfies that*

(i) *H is proper with respect to $V \times F(\mathcal{U} \times \Re^m)$,*
(ii) *H maps $\mathcal{U} \times \Re^m$ homeomorphically onto $\mathrm{int}K \times F(\mathcal{U} \times \Re^m)$, and*
(iii) *$H(K \times K \times \Re^m) \supseteq K \times F(\mathcal{U} \times \Re^m)$.*

Proof. Define the sets M, N, M_0, and N_0 and the map G as in (3.1), and the set E by

$$E := V \times F(\mathcal{U} \times \Re^m).$$

(i) The continuity assumption on F is stricter than the one in Theorem 3.10. Thus the local homeomorphism of $H|_{(M_0,N)} = G|_{(M_0,N)}$ is similarly obtained from Lemma 3.9. Using Lemma 4.1 instead of Lemma 3.5, we can see that H is proper with respect to $E = V \times F(\mathcal{U} \times \Re^m)$, i.e., the assertion (i) holds.

(ii) In the above discussion, we have shown that (d) of Proposition 3.3 holds. Since the sets M, N, and N_0 are the same as in Theorem 3.10, we can see that the assumptions (a)–(c) of Proposition 3.3 are also satisfied and that

$$H(\mathcal{U} \times \Re^m) = G(M_0) = N_0 = \mathrm{int}K \times F(\mathcal{U} \times \Re^m)$$

holds.

(iii) Since $\mathcal{U}$ is star-shaped ((vii) of Lemma 2.6), $\mathcal{U} \times \Re^m$ is connected, and N_0 is also connected by the continuity of F. Combining this with the facts

$$K \times K \times \Re^m = \mathrm{cl}(M) \supseteq \mathrm{cl}(M_0) \quad \text{and} \quad E \cap \mathrm{cl}(N_0) = K \times F(\mathcal{U} \times \Re^m),$$

the assertion (iii) follows from (ii) of Proposition 3.3. □

The following corollary is a direct consequence of Theorem 3.12.

COROLLARY 4.3 (cf. Theorem 3 of [19] and Theorem 3.12). *Suppose that a continuous map $F : K \times K \times \Re^m \to V \times \Re^m$ satisfies Assumptions* 1.1 *and* 3.11. *Then the set $F(\mathcal{U} \times \Re^m)$ is an open convex set.*

COROLLARY 4.4 (cf. Corollary 1 of [19]). *Suppose that the map $F : K \times K \times \Re^m \to V \times \Re^m$ is continuous and that $0 \in F(\mathcal{U} \times \Re^m)$, which implies that the CP has an interior feasible point $(\bar{x}, \bar{y}, \bar{z}) \in \mathrm{int}K \times \mathrm{int}K \times \Re^m$ satisfying $\bar{x} \circ \bar{y} \in \mathrm{int}K$ and $F(\bar{x}, \bar{y}, \bar{z}) = 0$. Let $p : [0,1] \to K \times F(\mathcal{U} \cap \Re^m)$ be a path for which $p(0) = 0$ and $p(t) \in \mathrm{int}K \times F(\mathcal{U} \times \Re^m)$ hold.*

(i) *If the map F satisfies Assumption* 1.1 *then there exists a unique path $(x, y, z) : (0,1] \to \mathrm{int}K \times \mathrm{int}K \times \Re^m$ such that $H(x(t), y(t), z(t)) = p(t)$ for all $t \in (0,1]$ and $\{(x(t), y(t), z(t)) : t \in (0,1]\}$ is bounded. Thus the CP is asymptotically solvable. In addition, any accumulation point of $\{(x(t), y(t), z(t)) : t \in (0,1]\}$ is a solution of the CP.*
(ii) *If F satisfies Assumptions* 1.1 *and* 3.11 *then for each $h \in \mathrm{int}K \times F(\mathcal{U} \cap \Re^m)$, we can take $p(t) = th$ for all $t \in [0,1]$ as a path $p : [0,1] \to K \times F(\mathcal{U} \cap \Re^m)$ satisfying the requirements above.*

Proof. (i) Note that $\{p(t) : t \in [0,1]\}$ is bounded by (ii) of Definition 3.1 of a path. The assertion follows from Theorem 4.2: (ii) of the theorem implies the unique existence of $(x(t), y(t), z(t))$ for any $t \in (0,1]$, (i) implies the boundedness of

$(x(t), y(t), z(t))$, and (iii) implies that any accumulation point of $\{(x(t), y(t), z(t))\}$ is a solution of the CP.

(ii) Since $(0,0) \in K \times F(\mathcal{U} \times \Re^m)$, Corollary 4.3 ensures that $\{p(t) : p(t) = th,\ t \in [0,1]\} \subseteq K \times F(\mathcal{U} \times \Re^m)$ for any $h \in \mathrm{int}K \times F(\mathcal{U} \cap \Re^m)$. It is obvious that $\{p(t) :\ t \in [0,1]\}$ is bounded, $p(0) = 0$, and $p(t) \in \mathrm{int}K \times F(\mathcal{U} \times \Re^m)$ for all $t \in (0,1]$. □

5. A homogeneous model for the CP. In this section, we give a homogenous model for solving a special class of CPs where the map $F : K \times K \to V$ is of the form

$$F(x,y) = y - \psi(x) \tag{5.1}$$

for a continuous map $\psi : K \to V$.

Our model is a natural extension of the homogeneous model proposed by Andersen and Ye [2] for solving CPs where K is the positive orthant in $\Re^n$. One of the remarkable features of their model is that the associated trajectory gives certifications on the strong feasibility or the strong infeasibility of the original problem. Our results, Theorems 5.4 and 5.5, show that our homogeneous model inherits the property even for the case of symmetric cones.

Throughout this section, we impose the following assumption on ψ.

ASSUMPTION 5.1. *The map $\psi : K \to V$ in (5.1) is monotone on the set K, i.e.,*

$$\langle \psi(x) - \psi(x'), x - x' \rangle \geq 0 \quad \textit{for all } x, x' \in K.$$

PROPOSITION 5.2. *Suppose that $S \subseteq K$ and $\psi : S \to V$ is monotone on the set S. Then the map $F : S \times K \to V$ given by (5.1) is (x,y)-everywhere-monotone on the set $S \times K$ with $m = 0$.*

Proof. Define $\phi : S \times K \to V$ and $c : V \times V \to \Re$ by $\phi(x,y) := x$ and $c := 0$. Let $r := F(x,y)$ and $r' := F(x',y')$, where $(x,y), (x',y') \in S \times K$. Then we see that $\psi(x) - \psi(x') = (y - y') - (r - r')$, and the monotonicity of ψ implies that

$$\begin{aligned} 0 &\leq \langle \psi(x) - \psi(x'), x - x' \rangle \\ &= \langle (y - y') - (r - r'), x - x' \rangle \\ &= \langle y - y', x - x' \rangle - \langle r - r', x - x' \rangle \\ &= \langle y - y', x - x' \rangle - \langle r - r', \phi(x,y) - \phi(x',y') \rangle + c(r, r'). \end{aligned}$$

Thus, by the definition of (x,y)-everywhere-monotonicity in Assumption 3.11, the map F is (x,y)-everywhere-monotone on the set $S \times K$ with $m = 0$. □

Define the sets $\Re_+ := \{\tau \in \Re :\ \tau \geq 0\}$ and $\Re_{++} := \{\tau \in \Re :\ \tau > 0\}$. For a given CP with a map F of the form (5.1), we consider the homogeneous model

$$\begin{array}{lll} \text{(HCP)} & \text{find} & (x, \tau, y, \kappa) \in (K \times \Re_{++}) \times (K \times \Re_+) \\ & \text{s.t.} & F_{\mathrm{H}}(x, \tau, y, \kappa) = 0,\ (x, \tau) \circ_{\mathrm{H}} (y, \kappa) = 0, \end{array} \tag{5.2}$$

where $F_{\mathrm{H}} : (K \times \Re_{++}) \times (K \times \Re_+) \to (V \times \Re)$ and $(x, \tau) \circ_{\mathrm{H}} (y, \kappa)$ are given by

$$F_{\mathrm{H}}(x, \tau, y, \kappa) := \begin{pmatrix} y - \tau\psi(x/\tau) \\ \kappa + \langle \psi(x/\tau), x \rangle \end{pmatrix} \tag{5.3}$$

and

$$(x, \tau) \circ_{\mathrm{H}} (y, \kappa) := \begin{pmatrix} x \circ y \\ \tau\kappa \end{pmatrix}. \tag{5.4}$$

For ease of notation, we use the following symbols:

$$V_{\rm H} := V \times \Re, \quad K_{\rm H} := K \times \Re_+, \quad x_{\rm H} := (x, \tau) \in V_{\rm H}, \quad y_{\rm H} := (y, \kappa) \in V_{\rm H}. \tag{5.5}$$

It should be noted that the set $K_{\rm H}$ is a Cartesian product of two symmetric cones K and $\Re_+$ and is given by

$$K_{\rm H} = \left\{ x_{\rm H}^2 = \begin{pmatrix} x^2 \\ \tau^2 \end{pmatrix} : \ x_{\rm H} \in V_{\rm H} \right\}.$$

Thus the closed convex cone $K_{\rm H}$ is the symmetric cone of $V_{\rm H}$. It can be easily seen that $\mathrm{int}K_{\rm H} = \mathrm{int}K \times \Re_{++}$.

The scalar product $\langle (x, \tau), (y, \kappa) \rangle_{\rm H}$ associated to the bilinear product $\circ_{\rm H}$ is given by

$$\langle (x, \tau), (y, \kappa) \rangle_{\rm H} := \langle x, y \rangle + \tau\kappa. \tag{5.6}$$

For any $x_{\rm H} = (x, \tau) \in V_{\rm H}$, the linear operator

$$L_{\rm H}(x_{\rm H}) := \begin{pmatrix} L(x) & 0 \\ 0^T & \tau \end{pmatrix}$$

is self-adjoint with respect to $\langle \cdot, \cdot \rangle$, i.e., $\langle v_{\rm H}, L_{\rm H}(x_{\rm H}) w_{\rm H} \rangle = \langle L_{\rm H}(x_{\rm H}) v_{\rm H}, w_{\rm H} \rangle$ for any $v_{\rm H}, w_{\rm H} \in V_{\rm H}$.

Let us define the map $\psi_{\rm H}$ by

$$\psi_{\rm H}(x_{\rm H}) = \psi_{\rm H}(x, \tau) := \begin{pmatrix} \tau\psi(x/\tau) \\ -\langle \psi(x/\tau), x \rangle \end{pmatrix} \tag{5.7}$$

for any $x_{\rm H} = (x, \tau) \in K \times \Re_{++}$. Then the map $F_{\rm H}$ is given by

$$F_{\rm H}(x_{\rm H}, y_{\rm H}) = y_{\rm H} - \psi_{\rm H}(x_{\rm H}). \tag{5.8}$$

We also define the set

$$\mathcal{U}_{\rm H} := \{ (x_{\rm H}, y_{\rm H}) \in \mathrm{int}K_{\rm H} \times \mathrm{int}K_{\rm H} : \ x_{\rm H} \circ_{\rm H} y_{\rm H} \in \mathrm{int}K_{\rm H} \}. \tag{5.9}$$

It is clear that the set $\mathcal{U}_{\rm H}$ has the properties described in Lemma 2.6 with $\mathcal{U} = \mathcal{U}_{\rm H}$.

The following proposition shows that a monotonicity of the map $F_{\rm H}$ on the set $\mathrm{int}K_{\rm H} \times \mathrm{int}K_{\rm H}$ can be obtained if the map ψ is monotone on the set K.

PROPOSITION 5.3. *Suppose that $\psi : K \to V$ satisfies Assumption* 5.1. *Then*

(i) *the map $\psi_{\rm H}$ is monotone on* $\mathrm{int}K_{\rm H}$,

(ii) *the map $F_{\rm H}$ is $(x_{\rm H}, y_{\rm H})$-everywhere-monotone on* $\mathrm{int}K_{\rm H} \times \mathrm{int}K_{\rm H}$.

Thus, $F_{\rm H}$ with the domain $\mathrm{int}K_{\rm H} \times \mathrm{int}K_{\rm H}$ *satisfies Assumptions* 1.1 *and* 3.11 *with $m = 0$ whenever ψ is monotone on K.*

Proof. (i) For any $x_{\rm H}, x'_{\rm H} \in \mathrm{int}K_{\rm H}$, it follows from the definition (5.7) that

$$\begin{aligned}
&\langle \psi_{\rm H}(x_{\rm H}) - \psi_{\rm H}(x'_{\rm H}), x_{\rm H} - x'_{\rm H} \rangle_{\rm H} \\
&= \langle \tau\psi(x/\tau) - \tau'\psi(x'/\tau'), x - x' \rangle - (\tau - \tau')[\langle \psi(x/\tau), x \rangle - \langle \psi(x'/\tau'), x' \rangle] \\
&= \langle \tau\psi(x/\tau), x - x' \rangle - \langle \tau'\psi(x'/\tau'), x - x' \rangle - (\tau - \tau')\langle \psi(x/\tau), x \rangle + (\tau - \tau')\langle \psi(x'/\tau'), x' \rangle \\
&= -\tau\langle \psi(x/\tau), x' \rangle - \tau'\langle \psi(x'/\tau'), x \rangle + \tau'\langle \psi(x/\tau), x \rangle + \tau\langle \psi(x'/\tau'), x' \rangle \\
&= -\tau\tau'\langle \psi(x/\tau), x'/\tau' \rangle - \tau\tau'\langle \psi(x'/\tau'), x/\tau \rangle + \tau\tau'\langle \psi(x/\tau), x/\tau \rangle + \tau\tau'\langle \psi(x'/\tau'), x'/\tau' \rangle \\
&= \tau\tau'\langle \psi(x/\tau), (x/\tau) - (x'/\tau') \rangle - \tau\tau'\langle \psi(x'/\tau'), (x/\tau) - (x'/\tau') \rangle \\
&= \tau\tau'\langle \psi(x/\tau) - \psi(x'/\tau'), (x/\tau) - (x'/\tau') \rangle \\
&\geq 0,
\end{aligned}$$

where the last inequality follows from the monotonicity of ψ.

(ii) The assertion follows from (i) above and Proposition 5.2 with $S = \mathrm{int}\, K_{\mathrm{H}}$. □

In the theorem below, the assertions (i)–(iii) follow only from the construction (5.7) of the map ψ_{H}. Note that the assertion (v) in the theorem ensures that if the original CP is strongly infeasible, then, under the assumption of Lipschitz continuity, a finite certificate proving infeasibility is given by an asymptotic solution of the HCP.

THEOREM 5.4 (cf. Theorem 1 of [2]). *Suppose that $\psi : K \to V$ satisfies Assumption* 5.1.

(i) *For any $x_{\mathrm{H}} \in \mathrm{int}K_{\mathrm{H}}$, $\langle x_{\mathrm{H}}, \psi_{\mathrm{H}}(x_{\mathrm{H}})\rangle_{\mathrm{H}} = 0$.*

(ii) *Any asymptotically feasible solution $(\hat{x}_{\mathrm{H}}, \hat{y}_{\mathrm{H}})$ of the HCP is an asymptotically complementary solution.*

(iii) *The HCP is asymptotically feasible.*

(iv) *The CP has a solution if and only if the HCP has an asymptotic solution $(x^*_{\mathrm{H}}, y^*_{\mathrm{H}}) = (x^*, \tau^*, y^*, \kappa^*)$ with $\tau^* > 0$. In this case, $(x^*/\tau^*, y^*/\tau^*)$ is a solution of the CP.*

(v) *Suppose that ψ satisfies the Lipschitz condition on K; i.e., there exists a constant $\gamma \geq 0$ such that*

$$\|\psi(x+h) - \psi(x)\| \leq \gamma\|h\| \quad \text{for any} \quad x \in K \text{ and } h \in V.$$

If the CP is strongly infeasible then the HCP has an asymptotic solution $(x^*, \tau^*, y^*, \kappa^*)$ with $\kappa^* > 0$. Conversely, if the HCP has an asymptotic solution $(x^*, \tau^*, y^*, \kappa^*)$ with $\kappa^* > 0$ then the CP is infeasible. In the latter case, $(x^*/\kappa^*, y^*/\kappa^*)$ is a certificate to prove infeasibility of the CP.

Proof. (i) By a simple calculation, we have $\langle x_{\mathrm{H}}, \psi_{\mathrm{H}}(x_{\mathrm{H}})\rangle_{\mathrm{H}} = \langle x, \tau\psi(x/\tau)\rangle - \tau\langle \psi(x/\tau), x\rangle = 0$.

(ii) Suppose that $(\hat{x}_{\mathrm{H}}, \hat{y}_{\mathrm{H}})$ is an asymptotically feasible solution. Then there exists a bounded sequence $(x^{(k)}_{\mathrm{H}}, y^{(k)}_{\mathrm{H}}) \in \mathrm{int}K_{\mathrm{H}} \times \mathrm{int}K_{\mathrm{H}}$ such that

$$\lim_{k\to\infty} F_{\mathrm{H}}(x^{(k)}_{\mathrm{H}}, y^{(k)}_{\mathrm{H}}) = \lim_{k\to\infty} (y^{(k)}_{\mathrm{H}} - \psi_{\mathrm{H}}(x^{(k)}_{\mathrm{H}})) = 0.$$

Here (i) above implies that

$$\langle x^{(k)}_{\mathrm{H}}, y^{(k)}_{\mathrm{H}}\rangle_{\mathrm{H}} = \langle x^{(k)}_{\mathrm{H}}, y^{(k)}_{\mathrm{H}}\rangle_{\mathrm{H}} - \langle x^{(k)}_{\mathrm{H}}, \psi_{\mathrm{H}}(x^{(k)}_{\mathrm{H}})\rangle_{\mathrm{H}} = \langle x^{(k)}_{\mathrm{H}}, y^{(k)}_{\mathrm{H}} - \psi_{\mathrm{H}}(x^{(k)}_{\mathrm{H}})\rangle_{\mathrm{H}}$$

holds for all $k \geq 0$. Thus, we see that $\lim_{k\to\infty}\langle x^{(k)}_{\mathrm{H}}, y^{(k)}_{\mathrm{H}}\rangle_{\mathrm{H}} = 0$. By the definition (5.6) of $\langle\cdot,\cdot\rangle_{\mathrm{H}}$ and (ii) of Lemma 2.6, we obtain that $(\hat{x}_{\mathrm{H}}, \hat{y}_{\mathrm{H}})$ is an asymptotically complementary solution.

(iii) For each $k \geq 0$, define

$$x^{(k)} := (1/2)^k e \in \mathrm{int}K,\ \tau_k := (1/2)^k \in \Re_{++},\ y^{(k)} := (1/2)^k e \in \mathrm{int}K,$$
$$\kappa_k := (1/2)^k \in \Re_{++}.$$

Then the sequence $\{(x^{(k)}_{\mathrm{H}}, y^{(k)}_{\mathrm{H}})\} = \{(x^{(k)}, \tau_k, y^{(k)}, \kappa_k)\}$ is bounded and

$$\lim_{k\to\infty} (y^{(k)} - \tau_k\psi(x^{(k)}/\tau_k)) = 0 \quad \text{and} \quad \lim_{k\to\infty} (\kappa_k + \langle\psi(x^{(k)}/\tau_k), x^{(k)}\rangle) = 0.$$

Thus, the bounded sequence $\{(x^{(k)}_{\mathrm{H}}, y^{(k)}_{\mathrm{H}})\} \subseteq \mathrm{int}K_{\mathrm{H}} \times \mathrm{int}K_{\mathrm{H}}$ satisfies

$$\lim_{k\to\infty} (y^{(k)}_{\mathrm{H}} - \psi_{\mathrm{H}}(x^{(k)}_{\mathrm{H}})) = \lim_{k\to\infty} F_{\mathrm{H}}(x^{(k)}, y^{(k)}_{\mathrm{H}}) = 0,$$

which implies that the HCP is asymptotically feasible.

(iv) It is easy to see that if $(x^*, \tau^*, y^*, \kappa^*) \in (K \times \Re_+) \times (K \times \Re_+)$ is a solution of the HCP with $\tau^* > 0$, then $y^*/\tau^* - \psi(x^*/\tau^*) = 0$, $x^* \circ y^* = 0$ and $(x^*/\tau^*, y^*/\tau^*) \in K \times K$ is a solution of the CP.

Conversely, let $(\hat{x}, \hat{y}) \in K \times K$ be a solution of the CP. Then $\hat{y} - 1 \cdot \psi(\hat{x}/1) = 0$, $(\hat{x}, 1) \circ_{\mathrm{H}} (\hat{y}, 0) = 0$, and $(\hat{x}, 1, \hat{y}, 0) \in (K \times \Re_{++}) \times (K \times \Re_+)$ is a solution of the HCP.

(v) By Proposition 5.2, the monotonicity of the map ψ on the set K implies that the map F defined by (5.1) satisfies Assumptions 1.1 and 3.11 with the domain $K \times K$ and $m = 0$. Thus the set $F(\mathcal{U})$ is open and convex by Theorem 3.12.

If the CP is strongly infeasible, then we must have $0 \notin \mathrm{cl}(F(\mathcal{U}))$. Since the set $\mathrm{cl}(F(\mathcal{U}))$ is a closed convex set, by the separating hyperplane theorem, there exist $a \in V$ with $\|a\| = 1$ and $\xi \in \Re$ such that

$$\langle a, b \rangle \geq \xi > 0 \quad \text{for all } b \in \mathrm{cl}(F(\mathcal{U})). \tag{5.10}$$

Since F is continuous on the set $\mathrm{cl}(\mathcal{U}) \subseteq K \times K$, we can see that $F(\mathrm{cl}(\mathcal{U})) \subseteq \mathrm{cl}(F(\mathcal{U}))$. In fact, if $b \in F(\mathrm{cl}(\mathcal{U}))$ then there exists a sequence satisfying

$$(x^{(k)}, y^{(k)}) \in \mathcal{U}, \quad \lim_{k \to \infty} (x^{(k)}, y^{(k)}) = (\bar{x}, \bar{y}) \in \mathrm{cl}(\mathcal{U}), \quad F(\bar{x}, \bar{y}) = b,$$

and the continuity of F on the set $\mathrm{cl}(\mathcal{U})$ implies that $\lim_{k \to \infty} F(x^{(k)}, y^{(k)}) = F(\bar{x}, \bar{y}) = b$. Therefore (5.10) implies that

$$\langle a, F(x, y) \rangle = \langle a, y - \psi(x) \rangle = \langle a, y \rangle - \langle a, \psi(x) \rangle \geq \xi > 0 \tag{5.11}$$

for any $(x, y) \in \mathrm{cl}(\mathcal{U})$. Note that (ix) of Lemma 2.6 ensures that the above relation (5.11) holds at $(x, y) = (0, \alpha\bar{y})$ for any fixed $\bar{y} \in K$ and any $\alpha > 0$. Thus, it must be true that $\langle a, \bar{y} \rangle \geq 0$ for all $\bar{y} \in K$, which implies that $a \in K$. Similarly, since $(x, 0) \in \mathrm{cl}(\mathcal{U})$ for all $x \in K$, it follows from (5.11) that

$$-\langle a, \psi(x) \rangle \geq \xi > 0 \tag{5.12}$$

for any $x \in K$. Thus, combining this with the fact that $a \in K$, we see that

$$-\langle a, \psi(\beta a) \rangle \geq \xi > 0 \quad \text{for all } \beta \geq 0. \tag{5.13}$$

From the monotonicity of the map ψ on the set $K \times K$, we also see that for any $x \in K$ and $\beta \geq 0$, $0 \leq \langle \beta x - x, \psi(\beta x) - \psi(x) \rangle = (\beta - 1)\langle x, \psi(\beta x) - \psi(x) \rangle$. Thus, for $\beta \geq 1$,

$$\langle x, \psi(\beta x) - \psi(x) \rangle \geq 0, \tag{5.14}$$

which implies that

$$\liminf_{\beta \to \infty} \langle x, \psi(\beta x) \rangle / \beta \geq 0. \tag{5.15}$$

Let $\{\beta_k\}$ be a sequence such that $\beta_k \to +\infty$, and let

$$\frac{\psi(\beta_k a)}{\beta_k} = \sum_{i=1}^{r} \lambda_i^{(k)} c_i^{(k)} \tag{5.16}$$

be a decomposition given by (i) of Theorem 2.2 for each k. We also define

$$\begin{aligned} &\lambda_k := \min\{\lambda_i^{(k)} \ (i = 1, 2, \ldots, r)\}, \\ &j_k \in \arg\min\{\lambda_i^{(k)} \ (i = 1, 2, \ldots, r)\}, \quad c^{(k)} := c_{j_k}^{(k)}. \end{aligned} \tag{5.17}$$

Note that $\{c^{(k)}\}$ is a sequence of primitive (i.e., nonzero) idempotents of a Euclidean Jordan algebra $(V, \circ)$. Thus, by (iv) of Proposition 2.7, there exist $\omega_1, \omega_2 > 0$ such that

$$0 < \omega_1 \leq \|c^{(k)}\| \leq \omega_2 \quad \text{for all } k. \tag{5.18}$$

We first claim that $\liminf_{k\to\infty} \lambda_k \geq 0$. Suppose that $\liminf_{k\to\infty} \lambda_k < 0$. Then, by taking a subsequence if necessary, we may assume that there exists a $\delta > 0$ satisfying $\lambda_k \leq -\delta$ for sufficiently large k's. Define $x^{(k)} := a + \epsilon c^{(k)}$ for $\epsilon > 0$. We can see that

$$\begin{aligned}
\langle x^{(k)}, \psi(\beta_k x^{(k)})\rangle/\beta_k &= \langle a + \epsilon c^{(k)}, \psi(\beta_k x^{(k)})\rangle/\beta_k \\
&= \langle a, \psi(\beta_k x^{(k)})\rangle/\beta_k + \epsilon\langle c^{(k)}, \psi(\beta_k x^{(k)})\rangle/\beta_k \\
&< \epsilon\langle c^{(k)}, \psi(\beta_k x^{(k)})\rangle/\beta_k \qquad \text{(by (5.12))} \\
&= \epsilon\left(\langle c^{(k)}, \psi(\beta_k x^{(k)}) - \psi(\beta_k a)\rangle/\beta_k + \langle c^{(k)}, \psi(\beta_k a)\rangle/\beta_k\right). \qquad (5.19)
\end{aligned}$$

The definitions (5.16) and (5.17) and the boundedness (5.18) of $\{c^{(k)}\}$ ensure that there exists $\bar{k}$ for which

$$\langle c^{(k)}, \psi(\beta_k a)\rangle/\beta_k = \lambda_k\langle c^{(k)}, c^{(k)}\rangle \leq -\delta\omega_1^2 < 0 \tag{5.20}$$

holds for any $k \geq \bar{k}$. In addition, since we set $x^{(k)} = a + \epsilon c^{(k)}$, by the Lipschitz continuity of ψ and the boundedness of $\{c^{(k)}\}$, there exist $\bar{\gamma} > 0$ and $\bar{\epsilon} > 0$ independent from k, for which

$$\langle c^{(k)}, \psi(\beta_k x^{(k)}) - \psi(\beta_k a)\rangle/\beta_k \leq \bar{\gamma}\epsilon \leq \delta\omega_1^2/2 \tag{5.21}$$

holds for any $\epsilon \leq \bar{\epsilon}$. Thus, by (5.20) and (5.21),

$$\langle c^{(k)}, \psi(\beta_k x^{(k)}) - \psi(\beta_k a)\rangle/\beta_k + \langle c^{(k)}, \psi(\beta_k a)\rangle/\beta_k \leq -\delta\omega_1^2/2 < 0$$

holds for any $k \geq \bar{k}$ and $\epsilon \leq \bar{\epsilon}$. Therefore, by (5.19), we obtain

$$\langle x^{(k)}, \psi(\beta_k x^{(k)})\rangle/\beta_k \leq -\epsilon\delta\omega_1^2/2 < 0$$

for all such k's and ϵ's. Since $x^{(k)} = a + \epsilon c^{(k)} \in K$, by fixing a suitably small $\epsilon \in (0, \bar{\epsilon}]$, the above inequality contradicts (5.15) and we must have

$$\liminf_{k\to\infty} \lambda_k = \liminf_{k\to\infty} \left[\min\{\lambda_i^{(k)}(i = 1, 2, \ldots, r)\}\right] \geq 0. \tag{5.22}$$

Next we claim that $\{\lambda_i^{(k)}\}$ is bounded for any $i = 1, 2, \ldots, r$. By the facts $\beta_k a \in K$ for any k, $e \in K$, and ψ is monotone on K, we see that

$$\begin{aligned}
0 &\leq \langle \beta_k a - e, \psi(\beta_k a) - \psi(e)\rangle/\beta_k \\
&= \langle a, \psi(\beta_k a)\rangle - \langle e, \psi(\beta_k a)/\beta_k\rangle - \langle a, \psi(e)\rangle + \langle e, \psi(e)/\beta_k\rangle \\
&< -\langle e, \psi(\beta_k a)/\beta_k\rangle - \langle a, \psi(e)\rangle + \langle e, \psi(e)/\beta_k\rangle,
\end{aligned}$$

where the last inequality follows from (5.12). This implies the existence of a constant $\sigma > 0$ such that

$$\langle e, \psi(\beta_k a)/\beta_k\rangle \leq \sigma \tag{5.23}$$

holds for sufficiently large k's. By the definition (5.16), the left-hand side of the above inequality is given by

$$\langle e, \psi(\beta_k a)/\beta_k \rangle = \left\langle e, \sum_{i=1}^{r} \lambda_i^{(k)} c_i^{(k)} \right\rangle = \sum_{i=1}^{r} \lambda_i^{(k)} \langle e, c_i^{(k)} \rangle.$$

Combining the above with (5.23), we have

$$\sum_{i=1}^{r} \lambda_i^{(k)} \langle e, c_i^{(k)} \rangle \leq \sigma.$$

Here, (iv) of Proposition 2.7 ensures the existence of $\omega_1, \omega_2 > 0$ such that

$$0 < \omega_1^2 \leq \langle e, c_i^{(k)} \rangle \leq \omega_2^2 \ \text{ for all } i \text{ and } k.$$

Since we have shown the inequality (5.22), the set $\{\lambda_i^{(k)}\}$ must be bounded for any $i = 1, 2, \ldots, r$.

By using (iv) of Proposition 2.7 again, we also see that $\{c_i^{(k)}\}$ is bounded for any $i = 1, 2, \ldots, r$. Thus, $\{\psi(\beta_k a)/\beta_k = \sum_{i=1}^{r} \lambda_i^{(k)} c_i^{(k)}\}$ is bounded, and by (5.22), any accumulation point $\psi^\infty(a)$ of the sequence must satisfy $\psi^\infty(a) \in K$.

Note that $\langle a, \psi(\beta_k a) \rangle \leq -\xi$ from (5.13) and $\langle a, \psi(\beta_k a) \rangle \geq \langle a, \psi(a) \rangle$ from (5.14). Thus $\{\langle a, \psi(\beta_k a) \rangle\}$ is also bounded. To summarize, by taking an appropriate subsequence and setting

$$\begin{aligned}
&\hat{x}^{(k)} := a + (1/\beta_k) e \in \mathrm{int} K, \quad \hat{\tau}_k := 1/\beta_k, \\
&\hat{y}^{(k)} := \sum_{i=1}^{r} (\max\{\lambda_i^{(k)},\ 1/\beta_k\} c_i^{(k)}) \in \mathrm{int} K, \quad \hat{\kappa}_k := -\langle a, \psi(\beta_k a) \rangle \geq \xi > 0,
\end{aligned}$$

then, by (5.18), (5.22), and the boundedness of $\{\lambda_i^{(k)}\}$ and $\{c_i^{(k)}\}$, we have

$$\lim_{k \to \infty} \sum_{i=1}^{r} (\max\{\lambda_i^{(k)},\ 1/\beta_k\} c_i^{(k)}) = \psi^\infty(a) \in K$$

and conclude that $(x^*, \tau^*, y^*, \kappa^*)$ given by

$$\begin{aligned}
&x^* := a = \lim_{k \to \infty} \hat{x}^{(k)} \in K, \quad \tau^* := 0 = \lim_{\beta_k \to \infty} \hat{\tau}_k, \\
&y^* := \psi^\infty(a) = \lim_{k \to \infty} \hat{y}^{(k)} \in K, \quad \kappa^* := \lim_{k \to \infty} \hat{\kappa}_k \geq \xi > 0
\end{aligned}$$

is an asymptotic solution of the HCP with $\kappa^* > 0$.

Conversely, suppose that there exists a bounded sequence $(x^{(k)}, \tau_k, y^{(k)}, \kappa_k) \in (\mathrm{int} K \times \Re_{++}) \times (\mathrm{int} K \times \Re_{++})$ such that

$$\lim_{k \to \infty} y^{(k)} = \lim_{k \to \infty} \tau_k \psi(x^{(k)}/\tau_k) \in K, \quad \lim_{k \to \infty} \kappa_k = \lim_{k \to \infty} -\langle x^{(k)}, \psi(x^{(k)}/\tau_k) \rangle \geq \xi > 0.$$

Let us show that there is no feasible point $(x, y) \in K \times K$ satisfying $y - \psi(x) = 0$. Suppose that $(x, y) \in K \times K$ and $y - \psi(x) = 0$. Since ψ_{H} is monotone on

$(K \times \Re_{++}) \times (K \times \Re_+)$, by the definition (5.7), we have

$$\begin{aligned}
0 &\le \langle (x^{(k)}, \tau_k) - (x, 1), \psi_{\mathrm{H}}(x^{(k)}, \tau_k) - \psi_{\mathrm{H}}(x, 1) \rangle \\
&= \langle x^{(k)} - x, \tau_k \psi(x^{(k)}/\tau_k) - \psi(x) \rangle + (\tau_k - 1)\big(\langle x, \psi(x) \rangle - \langle x^{(k)}, \psi(x^{(k)}/\tau_k) \rangle\big) \\
&= \langle x^{(k)}, \tau_k \psi(x^{(k)}/\tau_k) \rangle + \langle x, \psi(x) \rangle \\
&\quad + (\tau_k - 1)\langle x, \psi(x) \rangle - (\tau_k - 1)\langle x^{(k)}, \psi(x^{(k)}/\tau_k) \rangle \\
&\quad - \langle x^{(k)}, \psi(x) \rangle - \langle x, \tau_k \psi(x^{(k)}/\tau_k) \rangle \\
&= \tau_k \langle x, \psi(x) \rangle + \langle x^{(k)}, \psi(x^{(k)}/\tau_k) \rangle - \langle x^{(k)}, \psi(x) \rangle - \langle x, \tau_k \psi(x^{(k)}/\tau_k) \rangle
\end{aligned}$$

and hence

$$\begin{aligned}
\langle x^{(k)}, \psi(x^{(k)}/\tau_k) \rangle &\ge \langle x^{(k)}, \psi(x) \rangle + \langle x, \tau_k \psi(x^{(k)}/\tau_k) \rangle - \tau_k \langle x, \psi(x) \rangle \\
&= \langle x^{(k)}, y \rangle + \langle x, \tau_k \psi(x^{(k)}/\tau_k) \rangle - \tau_k \langle x, y \rangle. \qquad (5.24)
\end{aligned}$$

Here $\lim_{k\to\infty} \tau_k = 0$ since $\lim_{k\to\infty} \kappa_k \ge \xi > 0$. In addition, it follows from the assumption that $\langle x^{(k)}, y \rangle \ge 0$ and that

$$\lim_{k\to\infty} \langle x, y^{(k)} \rangle = \langle x, \lim_{k\to\infty} y^{(k)} \rangle = \langle x, \lim_{k\to\infty} \tau_k \psi(x^{(k)}/\tau_k) \rangle \ge 0.$$

Thus the relation (5.24) ensures that

$$\lim_{k\to\infty} \langle x^{(k)}, \psi(x^{(k)}/\tau_k) \rangle \ge 0,$$

which contradicts

$$\kappa_k := -\langle x^{(k)}, \psi(x^{(k)}/\tau_k) \rangle \ge \xi > 0.$$

In addition, any limit of $x^{(k)}$ gives a separating hyperplane, i.e., a certificate proving infeasibility. □

Note that the Lipschitz continuity of ψ on K in (v) of the above theorem seems to be indispensable for showing the existence of a uniform bound $\bar{\gamma}$ satisfying (5.21), while only the differentiability (or the scaled Lipschitz continuity) of ψ is assumed in [2]. The differentiability of ψ on K implies the Lipschitz continuity of ψ at $a \in K$. However, the local Lipschitz continuity of ψ might be not enough since the sequence $\{\beta_k a\} \subseteq K$ is never bounded along $\beta_k \to +\infty$.

We are going to show that a central path of the homogeneous model HCP is well defined. As we will see in (iii) of Theorem 5.5, any limit point of the path lets us know if the HCP has an asymptotically complementary solution $(x_{\mathrm{H}}^*, y_{\mathrm{H}}^*) = (x^*, \tau^*, y^*, \kappa^*)$ with $\tau^* > 0$ or if it has such a solution with $\kappa^* > 0$.

Therefore, in view of (iv) and (v) of Theorem 5.4, if we find a limit of the path then we can determine whether the CP is (asymptotically) solvable, strongly infeasible, or of some other type.

Let us consider the map

$$H_{\mathrm{H}}(x_{\mathrm{H}} y_{\mathrm{H}}) := \begin{pmatrix} x_{\mathrm{H}} \circ_{\mathrm{H}} y_{\mathrm{H}} \\ F_{\mathrm{H}}(x_{\mathrm{H}}, y_{\mathrm{H}}) \end{pmatrix} \qquad (5.25)$$

and choose an initial point $(x_{\mathrm{H}}^{(0)}, y_{\mathrm{H}}^{(0)})$ such that $(x_{\mathrm{H}}^{(0)}, y_{\mathrm{H}}^{(0)}) \in \mathrm{int}K_{\mathrm{H}} \times \mathrm{int}K_{\mathrm{H}}$ and $x_{\mathrm{H}} \circ_{\mathrm{H}} y_{\mathrm{H}} \in \mathrm{int}K_{\mathrm{H}}$. For simplicity, we set $(x_{\mathrm{H}}^{(0)}, y_{\mathrm{H}}^{(0)}) = (x^{(0)}, \tau_0, y^{(0)}, \kappa_0) = (e, 1, e, 1) \in \mathrm{int}K_{\mathrm{H}} \times \mathrm{int}K_{\mathrm{H}}$. Define

$$h_{\mathrm{H}}^{(0)} := \begin{pmatrix} p_{\mathrm{H}}^{(0)} \\ f_{\mathrm{H}}^{(0)} \end{pmatrix} := \begin{pmatrix} x_{\mathrm{H}}^{(0)} \circ_{\mathrm{H}} y_{\mathrm{H}}^{(0)} \\ F_{\mathrm{H}}(x_{\mathrm{H}}^{(0)}, y_{\mathrm{H}}^{(0)}) \end{pmatrix} = \begin{pmatrix} e_{\mathrm{H}} \\ y_{\mathrm{H}}^{(0)} - \psi_{\mathrm{H}}(x_{\mathrm{H}}^{(0)}) \end{pmatrix}, \qquad (5.26)$$

where $e_{\mathrm{H}} = (e, 1) \in \mathrm{int}K_{\mathrm{H}}$ is the identity element in V_{H}.

THEOREM 5.5 (cf. Theorem 2 of [2]). *Suppose that $\psi : K \to V$ satisfies Assumption* 5.1. *Define $h_{\mathrm{H}}^{(0)}$ by* (5.26).

(i) *For any $t \in (0,1]$, there exists a point $(x_{\mathrm{H}}(t), y_{\mathrm{H}}(t)) \in \mathrm{int}K_{\mathrm{H}} \times \mathrm{int}K_{\mathrm{H}}$ such that $H_{\mathrm{H}}(x_{\mathrm{H}}(t), y_{\mathrm{H}}(t)) = th_{\mathrm{H}}^{(0)}$.*

(ii) *The set*

$$P := \{(x_{\mathrm{H}}, y_{\mathrm{H}}) \in \mathrm{int}K_{\mathrm{H}} \times \mathrm{int}K_{\mathrm{H}} : \ H_{\mathrm{H}}(x_{\mathrm{H}}(t), y_{\mathrm{H}}(t)) = th_{\mathrm{H}}^{(0)}, \ t \in (0,1]\}$$

forms a bounded path in $\mathrm{int}K_{\mathrm{H}} \times \mathrm{int}K_{\mathrm{H}}$. Any accumulation point $(x_{\mathrm{H}}(0), y_{\mathrm{H}}(0))$ is an asymptotically complementary solution of the HCP.

(iii) *If the HCP has an asymptotically complementary solution $(x_{\mathrm{H}}^*, y_{\mathrm{H}}^*) = (x^*, \tau^*, y^*, \kappa^*)$ with $\tau^* > 0$ ($\kappa^* > 0$, respectively), then any accumulation point $(x_{\mathrm{H}}(0), y_{\mathrm{H}}(0)) = (x(0), \tau(0), y(0), \kappa(0))$ of the bounded path P satisfies $\tau(0) > 0$ ($\kappa(0) > 0$, respectively).*

Proof. (i) It follows from Proposition 5.3 that the map F_{H} defined by (5.3) satisfies Assumptions 1.1 and 3.11. Thus, by Theorem 3.12, the set $H_{\mathrm{H}}(\mathcal{U}_{\mathrm{H}})$ with

$$\mathcal{U}_{\mathrm{H}} := \{(x_{\mathrm{H}}, y_{\mathrm{H}}) \in \mathrm{int}K_{\mathrm{H}} \times \mathrm{int}K_{\mathrm{H}} : \ x_{\mathrm{H}} \circ_{\mathrm{H}} y_{\mathrm{H}} \in \mathrm{int}K_{\mathrm{H}}\}$$

is an open convex subset of $V_{\mathrm{H}} \times V_{\mathrm{H}}$. Here we have already seen that $0 \in \mathrm{cl}(H_{\mathrm{H}}(\mathcal{U}_{\mathrm{H}}))$ in (ii) and (iii) of Theorem 5.4.

Since the set $H_{\mathrm{H}}(\mathcal{U}_{\mathrm{H}})$ is convex, the fact above implies that $th_{\mathrm{H}}^{(0)} \in H_{\mathrm{H}}(\mathcal{U}_{\mathrm{H}})$ for any $t \in (0,1]$. Combining this with the homeomorphism of the map H_{H} in Theorem 3.10, we obtain the assertion (i).

(ii) The homeomorphism of the map H_{H} also ensures that the set P forms a path in $\mathrm{int}K_{\mathrm{H}} \times \mathrm{int}K_{\mathrm{H}}$. It suffices to show that the path P is bounded.

Let $(x_{\mathrm{H}}, y_{\mathrm{H}}) = (x, \tau, y, \kappa) \in P$. Then there exists a $t \in (0,1]$ for which

$$x_{\mathrm{H}} \circ_{\mathrm{H}} y_{\mathrm{H}} = te_{\mathrm{H}} \ \text{ and } \ y_{\mathrm{H}} - \psi_{\mathrm{H}}(x_{\mathrm{H}}) = tf_{\mathrm{H}}^{(0)} \tag{5.27}$$

hold and

$$\begin{aligned}
&\langle x_{\mathrm{H}}, f_{\mathrm{H}}^{(0)} \rangle_{\mathrm{H}} \\
&= \langle x_{\mathrm{H}}, y_{\mathrm{H}}^{(0)} \rangle_{\mathrm{H}} - \langle x_{\mathrm{H}}, \psi_{\mathrm{H}}(x_{\mathrm{H}}^{(0)}) \rangle_{\mathrm{H}} && \text{(by (5.26))} \\
&= \langle x_{\mathrm{H}}, y_{\mathrm{H}}^{(0)} \rangle_{\mathrm{H}} + \langle y_{\mathrm{H}}, x_{\mathrm{H}}^{(0)} \rangle_{\mathrm{H}} - \langle y_{\mathrm{H}}, x_{\mathrm{H}}^{(0)} \rangle_{\mathrm{H}} - \langle x_{\mathrm{H}}, \psi_{\mathrm{H}}(x_{\mathrm{H}}^{(0)}) \rangle_{\mathrm{H}} \\
&= \langle x_{\mathrm{H}}, y_{\mathrm{H}}^{(0)} \rangle_{\mathrm{H}} + \langle y_{\mathrm{H}}, x_{\mathrm{H}}^{(0)} \rangle_{\mathrm{H}} - \langle x_{\mathrm{H}}^{(0)}, tf_{\mathrm{H}}^{(0)} + \psi_{\mathrm{H}}(x_{\mathrm{H}}) \rangle_{\mathrm{H}} - \langle x_{\mathrm{H}}, \psi_{\mathrm{H}}(x_{\mathrm{H}}^{(0)}) \rangle_{\mathrm{H}} && \text{(by (5.27))} \\
&= \langle x_{\mathrm{H}}, y_{\mathrm{H}}^{(0)} \rangle_{\mathrm{H}} + \langle y_{\mathrm{H}}, x_{\mathrm{H}}^{(0)} \rangle_{\mathrm{H}} - t\langle x_{\mathrm{H}}^{(0)}, f_{\mathrm{H}}^{(0)} \rangle_{\mathrm{H}} - \langle x_{\mathrm{H}}^{(0)}, \psi_{\mathrm{H}}(x_{\mathrm{H}}) \rangle_{\mathrm{H}} - \langle x_{\mathrm{H}}, \psi_{\mathrm{H}}(x_{\mathrm{H}}^{(0)}) \rangle_{\mathrm{H}} \\
&\geq \langle x_{\mathrm{H}}, y_{\mathrm{H}}^{(0)} \rangle_{\mathrm{H}} + \langle y_{\mathrm{H}}, x_{\mathrm{H}}^{(0)} \rangle_{\mathrm{H}} - t\langle x_{\mathrm{H}}^{(0)}, f_{\mathrm{H}}^{(0)} \rangle_{\mathrm{H}} - \langle x_{\mathrm{H}}, \psi_{\mathrm{H}}(x_{\mathrm{H}}) \rangle_{\mathrm{H}} - \langle x_{\mathrm{H}}^{(0)}, \psi_{\mathrm{H}}(x_{\mathrm{H}}^{(0)}) \rangle_{\mathrm{H}} \\
&\qquad \text{(by the monotonicity of } \psi_{\mathrm{H}}) \\
&= \langle x_{\mathrm{H}}, y_{\mathrm{H}}^{(0)} \rangle_{\mathrm{H}} + \langle y_{\mathrm{H}}, x_{\mathrm{H}}^{(0)} \rangle_{\mathrm{H}} - t\langle x_{\mathrm{H}}^{(0)}, f_{\mathrm{H}}^{(0)} \rangle_{\mathrm{H}} && \text{(by (i) of Theorem 5.4)} \\
&= \langle x_{\mathrm{H}}, y_{\mathrm{H}}^{(0)} \rangle_{\mathrm{H}} + \langle y_{\mathrm{H}}, x_{\mathrm{H}}^{(0)} \rangle_{\mathrm{H}} - t\langle x_{\mathrm{H}}^{(0)}, y_{\mathrm{H}}^{(0)} - \psi_{\mathrm{H}}(x_{\mathrm{H}}^{(0)}) \rangle_{\mathrm{H}} && \text{(by (5.26))} \\
&= \langle x_{\mathrm{H}}, y_{\mathrm{H}}^{(0)} \rangle_{\mathrm{H}} + \langle y_{\mathrm{H}}, x_{\mathrm{H}}^{(0)} \rangle_{\mathrm{H}} - t\langle x_{\mathrm{H}}^{(0)}, y_{\mathrm{H}}^{(0)} \rangle_{\mathrm{H}} && \text{(by (i) of Theorem 5.4).}
\end{aligned}$$

In addition, for the same $t \in (0,1]$, we have

$$\begin{aligned} t\langle x_{\mathrm{H}}, f_{\mathrm{H}}^{(0)}\rangle_{\mathrm{H}} &= \langle x_{\mathrm{H}}, tf_{\mathrm{H}}^{(0)}\rangle_{\mathrm{H}} \\ &= \langle x_{\mathrm{H}}, y_{\mathrm{H}} - \psi_{\mathrm{H}}(x_{\mathrm{H}})\rangle_{\mathrm{H}} \qquad \text{(by (5.27))} \\ &= \langle x_{\mathrm{H}}, y_{\mathrm{H}}\rangle_{\mathrm{H}} \qquad \text{(by (i) of Theorem 5.4)} \\ &= t\langle e_{\mathrm{H}}, e_{\mathrm{H}}\rangle_{\mathrm{H}} \qquad \text{(by (5.27))} \\ &= t\langle x_{\mathrm{H}}^{(0)}, y_{\mathrm{H}}^{(0)}\rangle_{\mathrm{H}}. \end{aligned}$$

Therefore, we obtain that

$$\begin{aligned} \langle x_{\mathrm{H}}, y_{\mathrm{H}}^{(0)}\rangle_{\mathrm{H}} + \langle y_{\mathrm{H}}, x_{\mathrm{H}}^{(0)}\rangle_{\mathrm{H}} &\le \langle x_{\mathrm{H}}, f_{\mathrm{H}}^{(0)}\rangle_{\mathrm{H}} + t\langle x_{\mathrm{H}}^{(0)}, y_{\mathrm{H}}^{(0)}\rangle_{\mathrm{H}} \\ &= \langle x_{\mathrm{H}}^{(0)}, y_{\mathrm{H}}^{(0)}\rangle_{\mathrm{H}} + t\langle x_{\mathrm{H}}^{(0)}, y_{\mathrm{H}}^{(0)}\rangle_{\mathrm{H}} \\ &= (1+t)\langle x_{\mathrm{H}}^{(0)}, y_{\mathrm{H}}^{(0)}\rangle_{\mathrm{H}} = (1+t)\langle e_{\mathrm{H}}, e_{\mathrm{H}}\rangle_{\mathrm{H}} \le 2\langle e_{\mathrm{H}}, e_{\mathrm{H}}\rangle_{\mathrm{H}}. \end{aligned}$$

Thus, by (ii) of Proposition 2.1, the set P is bounded.

(iii) Let $(x_{\mathrm{H}}^*, y_{\mathrm{H}}^*) = (x^*, \tau^*, y^*, \kappa^*)$ be an asymptotic solution for the HCP. Then there exists a bounded sequence

$$\{(x_{\mathrm{H}}^{(k)}, y_{\mathrm{H}}^{(k)})\} = \{(x^{(k)}, \tau_k, y^{(k)}, \kappa_k)\} \subseteq \mathrm{int}K_{\mathrm{H}} \times \mathrm{int}K_{\mathrm{H}}$$

such that

$$\lim_{k\to\infty}(x_{\mathrm{H}}^{(k)}, y_{\mathrm{H}}^{(k)}) = (x_{\mathrm{H}}^*, y_{\mathrm{H}}^*), \quad \lim_{k\to\infty} y_{\mathrm{H}}^{(k)} - \psi_{\mathrm{H}}(x_{\mathrm{H}}^{(k)}) = 0, \quad \lim_{k\to\infty} x_{\mathrm{H}}^{(k)} \circ_{\mathrm{H}} y_{\mathrm{H}}^{(k)} = 0.$$

Let $(x_{\mathrm{H}}(t), y_{\mathrm{H}}(t)) = (x(t), \tau(t), y(t), \kappa(t))$ be any point on the path P. Then,

$$x_{\mathrm{H}}(t) \circ_{\mathrm{H}} y_{\mathrm{H}}(t) = te_{\mathrm{H}} \ \text{ and } \ y_{\mathrm{H}}(t) - \psi_{\mathrm{H}}(x_{\mathrm{H}}(t)) = tf_{\mathrm{H}}^{(0)}. \tag{5.28}$$

By the boundedness of the set P (see (ii) above), there exists an $\epsilon \in (0,1]$ such that

$$\|x_{\mathrm{H}}(t)\| \le 1/\epsilon \ \text{ and } \ \|y_{\mathrm{H}}(t)\| \le 1/\epsilon \tag{5.29}$$

hold for any $t \in (0,1]$. In addition, for each $t \in (0,1]$, there exists an index $k(t)$ such that for any $k \ge k(t)$, we have

$$\|x_{\mathrm{H}}^{(k)} - x_{\mathrm{H}}^*\| \le \epsilon, \ \ \|y_{\mathrm{H}}^{(k)} - y_{\mathrm{H}}^*\| \le \epsilon, \text{ and } \|y_{\mathrm{H}}^{(k)} - \psi_{\mathrm{H}}(x_{\mathrm{H}}^{(k)})\| \le t\epsilon. \tag{5.30}$$

Here, by the monotonicity of ψ_{H},

$$\begin{aligned} &\langle x_{\mathrm{H}}(t) - x_{\mathrm{H}}^{(k)}, y_{\mathrm{H}}(t) - y_{\mathrm{H}}^{(k)}\rangle_{\mathrm{H}} \\ &\quad = \langle x_{\mathrm{H}}(t) - x_{\mathrm{H}}^{(k)}, \psi_{\mathrm{H}}(x_{\mathrm{H}}(t)) - \psi_{\mathrm{H}}(x_{\mathrm{H}}^{(k)})\rangle_{\mathrm{H}} \\ &\qquad + \langle x_{\mathrm{H}}(t) - x_{\mathrm{H}}^{(k)}, y_{\mathrm{H}}(t) - \psi_{\mathrm{H}}(x_{\mathrm{H}}(t))\rangle_{\mathrm{H}} - \langle x_{\mathrm{H}}(t) - x_{\mathrm{H}}^{(k)}, y_{\mathrm{H}}^{(k)} - \psi_{\mathrm{H}}(x_{\mathrm{H}}^{(k)})\rangle_{\mathrm{H}} \\ &\quad \ge \langle x_{\mathrm{H}}(t) - x_{\mathrm{H}}^{(k)}, y_{\mathrm{H}}(t) - \psi_{\mathrm{H}}(x_{\mathrm{H}}(t))\rangle_{\mathrm{H}} - \langle x_{\mathrm{H}}(t) - x_{\mathrm{H}}^{(k)}, y_{\mathrm{H}}^{(k)} - \psi_{\mathrm{H}}(x_{\mathrm{H}}^{(k)})\rangle_{\mathrm{H}} \end{aligned}$$

and hence, for any $t \in (0,1]$ and any $k \geq k(t)$,

$$
\begin{aligned}
&\langle x_{\mathrm{H}}(t), y_{\mathrm{H}}^{(k)}\rangle_{\mathrm{H}} + \langle y_{\mathrm{H}}(t), x_{\mathrm{H}}^{(k)}\rangle_{\mathrm{H}} \\
&\leq \langle x_{\mathrm{H}}(t), y_{\mathrm{H}}(t)\rangle_{\mathrm{H}} + \langle x_{\mathrm{H}}^{(k)}, y_{\mathrm{H}}^{(k)}\rangle_{\mathrm{H}} \\
&\quad - \langle x_{\mathrm{H}}(t) - x_{\mathrm{H}}^{(k)}, y_{\mathrm{H}}(t) - \psi_{\mathrm{H}}(x_{\mathrm{H}}(t))\rangle_{\mathrm{H}} \\
&\quad + \langle x_{\mathrm{H}}(t) - x_{\mathrm{H}}^{(k)}, y_{\mathrm{H}}^{(k)} - \psi_{\mathrm{H}}(x_{\mathrm{H}}^{(k)})\rangle_{\mathrm{H}} \\
&= \langle x_{\mathrm{H}}(t), y_{\mathrm{H}}(t)\rangle_{\mathrm{H}} + \langle x_{\mathrm{H}}^{(k)}, y_{\mathrm{H}}^{(k)}\rangle_{\mathrm{H}} \\
&\quad - \langle x_{\mathrm{H}}(t), y_{\mathrm{H}}(t) - \psi_{\mathrm{H}}(x_{\mathrm{H}}(t))\rangle_{\mathrm{H}} - \langle x_{\mathrm{H}}^{(k)}, y_{\mathrm{H}}^{(k)} - \psi_{\mathrm{H}}(x_{\mathrm{H}}^{(k)})\rangle_{\mathrm{H}} \\
&\quad + \langle x_{\mathrm{H}}^{(k)}, y_{\mathrm{H}}(t) - \psi_{\mathrm{H}}(x_{\mathrm{H}}(t))\rangle_{\mathrm{H}} + \langle x_{\mathrm{H}}(t), y_{\mathrm{H}}^{(k)} - \psi_{\mathrm{H}}(x_{\mathrm{H}}^{(k)})\rangle_{\mathrm{H}} \\
&= \langle x_{\mathrm{H}}^{(k)}, y_{\mathrm{H}}(t) - \psi_{\mathrm{H}}(x_{\mathrm{H}}(t))\rangle_{\mathrm{H}} + \langle x_{\mathrm{H}}(t), y_{\mathrm{H}}^{(k)} - \psi_{\mathrm{H}}(x_{\mathrm{H}}^{(k)})\rangle_{\mathrm{H}} \qquad \text{(by (i) of Theorem 5.4)} \\
&= \langle x_{\mathrm{H}}^{(k)}, t f_{\mathrm{H}}^{(0)}\rangle_{\mathrm{H}} + \langle x_{\mathrm{H}}(t), y_{\mathrm{H}}^{(k)} - \psi_{\mathrm{H}}(x_{\mathrm{H}}^{(k)})\rangle_{\mathrm{H}} \qquad \text{(by (5.28))} \\
&\leq t\|x_{\mathrm{H}}^{(k)}\|\|f_{\mathrm{H}}^{(0)}\| + \|x_{\mathrm{H}}(t)\|\|y_{\mathrm{H}}^{(k)} - \psi_{\mathrm{H}}(x_{\mathrm{H}}^{(k)})\| \\
&\leq t(\|x_{\mathrm{H}}^{*}\| + \epsilon)\|h_{\mathrm{H}}^{(0)}\| + t \qquad \text{(by (5.29) and (5.30))} \\
&\leq t\delta,
\end{aligned}
$$

where $\delta := 1 + (\|x_{\mathrm{H}}^{*}\| + 1)\|h_{\mathrm{H}}^{(0)}\| > 0$. Note that (5.28) implies $x_{\mathrm{H}}(t) = t y_{\mathrm{H}}(t)^{-1}$ and $y_{\mathrm{H}}(t) = t x_{\mathrm{H}}(t)^{-1}$. Combining them, it must hold that for any $t \in (0,1]$ and $k \geq k(t)$,

$$
\begin{aligned}
t\delta &\geq \langle x_{\mathrm{H}}(t), y_{\mathrm{H}}^{(k)}\rangle_{\mathrm{H}} + \langle y_{\mathrm{H}}(t), x_{\mathrm{H}}^{(k)}\rangle_{\mathrm{H}} \\
&= \langle t y_{\mathrm{H}}(t)^{-1}, y_{\mathrm{H}}^{(k)}\rangle_{\mathrm{H}} + \langle t x_{\mathrm{H}}(t)^{-1}, x_{\mathrm{H}}^{(k)}\rangle_{\mathrm{H}} \\
&= t\left\{\langle y(t)^{-1}, y^{(k)}\rangle + \frac{\kappa_k}{\kappa(t)} + \langle x(t)^{-1}, x^{(k)}\rangle + \frac{\tau_k}{\tau(t)}\right\}.
\end{aligned}
$$

Since $\langle y(t)^{-1}, y^{(k)}\rangle > 0$ and $\langle x(t)^{-1}, x^{(k)}\rangle > 0$, we finally obtain that

$$
\frac{\kappa_k}{\kappa(t)} < \delta \quad \text{and} \quad \frac{\tau_k}{\tau(t)} < \delta
$$

for any $t \in (0,1]$ and $k \geq k(t)$. Thus, the assertion (iii) follows from the facts $\kappa_k \to \kappa^*$, $\tau_k \to \tau^*$, and $\delta > 0$. □

6. Concluding remarks. In this paper, we studied the homeomorphism of the interior point map defined by (1.4) for monotone complementarity problems (CPs) over symmetric cones associated with Euclidean Jordan algebras. As an application of our results, we provided a homogeneous model (HCP) for the problems and showed the existence of a trajectory. In Theorem 5.4, we have shown that the following implications hold under the Lipschitz continuity of the map:

[The original CP has a solution.]
$\Longleftrightarrow$ [The HCP has an asymptotic solution $(x^*, \tau^*, y^*, \kappa^*)$ with $\tau^* > 0$.]
[The original CP is strongly infeasible.]
$\Longrightarrow$ [The HCP has an asymptotic solution $(x^*, \tau^*, y^*, \kappa^*)$ with $\kappa^* > 0$.]
$\Longrightarrow$ [The original CP is infeasible (with a finite certification of the infeasibility).]

To our knowledge, no better results have been provided even for the case of $K = \Re^n_{++}$. It is still an open problem to find an exact certification proving strong infeasibility of the original problem.

Other issues to be investigated are the development of numerical algorithms and their complexity analyses and discussing the existence of maximal complementarity solutions for the problems. Concerning the second subject, we have already obtained a partial result in the proof of (iii) of Theorem 5.5. As a related result, we should refer to Chua's work [3] for homogeneous conic programming: The author showed that the paths defined by a class of optimal barriers converge to analytical centers of optimal faces whenever the primal-dual pair of problems has strictly complementarity solutions.

Recently, the concept of P-properties on Euclidean Jordan algebra was introduced by Gowda, Sznajder, and Tao [8] and by Tao and Gowda [31], aiming to provide nonmonotone properties on the algebra. For the case of the n-dimensional positive orthant, a homogeneous model for P_0 complementarity problems has been proposed in [33]. In contrast to our results, however, the lack of the monotonicity of the map prevents us from obtaining the convexity of the image as in Theorem 3.12. The convexity of the set is a key property in obtaining a certificate proving strong infeasibility of the original problem.

Acknowledgments. The author would like to thank Leonid Faybusovich for many helpful comments and valuable suggestions. The author is also grateful to the anonymous referees whose comments helped improve the paper considerably.

REFERENCES

[1] F. Alizadeh and S. H. 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, MA, 2000, pp. 195–233.

[2] E. Andersen and Y. Ye, *On a homogeneous algorithm for the monotone complementarity problems*, Math. Program., 84 (1999), pp. 375–400.

[3] C. B. Chua, *A New Notion of Weighted Centers for Semidefinite Programming*, Department of Combinatorics and Optimization, University of Waterloo, Waterloo, Canada, 2004.

[4] E. de Klerk, C. Roos, and T. Terlaky, *Initialization in semidefinite programming via a self-dual skew-symmetric embedding*, Oper. Res. Lett., 20 (1997), pp. 213–221.

[5] J. Faraut and A. Korányi, *Analysis on Symmetric Cones*, Oxford Science Publishers, Oxford, UK, 1994.

[6] L. Faybusovich, *Euclidean Jordan algebras and interior-point algorithms*, Positivity, 1 (1997), pp. 331–357.

[7] L. Faybusovich, *Linear systems in Jordan algebras and primal-dual interior point algorithms*, J. Comput. Appl. Math., 86 (1997), pp. 149–175.

[8] M. S. Gowda, R. Sznajder, and J. Tao, *Some P-properties for linear transformations on Euclidean Jordan algebras*, Linear Algebra Appl., 393 (2004), pp. 203–232.

[9] O. Güler, *Existence of interior points and interior paths in nonlinear monotone complementarity problems*, Math. Oper. Res., 18 (1993), pp. 128–148.

[10] O. Güler, *Barrier functions in interior-point methods*, Math. Oper. Res., 21 (1996), pp. 860–885.

[11] R. A. Hauser and O. Güler, *Self-scaled barrier functions on symmetric cones and their classification*, Found. Comput. Math., 2 (2002), pp. 121–143.

[12] S. Hayashi, N. Yamashita, and M. Fukushima, *Robust Nash equilibria and second-order cone complementarity problems*, J. Nonlinear Convex Anal., 6 (2005), pp. 283–296.

[13] M. Kojima, N. Megiddo, and T. Noma, *Homotopy continuation methods for nonlinear complementarity problems*, Math. Oper. Res., 16 (1991), pp. 754–774.

[14] M. Kojima, S. Mizuno, and A. Yoshise, *A primal-dual interior point algorithm for linear programming*, in Progress in Mathematical Programming, N. Megiddo, ed., Springer-Verlag, New York, 1989, pp. 29–47.

[15] 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.

[16] Z.-Q. Luo, J. F. Sturm, and S. Zhang, *Conic convex programming and self-dual embedding*, Optim. Methods Softw., 14 (2000), pp. 196–218.
[17] R.D.C. Monteiro and I. Adler, *Interior path following primal-dual algorithms.* I. *Linear programming*, Math. Programming, 44 (1989), pp. 27–41.
[18] R.D.C. Monteiro and J.-S. Pang, *Properties of an interior-point mapping for mixed complementarity problems*, Math. Oper. Res., 21 (1996), pp. 629–654.
[19] R.D.C. Monteiro and J.-S. Pang, *On two interior-point mappings for nonlinear semidefinite complementarity problems*, Math. Oper. Res., 23 (1998), pp. 39–60.
[20] R.D.C. Monteiro and P. Zanjacomo, *General interior-point maps and existence of weighted paths for nonlinear semidefinite complementarity problems*, Math. Oper. Res., 25 (2000), pp. 381–399.
[21] Yu. E. Nesterov and M. J. Todd, *Self-scaled barriers and interior-point for convex programming*, Math. Oper. Res., 22 (1997), pp. 1–42.
[22] Yu. E. Nesterov and M. J. Todd, *Primal-dual interior-point methods for self-scaled cones*, SIAM J. Optim., 8 (1998), pp. 324–364.
[23] Yu. 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.
[24] B. K. Rangarajan, *Polynomial convergence of infeasible-interior-point methods over symmetric cones*, SIAM J. Optim., 16 (2006), pp. 1211–1229.
[25] B. K. Rangarajan and M. J. Todd, *Convergence of Infeasible-Interior-Point Methods for Self-Scaled Conic Programming*, Technical report TR1388, Department of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY, 2003.
[26] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer-Verlag, New York, 1998.
[27] S. H. Schmieta, *Complete Classification of Self-Scaled Barrier Functions*, Technical report, Department of IEOR, Columbia University, New York, NY, 2000.
[28] S. H. Schmieta and F. Alizadeh, *Extension of primal-dual interior-point algorithm to symmetric cones*, Math. Program., 96 (2003), pp. 409–438.
[29] M. Shida, S. Shindoh, and M. Kojima, *Centers of monotone generalized complementarity problems*, Math. Oper. Res., 22 (1997), pp. 969–976.
[30] J. F. Sturm, *Using SeDuMi* 1.02*, a MATLAB toolbox for optimization over symmetric cones*, Optim. Methods Softw., 11–12 (1999), pp. 625-653.
[31] J. Tao and M. S. Gowda, *Some P-Properties for Nonlinear Transformations on Euclidean Jordan Algebras*, Technical report TRGOW04-01, Department of Mathematics and Statistics, University of Maryland, Baltimore, MD, 2004.
[32] Y. Ye, *On homogeneous and self-dual algorithms for LCP*, Math. Programming, 76 (1997), pp. 211–221.
[33] A. Yoshise, *A Homogeneous Model for P_0 and P_* Nonlinear Complementarity Problems*, DPS 1059, Institute of Policy and Planning Sciences, The University of Tsukuba, Tsukuba, Japan, 2003.

SIAM J. OPTIM.
Vol. 17, No. 4, pp. 1154–1175
© 2006 Society for Industrial and Applied Mathematics

THE LAGRANGE MULTIPLIER RULE FOR MULTIFUNCTIONS IN BANACH SPACES*

XI YIN ZHENG† AND KUNG FU NG‡

Abstract. We study general constrained multiobjective optimization problems with objectives being closed multifunctions in Banach spaces. In terms of the coderivatives and normal cones, we provide generalized Lagrange multiplier rules as necessary optimality conditions of the above problems. In an Asplund space setting, sharper results are presented.

Key words. multifunction, normal cone, coderivative, Pareto solution

AMS subject classifications. 49J52, 90C29

DOI. 10.1137/060651860

1. Introduction.

Let X be a Banach space and $f_i : X \to R \cup \{+\infty\}$ be proper lower semicontinuous functions $(i = 0, 1, \ldots, m)$. Many authors (see [2, 3, 4, 16, 29, 30]) studied the following optimization problem with inequality and equality constraints:

$$\begin{aligned} &\min f_0(x), \\ &f_i(x) \le 0, \quad i = 1, \ldots, n, \\ &f_i(x) = 0, \quad i = n+1, \ldots, m, \\ &x \in \Omega. \end{aligned} \tag{1.1}$$

Under some restricted conditions (e.g., each f_i is locally Lipschitz), it is well known, as the Lagrange multiplier rule, that if $\bar{x}$ is a local solution of (1.1), then there exists $\lambda_i \in R$ $(0 \le i \le m)$ such that

$$\begin{aligned} &0 \in \sum_{i=0}^{m} \partial(\lambda_i f_i)(\bar{x}) + N(\Omega, \bar{x}), \\ &\sum_{i=0}^{m} |\lambda_i| = 1 \ \text{ and } \ \lambda_i \ge 0, \quad 0 \le i \le n, \end{aligned} \tag{1.2}$$

where $\partial(\lambda_i f_i)$ and $N(\Omega, \bar{x})$ denote the subdifferential and the normal cone (see section 2 for their definitions). Some authors established the so-called fuzzy Lagrange multiplier rule (see [3, 14, 20] and the references therein). The main aim of this paper is to establish the corresponding rules for multifunctions in Banach spaces.

Let X, $Y_0, Y_1, \ldots, Y_m$ be Banach spaces, Ω be a closed subset of X, and $F_i : X \to 2^{Y_i}$ $(i = 0, 1, \ldots, m)$ be closed multifunctions. Let $C_0 \subset Y_0$ be a closed convex cone

*Received by the editors February 10, 2006; accepted for publication (in revised form) May 26, 2006; published electronically December 26, 2006.
http://www.siam.org/journals/siopt/17-4/65186.html

†Department of Mathematics, Yunnan University, Kunming 650091, People's Republic of China (xyzheng@ynu.edu.cn). The research of this author was supported by the National Natural Science Foundation of the People's Republic of China (grant 10361008) and the Natural Science Foundation of Yunnan Province, China (grant 2003A0002M).

‡Department of Mathematics, The Chinese University of Hong Kong, Shatin, New Territory, Hong Kong (kfng@math.cuhk.edu.hk). The research of this author was supported by a direct grant (CUHK) and an earmarked grant from the Research Grant Council of Hong Kong.

such that $C_0 \neq C_0 \cap -C_0$ (i.e, C_0 is not a linear subspace), which specifies a preorder $\leq_{C_0}$ on Y_0 as follows: for $y_1, y_2 \in Y_0$,

$$y_1 \leq_{C_0} y_2 \text{ if and only if } y_2 - y_1 \in C_0.$$

For $i = 1, \ldots, m$, let C_i be a closed convex cone in Y_i. Consider the following constrained multiobjective optimization problem:

$$\begin{aligned} & C_0 - \min F_0(x), \\ & F_i(x) \cap -C_i \neq \emptyset, \quad i = 1, \ldots, m, \\ & x \in \Omega. \end{aligned} \tag{1.3}$$

Recall that $\bar{a} \in A$ is said to be a Pareto efficient point if $\bar{a} \leq_{C_0} a$ whenever $a \in A$ and $a \leq_{C_0} \bar{a}$, that is,

$$A \cap (\bar{a} - C_0) \subset \bar{a} + C_0 \cap -C_0.$$

We use $\mathrm{E}(A, C_0)$ to denote the set of all Pareto efficient points of A. In the case when C_0 is pointed (i.e., $C_0 \cap -C_0 = \{0\}$),

$$\bar{a} \in E(A, C_0) \Longleftrightarrow A \cap (\bar{a} - C_0) = \{\bar{a}\}.$$

For $\bar{x} \in X$ and $\bar{y} \in F_0(\bar{x})$, we say that $(\bar{x}, \bar{y})$ is a local Pareto solution of the multiobjective optimization problem (1.3) if there exists a neighborhood U of $\bar{x}$ such that

$$\bar{y} \in \mathrm{E}\left(F_0\left[U \cap \Omega \cap \left(\bigcap_{i=1}^{m} F_i^{-1}(-C_i)\right)\right], C_0\right).$$

In the case when each F_i is single-valued, many authors have established sufficient or necessary optimality conditions for Pareto solutions and weak Pareto solutions under some restricted conditions; e.g., the ordering cone has a nonempty interior, the spaces are finite dimensional, and $C_i = R_+^n$ (see [1, 5, 7, 9, 10, 11, 12, 13, 22, 23, 24, 26, 27] and the references therein). In the set-valued setting, in terms of cotangent derivatives Götz and Jahn [8] provided the Lagrange multiplier rule for (1.3) under the convexity assumption. Ye and Zhu [25] and Mordukhovich, Treiman, and Zhu [19] gave some necessary optimality conditions for multiobjective optimization problems with respect to an abstract order in a Euclidean space or Asplund space setting. Recently, the authors [28] studied a unconstrained multiobjective problem with the objective being multifunctions in Banach spaces and, as generalizations of the Fermat rule, presented necessary optimization conditions. In this paper, in a general setting we provide the following fuzzy Lagrange multiplier rule for constrained multiobjective optimization problem (1.3).

Let X, Y_i be Banach spaces, Ω be a closed subset of X, and $F_i : X \to 2^{Y_i}$ be a closed multifunction $(i = 0, 1, \ldots, m)$. Suppose that $(\bar{x}, \bar{y}_0)$ is a local Pareto solution of the constrained multiobjective optimization problem (1.3), and let $\bar{y}_i \in F_i(\bar{x}) \cap -C_i$ $(i = 1, \ldots, m)$. Then one of the following two assertions holds.

(i) For any $\varepsilon > 0$ there exist $x_i \in \bar{x} + \varepsilon B_X$, $w \in \Omega \cap (\bar{x} + \varepsilon B_X)$, $y_i \in F_i(x_i) \cap (\bar{y}_i + \varepsilon B_{Y_i})$, and $c_i^* \in C_i^+$ such that

$$\sum_{i=0}^{m} \|c_i^*\| = 1 \text{ and } 0 \in \sum_{i=0}^{m} D_c^* F_i(x_i, y_i)(c_i^* + \varepsilon B_{Y_i^*}) + N_c(\Omega, w) + \varepsilon B_{X^*},$$

where B_X denotes the closed unit ball of X, $C_i^+ := \{y^* \in Y_i^* : \langle y^*, c\rangle \geq 0\ \forall c \in C_i\}$, $N_c(\cdot,\cdot)$ denotes the Clarke normal cone, and $D_c^*F_i(\cdot,\cdot)$ denotes the Mordukhovich coderivative with respect to the Clarke normal cone.

(ii) For any $\varepsilon > 0$ there exist $x_i \in \bar{x} + \varepsilon B_X$, $w \in \Omega \cap (\bar{x} + \varepsilon B_X)$, $y_i \in F_i(x_i) \cap (\bar{y}_i + \varepsilon B_{Y_i})$, $x_i^* \in D_c^*F_i(x_i, y_i)(\varepsilon B_{Y_i^*})$, and $w^* \in N_c(\Omega, w) + \varepsilon B_{X^*}$ such that

$$\|w^*\| + \sum_{i=0}^{m} \|x_i^*\| = 1 \ \text{ and } \ w^* + \sum_{i=0}^{m} x_i^* = 0. \tag{1.4}$$

Using this result, we give some exact Lagrange multiplier rules for (1.3). In the case when X, Y_i are Asplund spaces, these results are sharpened; in particular, we prove the following result (see section 2 for terms undefined).

Let $(\bar{x}, \bar{y}_0)$ be a local Pareto solution of (1.3), and let $\bar{y}_i \in F_i(\bar{x}) \cap -C_i$. Suppose that each F_i is pseudo-Lipschitz around $(\bar{x}, \bar{y}_i)$ and that each C_i is dually compact (e.g., C_i has a nonempty interior). Then there exists $c_i^* \in C_i^+$ such that

$$\sum_{i=0}^{m} \|c_i^*\| = 1 \ \text{ and } \ 0 \in \sum_{i=0}^{m} D^*F_i(\bar{x}, \bar{y}_i)(c_i^*) + N(\Omega, \bar{x}), \tag{1.5}$$

where $D^*F_i(\cdot,\cdot)$ denotes the Mordukhovich coderivative with respect to the limiting normal cone (see section 2 for its definition). Under the condition that X, Y_i are finite dimensional, we provide the following necessity optimality condition of constrained multiobjective optimization problem (1.3).

Let each F_i be a closed multifunction and each C_i be a closed convex cone. Suppose that $(\bar{x}, \bar{y}_0)$ is a local Pareto solution of (1.3). Then, for any $\bar{y}_i \in F_i(\bar{x}) \cap -C_i$, one of the following assertions holds.

(a) There exists $c_i^* \in C_i^+$ such that (1.5) holds.

(b) There exist $x_i^* \in D^*F_i(\bar{x}, \bar{y}_i)(0)$ and $w^* \in N(\Omega, \bar{x})$ such that (1.4) holds.

Let $f_0, f_1, \ldots, f_m$ be as in (1.1). In the special case when $Y_i = R$, $C_i = \{0\}$ for $0 \leq i \leq m$, $F_i(x) = [f_i(x),\ +\infty)$ for $0 \leq i \leq n$, and $F_i(x) = f_i(x)$ for $n+1 \leq i \leq m$. The above results can be applied to (1.1). In particular, under the assumption that X is an Asplund space and that $f_0, f_1, \cdot, f_n$ are lower semicontinuous and $f_{n+1}, \ldots, f_m$ are continuous, we prove that if $\bar{x}$ is a local solution of (1.1), then one of the following assertions holds.

(i) For any $\varepsilon > 0$ there exist $\lambda_i \in R \setminus \{0\}$, $w \in (\bar{x} + \varepsilon B_X) \cap \Omega$, and $x_i \in \bar{x} + \varepsilon B_X$ with $|f_i(x_i) - f_i(\bar{x})| < \varepsilon$ such that $\lambda_i \geq 0$ for $0 \leq i \leq n$, $\sum_{i=0}^{m} |\lambda_i| = 1$, and

$$0 \in \sum_{i=0}^{m} \hat{\partial}(\lambda_i f_i)(x_i) \cap MB_{X^*} + \hat{N}(\Omega, w) \cap MB_{X^*} + \varepsilon B_{X^*},$$

where $M > 0$ is a constant independent of ε.

(ii) For any $\varepsilon > 0$ there exist $w \in (\bar{x} + \varepsilon B_X) \cap \Omega$, $x_i \in \bar{x} + \varepsilon B_X$ with $|f_i(x_i) - f_i(\bar{x})| < \varepsilon$, $\varepsilon_i \in (-\varepsilon,\ \varepsilon)$, $w^* \in \hat{N}(\Omega, w) + \varepsilon B_{X^*}$, and $x_i^* \in \hat{\partial}(\varepsilon_i f_i)(x_i)$ such that (1.4) holds and $\varepsilon_i > 0$ for $0 \leq i \leq n$.

2. Preliminaries. Throughout this section, we assume that Y is a Banach space. Let $f : Y \to R \cup \{+\infty\}$ be a proper lower semicontinuous function, and let $\mathrm{epi}(f)$ denote the epigraph of f, that is,

$$\mathrm{epi}(f) := \{(y, t) \in Y \times R :\ f(y) \leq t\}.$$

Let $y \in \mathrm{dom}(f)$, let $h \in Y$, and let $f^\circ(y,h)$ denote the generalized directional derivative given by Rockafellar (see [4]), that is,

$$f^\circ(y,h) := \lim_{\varepsilon \downarrow 0} \limsup_{\substack{f \\ z \to y, t \downarrow 0}} \inf_{w \in h + \varepsilon B_Y} \frac{f(z+tw) - f(z)}{t},$$

where B_Y denotes the closed unit ball of Y, and the expression $z \overset{f}{\to} y$ means $z \to y$ and $f(z) \to f(y)$. It is known that $f^\circ(y,h)$ reduces to Clarke's directional derivative when f is locally Lipschitzian (see [4]). Let

$$\partial_c f(y) := \{y^* \in Y^* : \ \langle y^*, h\rangle \le f^\circ(y,h) \quad \forall h \in Y\}.$$

Let A be a closed subset of Y, and let $N_c(A,a)$ denote Clarke's normal cone of A at a, that is,

$$N_c(A,a) := \begin{cases} \partial_c \delta_A(a), & a \in A, \\ \emptyset, & a \notin A, \end{cases}$$

where δ_A denotes the indicator function of A: $\delta_A(y) = 0$ if $y \in A$ and $\delta_A(y) = +\infty$ otherwise. The following result (see [4, Corollary, p. 52]) presents an important necessity optimality condition in terms of Clarke's subdifferential and normal cone for a nonsmooth constrained optimization problem.

PROPOSITION 2.1. *Let $f : Y \to R$ be a locally Lipschitz function and A be a closed subset of Y. Suppose that f attains its minimum over A at $a \in A$. Then $0 \in \partial_c f(a) + N_c(A,a)$.*

We also need the notion of Fréchet normal cones and that of limiting normal cones. For $\varepsilon \ge 0$, the set of ε-normals to A at a is defined by

$$\hat{N}_\varepsilon(A,a) := \left\{ y^* \in Y^* : \limsup_{y \overset{A}{\to} a} \frac{\langle y^*, y-a\rangle}{\|y-a\|} \le \varepsilon \right\},$$

where $y \overset{A}{\to} a$ means that $y \to a$ with $y \in A$. The set $\hat{N}_0(A,a)$ is simply denoted by $\hat{N}(A,a)$ and is called the Fréchet normal cone to A at a. The limiting Fréchet normal cone to A at a is defined by

$$N(A,a) := \{y^* \in Y^* : \ \exists \varepsilon_n \to 0^+, \ y_n \overset{A}{\to} a, \ y_n^* \overset{w^*}{\to} y^* \text{ with } y_n^* \in \hat{N}_{\varepsilon_n}(A,y_n)\}.$$

In the case when A is convex, it is well known that

$$N_c(A,a) = N(A,a) = \hat{N}(A,a).$$

Recall that the Fréchet subdifferential $\hat{\partial} f(y)$ and the limiting subdifferential $\partial f(y)$ of f at $y \in \mathrm{dom}(f)$ are defined by

$$\hat{\partial} f(y) = \{y^* : \ (y^*, -1) \in \hat{N}(\mathrm{epi}(f), (y, f(y))\}$$

and

$$\partial f(y) := \{y^* \in Y^* : \ (y^*, -1) \in N(\mathrm{epi}(f), (y, f(y)))\},$$

respectively. It is known (see [18]) that

$$\hat{\partial} f(y) := \left\{ y^* \in Y^* : \liminf_{v \to y} \frac{f(v) - f(y) - \langle y^*, v - y \rangle}{\|v - y\|} \geq 0 \right\}.$$

Let $\hat{\partial}^\infty f(y)$ and $\partial^\infty f(y)$ denote, respectively, the singular Fréchet subdifferential and the singular limiting subdifferential of f at y, that is,

$$\hat{\partial}^\infty f(y) = \{ y^* : \ (y^*, 0) \in \hat{N}(\mathrm{epi}(f), (y, f(y))) \}$$

and

$$\partial^\infty f(y) := \{ y^* \in Y^* : \ (y^*, 0) \in N(\mathrm{epi}(f), (y, f(y))) \}.$$

Recall that a Banach space Y is called an Asplund space if every continuous convex function defined on an open convex subset D of Y is Fréchet differentiable at each point of a dense G_δ subset of D. It is well known that Y is an Asplund space if and only if every separable subspace of Y has a separable dual. The class of Asplund spaces is well investigated in geometric theory of Banach spaces; see [21] and the references therein. In the case when Y is an Asplund space, Mordukhovich and Shao [18] proved that $\partial f(y) = \limsup_{v \xrightarrow{f} y} \hat{\partial} f(v)$,

$$N(A, a) := \{ y^* \in Y^* : \ \exists y_n \xrightarrow{A} a, \ y_n^* \xrightarrow{w^*} y^* \text{ with } y_n^* \in \hat{N}(A, y_n) \}, \tag{2.1}$$

and $N_c(A, a)$ is the weak* closed convex hull of $N(A, a)$.

In the Asplund space setting, in terms of the Fréchet subdifferential and Fréchet normal cone one has the following necessity optimality condition similar to Proposition 2.1.

PROPOSITION 2.2. *Let Y be an Asplund space and $f : Y \to R$ a locally Lipschitz function, and let A be a closed subset of Y. Suppose that f attains its minimum over A at $a \in A$. Then for any $\varepsilon > 0$ there exist $a_\varepsilon \in a + \varepsilon B_Y$ and $u_\varepsilon \in A \cap (a + \varepsilon B_Y)$ such that*

$$0 \in \hat{\partial} f(a_\varepsilon) + \hat{N}(A, u_\varepsilon) + \varepsilon B_{Y^*}.$$

Proposition 2.2 is due to Fabian [6] (also see [18] for the details).

For $\Phi : X \to 2^Y$, a multifunction from another Banach space X to Y, let $\mathrm{Gr}(\Phi)$ denote the graph of Φ, that is,

$$\mathrm{Gr}(\Phi) := \{ (x, y) \in X \times Y : \ y \in \Phi(x) \}.$$

We say that Φ is closed if $\mathrm{Gr}(\Phi)$ is a closed subset of $X \times Y$ and that Φ is convex if $\mathrm{Gr}(\Phi)$ is a convex subset of $X \times Y$. Recall (see [15, 17]) that Φ is pseudo-Lipschitz at $(\bar{x}, \bar{y}) \in \mathrm{Gr}(\Phi)$ if there exist a constant $L > 0$, a neighborhood U of $\bar{x}$, and a neighborhood V of $\bar{y}$ such that

$$\Phi(x_1) \cap V \subset \Phi(x_2) + \|x_1 - x_2\| L B_Y \quad \forall x_1, x_2 \in U.$$

For $x \in X$ and $y \in \Phi(x)$, let $\hat{D}^* \Phi(x, y)$, $D^* \Phi(x, y)$ and $D_c^* \Phi(x, y) : Y^* \to 2^{X^*}$ denote the Mordukhovich coderivatives of Φ at (x, y) with respect to the Fréchet, limiting, and Clarke normal cones, respectively, that is,

$$\hat{D}^* \Phi(x, y)(y^*) := \{ x^* \in X^* : \ (x^*, -y^*) \in \hat{N}(\mathrm{Gr}(\Phi), (x, y)) \} \quad \forall y^* \in Y^*, \tag{2.2}$$

$$D^* \Phi(x, y)(y^*) := \{ x^* \in X^* : \ (x^*, -y^*) \in N(\mathrm{Gr}(\Phi), (x, y)) \} \quad \forall y^* \in Y^*, \tag{2.3}$$

and

$$D_c^*\Phi(x,y)(y^*) := \{x^* \in X^* :\ (x^*,-y^*) \in N_c(\mathrm{Gr}(\Phi),(x,y))\} \quad \forall y^* \in Y^*$$

(see [17, 18]). We will need the following known result.

PROPOSITION 2.3. *Let $\Phi : X \to 2^Y$ be a closed multifunction. Suppose that Φ is pseudo-Lipschitz at $(\bar{x},\bar{y}) \in \mathrm{gr}(\Phi)$. Then there exist constants $L,\delta > 0$ such that*

$$\sup\{\|x^*\| :\ x^* \in \hat{D}^*\Phi(x,y)(y^*)\} \le L\|y^*\|$$

for any $(x,y) \in \mathrm{Gr}(\Phi) \cap (B(\bar{x},\delta) \times B(\bar{y},\delta))$ and any $y^* \in Y^*$.

Proposition 2.3 can be found in Mordukhovich [15]. Moreover, readers can find a simpler proof of Proposition 2.3 in Jourani and Thibault [11].

Let $S_i : M_i \to 2^Y$ $(i = 1,\dots,n)$ be multifunctions from metric spaces M_i with metrics d_i. Recall (see [19]) that $\bar{x}$ is called an extremal point of the system $(S_1,\dots,S_n)$ at $(\bar{s}_1,\dots,\bar{s}_n)$, provided that $\bar{x} \in \bigcap_{i=1}^n S_i(\bar{s}_i)$ and there exists $r > 0$ such that for any $\varepsilon > 0$ there exists $(s_1,\dots,s_n) \in M_1 \times \cdots \times M_n$ with

$$d_i(s_i,\bar{s}_i) \le \varepsilon,\ d(\bar{x},S_i(s_i)) \le \varepsilon, \quad i = 1,\dots,n, \text{ and } \bigcap_{i=1}^n S_i(s_i) \cap (\bar{x} + rB_Y) = \emptyset.$$

Mordukhovich, Treiman, and Zhu [19] proved the following extended extremal principle.

Theorem MTZ. Let $S_i : M_i \to 2^Y$ be multifunctions from metric spaces (M_i,d_i) to an Asplund space Y, $i = 1,\dots,n$. Assume that $\bar{x}$ is an extremal point of the system $(S_1,\dots,S_n)$ at $(\bar{s}_1,\dots,\bar{s}_n)$, where each S_i is closed-valued around $\bar{s}_i$. Then for any $\sigma > 0$ there exist $s_i \in M_i$, $x_i \in S_i(s_i)$, and $x_i^* \in Y^*$, $i = 1,\dots,n$, such that

$$d_i(s_i,\bar{s}_i) \le \sigma,\ \|x_i - \bar{x}\| \le \sigma,\ x_i^* \in \hat{N}(S_i(s_i),x_i) + \sigma B_{Y^*},\ \sum_{i=1}^n \|x_i^*\| = 1,\ \text{and}\ \sum_{i=1}^n x_i^* = 0.$$

Next we provide a slight improvement of Theorem MTZ, which will be used in the proofs of the main results.

For a natural number n and subsets $A_1,\dots,A_n$ of Y, we define the nonintersection index $\gamma(A_1,\dots,A_n)$ of $A_1,\dots,A_n$ as

$$\gamma(A_1,\dots,A_n) := \inf\left\{\sum_{i=1}^{n-1} \|a_i - a_n\| :\ (a_1,\dots,a_n) \in A_1 \times \cdots \times A_n\right\}.$$

LEMMA 2.1. *Let Y be an Asplund space and $A_1,\dots,A_n$ be closed subsets of Y with $\bigcap_{i=1}^n A_i = \emptyset$. Let $a_i \in A_i$ $(i = 1,\dots,n)$ and $\varepsilon > 0$ such that*

$$\sum_{i=1}^{n-1} \|a_i - a_n\| < \gamma(A_1,\dots,A_n) + \varepsilon.$$

Then for any $\lambda > 0$ there exist $\tilde{a}_i \in A_i$ and $a_i^* \in Y^*$ such that

$$\sum_{i=1}^n \|a_i - \tilde{a}_i\| < \lambda, \quad a_i^* \in \hat{N}(A_i,\tilde{a}_i) + \frac{\varepsilon}{\lambda}B_{Y^*},$$

$$\sum_{i=1}^n \|a_i^*\| = 1 \ \text{ and } \ \sum_{i=1}^n a_i^* = 0.$$

Proof. Let the product Y^n be equipped with the norm $|||(x_1,\dots,x_n)||| = \sum_{i=1}^n \|x_i\|$ for any $x_i \in Y$ $(i=1,\dots,n)$, and define $f : Y^n \to R \cup \{+\infty\}$ by

$$f(x_1,\dots,x_n) := \sum_{i=1}^{n-1} \|x_i - x_n\| + \delta_{A_1\times\cdots\times A_n}(x_1,\dots,x_n) \quad \forall (x_1,\dots,x_n) \in Y^n.$$

Then

$$\inf\{f(x_1,\dots,x_n) :\ (x_1,\dots,x_n) \in Y^n\} = \gamma(A_1,\dots,A_n),$$

and so, by the assumption,

$$f(a_1,\dots,a_n) < \inf\{f(x_1,\dots,x_n) :\ (x_1,\dots,x_n) \in Y^n\} + \varepsilon.$$

Take $\eta \in (0,\ \varepsilon)$ and $\beta \in (0,\ \lambda)$ such that

$$\frac{\eta}{\beta} < \frac{\varepsilon}{\lambda} \ \text{ and } \ f(a_1,\dots,a_n) < \inf\{f(x_1,\dots,x_n) :\ (x_1,\dots,x_n) \in Y^n\} + \eta.$$

Then, by the Ekeland variational principle, there exists $\tilde{x}_i \in A_i$ such that

$$\sum_{i=1}^n \|a_i - \tilde{x}_i\| \le \beta \tag{2.4}$$

and

$$f(\tilde{x}_1,\dots,\tilde{x}_n) \le f(x_1,\dots,x_n) + \frac{\eta}{\beta}\sum_{i=1}^n \|x_i - \tilde{x}_i\| \quad \forall (x_1,\dots,x_n) \in Y^n. \tag{2.5}$$

This and the definition of f imply that $(\tilde{x}_1,\dots,\tilde{x}_n) \in A_1 \times \cdots \times A_n$. It follows from $\bigcap_{i=1}^n A_i = \emptyset$ that

$$\sum_{i=1}^{n-1} \|\tilde{x}_i - \tilde{x}_n\| > 0. \tag{2.6}$$

We define a continuous convex function ψ by

$$\psi(x_1,\dots,x_n) := \sum_{i=1}^{n-1} \|x_i - x_n\| + \frac{\eta}{\beta}\sum_{i=1}^n \|x_i - \tilde{x}_i\| \quad \forall (x_1,\dots,x_n) \in Y^n.$$

It follows from (2.5) that ψ attains its minimum over $A_1 \times \cdots \times A_n$ at $(\tilde{x}_1,\dots,\tilde{x}_n)$. By (2.6) and Proposition 2.2, there exist $\bar{x}_i \in Y$ and $\tilde{a}_i \in A_i$ $(i=1,\dots,n)$ such that

$$\sum_{i=1}^{n-1} \|\bar{x}_i - \bar{x}_n\| > 0, \quad \sum_{i=1}^n \|\tilde{a}_i - \tilde{x}_i\| < \lambda - \beta$$

and

$$0 \in \partial\psi(\bar{x}_1,\dots,\bar{x}_n) + \hat{N}(A_1 \times \cdots \times A_n, (\tilde{a}_1,\dots,\tilde{a}_n)) + \left(\frac{\varepsilon}{\lambda} - \frac{\eta}{\beta}\right) B^n_{Y^*}. \tag{2.7}$$

It follows from (2.4) that $\sum_{i=1}^n \|\tilde{a}_i - a_i\| \le \sum_{i=1}^n \|\tilde{a}_i - \tilde{x}_i\| + \sum_{i=1}^n \|\tilde{x}_i - a_i\| < \lambda$. Let

$$\phi(x_1, \dots, x_n) := \sum_{i=1}^{n-1} \|x_i - x_n\| \quad \forall (x_1, \dots, x_n) \in Y^n.$$

Then

$$\partial\psi(\bar{x}_1, \dots, \bar{x}_n) \subset \partial\phi(\bar{x}_1, \dots, \bar{x}_n) + \frac{\eta}{\beta} B_{(Y^n)^*}.$$

This and (2.7) imply that

$$0 \in \partial\phi(\bar{x}_1, \dots, \bar{x}_n) + \hat{N}(A_1 \times \cdots \times A_n, (\tilde{a}_1, \dots, \tilde{a}_n)) + \frac{\varepsilon}{\lambda} B_{(Y^n)^*}. \tag{2.8}$$

We claim that

$$\partial\phi(\bar{x}_1, \dots, \bar{x}_n) \subset \left\{ (x_1^*, \dots, x_n^*) \in (Y^*)^n : \sum_{i=1}^n x_i^* = 0 \text{ and } \sum_{i=1}^n \|x_i^*\| \ge 1 \right\}. \tag{2.9}$$

Granting this and noting that

$$\hat{N}(A_1 \times \cdots \times A_n, (\tilde{a}_1, \dots, \tilde{a})) = \hat{N}(A_1, \tilde{a}_1) \times \cdots \times \hat{N}(A_n, \tilde{a}_n)$$

is a cone, it follows from (2.8) that there exists $(a_1^*, \dots, a_n^*) \in (Y^*)^n$ such that

$$a_i^* \in \hat{N}(A_i, \tilde{a}_i) + \frac{\varepsilon}{\lambda} B_{Y^*}, \ \sum_{i=1}^n \|a_i^*\| = 1, \text{ and } \sum_{i=1}^n a_i^* = 0.$$

It remains to show that (2.9) holds. Let $(x_1^*, \dots, x_n^*) \in \partial\phi(\bar{x}_1, \dots, \bar{x}_n)$. It follows from the convexity of ϕ that for any $h \in Y$,

$$\sum_{i=1}^n \langle x_i^*, h\rangle \le \phi(\bar{x}_1 + h, \dots, \bar{x}_n + h) - \phi(\bar{x}_1, \dots, \bar{x}_n) = 0.$$

This means that $\sum_{i=1}^n x_i^* = 0$. On the other hand,

$$-\sum_{i=1}^{n-1} \langle x_i^*, \bar{x}_i - \bar{x}_n\rangle = \sum_{i=1}^n \langle x_i^*, -\bar{x}_i\rangle \le \phi(0, \dots, 0) - \phi(\bar{x}_1, \dots, \bar{x}_n) = -\sum_{i=1}^{n-1} \|\bar{x}_i - \bar{x}_n\|.$$

Since, as in (2.6), $\sum_{i=1}^{n-1} \|\bar{x}_i - \bar{x}_n\| > 0$, it follows that $\sum_{i=1}^n \|x_i^*\| \ge 1$. This completes the proof. □

Remark. Lemma 2.1 recaptures Theorem MTZ. Indeed, by the assumption of Theorem MTZ, there exists $r > 0$ such that for any $\sigma \in (0, \min\{\frac{r}{2}, r^{\frac{1}{2}}\})$ there exists $(s_1, \dots, s_n) \in M_1 \times \cdots \times M_n$ such that each $S_i(s_i)$ is closed,

$$d_i(s_i, \bar{s}_i) < \sigma, \ d(\bar{x}, S_i(s_i)) < \frac{\sigma^2}{2n}, \quad i = 1, \dots, n, \text{ and } \bigcap_{i=1}^n S_i(s_i) \cap (\bar{x} + rB_Y) = \emptyset.$$

Hence, there exists $u_i \in S_i(s_i)$ such that $\|u_i - \bar{x}\| < \frac{\sigma^2}{2n}$. This implies that

$$\sum_{i=1}^{n-1} \|u_i - u_n\| \le \sum_{i=1}^{n-1} (\|u_i - \bar{x}\| + \|\bar{x} - u_n\|) < \sigma^2,$$

and so

$$\sum_{i=1}^{n-1} \|u_i - u_n\| < \gamma(S_1(s_1) \cap (\bar{x} + rB_Y), \ldots, S_n(s_n) \cap (\bar{x} + rB_Y)) + \sigma^2.$$

Now with $A_i = S_i(s_i) \cap (\bar{x} + rB_Y)$, $a_i = u_i$, $\varepsilon = \sigma^2$, and $\lambda = \sigma$, there exist $\tilde{a}_i \in A_i$ and $a_i^* \in Y^*$ satisfying the properties as stated in Lemma 2.1. Note that $\tilde{a}_i$ lies in the interior of $\bar{x} + rB_Y$, and it follows that $a_i^* \in \hat{N}(S_i(s_i), \tilde{a}_i)$. Thus Theorem MTZ is seen to hold.

Similar to the proof of Lemma 2.1 but applying Proposition 2.1 in place of Proposition 2.2, we have the following result applicable to the case when Y is a general Banach space.

LEMMA 2.2. *Let Y be a Banach space and $A_1, \ldots, A_n$ be closed subsets of Y with $\bigcap_{i=1}^n A_i = \emptyset$. Let $a_i \in A_i$ $(i = 1, \ldots, n)$ and $\varepsilon > 0$ such that*

$$\sum_{i=1}^{n-1} \|a_i - a_n\| \le \gamma(A_1, \ldots, A_n) + \varepsilon.$$

Then for any $\lambda > 0$ there exist $\tilde{a}_i \in A_i$ and $a_i^* \in Y^*$ such that

$$\sum_{i=1}^{n} \|a_i - \tilde{a}_i\| < \lambda, \quad a_i^* \in N_c(A_i, \tilde{a}_i) + \frac{\varepsilon}{\lambda} B_{Y^*},$$

$$\sum_{i=1}^{n} \|a_i^*\| = 1 \text{ and } \sum_{i=1}^{n} a_i^* = 0.$$

3. Fuzzy Lagrange multiplier rules. In this section, we always assume that X, Y_i are Banach spaces (unless stated otherwise), that $C_i \subset Y_i$ is a closed convex cone, and that each multifunction $F_i : X \to 2^{Y_i}$ is closed. Further we assume that the ordering cone C_0 in Y_0 is nontrivial (i.e., C_0 is not a linear subspace). For convenience we define the norm on the product $X \times \prod_{i=0}^m Y_i$ by

$$\|(x, y_0, y_1, \ldots, y_m)\| = \|x\| + \sum_{i=0}^{m} \|y_i\|.$$

In this section we present three fuzzy Lagrange multiplier rules. The first one works on general Banach spaces, while the last two work on Asplund spaces dealing, respectively, with the set-valued and the numeral-valued functions.

THEOREM 3.1. *Let $(\bar{x}, \bar{y}_0)$ be a local Pareto solution of the constrained multiobjective optimization problem* (1.3) *and $\bar{y}_i$ be a point in $F_i(\bar{x}) \cap -C_i$ $(i = 1, \ldots, m)$. Then one of the following assertions holds.*

(i) *For any $\varepsilon > 0$ there exist $x_i \in \bar{x} + \varepsilon B_X$, $w \in \Omega \cap (\bar{x} + \varepsilon B_X)$, $y_i \in F_i(x_i) \cap (\bar{y}_i + \varepsilon B_{Y_i})$, and $c_i^* \in C_i^+$ such that*

$$\sum_{i=0}^{m} \|c_i^*\| = 1 \text{ and } 0 \in \sum_{i=0}^{m} D_c^* F_i(x_i, y_i)(c_i^* + \varepsilon B_{Y_i^*}) \cap M B_{X^*} + N_c(\Omega, w) \cap M B_{X^*} + \varepsilon B_{X^*},$$

where $M > 0$ is a constant independent of ε.

(ii) *For any* $\varepsilon > 0$ *there exist* $x_i \in \bar{x} + \varepsilon B_X$, $w \in \Omega \cap (\bar{x} + \varepsilon B_X)$, $y_i \in F_i(x_i) \cap (\bar{y}_i + \varepsilon B_{Y_i})$, $x_i^* \in D_c^* F_i(x_i, y_i)(\varepsilon B_{Y_i^*})$, *and* $w^* \in N_c(\Omega, w) + \varepsilon B_{X^*}$ *such that*

$$\|w^*\| + \sum_{i=0}^{m} \|x_i^*\| = 1 \ \text{ and } \ w^* + \sum_{i=0}^{m} x_i^* = 0.$$

Proof. By the assumption there exists $\delta > 0$ such that

$$\bar{y}_0 \in E\left(F_0\left[(\bar{x} + \delta B_X) \cap \Omega \cap \left(\bigcap_{i=1}^{m} F_i^{-1}(-C_i)\right)\right], C_0\right). \tag{3.1}$$

Since the ordering cone C_0 is not a subspace of Y_0, there exists $c_0 \in C_0$ with $\|c_0\| = 1$ such that

$$c_0 \notin -C_0. \tag{3.2}$$

For any natural number k, let $s_k := \frac{1}{(m+2)k^2}$, and consider the following sets in the product space $X \times \prod_{j=0}^{m} Y_j$:

$$A_i := \left\{(x, y_0, y_1, \ldots, y_m) \in X \times \prod_{j=0}^{m} Y_j : \ (x, y_i) \in \mathrm{Gr}(F_i)\right\}, \quad i = 0, 1, \ldots, m,$$

and

$$A_{m+1} := ((\bar{x} + \delta B_X) \cap \Omega) \times (\bar{y}_0 - s_k c_0 - C_0) \times \prod_{i=1}^{m} (\bar{y}_i - C_i).$$

Then $\bigcap_{i=0}^{m+1} A_i = \emptyset$. Indeed, if this is not the case, then there exist $x' \in X$ and $y_i' \in F_i(x')$ $(i = 0, 1, \ldots, m)$ such that

$$x' \in (\bar{x} + \delta B_X) \cap \Omega, \ y_0' \leq_{C_0} \bar{y}_0 - s_k c_0, \ \text{ and } \ y_i' \in \bar{y}_i - C_i (\subset -C_i), \quad i = 1, \ldots, m.$$

Hence, $x' \in (\bar{x} + \delta B_X) \cap \Omega \cap \left(\bigcap_{i=1}^{m} F_i^{-1}(-C_i)\right)$, and so

$$y_0' \in F_0\left[(\bar{x} + \delta B_X) \cap \Omega \cap \left(\bigcap_{i=1}^{m} F_i^{-1}(-C_i)\right)\right].$$

It follows from (3.1) that $\bar{y}_0 \leq_{C_0} y_0'$, and so $\bar{y}_0 \leq_{C_0} \bar{y}_0 - s_k c_0$. This implies that $c_0 \in -C_0$, contradicting (3.2). Let

$$a_0 = a_1 = \cdots = a_m = (\bar{x}, \bar{y}_0, \bar{y}_1, \ldots, \bar{y}_m) \ \text{ and } \ a_{m+1} = (\bar{x}, \bar{y}_0 - s_k c_0, \bar{y}_1, \ldots, \bar{y}_m).$$

Then

$$\sum_{i=0}^{m} \|a_i - a_{m+1}\| = (m+1)s_k < \frac{1}{k^2} \leq \gamma(A_0, A_1, \ldots, A_{m+1}) + \frac{1}{k^2}.$$

By Lemma 2.2 (applied to the family $\{A_0, A_1, \ldots, A_{m+1}\}$ and the constants $\varepsilon = \frac{1}{k^2}$, $\lambda = \frac{1}{k}$), there exist

$$\tilde{a}_i(k) := (x_i(k), y_{i,0}(k), y_{i,1}(k), \ldots, y_{i,m}(k)) \in X \times \prod_{j=0}^{m} Y_j$$

and

$$(x_i^*(k), y_{i,0}^*(k), y_{i,1}^*(k), \dots, y_{i,m}^*(k)) \in X^* \times \prod_{j=0}^{m} Y_j^*$$

$(i = 0, 1, \dots, m+1)$ such that

$$\sum_{i=0}^{m+1} \|\tilde{a}_i(k) - a_i\| = \sum_{i=0}^{m} \left(\|x_i(k) - \bar{x}\| + \sum_{j=0}^{m} \|y_{i,j}(k) - \bar{y}_j\| \right) \tag{3.3}$$

$$+ \ \|x_{m+1}(k) - \bar{x}\| + \|y_{m+1,0}(k) - (\bar{y}_0 - s_k c_0)\| + \sum_{j=1}^{m} \|y_{m+1,j}(k) - \bar{y}_j\| < \frac{1}{k},$$

$$(x_i^*(k), y_{i,0}^*(k), \dots, y_{i,m}^*(k)) \in N_c(A_i, \tilde{a}_i(k)) + \frac{1}{k} \left(B_{X^*} \times \prod_{j=0}^{m} B_{Y_j^*} \right), \tag{3.4}$$

$$\sum_{i=0}^{m+1} \max\{\|x_i^*(k)\|, \ \max\{\|y_{i,j}^*(k)\| : \ j = 0, 1, \dots, m\}\} = 1, \tag{3.5}$$

and

$$\sum_{i=0}^{m+1} (x_i^*(k), y_{i,0}^*(k), y_{i,1}^*(k), \dots, y_{i,m}^*(k)) = 0. \tag{3.6}$$

By the definitions of A_{m+1} and $\tilde{a}_{m+1}(k)$, we see that $N_c(A_{m+1}, \tilde{a}_{m+1}(k))$ is equal to the following product:

$$N_c((\bar{x}+\delta B_X)\cap\Omega, x_{m+1}(k)) \times N_c(\bar{y}_0 - s_k c_0 - C_0, y_{m+1,0}(k)) \times \prod_{j=1}^{m} N_c(\bar{y}_j - C_j, y_{m+1,j}(k)).$$

By well-known relations

$$N_c(\bar{y}_0 - s_k c_0 - C_0, y_{m+1,0}(k)) \subset C_0^+ \ \text{ and } \ N_c(\bar{y}_j - C_j, y_{m+1,j}(k)) \subset C_j^+ \ (1 \le j \le m),$$

it follows that

$$N_c(A_{m+1}, \tilde{a}_{m+1}(k)) \subset N_c((\bar{x} + \delta B_X) \cap \Omega, x_{m+1}(k)) \times \prod_{j=0}^{m} C_j^+.$$

We do the above for every natural number k, and by (3.3) we assume without loss of generality that $\bar{x} + \delta B_X$ is a neighborhood of $x_{m+1}(k)$, and so $N_c((\bar{x} + \delta B_X) \cap \Omega, x_{m+1}(k)) = N_c(\Omega, x_{m+1}(k))$. Hence,

$$N_c(A_{m+1}, \tilde{a}_{m+1}(k)) \subset N_c(\Omega, x_{m+1}(k)) \times \prod_{j=0}^{m} C_j^+.$$

This and (3.4) imply that there exists $(c_0^*(k), c_1^*(k), \dots, c_m^*(k)) \in \prod_{j=0}^{m} C_j^+$ such that

$$x_{m+1}^*(k) \in N_c(\Omega, x_{m+1}(k)) + \frac{1}{k} B_{X^*} \tag{3.7}$$

and

$$\|y^*_{m+1,j}(k) - c^*_j(k)\| \le \frac{1}{k}, \quad j = 0, 1, \dots, m. \tag{3.8}$$

Moreover, for $0 \le i \le m$, we have by the definition of A_i and $\tilde{a}_i(k)$ that

(3.9)
$$N_c(A_i, \tilde{a}_i(k)) = \{(x^*, y^*_0, \dots, y^*_m) : \ (x^*, y^*_i) \in N_c(\mathrm{Gr}(F_i), (x_i(k), y_{i,i}(k))) \text{ and } y^*_j = 0 \quad \forall j \ne i\}.$$

This and (3.4) imply that for $0 \le i \le m$,

$$x^*_i(k) \in D^*_c F_i(x_i(k), y_{i,i}(k)) \left(-y^*_{i,i}(k) + \frac{1}{k} B_{Y^*_i}\right) + \frac{1}{k} B_{X^*} \tag{3.10}$$

and

$$\|y^*_{i,j}(k)\| \le \frac{1}{k}, \quad 0 \le j \le m \ \text{ and } \ j \ne i. \tag{3.11}$$

By (3.6), (3.8), and (3.11), one has

$$-y^*_{i,i}(k) = y^*_{m+1,i}(k) + \sum_{l=0, l\ne i}^{m} y^*_{l,i}(k) \in c^*_i(k) + \frac{m+1}{k} B_{Y^*_i}, \quad i = 0, 1, \dots, m. \tag{3.12}$$

This and (3.10) imply that for $i = 0, 1, \dots, m$,

$$x^*_i(k) \in D^*_c F_i(x_i(k), y_{i,i}(k)) \left(c^*_i(k) + \frac{m+2}{k} B_{Y^*_i}\right) + \frac{1}{k} B_{X^*}. \tag{3.13}$$

In the case when $\{\sum_{j=0}^m \|c^*_j(k)\|\}$ does not converge to 0, without loss of generality we assume that there exists $r > 0$ such that $\sum_{j=0}^m \|c^*_j(k)\| > r$ for all k (passing to subsequences if necessary). It follows from (3.13), (3.7), and (3.6) that

$$\frac{x^*_i(k)}{\sum\limits_{j=0}^m \|c^*_j(k)\|} \in D^*_c F_i(x_i(k), y_{i,i}(k)) \left(\frac{c^*_i(k)}{\sum\limits_{j=0}^m \|c^*_j(k)\|} + \frac{m+2}{rk} B_{Y^*_i} \right) + \frac{1}{rk} B_{X^*}, \quad 0 \le i \le m,$$

$$\frac{x^*_{m+1}(k)}{\sum\limits_{j=0}^m \|c^*_j(k)\|} \in N_c(\Omega, x_{m+1}(k)) + \frac{1}{rk} B_{X^*} \quad \text{and} \quad \sum_{i=0}^{m+1} \frac{x^*_i(k)}{\sum\limits_{j=0}^m \|c^*_j(k)\|} = 0.$$

By virtue of (3.3) and (3.5) and by considering large enough k, it follows that (i) holds with $M = \frac{m+2}{r}$.

Next we consider the case when $t_k := \sum_{j=0}^m \|c^*_j(k)\| \to 0$. In this case, (3.8) implies that

$$y^*_{m+1,j}(k) \to 0 \ \text{ for } j = 0, 1, \dots, m.$$

It follows from (3.11), (3.12), and (3.5) that $\sum_{i=0}^{m+1} \|x^*_i(k)\| \to 1$. Thus, by (3.13), (3.7), and (3.6), there exist

$$\tilde{x}^*_i(k) \in D^*_c F_i(x_i(k), y_{i,i}(k)) \left(c^*_i(k) + \frac{m+2}{k} B_{Y^*_i}\right) \ \text{ for } i = 0, 1, \dots, m$$

and

$$\tilde{x}^*_{m+1}(k) \in N_c(\Omega, x_{m+1}(k)) + \frac{m+2}{k} B_{X^*}$$

such that

$$r_k := \sum_{i=0}^{m+1} \|\tilde{x}^*_i(k)\| \to 1 \text{ and } \sum_{i=0}^{m+1} \tilde{x}^*_i(k) = 0.$$

Therefore, for all k large enough,

$$\frac{\tilde{x}^*_i(k)}{r_k} \in D^*_c F_i(x_i(k), y_{i,i}(k)) \left(\left(\frac{c^*_i(k)}{r_k} + \frac{m+2}{kr_k} \right) B_{Y_i^*} \right),$$

$$\frac{\tilde{x}^*_{m+1}(k)}{r_k} \in N_c(\Omega, x_{m+1}(k)) + \frac{m+2}{kr_k} B_{X^*},$$

$$\sum_{i=0}^{m+1} \left\| \frac{\tilde{x}^*_i(k)}{r_k} \right\| = 1 \text{ and } \sum_{i=0}^{m+1} \frac{\tilde{x}^*_i(k)}{r_k} = 0.$$

Noting that $r_k \to 1$ and $\|c^*_i(k)\| \le t_k \to 0$, this implies that (ii) holds, and the proof is completed. □

In the special case when $F_i(x) = 0$ for all $x \in X$ and $i = 1, \ldots, m$, (1.3) reduces to the following problem:

$$C_0 - \min F_0(x), \quad x \in \Omega, \tag{3.14}$$

and $D^*_c F_i(x, 0)(y^*_i) = 0$ for all $(x, y^*_i) \in X \times Y_i^*$ and $i = 1, \ldots, m$. Thus, the following corollary is an immediate consequence of Theorem 3.1 and recaptures [28, Theorem 3.1] by putting our $\Omega = X$.

COROLLARY 3.1. *Let $(\bar{x}, \bar{y})$ be a local Pareto solution of the constrained multiobjective optimization problem* (3.14). *Then one of the following two assertions holds.*

(i) *For any $\varepsilon > 0$ there exist $u \in \bar{x} + \varepsilon B_X$, $w \in \Omega \cap (\bar{x} + \varepsilon B_X)$, $y \in F_0(u) \cap (\bar{y} + \varepsilon B_Y)$, and $c^* \in C^+$ with $\|c^*\| = 1$ such that*

$$0 \in D^*_c F_0(u, y)(c^* + \varepsilon B_{Y^*}) \cap M B_{X^*} + N_c(\Omega, w) \cap M B_{X^*} + \varepsilon B_{X^*},$$

where $M > 0$ is a constant independent of ε.

(ii) *For any $\varepsilon > 0$ there exist $u \in \bar{x} + \varepsilon B_X$, $w \in \Omega \cap (\bar{x} + \varepsilon B_X)$, $y \in F_0(u) \cap (\bar{y} + \varepsilon B_Y)$, and $x^* \in X^*$ with $\|x^*\| = 1$ such that*

$$x^* \in D^*_c F_0(u, y)(\varepsilon B_{Y^*}) \cap (-N_c(\Omega, w) + \varepsilon B_{X^*}).$$

When X and each Y_i are Asplund spaces, Theorem 3.1 can be strengthened to the following theorem, Theorem 3.2, in which D^*_c and $N_c(\Omega, \cdot)$ are replaced, respectively, by the Fréchet coderivative $\hat{D}^*$ and the Fréchet normal cone $\hat{N}(\Omega, \cdot)$ (recall that $\hat{N}(A, a) \subset N(A, a)$ and $N_c(A, a)$ is the weak*-closed convex hull of $N(A, a)$). The proof is the same as the proof of Theorem 3.1, but use Lemma 2.1 in place of Lemma 2.2.

THEOREM 3.2. *Let $(\bar{x}, \bar{y}_0)$ be a local Pareto solution of the constrained multiobjective optimization problem* (1.3) *and $\bar{y}_i$ be a point in $F_i(\bar{x}) \cap -C_i$ $(i = 1, \ldots, m)$.*

Suppose that X and Y_i $(i = 0, 1, \ldots, m)$ are Asplund spaces. Then one of the following assertions holds.

(i) *For any $\varepsilon > 0$ there exist $x_i \in \bar{x} + \varepsilon B_X$, $w \in \Omega \cap (\bar{x} + \varepsilon B_X)$, $y_i \in F_i(x_i) \cap (\bar{y}_i + \varepsilon B_{Y_i})$, and $c_i^* \in C_i^+$ such that*

$$\sum_{i=0}^{m} \|c_i^*\| = 1 \text{ and } 0 \in \sum_{i=0}^{m} \hat{D}^* F_i(x_i, y_i)(c_i^* + \varepsilon B_{Y_i^*}) \cap M B_{X^*} + \hat{N}(\Omega, w) \cap M B_{X^*} + \varepsilon B_{X^*},$$

where $M > 0$ is a constant independent of ε.

(ii) *For any $\varepsilon > 0$ there exist $x_i \in \bar{x} + \varepsilon B_X$, $w \in \Omega \cap (\bar{x} + \varepsilon B_X)$, $y_i \in F_i(x_i) \cap (\bar{y}_i + \varepsilon B_{Y_i})$, $x_i^* \in \hat{D}^* F_i(x_i, y_i)(\varepsilon B_{Y_i^*})$, and $w^* \in \hat{N}(\Omega, w) + \varepsilon B_{X^*}$ such that*

$$\|w^*\| + \sum_{i=0}^{m} \|x_i^*\| = 1 \text{ and } w^* + \sum_{i=0}^{m} x_i^* = 0.$$

Next we prove that (ii) in Theorem 3.2 cannot happen when each F_i is pseudo-Lipschitz at $(\bar{x}, \bar{y}_i)$.

COROLLARY 3.2. *Let $(\bar{x}, \bar{y}_0)$ be a local Pareto solution of the constrained multiobjective optimization problem* (1.3) *and $\bar{y}_i$ be a point in $F_i(\bar{x}) \cap -C_i$ $(i = 1, \ldots, m)$. Suppose that X and Y_i $(i = 0, 1, \ldots, m)$ are Asplund spaces and that each F_i is pseudo-Lipschitz at $(\bar{x}, \bar{y}_i)$. Then for any $\varepsilon > 0$ there exist $x_i \in \bar{x} + \varepsilon B_X$, $w \in \Omega \cap (\bar{x} + \varepsilon B_X)$, $y_i \in F_i(x_i) \cap (\bar{y}_i + \varepsilon B_{Y_i})$, and $c_i^* \in C_i^+$ such that*

$$\sum_{i=0}^{m} \|c_i^*\| = 1 \text{ and } 0 \in \sum_{i=0}^{m} \hat{D}^* F_i(x_i, y_i)(c_i^* + \varepsilon B_{Y_i^*}) \cap M B_{X^*} + \hat{N}(\Omega, w) \cap M B_{X^*} + \varepsilon B_{X^*},$$

where $M > 0$ is a constant independent of ε.

Proof. Since each F_i is pseudo-Lipschitz at $(\bar{x}, \bar{y}_i)$, Proposition 2.3 implies that there exist constants $L, \delta > 0$ such that for any $(x, y_i) \in \mathrm{Gr}(F_i) \cap (B(\bar{x}, \delta) \times B(\bar{y}_i, \delta))$ and $y_i^* \in Y^*$,

$$\sup\{\|x^*\| : x^* \in \hat{D}^* F_i(x, y_i)(y_i^*)\} \le L\|y_i^*\|. \tag{3.15}$$

We need only show that (i) of Theorem 3.2 holds. If this is not the case, Theorem 3.2 implies that there exist

$$x_i \in B(\bar{x}, \delta),\ w \in \Omega \cap B(\bar{x}, \delta),\ y_i \in F_i(x_i) \cap B(\bar{y}_i, \delta), \tag{3.16}$$

$$x_i^* \in \hat{D}^* F_i(x_i, y_i)\left(\frac{B_{Y_i^*}}{4(m+1)L}\right) \text{ and } w^* \in \hat{N}(\Omega, w) + B_{X^*} \tag{3.17}$$

such that

$$\|w^*\| + \sum_{i=0}^{m} \|x_i^*\| = 1 \text{ and } w^* + \sum_{i=0}^{m} x_i^* = 0. \tag{3.18}$$

By (3.15), (3.16), and (3.17), one has

$$\left\| \sum_{i=0}^{m} x_i^* \right\| \le \sum_{i=0}^{m} \|x_i^*\| \le \frac{1}{4},$$

contradicting (3.18). This completes the proof. ∎

Let $f : X \to R \cup \{+\infty\}$ be a proper lower semicontinuous function and $F(x) = [f(x), +\infty)$ for all $x \in X$. Then F is closed and $\mathrm{Gr}(F) = \mathrm{epi}(f)$. Recall (see [14, Lemma 2.2]) that if $r \in F(\bar{x})$ and X is an Asplund space, then the following assertions hold.

(α) $\lambda \neq 0$ and $x^* \in \hat{D}^*F(\bar{x}, r)(\lambda) \Longleftrightarrow \lambda > 0$, $r = f(\bar{x})$, and $x^* \in \hat{\partial}(\lambda f)(\bar{x})$.

(β) For any $x^* \in \hat{D}^*F(\bar{x}, r)(0)$ there exist sequences $\{x_k\}$, $\{x_k^*\}$, and $\{\lambda_k\}$ such that

$$x_k^* \in \hat{\partial}(\lambda_k f)(x_k),\ (x_k, f(x_k)) \to (\bar{x}, f(\bar{x})),\ \lambda_k \downarrow 0, \text{ and } \|x_k^* - x^*\| \to 0.$$

Let $g : X \to R$ be a continuous function and $G(x) = \{g(x)\}$ for all $x \in X$. The following assertions are known (see [14, Lemma 2.3]).

(α') $\hat{D}^*G(x, g(x))(\lambda) = \partial(\lambda g)(x)$ for any $\lambda \neq 0$.

(β') $x^* \in \hat{D}^*G(x, g(x))(0)$ if and only if there exist sequences $\{x_k\}$, $\{x_k^*\}$, and $\{t_k\}$ such that

$$x_k^* \in \hat{\partial}(t_k g)(x_k) \cup \hat{\partial}(-t_k g)(x_k),\ (x_k, g(x_k)) \to (\bar{x}, g(\bar{x})),\ t_k \downarrow 0, \text{ and } \|x_k^* - x^*\| \to 0.$$

As an application of Theorem 3.2, now we can establish fuzzy necessary optimality conditions for scalar-objective optimization problem (1.1).

THEOREM 3.3. *Let X be an Asplund space and Ω be a closed subset of X. Let $f_0, f_1, \ldots, f_n : X \to R \cup \{+\infty\}$ be proper lower semicontinuous and $f_{n+1}, \ldots, f_m : X \to R$ be continuous. Suppose that $\bar{x}$ is a local solution of* (1.1). *Then one of the following assertions hold.*

(i) *For any $\varepsilon > 0$ there exist $\lambda_i \in R \setminus \{0\}$, $w \in (\bar{x} + \varepsilon B_X) \cap \Omega$, and $x_i \in \bar{x} + \varepsilon B_X$ with $|f_i(x_i) - f_i(\bar{x})| < \varepsilon$ such that $\lambda_i > 0$ for $0 \le i \le n$, $\sum_{i=0}^m |\lambda_i| = 1$, and*

$$0 \in \sum_{i=0}^m \hat{\partial}(\lambda_i f_i)(x_i) \cap MB_{X^*} + \hat{N}(\Omega, w) \cap MB_{X^*} + \varepsilon B_{X^*},$$

where $M > 0$ is a constant independent of ε.

(ii) *For any $\varepsilon > 0$ there exist $w \in (\bar{x} + \varepsilon B_X) \cap \Omega$, $x_i \in \bar{x} + \varepsilon B_X$ with $|f_i(x_i) - f_i(\bar{x})| < \varepsilon$, $\varepsilon_i \in (-\varepsilon, \varepsilon) \setminus \{0\}$, $w^* \in \hat{N}(\Omega, w) + \varepsilon B_{X^*}$, and $x_i^* \in \hat{\partial}(\varepsilon_i f_i)(x_i)$ such that $\varepsilon_i > 0$ for $0 \le i \le n$,*

$$\|w^*\| + \sum_{i=0}^m \|x_i^*\| = 1 \text{ and } w^* + \sum_{i=0}^m x_i^* = 0.$$

Proof. Let ε be an arbitrary positive number. By the lower semicontinuity assumption, there exists $\delta \in (0, \frac{1}{2})$ such that

$$f_i(\bar{x}) - \varepsilon < f_i(x) \text{ for any } x \in \bar{x} + \delta B_X \text{ and } i = 0, 1, \ldots, n. \tag{3.19}$$

Let $Y_0 = Y_1 = \cdots = Y_m = R$. Let $C_i = R^+$, $F_i(x) = [f_i(x), +\infty)$ for $i = 0, 1, \ldots, n$ and $C_i = \{0\}$, $F_i(x) = \{f_i(x)\}$ for $i = n+1, \ldots, m$. Then, each F_i is closed, $(\bar{x}, \bar{y}_0)$ is a local Pareto solution of (1.3), and $\bar{y}_i := f_i(\bar{x}) \in F_i(\bar{x}) \cap -C_i$ for $i = 1, \ldots, m$. Hence, one of the assertions (i) and (ii) in Theorem 3.2 holds. It suffices to show that (i) in Theorem 3.2$\Longrightarrow$(i) and (ii) in Theorem 3.2$\Longrightarrow$(ii). As the arguments are similar, we shall prove only that the implication (i) in Theorem 3.2$\Longrightarrow$(i). Suppose that (i) in Theorem 3.2 holds. Let $\sigma \in (0, \min\{\frac{\varepsilon}{4}, \delta\})$, and take ($\alpha$) into account. Then there

exist $w \in (\bar{x}+\sigma B_X)\cap\Omega$, $(u_i, r_i) \in (\bar{x}+\sigma B_X)\times(f_i(\bar{x})-\sigma,\ f_i(\bar{x})+\sigma)$, and $s_i \in R$ such that

$$r_i \geq f_i(u_i) \ \text{ for } 0 \leq i \leq n, \quad r_i = f_i(u_i) \ \text{ for } n+1 \leq i \leq m,$$

$$s_i \geq 0 \ \text{ for } i = 0, 1, \ldots, n, \quad \sum_{i=0}^{m} |s_i| \geq 1-\sigma, \tag{3.20}$$

and

$$0 \in \sum_{i=0}^{m} \hat{D}^* F_i(u_i, r_i)(s_i) \cap KB_{X^*} + \hat{N}(\Omega, w) \cap KB_{X^*} + \sigma B_{X^*}, \tag{3.21}$$

where $K > 0$ is a constant. By (3.19), one has

$$f_i(\bar{x}) - \varepsilon < f(u_i) \leq r_i < f_i(\bar{x}) + \sigma < f_i(\bar{x}) + \varepsilon \ \text{ for } i = 0, 1, \ldots, n. \tag{3.22}$$

Take $u_i^* \in \hat{D}F_i(u_i, r_i)(s_i) \cap KB_{X^*}$ (by (3.21)) such that

$$-\sum_{i=0}^{m} u_i^* \in \hat{N}(\Omega, w) \cap KB_{X^*} + \sigma B_{X^*}. \tag{3.23}$$

Let $I_0 := \{0 \leq i \leq m : \ s_i = 0\}$. It follows from (α) and (α') that

$$u_i^* \subset \hat{\partial}(s_i f_i)(u_i) \cap KB_{X^*} \ \text{ for any } i \in \{0, 1, \ldots, m\} \setminus I_0. \tag{3.24}$$

For any $i \in \{0, 1, \ldots, n\} \cap I_0$, (3.22) and (β) imply that there exist $\tilde{u}_i \in u_i + \sigma B_X$ with $|f_i(\tilde{u}_i) - f_i(u_i)| < \varepsilon - |f_i(u_i) - f_i(\bar{x})|$, $t_i > 0$, and $x_i^* \in \hat{\partial}(t_i f_i)(\tilde{u}_i)$ such that $\|x_i^* - u_i^*\| < \frac{\sigma}{m}$. Hence, for any $i \in \{0, 1, \ldots, n\} \cap I_0$,

$$\|\tilde{u}_i - \bar{x}\| \leq \|\tilde{u}_i - u_i\| + \|u_i - \bar{x}\| \leq 2\sigma < \varepsilon, \quad |f_i(\tilde{u}_i) - f_i(\bar{x})| < \varepsilon \tag{3.25}$$

and

$$u_i^* \subset \hat{\partial}(t_i f_i)(\tilde{u}_i) \cap \left(K + \frac{1}{m}\right) B_{X^*} + \frac{\sigma}{m} B_{X^*}. \tag{3.26}$$

Moreover, for any $j \in \{n+1, \ldots, m\} \cap I_0$, (β') implies that there exist $\tilde{u}_j \in u_j + \sigma B_X$ with $|f_j(\tilde{u}_j) - f_j(u_j)| < \sigma$, $t_j \in R \setminus \{0\}$, and $x_j^* \in \hat{\partial}(t_j f_j)(\tilde{u}_j)$ such that $\|x_j^* - u_j^*\| < \frac{\sigma}{m}$. Hence, for any $j \in \{n+1, \ldots, m\} \cap I_0$,

$$\|\tilde{u}_j - \bar{x}\| < 2\sigma < \varepsilon, \quad |f_j(\tilde{u}_j) - f_j(\bar{x})| < 2\sigma < \varepsilon \tag{3.27}$$

and

$$u_j^* \subset \hat{\partial}(t_j f_j)(\tilde{u}_j) \cap \left(K + \frac{1}{m}\right) B_{X^*} + \frac{\sigma}{m} B_{X^*}. \tag{3.28}$$

Let $\eta := \sum_{i=0}^{m} |s_i| + \sum_{i \in I_0} |t_i|$, $\lambda_i := \frac{s_i}{\eta}$ if $i \in \{0, 1, \ldots, m\} \setminus I_0$, and $\lambda_i := \frac{t_i}{\eta}$ if $i \in I_0$, and let $x_i := u_i$ if $i \in \{0, 1, \ldots, m\} \setminus I_0$ and $x_i := \tilde{u}_i$ if $i \in I_0$. Then

$$\eta \geq 1 - \sigma > \frac{1}{2}, \ \lambda_i > 0 \ \text{ for } 0 \leq i \leq n, \ \sum_{i=0}^{m} |\lambda_i| = 1,$$

and dividing (3.23), (3.24), (3.26), and (3.28) by η, it follows that

$$0 \in \sum_{i=0}^{m} \hat{\partial}(\lambda_i f_i)(u_i) \cap \left(2K + \frac{2}{m}\right) B_{X^*} + \hat{N}(\Omega, w) \cap 2KB_{X^*} + \varepsilon B_{X^*}.$$

It follows from (3.25) and (3.27) that (i) holds with $M = 2K + \frac{2}{m}$. The proof is completed. ∎

4. Lagrange multiplier rules. In this section, we provide some exact Lagrange multiplier rules for the constrained multiobjective optimization problem (1.3). We will need the following notions. Recall (see [28]) that a closed convex cone C in X is dually compact if there exists a compact subset K of X such that

$$C^+ \subset \{x^* \in X^* : \ \|x^*\| \leq \max\{\langle x^*, x\rangle : \ x \in K\}\}. \tag{4.1}$$

This condition is trivially satisfied if X is finite dimensional (because one can then take $K = B_X$). Note that if C has a nonempty interior, then there exists $c_0 \in C$ such that

$$C^+ \subset \{x^* \in X^* : \ \|x^*\| \leq \langle x^*, c_0\rangle\}.$$

Thus,

$$\operatorname{int}(C) \neq \emptyset \Longrightarrow C \text{ is dually compact.}$$

It is known that if C is dually compact, then

$$c_n^* \in C^+ \ \text{ and } \ c_n^* \overset{w^*}{\to} 0 \Longrightarrow c_n^* \to 0. \tag{4.2}$$

The concept C being dually compact is closely related to the locally compact concept introduced in Loewen [12] (see [28, Proposition 3.1] for the details).

Following Mordukhovich [15] and Mordukhovich and Shao [17], we say that a multifunction Φ from X to another Banach space Y is partially sequentially normally compact at $(x, y) \in \operatorname{Gr}(\Phi)$ if for any (generalized) sequence $\{(x_n, y_n, x_n^*, y_n^*)\}$ satisfying

$$x_n^* \in \hat{D}^*\Phi(x_n, y_n)(y_n^*), \ \ (x_n, y_n) \to (x, y), \ \|y_n^*\| \to 0, \text{ and } x_n^* \overset{w^*}{\to} 0$$

one has $\|x_n^*\| \to 0$.

Clearly, Φ is automatically partially sequentially normally compact at each point of $\operatorname{Gr}(\Phi)$ if X is finite dimensional. Moreover, Proposition 2.3 implies that Φ is partially sequentially normally compact at $(x, y) \in \operatorname{Gr}(\Phi)$ if Φ is pseudo-Lipschitz at (x, y).

In the remainder of this paper, we make the following blanket assumptions.

Assumption 4.1. Each F_i is a closed multifunction.

Assumption 4.2. $(\bar{x}, \bar{y}_0) \in \operatorname{Gr}(F_0)$ is a local Pareto solution of the constrained multiobjective optimization problem (1.3) and $\bar{y}_i \in F_i(\bar{x}) \cap -C_i$ $(1 \leq i \leq m)$.

We first consider the case when X, Y_i are Asplund spaces (thus, in particular (2.1) is valid in these spaces).

THEOREM 4.1. *Let Assumptions* 4.1 *and* 4.2 *hold and X, Y_i be Asplund spaces. Suppose that each C_i is dually compact and that each F_i is partially sequentially normally compact at $(\bar{x}, \bar{y}_i)$. Then one of the following assertions holds.*

(i) *There exists $c_i^* \in C_i^+$ such that*

$$\sum_{i=0}^{m} \|c_i^*\| = 1 \ \text{ and } \ 0 \in \sum_{i=0}^{m} D^*F_i(\bar{x}, \bar{y}_i)(c_i^*) + N(\Omega, \bar{x}).$$

(ii) *There exist $x_i^* \in D^*F_i(\bar{x}, \bar{y}_i)(0)$ and $w^* \in N(\Omega, \bar{x})$ such that*

$$\|w^*\| + \sum_{i=0}^{m} \|x_i^*\| = 1 \ \text{ and } \ w^* + \sum_{i=0}^{m} x_i^* = 0.$$

Proof. Since X, Y_i are Asplund spaces, Assumptions 4.1 and 4.2 imply that one of the assertions (i) and (ii) in Theorem 3.2 holds. Suppose that the assertion (i) in Theorem 3.2 holds. Then, for any natural number k there exist

$$(x_i(k), y_i(k)) \in \mathrm{Gr}(F_i) \cap \left(\left(\bar{x} + \frac{1}{k} B_X \right) \times \left(\bar{y}_i + \frac{1}{k} B_{Y_i} \right) \right), \tag{4.3}$$

$$w(k) \in \left(\bar{x} + \frac{1}{k} B_X \right) \cap \Omega \ \text{ and } \ c_i^*(k) \in C_i^+ \tag{4.4}$$

such that

$$\sum_{i=0}^{m} \|c_i^*(k)\| = 1 \tag{4.5}$$

and

$$\begin{aligned} 0 \in \sum_{i=0}^{m} \hat{D}^* F_i(x_i(k), y_i(k)) \left(c_i^*(k) + \frac{1}{k} B_{Y_i^*} \right) \cap M B_{X^*} \\ + \hat{N}(\Omega, w(k)) \cap M B_{X^*} + \frac{1}{k} B_{X^*}, \end{aligned} \tag{4.6}$$

where $M > 0$ is a constant independent of k. Hence there exist bounded sequences $\{x_i^*(k)\}$ and $\{x^*(k)\}$ such that

$$x_i^*(k) \in \hat{D}^* F_i(x_i(k), y_i(k)) \left(c_i^*(k) + \frac{1}{k} B_{Y_i^*} \right),$$

$$x^*(k) \in \hat{N}(\Omega, w(k)) \ \text{ and } \ x^*(k) + \sum_{i=0}^{m} x_i^*(k) \to 0.$$

Since a bounded set in a dual space is relatively weak* compact, without loss of generality we can assume that

$$x_i^*(k) \overset{w^*}{\to} x_i^* \ \text{ and } \ c_i^*(k) \overset{w^*}{\to} c_i^* \ (i = 0, 1, \ldots, m).$$

It follows from (2.1), (4.3), and (4.4) that

$$0 \in \sum_{i=0}^{m} D^* F_i(\bar{x}, \bar{y}_i)(c_i^*) + N(\Omega, \bar{x}).$$

Noting that $\sum_{i=0}^{m} \|c_i^*\| \neq 0$ by (4.2) and (4.5), this implies that (i) is true.

Next suppose that assertion (ii) in Theorem 3.2 holds. Then for any natural number k there exist

(4.7)

$$(x_i(k), y_i(k)) \in \mathrm{Gr}(F_i) \cap \left(\left(\bar{x} + \frac{1}{k} B_X \right) \times \left(\bar{y}_i + \frac{1}{k} B_{Y_i} \right) \right), \ w(k) \in \left(\bar{x} + \frac{1}{k} B_X \right) \cap \Omega,$$

$$x_i^*(k) \in \hat{D}^* F_i(x_i(k), y_i(k)) \left(\frac{1}{k} B_{Y_i^*} \right) \ \text{ and } \ x^*(k) \in \hat{N}(\Omega, w(k)) \tag{4.8}$$

such that

$$\|x^*(k)\| + \sum_{i=0}^{m} \|x_i^*(k)\| \to 1 \ \text{ and } \ x^*(k) + \sum_{i=0}^{m} x_i^*(k) \to 0. \tag{4.9}$$

Without loss of generality we assume that

$$x^*(k) \overset{w^*}{\to} x^* \text{ and } x_i^*(k) \overset{w^*}{\to} x_i^* \ (i = 0, 1, \ldots, m),$$

and hence it follows from (2.1) that

$$x_i^* \in D^*F_i(\bar{x}, \bar{y}_i)(0),\ x^* \in N(\Omega, \bar{x}),\ \text{ and } \ x^* + \sum_{i=0}^{m} x_i^* = 0.$$

Further $\|x^*\| + \sum_{i=0}^{m} \|x_i^*\| \neq 0$ by (4.9) and thanks to the assumption that each F_i is partially sequentially normally compact at $(\bar{x}, \bar{y}_i)$. Thus (ii) holds, and the proof is completed. □

As already noted, every closed multifunction between two finite dimensional spaces is partially sequentially normally compact at each point in its graph, and every closed convex cone in a finite dimensional space is dually compact. Thus, the following corollary is a consequence of Theorem 4.1.

COROLLARY 4.1. *Let Assumptions* 4.1 *and* 4.2 *hold, and suppose that* X, Y_i *are finite dimensional. Then one of* (i) *and* (ii) *in Theorem* 4.1 *holds.*

In the case when each F_i is pseudo-Lipschitz, we have the following sharp Lagrange multiplier rule.

THEOREM 4.2. *Let Assumptions* 4.1 *and* 4.2 *hold and* X, Y_i *be Asplund spaces. Suppose that each* C_i *is dually compact and that each* F_i *is pseudo-Lipschitz at* $(\bar{x}, \bar{y}_i)$. *Then there exists* $c_i^* \in C_i^+$ *such that*

$$\sum_{i=0}^{m} \|c_i^*\| = 1 \text{ and } 0 \in \sum_{i=0}^{m} D^*F_i(\bar{x}, \bar{y}_i)(c_i^*) + N(\Omega, \bar{x}). \tag{4.10}$$

Proof. By Corollary 3.2, for any natural number k there exist $x_i(k) \in \bar{x} + \frac{1}{k}B_X$, $w(k) \in \Omega \cap (\bar{x} + \frac{1}{k}B_X)$, $y_i(k) \in F_i(x_i) \cap (\bar{y}_i + \frac{1}{k}B_{Y_i})$, and $c_i^*(k) \in C_i^+$ such that

$$\sum_{i=0}^{m} \|c_i^*(k)\| = 1 \tag{4.11}$$

and

$$0 \in \sum_{i=0}^{m} \hat{D}^*F_i(x_i(k), y_i(k)) \left(c_i^*(k) + \frac{1}{k}B_{Y_i^*}\right) \cap MB_{X^*} + \hat{N}(\Omega, w(k)) \cap MB_{X^*} + \frac{1}{k}B_{X^*},$$

where $M > 0$ is a constant independent of k. Hence there exist

$$x_i^*(k) \in \hat{D}^*F_i(x_i(k), y_i(k)) \left(c_i^*(k) + \frac{1}{k}B_{Y_i^*}\right) \text{ and } x^*(k) \in \hat{N}(\Omega, w(k))$$

such that

$$\max\{\|x^*(k)\|,\ \max\{\|x_i^*(k)\| :\ 0 \le i \le m\}\} \le M \text{ and } x^*(k) + \sum_{i=0}^{m} x_i^*(k) \to 0.$$

Without loss of generality, we can assume that

$$x^*(k) \overset{w^*}{\to} x^*,\ x_i^*(k) \overset{w^*}{\to} x_i^*, \text{ and } c_i^*(k) \overset{w^*}{\to} \tilde{c}_i^* \quad \text{for } i = 0, 1, \ldots, m. \tag{4.12}$$

Hence,

$$x^* \in N(\Omega, \bar{x}),\ x_i^* \in D^*F_i(\bar{x}, \bar{y}_i)(\tilde{c}_i^*)\ (i = 0, 1, \ldots, m), \text{ and } x^* + \sum_{i=0}^{m} x_i^* = 0,$$

and so

$$0 \in \sum_{i=0}^{m} D^* F_i(\bar{x}, \bar{y}_i)(\tilde{c}_i^*) + N(\Omega, \bar{x}). \tag{4.13}$$

Since each C_i is dually compact, (4.11), (4.12), and (4.2) imply that $\sum_{i=0}^{m} \|\tilde{c}_i^*\| \neq 0$. It follows from (4.13) that (4.10) holds with $c_i^* = \frac{\tilde{c}_i^*}{\sum_{j=0}^{m} \|\tilde{c}_j^*\|}$. The proof is completed. □

Let $\bar{x}$ be a local solution of single-objective optimization problem (1.1), and suppose that each f_i is locally Lipschitz at $\bar{x}$. Let F_i and C_i be as in the proof of Theorem 3.3. Then $\bar{x}$ is a local Pareto solution of (1.3), and each F_i is pseudo-Lipschitz at $(\bar{x}, f_i(\bar{x}))$. It is routine to verify that

$$D^* F_i(\bar{x}, f_i(\bar{x}))(s) = \partial(sf_i)(\bar{x}) \text{ for } 0 \leq i \leq n, \quad s \geq 0,$$

and

$$D^* F_i(\bar{x}, f_i(\bar{x}))(t) = \partial(tf_i)(\bar{x}) \text{ for } n+1 \leq i \leq m, \quad t \in R.$$

Thus, (4.10) reduces to (1.2).

In the remainder of this section, we consider the case when X, Y_i are general Banach spaces. In this case we need the notion of the normal closedness.

We say that Ω is normally closed at $x \in \Omega$ if for (generalized) sequences

$$x_n \to x,\ x_n^* \in N_c(\Omega, x_n),\ x_n^* \xrightarrow{w^*} x^* \text{ implies } x^* \in N_c(\Omega, x)$$

(see [4, Corollary, p. 58]).

It is known that Ω is normally closed at each point of Ω if Ω is convex. Moreover, if Ω is epi-Lipschitz around $x \in \Omega$, then Ω is normally closed at x. We say that a closed multifunction $\Phi : X \to 2^Y$ is normally closed at $(x, y) \in \mathrm{Gr}(\Phi)$ if $\mathrm{Gr}(\Phi)$ is normally closed at (x, y) (see [28]).

Mimicking a corresponding notion introduced in [17], we say that $\Phi : X \to 2^Y$ is partially sequentially normally compact at $(x, y) \in \mathrm{Gr}(\Phi)$ in the Clarke sense if for any (generalized) sequence $\{(x_n, y_n, x_n^*, y_n^*)\}$ satisfying

$$x_n^* \in D_c^* \Phi(x_n, y_n)(y_n^*),\ \ (x_n, y_n) \to (x, y),\ \|y_n^*\| \to 0, \text{ and } x_n^* \xrightarrow{w^*} 0$$

one has $\|x_n^*\| \to 0$.

The following result can be proved in the same way as for Theorem 4.1 (but apply Theorem 3.2 in place of Theorem 3.1).

THEOREM 4.3. *Let Assumptions* 4.1 *and* 4.2 *hold, and suppose that each C_i is dually compact. Suppose that each F_i is partially sequentially normally compact at $(\bar{x}, \bar{y}_i)$ in the Clarke sense and that Ω and F_i are normally closed at $\bar{x}$ and $(\bar{x}, \bar{y}_i)$, respectively. Then one of the following assertions holds.*

(i) *There exist $c_i^* \in C_i^+$ such that*

$$\sum_{i=0}^{m} \|c_i^*\| = 1 \text{ and } 0 \in \sum_{i=0}^{m} D_c^* F_i(\bar{x}, \bar{y}_i)(c_i^*) + N_c(\Omega, \bar{x}).$$

(ii) *There exist $x_i^* \in D_c^* F_i(\bar{x}, \bar{y}_i)(0)$ and $w^* \in N_c(\Omega, \bar{x})$ such that*

$$\|w^*\| + \sum_{i=0}^{m} \|x_i^*\| = 1 \text{ and } w^* + \sum_{i=0}^{m} x_i^* = 0.$$

As in many classical situations, one can also provide a sufficient condition for $(\bar{x}, \bar{y}_0)$ to be a Pareto solution of (1.3), provided that a suitable convexity assumption is made.

PROPOSITION 4.1. *Let each F_i be a closed convex multifunction and Ω be a closed convex subset of X. Let $\bar{y}_0 \in F_0(\bar{x})$ and $\bar{y}_i \in F_i(\bar{x}) \cap -C_i$ for $i = 1, \ldots, m$. Assume that there exists $c_i^* \in C_i^+$ such that*

$$\langle c_0^*, c\rangle > 0 \quad \forall c \in C_0 \setminus \{0\}, \quad \sum_{i=1}^{m} \langle c_i^*, \bar{y}_i\rangle = 0 \tag{4.14}$$

and

$$0 \in \sum_{i=0}^{m} D^* F_i(\bar{x}, \bar{y}_i)(c_i^*) + N(\Omega, \bar{x}). \tag{4.15}$$

Then $(\bar{x}, \bar{y}_0)$ is a Pareto solution of the constrained multiobjective optimization problem (1.3).

Proof. By (4.15) there exists $x_i^* \in X^*$ such that

$$x_i^* \in D^* F_i(\bar{x}, \bar{y}_i)(c_i^*) \quad \text{and} \quad -\sum_{i=0}^{m} x_i^* \in N(\Omega, \bar{x}).$$

It follows from the convexity of F_i and Ω that

$$\langle x_i^*, x\rangle - \langle c_i^*, y_i\rangle \le \langle x_i^*, \bar{x}\rangle - \langle c_i^*, \bar{y}_i\rangle \quad \forall (x, y_i) \in \mathrm{Gr}(F_i) \ \text{ and } \ i = 0, 1, \ldots, m \tag{4.16}$$

and

$$\left\langle -\sum_{i=0}^{m} x_i^*, x \right\rangle \le \left\langle -\sum_{i=0}^{m} x_i^*, \bar{x} \right\rangle \quad \forall x \in \Omega. \tag{4.17}$$

Summing up (4.16) over all i and making use of (4.17) and (4.14) we have

$$\langle c_0^*, \bar{y}_0\rangle \le \sum_{i=0}^{m} \langle c_i^*, y_i\rangle \ \text{ for any } x \in \Omega, \ y_i \in F_i(x), \text{ and } i = 0, 1, \ldots, m.$$

Since $c_i^* \in C_i^+$, it follows that

$$\langle c_0^*, \bar{y}_0\rangle \le \langle c_0^*, y_0\rangle \quad \forall y_0 \in F_0\left(\Omega \cap \bigcap_{i=1}^{m} F_i^{-1}(-C_i)\right).$$

This and the inequality in (4.14) imply that $\bar{y}_0 \in E\left(F_0\left(\Omega \cap \bigcap_{i=1}^{m} F_i^{-1}(-C_i)\right), C_0\right)$. The proof is completed. ∎

Acknowledgment. We thank the referee for his helpful comments and for [6, 11, 12].

REFERENCES

[1] J. M. Borwein, *On the existence of Pareto efficient points*, Math. Oper. Res., 18 (1983), pp. 64–73.

[2] J. M. Borwein, J. S. Treiman, and Q. J. Zhu, *Necessary conditions for constrained optimization problems with semicontinuous and continuous data.*, Trans. Amer. Math. Soc., 350 (1998), pp. 2409–2429.

[3] J. M. Borwein and Q. J. Zhu, *A survey of subdifferential calculus with applications*, Nonlinear Anal., 38 (1999), pp. 687–773.

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

[5] L. Cesari and M. B. Suryanarayana, *Existence theorems for Pareto optimization: Multivalued and Banach space-valued functionals*, Trans. Amer. Math. Soc., 244 (1978), pp. 37–65.

[6] M. Fabian, *Subdifferentiability and trustworthiness in light of a new variational principle of Borwein and Preiss*, Acta Univ. Carolin. Math. Phys., 30 (1989), pp. 51–56.

[7] F. Flores-Bazan, *Ideal, weakly efficient solutions for vector optimzation problems*, Math. Program., 93 (2002), pp. 453–475.

[8] A. Götz and J. Jahn, *The Lagrange multiplier rule in set-valued optimization*, SIAM J. Optim., 10 (1999), pp. 331–344.

[9] M. I. Henig, *Existence and characterization of efficient decisions with respect to cones*, Math. Programming, 23 (1982), pp. 111–116.

[10] G. Isac, *Pareto optimization in infinite-dimensional spaces: The importance of nuclear cones*, J. Math. Anal. Appl., 182 (1994), pp. 393–404.

[11] A. Jourani and L. Thibault, *Qualification conditions for calculus rules of coderivatives of multivalued mappings*, J. Math. Anal. Appl., 218 (1998), pp. 66–81.

[12] P. D. Loewen, *Limits of Fréchet normals in nonsmooth analysis*, in Optimization and Nonlinear Analysis, A. D. Ioffe, ed., Pitman Res. Notes Math. Ser. 244, Longman Scientific and Technical, Harlow, UK, 1992, pp. 178–188.

[13] M. Minami, *Weak Pareto-optimal necessary conditions in a nondifferential multiobjective program on a Banach space*, J. Optim. Theory Appl., 41 (1983), pp. 451–461.

[14] H. V. Ngai and M. Théra, *A fuzzy necessary optimality condition for non-Lipschitz optimization in Asplund spaces*, SIAM J. Optim., 12 (2002), pp. 656–668.

[15] B. S. Mordukhovich, *Coderivative of set-valued mappings: Calculus and application*, Nonlinear Anal., 30 (1997), pp. 3059–3070.

[16] B. S. Mordukhovich, *Necessary conditions in nonsmooth minimization via lower and upper subgradients*, Set-Valued Anal., 12 (2004), pp. 163–193.

[17] B. S. Mordukhovich and Y. Shao, *Nonconvex differential calculus for infinite-dimensional multifunctions*, Set-Valued Anal., 4 (1996), pp. 205–236.

[18] B. S. Mordukhovich and Y. Shao, *Nonsmooth sequential analysis in Asplund spaces*, Trans. Amer. Math. Soc., 348 (1996), pp. 1235–1280.

[19] B. S. Mordukhovich, J. S. Treiman, and Q. J. Zhu, *An extended extremal principle with applications to multiobjective optimization*, SIAM J. Optim., 14 (2003), pp. 359–379.

[20] B. S. Mordukhovich and B. Wang, *Necessary suboptimality and optimality conditions via variational principles*, SIAM J. Control Optim., 41 (2002), pp. 623–640.

[21] R. R. Phelps, *Convex Functions, Monotone Operators and Differentiability*, 2nd ed., Springer-Verlag, Berlin, 1993.

[22] Y. Sonntag and C. Zalinescu, *Comparison of existence results for efficient points*, J. Optim. Theory Appl., 105 (2000), pp. 161–188.

[23] A. Sterna-Karwat, *On existence of cone maximal points in real topological linear spaces*, Israel J. Math., 54 (1986), pp. 33–41.

[24] X. Q. Yang and V. Jeyakumar, *First and second-order optimality conditions for composite multiobjective optimization*, J. Optim. Theory Appl., 95 (1997), pp. 209–224.

[25] J. J. Ye and Q. J. Zhu, *Multiobjective optimization problem with variational inequality constraints*, Math. Program., 96 (2003), pp. 139–160.

[26] A. Zaffaroni, *Degrees of efficiency and degree of minimality*, SIAM J. Control Optim., 42 (2003), pp. 1071–1086.

[27] C. Zalinescu, *Convex Analysis in General Vector Spaces*, Word Scientific, Singapore, 2002.

[28] X. Y. Zheng and K. F. Ng, *The Fermat rule for multifunctions on Banach spaces*, Math. Program., 104 (2005), pp. 69-90.

[29] Q. J. Zhu, *Hamiltonian necessary conditions for a multiobjective optimal control problem with endpoint constraints*, SIAM J. Control Optim., 39 (2000), pp. 97–112.

[30] Q. J. Zhu, *Necessary conditions for constrained optimization problems in smooth Banach spaces and applications*, SIAM J. Optim., 12 (2002), pp. 1032–1047.

SIAM J. OPTIM.
Vol. 17, No. 4, pp. 1176–1188
© 2006 Society for Industrial and Applied Mathematics

NEW REDUCTION TECHNIQUES FOR THE GROUP STEINER TREE PROBLEM*

CARLOS EDUARDO FERREIRA† AND FERNANDO M. DE OLIVEIRA FILHO†

Abstract. The group Steiner tree problem consists of, given a graph G, a collection $\mathcal{R}$ of subsets of $V(G)$, and a positive cost c_e for each edge e of G, finding a minimum-cost tree in G that contains at least one vertex from each $R \in \mathcal{R}$. We call the sets in $\mathcal{R}$ *groups*. The well-known Steiner tree problem is the special case of the group Steiner tree problem in which each set in $\mathcal{R}$ is unitary. In this paper, we present a general reduction test designed to remove group vertices, that is, vertices belonging to some group. Through the use of these tests we can conclude that a given group vertex can be considered a nonterminal and hence can be removed from its group. We also present some computational results on instances from SteinLib [T. Koch, A. Martin, and S. Voss, *SteinLib*: *An updated library on Steiner tree problems in graphs*, in Steiner Trees in Industry, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2001, pp. 285–325].

Key words. Steiner trees, preprocessing, reduction techniques, branch-and-cut

AMS subject classifications. 90C27, 90C35, 90C57

DOI. 10.1137/040610891

1. Introduction.

1.1. The problem and known results.

The Steiner tree problem consists of, given a graph G with costs on its edges and a set of vertices called *terminals*, finding a minimum-cost tree in G that connects all the terminals.

The group Steiner tree problem is a generalization of the Steiner tree problem and was proposed by Reich and Widmayer [12] with applications to VLSI design in mind. It is as follows:

PROBLEM GST $(G, \mathcal{R}, c)$. *Given a graph G, a collection $\mathcal{R}$ of subsets of $V(G)$, and a positive cost c_e for each $e \in E(G)$, find a minimum-cost tree T in G containing at least one vertex of each $R \in \mathcal{R}$.*

The elements of $\mathcal{R}$ are called *groups*. Any vertex belonging to a group is said to be a *group vertex*. All other vertices are *nonterminals* or *Steiner vertices*. We denote by $\hat{\mathcal{R}}$ the set of all group vertices. A tree in G that contains at least one vertex from each group is called a *group Steiner tree*.

We assume that the costs are all positive and that the groups are pairwise disjoint. Such assumptions are reasonable in practice and make our exposition much simpler. So, for a group vertex u, we denote by $\mathrm{gr}(u)$ the group which contains u. Also, we assume that there are at least two groups and that at least one component of G contains at least one vertex from each group (in other words, we assume that the problem is feasible).

Notice that the Steiner tree problem is the special case of the group Steiner tree problem in which every group has only one vertex. Therefore, since the Steiner tree

*Received by the editors July 1, 2004; accepted for publication (in revised form) May 30, 2006; published electronically December 26, 2006.

http://www.siam.org/journals/siopt/17-4/61089.html

†Department of Computer Science, Institute of Mathematics and Statistics, University of São Paulo, 05508-970 São Paulo, Brazil (cef@ime.usp.br, fmario@ime.usp.br). The first author was partially supported by CNPQ 300752/94-6 and Pronex 107/97. The second author was supported by FAPESP grant 03/10045-0.

problem is NP-hard, the group Steiner tree problem is also NP-hard. There are also stronger complexity results related to the group Steiner tree problem. Ihler [8] considered the complexity of several special cases of the problem. From a reduction of set-cover to the GST [6, 8] and from strong results known for set-cover [5], it is known that the group Steiner tree problem, even when G is restricted to a tree, cannot be approximated in polynomial time by a factor better than $(1 + o(1)) \ln |\mathcal{R}|$, where $\mathcal{R}$ is the collection of groups, unless NP $\subseteq$ DTIME$(n^{O(\log \log n)})$. Stronger results on the hardness of approximating the group Steiner tree problem have been obtained by Halperin and Krauthgamer [7], though under a different hypothesis.

Rohe and Zachariasen [13] considered the rectilinear case of the group Steiner tree problem. They developed an exact algorithm for this special case of the problem and presented reduction techniques, some of them based on existing reductions for the Steiner tree problem. Duin, Volgenant, and Voss [4] use a simple reduction to transform an instance of the group Steiner tree problem into an instance of the Steiner tree problem and solve the latter problem. Details on how the group Steiner tree problem arises in VLSI design can be found in the papers of Reich and Widmayer [12] and Rohe and Zachariasen [13].

1.2. Our results. The group Steiner tree problem can be reduced to the Steiner tree problem as follows: given an instance $(G, \mathcal{R}, c)$ of the GST, create a new vertex v_R in G for each $R \in \mathcal{R}$, connecting v_R to all the vertices in R by edges with a conveniently large cost.

Examples can be constructed in which the addition of the artificial edges has a negative impact on the lower bounds provided by formulations for the problem. The performance of primal heuristics may also be poorer on the extended graph.

For that reason, we present in this paper some reduction techniques designed specifically for the group Steiner tree problem. We concentrate on techniques for removing group vertices, that is, methods that can allow us to conclude that a given group vertex can be considered a nonterminal. The quality of many formulations for the group Steiner tree problem depends on the size of the groups in question (cf. [1]), and we have observed in practice that with smaller groups one can really obtain better bounds.

The rest of the paper is structured as follows: in section 2 we present the concept of bottleneck Steiner distances for the GST, which was generalized from the concept of bottleneck distances for the Steiner tree problem by Rohe and Zachariasen [13]. We also present in this section a simple test using the bottleneck distances. In section 3 we present a more powerful test based on expansion techniques. Finally, in section 4 we present some computational results obtained.

2. Bottleneck distances and a simple test.

2.1. Group bottleneck Steiner distances. In this section we present a generalization of the concept of bottleneck distances. The concept of bottleneck Steiner distances was introduced by Duin and Volgenant [3] and still forms the base of the most successful reduction tests for the Steiner tree problem. This concept was then generalized by Rohe and Zachariasen [13] to deal with the GST.

If H is any graph with costs on its edges given by a function l and $u, v \in V(H)$, the *bottleneck* between u and v is given by

$$b(u, v) := \min_{P} \max_{e \in E(P)} l_e,$$

where the minimum is taken over all (u, v)-paths P in H.

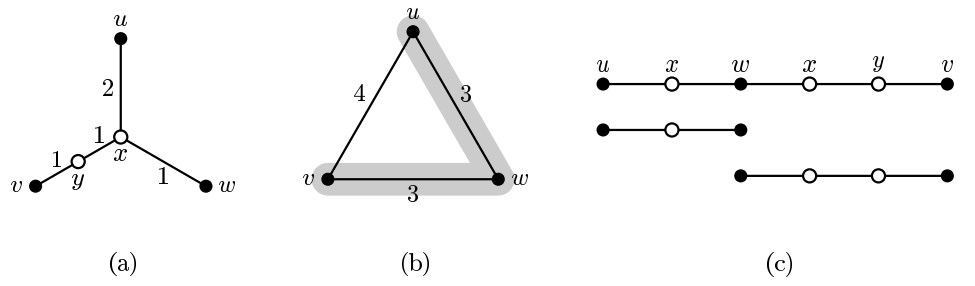


FIG. 1. (a) *A graph G with costs on its edges. The black vertices form the set X.* (b) *The graph $K_G[X]$. On each edge we have the distance, according to the costs in* (a)*, between each pair of vertices of X. In gray we have the path in $K_G[X]$ that gives the bottleneck Steiner distance for X between u and v.* (c) *A bottleneck Steiner walk for X between u and v. Below the walk we have its elementary paths.*

Consider now an instance $(G, \mathcal{R}, c)$ of the GST. We say that a set $X \subseteq V(G)$ is *valid* for $\mathcal{R}$ if X contains exactly one vertex of each group in $\mathcal{R}$. For any collection $\mathcal{R}$ of groups, we denote by $\nu(\mathcal{R})$ the collection of all sets that are valid for $\mathcal{R}$. If we know a valid set $X \in \nu(\mathcal{R})$ contained in some optimal solution of the problem GST $(G, \mathcal{R}, c)$, solving the latter problem is the same as solving the Steiner tree problem in the same graph, with the same costs, with X as the terminal set.

Now, denote by K_G the complete graph with vertex set $V(G)$. Let $d_G(u, v)$ denote the cost of a minimum-cost (u, v)-path in G, with edge costs given by c. We can view $d_G(u, v)$ as being the cost of the edge uv of K_G. In this context, given a valid set X, the *bottleneck Steiner distance* for X between vertices $u, v \in V(G)$ is the bottleneck between u and v in $K_G[X \cup \{u, v\}]$, with edge costs in K_G given by d_G, and is denoted by $s_X(u, v)$. Here, we denote by $K_G[A]$, where $A \subseteq V(G)$, the subgraph of K_G induced by the vertices in A, that is, the graph having A as its vertex set and containing every edge of K_G that has both its endpoints in A.

From each (u, v)-path P in $K_G[X \cup \{u, v\}]$ we can obtain a (u, v)-walk W in G by concatenating minimum-cost paths in G corresponding to the edges of P. Such a walk is divided into paths between u, consecutive vertices of X, and v; such paths are called *elementary*. A (u, v)-walk W obtained from a (u, v)-path P in $K_G[X \cup \{u, v\}]$ such that $s_X(u, v) = \max\{d_G(x, y) \colon xy \in E(P)\}$ is called a *bottleneck Steiner walk* for X. Note that, if W is a bottleneck Steiner walk for X, the cost of the greatest elementary path of W is exactly $s_X(u, v)$. Figure 1 illustrates these definitions.

We are now ready to define the concept of group bottleneck Steiner distance. Given an instance $(G, \mathcal{R}, c)$ of the GST, the *group bottleneck Steiner distance* between vertices $u, v \in V(G)$ is given by

$$s(u, v) := \max_{X \in \nu(\mathcal{R})} s_X(u, v).$$

Given a valid set X and vertices $u, v \in V(G)$, we can compute $s_X(u, v)$ in polynomial time (cf. Duin [2]). Rohe and Zachariasen [13] proved that the problem of computing $s(u, v)$ is NP-hard. In the same work, the authors presented two heuristics for computing upper bounds for $s(u, v)$.

2.2. A simple test. We now illustrate the use of the group bottleneck Steiner distances introduced in the previous section. We first note that many of the reduction methods devised for the Steiner tree problem that use bottleneck Steiner distances

can be easily adapted to be used with the GST. Therefore, we do not concentrate on tests that remove nonterminals or edges, but instead we concentrate on tests that remove group vertices.

Given an instance $(G, \mathcal{R}, c)$ of the GST and a group vertex r, we wish to remove r from its group. Then we may use other reduction methods to try to remove r from the graph. To show that r may be regarded as a nonterminal, we need only show an optimal solution of the problem that contains a vertex from the same group of r different from r. If all the neighbors of r belong to the same group of it, then that is easy: since we assume that there are at least two groups, every optimal solution that uses r also uses some of its neighbors, that is, uses some vertex of $\mathrm{gr}(r) \setminus \{r\}$.

The test we will present in this section generalizes the previous observation. To present it, though, we need some notation. Consider a graph G and a set A of vertices of G. We denote by $N_G(A)$ the set of neighbors of A in G, i.e.,

$$N_G(A) := \{v \in V(G) \setminus A \colon v \text{ is adjacent to some vertex of } A\}.$$

If there is no risk of ambiguity, we drop the graph from the symbol, denoting the set of neighbors simply by $N(A)$. If u is a vertex, we may also write $N(u)$ instead of $N(\{u\})$. Given two disjoint sets of vertices A and B of a graph G, we denote by $\delta_G(A, B)$ the set of edges of G with one endpoint in A and the other in B. Analogously, if there is no risk of ambiguity, we drop the graph from the symbol. Moreover, if u is a vertex we may, for instance, write $\delta(u, B)$ instead of $\delta(\{u\}, B)$.

Consider now an instance $(G, \mathcal{R}, c)$ of the GST. Again, denote by K_G the complete graph with vertex set $V(G)$. Given a set A of vertices of G, we shall denote by $\tau(A)$ the cost of a minimum spanning tree in $K_G[A]$, where the cost of the edge uv of $K_G[A]$ equals $s(u, v)$. Given a group vertex r and a set A of vertices such that $r \notin A$, let

$$\pi(r, A) := \min\{d_G(u, v) \colon u \in A,\ v \in \mathrm{gr}(r) \setminus \{r\}\};$$

that is, $\pi(r, A)$ is the distance from A to the closest vertex of the same group of r distinct from r itself. As is usual, if $A \subseteq E(G)$, we write $c(A) := \sum_{e \in A} c_e$. Finally, we have the following result.

THEOREM 1. *Consider an instance $(G, \mathcal{R}, c)$ of the* GST. *Let $R \in \mathcal{R}$ and $r \in R$. If, for every $U \subseteq N(r) \setminus R$, $|U| \geq 1$, we have $\tau(U) + \pi(r, U) \leq c(\delta(r, U))$, then r may be regarded as a nonterminal.*

To prove Theorem 1, we will need the following lemma, which will also be useful later.

LEMMA 2. *Consider an instance $(G, \mathcal{R}, c)$ of the* GST. *Let F be a forest in G and suppose F contains a valid set X. Let $U \subseteq V(F)$ be a set containing at least one vertex from each component of F. There is a group Steiner tree T in G that contains F, and its cost is at most $c(E(F)) + \tau(U)$.*

Proof. Consider a minimum spanning tree in $K_G[U]$ with cost $\tau(U)$. Let yz be the smallest edge of this tree and consider a (y, z)-bottleneck Steiner walk W for X. Traverse W from y to z. During the traversal, subpaths of W connecting vertices in different components of F may be found. Since F contains a valid set, such subpaths are subpaths of elementary paths of W and have cost at most $s_X(y, z) \leq s(y, z)$. When such a subpath is found, add it to F and continue with the traversal. After doing this, if F is not yet connected, repeat the same procedure, choosing the second smallest edge from the minimum spanning tree, and so on.

After we add at most $|U| - 1$ paths to F it becomes connected and hence contains a spanning tree T containing the original forest F. The total cost of the paths added to

connect F is at most $\tau(U)$, and we have that the cost of T is at most $c(E(F)) + \tau(U)$. Since F contains a valid set, T also contains a valid set and is thus a group Steiner tree. □

With this, we can easily prove Theorem 1.

Proof of Theorem 1. To show that r may be regarded as a nonterminal we show an optimal solution of GST $(G, \mathcal{R}, c)$ that contains some vertex of R other than r.

To this end, let T^* be an optimal solution of GST$(G, \mathcal{R}, c)$ and suppose that r is the only vertex of R that occurs in T^*; otherwise we have nothing to do. Let $U := N_{T^*}(r)$ and $T' := T^* - r$. Let $u \in U$ and $r' \in R \setminus \{r\}$ be such that $d_G(u, r') = \pi(r, U)$. Connect r' to T' using a subpath of a minimum-cost (u, r')-path such that the resulting graph, say F, is a forest.

Note that F has at most $|U|$ components and that each such component contains at least one vertex of U. Notice, moreover, that F contains a valid set. From Lemma 2 we know that there is a group Steiner tree $\hat{T}$ that contains F and whose cost is at most

$$c(E(F)) + \tau(U) \leq c(E(T^*)) + \pi(r, U) - c(\delta(r, U)) + \tau(U) \leq c(E(T^*)),$$

and hence $\hat{T}$ is also an optimal solution. But it contains a vertex in $R \setminus \{r\}$, namely r', and we are done. □

Notice that the test above is indeed a generalization of the observation we made before. Moreover, it is similar to a test introduced by Duin and Volgenant [3] to remove nonterminals in the Steiner tree problem.

3. An expansion approach.

3.1. An outline of the expansion test. In recent years, many reduction tests have been designed for the Steiner tree problem using the idea of expansion [14, 11]. While the classical tests inspect only simple graph structures, like edges or vertices, those tests inspect more general structures, like paths or trees.

Polzin and Daneshmand [11] give a general framework for expansion tests for the Steiner tree problem (also described in the thesis [10]). Their test can be used to eliminate edges and nonterminals, and if we use the group bottleneck Steiner distances defined in section 2, we can use the same tests with the GST after some minor adjustments. In this section we investigate how to use the same idea of expansion presented in [11] in order to remove group vertices. The resulting test is much more powerful than that presented in section 2.

To present our test, we first need to introduce some notation from [11]. Given a tree T, we denote by $L(T)$ the set of leaves of T. Let T' be a subtree of T. The *linking set* between T and T' is the set of vertices v of T' with at least one path from itself to a leaf of T not containing any edge of T'. In other words, the linking set is the set of vertices that connects T' and T. If the linking set between T and T' is $L(T')$, then T' is said to be *peripherally contained* in T. Figure 2 illustrates these definitions.

Let $(G, \mathcal{R}, c)$ be an instance of the GST and consider a group vertex $r \in R$ for some $R \in \mathcal{R}$. Let T be a tree in G. We shall say that the tree T *respects* r if T contains r and every other group vertex that occurs in T is a leaf. Given a leaf u of T, let

$$\mathcal{X}(u, T) := \{uv \in E(G) \colon v \notin V(T)\}.$$

We call any nonempty subset of $\mathcal{X}(u, T)$ an *expansion of* T *through* u.

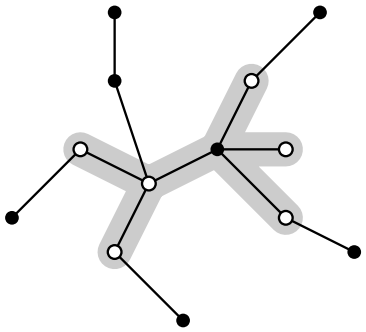


FIG. 2. *The subtree in gray is not peripherally contained in the larger tree. The linking set, which is composed in this case by the white vertices, is not entirely composed of leaves of the gray subtree.*

Our test is comprised of two recursive functions. The first function, EXPANSION-TEST-LEAF, receives a group vertex r and a tree T that respects r and of which r is a leaf. It returns YES if the following assertion is true:

(3.1)
If T is peripherally contained in some optimal solution in which r is a leaf, then there is an optimal solution containing a vertex of $\mathrm{gr}(r)$ *other than r.*

The other function is called EXPANSION-TEST-INT. It receives a group vertex r and a tree T that respects r and returns YES if the following assertion is true:

(3.2)
If T is peripherally contained in some optimal solution in which r is not a leaf, then there is an optimal solution containing a vertex of $\mathrm{gr}(r)$ *other than r.*

The functions described above are very similar; thus we will present only an outline of their code and comment on the differences later. The function TEST-TREE called in line 1 of the code below contains the specific test conditions: the call TEST-TREE (r, T) returns YES if (3.1) (in case of EXPANSION-TEST-LEAF), or (3.2) (in case of EXPANSION-TEST-INT), is satisfied by T. The implementation of such a function will be discussed later.

Function EXPANSION-TEST-OUTLINE (r, T)
1 if TEST-TREE $(r, T) =$ YES
2 then return YES
3 for each leaf u of T that is r or a nonterminal do
4 $S \leftarrow$ YES
5 for each nonempty $U \subseteq \mathcal{X}(u, T)$ do
6 if EXPANSION-TEST-OUTLINE $(r, T + U) =$ NO
7 then $S \leftarrow$ NO
8 if $S =$ YES
9 then return YES
10 return NO

We will delay the proof of correctness of our test until section 3.3. For now, it is

convenient to observe that, throughout execution, the property that T respects r is preserved.

Both functions have almost the same code. The only difference is that, in function Expansion-Test-Leaf, we can only choose to expand through r if it has degree zero in T (i.e., if T is only one vertex, r), as r must be a leaf in $T + U$. Moreover, the expansions considered in this case must all be unitary. In function Expansion-Test-Int, no such care is needed. Another important observation is that in line 5 we may consider only expansions that do not introduce any other vertices of the same group as r, as the test is trivially true for $T + U$ if U contains an edge incident to a vertex in $\mathrm{gr}(r) \setminus \{r\}$.

It is now easy to combine both functions described in a test that determines if a group vertex can be regarded as a nonterminal. Given a group vertex r, we call both functions with r being the initial tree. If both functions return yes, then we know that if r is contained in an optimal solution, there is an optimal solution containing another vertex of $\mathrm{gr}(r)$ distinct from r itself; therefore r can be considered a nonterminal.

As for the running time of our test, it is heavily dependent on the degree of the vertices through which we expand, as well as the maximum depth of the recursion. In practice, we limit the depth of the recursion, truncating the test if necessary, and we also choose not to expand through vertices of higher degree. A more detailed discussion of such issues will be made in section 4.

3.2. Test conditions. In this section we discuss some possible implementations of the function Test-Tree used in the expansion procedure.

From now on, suppose we are given an instance $(G, \mathcal{R}, c)$ of the GST, a group vertex $r \in R$ for some group R, and a tree T that respects r. One of those implementations is based on the following result.

Proposition 3. *If $\tau(L(T)) + \pi(r, L(T) \setminus \{r\}) \leq c(E(T))$, then if T is peripherally contained in some optimal solution, there is an optimal solution containing some vertex of $R \setminus \{r\}$.*

Proof. Suppose T is peripherally contained in some optimal solution T^*. Let I be the set of internal vertices of T and $T' := (T^* - E(T)) - I$. Let $u \in L(T) \setminus \{r\}$ and $r' \in R \setminus \{r\}$ be such that $d_G(u, r') = \pi(r, L(T) \setminus \{r\})$. Connect r' to T' by means of a subpath of a minimum-cost (u, r')-path so that the resulting graph, which we will denote by F, is a forest.

Note that F has at most $|L(T)|$ components and that each such component contains some vertex of $L(T)$ (this is true, since T is peripherally contained in T^*). Moreover, F contains a valid set, since it respects r. Using Lemma 2, it follows that there is a group Steiner tree $\hat{T}$ that contains F and has cost at most

$$c(E(F)) + \tau(L(T)) \leq c(E(T^*)) + \tau(L(T)) + \pi(r, L(T) \setminus \{r\}) - c(E(T)) \leq c(E(T^*)),$$

being, then, an optimal solution to our problem. Since it contains a vertex of $R \setminus \{r\}$, we are done. ∎

If, additionally, r is a leaf in the tree in which T is peripherally contained, we can weaken the hypothesis. In this case, we need not consider r a leaf of T; that is, we may use $\tau(L(T) \setminus \{r\})$ instead of $\tau(L(T))$. This version of the test can be used together with the function Expansion-Test-Leaf.

Another way of proving that assertion (3.1) or (3.2) is true is to show that T is not peripherally contained in any optimal solution. Obviously, this assertion can be modified in accordance with the function being considered. For instance, if we are considering the function Expansion-Test-Leaf, we can show that T is not

peripherally contained in any optimal solution in which r is a leaf. In the case r is a leaf of T, we can use the tests presented in [11] with some minor modifications, including the bound-based tests. For instance, we have the following result.

PROPOSITION 4. *If r is a leaf of T and $\tau(L(T)) < c(E(T))$, then T is not peripherally contained in any optimal solution.*

Proof. Suppose that T is peripherally contained in an optimal solution T^*. Let I be the set of internal vertices of T and $F := (T^* - E(T)) - I$.

Now, since T is peripherally contained in T^*, F is composed of exactly $|L(T)|$ components, each one containing exactly one vertex from $L(T)$. Moreover, since r is a leaf of T and since T respects r, F contains a valid set. It then follows from Lemma 2 that there exists a group Steiner tree $\hat{T}$ containing F, the cost of which is at most

$$c(E(T^*)) + \tau(L(T)) - c(E(T)) < c(E(T^*)),$$

a contradiction. □

Many tests presented in [11] allow one to conclude that there is an optimal solution that does not contain T peripherally. It is important to notice again that Proposition 4 is stronger, allowing one to conclude that T is not peripherally contained in *any* optimal solution. However, several of the tests presented in this same paper can be altered as to allow us to reach the desired conclusion, as is the case of the result just presented.

In the case where r is not a leaf of T we can use a modified version of Proposition 4.

PROPOSITION 5. *Suppose r is not a leaf of T and let p be the cost of a shortest path in G connecting some leaf of T with r. If $\tau(L(T)) + p < c(E(T))$, then T is not peripherally contained in any optimal solution.*

Proof. It is analogous to the proof of Proposition 4. □

It is interesting to observe that we could, instead of using a path of cost p to connect r with a leaf of T, require $\tau(L(T) \cup \{r\}) < c(E(T))$, in which case Lemma 2 could be applied to reconnect the leaves of T and r after we remove T from some optimal solution.

It is also worth pointing out how important it is to assume that T is peripherally contained (instead of only contained) in some optimal solution. This allow us to remove the edges of T and its internal vertices and even then have some information about the structure of the remaining graph. Moreover, since we need only care about reconnecting the leaves of T, we get stronger tests.

Figure 3 contains an example of an instance that can be reduced using the expansion test.

3.3. Correctness of the expansion test. The correctness of the expansion test follows directly from the following proposition.

PROPOSITION 6. *Let T be a tree in G. If the tree T is peripherally contained in some optimal solution T^* of the* GST $(G, \mathcal{R}, c)$*, then for each vertex $u \in L(T)$ which is not a leaf of T^* there exists $U \subseteq \mathcal{X}(u, T)$, $U \neq \emptyset$, such that $T + U$ is peripherally contained in T^*.*

Proof. Let T^* be an optimal solution and assume it peripherally contains T. Choose $u \in L(T)$ such that u is not a leaf of T^* and let $U := \{uv \in E(T^*) \colon v \notin V(T)\}$. Since u is not a leaf in T^*, $U \neq \emptyset$. We claim that $T + U$ is peripherally contained in T^*.

To prove the claim, consider an internal vertex of $T + U$. If this vertex is u, it is immediate that any path from u to a leaf of T^* uses some edge of $T + U$ (since it uses some edge of T or some edge of U). If the internal vertex chosen is not u, then

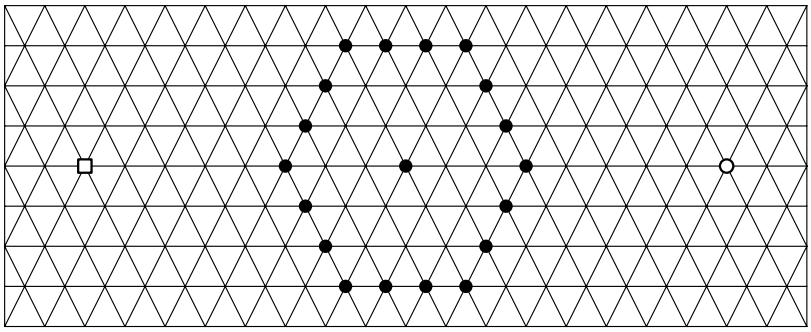


FIG. 3. *In this figure we have an instance of the group Steiner tree problem. The intersections between the segments are vertices and the segments themselves edges of our graph. All costs are unitary. We have tree groups:* •, ∘, *and* □. *The vertex in the center, which belongs to the group* •, *can be removed by the expansion test but not by the simple test of section* 2.2.

it is an internal vertex of T and, since T is peripherally contained in T^*, every path from this vertex to a leaf of T^* uses some edge of T and hence some edge of $T + U$.

Consider now a leaf u of $T + U$. Since $T + U$ is a subtree of T^*, and since T^* is a tree, it is easy to see that there must be a path from u to a leaf of T^* using only edges not in $T + U$. Therefore, $T + U$ is peripherally contained in T^* and the claim is proved. □

Note that the above proposition makes it possible for us to expand through any nonterminal leaf of a tree, since such a vertex cannot be a leaf in any optimal solution. Note also that we do not mention the fact that T respects some group vertex r, although such an assumption is important in the results of the previous section.

3.4. A supporting test. In this section we present a test that follows the same idea of the test presented in Theorem 1. We will then show how to combine this test with the idea of expansion to achieve more reductions. To present the test, we first introduce some notation. Given an instance $(G, \mathcal{R}, c)$ of the GST and disjoint sets A, B of vertices of G, we denote by $\underline{\delta}_G(A, B)$ the set containing for each $u \in B$ a least cost edge from $\delta_G(A, B)$ incident to u, if such an edge exists.

PROPOSITION 7. *Let* $(G, \mathcal{R}, c)$ *be an instance of the* GST, $R \in \mathcal{R}$, *and* $r \in R$. *Let* $C \subseteq V(G)$ *be such that* $C \cap \hat{\mathcal{R}} = \{r\}$. *If for every* $U \subseteq N(C)$ *with* $|U| \geq 1$ *we have* $\tau(U) + \pi(r, U) \leq c(\underline{\delta}(C, U))$, *then there exists an optimal solution containing a vertex of* $R \setminus \{r\}$.

Proof. Let T^* be an optimal solution in which r occurs. Let $C' := C \cap V(T^*)$ and $U := N_{T^*}(C')$. Note that $U \subseteq N_G(C)$, and since $C \cap \hat{\mathcal{R}} = \{r\}$, U is nonempty. Let $T' := T^* - C'$. Choose $u \in U$ and $r' \in R \setminus \{r\}$ such that $d_G(u, r') = \pi(r, U)$. Connect r' to T' by means of a subpath of a minimum-cost (u, r')-path, such that the resulting graph, say F, is a forest.

The graph F has at most $|U|$ components and each such component contains a vertex of U. Moreover, since $C \cap \hat{\mathcal{R}} = \{r\}$, F contains a valid set. Using Lemma 2, we know that there is a group Steiner tree $\hat{T}$ that contains F and that its cost is at most $c(E(F)) + \tau(U)$.

Now, we know that $c(\delta_{T^*}(C')) \geq c(\underline{\delta}_G(C, U))$. Hence, the cost of $\hat{T}$ is at most $c(E(T^*)) + \tau(U) + \pi(r, U) - c(\delta_{T^*}(C')) \leq c(E(T^*))$. Therefore, $\hat{T}$ is also an optimal solution. Now, since it contains a vertex of $R \setminus \{r\}$, we are done. □

We may combine this test with the expansion approach using the following idea. In each call of EXPANSION-TEST-INT we have a tree T and we wish to assert that (3.2)

is true for T. If the tree T is such that $V(T) \cap \hat{\mathcal{R}} = \{r\}$, then we can also check whether $V(T)$ verifies the hypothesis of Proposition 7. If this is the case, then r can be removed and there is no need to continue the expansion or to call the function Expansion-Test-Leaf. In fact, in this case we can end the recursion right away.

4. Computational results. We have implemented all the above methods. We did not implement any of the bound-based tests of [11] mainly because time-efficient heuristics for finding lower bounds were unable to find good lower bounds for the problem.

In our implementation, we observed that the time consumed by the expansion test is highly dependent on the degree of the vertices through which we choose to expand and on the depth of the recursion itself. Therefore, we limit these parameters to a small constant. In our case, we expand only through vertices of degree at most 8 and we limit the depth of the recursion to 8. When using the test of section 3.4, we consider only sets C for which $|N(C)|$ is limited by a small constant (8, in our case).

The test described in section 3.4 is effective in many cases, and reductions can be achieved by using it. We also attempted to use this test alone, together with another mechanism of expansion. In this case the expansion can be carried out in many ways; the only thing needed is to guarantee the necessary properties of the subset of vertices being inspected in each step. For instance, we may run a breadth-first search from r and, in each step of the search, test the set of vertices visited by the search. In all cases, we observed that the combination of this test with the expansion scheme presented earlier is far more effective.

The most important practical issue to consider, however, is the calculation of the group bottleneck Steiner distances. If we have a fixed valid set X, we can compute $s_X(u, v)$ for each pair u, v of vertices in time $O(n^2|X|)$ (cf. Duin [2]), given that we have the distances in G already computed. Since computing the distances takes time $O(n^3)$ (if the original graph is sparse, there are better algorithms), the whole process takes time $O(n^3)$.

Rohe and Zachariasen [13] show that computing $s(u, v)$ is NP-hard. They also propose heuristics to compute upper bounds for $s(u, v)$. We used those heuristics in our implementation and verified that the bounds produced by them are good, especially when the diameter of the groups is small. After removing group vertices, the group bottleneck Steiner distances only decrease. This means that we need not update the distances, though updates could lead to more reductions. In our implementation, we noticed that frequent updates did little to improve the number of reductions, so we try to remove as many group vertices as we can before updating. More specifically, we run our reduction algorithm until we can no longer remove any vertex; then we update the distances and try again.

Our results are summarized in Tables 1 and 2. The instances used are the same as in [13]. We did not have access to the original instances, however, so our reduction algorithm was tested upon instances that had already gone through a reduction process. Therefore, the reductions we achieved are in addition to those already achieved by the tests in [13]. The average number of group vertices removed was 6.5% for the WRP3 collection and 5.98% for the WRP4 collection.

In our tables we do not show time consumption, though it was never above 10 minutes. By changing the parameters of our program, like the maximum depth of the recursion, we observed that the time consumed increases too much, whereas there is no significant increase in the number of reductions achieved. The tests were carried out on a 1.1 GHz AMD Athlon with 512 Mb RAM memory.

TABLE 1

This table shows the results of the reduction procedure for the WRP3 collection. Column "$|\hat{\mathcal{R}}|$" has the number of group vertices of the original instance. Column "Exp." contains the number of group vertices removed by the expansion procedure alone. Column "Supp." contains the number of group vertices removed by the supporting test presented in section 3.4. *The last column, "Total," has the total number of group vertices removed.*

Reductions achieved: WRP3 collection									
Instance	$\|\hat{\mathcal{R}}\|$	Exp.	Supp.	Total	Instance	$\|\hat{\mathcal{R}}\|$	Exp.	Supp.	Total
wrp3-11	34	1	5	6	wrp3-50	229	3	7	10
wrp3-12	37	5	9	14	wrp3-52	144	4	8	12
wrp3-13	64	1	6	7	wrp3-53	161	2	15	17
wrp3-14	64	0	13	13	wrp3-55	185	3	9	12
wrp3-15	57	12	16	28	wrp3-56	149	0	8	8
wrp3-16	46	4	6	10	wrp3-60	296	0	3	3
wrp3-17	82	0	2	2	wrp3-62	174	2	6	8
wrp3-19	67	3	10	13	wrp3-64	244	3	5	8
wrp3-20	55	4	6	10	wrp3-66	216	1	2	3
wrp3-21	76	5	14	19	wrp3-67	215	5	14	19
wrp3-22	69	2	10	12	wrp3-69	192	1	7	8
wrp3-23	51	0	0	0	wrp3-70	294	6	16	22
wrp3-24	83	14	17	31	wrp3-71	266	4	11	15
wrp3-25	81	0	8	8	wrp3-73	192	3	9	12
wrp3-26	91	1	2	3	wrp3-74	183	2	10	12
wrp3-27	106	3	7	10	wrp3-75	213	2	6	8
wrp3-28	77	2	8	10	wrp3-76	255	6	18	24
wrp3-29	65	2	3	5	wrp3-78	367	3	12	15
wrp3-30	108	2	5	7	wrp3-79	175	1	1	2
wrp3-31	80	1	8	9	wrp3-80	224	4	11	15
wrp3-33	104	0	2	2	wrp3-83	336	1	5	6
wrp3-34	170	6	9	15	wrp3-84	290	2	4	6
wrp3-36	101	2	8	10	wrp3-85	197	0	0	0
wrp3-37	164	2	5	7	wrp3-86	235	4	13	17
wrp3-38	153	1	3	4	wrp3-88	253	2	6	8
wrp3-39	363	0	2	2	wrp3-91	266	7	11	18
wrp3-41	83	0	1	1	wrp3-92	409	1	6	7
wrp3-42	134	2	10	12	wrp3-94	351	4	16	20
wrp3-43	85	0	0	0	wrp3-95	327	2	48	50
wrp3-45	242	2	7	9	wrp3-96	396	2	10	12
wrp3-48	140	2	7	9	wrp3-98	420	12	23	35
wrp3-49	251	4	18	22	wrp3-99	348	12	17	29

TABLE 2
Test results for the WRP4 collection. The meaning of the columns is the same as for Table 1.

Reductions achieved: WRP4 collection									
Instance	$\lvert\hat{\mathcal{R}}\rvert$	Exp.	Supp.	Total	Instance	$\lvert\hat{\mathcal{R}}\rvert$	Exp.	Supp.	Total
wrp4-11	41	1	2	3	wrp4-44	174	4	14	18
wrp4-13	41	0	9	9	wrp4-45	180	0	3	3
wrp4-14	52	3	4	7	wrp4-46	183	0	0	0
wrp4-15	57	1	4	5	wrp4-47	186	8	10	18
wrp4-16	47	2	8	10	wrp4-48	140	3	12	15
wrp4-17	64	12	20	32	wrp4-49	195	10	21	31
wrp4-18	53	1	6	7	wrp4-50	198	1	14	15
wrp4-19	51	1	4	5	wrp4-51	202	2	6	8
wrp4-21	107	10	17	27	wrp4-52	208	1	4	5
wrp4-22	86	0	3	3	wrp4-53	211	0	4	4
wrp4-23	91	0	0	0	wrp4-54	215	2	2	4
wrp4-24	120	19	27	46	wrp4-55	222	12	19	31
wrp4-25	88	3	8	11	wrp4-56	238	20	29	49
wrp4-26	104	2	6	8	wrp4-58	233	7	8	15
wrp4-27	107	0	0	0	wrp4-59	239	2	5	7
wrp4-28	110	0	0	0	wrp4-60	238	4	9	13
wrp4-29	114	0	0	0	wrp4-61	246	8	10	18
wrp4-30	118	0	3	3	wrp4-62	281	11	24	35
wrp4-31	124	1	3	4	wrp4-63	250	1	8	9
wrp4-32	130	1	5	6	wrp4-64	257	3	9	12
wrp4-33	97	1	5	6	wrp4-66	280	4	13	17
wrp4-34	136	0	0	0	wrp4-67	376	16	57	73
wrp4-35	139	0	0	0	wrp4-68	273	6	6	12
wrp4-36	143	0	0	0	wrp4-69	274	17	24	41
wrp4-37	148	0	0	0	wrp4-70	278	8	14	22
wrp4-38	150	0	0	0	wrp4-71	285	0	4	4
wrp4-39	200	24	55	79	wrp4-72	295	6	16	22
wrp4-40	158	5	5	10	wrp4-73	303	5	7	12
wrp4-41	164	0	0	0	wrp4-74	295	0	6	6
wrp4-42	168	0	0	0	wrp4-75	299	5	8	13
wrp4-43	170	15	16	31	wrp4-76	305	4	10	14

Acknowledgment. We wish to thank the anonymous referees for numerous suggestions that helped improve the quality of the text.

REFERENCES

[1] F. M. DE OLIVEIRA FILHO, *O problema de Steiner em grafos*, Master's dissertation, Universidade de São Paulo, São Paulo, Brazil, 2005.
[2] C. W. DUIN, *Steiner's Problem in Graphs*, Ph.D. thesis, Universiteit van Amsterdam, Amsterdam, The Netherlands, 1994.
[3] C. W. DUIN AND A. VOLGENANT, *Reduction tests for the Steiner problem in graphs*, Networks, 19 (1989), pp. 549–567.
[4] C. W. DUIN, A. VOLGENANT, AND S. VOSS, *Solving group Steiner problems as Steiner problems*, European J. Oper. Res., 154 (2004), pp. 323–329.
[5] U. FEIGE, *A threshold of* $\ln n$ *for approximating set cover*, J. ACM, 45 (1998), pp. 634–652.
[6] N. GARG, G. KONJEVOD, AND R. RAVI, *A polylogarithmic approximation algorithm for the group Steiner tree problem*, J. Algorithms, 37 (2000), pp. 66–84.
[7] E. HALPERIN AND R. KRAUTHGAMER, *Polylogarithmic inapproximability*, in STOC '03: Proceedings of the 35th Annual ACM Symposium on Theory of Computing, ACM, New York, 2003, pp. 585–594.
[8] E. IHLER, *The complexity of approximating the class Steiner tree problem*, in Graph-Theoretic Concepts in Computer Science, Lecture Notes in Comput. Sci. 570, Springer, Berlin, 1992, pp. 85–96.
[9] T. KOCH, A. MARTIN, AND S. VOSS, *SteinLib: An updated library on Steiner tree problems in graphs*, in Steiner Trees in Industry, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2001, pp. 285–325.
[10] T. POLZIN, *Algorithms for Steiner Problems in Networks*, Ph.D. thesis, University of Saarland, Saarbrücken, Germany, 2003.
[11] T. POLZIN AND S. V. DANESHMAND, *Extending reduction techniques for the Steiner tree prob-*

lem, in Algorithms—ESA 2002, Springer, Berlin, 2002, pp. 795–807.

[12] G. Reich and P. Widmayer, *Beyond Steiner's problem: A VLSI oriented generalization*, in Graph-Theoretic Concepts in Computer Science, Lecture Notes in Comput. Sci. 411, Springer, Berlin, 1990, pp. 196–210.

[13] A. Rohe and M. Zachariasen, *Rectilinear group Steiner trees and applications in VLSI design*, Math. Program., 94 (2003), pp. 407–433.

[14] E. Uchoa, M. P. Aragao, and C. Ribeiro, *Preprocessing Steiner problems from VLSI layout*, Networks, 40 (2002), pp. 38–50.

SIAM J. OPTIM.
Vol. 17, No. 4, pp. 1189–1204
© 2006 Society for Industrial and Applied Mathematics

THE IMAGE CONTAINMENT PROBLEM AND SOME CLASSES OF POLYNOMIAL INSTANCES*

RAFFAELE PESENTI† AND FRANCA RINALDI‡

Abstract. The image containment problem (ICP) is a minimum cost design problem concerning the containment of particular polyhedra, called zonotopes, that are images of boxes through linear transformations. The ICP is NP-hard. Here we study a family of nontrivial ICP instances, called worst case demand (WCD) instances. We prove that such instances can be recognized and solved in polynomial time via linear programming. Then we characterize the classes of instances that are WCD independently on the choice of the cost vector (structurally worst case demand classes (SWCD)) and we show that recognizing whether a class of instances is SWCD is a coNP-complete problem. Finally, we describe two families of SWCD classes that are interesting from an applicative point of view: the classes defined by the incidence matrices of particular directed graphs and those defined by pre-Leontief matrices.

Key words. containment of polyhedra, zonotopes, parallelotopes, pre-Leontief matrices, network design

AMS subject classifications. 90C08, 90C35, 68Q25

DOI. 10.1137/040606843

1. Introduction.

The *image containment problem (ICP)* is a design problem concerning the containment of polyhedra and, more specifically, of *zonotopes*.

Given two vectors $d^+ \in \mathbb{Q}_+^q$ and $u^+ \in \mathbb{Q}_+^m$, define the *boxes* $D(d^+) \doteq \{d \in \mathbb{R}^q : 0 \le d \le d^+\}$ and $U(u^+) \doteq \{u \in \mathbb{R}^m : 0 \le u \le u^+\}$.

Image containment problem (ICP). *Let $B \in \mathbb{Q}^{n\times m}$ and $F \in \mathbb{Q}^{n\times q}$ be two matrices and $d^+ \in \mathbb{Q}_+^q$ and $c \in \mathbb{Q}_+^m$ be two vectors with strictly positive components. Find, when it exists, a vector $u^+ \in \mathbb{Q}_+^m$ of minimum cost cu^+ such that*

$$\forall d \in D(d^+) \;\; \exists u \in U(u^+) : \quad Fd = Bu. \tag{1.1}$$

Denoting by $FD(d^+) = \{x \in \mathbb{R}^n : \exists d \in D(d^+) \text{ s.t. } x = Fd\}$ and $BU(u^+) = \{x \in \mathbb{R}^n : \exists u \in U(u^+) \text{ s.t. } x = Bu\}$ the images of $D(d^+)$ and $U(u^+)$ through the linear transformations induced by F and B, respectively, the ICP may be equivalently expressed as

$$\begin{aligned} z_{ICP} &= \min cu^+, \\ FD(d^+) &\subseteq BU(u^+), \\ u^+ &\ge 0. \end{aligned}$$

So, the ICP is the problem of determining the minimum cost box $U(u^+)$ such that its image in $\mathbb{R}^n$ through the linear transformation B contains the image $FD(d^+)$ of the box $D(d^+)$. Note that the assumption $d^+ > 0$ is not restrictive, since one can

*Received by the editors April 16, 2004; accepted for publication (in revised form) June 12, 2006; published electronically December 26, 2006. This work was supported by MIUR project PRIN "Analysis, Optimization, and Coordination of Logistic and Production Systems."
http://www.siam.org/journals/siopt/17-4/60684.html
†Dipartimento di Ingegneria Informatica, Università di Palermo, v.le delle Scienze, 90128 Palermo, Italy (pesenti@unipa.it).
‡Dipartimento di Matematica e Informatica, Università di Udine, via delle Scienze 206, 33100 Udine, Italy (rinaldi@dimi.uniud.it).

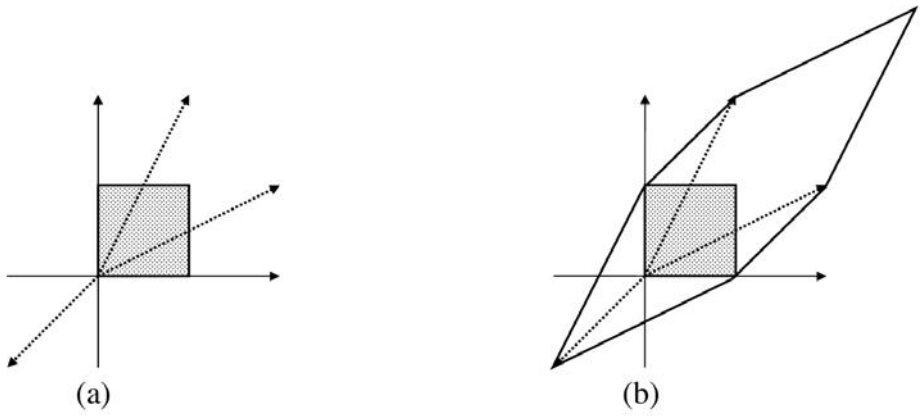


FIG. 1.1. *An ICP instance and its optimal solution.*

always eliminate possible null components of d^+ (and the corresponding columns of F) without modifying the set $FD(d^+)$.

Both sets $BU(u^+)$ and $FD(d^+)$ are zonotopes, i.e., centrally symmetric polytopes that are vector sums (Minkowsky sums) of a finite number of line segments (see [9, 14, 18]). Indeed we may write $BU(u^+) = \sum_{i=1}^{m} [0, u_i^+] B^i$ and $FD(d^+) = \sum_{j=1}^{q} [0, d_j^+] F^j$, where B^i and F^j denote the generic columns of B and F, respectively, and $[0, \alpha]$, $\alpha \in \mathbb{R}_+$, denotes the interval with extremes 0 and α. Therefore, $BU(u^+)$ is a sum of m line segments (called *generators*), where the ith segment has an end in the origin, the direction of the column B^i of B, and length equal to $u_i^+ \|B^i\|$. In this perspective, the cost cu^+ is a weighted sum of the lengths of the generators of $BU(u^+)$. In particular, the sum of the lengths of the generators, $\sum_{i=1}^{m} u_i^+ \|B^i\|$, is called the total length of the zonotope [10]. Figure 1.1 reports an ICP instance (Figure 1.1(a)) and its minimum total length solution (Figure 1.1(b)) where the zonotope $FD = ID$ is represented as a shaded square and the unit directions B^i, $i = 1, 2, 3$, as dotted vectors.

Since in the definition of the ICP the vector d^+ is assigned, in the rest of the paper we also use the notation D instead of $D(d^+)$. In addition, we refer to vectors $d \in D$ as *demands*.

The ICP is a particular case of the *circumbody* containment problem, a relevant subject of *computational convexity* (see [7]). In this setting, the *circumbody* containment problem requires finding, given a class $\mathcal{C}$ of closed convex sets, an element of $\mathcal{C}$ that contains a given convex body K and minimizes an assigned functional $\omega : \mathcal{C} \to \mathbb{R}$, where ω is assumed nonnegative and monotone with respect to inclusion. The ICP corresponds to the case where the elements of $\mathcal{C}$ are zonotopes of the form $BU(u^+)$, $u^+ \in \mathbb{R}_+^m$, the function ω is defined by $\omega(BU(u^+)) = cu^+$, and K is a zonotope of the form $K = FD(d^+)$.

Circumbody containment problems involving zonotopes were studied, for instance, in [10] and [16]. In [10] the authors consider the problem of finding a minimum total length zonotope containing k assigned points of $\mathbb{R}^n$ when the m directions of the generators are given and the center of the zonotope can be chosen arbitrarily. Assuming k greater than m^{n-1}, they show that the problem can be solved via linear programming in $O(km^{n-1} + m^{O(n)})$. In [16] only zonotopes in $\mathbb{R}^2$ centered in the origin are considered and a method to approximate a given zonotope by means of zonotopes

with assigned unit generators is proposed. As stated by the authors, this method converges to an optimal approximation with respect to the Hausdorff/integral metric. Differently from the above cases, in the ICP the center $\frac{Bu^+}{2}$ of the zonotope $BU(u^+)$ depends on the optimal solution u^+ and thus cannot be chosen arbitrarily nor is a priori known.

The ICP may arise in several applicative contexts. It models, for instance, those situations in which a system is subject to two kinds of actions: exogenous inputs and internal decisions. In this context, the vectors $d \in D$ represent exogenous inputs (demands, disturbances, etc.) that are supposed *unknown but bounded*; i.e., they can assume any nonnegative value not exceeding an assigned upper bound d^+. Any input d induces the effect Fd on the system. The vectors $u \in U(u^+)$ represent the internal decisions that have to *counteract* the demand action by producing an effect $Bu = Fd$. So the ICP consists of finding a minimum cost box $U(u^+)$ of decision vectors that can counteract any possible demand. As shown in [4], the condition $FD \subseteq BU(u^+)$ is a necessary and sufficient condition for the existence of feedback stabilizing strategies for the system on an infinite time horizon. That paper also contains an applicative example of the ICP in the context of manufacturing systems.

A particular case of the ICP, where B and F are the incidence matrices of two networks G and H with the same node set, was introduced in [3] and called the *minimum cost network containment problem (MCNCP)*. Given a capacity vector d^+ on the arcs of H, the MCNCP is the problem of determining a minimum cost capacity vector u^+ on the arcs of G with the property that, for any flow $d \in D(d^+)$, there exists a feasible flow $u \in U(u^+)$ in G that counteracts d, i.e., causes the same imbalance $Bu = Fd$ at the nodes. Actually, the MCNCP is coNP-hard [13]. An exact algorithm for the MCNCP was proposed in [15], where the role of this problem in the management of power transmission networks was also outlined.

The ICP may also be interpreted as a minimum cost substitution problem as in the following example. Given a production-distribution system, let F be the incidence matrix of the hypergraph representing the production and/or the distribution lines and d^+ be the vector whose components are the upper capacities of these lines. Then the feasible system outputs y have the form $y = Fd$, $0 \leq d \leq d^+$, where d is a vector of material flows. Suppose that the system has to be renewed and that the new structure of the processes is represented by the matrix B. Then the condition $FD(d^+) \subseteq BU(u^+)$ requires that the capacities u^+ of the new lines allow the production of at least the same set of outputs as the old system.

The ICP extends the MCNCP to the case of general matrices B and F and thus it is coNP-hard as well. In this paper, we deal with a particular subset of ICP instances for which a *worst case demand (WCD)* exists, in the sense stated in the following definition.

DEFINITION 1.1. *An ICP instance is called* worst case demand (WCD) *instance if there exists a vector* $\hat{d} \in D$ *and an optimal solution* $\hat{u}$ *of* $\{\min cu : Bu = F\hat{d}, u \geq 0\}$ *such that* $FD \subseteq BU(\hat{u})$. *In this case the vector* $\hat{d}$ *and the solution* $\hat{u}$ *are called* worst case demand *and* worst case hyperflow*, respectively.*

In other words, for WCD instances, the minimum cost one has to pay to counteract the single demand $\hat{d}$ is the same one has to pay to be able to counteract any demand in D and one can choose $u^+ = \hat{u}$ as the optimal solution of the ICP. In general, an ICP instance is not WCD. For example, Figure 1.2 (a), (b), and (c) report, respectively, the three minimum total length zonotopes that include, separately, each of the three nontrivial vertices $(0, 1)$, $(1, 1)$, and $(1, 0)$ of $FD = ID$ for the instance

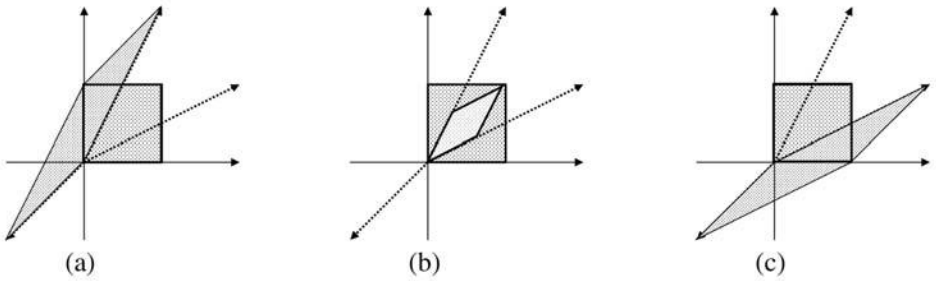


FIG. 1.2. *Optimal zonotopes including, separately, the three vertices of FD for the instance in Figure* 1.1(a).

in Figure 1.1(a). Clearly, FD is not contained in any of these three sets (nor in the minimum total length zonotope that contains any assigned point of FD), and in fact the instance is not WCD.

The main results of the paper are the following. First, we prove that the ICP can be solved in polynomial time on the set of the WCD instances and that the problem of recognizing whether an ICP instance is WCD can be solved in polynomial time as well. In general, for assigned matrices B and F, we show that the WCD property may depend on the choice of the cost vector c while it does not depend on the vector d^+. Then, as a second result, we provide some necessary and sufficient conditions on the structure of the matrices B and F that make the corresponding instances WCD independently of the particular cost vector c. In this case, we say that B and F define a *structurally worst case demand (SWCD)* class of instances. We first use the above conditions to show that any ICP class defined by a matrix B whose columns are linearly independent is SWCD. Then we identify two other families of SWCD classes that have significant interest from an applicative point of view. The first family contains all the MCNCP classes of instances that are SWCD. More precisely, we characterize the structure of the networks G and H that define a class with this property. This result enables us to show, in particular, that the problem of deciding whether a given class of MCNCP instances is SWCD is coNP-complete. Thus, the more general problem of deciding whether an ICP class is SWCD is coNP-complete as well. The second family is defined by matrices B that are pre-Leontief [2] and nonnegative matrices F. This family includes, in particular, the case when B describes either a network system or a system of generalized flows. We also prove that the instances of the two families admit an integral optimal solution whenever d^+ is an integral vector. This property does not in general hold, even for the MCNCP instances [3].

The outline of the paper is the following. In section 2, we present some preliminary results and assumptions concerning the ICP. In section 3, we show how to recognize and solve WCD instances in polynomial time. In section 4, we introduce the SWCD classes of instances and provide some necessary and sufficient conditions on the pairs (B, F) of matrices that define such classes. We also show that these conditions are satisfied when B is a nonsingular matrix. In section 5, we analyze the two families of SWCD classes mentioned above. Finally, in section 6, we discuss the possible extension of the previous results to a particular generalization of the ICP.

Notation. The following notation will be used throughout the paper. Given an $n \times m$ matrix A, we write $A \geq 0$ and $A > 0$ to mean that all the entries of A are

nonnegative and strictly positive, respectively. Given two integers i, k and a set of integers β, we denote by A_{ik}, A^k, and A^β the (i,k)-entry of A, the kth column of A, and the submatrix of A formed by the columns A^k, $k \in \beta$, respectively. Moreover, we denote by $lin\{A\} = \langle A^k : k = 1, \dots, m\rangle$ and $cone\{A\} = cone\{A^k : k = 1, \dots, m\}$ the linear space and the convex cone, respectively, generated by the columns of A. If $m \geq n$, we call any $n \times n$ nonsingular submatrix $\mathcal{B}$ of A a basis submatrix. Given two vectors a and b we write $a \gneq b$ ($a \lneq b$) to mean that, componentwise, $a_k \geq b_k$ ($a_k \leq b_k$) and at least one inequality strictly holds, i.e., $a \neq b$. Given a vector space $\mathbb{R}^q$, we denote by e^j, $j = 1, \dots, q$, the vector defined by $e^j_k = 0$ if $k \neq j$, $e^j_j = 1$ and by e the vector with components $e_k = 1$ for each k. Given a directed hypergraph H with n nodes and m hyperarcs, we call the $n \times m$ matrix A whose kth column corresponds to the hyperarc $h_k = (T, H)$ and has entries $A_{ik} = -1$ if $i \in T$, $A_{jk} = 1$ if $j \in H$, 0 otherwise, the incidence matrix of H.

2. Preliminary results. Denote the generic ICP instance by the quadruple (B, F, c, d^+). The following lemma states a necessary and sufficient condition for the feasibility of the ICP.

LEMMA 2.1. *An ICP instance (B, F, c, d^+) admits a feasible solution if and only if*

$$cone\{F^j : j = 1, \dots, q\} \subseteq cone\{B^i : i = 1, \dots, m\}. \tag{2.1}$$

Proof. If (B, F, c, d^+) is feasible, then for each $j = 1, \dots, q$ there exists $u^j \geq 0$ such that $F(0, \dots, 0, d^+_j, 0, \dots, 0)^T = d^+_j F^j = Bu^j$; thus each F^j is a conic combination of the columns of B. This implies condition (2.1). Conversely, if condition (2.1) holds, then FD is a bounded subset of $cone\{B^i : i = 1, \dots, m\}$ and thus there exists $\bar{u} \geq 0$ such that $FD \subseteq \{Bu : 0 \leq u \leq \bar{u}\}$. Clearly, the vector $\bar{u}$ is a feasible solution of the ICP. ∎

LEMMA 2.2. *Any feasible ICP instance (B, F, c, d^+) with $B \in \mathbb{Q}^{n \times m}$ can be transformed in an equivalent instance (B', F', c, d^+), $B' \in \mathbb{Q}^{k \times m}$, such that* $\operatorname{rank}(B) = \operatorname{rank}(B') = k$.

Proof. Assume $\operatorname{rank}(B) = k < n$ and let $V = \begin{bmatrix} W \\ Z \end{bmatrix}$ be an $n \times n$ matrix where the rows of the $k \times n$ matrix W span $lin\{B\}$, and the rows of the $(n-k) \times n$ matrix Z span the orthogonal space $lin\{B\}^\perp$. Since V is nonsingular, $Bu = Fd$ if and only if $VBu = VFd$. Moreover, $ZBu = 0$ for each $u \in \mathbb{R}^m$ and this implies, by Lemma 2.1, $ZFd = 0$ for each $d \in \mathbb{R}^q_+$. Therefore the condition $FD \subseteq BU(u^+)$ is equivalent to the condition $WFD \subseteq WBU(u^+)$ and the instance (B, F, c, d^+) is equivalent to the instance (WB, WF, c, d^+) where WB is a $k \times m$ matrix of rank k. ∎

Since in general the orthogonal space $lin\{F\}^\perp$ is not contained in $lin\{B\}^\perp$, the argument followed in the proof of Lemma 2.2 cannot be used to assume that F is a full rank matrix. For example, consider the case where B is the 2×2 identity matrix I, $F = [1, 1]^T$, and $d^+ = 1$. For the choice $u^+ = [2, 0]$ the zonotope $BU(u^+) = [0, 2]e^1 + [0, 0]e^2$ does not contain the segment $FD(d^+) = [0, 1]e$ while the projection of $BU(u^+)$ onto the linear space $lin\{F\}$ is equal to (and hence includes) $FD(d^+)$.

By Lemma 2.2 we can always assume that B is an $n \times m$ matrix with $n \leq m$ and full rank. When B is a nonsingular square matrix, the solutions $BU(u^+)$ are parallelotopes, i.e., zonotopes whose generators are linearly independent. If, in particular, the columns of B form an orthogonal basis, these parallelotopes are, possibly not axis parallel, boxes. Then, if B is nonsingular, the ICP is the problem of finding the minimum cost parallelotope with assigned generator directions and a vertex in the origin that encloses the zonotope FD. The following result holds.

THEOREM 2.3. *Any ICP instance* (B, F, c, d^+) *where* B *is a nonsingular square matrix can be solved in polynomial time.*

Proof. For each $u \in \mathbb{R}^m$, we can obtain the representation of the parallelotope $BU(u)$ in terms of a set of $2n$ hyperplanes as $BU(u) = \{x \in \mathbb{R}^n : A_i x \leq b_i, i = 1, \ldots, 2n\}$ in polynomial time [14]. Therefore, the separation problem of finding a vector $d \in D$ such that $Fd \notin BU(u)$, if such a vector exists, can be solved in polynomial time by solving the $2n$ linear programming problems $\max\{A_i F d : d \in D\}$, $i = 1, \ldots, 2n$. Then the thesis follows by a well-known result of Grötschel, Lovász, and Schrijver [8], stating the equivalence between the complexity of an optimization problem and the corresponding separation problem. □

3. WCD instances. In this section, we first reformulate the ICP as a linear programming problem with an exponential number of variables and constraints. Then, we introduce a lower bound and an upper bound for the latter problem and show that an ICP instance is WCD if and only if these two bounds are equal.

Due to the convexity of the sets $BU(u^+)$ and FD, the ICP may be equivalently expressed as a linear programming problem (with an exponential number of variables and constraints) in the following form.

ICP_{LP}:

$$\begin{aligned} & z_{ICP} = \min cu^+, \\ & Bu^r = Fd^r \quad \forall r : d^r \in \text{Ext}\{D\}, \\ & 0 \leq u^r \leq u^+ \quad \forall r : d^r \in \text{Ext}\{D\}, \end{aligned} \tag{3.1}$$

where $\text{Ext}\{D\}$ is the set of the vertices of the set D.

A lower bound on the optimal value z_{ICP} can be found by solving, for any $d \in D$, the following relaxation of (3.1).

Lower bound problem ($\text{LBP}(d)$):

$$\begin{aligned} z_{LBP}(d) &= \min cu, \\ Bu &= Fd, \\ u &\geq 0. \end{aligned} \tag{3.2}$$

According to Definition 1.1, an ICP instance is WCD if there exists a worst case demand $\hat{d} \in D$ such that problems ICP_{LP} and $\text{LBP}(\hat{d})$ have the same optimal value and at least one common optimal solution (the so-called worst case hyperflow) $\hat{u}$. However, as shown in the following example, even in the WCD case, an optimal solution of the problem $\text{LBP}(\hat{d})$ may not be a worst case hyperflow.

Example 3.1. Consider the ICP instance (B, F, c, d^+) where B and F are, respectively, the incidence matrices of the networks $G = (V, E_G)$ and $H = (V, E_H)$ shown in Figure 3.1. The two networks share the node set $V = \{1, 2, 3, 4, 5, 6\}$, while the two arc sets are $E_G = \{(1,2), (1,5), (2,4), (3,2), (3,6), (4,5), (4,6)\}$ and $E_H = \{(1,6), (3,5)\}$. Let $c_{ij} = 1$ for all $(i,j) \in E_G$ except for $c_{15} = c_{36} = 3$ and let D be the box defined by $d^+ = (1,1)$.

Both the settings

$$\hat{u}_{12} = \hat{u}_{32} = \hat{u}_{45} = \hat{u}_{46} = 1, \qquad \hat{u}_{24} = 2, \qquad \hat{u}_{ij} = 0 \ \text{ otherwise} \tag{3.3}$$

and

$$\bar{u}_{15} = \bar{u}_{36} = 1, \qquad \bar{u}_{ij} = 0 \ \text{ otherwise} \tag{3.4}$$

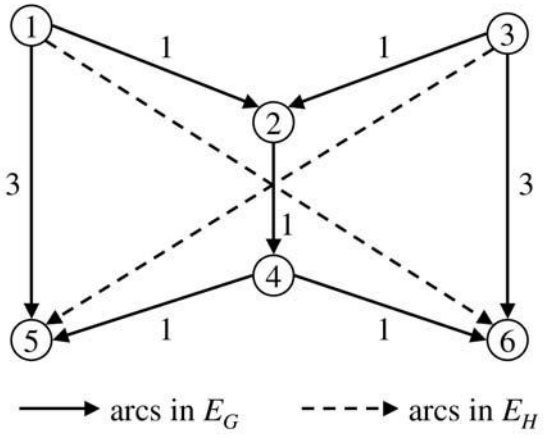


FIG. 3.1. *Networks G and H in Example* 3.1.

define optimal solutions (of cost 6) to problem (3.2) with $\hat{d} = d^+$. Nevertheless, while the solution $\hat{u}$ is feasible for the ICP, the solution $\bar{u}$ is not, since no flow in $U(\bar{u})$ can counteract, for instance, the demand $d = (0, 1)$.

Now consider, for any $d \in D$, the following linear programming problem.

Upper bound problem (UBP(d)):

$$
\begin{aligned}
z_{UBP}(d) &= \min cu, \\
Bu^j &= Fd^j, \quad j = 1, \ldots, q, \\
u &= \sum_{j=1}^{q} u^j, \\
u^j &\geq 0, \quad j = 1, \ldots, q,
\end{aligned}
\tag{3.5}
$$

where d^j, $j = 1, \ldots, q$, denotes the vector of D defined by $d^j_k = 0$ if $k \neq j$ and $d^j_j = d_j$.

It is immediate to verify that, for each $d \in D$, problem UBP(d) can be decomposed in the sum of q independent problems LBP(d^j), $j = 1, \ldots, q$, and that $z_{UBP}(d) = \sum_{j=1}^{q} z_{LBP}(d^j)$. Moreover, the following lemma holds.

LEMMA 3.2. *Given an ICP instance* (B, F, c, d^+), *any feasible solution* $\hat{u} = \sum_{j=1}^{q} \hat{u}^j$ *of UBP*(d^+) *is a feasible solution of the* ICP_{LP}. *Thus,* $z_{ICP} \leq z_{UBP}(d^+)$.

Proof. Any vertex $\bar{d}$ of D may be written as a sum of vertices d^{+j} for j in a suitable set $J \subseteq \{1, \ldots, q\}$. Then the vector $\bar{u} = \sum_{j \in J} \hat{u}^j$ is such that $B\bar{u} = \sum_{j \in J} B\hat{u}^j = \sum_{j \in J} Fd^{+j} = F\bar{d}$ and $0 \leq \bar{u} \leq \hat{u}$. Hence, $F\bar{d} \in BU(\hat{u})$ for each $\bar{d} \in \text{Ext}\{D\}$ and $FD \subseteq BU(\hat{u})$. ∎

A particular consequence of the above lemma is that the existence of a feasible solution u^j of the problem LBP(d^{+j}) for each $j = 1, \ldots, q$ is a necessary and sufficient condition for the existence of a feasible solution of the correspondent ICP problem. It is easy to realize that this condition is an alternative formulation of condition (2.1).

We have proved that for any ICP instance it holds that

$$
z_{LBP}(d^+) \leq z_{ICP} \leq z_{UBP}(d^+). \tag{3.6}
$$

The rest of this section is devoted to characterizing the WCD instances as those instances for which both conditions in (3.6) hold as equalities. To this aim we need the following lemma.

LEMMA 3.3. *If an ICP instance (B, F, c, d^+) is WCD, then $F\hat{d} = Fd^+$ for every worst case demand $\hat{d}$. In particular, d^+ is a worst case demand.*

Proof. Assume that $\hat{d}$ is a worst case demand with $F\hat{d} \neq Fd^+$ and define the vector $\bar{d}$ as $\bar{d} \doteq d^+ - \hat{d}$. Since $\hat{d} \neq d^+$, $\bar{d}$ satisfies the relations $0 \lneq \bar{d} \leq d^+$. If $\hat{u}$ is a worst case hyperflow, there exist two vectors u° and $\bar{u}$, $0 \leq u^\circ, \bar{u} \leq \hat{u}$, such that $Bu^\circ = Fd^+$ and $B\bar{u} = F\bar{d}$. From $F\hat{d} \neq Fd^+$ it follows that $\hat{u} \neq u^\circ$, and, in particular, $\hat{u} \gneq u^\circ$. The latter condition, together with $\bar{u} \geq 0$, implies $\bar{\bar{u}} \doteq \bar{u} + \hat{u} - u^\circ \gneq 0$ and $\bar{\bar{u}}_j = 0$ if $\hat{u}_j = 0$. Moreover, from $B(u^\circ - \hat{u}) = F(d^+ - \hat{d}) = F\bar{d} = B\bar{u}$ we obtain $B\bar{\bar{u}} = 0$. Then, there exists an $\alpha > 0$ such that $\alpha\bar{\bar{u}} \leq \hat{u}$, $\alpha\bar{\bar{u}} \neq 0$, and $B\alpha\bar{\bar{u}} = \alpha B\bar{\bar{u}} = 0$. From $B(\hat{u} - \alpha\bar{\bar{u}}) = B\hat{u} = F\hat{d}$ and $0 \lneq \hat{u} - \alpha\bar{\bar{u}} \lneq \hat{u}$ it follows that $\hat{u} - \alpha\bar{\bar{u}}$ is a feasible solution of problem LBP($\hat{d}$). In addition, since c is a strictly positive vector, we have $c(\hat{u} - \alpha\bar{\bar{u}}) < c\hat{u}$ in contradiction with the optimality of $\hat{u}$ to problem LBP($\hat{d}$). ∎

We now prove the main result of this section.

THEOREM 3.4. *An ICP instance (B, F, c, d^+) is WCD if and only if problems LBP(d^+) and UBP(d^+) have the same optimal value. In this case, any optimal solution u^* of UBP(d^+) is also an optimal solution of ICP_{LP}.*

Proof. Necessity. Let $\hat{u}$ denote a worst case hyperflow. By Lemma 3.3, $B\hat{u} = Fd^+$. In addition, there exist vectors u^j, $j = 1, \ldots, q$, with $0 \leq u^j \leq \hat{u}$ and $Bu^j = Fd^{+j}$. Since $\sum_{j=1}^q u^j$ is a feasible solution of UBP(d^+), it is now sufficient to show that $c\hat{u} = c\sum_{j=1}^q u^j$. Assume, by contradiction, that $c\hat{u} < c\sum_{j=1}^q u^j$, and consider the vector $\bar{u} = \hat{u} - \sum_{j=1}^q u^j$. Then $c\bar{u} < 0$ and $B\bar{u} = F(d^+ - \sum_{j=1}^q d^{+j}) = 0$. Moreover, the inequalities $u^j \leq \hat{u}$, $j = 1, \ldots, q$, imply that, componentwise, $\bar{u}_k < 0$ only if $\hat{u}_k > 0$. Hence, there exists a sufficiently small $\alpha > 0$ such that $\hat{u} + \alpha\bar{u} \geq 0$, $c(\hat{u} + \alpha\bar{u}) < c\hat{u}$, and $B(\hat{u} + \alpha\bar{u}) = Fd^+$, in contradiction with the hypothesis that $\hat{u}$, being a worst case hyperflow, is an optimal solution of problem LBP(d^+).

Sufficiency. The sufficiency follows immediately from Lemma 3.2. ∎

As a consequence of the above theorem, we may state the following complexity result.

COROLLARY 3.5. (i) *The problem of verifying whether an ICP instance (B, F, c, d^+) is WCD can be solved in polynomial time.*

(ii) *The ICP can be solved in polynomial time on the set of the WCD instances.*

Proof. Problems LBP(d^+) and UBP(d^+) are linear programming problems with a polynomial number of variables and constraints; thus they can be solved in polynomial time. Then, statements (i) and (ii) follow from Theorem 3.4. ∎

To conclude this section, we analyze the quality of the upper bound $z_{UBP}(d^+)$ and of the lower bound $z_{LBP}(d^+)$ in the general case.

COROLLARY 3.6. *The optimal value of problem UBP(d^+) is at most q times the optimal value of problem ICP_{LP}.*

Proof. The thesis follows trivially from the inequalities $z_{UBP} \leq q \max_j z_{LBP}(d^{+j})$ and $\max_j z_{LBP}(d^{+j}) \leq z_{ICP}$. ∎

On the other side, as shown in the following example, the worst case ratio $\frac{z_{ICP}}{z_{LBP}(d^+)}$ can be arbitrarily large.

Example 3.7. Consider again the WCD instance in Example 3.1. If the costs c_{15} and c_{36} change from 3 to ε with $0 < \varepsilon < 3$, the instance is not WCD anymore. Indeed, in this case, the only optimal solution of LBP(d^+) is $\bar{u}_{15} = \bar{u}_{36} = 1$, $\bar{u}_{ij} = 0$

otherwise, with $z_{LBP}(d^+) = 2\varepsilon$. On the other hand, the optimal solution of the ICP is still the vector $\hat{u}$ in (3.3) and thus $z_{ICP} = 6$.

4. SWCD classes of instances. In this section, we consider the classes of ICP instances that are WCD independently of the choice of the cost vector c and the demand vector d^+. In particular, we determine some necessary and sufficient conditions that characterize this property.

First note that, as Examples 3.1 and 3.7 show, an ICP instance may or may not be WCD depending on the cost vector c. The next lemma points out that, under the assumption $d^+ > 0$, the WCD property does not depend on the demand d^+.

LEMMA 4.1. *If the ICP instance (B, F, c, d^+), $d^+ > 0$, is WCD, then each instance $(B, F, c, \hat{d})$, $\hat{d} \geq 0$, is WCD.*

Proof. By Theorem 3.4 there exists a common optimal solution u^+ of problems LBP(d^+) and UBP(d^+). Then, by linearity, αu^+, $\alpha > 0$, is a common optimal solution of problems LBP(αd^+) and UBP(αd^+). As a consequence, any instance $(B, F, c, \alpha d^+)$, $\alpha > 0$, is WCD and this observation allows us to assume without loss of generality that $\hat{d} \lneq d^+$. Assume that the instance $(B, F, c, \hat{d})$ is not WCD for some $0 \leq \hat{d} \lneq d^+$ and define $\bar{d} = d^+ - \hat{d} \gneq 0$. Since $(B, F, c, \hat{d})$ is not WCD, the inequalities $z_{LBP}(\bar{d}) \leq z_{UBP}(\bar{d})$ and $z_{LBP}(\hat{d}) < z_{UBP}(\hat{d})$ hold. Moreover, since the sum of any two optimal solutions of problems $LBP(\bar{d})$ and $LBP(\hat{d})$ is feasible for LBP(d^+), we have $z_{LBP}(d^+) \leq z_{LBP}(\bar{d}) + z_{LBP}(\hat{d})$ and, consequently, $z_{LBP}(d^+) < z_{UBP}(\bar{d}) + z_{UBP}(\hat{d})$. On the other hand, for each $d \geq 0$, $z_{UBP}(d) = \sum_{j=1}^{q} z_{LBP}(d^j)$ and, by linearity, $z_{LBP}(d^j) = d_j z_{LBP}(e^j)$ for each $j = 1, \ldots, q$. As a consequence, $z_{UBP}(d^+) = \sum_{j=1}^{q} d_j^+ z_{LBP}(e^j) = \sum_{j=1}^{q} (\bar{d}_j + \hat{d}_j) z_{LBP}(e^j) = \sum_{j=1}^{q} z_{LBP}(\bar{d}^j) + \sum_{j=1}^{q} z_{LBP}(\hat{d}^j) = z_{UBP}(\bar{d}) + z_{UBP}(\hat{d})$. Therefore $z_{LBP}(d^+) < z_{UBP}(d^+)$, in contradiction with the fact that (B, F, c, d^+) is a WCD instance. ∎

Let (B, F) denote the class of all the feasible ICP instances defined by the same pair of matrices B and F and vectors c and d^+ with strictly positive components. By Lemma 2.1, (B, F) either is empty or contains any instance (B, F, c, d^+).

DEFINITION 4.2. *The class (B, F) is a* structurally worst case demand (SWCD) *class if all the instances $(B, F, c, d^+) \in (B, F)$ are WCD.*

As an immediate consequence of Lemma 4.1, a nonempty class (B, F) is SWCD if and only if any instance (B, F, c, e) is WCD.

The next theorem states some necessary and sufficient conditions that characterize the pairs (B, F) of matrices that define SWCD classes.

THEOREM 4.3. *Given a nonempty class (B, F), the following statements are equivalent:*

(i) *(B, F) is SWCD.*

(ii) *Each extreme ray $(\bar{u}, \bar{d})$ of the cone $C = \{(u, d) \in \mathbb{R}^{m+q} : Bu = Fd,\ u \geq 0,\ d \geq 0\}$ is such that $\bar{d}_j > 0$ for at most one j, $j = 1, \ldots, q$.*

(iii) *Each vertex of the polyhedron $P = \{u \in \mathbb{R}^m : Bu = Fe,\ u \geq 0\}$ is determined by a basis submatrix $\mathcal{B}$ of B such that $\mathcal{B}^{-1} F e^j \geq 0$ for each $j = 1, \ldots, q$.*

Proof. The statement holds trivially if $q = 1$, so we can assume $q > 1$. For $d \geq 0$, denote by $\gamma(d)$ the number of nonnull components of d.

(i) ⇒ (ii) We show that, if there exists an extreme ray $(\bar{u}, \bar{d})$ of C with $\gamma(\bar{d}) > 1$, then there exists an instance in (B, F) that is not WCD. Let $(\bar{u}, \bar{d})$ be an extreme ray of C with $\gamma(\bar{d}) > 1$. Then $\bar{u}$ is a vertex of the polyhedron $P(\bar{d}) = \{u \in \mathbb{R}^m : Bu = F\bar{d},\ u \geq 0\}$ and for each vertex $\hat{u}$ of $P(\bar{d})$, $\hat{u} \neq \bar{u}$, we have $\hat{u}_k > 0$ for at least one k such that $\bar{u}_k = 0$. In addition, $r \gneq 0$ for each possible extreme ray r of $P(\bar{d})$. So we can

define a cost vector $c > 0$ such that $c\hat{u} > c\bar{u}$ for each vertex $\hat{u} \neq \bar{u}$ and $cr > 0$ for each extreme ray r by setting $c_k = \varepsilon > 0$ if $\bar{u}_k > 0$, $c_k = 1$ otherwise, and choosing ε sufficiently small. Consider now the ICP instance $(B, F, c, \bar{d})$. By the structure of problem UBP($\bar{d}$), if $\hat{u}$ is an optimal solution of UBP($\bar{d}$) the vector $(\hat{u}, \bar{d})$ is the sum of $\gamma(\bar{d}) > 1$ elements of C. This implies that $(\hat{u}, \bar{d})$ is not an extreme ray of C and $\hat{u} \neq \bar{u}$. Since $\bar{u}$ is the unique optimal solution of problem LBP($\bar{d}$) it follows that $c\bar{u} < c\hat{u}$ and the instance $(B, F, c, \bar{d})$ is not WCD by Theorem 3.4.

(ii) $\Rightarrow$ (iii) We show that the thesis holds for each basis submatrix $\mathcal{B} = B^\beta$ such that $\bar{u}_\beta = \mathcal{B}^{-1}Fe \geq 0$. Given a basis submatrix $\mathcal{B} = B^\beta$, define $\bar{u} = [\bar{u}_\beta, \bar{u}_{\{1,\dots,m\}\setminus\beta}] = [\mathcal{B}^{-1}Fe, 0]$ and $\bar{u}^j = [\bar{u}^j_\beta, \bar{u}^j_{\{1,\dots,m\}\setminus\beta}] = [\mathcal{B}^{-1}Fe^j, 0]$, $j = 1, \dots, q$. We prove the statement by showing that $(\bar{u}, e) \in C$ implies $(\bar{u}^j, e^j) \in C$ for each $j = 1, \dots, q$. Since $\bar{u} \geq 0$, $(\bar{u}, e)$ belongs to C. Moreover, since $\gamma(e) = q > 1$, $(\bar{u}, e)$ is not an extreme ray of C and we can write $(\bar{u}, e) = \sum_{j=1}^q (\hat{u}^j, e^j)$, where each $(\hat{u}^j, e^j)$ is an extreme ray of C. It remains to prove that $\hat{u}^j = \bar{u}^j$, $j = 1, \dots, q$. To this aim, observe that, since $\hat{u}^j \geq 0$ and $\bar{u}_{\{1,\dots,m\}\setminus\beta} = 0$, we have $\hat{u}^j_{\{1,\dots,m\}\setminus\beta} = \bar{u}^j_{\{1,\dots,m\}\setminus\beta} = 0$. In this case, $B\hat{u}^j = \mathcal{B}\hat{u}^j_\beta = Fe^j$ and finally $\hat{u}^j_\beta = \mathcal{B}^{-1}Fe^j = \bar{u}^j_\beta$ for each j.

(iii) $\Rightarrow$ (i) For any $c > 0$, let $\mathcal{B}$ be an optimal basis submatrix of problem LBP(e) as in statement (iii). Then, by hypothesis, $\bar{u} = [\bar{u}_\beta, \bar{u}_{\{1,\dots,m\}\setminus\beta}] = [\mathcal{B}^{-1}Fe, 0] \geq 0$ implies $\bar{u}^j = [\bar{u}^j_\beta, \bar{u}^j_{\{1,\dots,m\}\setminus\beta}] = [\mathcal{B}^{-1}Fe^j, 0] \geq 0$ for $j = 1, \dots, q$, and each $\bar{u}^j$ is a feasible solution of the corresponding problem LBP(e^j). Now, by linearity arguments, $z_{UBP}(e) \leq c\sum_{j=1}^q \bar{u}^j = c\sum_{j=1}^q \mathcal{B}^{-1}Fe^j = c\mathcal{B}^{-1}Fe = z_{LBP}(e)$, which implies $z_{UBP}(e) = z_{LBP}(e)$. Then the instance (B, F, c, e) is WCD by Theorem 3.4. Since the previous argument holds for any $c > 0$, the class (B, F) is SWCD. ∎

From a geometric perspective, Theorem 4.3 in point (iii) states that the class (B, F) is SWCD if and only if for each basis submatrix $\mathcal{B}$ of B the condition $Fe \in cone\{\mathcal{B}\}$ implies $cone\{F\} \subseteq cone\{\mathcal{B}\}$. Moreover, as shown in the proof of part (ii) $\Rightarrow$ (iii), the following property holds.

COROLLARY 4.4. *Let (B, F) be a nonempty SWCD class. Then for every instance $(B, F, c, d^+) \in (B, F)$ any feasible basic solution of problem $LBP(d^+)$ is feasible for the ICP. As a consequence, any instance in (B, F) admits an optimal solution that is a parallelotope.*

For a generic WCD instance, the optimal solution of the ICP obtained by solving problem $UBP(d^+)$ is not in general a parallelotope. Instead, by the above result, if the instance belongs to an SWCD class, an optimal parallelotope solution of the ICP can always be found by solving problem $LBP(d^+)$.

Other consequences of Theorem 4.3 concern the properties of the ICP instances defined by a nonsingular square matrix B. The next corollary strengthens the result in Theorem 2.3; the following one considers the case of more general objective functions.

COROLLARY 4.5. *Let B be a matrix whose columns are linearly independent. Then any nonempty class (B, F) is SWCD.*

Proof. By Lemma 2.2, we can assume that B is a nonsingular square matrix, so B contains itself as a unique basis submatrix. As each problem $LBP(e^j)$, $j = 1, \dots, q$, is feasible by Lemma 2.1, this implies $B^{-1}Fe^j \geq 0$ for each j and the thesis follows by Theorem 4.3 (iii). ∎

COROLLARY 4.6. *Given a nonempty class (B, F), where B is a nonsingular square matrix, let $g(u)$, $g : \mathbb{R}^m \to \mathbb{R}$, be a function not decreasing in any component*

of u. Then $u^+ = B^{-1}Fd^+$ is an optimal solution of the problem

$$\min\{g(u^+) : \ FD(d^+) \subseteq BU(u^+), \ u^+ \geq 0\}. \tag{4.1}$$

Proof. The vector $\bar{u} = B^{-1}Fd^+$ is the unique feasible solution of $LBP(d^+)$ and thus $\bar{u} \leq u^+$ for each feasible solution u^+ of problem (4.1). On the other hand, since $\bar{u}$ is a basic solution, $\bar{u}$ is feasible for (4.1) by Corollary 4.4. Then the thesis follows by the monotonicity of the function g. ∎

A final consequence of Theorem 4.3 (iii) and Lemma 4.1 is that if (B, F) is an SWCD class, then any nonempty class $(\hat{B}, \hat{F})$, where $\hat{B}$ and $\hat{F}$ are obtained from B and F, respectively, by deleting some of their columns, is SWCD as well.

5. Two families of SWCD classes. In this section, we introduce two particular families of SWCD classes. First, we characterize the classes of instances of the minimum cost network containment problem (MCNCP) that enjoy the SWCD property. As a corollary, we prove that the problem of verifying whether a given MCNCP (ICP) class is SWCD is coNP-complete. Next, we prove that pre-Leontief matrices B and nonnegative matrices F define SWCD classes. In both the cases the SWCD property implies the integrality of the optimal basic solutions of the ICP problem.

5.1. The SWCD classes of MCNCP instances. As mentioned in the introduction, the MCNCP instances are particular ICP instances defined by matrices B and F that are the node-arc incidence matrices of two directed graphs with the same set of nodes.

In the following, given a directed graph $G = (V, E)$, a *directed path* in G is a sequence $i_1 a_1 i_2 \dots a_{k-1} i_k$ of nodes and arcs without any repetition of nodes and such that $a_r = (i_r, i_{r+1}) \in E$ for each $1 \leq r \leq k$. A *directed cycle* in G is a directed path together with the arc (i_k, i_1).

DEFINITION 5.1. *Given a class (B, F) of MCNCP instances, let G be the directed graph with incidence matrix $[B| - F]$. We say that the class is* $2F$-cycle free *if each directed cycle of G includes at most one F-arc, i.e., at most one arc corresponding to a column of $-F$.*

THEOREM 5.2. *A class (B, F) of MCNCP instances is SWCD if and only if it is* $2F$-cycle free. *In this case, every instance in (B, F) defined by an integral vector d^+ admits an integral optimal solution.*

Proof. In the case of MCNCP instances, any element of the cone $C = \{(u, d) \in \mathbb{R}^{m+q} : Bu = Fd, \ u \geq 0, \ d \geq 0\}$ corresponds to a circulation in G and, by the flow decomposition theorem [1], it is the sum of cycle flows, that is, flows of the type αx, where x is the incidence vector of a directed cycle of G and $\alpha \in \mathbb{R}_+$. As a consequence, (u, d) is an extreme ray of C if and only if it is a cycle flow. Then, by Theorem 4.3 (ii), (B, F) is SWCD if and only if any cycle in G contains at most one F-arc, that is, G is $2F$-cycle free. The second statement follows by the integrality property of the optimal basic solutions of the min cost flow problem LBP(d^+). ∎

We remark that, as shown in [3], in general an MCNCP instance with integral data may not admit an integral optimal solution.

As a consequence of the above theorem we obtain the following complexity result.

THEOREM 5.3. *Determining whether a given class (B, F) of MCNCP instances is SWCD is a coNP-complete problem.*

Proof. We prove the statement by showing that the problem of deciding whether a class (B, F) of MCNCP instances is not SWCD is an NP-complete problem. We first show that this problem is in NP. By definition, a class (B, F) is not SWCD if

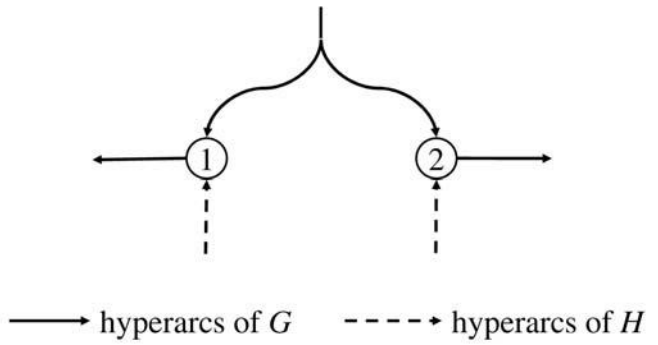


FIG. 5.1. *Hypergraphs G and H in Example* 5.5.

it includes an instance that is not WCD. Then, any non-WCD instance in (B, F) is a certificate that (B, F) is not SWCD which can be checked in polynomial time by Corollary 3.5. In order to prove that the problem is NP-complete, let us consider a polynomial transformation from the 2-disjoint paths problem for directed graphs (2DPP) defined as follows. Given a directed graph $H = (V, E)$ and two assigned pairs of nodes (s_1, t_1) and (s_2, t_2), find if there exist two node-disjoint paths connecting s_1 to t_1 and s_2 to t_2. As shown in [6], 2DPP is NP-complete. Given a 2DPP instance, consider the MCNCP class of instances (B, F) where B is the incidence matrix of the graph H and F is the incidence matrix of the graph having node set V and arc set $\{(s_1, t_2), (s_2, t_1)\}$. By Theorem 5.2, (B, F) is not SWCD if and only if the directed graph G having $[B| - F]$ as incidence matrix contains a directed cycle including both the arcs (t_2, s_1) and (t_1, s_2). It is immediate to verify that this condition is equivalent to the existence of two node-disjoint paths in H connecting s_1 to t_1 and s_2 to t_2. Since the above transformation is polynomial, the thesis follows. □

Since the MCNCP instances are particular ICP instances, the above result trivially generalizes to the ICP problem.

COROLLARY 5.4. *Determining whether a given class (B, F) of ICP instances is SWCD is a coNP-complete problem.*

5.2. The classes of pre-Leontief instances. Let us now consider the case where F is a nonnegative matrix. As the following simple example shows, this condition does not in general imply the SWCD property.

Example 5.5. Consider any instance (B, F, c, d^+) where B is the incidence matrix of the hypergraph G with node set $V = \{1, 2\}$ and hyperarcs $(\emptyset, \{1, 2\}), (1, \emptyset), (2, \emptyset)$, $F = I$ is the incidence matrix of the hypergraph H having hyperarcs $(\emptyset, 1)$ and $(\emptyset, 2)$, $d^+ = (1, 1)$, and $c > 0$ is any cost vector (see Figure 5.1). The solution $(1, 0, 0)$ is the unique feasible solution of problem LBP(d^+). Nevertheless, there exists no u, $0 \leq u \leq (1, 0, 0)$, such that either $Bu = (1, 0)$ or $Bu = (0, 1)$. Indeed, for each cost vector $c > 0$, the optimal solution of the ICP is given by $\hat{u} = (1, 1, 1)$.

Notwithstanding the previous example, Theorem 4.3 implies the following result.

COROLLARY 5.6. *If F is a nonnegative matrix and each vertex of the polyhedron $P = \{u \in \mathbb{R}^m : Bu = Fe, u \geq 0\}$ is determined by a basis submatrix $\mathcal{B}$ that is inverse nonnegative, i.e., $\mathcal{B}^{-1} \geq 0$, then the class (B, F) is SWCD.*

Proof. The thesis follows by Theorem 4.3 (iii), since the vectors $\mathcal{B}^{-1}Fe^j$, $j =$

$1, \ldots, q$, as products of nonnegative matrices, are nonnegative. □

From a geometric point of view, the conditions stated in Corollary 5.6 simply guarantee that the columns of F are contained in $\mathbb{R}^n_+$ and $\mathbb{R}^n_+ \subseteq cone\{\mathcal{B}\}$ for each basis submatrix $\mathcal{B}$ of B that is feasible for problem LBP(e). So $cone\{F\} \subseteq \mathbb{R}^n_+ \subseteq cone\{\mathcal{B}\}$.

Due to the previous corollary, we are interested in matrices B whose basis submatrices are inverse nonnegative. A fundamental class of matrices that enjoy this property is defined by the pre-Leontief matrices [17].

DEFINITION 5.7. *A matrix B is said to be* pre-Leontief *if each column of B contains at most one positive entry (which can be assumed without loss of generality to be equal to 1). Moreover, B is said to be* Leontief *if each column has exactly one positive element and there exists $\bar{u} \geq 0$ such that $B\bar{u} > 0$.*

Pre-Leontief matrices have been extensively studied for their interest in several fields, such as, for instance, operations management, polyhedral combinatorics, logic, and expert systems. An overview on their applications can be found, e.g., in [11, 12]. Any pre-Leontief matrix B can be seen as the incidence matrix of a directed hypergraph whose hyperarcs have at most one head [11]. So this class includes, in particular, the incidence matrices of directed graphs with generalized flows [1] and hyperarborescences.

The algebraic properties of the linear programming problems defined by a pre-Leontief constraint matrix and their algorithmic consequences were thoroughly investigated in [5, 11, 12, 17].

DEFINITION 5.8. *A problem of the form $\{\min cu : Bu = b,\ u \geq 0\}$ where B is a pre-Leontief (Leontief) matrix is called a* pre-Leontief (Leontief) flow problem. *If, in addition, $b \geq 0$, then the problem is called a* pre-Leontief (Leontief) substitution flow problem.

The following lemma reports some results concerning the structure of the vertices of a pre-Leontief substitution flow problem obtained in [17] (and appearing also in [11]).

LEMMA 5.9. (i) *If A is a pre-Leontief matrix, then after permuting the rows and columns appropriately, A can be written as*

$$\begin{pmatrix} A_1 & A_2 \\ 0 & A_3 \end{pmatrix}, \tag{5.1}$$

where A_1 is Leontief and $A_3x = 0$ for each $x \geq 0$ such that $A_3x \geq 0$.

(ii) *If x is a vertex of a pre-Leontief flow problem defined by the constraint matrix A written as in* (5.1) *and right-hand side $b = (b_1, b_2) \geq 0$, then $x = (x_1, x_2)$, where x_1 is a vertex of $\{x_1 \geq 0 : A_1x = b_1\}$ and $x_2 = 0$ (so $b_2 = 0$).*

(iii) *Any vertex x of a Leontief substitution flow problem is determined by a basis submatrix $\mathcal{B}$ that is Leontief.*

(iv) *Any nonsingular square Leontief matrix has a nonnegative inverse.*

(v) *If a pre-Leontief substitution flow problem with integral constraint matrix and integral right-hand side vector has an optimal solution, then it has an integral optimal solution.*

The above properties of pre-Leontief substitution flow problems carry to the following results.

THEOREM 5.10. *Let (B, F) be a class of ICP instances where B is a pre-Leontief matrix and F is a nonnegative matrix. Then (B, F) is an SWCD class. Moreover, every instance in (B, F) defined by integral B, F, and d^+ admits an integral optimal solution.*

Proof. Since B is a pre-Leontief matrix and $F \geq 0$, problems $LBP(e)$ and $LBP(e^j)$, $j = 1, \ldots, q$, are pre-Leontief substitution flow problems. If B is Leontief, then the thesis follows by Lemma 5.9 (iii)–(iv) and Corollary 5.6. In the opposite case, let B be partitioned as in (5.1) and denote by F_1 and F_2 the matrices formed by the rows of F corresponding to the rows of A_1 and A_3, respectively. Since $F \geq 0$ and $e > 0$, if problem LBP(e) is feasible, then $F_2 = 0$ by Lemma 5.9 (ii). In this case, by Lemma 5.9 (ii) and Theorem 4.3 (iii), the class (B, F) is SWCD if and only if the class (A_1, F_1) is SWCD and, A_1 being a Leontief matrix, the thesis follows by the argument above. The second statement is an immediate consequence of Lemma 5.9 (v). ∎

6. Extensions to more general polyhedra. In this section, we discuss which results of sections 3 and 4 can be generalized to the case where the set $U(u^+)$ introduced in (1.1) has the more general form

$$U_A(u^+) = \{u \in \mathbb{R}^m : u \geq 0,\ Au \leq u^+\}, \tag{6.1}$$

where A is a given nonnegative matrix with at least a nonzero entry in each column. Note that, when A is the identity matrix, $U_A(u^+) = U(u^+)$. This generalization was already considered in [15] for the MCNCP. In this context, a set as in (6.1) defines constraints, called *bundle constraints*, which impose upper bounds on the weighted sum of flows on assigned subsets of arcs, and the corresponding upper bounds are called *generalized capacities*.

Denote by (B, F, c, d^+, A) the generic instance of the ICP where a set $U_A(u^+)$ is considered. We redefine the ICP$_{LP}$ problem as

$$\begin{aligned} z^A_{ICP} &= \min cu^+, \\ Bu^r &= Fd^r \quad \forall r : d^r \in \mathrm{Ext}\{D\}, \\ Au^r &\leq u^+ \quad \forall r : d^r \in \mathrm{Ext}\{D\}, \\ u^r &\geq 0 \quad \forall r : d^r \in \mathrm{Ext}\{D\}. \end{aligned} \tag{6.2}$$

Moreover, we say that (B, F, c, d^+, A) is a WCD instance with respect to the matrix A (WCD(A)) if there exists a demand $\bar{d}$ such that the LPB(d) problem

$$\begin{aligned} z^A_{LBP}(d) &= \min c\tilde{u}, \\ Bu &= Fd, \\ Au &\leq \tilde{u}, \\ u &\geq 0 \end{aligned} \tag{6.3}$$

when $d = \bar{d}$ has the same optimal value of (6.2) and at least one common optimal solution $u^+ = \tilde{u}$.

Unfortunately, as the following example shows, one cannot guarantee that, when an instance is WCD(A), d^+ is a worst case demand.

Example 6.1. Consider the ICP instance (B, F, c, e, A), where B and F are, respectively, the incidence matrices of the networks $G = (V, E_G)$ and $H = (V, E_H)$ with node set $V = \{1, 2, 3, 4\}$ and arc sets $E_G = \{(1, 2), (1, 4), (3, 2), (3, 4)\}$ and $E_H = \{(1, 2), (3, 4)\}$. Let A be the matrix

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix},$$

so that $u^+ \in \mathbb{R}^4$, and consider the cost vector $c = (1,1,1,1)$. First note that the instance is WCD(A) and $\bar{d} = (1,0)$ is a worst case demand. Indeed, the optimal solution of problem (6.3) when $d = \bar{d}$ is $u^* = (1,0,0,0), \tilde{u}^* = (1,1,1,1)$ of cost $z^A_{LBP}(1,0) = 4$. Moreover, $u^+ = \tilde{u}^* = (1,1,1,1)$ is feasible for problem (6.2) and thus $z^A_{ICP} = z^A_{LBP}(1,0) = 4$. Now consider the maximal demand $d^+ = (1,1)$. When $d = d^+$, the optimal solution of problem (6.3) is $u = (0,1,1,0), \tilde{u} = (0,1,1,0)$ of cost $z^A_{LBP}(1,1) = 2 < z^A_{ICP}$, and thus d^+ is not a worst case demand.

Notwithstanding the previous negative result, let us redefine problem UBP(d) as

$$\begin{aligned} z^A_{UBP}(d) &= \min c\tilde{u}, \\ Bu^j &= Fd^j, \quad j = 1,\dots,q, \\ u &= \sum_{j=1}^{q} u^j, \\ Au &\le \tilde{u}, \\ u^j &\ge 0, \quad j = 1,\dots,q. \end{aligned} \tag{6.4}$$

When $d = d^+$, $z^A_{UBP}(d^+)$ is still an upper bound for z^A_{ICP}. As a consequence, if $z^A_{UBP}(d^+) = z^A_{LBP}(d^+)$, then d^+ is a worst case demand and the instance is WCD(A). This observation allows us to show the following result, which links SWCD classes of ICP instances (as defined in section 4) with classes of WCD(A) instances.

THEOREM 6.2. *If a class (B,F) is SWCD, then any instance (B,F,c,d^+,A) where A is a nonnegative matrix without null columns and $c > 0$ is WCD(A). Moreover, d^+ is a worst case demand.*

Proof. Given a nonnegative matrix A without null columns, an SWCD class (B,F), and a cost vector $c > 0$, let $(u^*_{LBP}, \tilde{u}^*_{LBP})$ and $((u^{j*}_{UBP})^q_{j=1}, u^*_{UBP}, \tilde{u}^*_{UBP})$ denote the optimal solutions of problems (6.3) and (6.4), respectively, for the instance (B,F,c,d^+,A). Then, as it is easy to verify, $\tilde{u}^*_{LBP} = Au^*_{LBP}$ and $\tilde{u}^*_{UBP} = Au^*_{UBP}$. This implies that $z^A_{LBP}(d^+)$ and $z^A_{UBP}(d^+)$ are, respectively, equal to the optimal values $z_{LBP}(d^+)$ and $z_{UBP}(d^+)$ of problems LBP(d^+) and UBP(d^+) for the instance (B,F,cA,d^+). Since $cA > 0$, the instance $(B,F,cA,d^+) \in (B,F)$ and thus is WCD. So, by Theorem 3.4, the equality $z^A_{LBP}(d^+) = z_{LBP}(d^+) = z_{UBP}(d^+) = z^A_{UBP}(d^+)$ holds and the instance (B,F,c,d^+,A) is WCD(A). ∎

The above theorem shows that an SWCD class of instances defines a correspondent class of WCD(A) instances. However, the converse does not hold as the following example shows.

Example 6.3. Consider the ICP instance (B,F,c,e,A), where B and F are, respectively, the incidence matrices of the networks $G = (V,E_G)$ and $H = (V,E_H)$ introduced in Example 6.1 and

$$A = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix}.$$

Let $c = (c_1,c_2,c_3,c_4) > 0$ be any nonnegative cost vector and $d^+ = e = (1,1)$. Since the optimal value of both problems (6.3) and (6.4) is $z^A_{LBP}(1,1) = z^A_{UBP}(1,1) = c_1 + c_4$ (with optimal solution $u = (1,0,0,1), \tilde{u} = (1,0,0,1)$), the maximal demand e is a WCD and the instance is WCD(A) for any $c > 0$. However, for $c = (2,1,1,2)$, the ICP instance (B,F,c,e) is not WCD since $z_{LBP}(1,1) = 2 < z_{ICP} = 4$.

Acknowledgment. We are grateful to an anonymous referee for useful suggestions concerning the geometrical aspects of this work.

REFERENCES

[1] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin, *Network Flows*, Prentice–Hall, Englewood Cliffs, NJ, 1993.
[2] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979.
[3] F. Blanchini, F. Rinaldi, and W. Ukovich, *A network design problem for a distribution system with uncertain demands*, SIAM J. Optim., 7 (1997), pp. 560–578.
[4] F. Blanchini, F. Rinaldi, and W. Ukovich, *Least inventory control of multi-storage systems with non-stochastic unknown input*, IEEE Trans. Robotics and Automation, 13 (1997), pp. 633–645.
[5] R. Cambini, G. Gallo, and M. G. Scutellà, *Flows on hypergraphs*, Math. Programming, 78 (1997), pp. 195–217.
[6] S. Fortune, J. Hopcroft, and J. Wyllie, *The directed graph homeomorphism problem*, Theoret. Comput. Sci., 10 (1980), pp. 111–121.
[7] P. Gritzmann and V. Klee, *On the complexity of some basic problems in computational convexity:* I. *Containment problems*, Discrete Math., 136 (1994), pp. 129–174.
[8] M. Grötschel, L. Lovász, and A. Schrijver, *Geometric Algorithms and Combinatorial Optimization*, Springer-Verlag, Berlin, 1988.
[9] B. Grünbaum, *Convex Polytopes*, Springer-Verlag, New York, 2003.
[10] L. J. Guibas, A. Nguyen, and L. Zhang, *Zonotopes as bounding volumes*, in Proceedings of the 14th Annual ACM–SIAM Symposium on Discrete Algorithms, Baltimore, MD, 2003, pp. 803–312.
[11] R. G. Jeroslow, K. Martin, R. L. Rardin, and J. Wang, *Gainfree Leontief substitution flow problems*, Math. Programming, 57 (1992), pp. 375–414.
[12] G. J. Koehler, A. B. Whinston, and G. P. Wright, *Optimization over Leontief Substitution Systems*, North–Holland, Amsterdam, 1975.
[13] S. T. McCormick, *Submodular containment is hard, even for networks*, Oper. Res. Lett., 19 (1996), pp. 95–99.
[14] P. McMullen, *On zonotopes*, Trans. Amer. Math. Soc., 159 (1971), pp. 91–109.
[15] R. Pesenti, F. Rinaldi, and W. Ukovich, *An exact algorithm for the solution of a network design problem*, Networks, 43 (2004), pp. 87–102.
[16] M. Svetoslav and C. Dalcidio, *On the approximation of the centered zonotopes in the plane*, in Proceedings of the 4th International Conference on Large Scale Scientific Computing 2003, Lecture Notes in Comput. Sci. 2907, Springer-Verlag, Berlin, 2004, pp. 246–253.
[17] A. F. Veinott, *Extreme points of Leontief substitution systems*, Linear Algebra Appl., 1 (1968), pp. 181–194.
[18] G. M. Ziegler, *Lectures on Polytopes*, Springer-Verlag, New York, 1995.

SIAM J. OPTIM.
Vol. 17, No. 4, pp. 1205–1223
© 2007 Society for Industrial and Applied Mathematics

THE ŁOJASIEWICZ INEQUALITY FOR NONSMOOTH SUBANALYTIC FUNCTIONS WITH APPLICATIONS TO SUBGRADIENT DYNAMICAL SYSTEMS*

JÉRÔME BOLTE†, ARIS DANIILIDIS‡, AND ADRIAN LEWIS§

Abstract. Given a real-analytic function $f : \mathbb{R}^n \to \mathbb{R}$ and a critical point $a \in \mathbb{R}^n$, the Łojasiewicz inequality asserts that there exists $\theta \in [\frac{1}{2}, 1)$ such that the function $|f - f(a)|^\theta \, \|\nabla f\|^{-1}$ remains bounded around a. In this paper, we extend the above result to a wide class of nonsmooth functions (that possibly admit the value $+\infty$), by establishing an analogous inequality in which the derivative $\nabla f(x)$ can be replaced by any element x^* of the subdifferential $\partial f(x)$ of f. Like its smooth version, this result provides new insights into the convergence aspects of subgradient-type dynamical systems. Provided that the function f is sufficiently regular (for instance, convex or lower-C^2), the bounded trajectories of the corresponding subgradient dynamical system can be shown to be of finite length. Explicit estimates of the rate of convergence are also derived.

Key words. Łojasiewicz inequality, subanalytic function, nonsmooth analysis, subdifferential, dynamical system, descent method

AMS subject classifications. Primary, 26D10; Secondary, 32B20, 49K24, 49J52, 37B35, 14P15

DOI. 10.1137/050644641

1. Introduction.

Let U be a nonempty open subset of $\mathbb{R}^n$ equipped with its canonical Euclidean norm $||\cdot||$, and let $f : U \to \mathbb{R}$ be a real-analytic function. According to the Łojasiewicz gradient inequality [16, 17, 18], if $a \in U$ is a *critical* point of f, that is, $\nabla f(a) = 0$, then there exists $\theta \in [0, 1)$ such that the function

$$\frac{|f - f(a)|^\theta}{\|\nabla f\|} \tag{1}$$

remains bounded around the point a. (Throughout this work we set $0^0 = 1$, and we interpret $\lambda/0$ as $+\infty$ if $\lambda > 0$ and 0 if $\lambda = 0$.)

Recently, Kurdyka [13, Theorem 1] has extended the above result to C^1 functions whose graphs belong to an o-minimal structure (see [8], for example), and thus in particular to "globally subanalytic" functions. On the other hand, (1) might fail for C^∞ functions with no "adequate" geometric structure. Such functions can either satisfy a weaker condition (i.e., $\theta = 1$) or present wild oscillations around their critical point, preventing any comparison between their value and the norm of their gradient.

*Received by the editors November 9, 2005; accepted for publication (in revised form) June 6, 2006; published electronically January 12, 2007. This research was supported by NSERC.
http://www.siam.org/journals/siopt/17-4/64464.html

†Equipe Combinatoire et Optimisation (UMR 7090), Case 189, Université Pierre et Marie Curie, 4 Place Jussieu, 75252 Paris Cedex 05, France (bolte@math.jussieu.fr; http://www.ecp6.jussieu.fr/pageperso/bolte/). This author's work was partially supported by the CECM (Simon Fraser University) and the C.M.M. (Universidad de Chile).

‡Departament de Matemàtiques, C1/320, Universitat Autònoma de Barcelona, E-08193 Bellaterra (Cerdanyola del Vallès), Spain (arisd@mat.uab.es; http://mat.uab.es/∼arisd). This author's research was partially supported by the CECM (Simon Fraser University), the C.M.M. (Universidad de Chile), and the MEC grant MTM2005–08572–C03-03 (Spain).

§School of Operations Research and Industrial Engineering, Cornell University, 234 Rhodes Hall, Ithaca, NY 14853 (aslewis@orie.cornell.edu; http://www.orie.cornell.edu/∼aslewis). This author's research was partially supported by National Science Foundation grant DMS-0504032.

The following one-dimensional examples illustrate failures of these two types (around the critical point $a = 0$):

$$f(x) = \begin{cases} \exp(-1/x^2) & \text{if } x \neq 0, \\ 0 & \text{if } x = 0, \end{cases} \quad \text{and} \quad g(x) = \begin{cases} \exp(-1/x^2)\sin(1/x) & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

The aim of this note is to establish a *nonsmooth* version of the Łojasiewicz inequality (1) for lower semicontinuous convex subanalytic functions (Theorem 3.3) and for continuous subanalytic functions (Theorem 3.1). A first and simple illustration is given by the example of the Euclidean norm function $h(x) = \|x\|$, which satisfies (1) for every $\theta \in [0, 1)$ around zero (which is a "generalized" critical point; see Definition 2.11) but is not differentiable at 0. Behavior of this type is hereby shown to hold true for a large class of nonsmooth functions, leading to the conclusion that the Łojasiewicz inequality is more linked to the underlying geometrical structure of f than to its smoothness.

Given an extended-real-valued subanalytic function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, our approach to generalizing property (1) relies on a one-sided notion of generalized gradients called subgradients. For both a mathematical and a historical account on this notion, as well as for classical results in nonsmooth analysis, one is referred to the monographs of Clarke et al. [7] and Rockafellar and Wets [20].

Subgradients are obtained according to a two-stage process. First the equality in the definition of the usual gradient is relaxed into an inequality (Definition 2.10(i)): this gives rise to the notion of Fréchet subgradients. Then, by a closure operation, the so-called limiting subdifferential ∂f can be defined (Definition 2.10(ii)). This notion constitutes the basis for the generalization of the Łojasiewicz inequality to nonsmooth functions. Let us also mention that in this formalism Fermat's rule reads as follows: if a is a local minimizer of f, then $\partial f(a) \ni 0$; conversely, if $a \in \mathbb{R}^n$ is such that $\partial f(a) \ni 0$, the point a is called a critical point.

Variational analysis and subdifferential calculus provide a framework for the modeling of unilateral constraints in mechanics and in partial differential equations [11, 6, 9]. Such a calculus is also central in optimization. In particular it provides variational tools to treat constrained and unconstrained minimization problems on an equal theoretical level. This stems from the simple fact that minimizing f over a closed set C amounts to minimizing $f + \delta_C$ over $\mathbb{R}^n$, where δ_C is the indicator function of C; that is

$$\delta_C(x) = \begin{cases} 0 & \text{if } x \in C, \\ +\infty & \text{otherwise.} \end{cases} \tag{2}$$

Those domains have as a common topic the behavior at infinity of dynamical systems governed by subdifferential operators; see [15] for an insight in optimization. An important motivation that drove us to transpose the Łojasiewicz result into a nonsmooth context is precisely its expected consequences in the asymptotic analysis of such subgradient-type dynamical systems. Those are modeled on the following type of differential inclusion:

$$\dot{x}(t) \in -\partial f(x(t)), \quad t \geq 0, \quad x(0) \in \mathbb{R}^n,$$

where for any $x \in \mathbb{R}^n$, $\partial f(x)$ denotes the set of limiting subgradients. The above differential inclusion generalizes the classical gradient dynamical system

$$\dot{x}(t) = -\nabla f(x(t)), \quad t \geq 0, \quad x(0) \in \mathbb{R}^n. \tag{3}$$

In his pioneering work on real-analytic functions [16, 17], Łojasiewicz provided the main ingredient—namely, (1)—that allows us to derive the convergence of all bounded trajectories of (3) to critical points. As can be seen from a counterexample due to Palis and De Melo [19, p. 14], the set of cluster points of a bounded trajectory generated by the gradient of a C^∞ function is, in general, far from being a singleton. Those famous results illustrate the importance of gradient vector fields of functions satisfying the Łojasiewicz inequality. An even more striking feature is that the trajectories converge "in direction" when approaching critical points. This fact had been conjectured by Thom (around 1972; see [22]) for real-analytic functions, and established by Kurdyka, Mostowski, and Parusiński in [14]. The subanalytic generalized Thom conjecture remains open even in the smooth case (see [13, Conjecture F]).

In section 4 we extend Łojasiewicz results to a nonsmooth setting (f is a subanalytic proximal retract), by showing that all bounded trajectories have a finite length (Theorem 4.5). We also provide estimates of the asymptotic convergence rate towards the critical points (Theorem 4.7).

For related results on this topic, see [1] ; for other applications to partial differential equations, see the works of Simon [21] and Haraux [12].

2. Preliminaries. The key ingredients for the nonsmooth extension of the Łojasiewicz inequality are subanalyticity of the function f and notions of generalized differentiation provided by variational analysis.

2.1. Subanalytic sets and stability properties. We recall the following definition.

DEFINITION 2.1 (subanalyticity). (i) *A subset A of $\mathbb{R}^n$ is called* semianalytic *if each point of $\mathbb{R}^n$ admits a neighborhood V for which $A \cap V$ assumes the following form:*

$$\bigcup_{i=1}^{p} \bigcap_{j=1}^{q} \{x \in V : f_{ij}(x) = 0, g_{ij}(x) > 0\},$$

where the functions $f_{ij}, g_{ij} : V \mapsto \mathbb{R}$ are real-analytic for all $1 \le i \le p, 1 \le j \le q$.

(ii) *The set A is called* subanalytic *if each point of $\mathbb{R}^n$ admits a neighborhood V such that*

$$A \cap V = \{x \in \mathbb{R}^n : (x, y) \in B\},$$

where B is a bounded semianalytic subset of $\mathbb{R}^n \times \mathbb{R}^m$ for some $m \ge 1$.

(iii) *Given $m, n \in \mathbb{N}^*$, a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ (respectively, a point-to-set operator $T : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$) is called* subanalytic *if its graph is a subanalytic subset of $\mathbb{R}^n \times \mathbb{R}$ (respectively, of $\mathbb{R}^n \times \mathbb{R}^m$).*

Recall that the graphs of f and T, denoted respectively by $\operatorname{Gr} f$ and $\operatorname{Gr} T$, are defined by

$$\operatorname{Gr} f := \{(x, \lambda) \in \mathbb{R}^n \times \mathbb{R} : f(x) = \lambda\}, \qquad \operatorname{Gr} T := \{(x, y) \in \mathbb{R}^n \times \mathbb{R}^m : y \in T(x)\}.$$

Some of the elementary properties of subanalytic sets have been gathered below (see, e.g., [4, 10, 18]):

- Subanalytic sets are closed under locally finite union and intersection. The complement of a subanalytic set is subanalytic (Gabrielov theorem).
- If A is subanalytic, then so are its closure $\operatorname{cl} A$, its interior $\operatorname{int} A$, and its boundary $\operatorname{bd} A$.

- Given a subanalytic set S, the distance $d_S(x) := \inf\{\|x-a\| : a \in S\}$ is a subanalytic function.
- *Path connectedness* (see, e.g., [10, Facts 1.10–1.12]): Any subanalytic set has a locally finite number of connected components. Each component is subanalytic and subanalytically path connected; that is, every two points can be joined by a continuous subanalytic path that lies entirely in the set.
- *Curve selection lemma* (see, e.g., [4, Lemma 6.3]): If A is a subanalytic subset of $\mathbb{R}^n$ and $a \in \operatorname{bd} A$, then there exists an analytic path $z : (-1,1) \to \mathbb{R}^n$ satisfying $z(0) = a$ and $z((0,1)) \subset A$.

The image and the preimage of a subanalytic set are not in general subanalytic sets. This is essentially due to the fact that the image of an unbounded subanalytic set by a linear projection may fail to be subanalytic. Consider, for instance, the set $\{(\frac{1}{n+1}, n) : n \in \mathbb{N}\}$, whose projection onto $\mathbb{R} \times \{0\}$ is not subanalytic at 0.

To remedy to this lack of stability, let us introduce a stronger analytic-like notion called global subanalyticity (see [10] and references therein).

For each $n \in \mathbb{N}$, set $C_n = (-1,1)^n$ and define τ_n by

$$\tau_n(x_1, \dots, x_n) = \left(\frac{x_1}{1+x_1^2}, \dots, \frac{x_n}{1+x_n^2} \right) \in C_n.$$

DEFINITION 2.2 (global subanalyticity; see, e.g., [10, p. 506]). (i) *A subset S of $\mathbb{R}^n$ is called* globally subanalytic *if its image under τ_n is a subanalytic subset of $\mathbb{R}^n$.*

(ii) *An extended-real-valued function (respectively, a multivalued mapping) is called globally subanalytic if its graph is globally subanalytic.*

Globally subanalytic sets are subanalytic, and conversely any *bounded* subanalytic set is globally subanalytic. Typical examples of subanalytic sets which are not globally subanalytic are the set of integers $\mathbb{Z}$, the graph of the sinus function, the spiral $\{(t\cos t, t\sin t) \in \mathbb{R}^2 : t \geq 0\}$, etc. The class of semialgebraic sets (e.g., [3, 8]) provides an important subclass of globally subanalytic sets. Recall that a set $A \subset \mathbb{R}^n$ is called *semialgebraic* if it assumes the following form:

$$A = \bigcup_{i=1}^{p} \bigcap_{j=1}^{q} \{x \in V : f_{ij}(x) = 0, g_{ij}(x) > 0\},$$

where $f_{ij}, g_{ij} : \mathbb{R}^n \mapsto \mathbb{R}$ are polynomial functions for all $1 \leq i \leq p, 1 \leq j \leq q$. (Readers who are unfamiliar with subanalytic geometry might in a first reading replace "subanalytic" and "globally subanalytic" by "semialgebraic" in the statements that follow.)

A major fact concerning the class of globally subanalytic sets is its stability under linear projections.

THEOREM 2.3 (projection theorem; see, e.g., [10, Example 4, p. 505]). *Let $\Pi(x_1, \dots, x_{n+1}) = (x_1, \dots, x_n)$ be the canonical projection from $\mathbb{R}^{n+1}$ onto $\mathbb{R}^n$. If S is a globally subanalytic subset of $\mathbb{R}^{n+1}$, then so is $\Pi(S)$ in $\mathbb{R}^n$.*

Among the numerous consequences of the above result in terms of stability, the following properties are crucial to our main results:

- The image or the preimage of a globally subanalytic set by a globally subanalytic function (respectively, globally subanalytic multivalued operator) is globally subanalytic (see, e.g., [10, p. 504]).
- *Monotonicity lemma* (e.g., [10, Fact 4.1]): Take $\alpha < \beta$ in $\mathbb{R}$. If $\varphi : (\alpha, \beta) \to \mathbb{R}$ is a globally subanalytic function, then there is a partition $t_0 := \alpha < t_1 <$

$\cdots < t_{l+1} := \beta$ of (α, β) such that $\varphi|_{(t_i, t_{i+1})}$ is C^∞ and either constant or strictly monotone, for $i \in \{0, \dots, l\}$. Moreover ([13], e.g.), φ admits a *Puiseux development* at $t = \alpha$; that is, there exists $\delta > 0$, a positive integer k, $l \in \mathbb{Z}$, and $\{a_n\}_{n \geq l} \subset \mathbb{R}$ such that

$$\varphi(t) = \sum_{n \geq l} a_n (t - \alpha)^{n/k} \quad \text{for all } t \in (\alpha, \alpha + \delta).$$

– *Łojasiewicz factorization lemma* (e.g., [4, Theorem 6.4]): Let $K \subset \mathbb{R}^n$ be a compact set and $f, g : K \to \mathbb{R}$ two continuous (globally) subanalytic functions. If $f^{-1}(0) \subset g^{-1}(0)$, then there exist $c > 0$ and a positive integer r such that $|g(x)|^r \leq c\,|f(x)|$ for all $x \in K$.

2.2. Notions from nonsmooth analysis and further stability results. Throughout this paper, we essentially deal with nondifferentiable functions defined on $\mathbb{R}^n$ with values in $\mathbb{R} \cup \{+\infty\}$. We denote by $\operatorname{dom} f$ the domain of the function, that is, the subset of $\mathbb{R}^n$ on which f is finite. In a similar way, the domain of a point-to-set operator $T : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$, denoted by $\operatorname{dom} T$, is defined as the subset of $\mathbb{R}^n$ on which T is nonempty. The epigraph and the strict epigraph of f are respectively defined by

$$\operatorname{epi} f := \{(x, \lambda) \in \mathbb{R}^n \times \mathbb{R} : \lambda \geq f(x)\}, \qquad \operatorname{epi}_s f := \{(x, \lambda) \in \mathbb{R}^n \times \mathbb{R} : \lambda > f(x)\},$$

while the *epigraphical sum* of two extended-real-valued functions $f, g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is the function defined by

$$\mathbb{R}^n \ni u \longmapsto h(u) \;=\; \inf\{f(v) + g(v - u) : v \in \mathbb{R}^n\} \in [-\infty, +\infty].$$

The terminology stems from the fact that the strict epigraph of h is the Minkowski sum of the strict epigraphs of f and g.

Even if $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is subanalytic, its domain and its epigraph may fail to be subanalytic sets.

Example 2.4. Consider the function $f : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ whose graph is given by the set $S := \{(\frac{1}{n}, n)\}$. Then the domain of f is not subanalytic, whereas its graph is. If $g : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ has $-S := \{(-\frac{1}{n}, -n) : n \in \mathbb{N}\}$ as its graph, both its domain and epigraph are not subanalytic.

Additional geometrical properties like convexity are also not sufficient to obtain regularity on the domain. This is shown in the example below.

Example 2.5. Let $\{q_n\}_{n \geq 1}$ be an enumeration of the rationals $\{q_n\}$, and define $h : \mathbb{R}^2 \to \mathbb{R} \cup \{+\infty\}$ in polar coordinates by

$$h(r, \theta) = \begin{cases} 0 & \text{if } r \in [0, 1), \\ n & \text{if } r = 1 \text{ and } \theta = q_n \pmod{2\pi}, \\ +\infty & \text{otherwise.} \end{cases}$$

Then h is convex and subanalytic, but its domain is not subanalytic.

As expected, such a behavior can be avoided by requiring the function to be globally subanalytic. The following two results are basic consequences of the projection theorem.

PROPOSITION 2.6. *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a globally subanalytic function. Then the domain, the epigraph, and the strict epigraph of f are globally subanalytic sets.*

PROPOSITION 2.7. *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a subanalytic function which is relatively bounded on its domain; that is, $\{f(x) : x \in dom\, f \cap B\}$ is bounded for every bounded subset B of $\mathbb{R}^n$. Then the domain, the epigraph, and the strict epigraph of f are subanalytic sets.*

Remark 2.8. Observe that Propositions 2.6 and 2.7 involve distinct assumptions and provide different results. This can be seen by considering, for instance, the subanalytic functions $f(x) = x^{-1}$ with dom $f = (0, +\infty)$ and $g := \delta_{\mathbb{N}}$.

The case for which functions under consideration are convex but not necessarily continuous requires more attention.

PROPOSITION 2.9. *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous convex and subanalytic function such that $\inf_{\mathbb{R}^n} f \in \mathbb{R}$. Define $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ as the epigraphical sum of f and the square function $\frac{1}{2}||\cdot||^2$, that is,*

$$h(x) = \inf\left\{f(u) + \frac{1}{2}||x-u||^2 : u \in \mathbb{R}^n\right\}, \quad x \in \mathbb{R}^n.$$

Then h is a C^1 subanalytic function.

Proof. The proof consists mainly of showing that the epigraphical sum of a convex function with a coercive function is a "graphically local" operation. The fact that h takes finite values and is a C^1 function is a classical result (see [20], for example). Therefore it suffices to prove that $h + \delta_B$ is subanalytic for every bounded subset B of $\mathbb{R}^n$. Let us fix some nonempty bounded set B of $\mathbb{R}^n$, and let us set $M = \sup\{h(x) : x \in B\}$. Thanks to the continuity of h we have $M < +\infty$.

The infimum in the definition of $h(x)$ is always attained at a unique point denoted $J(x)$, and the mapping $J : \mathbb{R}^n \to \mathbb{R}^n$ so defined is a nonexpansive mapping (see [6]). Moreover, the function f is bounded on the bounded set $J(B)$. Indeed, if $u = J(b)$ for some $b \in B$, the definition of J implies that

$$f(u) = f(J(b)) = h(b) - \frac{1}{2}||b - J(b)||^2 < M.$$

Let C be some ball containing the bounded set $J(B)$, and let $f^M : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be the function whose graph is given by Gr $f \cap (C \times [\inf_{\mathbb{R}^n} f, M])$. By definition the function f^M has a bounded subanalytic graph, and it is therefore globally subanalytic. According to the above considerations the values of h on B coincide with those of the function

$$\hat{h}(x) := \inf\left\{f^M(u) + \frac{1}{2}||x-u||^2 : u \in \mathbb{R}^n\right\}, \quad x \in \mathbb{R}^n.$$

The strict epigraph of $\hat{h}$ is the sum of the strict epigraphs of the bounded subanalytic function f^M and the square function $u \mapsto \frac{1}{2}||x-u||^2$ (which is globally subanalytic for it is semialgebraic). This yields that $\hat{h}$ (and consequently $h + \delta_B$) is globally subanalytic; hence h is subanalytic. ∎

The notion of subdifferential—that is, an appropriate multivalued operator playing the role of the usual gradient mapping—is crucial for our considerations. In what follows we denote by $\langle \cdot, \cdot \rangle$ the usual Euclidean product of $\mathbb{R}^n$.

DEFINITION 2.10 (subdifferential; see, e.g., [20, Definition 8.3]). (i) *The* Fréchet subdifferential $\hat{\partial} f(x)$ *of a lower semicontinuous function f at $x \in \mathbb{R}^n$ is given by*

$$\hat{\partial} f(x) = \left\{x^* \in \mathbb{R}^n : \liminf_{y \to x, y \neq x} \frac{f(y) - f(x) - \langle x^*, y - x\rangle}{\|y - x\|} \geq 0\right\}$$

whenever $x \in dom\, f$, and by $\hat{\partial} f(x) = \emptyset$ otherwise.

(ii) *The* limiting subdifferential *at $x \in \mathbb{R}^n$, denoted by $\partial f(x)$, is the set of all cluster points of sequences $\{x_n^*\}_{n\geq 1}$ such that $x_n^* \in \hat{\partial} f(x_n)$ and $(x_n, f(x_n)) \to (x, f(x))$ as $n \to +\infty$.*

If the function f is of class C^1, the above notion coincides with the usual concept of gradient; that is, $\partial f(x) = \hat{\partial} f(x) = \{\nabla f(x)\}$. For a general lower semicontinuous function, the limiting subdifferential $\partial f(x)$ (thus, a fortiori the Fréchet subdifferential $\hat{\partial} f(x)$) can possibly be empty at several points $x \in \operatorname{dom} f$. Nevertheless (see, e.g., [20, Chapter 8]), both the domain of $\hat{\partial} f$ and (a fortiori) the domain of ∂f are dense in the domain of f.

Using the limiting subdifferential ∂f, we define the *nonsmooth slope* of f by

$$m_f(x) := \inf\{\|x^*\| : x^* \in \partial f(x)\}. \tag{4}$$

By definition, $m_f(x) = +\infty$ whenever $\partial f(x) = \emptyset$.

Let us recall that if f is continuous, the operator $\partial f : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ has a closed graph. This is also the case for a lower semicontinuous convex function, where both $\partial f(x)$ and $\hat{\partial} f(x)$ coincide with the classical subdifferential of convex analysis; that is,

$$\partial f(x) = \hat{\partial} f(x) = \{x^* \in \mathbb{R}^n : f(\cdot) - \langle x^*, \cdot\rangle \text{ has a global minimum at } x\}. \tag{5}$$

We are ready to state the notion of generalized critical point (in the sense of variational analysis).

Definition 2.11 (critical point). *A point $a \in \mathbb{R}^n$ is said to be a (generalized)* critical point *of the function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ if it belongs to the set*

$$\operatorname{crit} f := \{x \in \mathbb{R}^n : 0 \in \partial f(x)\}.$$

Remark 2.12. If f is lower semicontinuous convex or if $\operatorname{dom} f$ is closed and $f\,|_{\operatorname{dom} f}$ is continuous, then the graph of ∂f is closed, which implies that the set $\operatorname{crit} f$ of the critical points of f is closed. In that case, let us also observe that the slope $m_f(x)$ is a lower semicontinuous function, and that

$$\operatorname{crit} f = m_f^{-1}(0).$$

The following result illustrates further the properties of stability of subanalytic sets recalled in subsections 2.1 and 2.2.

Proposition 2.13. *Let f be an extended-real-valued function.*

(i) *If f is globally subanalytic, then the operators $\hat{\partial} f$ and ∂f, the function m_f, and the set* $\operatorname{crit} f$ *are globally subanalytic.*

(ii) *If f is subanalytic and relatively bounded on its domain, then the operators $\hat{\partial} f$ and ∂f, the function m_f, and the set* $\operatorname{crit} f$ *are subanalytic.*

Proof. The local nature of the Fréchet and the limiting subdifferential allows us to restrict our proof to the globally subanalytic function $f_B := f + \delta_B$, where B is some nontrivial ball. It suffices therefore to establish (i).

Thanks to the projection theorem (Theorem 2.3), the proof becomes a routine application of [8, Theorem 1.13], which asserts that if $\Phi(x_1, \ldots, x_n)$ is a first order formula (in the language of the subanalytic structure of $\mathbb{R}^n$), then the set $\{(x_1, \ldots, x_n) \in \mathbb{R}^n : \Phi(x_1, \ldots, x_n)\}$ is definable, or in other words, it belongs to the structure.[1]

[1] Global subanalytic sets form a model-complete first order theory. In fact, whether or not a structure is "model complete" depends only on the theory of the structure, that is, the set of the sentences (i.e., quantifier-free formulas) of its language which are true in this theory. We refer to [23, p. 1052] for more details.

As an illustration of this standard technique, let us prove that the operator $\hat{\partial} f : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is globally subanalytic. To this end, set $A = \operatorname{epi} f$, $\Gamma = \operatorname{Gr} f$, and $D = \operatorname{dom} f$, which are all globally subanalytic sets. According to Definition 2.10(ii) the graph $\operatorname{Gr} \hat{\partial} f$ of the Fréchet subdifferential $\hat{\partial} f(x)$ is the set of $(x, x^*) \in \mathbb{R}^n \times \mathbb{R}^n$ such that

$$\forall \varepsilon > 0, \exists \delta > 0, \forall (y, \beta) \in (B(x, \delta) \times \mathbb{R}) \cap A \Rightarrow (x, \beta - \langle x^*, y - x \rangle + \varepsilon \|y - x\|) \in A,$$

where $B(x, \delta)$ denotes the open ball of center x and radius $\delta > 0$. Since the above first order formula involves only globally subanalytic sets (namely, the subanalytic sets $B(x, \delta), \mathbb{R}$, and A), it follows that $\operatorname{Gr} \hat{\partial} f$ is subanalytic.

Subanalyticity of the graphs of the operator ∂f and of the function m_f can be proved similarly. Finally, $\operatorname{crit} f$ being the inverse image of (the subanalytic set) $\{0\}$ by m_f, it is a subanalytic set. □

Similarly one obtains the following corollary.

COROLLARY 2.14. *Under the assumptions of Proposition* 2.13(ii), *the restrictions of the multivalued mappings* $\hat{\partial} f$, ∂f, *and of the slope function* m_f *to any bounded subanalytic subset of* $\mathbb{R}^n$ *are globally subanalytic.*

Remark 2.15. The assumptions (and consequently the results) of the statements (i) and (ii) of Proposition 2.13 are of different natures. For example, let us consider the lower semicontinuous convex function $f : \mathbb{R}^2 \to \mathbb{R} \cup \{+\infty\}$, defined by

$$f(x, y) = \begin{cases} x^2/y & \text{if } y > 0, \\ 0 & \text{if } x = y = 0, \\ +\infty & \text{elsewhere.} \end{cases}$$

Then Proposition 2.13(i) applies, but not (ii), since f is not relatively bounded on $\operatorname{dom} f$.

3. Main results.

3.1. The Łojasiewicz inequality for subanalytic continuous functions. Assuming f subanalytic, and having a closed domain relative to which it is continuous, the set $\operatorname{crit} f$ is closed (Remark 2.12) and subanalytic (Proposition 2.13), so it has a locally finite number of connected components (see subsection 2.1). For any a in $\operatorname{crit} f$, let us denote by $(\operatorname{crit} f)_a$ the connected component of $\operatorname{crit} f$ containing a. In [5, Theorem 13] it has been established that

$$f \text{ is constant on } (\operatorname{crit} f)_a . \tag{6}$$

The proof of (6) relies on a fundamental structural result about subanalytic functions (stratification) and on the Pawłucki generalization of the Puiseux lemma; see [5]. Nevertheless, (6) can be easily proved for continuous functions that also satisfy

$$\hat{\partial} f(x) = \partial f(x) \quad \text{for all } x \in \mathbb{R}^n. \tag{7}$$

Indeed, given x and y in some connected component S_i of $\operatorname{crit} f$, we consider the continuous subanalytic path $z : [0, 1] \to S_i$ with $z(0) = x$ and $z(1) = y$, and the subanalytic function $h(t) = (f \circ z)(t)$ (see subsection 2.1). Since $0 \in \hat{\partial} f(z(t))$ for all $t \in [0, 1]$, from the "monotonicity lemma" and the chain rule for the Fréchet subdifferential [20, Theorem 10.6] we get $h'(t) = 0$ for almost all t. It follows that h is constant on $[0, 1]$, whence $f(x) = f(y)$.

Examples of continuous functions that satisfy (7) are C^1 functions (for which $\partial f(x) = \hat{\partial} f(x) = \{\nabla f(x)\}$), proximal retracts (or lower-C^2 functions; see [20, Definition 10.29] and section 4), or more generally subdifferentially regular functions [20, Definition 7.25].

The main result of subsection 3.1 can now be stated as follows.

THEOREM 3.1. *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a subanalytic function with closed domain, and assume that $f\,|_{\mathrm{dom} f}$ is continuous. Let $a \in \mathbb{R}^n$ be a critical point of f. Then there exists an exponent $\theta \in [0, 1)$ such that the function*

$$\frac{|f - f(a)|^\theta}{m_f} \tag{8}$$

is bounded around a.

Note that we have adopted here the following conventions: $0^0 = 1$ and $\infty/\infty = 0/0 = 0$.

Proof. Let us set $S = \operatorname{crit} f$ and $S_a = (\operatorname{crit} f)_a$. Replacing if necessary f by $g(x) = f(x) - f(a)$, there is no loss of generality to assume $f(a) = 0$, so that (6) implies $S_a \subset f^{-1}(0)$.

We may also assume that f is globally subanalytic and that the set S_a is compact. Indeed, if this is not the case, then we replace the function f by the globally subanalytic function g defined (for some $R > 0$) by $g(x) = f(x) + \delta_{\bar{B}(a,R)}(x)$, where $\delta_{\bar{B}(a,R)}$ denotes the indicator function of the closed ball $\bar{B}(a, R)$. Then g has a closed domain relative to which it is continuous, a is a critical point for g, and $(\operatorname{crit} g)_a \cap B(a, R) = S_a \cap B(a, R)$. Establishing (8) for f is thus the equivalent of doing so for the globally subanalytic function g.

It is also sufficient to establish separately that the function $x \mapsto [m_f(x)]^{-1} |f(x)|^\theta$ is bounded when x varies inside the subanalytic set $f^{-1}((0, +\infty])$, and subsequently to do the same when x varies in $f^{-1}((-\infty, 0])$. Since this latter assertion will follow by reproducing essentially the same arguments, we may assume with no loss of generality that $f \geq 0$.

Let us choose $\Delta > 0$ so that the compact set $U = \{x \in \mathbb{R}^n : d_{S_a}(x) \leq \Delta\} \cap \mathrm{dom}\, f$ separates S_a from the other connected components of S. Note that U is a globally subanalytic set (see subsection 2.1). We claim that for all $\bar{x}$ in the boundary of S_a we have

$$\lim_{\substack{x \to \bar{x} \\ x \in U \setminus S_a}} \frac{f(x)}{m_f(x)} = 0. \tag{9}$$

If the above limit were not zero, there would exist a sequence $\{(x_p, x_p^*)\}$ in $\mathrm{Gr}\, \partial f$ and $r > 0$ with $x_p \to \bar{x}$ as $p \to +\infty$ and such that $f(x_p) > r||x_p^*|| > 0$ for all p. By the definition of the limiting subdifferential there exists a sequence $(y_p, y_p^*) \in \mathrm{Gr}\, \hat{\partial} f$ such that $f(y_p) > r||y_p^*|| > 0$, where y_p converges to $\bar{x}$. This proves that for some $r > 0$ the point $\bar{x}$ belongs to the closure of the set

$$F = \{x \in U \setminus S_a : \exists x^* \in \hat{\partial} f(x), f(x) > r\, ||x^*|| > 0\}.$$

Owing to Proposition 2.13 (i) the latter set is globally subanalytic, so by the "curve selection lemma" (subsection 2.1) there exists an analytic curve $z : (-1, 1) \to \mathbb{R}^n$ with $z(0) = \bar{x}$ and $z((0, 1)) \subset F$. Hence for all small $t > 0$ there exists a nonzero subgradient $z^*(t) \in \hat{\partial} f(z(t))$ satisfying

$$f(z(t)) > r\, \|z^*(t)\| > 0. \tag{10}$$

Thanks to the continuity of $f\,|_{\mathrm{dom} f}$ at $\bar{x} = z(0)$ the subanalytic function

$$[0,1) \ni t \mapsto h(t) = (f \circ z)(t)$$

is continuous at $t = 0$, and (6) implies that $h(0) = f(\bar{x}) = 0$. Applying the "monotonicity lemma" (subsection 2.1) to the globally subanalytic function h and the chain rule calculus for the Fréchet subdifferential [20, Theorem 10.6], we get for t small enough that $|h'(t)| \leq M \|z^*(t)\|$, where $M = \max\{\|\dot{z}(t)\| : t \in (-1/2, 1/2)\}$. Then by applying (10), it follows that

$$\frac{h(t)}{|\,h'(t)\,|} > r\,M^{-1} > 0 \qquad \text{for all small } t > 0. \tag{11}$$

Considering the Puiseux development of h around $t = 0$ (see subsection 2.1), we conclude that for some positive rational q and some $c > 0$ we have $h(t) = ct^q + o(t^q)$ for all small $t > 0$. By differentiating the Puiseux development of h at $t = 0$ and substituting into (11), we obtain a contradiction.

Let us now establish (8). To this end, let us consider the globally subanalytic function

$$\varphi(t) = \inf\{m_f(x) : x \in U \cap f^{-1}(t)\} \quad \text{if } t \in \mathbb{R}_+.$$

Clearly $\varphi(0) = 0$, while from the definition of U, it ensues that $0 < \varphi(t) \leq +\infty$ for all small $t > 0$. If for every $\delta > 0$ the function φ assumes at least one infinite value in the interval $(0, \delta)$, then the subanalyticity of dom φ guarantees that 0 is an isolated point in dom φ. In this case (8) holds trivially. We may thus assume that φ is finite around 0. Evoking again the "monotonicity lemma" (subsection 2.1), we deduce that

$$l = \lim_{t \to 0^+} \varphi \in [0, +\infty].$$

In case $l \neq 0$, equation (8) follows easily (with $\theta = 0$), so we may assume $l = 0$ and φ continuous. In this case, we consider the Puiseux expansion of φ, which has the form

$$\varphi(t) = \sum_{n=0}^{+\infty} a_n t^{\frac{n}{k}} \quad \text{for all small } t > 0,$$

where k is a positive integer. Let $n_0 \in \mathbb{N}^*$ be the first integer such that $a_{n_0} \neq 0$, and let us set $\eta = \frac{n_0}{k}$. Then

$$\varphi(t) = ct^\eta + o(t^\eta), \tag{12}$$

where $c := a_{n_0} > 0$. Unless (8) holds trivially, we may assume by (6) that there exists a sequence $\{x_\nu\}_\nu \subset U \smallsetminus S_a$ such that $x_\nu \to a$, $m_f(x_\nu) \to 0$, and $f(x_\nu) \geq 0$. Let us consider the globally subanalytic set

$$A = \{x \in U \smallsetminus S_a : m_f(x) = \varphi(f(x)),\ f(x) \geq 0\} \neq \emptyset.$$

We claim that

$$\mathrm{cl}\, A \cap S_a \neq \emptyset. \tag{13}$$

Indeed, if (13) were not true, then by a standard compactness argument, there would exist an open neighborhood V around S_a such that $S_a \subset V \cap \mathrm{dom}\, f \subset U$

and $A \cap V = \emptyset$. Setting $t_\nu = f(x_\nu)$ (for the sequence $\{x_\nu\}_\nu$ mentioned above) and considering $y_\nu \in U$ such that $m_f(y_\nu) = \varphi(t_\nu)$ (by Remark 2.12, m_f is lower semicontinuous) and $f(y_\nu) = t_\nu$, we would obtain $\{y_\nu\}_\nu \subset U \setminus V$. By compactness, we could then assume that $y_\nu \to y \in U \setminus V$, which would yield (by continuity of φ) that $m_f(y) = 0$, that is, $y \in S_a$, and a contradiction follows.

Thus (13) holds, and there exists an analytic curve $z : (-1, 1) \to \mathbb{R}^n$ with $z(0) := b \in S_a$ and $z((0, 1)) \subset A$. As $s \searrow 0^+$ we get (by continuity of f and φ) that $f(z(s)) \to 0$ and $m_f(z(s)) = \varphi(f(z(s))) \to 0$. We deduce from (12) that

$$m_f(z(s)) = c(f(z(s)))^\eta + o((f(z(s))^\eta),$$

so (9) implies that $\eta < 1$. Take $\theta \in (\eta, 1)$ and apply (12) to obtain the existence of $t_0 > 0$ such that $\varphi(t) \geq ct^\theta$ for all $t \in [0, t_0)$. By using the continuity of $f \mid_{\operatorname{dom} f}$ at a, it follows that there exists $\mu > 0$ such that $|f(x)| < t_0$ for all $x \in \operatorname{dom} f \cap B(a, \mu)$. Finally, to obtain (8), we simply observe that

$$m_f(x) \geq \varphi(f(x)) \geq cf(x)^\theta \quad \text{for all } x \in B(a, \mu).$$

The proof is complete. □

Remark 3.2. Let us note that (8) still holds around any point $a \in \operatorname{dom} f \setminus \operatorname{crit} f$. Indeed, if $a \notin \operatorname{crit} f$, then $m_f(x)$ is bounded below away from 0 in a neighborhood of a, so (8) follows from the continuity of f. In this case, the assumption of subanalyticity is obviously not needed.

3.2. The Łojasiewicz inequality for subanalytic lower semicontinuous convex functions. In this subsection we are interested in lower semicontinuous convex subanalytic functions $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ which are somewhere finite, that is, convex functions for which $\operatorname{dom} f \neq \emptyset$. In this case, in view of (5), the set of critical points $\operatorname{crit} f$ is closed and convex and coincides with the set of minimizers of f.

Before proceeding let us recall classical facts from convex analysis (e.g., [20]). Let us denote by g the epigraphical sum of f and $\frac{1}{2}\|\cdot\|^2$ (see Proposition 2.9). The function $g : \mathbb{R}^n \to \mathbb{R}$ is finite-valued, and C^1 and enjoys the following properties:

(a) $g \leq f$.

(b) The set of critical points of g is exactly the set of critical points of f.

(c) The infimum values of f and g coincide; i.e., $\inf_{\mathbb{R}^n} f = \inf_{\mathbb{R}^n} g$.

The properties of g are related to the so-called Moreau regularizing process; for more details and further results, see [20].

We are ready to state the main result of this subsection.

THEOREM 3.3. *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous convex subanalytic function with $\operatorname{crit} f \neq \emptyset$. For any bounded set K there exists an exponent $\theta \in [0, 1)$ such that the function*

$$\frac{|f - \min f|^\theta}{m_f} \tag{14}$$

is bounded on K.

Proof. By Proposition 2.9, the function g defined above is subanalytic and continuous. Applying (b) and the results of the preceding section, we see that $S := \operatorname{crit} f$ is subanalytic. Let us show how g may be used to derive a growth condition for f. For any $x \in K$, the equivalence

$$d_S(x) = 0 \iff |g(x) - \min g| = 0,$$

combined with the Łojasiewicz factorization lemma (subsection 2.1) for the continuous subanalytic functions $|g - \min g|$ and d_S (restricted to the bounded set K), yields the existence of $r > 1$ and $c > 0$ such that

$$c\,[d_S(x)]^r \;\leq\; |g(x) - \min g| \quad \text{for all } x \in K.$$

On the other hand, the properties (a), (b), (c) of g imply that

$$|f(x) - \min f| \geq |g(x) - \min g| \quad \text{for all } x \in \mathbb{R}^n,$$

so that

$$[d_S(x)] \;\leq\; c^{-1/r}\,|f(x) - \min f|^{1/r}. \tag{15}$$

Moreover, since f is convex we get for all a in S and all $(x, x^*) \in \operatorname{Gr} \partial f$

$$f(a) \geq f(x) + \langle x^*, a - x\rangle.$$

Thus for all $(x, x^*) \in \operatorname{Gr} \partial f$ it follows that $|f(x) - f(a)| \;\leq\; \|x^*\|\,\|x - a\|$, and by taking the infimum over all $a \in S$, we obtain

$$|f(x) - \min f| \;\leq\; \|x^*\|\,d_S(x). \tag{16}$$

We therefore deduce from (15) that for all $x \in K$ and all $(x, x^*) \in \operatorname{Gr} \partial f$

$$|f(x) - \min f| \;\leq\; c^{-1/r}\,\|x^*\| \cdot |f(x) - \min f|^{1/r}.$$

By setting $\theta = 1 - r^{-1}$, the latter inequality implies $|f(x) - \min f|^{\theta} \;\leq\; c^{-1/r}\,m_f(x)$ for all $x \in K$, and (14) follows. □

Remark 3.4. The lower semicontinuous convex function f considered in Remark 2.15 provides an example where Theorem 3.3 applies while Theorem 3.1 does not.

Remark 3.5. A careful examination of the proof of Theorem 3.3 shows that the important assumption is not subanalyticity of the function, but rather the growth condition near critical values that subanalyticity implies. Indeed, let K be a compact set and f be any lower semicontinuous convex function f that satisfies

$$|f(x) - \min f| \geq\; c\,d_S(x)^r \qquad \text{for all } x \in K, \tag{17}$$

where $c > 0$, $r \geq 1$ and with $S = \operatorname{crit} f \neq \emptyset$. The argument of Theorem 3.3 may be then slightly modified in order to derive a Łojasiewicz inequality around any critical point a belonging to the interior of K.

Remark 3.6. From relation (16), which is true for all lower semicontinuous convex functions, a weaker version of (14) can be deduced. Indeed, if f is convex (but not necessarily subanalytic), then the function

$$\frac{|f - \min f|}{m_f}$$

is bounded around any critical point of f.

Remark 3.7. By using elementary arguments it can be shown that f satisfies the Łojasiewicz inequality around any point $a \in \operatorname{dom} f$ (cf. Remark 3.2).

4. Applications to dynamical systems. Throughout this section, unless otherwise stated, we make the following assumptions:

($\mathcal{H}1$) f is either lower semicontinuous convex or lower-C^2 with $\operatorname{dom} f = \mathbb{R}^n$.

($\mathcal{H}2$) f is somewhere finite ($\operatorname{dom} f \neq \emptyset$) and bounded from below.

We recall (see [20, Definition 10.29], for example) that a function f is called *lower-C^2* if for every $x_0 \in \operatorname{dom} f$ there exist a neighborhood U of x_0, a compact topological space S, and a jointly continuous function $F : U \times S \to \mathbb{R}$ satisfying $f(x) = \max_{s \in S} F(x, s)$ for all $x \in U$ and such that the (partial) derivatives $\nabla_x F(\cdot, \cdot)$ and $\nabla_x^2 F(\cdot, \cdot)$ exist and are jointly continuous.

A lower-C^2 function f is locally Lipschitz and locally representable as a difference of a convex continuous and a convex quadratic function [20, Theorem 10.33]. In particular, it satisfies

$$\partial f = \hat{\partial} f. \tag{18}$$

Note that (18) is also true for a lower semicontinuous convex function (see relation (5)).

As mentioned in the introduction, an important motivation for establishing the Łojasiewicz inequality for classes of nonsmooth functions is the expected asymptotic properties of the corresponding subgradient dynamical systems. This latter term refers to differential inclusions of the form

$$\dot{x}(t) + \partial f(x(t)) \ni 0,$$

where $\partial f : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is the limiting subdifferential of f. A *trajectory* of the above dynamical system is any absolutely continuous curve $x : [0, T) \to \mathbb{R}^n$ that satisfies

$$(\mathcal{G}) \qquad \begin{cases} \dot{x}(t) + \partial f(x(t)) \ni 0 \quad \text{a.e. on } (0, T), \\ \\ \partial f(x(t)) \neq \emptyset \quad \text{for all } t \in [0, T), \end{cases}$$

where the notation "a.e." stands for "almost everywhere" in the sense of the Lebesgue measure of $\mathbb{R}$. Let us also recall that an absolutely continuous function (or curve) $x(t)$ is a.e. differentiable and can be entirely determined, up to a constant, by integration of its classical derivative. A trajectory $x(t)$ is called *maximal* if there is no possible extension of its domain compatible with $(\mathcal{G})$.

The following existence-uniqueness result is known to hold (see [6, Theorem 3.2, p. 57] or [2, Chapter 3.4] for the convex case, and [6, Proposition 3.12, p. 106] for the convex case with Lipschitz perturbation; see also [9] for related work).

Existence of trajectories. Under the assumptions ($\mathcal{H}1$) and ($\mathcal{H}2$), for every $x_0 \in \mathbb{R}^n$ such that $\partial f(x_0) \neq \emptyset$, there exists a unique trajectory $x : [0, T) \to \mathbb{R}^n$ of $(\mathcal{G})$ satisfying

$$(\mathcal{T}) \qquad x(0) = x_0.$$

In addition, the function $h := f \circ x$ is absolutely continuous.

Let us now recall some classical consequences of (18) and of the above existence result. For the sake of completeness, some elementary proofs are provided.

COROLLARY 4.1. *Let $x : [0,T) \to \mathbb{R}^n$ be a trajectory of $(\mathcal{G})$ satisfying $(\mathcal{T})$.*
(i) *For almost all $t \in (0,T)$*

$$\frac{d}{dt}(f \circ x)(t) = \langle \dot{x}(t), x^* \rangle \quad \textit{for all } x^* \in \partial f(x(t)).$$

(ii) *For almost all $t \in (0,T)$, the function $x^* \mapsto \langle \dot{x}(t), x^* \rangle$ is constant on $\partial f(x(t))$.*
(iii) *The trajectory x can be extended to a maximal trajectory $\hat{x} \in W^{1,2}(\mathbb{R}_+ ; \mathbb{R}^n)$.*

Proof. Set $h = f \circ x$ and note that the absolutely continuous functions h and x are simultaneously differentiable on $(0,T) \backslash N$, where N is a set of measure zero. Let $t \in (0,T) \backslash N$. Since $x(t) \in \operatorname{dom} \partial f$ and $\partial f(x(t)) = \hat{\partial} f(x(t))$, one may adapt the ideas of [6, Lemma 3.3, p. 73] (chain rule) to obtain

$$\partial h(t) = \{h'(t)\} = \left\{ \frac{d}{dt}(f \circ x)(t) \right\} = \{ \langle \dot{x}(t),\, x^* \rangle,\ x^* \in \partial f(x(t)) \}.$$

Thus (i) and (ii) follow.

To establish (iii) let us first prove that $x \in W^{1,2}((0,T) ; \mathbb{R}^n)$. Thanks to $(\mathcal{G})$, we deduce from (i) that

$$\frac{d}{dt}(f \circ x)(t) = -\|\dot{x}(t)\|^2 \quad \text{for all } (0,T).$$

Hence f is a Lyapunov function of the dynamical system $(\mathcal{G})$, and

$$\int_0^T \|\dot{x}(t)\|^2 dt = f(x_0) - f(x(T)) < +\infty;$$

that is, $\dot{x} \in L^2((0,T) ; \mathbb{R}^n)$. Note that $\dot{x}(t)$ remains bounded as t converges to T. (For a lower semicontinuous convex function f this is a classical result (see [2, p. 147], for example); if f is lower-C^2, this follows from $(\mathcal{G})$ and the fact that ∂f is locally bounded around T.) Since the graph of ∂f is closed (Remark 2.12) we get $x(T) \in \operatorname{dom} \partial f$. Thus, thanks to the existence result $(\mathcal{T})$, the initial trajectory is in fact extendible to a semiopen interval $[0, T+\delta)$, for some $\delta > 0$, containing $[0,T]$. A standard argument shows that the maximal extension of $x(t)$ is defined in $(0, +\infty)$. □

An interesting hidden property of $(\mathcal{G})$ is the following.

COROLLARY 4.2. *Let x be a maximal trajectory of $(\mathcal{G})$ satisfying $(\mathcal{T})$. Then for almost all $t \in \mathbb{R}_+$*

$$\|\dot{x}(t)\| = m_f(x(t)) \qquad \textit{and} \qquad \frac{d}{dt}(f \circ x)(t) = -\left[m_f(x(t))\right]^2.$$

Proof. From $(\mathcal{G})$, we obtain the existence of a curve $t \mapsto g(t) \in \partial f(x(t))$ such that

$$\dot{x}(t) = -g(t) \quad \text{a.e. on } \mathbb{R}_+.$$

Combining this with Corollary 4.1(ii), we get that for almost all t in $\mathbb{R}_+$

$$\|g(t)\|^2 = \langle g(t),\, x^* \rangle \quad \text{for all } x^* \in \partial f(x(t)),$$

which yields via a standard argument that $\|g(t)\| = m_f(x(t))$. Now evoking Corollary 4.1(i) finishes the proof. □

Remark 4.3. Corollary 4.2 says that the trajectories of $(\mathcal{G})$ (the existence of which is guaranteed under the assumptions $(\mathcal{H}1)$ and $(\mathcal{H}2)$) are necessarily "slow solutions"

(see [2, p. 139]) of the differential inclusion $(\mathcal{G})$. In particular, if the trajectory $x(t)$ meets a critical point of f, that is, if there exists $t_0 > 0$ such that $m_f(x(t_0)) = 0$, then Corollary 4.2 guarantees that the trajectory stops there; that is, $x(t) = x(t_0)$ for all $t \geq t_0$. In this case, the trajectory has a finite length equal to $\int_0^{t_0} \|\dot{x}(s)\| ds$.

Another consequence of Corollary 4.2 is that $(\mathcal{G})$ defines a descent method in the sense that f decreases along any trajectory. Although compactness implies that bounded trajectories have at least one cluster point as $t \to +\infty$, those might not converge towards one of them—and a fortiori, have an infinite length. The next result shows that this cannot happen if f is assumed subanalytic (or more generally, if f satisfies the Łojasiewicz inequality). Indeed, via a "Łojasiewicz-type" argument (e.g., [14]) we establish successively that the tail of the trajectory is trapped inside a convenient ball of its cluster point, that this tail necessarily has a finite length, and finally that the trajectory converges to this cluster point. In the remainder, in addition to $(\mathcal{H}1)$ and $(\mathcal{H}2)$, the following is also assumed:

$(\mathcal{H}3)$ f is a *subanalytic* function.

Let us give some examples of subanalytic functions related with optimization problems.

Example 4.4. - (supremum operations) Let $g : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ be an analytic function, and let K be a compact subanalytic subset of $\mathbb{R}^p$. Then

$$f(x) = \sup_{y \in K} g(x, y)$$

is a lower-C^2 subanalytic function (see [4], for example). If in addition $x \mapsto g(x, y)$ is convex for all y, then f is convex.

- (constraints sets) Let $g_i : \mathbb{R}^n \to \mathbb{R}, i \in \{1, \ldots, m\}$, be a family of analytic functions. The feasible set

$$C := \{x \in \mathbb{R}^n : g_i(x) \leq 0, \forall i \in \{1, \ldots, m\}\}$$

together with its indicator function are subanalytic objects.

- (Barrier and penalty functions) Those can be used to minimize convex functions via parametric versions of $(\mathcal{G})$. Typical examples on $\mathbb{R}$ are the functions $h_1 : x > 0 \mapsto x^{-p} (p \geq 1)$, $h_2 : x \geq 0 \mapsto -x^\nu$, $(\nu \in (0, 1))$, $h_3(x) = x^2$ if $x \leq 0$ and $h_3(x) = 0$ otherwise.

We are now ready to state the following result.

THEOREM 4.5. *Assume that a function f satisfies $(\mathcal{H}1)$–$(\mathcal{H}3)$. Then any bounded maximal trajectory of $(\mathcal{G})$ has a finite length and converges to some critical point of f.*

Proof. Let $\{x(t)\}_{t \geq 0}$ be a bounded maximal trajectory of $(\mathcal{G})$. By Corollary 4.1, the trajectory is defined over all $\mathbb{R}_+$. Using $(\mathcal{H}2)$ and Corollary 4.2(iii), we conclude that there exists $\beta \in \mathbb{R}$ such that $\lim_{t \to +\infty} f(x(t)) = \beta$. Replacing f by $f - \beta$ and using the basic rules of subdifferential calculus, we may assume that $\beta = 0$.

In view of Remark 4.3, we may also assume that $f(x(t)) \neq 0$ for all $t > 0$. Consequently, the function $t \mapsto (f \circ x)(t)$ is positive and strictly decreasing to 0 as $t \to +\infty$. Moreover, by compactness, there exists some cluster point $a \in \mathbb{R}^n$ for the trajectory $x(t)$. So there exists an increasing sequence $(t_n)_{n \geq 1}$ with $t_n \to +\infty$ such that

$$\lim_{t_n \to +\infty} x(t_n) = a. \tag{19}$$

By continuity of $(f \circ x)$ we deduce that $f(a) = 0$. Using $(\mathcal{T})$, (19), and the fact that ∂f has a closed graph (see Remark 2.12), we deduce that $a \in \operatorname{dom} \partial f$. We do not know yet whether a is critical or not, but nevertheless, the Łojasiewicz inequality holds around a. Indeed, if $a \in \operatorname{crit} f$, then use Theorem 3.1 or Theorem 3.3, and if $a \notin \operatorname{crit} f$, then just recall Remarks 3.2 and 3.7. It follows that there exist $c > 0$, $\theta \in [0, 1)$, and $\varepsilon > 0$ (defining an open neighborhood $B(a, \varepsilon)$ of a) such that

$$|f(x)|^\theta \leq c\, m_f(x) \quad \text{for all } x \in B(a, \varepsilon). \tag{20}$$

Let us consider the (positive, absolutely continuous) function $\tilde{h} = (f \circ x)^{1-\theta}$. Since $x(t) \to a$ and since the function $\tilde{h}$ is strictly decreasing and converges to 0 (as $t \to +\infty$), there exists $t_0 > 0$ such that for all $t \geq t_0$

$$\frac{|\tilde{h}(t) - \tilde{h}(t_0)|}{c^{-1}(1-\theta)} \leq \frac{\varepsilon}{3}, \tag{21}$$

with

$$\|x(t_0) - a\| \leq \frac{\varepsilon}{3}. \tag{22}$$

Let us set

$$T_{\text{out}} := \inf \{\, t \geq t_0,\, x(t) \notin B(a, \varepsilon) \,\}. \tag{23}$$

By continuity of the trajectory we have $t_0 < T_{\text{out}} \leq +\infty$.

Claim $T_{\text{out}} = +\infty$ (that is, the tail of the trajectory remains trapped in $B(a, \varepsilon)$).

Proof of the claim. For almost all $t \in [t_0, T_{\text{out}})$ we have

$$\begin{aligned} \frac{d}{dt}\tilde{h}(t) &= (1-\theta)\, f(x(t))^{-\theta}\, \frac{d}{dt}(f \circ x)(t) \\ &\leq -(1-\theta)\, f(x(t))^{-\theta}\, [m_f(x(t))]^2 \\ &\leq -(1-\theta)\, c^{-1}\, m_f(x(t)), \end{aligned}$$

where we have successively used Corollary 4.2 and (20). By integration, we obtain for all $t \in [t_0, T_{\text{out}})$

$$\int_{t_0}^{t} m_f(x(s))ds \leq -\left[\frac{\tilde{h}(t) - \tilde{h}(t_0)}{c^{-1}(1-\theta)}\right], \tag{24}$$

which according to (21) and Corollary 4.2, yields

$$\int_{t_0}^{t} \|\dot{x}(s)\|ds \leq \frac{\varepsilon}{3} \quad \text{for all } t \in [t_0, T_{\text{out}}]. \tag{25}$$

To see that $T_{\text{out}} = +\infty$, we just argue by contradiction. If $T_{\text{out}} < +\infty$, then using (22) and (25), we obtain

$$\|x(T_{\text{out}}) - a\| \leq \left| \left(x(t_0) + \int_{t_0}^{T_{\text{out}}} \|\dot{x}(s)\|ds \right) - a \right| \leq \frac{2\varepsilon}{3}.$$

The latter obviously contradicts (23). Thus $T_{\text{out}} = +\infty$, and the claim is proved.

Resorting to (25) again, we conclude that $\int_{t_0}^{+\infty} \|\dot{x}(s)\| ds \leq \frac{\varepsilon}{3}$, so $x(t)$ has a finite length and hence converges. Thus $\lim_{t\to+\infty} x(t) = a$, and $m_f(x(t))$ admits 0 as a limit point. By using the closedness of $\operatorname{Gr} \partial f$, we conclude that a is a critical point of f. □

Remark 4.6 (generalized gradient conjecture). The "gradient conjecture" of Thom [22] can obviously be reformulated in this nonsmooth setting. For any bounded trajectory $x(\cdot)$ of $(\mathcal{G})$, let us set $x_\infty := \lim_{t\to+\infty} x(t)$. Is it true that

$$t \mapsto \frac{x(t) - x_\infty}{||x(t) - x_\infty||}$$

has a limit as t goes to infinity? For real-analytic functions this conjecture has been proved by Kurdyka, Mostowski, and Parusiński [14].

Before we proceed to an estimate of the rate of convergence, let us introduce some terminology.

- We define

$$\sigma(t) = \int_t^{+\infty} \|\dot{x}(s)\| ds \quad \text{for all } t \in \mathbb{R}_+ \tag{26}$$

  to be the tail length function for the trajectory $x(t)$.
- A *Łojasiewicz exponent* of the function f at a point $a \in \mathbb{R}^n$ of its domain is any number $\theta \in [0, 1)$ for which the Łojasiewicz inequality holds around a.

Let us finally point out some facts arising from the proof of Theorem 4.5. Replacing $\tilde{h}(t)$ by $[f(x(t)]^{1-\theta}$ and $m_f(x(s))$ by $\|\dot{x}(s)\|$ (see Corollary 4.2) in (24) and letting $t \to +\infty$, we deduce

$$\int_{t_0}^{+\infty} \|\dot{x}(s)\| ds \;\leq\; \frac{c}{(1-\theta)}\, f(x(t_0))^{1-\theta}.$$

The above inequality remains true for every $t \geq t_0$ (in view of the Claim). Thus assuming $\theta > 0$ and evoking again (20) and Corollary 4.2, we obtain (for $k = c^{1/\theta}$)

$$\int_t^{+\infty} \|\dot{x}(s)\| ds \;\leq\; \frac{k}{(1-\theta)}\, \|\dot{x}(t)\|^{\frac{1-\theta}{\theta}} \qquad \text{for all } t \geq t_0. \tag{27}$$

We are now ready to state the following result.

THEOREM 4.7. *Under the assumptions* $(\mathcal{H}1)$–$(\mathcal{H}3)$, *let* $x(t)$ *be a bounded maximal trajectory of* $(\mathcal{G})$. *Then* $x(t)$ *converges to some critical point* $a \in \mathbb{R}^n$ *of* f. *Let* $\theta \in [0, 1)$ *be a Łojasiewicz exponent at this point. Then there exist* $k > 0$, $k' > 0$, *and* $t_0 \geq 0$ *such that for all* $t \geq t_0$ *the following estimates hold:*

- *If* $\theta \in (\frac{1}{2}, 1)$, *then* $\|x(t) - a\| \leq k(t+1)^{-(\frac{1-\theta}{2\theta-1})}$.
- *If* $\theta = \frac{1}{2}$, *then* $\|x(t) - a\| \leq k \exp(-k't)$.
- *If* $\theta \in [0, \frac{1}{2})$, *then* $x(t)$ *converges in finite time.*

Proof. We can always assume that $\theta > 0$. (If $\theta = 0$, we replace it by some $\theta' \in (0, 1/2)$, and we proceed as below.)

Let U be a neighborhood of a in which the Łojasiewicz inequality holds. Since $x(t)$ converges to a there exists $t_0 \geq 0$ such that $x(t) \in U$ for every $t \geq t_0$. In particular, (27) holds. Let us now consider the tail length function $\sigma(t)$ defined in (26). Note that

$$\|x(t) - a\| \leq \sigma(t). \tag{28}$$

Since $\dot{\sigma}(t) = -\|\dot{x}(s)\|$ for all $t \geq t_0$, inequality (27) yields

$$\sigma(t) \leq \frac{k}{(1-\theta)} \left[-\dot{\sigma}(t)\right]^{\frac{1-\theta}{\theta}}. \tag{29}$$

Thus $\sigma(t)$ is an absolutely continuous function and satisfies the following differential inequality:

$$\dot{\sigma}(t) \leq -L\left[\sigma(t)\right]^{\frac{\theta}{1-\theta}} \quad \text{for all } t \geq t_0, \tag{30}$$

where L is a positive constant. To obtain the announced estimates it suffices to solve the following differential equation—considering separately the cases $\theta \in (1/2, 1)$, $\theta = 1/2$, and $\theta \in (0, 1/2)$:

$$\begin{cases} \dot{y}(t) = -L\left[y(t)\right]^{\frac{\theta}{1-\theta}} & \text{for all } t \geq t_0, \\ \\ y(t_0) = \sigma(t_0). \end{cases} \tag{31}$$

The announced estimates then follow from (28) and the fact that $\sigma(t) \leq y(t)$ for all $t \geq t_0$. (Indeed, if $\sigma(\bar{t}) = y(\bar{t})$ for some $\bar{t} \geq t_0$, then a comparison of (30) and (31) shows that $\dot{\sigma}(\bar{t}) \leq \dot{y}(\bar{t})$.) The proof is complete. □

Remark 4.8. The results of this section can be generalized to a wider setting as follows. Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous function complying with the following requirements:

(i) $\operatorname{dom} f \neq \emptyset$ and $\hat{\partial} f = \partial f$.
(ii) either f is convex or $f\,|_{\operatorname{dom} f}$ is continuous.
(iii) f has the Łojasiewicz property; that is, property (8) holds around any critical point.

If we assume in addition that, for all initial conditions $x_0 \in \operatorname{dom} \partial f$, the differential inclusion $(\mathcal{G})$ has a (unique) global solution x such that $f \circ x$ is absolutely continuous, then both Theorems 4.5 and 4.7 can be extended in this wider setting.

Prominent examples of functions meeting the above-mentioned conditions are continuous subanalytic ϕ-convex functions [9], or lower semicontinuous convex functions satisfying some growth condition of the type (17).

Acknowledgments. The second author wishes to thank K. Kurdyka, M. Quincampoix, and P. Cardaliaguet for useful discussions. A major part of this work was accomplished during a research visit of the first and the second authors to the Center for Experimental and Constructive Mathematics (Simon Fraser University, Canada). These two authors wish to thank their hosts at Simon Fraser University for their hospitality. The authors would also like to thank the anonymous referees for their useful comments.

REFERENCES

[1] P.-A. Absil, R. Mahony, and B. Andrews, *Convergence of the iterates of descent methods for analytic cost functions,* SIAM J. Optim., 16 (2006), pp. 531–547.
[2] J.-P. Aubin and A. Cellina, *Differential Inclusions*, Grundlehren Math. Wiss. 264, Springer, New York, 1984.
[3] R. Benedetti and J.-J. Risler, *Real Algebraic and Semialgebraic Sets*, Hermann, Paris, 1990.
[4] E. Bierstone and P. Milman, *Semianalytic and subanalytic sets*, IHES Publ. Math., 67 (1988) pp. 5–42.

[5] J. Bolte, A. Daniilidis, and A.S. Lewis, *A Sard theorem for non-differentiable functions,* J. Math. Anal. Appl, 321 (2006), pp 729–740.
[6] H. Brézis, *Opérateurs maximaux monotones et semi-groupes de contraction dans des espaces de Hilbert*, North-Holland Math. Stud. 5, North–Holland, Amsterdam, 1973.
[7] F.H. Clarke, Yu. Ledyaev, R.I. Stern, and P.R. Wolenski, *Nonsmooth Analysis and Control Theory*, Graduate Texts in Math. 178, Springer-Verlag, New York, 1998.
[8] M. Coste, *An introduction to o-minimal geometry*, RAAG Notes, Institut de Recherche Mathématiques de Rennes, 1999.
[9] M. Degiovanni, A. Marino, and M. Tosques, *Evolution equations with lack of convexity,* Nonlinear Anal., 9 (1985), pp 1401–1443.
[10] L. van den Dries and C. Miller, *Geometric categories and o-minimal structures,* Duke Math. J., 84 (1996), pp. 497–540.
[11] M. Fremond, J. Haslinger, J.-J. Moreau, P.M. Suquet, and J.J. Telega, *Nonsmooth Mechanics and Applications*, J.-J. Moreau and P.D. Panagiotopoulos, eds., CISM Courses and Lectures 302, Springer-Verlag, Vienna, 1988.
[12] A. Haraux, *A hyperbolic variant of Simon's convergence theorem*, in Evolution Equations and Their Applications in Physical and Life Sciences (Bad Herrenalb, Germany, 1998), Lecture Notes in Pure and Appl. Math. 215, Dekker, New York, 2001, pp. 255–264.
[13] K. Kurdyka, *On gradients of functions definable in o-minimal structures,* Ann. Inst. Fourier, 48 (1998), pp. 769–783.
[14] K. Kurdyka, T. Mostowski, and A. Parusiński, *Proof of the gradient conjecture of R. Thom*, Ann. Math., 152 (2000), pp. 763–792.
[15] B. Lemaire, *The proximal algorithm,* in New Methods in Optimization and Their Industrial Uses (Pau/Paris, 1987), Internat. Schriftenreihe Numer. Math. 87, Birkhäuser, Basel, 1989, pp. 73–87.
[16] S. Łojasiewicz, *Une propriété topologique des sous-ensembles analytiques réels*, in Les Équations aux Dérivées Partielles, Éditions du Centre National de la Recherche Scientifique, Paris, 1963, pp. 87–89.
[17] S. Łojasiewicz, *Sur les trajectoires de gradient d'une fonction analytique*, Seminari di Geometria 1982-1983 (lecture notes), Dipartemento di Matematica, Universita di Bologna, 1984, pp. 115–117.
[18] S. Łojasiewicz, *Sur la géométrie semi- et sous-analytique,* Ann. Inst. Fourier, 43 (1993), pp. 1575–1595.
[19] J. Palis and W. De Melo, *Geometric Theory of Dynamical Systems. An Introduction,* Translated from the Portuguese by A.K. Manning, Springer-Verlag, New York/Berlin, 1982.
[20] R.T. Rockafellar and R. Wets, *Variational Analysis*, Grundlehren Math. Wiss. 317, Springer, New York, 1998.
[21] L. Simon, *Asymptotics for a class of non-linear evolution equations, with applications to geometric problems,* Ann. Math., 118 (1983), pp. 525–571.
[22] R. Thom, *Problèmes rencontrés dans mon parcours mathématique: Un bilan,* IHES Publ. Math., 70 (1989), pp. 199–214.
[23] A. Wilkie, *Model completeness results for expansions of the ordered field of real numbers by restricted Pfaffian functions and the exponential function*, J. Amer. Math. Soc., 9 (1996), pp. 1051–1094.

SIAM J. OPTIM.
Vol. 17, No. 4, pp. 1224–1238
© 2007 Society for Industrial and Applied Mathematics

QUADRATIC MATRIX PROGRAMMING*

AMIR BECK†

Abstract. We introduce and study a special class of nonconvex quadratic problems in which the objective and constraint functions have the form $f(\boldsymbol{X}) = \mathrm{Tr}(\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}) + 2\mathrm{Tr}(\boldsymbol{B}^T\boldsymbol{X}) + c, \boldsymbol{X} \in \mathbb{R}R^{n\times r}$. The latter formulation is termed *quadratic matrix programming* (QMP) of order r. We construct a specially devised semidefinite relaxation (SDR) and dual for the QMP problem and show that under some mild conditions strong duality holds for QMP problems with at most r constraints. Using a result on the equivalence of two characterizations of the nonnegativity property of quadratic functions of the above form, we are able to compare the constructed SDR and dual problems to other known SDRs and dual formulations of the problem. An application to robust least squares problems is discussed.

Key words. quadratic matrix programming, semidefinite relaxation, strong duality, nonconvex quadratic problems

AMS subject classifications. 90C20, 90C26, 90C46

DOI. 10.1137/05064816X

1. Introduction.

This work is concerned with nonconvex quadratic optimization problems of the form

$$\begin{aligned} &\min \mathrm{Tr}(\boldsymbol{X}^T\boldsymbol{A}_0\boldsymbol{X}) + 2\mathrm{Tr}(\boldsymbol{B}_0^T\boldsymbol{X}) + c_0 \\ &\text{s.t. } \mathrm{Tr}(\boldsymbol{X}^T\boldsymbol{A}_i\boldsymbol{X}) + 2\mathrm{Tr}(\boldsymbol{B}_i^T\boldsymbol{X}) + c_i \le \alpha_i, i \in \mathcal{I}, \\ &\quad \mathrm{Tr}(\boldsymbol{X}^T\boldsymbol{A}_j\boldsymbol{X}) + 2\mathrm{Tr}(\boldsymbol{B}_j^T\boldsymbol{X}) + c_j = \alpha_j, j \in \mathcal{E}, \\ &\quad \boldsymbol{X} \in \mathbb{R}^{n\times r}, \end{aligned} \tag{1}$$

with $\boldsymbol{A}_i = \boldsymbol{A}_i^T \in \mathbb{R}^{n\times n}, \boldsymbol{B}_i \in \mathbb{R}^{n\times r}, \alpha_i, c_i \in \mathbb{R}, i \in \{0\} \cup \mathcal{I} \cup \mathcal{E}$. Problems of the above type arise naturally in several applications such as robust least squares [9], and in problems involving orthogonal constraints such as the orthogonal procrustes problem [17] (see the discussion in section 2).

Problem (1) is called a *quadratic matrix programming* (QMP) problem of order r. Correspondingly, the objective and constraint functions are called *quadratic matrix (QM) functions*. It can be shown that every QM function is in particular a quadratic function with nr variables; see the discussion in section 2.1. Thus, the family of QMP problems is a special case of quadratically constrained quadratic programming (QCQP) problems. However, it is worthwhile to study these problems independently since, as we shall see, they enjoy stronger results than those currently known for the general QCQP problem. For example, we will establish strong duality results for QMP problems with at most r constraints (see section 3.2).

Strong duality is known to hold for only a few classes of nonconvex QCQP. The simplest and best-known example is the trust region problem, which consists of minimizing an indefinite quadratic function over a ball and admits an exact semidefinite relaxation (SDR); see [13, 8]. Extensions of this problem were considered in

*Received by the editors December 21, 2005; accepted for publication (in revised form) June 30, 2006; published electronically January 12, 2007.
http://www.siam.org/journals/siopt/17-4/64816.html
†Department of Industrial Engineering and Management, Technion—Israel Institute of Technology, Haifa 32000, Israel (becka@ie.technion.ac.il).

[12, 18, 5, 16]. In general these results cannot be extended to QCQP problems involving two constraints [19, 20]. An exception is the case in which all the functions involved (objective plus two constraints) are homogenous quadratic functions. In this case, it was proven in [19] that under mild conditions the semidefinite relaxation is tight. Another interesting tractable class of QCQP problems was considered in [1] in the context of quadratic problems with orthogonal constraints.

In this paper strong duality/tightness of the SDR is shown to hold for the class of QMP problems of order r with at most r constraints. In section 3 we construct an SDR and dual formulations for the QMP problem originating from a homogenization procedure specially devised to QMP problems. Using the SDR formulation combined with known results on the existence of low-rank solutions of semidefinite programs [3, 2, 14, 15], the strong duality result is shown to follow. Moreover, an algorithm for extracting a solution to the QMP problem from its associated SDR is described. In section 4 an alternative SDR and dual construction are discussed. These constructions stem from the standard construction of SDR and dual for QCQP problems. Using a result on the equivalence of two linear matrix inequality (LMI) representations of the claim on nonnegativity of a QM function, we are able to prove that the two SDR and dual formulations are equivalent. Finally, in section 5 we present an application of our results in the field of robust optimization.

Notation. For simplicity, instead of inf/sup we use min/max; however, this does not mean that we assume that the optimum is attained and/or finite. Vectors are denoted by boldface lowercase letters, e.g., $\boldsymbol{y}$, and matrices by boldface uppercase letters, e.g., $\boldsymbol{A}$. For two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, $\boldsymbol{A} \succ \boldsymbol{B}$ ($\boldsymbol{A} \succeq \boldsymbol{B}$) means that $\boldsymbol{A} - \boldsymbol{B}$ is positive definite (semidefinite). $\mathcal{S}^n = \{\boldsymbol{A} \in \mathbb{R}^{n\times n} : \boldsymbol{A} = \boldsymbol{A}^T\}$ is the set of symmetric $n \times n$ matrices, and $\mathcal{S}^n_+ = \{\boldsymbol{A} \in \mathbb{R}^{n\times n} : \boldsymbol{A} \succeq \boldsymbol{0}\}$ is the set all real $n \times n$ symmetric positive semidefinite matrices. $\boldsymbol{0}_{n\times m}$ is the $n \times m$ matrix of zeros, and $\boldsymbol{I}_r$ is the $r \times r$ identity matrix. For a matrix $\boldsymbol{M}$, $\mathrm{vec}(\boldsymbol{M})$ denotes the vector obtained by stacking the columns of $\boldsymbol{M}$. For a square matrix $\boldsymbol{U}$, $[\boldsymbol{U}]_r$ denotes the southeast $r \times r$ submatrix of $\boldsymbol{U}$; i.e., if $\boldsymbol{U} = (u_{ij})_{i,j=1}^{n+r}$, then $[\boldsymbol{U}]_r = (u_{ij})_{i,j=n+1}^{n+r}$. For two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, $\boldsymbol{A} \otimes \boldsymbol{B}$ denotes the corresponding Kronecker product. $\boldsymbol{E}^r_{ij}$ is the $r \times r$ matrix with 1 at the ijth component and 0 elsewhere, and δ_{ij} is the Kronecker delta, i.e., $\delta_{ii} = 1$ and $\delta_{ij} = 0$ for $i \neq j$. The value of the optimal objective function of an optimization problem

$$(\mathrm{P}) : \min\{f(\boldsymbol{x}) : \boldsymbol{x} \in C\}$$

is denoted by val(P). The optimization problem (P) is called *bounded below* if the minimum is finite, and termed *solvable* in the case where the minimum is finite and attained (similar definitions for maximum problems). We follow the MATLAB convention and use ";" for adjoining scalars, vectors, or matrices in a column. We also use some standard abbreviations such as SDP (semidefinite programming), LMI (linear matrix inequality), SDR (semidefinite relaxation), and QCQP (quadratically constrained quadratic programming), and some nonstandard abbreviations such as QM (quadratic matrix) and QMP (quadratic matrix programming).

2. Quadratic matrix problems.

2.1. Quadratic matrix functions: Definition and basic properties.

We begin by recalling that a quadratic function $g : \mathbb{R}^n \to \mathbb{R}$ is a function of the form

$$g(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} + 2\boldsymbol{b}^T \boldsymbol{x} + c, \tag{2}$$

where $\boldsymbol{A} \in \mathcal{S}^n, \boldsymbol{b} \in \mathbb{R}^n$, and $c \in \mathbb{R}$. We will also use the term "quadratic vector function" instead of "quadratic function" to distinguish it from the term "quadratic matrix function" defined below.

A *quadratic matrix (QM) function* of order r is a function $f : \mathbb{R}^{n\times r} \to \mathbb{R}$ of the form

$$f(\boldsymbol{X}) = \mathrm{Tr}(\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}) + 2\mathrm{Tr}(\boldsymbol{B}^T\boldsymbol{X}) + c, \quad \boldsymbol{X} \in \mathbb{R}^{n\times r}, \tag{3}$$

where $\boldsymbol{A} \in \mathcal{S}^n$, $\boldsymbol{B} \in \mathbb{R}^{n\times r}$, and $c \in \mathbb{R}$. If $\boldsymbol{B} = \boldsymbol{0}_{n\times r}, c = 0$, then f is called a *homogenous QM function* or a *QM form*. We note that every quadratic vector function is a QM function of order one. The opposite statement is also true: every QM function is in particular a quadratic vector function. Indeed, the function f from (3) can be written as follows:

$$f(\boldsymbol{X}) = f^V(\mathrm{vec}(\boldsymbol{X})), \tag{4}$$

where $f^V : \mathbb{R}^{nr} \to \mathbb{R}$ is defined by

$$f^V(\boldsymbol{z}) = \boldsymbol{z}^T(\boldsymbol{I}_r \otimes \boldsymbol{A})\boldsymbol{z} + 2vec(\boldsymbol{B})^T\boldsymbol{z} + c. \tag{5}$$

The function f^V is called the *vectorized function of* f. From the above relation we can immediately deduce that f is (strictly) convex if and only if $\boldsymbol{A} \succeq \boldsymbol{0}$ ($\boldsymbol{A} \succ \boldsymbol{0}$).[1]

2.2. QM problems. Our main objective is to study *quadratic matrix programming* (QMP) problems in which the goal is to minimize a QM objective function subject to equality and inequality QM constraints:

$$\begin{array}{ll} (\mathrm{QMP}) & \min f_0(\boldsymbol{X}) \\ & \text{s.t. } f_i(\boldsymbol{X}) \le \alpha_i, i \in \mathcal{I}, \\ & \quad f_j(\boldsymbol{X}) = \alpha_j, j \in \mathcal{E}, \\ & \quad \boldsymbol{X} \in \mathbb{R}^{n\times r}, \end{array} \tag{6}$$

where $f_i : \mathbb{R}^{n\times r} \to \mathbb{R}, i \in \mathcal{I} \cup \mathcal{E} \cup \{0\}$, are QM functions of order r given by

$$f_i(\boldsymbol{X}) = \mathrm{Tr}(\boldsymbol{X}^T\boldsymbol{A}_i\boldsymbol{X}) + 2\mathrm{Tr}(\boldsymbol{B}_i^T\boldsymbol{X}) + c_i, \quad \boldsymbol{X} \in \mathbb{R}^{n\times r},$$

with $\boldsymbol{A}_i \in \mathcal{S}^n, \boldsymbol{B}_i \in \mathbb{R}^{n\times r}$, and $c_i \in \mathbb{R}, i \in \{0\} \cup \mathcal{I} \cup \mathcal{E}$. The index sets $\{0\}, \mathcal{I}, \mathcal{E}$ are pairwise disjoint sets of nonnegative integers.

In the case where all the functions $f_i, i \in \mathcal{I} \cup \mathcal{E} \cup \{0\}$, are homogeneous QM functions of order r, the QMP problem (6) is called a *homogenous QMP problem* (of order r). By using the correspondence (4), we can represent the QMP problem as the QCQP problem:

$$\begin{array}{l} \min f_0^V(\boldsymbol{z}) \\ \text{s.t. } f_i^V(\boldsymbol{z}) \le \alpha_i, i \in \mathcal{I}, \\ \quad f_j^V(\boldsymbol{z}) = \alpha_j, j \in \mathcal{E}, \\ \quad \boldsymbol{z} \in \mathbb{R}^{nr}, \end{array} \tag{7}$$

which will be called the *vectorized QMP problem*.

[1] Indeed, $\boldsymbol{I}_r \otimes \boldsymbol{A}$ and $\boldsymbol{A}$ have the same eigenvalues (but with different multiplicities) [10].

The QMP problem appears in several fields of applications. Here we present two examples in which the QMP problem naturally arises.

Example 1. In the *orthogonal procrustes* problem [17] we seek to find a square matrix $\boldsymbol{X}$ which solves the following optimization problem:

$$\begin{aligned} &\min \|\boldsymbol{A}\boldsymbol{X} - \boldsymbol{B}\|_F^2 \\ &\text{s.t. } \boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}_r, \\ &\boldsymbol{X} \in \mathbb{R}^{r\times r}, \end{aligned}$$

where $\boldsymbol{A} \in \mathbb{R}^{n\times r}, \boldsymbol{B} \in \mathbb{R}^{n\times r}$. The orthogonal procrustes problem can be rewritten as a QMP problem with r^2 equality constraints:

$$\begin{aligned} &\min \mathrm{Tr}(\boldsymbol{X}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{X}) - 2\mathrm{Tr}(\boldsymbol{B}^T\boldsymbol{A}\boldsymbol{X}) + \|\boldsymbol{B}\|_F^2 \\ &\text{s.t. } \mathrm{Tr}(\boldsymbol{X}^T(\boldsymbol{E}_{ij}^r + \boldsymbol{E}_{ij}^r)\boldsymbol{X}) = 2\delta_{ij}, \quad 1 \le i,j \le r, \\ &\boldsymbol{X} \in \mathbb{R}^{r\times r}. \end{aligned}$$

We note that although the orthogonal procrustes problem can be solved efficiently [17], it is not clear whether the *unbalanced* orthogonal procrustes problem—in which $\boldsymbol{X}$ is not square—is tractable [7].

Example 2. The *robust least squares* (RLS) problem was introduced and studied in [9, 6].[2] Consider a linear system $\boldsymbol{A}\boldsymbol{x} \approx \boldsymbol{b}$, where $\boldsymbol{A} \in \mathbb{R}^{r\times n}, \boldsymbol{b} \in \mathbb{R}^r$, and $\boldsymbol{x} \in \mathbb{R}^n$. Assume that the matrix $\boldsymbol{A}$ is not fixed but rather given by a family of matrices[3] $\boldsymbol{A}+\boldsymbol{\Delta}^T$, where $\boldsymbol{A}$ is a known nominal value and $\boldsymbol{\Delta} \in \mathbb{R}^{n\times r}$ is an unknown perturbation matrix known to reside in a compact uncertainty set $\mathcal{U}$. The RLS approach to this problem is to seek a vector $\boldsymbol{x} \in \mathbb{R}^n$ that minimizes the worst case data error with respect to all possible values of $\boldsymbol{\Delta} \in \mathcal{U}$:

$$\min_{\boldsymbol{x}} \max_{\boldsymbol{\Delta}\in\mathcal{U}} \|\boldsymbol{b} - (\boldsymbol{A} + \boldsymbol{\Delta}^T)\boldsymbol{x}\|^2. \tag{8}$$

Now, by making some simple algebraic manipulation, we can rewrite the objective function in (8) as

$$\|\boldsymbol{b} - (\boldsymbol{A} + \boldsymbol{\Delta}^T)\boldsymbol{x}\|^2 = \mathrm{Tr}(\boldsymbol{\Delta}^T\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{\Delta}) + 2\mathrm{Tr}((\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x})\boldsymbol{x}^T\boldsymbol{\Delta}) + \mathrm{Tr}((\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x})(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x})^T),$$

so that the inner maximization problem in (8) takes the following form:

$$\max\{\mathrm{Tr}(\boldsymbol{\Delta}^T\mathbf{Q}\boldsymbol{\Delta}) + 2\mathrm{Tr}(\mathbf{F}^T\boldsymbol{\Delta}) + c : \boldsymbol{\Delta} \in \mathcal{U}\}, \tag{9}$$

where $\mathbf{Q}, \mathbf{F}$, and c depend on $\boldsymbol{x}$ and are given by

$$\mathbf{Q} = \boldsymbol{x}\boldsymbol{x}^T \in \mathcal{S}^n, \quad \mathbf{F} = \boldsymbol{x}(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x})^T \in \mathbb{R}^{n\times r}, \quad c = \|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|^2 \in \mathbb{R}. \tag{10}$$

In [9] the uncertainty set $\boldsymbol{\Delta}$ was chosen to be a simple Frobenius norm constraint, i.e.,

$$\mathcal{U} = \{\boldsymbol{\Delta} \in \mathbb{R}^{n\times r} : \mathrm{Tr}(\boldsymbol{\Delta}^T\boldsymbol{\Delta}) \le \rho\}.$$

[2]Here we study the unstructured case.

[3]The perturbation matrix appears in a transpose form so that the derived QM function will have the form (3). Furthermore, for the sake of simplicity we do not consider uncertainties in the RHS vector $\boldsymbol{b}$, although such uncertainties can be incorporated into our analysis in a straightforward manner.

The inner maximization problem (9) with the above choice of $\mathcal{U}$ is a QMP problem of order r with a single inequality constraint.

The fact that the uncertainty set $\mathcal{U}$ was given in [9] by a *single* quadratic constraint was a crucial element in establishing the tractability of the RLS problem. In fact, it is well known that in the structured case, the inner maximization problem of the RLS problem becomes NP-hard when the uncertainty set is given by an intersection of ellipsoids. Nonetheless, in section 5, using the results developed in sections 3 and 4, we will show that more complicated choices of $\mathcal{U}$ can be considered. In particular, we will prove in section 5 that the RLS problem remains tractable in the case where $\mathcal{U}$ is given by a set of at most r QM inequality constraints. The latter form of the uncertainty set can model, for example, the situation where each column of the perturbation matrix $\boldsymbol{\Delta}^T$ has a *separate* norm constraint.

3. Semidefinite relaxations of the QMP problem and strong duality results. We begin by constructing an SDR for the QMP problem. A natural approach for constructing such an SDR is to consider the SDR of the vectorized problem (7) (recall that problem (7) is a (QCQP)). However, this approach, which is discussed in detail in section 4, does not seem to offer useful theoretical insights into questions such as strong duality/tightness of SDR. For that reason we construct a *new* scheme, specifically devised to obtain an SDR for QMP problems (see section 3.1). Using the derived SDR, we will show in section 3.2 that, under some mild conditions, strong duality holds for QMP problems of order r with at most r constraints.

3.1. An SDR of the QMP problem. Recall that the homogenized version a quadratic vector function g given by (2) is the quadratic form $g^H : \mathbb{R}^{n+1} \to \mathbb{R}$ defined by

$$g^H(\boldsymbol{x};t) = \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} + 2\boldsymbol{b}^T\boldsymbol{x}t + ct^2. \tag{11}$$

The matrix associated with the quadratic form g^H is denoted by

$$\boldsymbol{M}(g) = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{b} \\ \boldsymbol{b}^T & c \end{pmatrix}. \tag{12}$$

We consider the following generalization of the above homogenization procedure to QM functions of order r: let f be the QM function given by (3); the *homogenized QM function* is denoted by $f^H : \mathbb{R}^{(n+r)\times r} \to \mathbb{R}$ and given by

$$f^H(\mathbf{Y};\boldsymbol{Z}) \equiv \mathrm{Tr}(\mathbf{Y}^T\boldsymbol{A}\mathbf{Y}) + 2\mathrm{Tr}(\boldsymbol{Z}^T\boldsymbol{B}^T\mathbf{Y}) + \frac{c}{r}\mathrm{Tr}(\boldsymbol{Z}^T\boldsymbol{Z}), \quad \mathbf{Y} \in \mathbb{R}^{n\times r}, \boldsymbol{Z} \in \mathbb{R}^{r\times r}, \tag{13}$$

which is a homogenous QM function of order r corresponding to the matrix

$$\boldsymbol{M}(f) \equiv \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}^T & \frac{c}{r}\boldsymbol{I}_r \end{pmatrix}. \tag{14}$$

In the case $r = 1$, definitions (13) and (14) coincide with the definitions of the homogenization of a quadratic function (11) and its associated matrix (12), respectively. The operator $\boldsymbol{M}$ will be used throughout the paper.

The homogenous function f^H satisfies the following easily verifiable properties, which will become useful in what follows:

$$f^H(\mathbf{Y};\boldsymbol{I}_r) = f(\mathbf{Y}) \quad \text{for every } \mathbf{Y} \in \mathbb{R}^{n\times r}, \tag{15}$$

$$f^H(\mathbf{Y};\boldsymbol{Z}) = f(\mathbf{Y}\boldsymbol{Z}^T) \text{ for every } \mathbf{Y} \in \mathbb{R}^{n\times r}, \boldsymbol{Z} \in \mathbb{R}^{r\times r} \text{ such that } \boldsymbol{Z}^T\boldsymbol{Z} = \boldsymbol{I}_r. \tag{16}$$

Using the above homogenization procedure for QM functions, we are able to construct (see Lemma 3.1 below) a homogeneous QMP problem of order r, equivalent to the (nonhomogeneous) QMP problem (6).

LEMMA 3.1. *Consider the following homogenized version of the QMP problem* (6):

$$\begin{aligned} &\min f_0^H(\mathbf{Y};\boldsymbol{Z}) \\ &s.t.\ f_i^H(\mathbf{Y};\boldsymbol{Z}) \le \alpha_i, i \in I, \\ &\quad f_j^H(\mathbf{Y};\boldsymbol{Z}) = \alpha_j, j \in \mathcal{E}, \\ &\quad \psi_{ij}(\mathbf{Y};\boldsymbol{Z}) = 2\delta_{ij}, 1 \le i \le j \le r, \\ &\quad \mathbf{Y} \in \mathbb{R}^{n\times r}, \boldsymbol{Z} \in \mathbb{R}^{r\times r}, \end{aligned} \tag{17}$$

where $\psi_{ij}(\mathbf{Y};\boldsymbol{Z}) = \mathrm{Tr}(\boldsymbol{Z}^T(\boldsymbol{E}_{ij}^r + \boldsymbol{E}_{ji}^r)\boldsymbol{Z})$ *and* δ_{ij} *is the Kronecker delta.*

1. *Suppose that the QMP problem* (6) *is solvable, and let* $\boldsymbol{X}^*$ *be an optimal solution of* (QMP). *Then problem* (17) *is solvable,* $(\boldsymbol{X}^*;\boldsymbol{I}_r)$ *is an optimal solution of* (17), *and* val(QMP) = val(17).
2. *Suppose that problem* (17) *is solvable, and let* $(\mathbf{Y}^*;\boldsymbol{Z}^*)$ *be an optimal solution of* (17). *Then problem* (QMP) *is solvable,* $\boldsymbol{X}^* = \mathbf{Y}^*(\boldsymbol{Z}^*)^T$ *is an optimal solution of* (QMP), *and* val(QMP) = val(17).

Proof. First note that the system of equalities

$$\mathrm{Tr}(\boldsymbol{Z}^T(\boldsymbol{E}_{ij}^r + \boldsymbol{E}_{ji}^r)\boldsymbol{Z}) = 2\delta_{ij}, \quad 1 \le i \le j \le r,$$

can be written as

$$\mathrm{Tr}((\boldsymbol{E}_{ij}^r + \boldsymbol{E}_{ji}^r)\boldsymbol{Z}\boldsymbol{Z}^T) = 2\delta_{ij}, \quad 1 \le i \le j \le r,$$

which, by using the symmetry of the matrix $\boldsymbol{Z}\boldsymbol{Z}^T$, is equivalent to

$$\boldsymbol{Z}^T\boldsymbol{Z} = \boldsymbol{Z}\boldsymbol{Z}^T = \boldsymbol{I}_r.$$

1. Let $\boldsymbol{X}^*$ be an optimal solution of (QMP). For every $(\mathbf{Y};\boldsymbol{Z})$, $(\mathbf{Y} \in \mathbb{R}^{n\times r}, \boldsymbol{Z} \in \mathbb{R}^{r\times r})$ in the feasible set of (17) (and in particular $\boldsymbol{Z}^T\boldsymbol{Z} = \boldsymbol{I}_r$) we have

$$f_0^H(\mathbf{Y},\boldsymbol{Z}) \overset{(16)}{=} f_0(\mathbf{Y}\boldsymbol{Z}^T) \ge f_0(\boldsymbol{X}^*) \overset{(15)}{=} f_0^H(\boldsymbol{X}^*;\boldsymbol{I}_r).$$

Therefore, $(\boldsymbol{X}^*;\boldsymbol{I}_r)$ is an optimal solution of (17) and val(QMP) = val(17).

2. Let $(\mathbf{Y}^*;\boldsymbol{Z}^*)$, $(\mathbf{Y} \in \mathbb{R}^{n\times r}, \boldsymbol{Z} \in \mathbb{R}^{r\times r})$ be an optimal solution of (17), and set $\boldsymbol{X}^* = \mathbf{Y}^*(\boldsymbol{Z}^*)^T$. Then for every $\boldsymbol{X} \in \mathbb{R}^{n\times r}$ which is in the feasible set of (QMP) we have

$$f_0(\boldsymbol{X}) \overset{(15)}{=} f_0^H(\boldsymbol{X};\boldsymbol{I}) \ge f_0^H(\mathbf{Y}^*;\boldsymbol{Z}^*) \overset{(16)}{=} f_0(\mathbf{Y}^*(\boldsymbol{Z}^*)^T) = f_0(\boldsymbol{X}^*),$$

and thus $\boldsymbol{X}^*$ is an optimal solution of (QMP) and val(QMP) = val(17). ∎

COROLLARY 3.2. *The QMP problem* (6) *is solvable if and only if problem* (17) *is solvable, and in that case* val(QMP) = val(17).

We will now exploit the homogenized QMP problem (17) in order to formulate a semidefinite relaxation. By denoting $W = (\mathbf{Y};\boldsymbol{Z}) \in \mathbb{R}^{(n+r)\times r}$, we conclude that

problem (17) can be written as

$$\begin{aligned}
&\min \operatorname{Tr}(\boldsymbol{M}(f_0)\boldsymbol{W}\boldsymbol{W}^T)\\
&\text{s.t. } \operatorname{Tr}(\boldsymbol{M}(f_i)\boldsymbol{W}\boldsymbol{W}^T) \le \alpha_i, i \in I,\\
&\quad \operatorname{Tr}(\boldsymbol{M}(f_j)\boldsymbol{W}\boldsymbol{W}^T) = \alpha_j, j \in \mathcal{E},\\
&\quad \operatorname{Tr}(\boldsymbol{N}_{ij}\boldsymbol{W}\boldsymbol{W}^T) = 2\delta_{ij}, 1 \le i \le j \le r,\\
&\quad \boldsymbol{W} \in \mathbb{R}^{(n+r)\times r},
\end{aligned}$$

where the operator $\boldsymbol{M}$ is defined in (14) and

$$\boldsymbol{N}_{ij} = \begin{pmatrix} \boldsymbol{0}_{n\times n} & \boldsymbol{0}_{n\times r} \\ \boldsymbol{0}_{r\times n} & \boldsymbol{E}_{ij}^r + \boldsymbol{E}_{ji}^r \end{pmatrix}, \quad 1 \le i \le j \le r.$$

Making the change of variables $\boldsymbol{U} = \boldsymbol{W}\boldsymbol{W}^T \in \mathcal{S}_+^{n+r}$, we conclude that problem (17) can be equivalently written as

$$\begin{aligned}
&\min \operatorname{Tr}(\boldsymbol{M}(f_0)\boldsymbol{U})\\
&\text{s.t. } \operatorname{Tr}(\boldsymbol{M}(f_i)\boldsymbol{U}) \le \alpha_i, i \in I,\\
&\quad \operatorname{Tr}(\boldsymbol{M}(f_j)\boldsymbol{U}) = \alpha_j, j \in \mathcal{E},\\
&\quad \operatorname{Tr}(\boldsymbol{N}_{ij}\boldsymbol{U}) = 2\delta_{ij}, 1 \le i \le j \le r,\\
&\quad \boldsymbol{U} \in \mathcal{S}_+^{n+r}, \operatorname{rank}(\boldsymbol{U}) \le r.
\end{aligned}$$

Omitting the "hard" constraint $\operatorname{rank}(\boldsymbol{U}) \le r$, we finally arrive at the following SDR of the QMP problem (6):

$$\begin{aligned}
\text{(SDRM)}\quad &\min \operatorname{Tr}(\boldsymbol{M}(f_0)\boldsymbol{U})\\
&\text{s.t. } \operatorname{Tr}(\boldsymbol{M}(f_i)\boldsymbol{U}) \le \alpha_i, i \in I,\\
&\quad \operatorname{Tr}(\boldsymbol{M}(f_j)\boldsymbol{U}) = \alpha_j, j \in \mathcal{E},\\
&\quad \operatorname{Tr}(\boldsymbol{N}_{ij}\boldsymbol{U}) = 2\delta_{ij}, 1 \le i \le j \le r,\\
&\quad \boldsymbol{U} \in \mathcal{S}_+^{n+r}.
\end{aligned} \tag{18}$$

The dual problem to the SDR problem (SDRM) is given by

$$\begin{aligned}
\text{(DM)}\quad &\max_{\lambda_i, \boldsymbol{\Phi}} -\sum_{i\in\mathcal{I}\cup\mathcal{E}} \lambda_i\alpha_i - \operatorname{Tr}(\boldsymbol{\Phi})\\
&\text{s.t. } \boldsymbol{M}(f_0) + \sum_{i\in\mathcal{I}\cup\mathcal{E}} \lambda_i \boldsymbol{M}(f_i) + \begin{pmatrix} \boldsymbol{0}_{n\times n} & \boldsymbol{0}_{n\times r} \\ \boldsymbol{0}_{r\times n} & \boldsymbol{\Phi} \end{pmatrix} \succeq \boldsymbol{0},\\
&\quad \boldsymbol{\Phi} \in \mathcal{S}^r,\\
&\quad \lambda_i \ge 0, i \in \mathcal{I}.
\end{aligned} \tag{19}$$

The symmetric matrix $\boldsymbol{\Phi} = (\phi_{ij})_{i,j=1}^r$ contains the Lagrange multipliers associated with the equality constraints $\operatorname{Tr}(\boldsymbol{N}_{ij}\boldsymbol{U}) = 2\delta_{ij}$. Specifically, for every $1 \le i \le r$, $\frac{1}{2}\phi_{ii}$ is the multiplier corresponding to the constraint $\operatorname{Tr}(\boldsymbol{N}_{ii}\boldsymbol{U}) = 2$, and $\phi_{ij} (= \phi_{ji})$ is the multiplier associated with $\operatorname{Tr}(\boldsymbol{N}_{ij}\boldsymbol{U}) = 0$ for $1 \le i < j \le r$. By the conic duality theorem [4] it follows that if (DM) is strictly feasible and bounded above, then (SDRM) is solvable and $\operatorname{val}(\text{SDRM}) = \operatorname{val}(\text{DM})$. For that reason we seek to find a simple condition under which (DM) is strictly feasible. The following lemma establishes such a condition.

LEMMA 3.3. *Suppose that the following condition is satisfied:*

$$\exists \gamma_i \in \mathbb{R}, i \in \mathcal{I} \cup \mathcal{E}, \text{ for which } \gamma_i \geq 0, i \in \mathcal{I}, \text{ such that } \boldsymbol{A}_0 + \sum_{i \in \mathcal{I} \cup \mathcal{E}} \gamma_i \boldsymbol{A}_i \succ \boldsymbol{0}. \tag{20}$$

Then problem (DM) *is strictly feasible.*

Proof. Let $\gamma_i \in \mathbb{R}, i \in \mathcal{I} \cup \mathcal{E}$, be numbers satisfying (20), and let $\epsilon > 0$ be a small enough number for which $\boldsymbol{A}_0 + \sum_{i \in \mathcal{I} \cup \mathcal{E}} (\gamma_i + \epsilon) \boldsymbol{A}_i \succ \boldsymbol{0}$. Define $\tilde{\gamma}_i \equiv \gamma_i + \epsilon$. Evidently, $\tilde{\gamma}_i > 0$ for $i \in \mathcal{I}$. Now, for every symmetric $r \times r$ matrix $\boldsymbol{\Phi}$ we have

$$\begin{aligned} \boldsymbol{M}(f_0) + \sum_{i \in \mathcal{I} \cup \mathcal{E}} \tilde{\gamma}_i \boldsymbol{M}(f_i) + \begin{pmatrix} \boldsymbol{0}_{n \times n} & \boldsymbol{0}_{n \times r} \\ \boldsymbol{0}_{r \times n} & \boldsymbol{\Phi} \end{pmatrix} \\ = \begin{pmatrix} \boldsymbol{A}_0 + \sum \tilde{\gamma}_i \boldsymbol{A}_i & \boldsymbol{B}_0 + \sum \tilde{\gamma}_i \boldsymbol{B}_i \\ (\boldsymbol{B}_0 + \sum \tilde{\gamma}_i \boldsymbol{B}_i)^T & \frac{1}{r} (c_0 + \sum \tilde{\gamma}_i c_i) \boldsymbol{I}_r + \boldsymbol{\Phi} \end{pmatrix}, \end{aligned} \tag{21}$$

where all the summations are over $i \in \mathcal{I} \cup \mathcal{E}$. Since $\boldsymbol{A}_0 + \sum \tilde{\gamma}_i \boldsymbol{A}_i \succ \boldsymbol{0}$, then by the Schur complement, the matrix on the RHS of (21) is positive definite if and only if

$$\boldsymbol{\Phi} \succ \left(\boldsymbol{B}_0 + \sum \tilde{\gamma}_i \boldsymbol{B}_i \right)^T \left(\boldsymbol{A}_0 + \sum \tilde{\gamma}_i \boldsymbol{A}_i \right)^{-1} \left(\boldsymbol{B}_0 + \sum \tilde{\gamma}_i \boldsymbol{B}_i \right) - \frac{1}{r} \left(c_0 + \sum \tilde{\gamma}_i c_i \right) \boldsymbol{I}_r.$$

Let $\tilde{\boldsymbol{\Phi}} \in \mathcal{S}^r$ be an arbitrary matrix satisfying the latter LMI. Thus, for $\lambda_i = \tilde{\gamma}_i, i \in \mathcal{I} \cup \mathcal{E}$, and $\boldsymbol{\Phi} = \tilde{\boldsymbol{\Phi}}$ we have that all the inequalities in (19) (regular and generalized) are strictly satisfied. ∎

Remark 3.1. Conditions similar to (20) are very common in the analysis of QCQP problems; see, e.g., [5, 18, 12, 19, 16]. This condition is automatically satisfied when at least one of the constraints or the objective function is strictly convex (see also [19, Proposition 2.1]).

3.2. Tightness of the SDR of the QMP problem. In this section we will show that, under some mild conditions, QMP problems of order r with at most r constraints have a tight SDR, and that strong duality holds. To show this, we need to verify that problem (SDRM) possesses a solution with rank smaller than or equal to r. This prompts us to consider questions concerning the existence of low-rank solutions to SDP problems—a subject extensively studied by Pataki [14, 15] and Barvinok [2, 3]; see also [11] for related results concerning the convexity of the image of several homogenous QMs.

Let us consider a general-form SDP problem:

$$\begin{aligned} &\min \operatorname{Tr}(\boldsymbol{C}_0 \boldsymbol{U}) \\ &\text{s.t. } \operatorname{Tr}(\boldsymbol{C}_i \boldsymbol{U}) \leq \alpha_i, i \in \mathcal{I}_1, \\ &\quad \operatorname{Tr}(\boldsymbol{C}_j \boldsymbol{U}) = \alpha_j, j \in \mathcal{E}_1, \\ &\quad \boldsymbol{U} \in \mathcal{S}^n_+, \end{aligned} \tag{22}$$

where $\mathcal{I}_1$ and $\mathcal{E}_1$ are disjoint index sets, $\boldsymbol{C}_i \in \mathcal{S}^n, i \in \{0\} \cup \mathcal{I}_1 \cup \mathcal{E}_1$, and $\alpha_i \in \mathbb{R}, i \in \mathcal{I}_1 \cup \mathcal{E}_1$. Pataki showed [15] that if the number of constraints is smaller than an upper bound which is a certain quadratic function of r, then there exists a solution with rank no larger than r (see Theorem 3.4 below). The proof of this result is constructive and is based on a simple rank reduction procedure[4] for finding extreme points of convex

[4]The SDP considered in [15] consists only of inequality constraints. However, the same analysis establishes the validity of Theorem 3.4.

sets of the form $\mathcal{S}_+^n \cap \mathcal{A}$, where $\mathcal{A}$ is an affine space. For the sake of completeness, and since the rank reduction procedure is a subroutine of the algorithm for solving the QMP problem, we recall both the claim (Theorem 3.4 below) and the rank reduction procedure (see Algorithm RED in the appendix).

THEOREM 3.4 (see [15]). *Suppose that problem* (22) *is solvable and that* $|\mathcal{I}_1| + |\mathcal{E}_1| \leq \binom{r+2}{2}$*, where* r *is a positive integer. Then problem* (22) *has a solution* $\boldsymbol{X}^*$ *for which* $\mathrm{rank}(\boldsymbol{X}^*) \leq r$.

Proof. Let $\boldsymbol{X}_0^*$ be an optimal solution of problem (22). Apply Algorithm RED (see the appendix) with input $\boldsymbol{X}_0^*$ and obtain an optimal solution $\boldsymbol{X}^*$ with $\mathrm{rank}(\boldsymbol{X}^*) \leq r$. ∎

Equipped with the latter result, we are now able to show that QMP problems of order r with at most r constraints possess a tight SDR under some mild conditions.

THEOREM 3.5 (tight SDR for the QMP problem). *If problem* (SDRM) *is solvable and* $|\mathcal{I}| + |\mathcal{E}| \leq r$*, then problem* (QMP) *is solvable and* $\mathrm{val(SDRM)} = \mathrm{val(QMP)}$.

Proof. It is sufficient to show that problem (SDRM) has a solution with rank smaller than or equal to r. The number of constraints in (SDRM) is equal to $|\mathcal{I}|+|\mathcal{E}|+\binom{r+1}{2}$, where the last term stands for the number of pairs (i,j) for which $1 \leq i \leq j \leq r$. Thus, using $|\mathcal{I}| + |\mathcal{E}| \leq r$, we conclude that the number of constraints in (SDRM) is bounded above by

$$r + \binom{r+1}{2} = \binom{r+2}{2} - 1.$$

Invoking Theorem 3.4, the result follows. ∎

As a conclusion from the conic duality theorem [4] we can now deduce the following corollary that guarantees tightness of the SDR and strong duality under the conditions that the QMP problem (6) is feasible and that condition (20) is valid.

COROLLARY 3.6 (strong duality for QMP problems). *Consider the QMP problem* (6) *with* $|\mathcal{I}| + |\mathcal{E}| \leq r$*, its semidefinite relaxation* (SDRM) *(problem* (18)*) and its dual* (DM) *(problem* (19)*). Suppose that condition* (20) *holds true and that the QMP problem is feasible. Then problems* (QMP) *and* (SDRM) *are solvable and* $\mathrm{val(QMP)} = \mathrm{val(SDRM)} = \mathrm{val(DM)}$.

Proof. By Lemma 3.3, the validity of condition (20) implies that the dual problem (DM) is strictly feasible. Moreover, since the primal SDP problem (SDRM) is feasible, it follows that the dual problem (DM) is bounded above. Thus, by the conic duality theorem [4], we conclude that problem (SDRM) is solvable and that $\mathrm{val(SDRM)} = \mathrm{val(DM)}$. Since problem (SDRM) is solvable we conclude, by Theorem 3.5, that $\mathrm{val(QMP)} = \mathrm{val(SDRM)}$. ∎

Remark 3.2. In the special case $r = 1$, Corollary 3.6 recovers the well-known strong duality/tightness of SDR results for QCQPs with a single quadratic constraint (see, e.g., [12, 5, 18, 16]).

It is interesting to note that we can also describe an algorithm for extracting the solution of a QMP problem (satisfying the condition in Corollary 3.6) from its SDR, which is based on the rank reduction algorithm of [15], as follows.

ALGORITHM SOL-QMP.

Step 1. Solve the SDP problem (SDRM) and obtain an optimal solution $\boldsymbol{U}^* \in \mathcal{S}_+^{n+r}$.

Step 2. Invoke Algorithm RED (see the appendix) with input $\boldsymbol{U}^*$, and produce an optimal solution $\boldsymbol{U}_1^* \in \mathcal{S}_+^{n+r}$ for which $\mathrm{rank}(\boldsymbol{U}_1^*) \leq r$.

Step 3. Calculate a decomposition: $\boldsymbol{U}_1^* = \boldsymbol{W}\boldsymbol{W}^T$, where $\boldsymbol{W} \in \mathbb{R}^{(n+r)\times r}$.

Step 4. Let $\boldsymbol{W} = (\mathbf{Y}; \boldsymbol{Z})$, where $\mathbf{Y} \in \mathbb{R}^{n\times r}$ and $\boldsymbol{Z} \in \mathbb{R}^{r\times r}$. Return an optimal solution $\boldsymbol{X}^* = \mathbf{Y}\boldsymbol{Z}^T$ to the QMP problem.

4. The vectorized semidefinite relaxation and dual of the QMP problem. In the previous section we considered a semidefinite relaxation that was based on a homogenization procedure specifically designed for QM functions. In this section we examine an alternative (and natural) approach in which we begin by transforming the problem into a "standard" QCQP and then use the usual relaxation technique. This approach produces the *vectorized SDR* and *vectorized dual* problems. We will prove that the two constructions are equivalent in some sense. In establishing this result we rely on the tight SDR result of section 3 and a result on two LMI representations of the property of nonnegativity of a QM function over $\mathbb{R}^{n\times r}$.

Our alternative SDR is constructed by following two steps.

Step 1. Transform the QMP problem (6) into the vectorized QMP problem (7).

Step 2. Formulate the corresponding SDR of the homogenized problem (7):

$$
\begin{aligned}
(SDRV)\quad & \min \operatorname{Tr}(\boldsymbol{M}(f_0^V)\boldsymbol{Z}) \\
& \text{s.t. } \operatorname{Tr}(\boldsymbol{M}(f_i^V))\boldsymbol{Z}) \le \alpha_i, i \in \mathcal{I}, \\
& \quad \operatorname{Tr}(\boldsymbol{M}(f_j^V)\boldsymbol{Z}) = \alpha_j, j \in \mathcal{E}, \\
& \quad Z_{nr+1,nr+1} = 1, \\
& \quad \boldsymbol{Z} \in \mathcal{S}_+^{nr+1}
\end{aligned} \tag{23}
$$

(recall that, since f_i^V is a QM function of order one, $\boldsymbol{M}(f_i^V) \equiv \left(\begin{smallmatrix} \boldsymbol{I}_r \otimes \boldsymbol{A}_i & \operatorname{vec}(\boldsymbol{B}_i) \\ \operatorname{vec}(\boldsymbol{B}_i)^T & c_i \end{smallmatrix}\right)$.

Problem (SDRV) is an SDP problem, and its dual is given by

$$
\begin{aligned}
\text{(DV)}\qquad & \max -\sum_{i\in\mathcal{I}\cup\mathcal{E}} \lambda_i\alpha_i - t \\
& \text{s.t. } \boldsymbol{M}(f_0^V) + \sum_{i\in\mathcal{I}\cup E} \lambda_i \boldsymbol{M}(f_i^V) + t\begin{pmatrix} \boldsymbol{0}_{nr,nr} & \boldsymbol{0}_{nr,1} \\ \boldsymbol{0}_{1,nr} & 1 \end{pmatrix} \succeq \boldsymbol{0}, \\
& \quad \lambda_i \ge 0, i \in \mathcal{I}.
\end{aligned} \tag{24}
$$

It can be shown that problem (DV) is in fact a Lagrangian dual of the QMP problem (6), and therefore the SDR (SDRV) can be interpreted as a bidual (i.e., dual of the dual) of the primal QMP problem. Problems (SDRV) and (DV) are called the *vectorized* semidefinite relaxation and dual of the QMP problem (respectively).

The pair of problems (SDRM)/(SDRV) and (DM)/(DV) seem quite different both with respect to the number of variables and the sizes of the related matrices. However, we will show in what follows (cf. Theorem 4.3) that these pairs of problems are equivalent in some sense.

Lemma 4.2 below presents two different LMI characterizations of the nonnegativity of a QM function over the entire space. This lemma is a key ingredient in proving the equivalence between the different dual/SDR problems. The proof of Lemma 4.2 relies on the following well-known result.

LEMMA 4.1 (see [4, p. 163]). *A quadratic inequality with a (symmetric) $n \times n$ matrix $\boldsymbol{A}$,*

$$\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} + 2\boldsymbol{b}^T\boldsymbol{x} + c \ge 0,$$

is valid for all $\boldsymbol{x} \in \mathbb{R}^n$ if and only if

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{b} \\ \boldsymbol{b}^T & c \end{pmatrix} \succeq \boldsymbol{0}.$$

LEMMA 4.2. *Let f be a QM function given in* (3). *Then the following three statements are equivalent:*

(i) $f(\boldsymbol{X}) \geq 0$ *for every* $\boldsymbol{X} \in \mathbb{R}^{n\times r}$.

(ii) *There exists* $\boldsymbol{\Phi} \in \mathcal{S}^r$ *for which* $Tr(\boldsymbol{\Phi}) \leq 0$ *such that*

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}^T & \frac{c}{r}\boldsymbol{I}_r + \boldsymbol{\Phi} \end{pmatrix} \succeq \boldsymbol{0}.$$

(iii)

$$\begin{pmatrix} \boldsymbol{I}_r \otimes \boldsymbol{A} & \mathrm{vec}(\boldsymbol{B}) \\ \mathrm{vec}(\boldsymbol{B})^T & c \end{pmatrix} \succeq \boldsymbol{0}.$$

Proof. (i⇔iii) By (4), the first statement is equivalent to the statement

$$f^V(\boldsymbol{z}) \geq 0 \text{ for every } \boldsymbol{z} \in \mathbb{R}^{nr},$$

which, by Lemma 4.1, is the same as the third statement.

(i⇔ii) We begin by showing the following identity between subsets of $\mathbb{R}$:

$$F = W, \tag{25}$$

where (recall that $[U]_r$ denotes the southeast $r \times r$ submatrix of $\boldsymbol{U}$)

$$\begin{aligned} F &= \{f(\boldsymbol{X}) : \boldsymbol{X} \in \mathbb{R}^{n\times r}\}, \\ W &= \{\mathrm{Tr}(\boldsymbol{M}(f)\boldsymbol{U}) : \boldsymbol{U} \in \mathcal{S}_+^{n+r}, [\boldsymbol{U}]_r = \boldsymbol{I}_r\}. \end{aligned}$$

The inclusion $F \subseteq W$ is clear. We will show that the reverse inclusion ($W \subseteq F$) holds true. Let $\alpha \in W$, and consider the QMP problem

$$\begin{aligned} &\min 0 \\ &\text{s.t. } f(\boldsymbol{X}) = \alpha, \\ &\quad \boldsymbol{X} \in \mathbb{R}^{n\times r}. \end{aligned} \tag{26}$$

Note that this is exactly the QMP problem (6) with $r = 1$, $\mathcal{I} = \emptyset$, $\mathcal{E} = \{1\}$, $\alpha_1 = \alpha, f_0 \equiv 0$, and $f_1 = f$. The corresponding SDR of the QMP problem (26) is given by

$$\begin{aligned} &\min 0 \\ &\text{s.t. } \mathrm{Tr}(\boldsymbol{M}(f)\boldsymbol{U}) = \alpha, \\ &\quad \boldsymbol{U} \in \mathcal{S}_+^{n+r},\ [\boldsymbol{U}]_r = \boldsymbol{I}_r. \end{aligned} \tag{27}$$

Since $\alpha \in W$ it follows that problem (27) is solvable (recall that the objective function is identically zero, and hence "solvability" is the same as "feasibility"). Invoking Theorem 3.5, we conclude that problem (26) is also feasible. Hence, $\alpha \in F$. The identity $F = W$ implies that statement (i) is the same as

$$\min\{\mathrm{Tr}(\boldsymbol{M}(f)\boldsymbol{U}) : \boldsymbol{U} \in \mathcal{S}_+^{n+r}, [\boldsymbol{U}]_r = \boldsymbol{I}_r\} \geq 0. \tag{28}$$

The latter SDP problem is strictly feasible ($\boldsymbol{U} = \boldsymbol{I}_{n+r} \succ \boldsymbol{0}$ is feasible) and bounded below (by zero) and thus, by the conic duality theorem, we conclude that the dual problem, given in this case by

$$\max_{\boldsymbol{\Phi}\in\mathcal{S}^r} \left\{ -\mathrm{Tr}(\boldsymbol{\Phi}) : \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}^T & \frac{c}{r}\boldsymbol{I}_r + \boldsymbol{\Phi} \end{pmatrix} \succeq \boldsymbol{0} \right\},$$

is solvable and has value equal to the value of the primal problem. Therefore, statement (28) is equivalent to the existence of a symmetric $r \times r$ matrix $\boldsymbol{\Phi}$ for which

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}^T & \frac{c}{r}\boldsymbol{I}_r + \boldsymbol{\Phi} \end{pmatrix} \succeq \mathbf{0}$$

and $\mathrm{Tr}(\boldsymbol{\Phi}) \leq 0$. □

We are now ready to prove the main result of this section, namely, that the values of the two dual problems (DM) and (DV) and the two SDR problems (SDRM) and (SDRV) are all equal to each other under some mild conditions.

THEOREM 4.3. *Consider the SDR*s (SDRM) *and* (SDRV) *(problems* (18) *and* (23)*) and the dual problems* (DM) *and* (DV) *(problems* (19) *and* (24)*) of the QMP problem* (6)*. Suppose that condition* (20) *is satisfied and that* (QMP) *is feasible. Then* (SDRM) *and* (SDRV) *are solvable and*

$$\mathrm{val(DM)} = \mathrm{val(DV)} = \mathrm{val(SDRM)} = \mathrm{val(SDRV)}.$$

Furthermore, if $\{\lambda_i\}_{i\in\mathcal{I}\cup\mathcal{E}}$ *and* $\boldsymbol{\Phi}$ *is an optimal solution of* (DM)*, then an optimal solution to* (DV) *is given by* $\{\lambda_i\}_{i\in\mathcal{I}\cup\mathcal{E}}, t$*, where* $t = \mathrm{Tr}(\boldsymbol{\Phi})$.

Proof. Since condition (20) is assumed to hold true then, by Lemma 3.3, the dual problem (DM) is strictly feasible, and an argument similar to the one used in the proof of Lemma 3.3 shows that (DV) is also strictly feasible. Thus, by the conic duality Theorem [4], both problems (SDRM) and (SDRV) are solvable, and we have the equality $\mathrm{val(DM)} = \mathrm{val(SDRM)}$ as well as $\mathrm{val(DV)} = \mathrm{val(SDRV)}$. We are left with the task of proving that $\mathrm{val(DM)} = \mathrm{val(DV)}$. Consider the LMI constraint in problem (DV), which can explicitly be written as follows:

$$\begin{pmatrix} \boldsymbol{I}_r \otimes (\boldsymbol{A}_0 + \sum \lambda_i \boldsymbol{A}_i) & \mathrm{vec}\left(\boldsymbol{B}_0 + \sum \lambda_i \boldsymbol{B}_i\right) \\ \mathrm{vec}\left(\boldsymbol{B}_0 + \sum \lambda_i \boldsymbol{B}_i\right)^T & c_0 + \sum \lambda_i c_i + t \end{pmatrix} \succeq \mathbf{0}, \tag{29}$$

where the summations are over $i \in \mathcal{I}\cup\mathcal{E}$. By the equivalence of the second and third part of Lemma 4.2 we have that the above LMI holds true if and only if there exists $\boldsymbol{Z} \in \mathcal{S}^r$ such that

$$\begin{pmatrix} \boldsymbol{A}_0 + \sum \lambda_i \boldsymbol{A}_i & \boldsymbol{B}_0 + \sum \lambda_i \boldsymbol{B}_i \\ (\boldsymbol{B}_0 + \sum \lambda_i \boldsymbol{B}_i)^T & \frac{1}{r}(c_0 + \sum \lambda_i c_i + t)\boldsymbol{I}_r + \boldsymbol{Z} \end{pmatrix} \succeq \mathbf{0},$$

and $\mathrm{Tr}(\boldsymbol{Z}) \leq 0$. Making the change of variables $\boldsymbol{\Phi} = \boldsymbol{Z} + \frac{t}{r}\boldsymbol{I}_r$, we deduce that the LMI (29) is equivalent to the existence of a matrix $\boldsymbol{\Phi} \in \mathcal{S}^r$ such that

$$\begin{pmatrix} \boldsymbol{A}_0 + \sum \lambda_i \boldsymbol{A}_i & \boldsymbol{B}_0 + \sum \lambda_i \boldsymbol{B}_i \\ (\boldsymbol{B}_0 + \sum \lambda_i \boldsymbol{B}_i)^T & \frac{1}{r}(c_0 + \sum \lambda_i c_i)\boldsymbol{I}_r + \boldsymbol{\Phi} \end{pmatrix} \succeq \mathbf{0}, \tag{30}$$

and

$$\mathrm{Tr}(\boldsymbol{\Phi}) \leq t. \tag{31}$$

Replacing the LMI in problem (24) with the LMIs (30) and (31), problem (DV) is transformed into

$$\begin{aligned} \max_{\lambda_i, \boldsymbol{\Phi}, t} \; & -\sum_{i\in\mathcal{I}\cup\mathcal{E}} \lambda_i \alpha_i - t \\ \text{s.t.} \; & \begin{pmatrix} \boldsymbol{A}_0 + \sum \lambda_i \boldsymbol{A}_i & \boldsymbol{B}_0 + \sum \lambda_i \boldsymbol{B}_i \\ (\boldsymbol{B}_0 + \sum \lambda_i \boldsymbol{B}_i)^T & \frac{1}{r}(c_0 + \sum \lambda_i c_i)\boldsymbol{I}_r + \boldsymbol{\Phi} \end{pmatrix} \succeq \mathbf{0}, \\ & \lambda_i \geq 0, i \in \mathcal{I}, \\ & \mathrm{Tr}(\boldsymbol{\Phi}) \leq t. \end{aligned}$$

It is clear that any optimal solution of the last problem satisfies $t = \mathrm{Tr}(\boldsymbol{\Phi})$, and thus the problem is the same as

$$\begin{array}{l} \max\limits_{\lambda_i,\Phi} -\sum\limits_{i\in\mathcal{I}\cup\mathcal{E}} \lambda_i\alpha_i - \mathrm{Tr}(\boldsymbol{\Phi}) \\ \text{s.t.} \begin{pmatrix} \boldsymbol{A}_0 + \sum \lambda_i \boldsymbol{A}_i & \boldsymbol{B}_0 + \sum \lambda_i \boldsymbol{B}_i \\ (\boldsymbol{B}_0 + \sum \lambda_i \boldsymbol{B}_i)^T & \frac{1}{r}(c_0 + \sum \lambda_i c_i)\boldsymbol{I}_r + \boldsymbol{\Phi} \end{pmatrix} \succeq \mathbf{0}, \\ \quad \lambda_i \geq 0, i \in \mathcal{I}, \end{array}$$

which is identical to problem (DM). ∎

Combining the latter result with the strong duality result, Corollary 3.6, the following corollary immediately follows.

COROLLARY 4.4. *Consider the QMP problem* (6) *with* $|\mathcal{I}|+|\mathcal{E}| \leq r$*, its vectorized semidefinite relaxation* (SDRV) *(problem* (23)*), and its vectorized dual* (DV) *(problem* (24)*). Suppose that condition* (20) *holds true and that the QMP problem is feasible. Then problems* (QMP) *and* (SDRV) *are solvable and* $\mathrm{val}(\mathrm{QMP}) = \mathrm{val}(\mathrm{SDRV}) = \mathrm{val}(\mathrm{DV})$.

5. An application to robust least squares. We continue the example from section 2.2. Suppose that the uncertainty set $\mathcal{U}$ associated with the matrix $\boldsymbol{A}$ is given by multiple norm constraints:

$$\mathcal{U} = \{\Delta \in \mathbb{R}^{n\times r} : \|\mathbf{L}_i\boldsymbol{\Delta}\|^2 \leq \rho_i, i = 1,\ldots,m\}, \tag{32}$$

where $\mathbf{L}_i \in \mathbb{R}^{k_i\times n}$ for some positive integers $k_1,\ldots,k_m$ and $\rho_i > 0, i = 1,\ldots,m$. The above form of the uncertainty set is more general then the standard single-constraint form, and it can thus be used to describe more complicated scenarios of uncertainties. For example, by setting $k_i = n, m = n$, and $\mathbf{L}_i = \boldsymbol{E}_{ii}^n$, we model the situation in which the uncertainty associated with each column of the matrix $\boldsymbol{A}$ has a separate norm constraint.

Assume that there exist nonnegative numbers $\gamma_1,\ldots,\gamma_m$ such that

$$\sum_{i=1}^m \gamma_i \mathbf{L}_i^T\mathbf{L}_i \succ \mathbf{0}.$$

If $m \leq r$, then the conditions of Corollary 4.4 are satisfied, and as a consequence the inner maximization problem (9) is equal to the value of the dual problem given by

$$\begin{array}{l} \min\limits_{t,\lambda_i} \sum\limits_{i=1}^m \lambda_i\rho_i + t \\ \text{s.t.} \begin{pmatrix} \boldsymbol{I}_r \otimes (-\mathbf{Q} + \sum_{i=1}^m \lambda_i \mathbf{L}_i^T\mathbf{L}_i) & -\mathrm{vec}(\mathbf{F}) \\ -\mathrm{vec}(\mathbf{F})^T & -c + t \end{pmatrix} \succeq \mathbf{0}, \\ \quad \lambda_i \geq 0, i = 1,2,\ldots,m. \end{array}$$

Here we considered the equivalent vectorized dual because it is not clear how to derive an SDP formulation from the nonvectorized dual. Now, using the identities (see [10])

$$\begin{aligned} \boldsymbol{I}_r \otimes \mathbf{Q} &\overset{(10)}{=} \boldsymbol{I}_r \otimes \boldsymbol{x}\boldsymbol{x}^T = (\boldsymbol{I}_r \otimes \boldsymbol{x})(\boldsymbol{I}_r \otimes \boldsymbol{x})^T, \\ \mathrm{vec}(\mathbf{F}) &\overset{(10)}{=} \mathrm{vec}(\boldsymbol{x}(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})^T) = (\boldsymbol{I}_r \otimes \boldsymbol{x})(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}), \end{aligned}$$

the dual problem is transformed into

$$\begin{aligned}
&\min_{t,\lambda_i} \sum_{i=1}^{m} \lambda_i \rho_i + t \\
&\text{s.t.} \begin{pmatrix} -(\boldsymbol{I}_r \otimes \boldsymbol{x})(\boldsymbol{I}_r \otimes \boldsymbol{x})^T + \sum_{i=1}^{m} \lambda_i (\boldsymbol{I}_r \otimes (\mathbf{L}_i^T \mathbf{L}_i)) & -(\boldsymbol{I}_r \otimes \boldsymbol{x})(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) \\ -(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})^T (\boldsymbol{I}_r \otimes \boldsymbol{x})^T & -\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2 + t \end{pmatrix} \succeq \mathbf{0}, \\
&\quad \lambda_i \geq 0, i = 1, 2, \ldots, m,
\end{aligned}$$

which, by the Schur complement can be written as

$$\begin{aligned}
&\min_{t,\lambda_i} \sum_{i=1}^{m} \lambda_i \rho_i + t \\
&\text{s.t.} \begin{pmatrix} \boldsymbol{I}_r & (\boldsymbol{I}_r \otimes \boldsymbol{x})^T & \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} \\ \boldsymbol{I}_r \otimes \boldsymbol{x} & \sum_{i=1}^{m} \lambda_i (\boldsymbol{I}_r \otimes (\mathbf{L}_i^T \mathbf{L}_i)) & \mathbf{0} \\ (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})^T & \mathbf{0} & t \end{pmatrix} \succeq \mathbf{0}, \\
&\quad \lambda_i \geq 0, i = 1, 2, \ldots, m.
\end{aligned}$$

Finally, we arrive at the following SDP formulation of the RLS problem (8):

$$\begin{aligned}
&\min_{t,\lambda_i,x} \sum_{i=1}^{m} \lambda_i \rho_i + t \\
&\text{s.t.} \begin{pmatrix} \boldsymbol{I}_r & (\boldsymbol{I}_r \otimes \boldsymbol{x})^T & \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} \\ \boldsymbol{I}_r \otimes \boldsymbol{x} & \sum_{i=1}^{m} \lambda_i (\boldsymbol{I}_r \otimes (\mathbf{L}_i^T \mathbf{L}_i)) & \mathbf{0} \\ (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})^T & \mathbf{0} & t \end{pmatrix} \succeq \mathbf{0}, \\
&\quad \lambda_i \geq 0, i = 1, 2, \ldots, m.
\end{aligned}$$

Appendix. A rank reduction algorithm for solvable semidefinite problems. We review here the rank reduction algorithm of [15] for solving SDP problems of the form (22).[5] The underlying assumption that guarantees the validity of the process is that problem (22) is solvable and that $|\mathcal{I}_1| + |\mathcal{E}_1| \leq \binom{r+2}{2} - 1$.

ALGORITHM RED.

Input: $\boldsymbol{X}_0$, an optimal solution to problem (22).

Output: An optimal solution $\boldsymbol{X}^*$ to problem (22) satisfying $\operatorname{rank}(\boldsymbol{X}^*) \leq r$.

1. **If** $\operatorname{rank}(\boldsymbol{X}_0) \leq r$, then go to step 3. **Else** go to step 2.
2. **While** $\operatorname{rank}(\boldsymbol{X}_0) > r$, **repeat** steps (a)–(e):
   (a) **Set** $d \leftarrow \operatorname{rank}(\boldsymbol{X}_0)$.
   (b) Compute a decomposition of $\boldsymbol{X}_0$: $\boldsymbol{X}_0 = \boldsymbol{U}\boldsymbol{U}^T$, where $\boldsymbol{U} \in \mathbb{R}^{n \times d}$.
   (c) Find a nontrivial solution[6] $\boldsymbol{Z}_0$ for the set of homogenous linear equations in the $d \times d$ *symmetric* variables matrix $\boldsymbol{Z}$ ($\boldsymbol{Z} = \boldsymbol{Z}^T$):
   $$\operatorname{Tr}(\boldsymbol{U}^T \boldsymbol{C}_i \boldsymbol{U} \boldsymbol{Z}) = 0, \quad i \in \mathcal{I}_1 \cup \mathcal{E}_1.$$
   (d) **If** $\boldsymbol{Z}_0 \succeq \mathbf{0}$, then **set** $\boldsymbol{W} \leftarrow -\boldsymbol{Z}_0$. **Else set** $\boldsymbol{W} \leftarrow \boldsymbol{Z}_0$.
   (e) **Set** $\boldsymbol{X}_0 \leftarrow \boldsymbol{U}(\boldsymbol{I} + \beta \boldsymbol{W})\boldsymbol{U}^T$, where $\beta = -1/\lambda_{\min}(\boldsymbol{W})$.
3. **Set** $\boldsymbol{X}^* \leftarrow \boldsymbol{X}_0$ and **STOP**.

[5] Note that in [15], the SDP problem contains only inequality constraints. However, it is immediately seen that exactly the same rank reduction algorithm also works here.

[6] Using the relations $|\mathcal{I}_1| + |\mathcal{E}_1| \leq \binom{r+2}{2} - 1, d > r$, it is easy to see that the homogenous system has more variables than equations and, as a result, has a nonzero solution.

Note that the algorithm does not make use of the matrix $\boldsymbol{C}_0$ corresponding to the objective function in (22). Indeed, it can be shown that since the input to the algorithm is an *optimal* solution of the SDP problem (22), then the value $\mathrm{Tr}(\boldsymbol{C}_0\boldsymbol{X}_0)$ remains constant throughout the process.

REFERENCES

[1] K. Anstreicher and H. Wolkowicz, *On Lagrangian relaxation of quadratic matrix constraints*, SIAM J. Matrix Anal. Appl., 22 (2000), pp. 41–55.

[2] A. Barvinok, *A remark on the rank of positive semidefinite matrices subject to affine constraints*, Discrete Comput. Geom., 25 (2001), pp. 23–31.

[3] A. Barvinok, *Problems of distance geometry and convex properties of quadratic maps*, Discrete Comput. Geom., 13 (1995), pp. 189–202.

[4] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization, Analysis, Algorithms, and Engineering Applications*, MPS-SIAM Ser. Optim. 2, SIAM, Philadelphia, 2001.

[5] A. Ben-Tal and M. Teboulle, *Hidden convexity in some nonconvex quadratically constrained quadratic programming*, Math. Programming, 72 (1996), pp. 51–63.

[6] S. Chandrasekaran, G. H. Golub, M. Gu, and A. H. Sayed, *Parameter estimation in the presence of bounded data uncertainties*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 235–252.

[7] L. Eldén and H. Park, *A Procrustes problem on the Stiefel manifold*, Numer. Math., 82 (1999), pp. 599–619.

[8] C. Fortin and H. Wolkowicz, *The trust region subproblem and semidefinite programming*, Optim. Methods Softw., 19 (2004), pp. 41–67.

[9] 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.

[10] A. Graham, *Kronecker Products and Matrix Calculus: With Applications*, Ellis Horwood Ser. Math. Appl., Ellis Horwood, Chichester, U.K., 1981.

[11] J. B. Hiriart-Urruty and M. Torki, *Permanently going back and forth between the "quadratic world" and the "convexity world" in optimization*, Appl. Math. Optim., 45 (2002), pp. 169–184.

[12] J. J. Moré, *Generalization of the trust region problem*, Optim. Methods Softw., 2 (1993), pp. 189–209.

[13] J. J. Moré and D. C. Sorensen, *Computing a trust region step*, SIAM J. Sci. Statist. Comput., 4 (1983), pp. 553–572.

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

[15] G. Pataki, *The geometry of semidefinite programming*, in Handbook of Semidefinite Programming, Internat. Ser. Oper. Res. Management Sci. 27, Kluwer Academic Publishers, Boston, 2000, pp. 29–65.

[16] B. T. Polyak, *Convexity of quadratic transformations and its use in control and optimization*, J. Optim. Theory Appl., 99 (1998), pp. 553–583.

[17] P. H. Schönemann, *A generalized solution of the orthogonal Procrustes problem*, Psychometrika, 31 (1966), pp. 1–10.

[18] R. J. Stern and H. Wolkowicz, *Indefinite trust region subproblems and nonsymmetric eigenvalue perturbations*, SIAM J. Optim., 5 (1995), pp. 286–313.

[19] Y. Ye and S. Zhang, *New results on quadratic minimization*, SIAM J. Optim., 14 (2003), pp. 245–267.

[20] Y. Yuan, *On a subproblem of trust region algorithms for constrained optimization*, Math. Programming, 47 (1990), pp. 53–63.

SIAM J. OPTIM.
Vol. 17, No. 4, pp. 1239–1252
© 2007 Society for Industrial and Applied Mathematics

MAXIMAL MONOTONICITY FOR THE PRECOMPOSITION WITH A LINEAR OPERATOR*

RADU IOAN BOŢ†, SORIN-MIHAI GRAD†, AND GERT WANKA†

Abstract. We give the weakest constraint qualification known to us that ensures the maximal monotonicity of the operator $A^* \circ T \circ A$ when A is a linear continuous mapping between two reflexive Banach spaces and T is a maximal monotone operator. As a special case we get the weakest constraint qualification that guarantees the maximal monotonicity of the sum of two maximal monotone operators on a reflexive Banach space. Then we give a weak constraint qualification assuring the Brézis–Haraux-type approximation of the range of the subdifferential of the precomposition to A of a proper convex lower semicontinuous function in nonreflexive Banach spaces, extending and correcting in a special case an older result due to Riahi.

Key words. maximal monotone operator, Fitzpatrick function, subdifferential, Brézis–Haraux-type approximation

AMS subject classifications. 47H05, 42A50, 90C25

DOI. 10.1137/050641491

1. Introduction.

The literature on maximal monotone operators is quite rich, especially in recent years when their connections to convex analysis, underlined with the help of some functions (see [12], [15], [16], [21], [23], [24]), were more and more intensively studied and used. One of the most interesting problems which involves both maximal monotone operators and convex analysis is the one of finding sufficient conditions that ensure the maximal monotonicity of the operator $A^* \circ T \circ A$ when A is a linear continuous mapping between two reflexive Banach spaces and T is a maximal monotone operator. From the papers dealing with this problem we refer here to [1], [6], [13], [15], and [24], the latter unifying the results concerning this issue from the others and giving four equivalent constraint qualifications, the weakest in the literature known to us. Finding a weaker sufficient condition under which the sum of two maximal monotone operators on reflexive Banach spaces is maximal monotone has been an older challenge for many mathematicians, the problem having existed for more than four decades. From Browder [3] and Rockafellar [20] in the 1960's to the recent papers of Simons and Zălinescu [23], Borwein [1], and Zălinescu [24], the conditions imposed on two maximal monotone operators in order to ensure the maximal monotonicity of their sum became weaker and weaker, the latter paper containing the weakest constraint qualification known to us so far that guarantees the mentioned result. We mention here also Simons' book [21], where many sufficient conditions for the mentioned problem are recalled, compared, and unified. This book and the lecture notes [17] due to Phelps are excellent references for anyone interested in maximal monotone operators. Within this paper we give a constraint qualification that guarantees the maximal monotonicity of $A^* \circ T \circ A$ and is satisfied also by some A and T that violate the other sufficient conditions known to us, already mentioned. This condition uses the so-called Fitzpatrick function and has been developed from the one

*Received by the editors September 29, 2005; accepted for publication (in revised form) August 6, 2006; published electronically January 12, 2007.
http://www.siam.org/journals/siopt/17-4/64149.html

†Faculty of Mathematics, Chemnitz University of Technology, D-09107 Chemnitz, Germany (radu.bot@mathematik.tu-chemnitz.de, sorin-mihai.grad@mathematik.tu-chemnitz.de, gert.wanka@mathematik.tu-chemnitz.de).

introduced by two of the authors in [2] for the Fenchel duality. For a special choice of A and T we obtain the weakest sufficient condition known to us that guarantees the maximal monotonicity of the sum of two maximal monotone operators.

Another result in maximal monotonicity for whose fulfillment we give a weaker sufficient condition is the one concerning the so-called Brézis–Haraux-type approximation of the range of $\partial(f \circ A)$, where f is a proper convex lower semicontinuous function defined on the image space of A with extended real values. Here we work in nonreflexive Banach spaces. Something similar has been done by Pennanen in [14] when the image space of A is reflexive. As a special case we recover and correct a result due to Riahi (see [18]) concerning the Brézis–Haraux-type approximation (see [21]) of the range of the sum of the subdifferentials of two lower semicontinuous functions by the sum of the ranges of the two subdifferentials, for which we give a weaker constraint qualification than in the original paper.

The paper is structured as follows. The next section contains necessary preliminaries, notions, and results used later; then we deal with the maximal monotonicity of $A^* \circ T \circ A$ and of the sum of two maximal monotone operators. Section 4 deals with the mentioned Brézis–Haraux-type approximations and is followed by a short summary of the results proved within the paper.

2. Preliminaries. In this section we introduce and recall some notions and results in order to make the paper self-contained. Even if the main results in the paper are given in (reflexive) Banach spaces, some of the preliminaries are valid also for more general spaces, and thus we begin by considering a nontrivial locally convex topological space X and its continuous dual space X^*, endowed with the weak* topology $w(X^*, X)$. By $\langle x^*, x\rangle$ we denote the value of the linear continuous functional $x^* \in X^*$ at $x \in X$. For a subset C of X we have the *indicator* function $\delta_C : X \to \overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$, defined by

$$\delta_C(x) = \begin{cases} 0 & \text{if } x \in C, \\ +\infty & \text{otherwise,} \end{cases}$$

and we denote by $\operatorname{int}(C)$ and $\operatorname{cl}(C)$ its *interior*, respectively, its *closure* in the corresponding topology. For C we define also the *linear hull* $\operatorname{lin}(C)$ as the intersection of all the linear subspaces of X containing C and the *affine hull* $\operatorname{aff}(C)$ which is the intersection of all the affine subsets of X containing C. For $C \subseteq X$ convex we denote the *intrinsic relative algebraic interior* of C by ${}^{ic}C$. One has $x \in {}^{ic} C$ if and only if $\cup_{\lambda>0} \lambda(C - x)$ is a closed linear subspace of X. We consider also the *first projection*, i.e., the function $\operatorname{pr}_1 : X \times Y \to X$, for Y some nontrivial locally convex space, defined as follows: $\operatorname{pr}_1(x, y) = x$ for any $(x, y) \in X \times Y$.

Given a function $f : X \to \overline{\mathbb{R}}$, we denote its *domain* by $\operatorname{dom}(f) = \{x \in X : f(x) < +\infty\}$ and its *epigraph* by $\operatorname{epi}(f) = \{(x, r) \in X \times \mathbb{R} : f(x) \le r\}$. For $x \in X$ such that $f(x) \in \mathbb{R}$ we define the *subdifferential* of f at x by $\partial f(x) = \{x^* \in X^* : f(y) - f(x) \ge \langle x^*, y - x\rangle\}$. We call f *proper* if $f(x) > -\infty\ \forall x \in X$ and $\operatorname{dom}(f) \ne \emptyset$. The *conjugate* of the function f is $f^* : X^* \to \overline{\mathbb{R}}$ introduced by

$$f^*(y) = \sup\big\{\langle y, x\rangle - f(x) : x \in X\big\}.$$

Between a function and its conjugate there is *Young's inequality*

$$f^*(y) + f(x) \ge \langle y, x\rangle \quad \forall x \in X\ y \in X^*.$$

Consider also the *identity* function on X defined as follows: $\operatorname{id}_X : X \to X$, $\operatorname{id}_X(x) = x$ $\forall x \in X$. When $f : X \to \overline{\mathbb{R}}$ and $g : Y \to \overline{\mathbb{R}}$, we define the function $f \times g : X \times Y \to \overline{\mathbb{R}} \times \overline{\mathbb{R}}$

through $f \times g(x,y) = (f(x), g(y))$, $(x,y) \in X \times Y$. When $f, g : X \to \overline{\mathbb{R}}$ are proper functions, we have the *infimal convolution* of f and g defined by

$$f \square g : X \to \overline{\mathbb{R}},\ f \square g(a) = \inf\{f(x) + g(a-x) : x \in X\}.$$

Given a linear continuous mapping $A : X \to Y$, we have its *image-set* $\mathrm{Im}(A) = AX = \{Ax : x \in X\} \subseteq Y$ and its *adjoint* $A^* : Y^* \to X^*$ given by $\langle A^*y^*, x\rangle = \langle y^*, Ax\rangle$ for any $(x, y^*) \in X \times Y^*$. For the proper function $f : X \to \overline{\mathbb{R}}$ we define also the *infimal function of f through A* as $Af : Y \to \overline{\mathbb{R}}$, $Af(y) = \inf\left\{f(x) : x \in X,\ Ax = y\right\}$, $y \in Y$. Throughout the present paper when an infimum or a supremum is attained we write min, respectively, max instead of inf and sup.

Further we give some results concerning the composition of a function with a linear continuous operator.

LEMMA 1 (cf. [8, Theorem 2.7]). *Let X and Y be nontrivial locally convex spaces, $A : X \to Y$ a linear continuous mapping, and $f : Y \to \overline{\mathbb{R}}$ a proper, convex, and lower semicontinuous function such that $f \circ A$ is proper on X. Then*

$$\mathrm{epi}((f \circ A)^*) = \mathrm{cl}(\mathrm{epi}(A^* f^*)), \tag{1}$$

where the closure is taken in the product topology of $(X^*, \tau) \times \mathbb{R}$, for every locally convex topology τ on X^* giving X as dual.

Remark 1. Significant choices for τ in the preceding lemma are the weak* topology $w(X^*, X)$ on X^* or the norm topology of X^* in case X is a reflexive Banach space.

LEMMA 2 (cf. [2, Theorem 2.4]). *Let X and Y be nontrivial locally convex spaces, τ a compatible topology on X^*, $A : X \to Y$ a linear continuous mapping, and $f : Y \to \overline{\mathbb{R}}$ a proper function. Then*

$$\mathrm{cl}(\mathrm{epi}(A^* f^*)) = \mathrm{cl}(A^* \times \mathrm{id}_{\mathbb{R}}(\mathrm{epi}(f^*))), \tag{2}$$

where the closure is taken in the product topology of $(X^*, \tau) \times \mathbb{R}$.

Taking into (1) and (2) the closure in the product topology of $(X^*, \tau) \times \mathbb{R}$, with τ any locally convex topology on X^* giving X as dual, we get

$$\mathrm{epi}((f \circ A)^*) = \mathrm{cl}(\mathrm{epi}(A^* f^*)) = \mathrm{cl}(A^* \times \mathrm{id}_{\mathbb{R}}(\mathrm{epi}(f^*))). \tag{3}$$

DEFINITION 1. *A set $M \subseteq X$ is said to be* closed regarding the subspace $Z \subseteq X$ *if $M \cap Z = \mathrm{cl}(M) \cap Z$.*

PROPOSITION 1. *Let X, Y, and U be nontrivial locally convex spaces, $A : X \to Y$ a linear continuous mapping, and $f : Y \to \overline{\mathbb{R}}$ a proper, convex, and lower semicontinuous function such that $f \circ A$ is proper on X. Consider, moreover, a linear mapping $M : U \to X^*$. Let τ be any locally convex topology on X^* giving X as dual. The following statements are equivalent.*

(a) *$A^* \times \mathrm{id}_{\mathbb{R}}(\mathrm{epi}(f^*))$ is closed regarding the subspace $\mathrm{Im}(M) \times \mathbb{R}$ in the product topology of $(X^*, \tau) \times \mathbb{R}$.*

(b) *$(f \circ A)^*(Mu) = \min\left\{f^*(y^*) : A^*y^* = Mu\right\}$ $\forall u \in U$.*

Proof. Because f is proper, convex, and lower semicontinuous, A linear and continuous, and $f \circ A$ proper it follows that $(f \circ A)^*$ is proper, convex, and lower semicontinuous.

(a) $\Rightarrow$ (b) Let $u \in U$. For any $y^* \in Y^*$ fulfilling $A^*y^* = Mu$ we have

$$f^*(y^*) \geq \sup_{x \in X}\{\langle y^*, Ax\rangle - f(Ax)\} = \sup_{x \in X}\{\langle Mu, x\rangle - (f \circ A)(x)\} = (f \circ A)^*(Mu),$$

and so we conclude that

$$\inf\left\{f^*(y^*) : A^*y^* = Mu\right\} \geq (f \circ A)^*(Mu). \tag{4}$$

If $(f \circ A)^*(Mu) = +\infty$, then (4) yields $f^*(y^*) = +\infty = (f \circ A)^*(Mu)$ for any $y^* \in Y^*$ such that $A^*y^* = Mu$. Consider further the case $(f \circ A)^*(Mu) \in \mathbb{R}$. It follows that $(Mu, (f \circ A)^*(Mu)) \in \operatorname{epi}((f \circ A)^*)$, and it is clear that it belongs also to $\operatorname{Im}(M) \times \mathbb{R}$. By (3), (a) gives

$$\begin{aligned}(A^* \times \operatorname{id}_{\mathbb{R}}(\operatorname{epi}(f^*))) \cap (\operatorname{Im}(M) \times \mathbb{R}) &= \operatorname{cl}(A^* \times \operatorname{id}_{\mathbb{R}}(\operatorname{epi}(f^*))) \cap (\operatorname{Im}(M) \times \mathbb{R}) \\ &= \operatorname{epi}((f \circ A)^*) \cap (\operatorname{Im}(M) \times \mathbb{R}),\end{aligned}$$

and so there is some $\bar{y}^* \in Y^*$ such that $A^*\bar{y}^* = Mu$ and $(\bar{y}^*, (f \circ A)^*(Mu)) \in \operatorname{epi}(f^*)$ or, equivalently, $f^*(\bar{y}^*) \leq (f \circ A)^*(Mu)$. Thus by (4) we get

$$\min\left\{f^*(y^*) : A^*y^* = Mu\right\} = (f \circ A)^*(Mu). \tag{5}$$

(b) $\Rightarrow$ (a) From (3) one gets $\operatorname{epi}((f \circ A)^*) \supseteq A^* \times \operatorname{id}_{\mathbb{R}}(\operatorname{epi}(f^*))$, followed by

$$\operatorname{epi}((f \circ A)^*) \cap (\operatorname{Im}(M) \times \mathbb{R}) \supseteq (A^* \times \operatorname{id}_{\mathbb{R}}(\operatorname{epi}(f^*))) \cap (\operatorname{Im}(M) \times \mathbb{R}).$$

For any pair $(x^*, r) \in \operatorname{epi}((f \circ A)^*) \cap (\operatorname{Im}(M) \times \mathbb{R})$ there is some $u \in U$ such that $x^* = Mu$ and we have $(f \circ A)^*(x^*) = (f \circ A)^*(Mu) \leq r$. The hypothesis (b) grants the existence of an $\bar{y}^* \in Y^*$ satisfying both $A^*\bar{y}^* = Mu = x^*$ and $f^*(\bar{y}^*) = (f \circ A)^*(Mu) \leq r$, i.e., $(\bar{y}^*, r) \in \operatorname{epi}(f^*)$. Thus $(x^*, r) = (A^*\bar{y}^*, r) \in A^* \times \operatorname{id}_{\mathbb{R}}(\operatorname{epi}(f^*))$, and as one thus gets

$$\operatorname{epi}((f \circ A)^*) \cap (\operatorname{Im}(M) \times \mathbb{R}) \subseteq (A^* \times \operatorname{id}_{\mathbb{R}}(\operatorname{epi}(f^*))) \cap (\operatorname{Im}(M) \times \mathbb{R}),$$

the conclusion follows by (3). □

COROLLARY 1 (cf. [2, Theorem 3.3]). *Let X and Y be nontrivial locally convex spaces, $A : X \to Y$ a linear continuous mapping, and $f : Y \to \overline{\mathbb{R}}$ a proper, convex, and lower semicontinuous function such that $f \circ A$ is proper. Let τ be any locally convex topology on X^* giving X as dual. Then*

(i) *$A^* \times \operatorname{id}_{\mathbb{R}}(\operatorname{epi}(f^*))$ is closed in the product topology of $(X^*, \tau) \times \mathbb{R}$ if and only if for any $x^* \in X^*$ one has*

$$(f \circ A)^*(x^*) = \min\{f^*(y^*) : A^*y^* = x^*\};$$

(ii) *if $A^* \times \operatorname{id}_{\mathbb{R}}(\operatorname{epi}(f^*))$ is closed in the product topology of $(X^*, \tau) \times \mathbb{R}$, then for any $x \in \operatorname{dom}(f \circ A)$ one has $\partial(f \circ A)(x) = A^*\partial f(Ax)$.*

Proof. (i) follows from Proposition 1 when taking $U = X^*$ and $M = \operatorname{id}_{X^*}$, while for (ii) we refer the reader to [2] and [11]. □

Remark 2. Let τ be any locally convex topology on X^* giving X as dual. We know that $A^* \times \operatorname{id}_{\mathbb{R}}(\operatorname{epi}(f^*)) \subseteq \operatorname{epi}(A^*f^*) \subseteq \operatorname{epi}((f \circ A)^*)$ (see Lemmas 1 and 2). From (3) it follows that $A^* \times \operatorname{id}_{\mathbb{R}}(\operatorname{epi}(f^*))$ is closed in the product topology of $(X^*, \tau) \times \mathbb{R}$ if and only if

$$A^* \times \operatorname{id}_{\mathbb{R}}(\operatorname{epi}(f^*)) = \operatorname{epi}(A^*f^*) = \operatorname{epi}((f \circ A)^*).$$

The second part of this section in devoted to monotone operators and some of their properties. Consider further X a Banach space equipped with the norm $\|\cdot\|$, while the norm on X^* is $\|\cdot\|_*$.

DEFINITION 2 (cf. [20]). *A multifunction $T : X \rightrightarrows X^*$ is called a* monotone operator*, provided that for any $x, y \in X$ one has*

$$\langle y^* - x^*, y - x\rangle \geq 0 \quad \textit{whenever } x^* \in T(x) \textit{ and } y^* \in T(y).$$

DEFINITION 3 (cf. [20]). *For any monotone operator $T : X \rightrightarrows X^*$ we have*

- *its* effective domain $D(T) = \{x \in X : T(x) \neq \emptyset\}$,
- *its* range $R(T) = \cup\{T(x) : x \in X\}$,
- *its* graph $G(T) = \{(x, x^*) : x \in X,\ x^* \in T(x)\}$.

DEFINITION 4 (cf. [20]). *A monotone operator $T : X \rightrightarrows X^*$ is called* maximal *when its graph is not properly included in the graph of any other monotone operator $T' : X \rightrightarrows X^*$.*

The subdifferential of a proper convex lower semicontinuous function on X is a typical example of a maximal monotone operator (see [19]). As we shall see in section 4, it belongs to many other classes of operators, too. We introduce also the *duality map* $J : X \rightrightarrows X^*$ defined as follows:

$$J(x) = \frac{1}{2}\partial\|x\|^2 = \left\{x^* \in X^* : \|x\|^2 = \|x^*\|^2 = \langle x^*, x\rangle\right\} \quad \forall x \in X,$$

because it gives the following criterion for the maximal monotonicity of a monotone operator $T : X \rightrightarrows X^*$.

PROPOSITION 2 (cf. [1], [21]). *A monotone operator T on a reflexive Banach space X is maximal if and only if the mapping $T(x + \cdot) + J(\cdot)$ is surjective $\forall x \in X$.*

As underlined by many authors (see [1], [12], [15], [16], [21], [23], [24]), there are strong connections between the maximal monotone operators and convex analysis. They are best noticeable by the *Fitzpatrick function* associated with the monotone operators (see [7]). Rediscovered after some years, it proved to be crucial in treating the problem of maximal monotonicity of the sum of maximal monotone operators within the latest papers on the subject [1], [16], [23], [24]. To a monotone operator $T : X \rightrightarrows X^*$ Fitzpatrick attached the function

$$\varphi_T : X \times X^* \to \overline{\mathbb{R}},\ \varphi_T(x, x^*) = \sup\left\{\langle y^*, x\rangle + \langle x^*, y\rangle - \langle y^*, y\rangle : y^* \in T(y)\right\}.$$

For any monotone operator T it is quite clear that φ_T is a convex lower semicontinuous function as a supremum of a family of continuous affine functions. An important result regarding the Fitzpatrick function and its conjugate in reflexive Banach spaces follows.

PROPOSITION 3 (cf. [5], [7], [15], [23]). *Let T be a maximal monotone operator on a reflexive Banach space X. Then for any pair $(x, x^*) \in X \times X^*$ we have*

$$\varphi_T^*(x^*, x) \geq \varphi_T(x, x^*) \geq \langle x^*, x\rangle.$$

Moreover, $\varphi_T^*(x^*, x) = \varphi_T(x, x^*) = \langle x^*, x\rangle$ if and only if $(x, x^*) \in G(T)$.

3. Maximal monotonicity for the precomposition with a linear operator. Within this section X and Y will be reflexive Banach spaces. Given the maximal monotone operator T on Y and the linear continuous mapping $A : X \to Y$, such that $A^{-1}\big(\mathrm{pr}_1(\mathrm{dom}(\varphi_T))\big) \neq \emptyset$, we introduce the operator $T_A : X \rightrightarrows X^*$ defined by $T_A(x) = A^* \circ T \circ A(x)$, $x \in X$, which is monotone but not always maximal monotone.

3.1. Maximal monotonicity for T_A. Various conditions which ensure the maximal monotonicity of T_A were given in many recent papers, among which we mention [1], [6], [13], [15], and [24]. We prove, using an idea due to Borwein [1], that T_A is maximal monotone, provided that the following constraint qualification is fulfilled:

(CQ) $A^* \times \mathrm{id}_Y \times \mathrm{id}_{\mathbb{R}}(\mathrm{epi}(\varphi_T^*))$ is closed regarding the subspace $X^* \times \mathrm{Im}(A) \times \mathbb{R}$.

THEOREM 1. *If (CQ) is fulfilled, then T_A is a maximal monotone operator.*

Proof. Let us fix some $z \in X$ and $z^* \in X^*$ and consider $f, g : X \times X^* \to \overline{\mathbb{R}}$, defined by

$$f(x, x^*) = \inf\{\varphi_T(A(x+z), y^*) - \langle y^*, Az\rangle : A^*y^* = x^* + z^*\}$$

and

$$g(x, x^*) = \frac{1}{2}\|x\|^2 + \frac{1}{2}\|x^*\|_*^2 - \langle z^*, x\rangle,\ (x, x^*) \in X \times X^*.$$

As f and g are convex and the latter is continuous we can apply Fenchel's duality theorem (see [25]) that guarantees the existence of some pair $(\bar{x}^*, \bar{x}) \in X^* \times X$ such that

$$\begin{aligned}\inf_{(x,x^*)\in X\times X^*}\{f(x, x^*) + g(x, x^*)\} &= \max_{(x^*,x)\in X^*\times X}\{-f^*(x^*, x) - g^*(-x^*, -x)\}\\ &= -f^*(\bar{x}^*, \bar{x}) - g^*(-\bar{x}^*, -\bar{x}). \qquad (6)\end{aligned}$$

Let us calculate the conjugates of f and g. Before this we introduce the linear continuous operator $B = A \times \mathrm{id}_{Y^*}$. For any $(w^*, w) \in X^* \times X$ we have

$$\begin{aligned}f^*(w^*, w) &= \sup_{\substack{x\in X,\\ x^*\in X^*}}\left\{\langle w^*, x\rangle + \langle x^*, w\rangle - \inf_{A^*y^*=x^*+z^*}\{\varphi_T(A(x+z), y^*) - \langle y^*, Az\rangle\}\right\}\\ &= \sup_{\substack{(x,x^*)\in X\times X^*,\ y^*\in Y^*,\\ A^*y^*=x^*+z^*}}\{\langle w^*, x\rangle + \langle x^*, w\rangle - \varphi_T(A(x+z), y^*) + \langle y^*, Az\rangle\}\\ &= \sup_{\substack{x\in X,\ y^*\in Y^*,\\ u=x+z\in X}}\{\langle w^*, u - z\rangle + \langle A^*y^* - z^*, w\rangle - \varphi_T(A(u), y^*) + \langle A^*y^*, z\rangle\}\\ &= \sup_{\substack{u\in X,\\ y^*\in Y^*}}\{\langle w^*, u\rangle + \langle y^*, A(w+z)\rangle - (\varphi_T \circ B)(u, y^*)\} - \langle w^*, z\rangle - \langle z^*, w\rangle\\ &= (\varphi_T \circ B)^*(w^*, A(w+z)) - \langle w^*, z\rangle - \langle z^*, w\rangle\end{aligned}$$

and

$$\begin{aligned}g^*(w^*, w) &= \sup_{\substack{x\in X,\\ x^*\in X^*}}\left\{\langle w^*, x\rangle + \langle x^*, w\rangle - \frac{1}{2}\|x\|^2 - \frac{1}{2}\|x^*\|_*^2 + \langle z^*, x\rangle\right\}\\ &= \frac{1}{2}\|w^* + z^*\|_*^2 + \frac{1}{2}\|w\|^2.\end{aligned}$$

Proposition 1 ensures that (CQ) is equivalent to the fact that for any $(w^*, w) \in X^* \times X$ one has

$$(\varphi_T \circ B)^*(w^*, Aw) = \min_{(y^*,y)\in Y^*\times Y}\{\varphi_T^*(y^*, y) : B^*(y^*, y) = (w^*, Aw)\}.$$

For any $(x, x^*) \in X \times X^*$ and $y^* \in Y^*$ such that $A^*y^* = x^* + z^*$, by Proposition 3, we have

$$\varphi_T(A(x+z), y^*) - \langle y^*, Az\rangle \geq \langle y^*, Ax + Az\rangle - \langle y^*, Az\rangle = \langle A^*y^*, x\rangle = \langle x^*, x\rangle + \langle z^*, x\rangle,$$

and so $f(x, x^*) \geq \langle x^*, x\rangle + \langle z^*, x\rangle$. Since $g(x, x^*) \geq -\langle x^*, x\rangle - \langle z^*, x\rangle$ we get $f(x, x^*) + g(x, x^*) \geq 0$. Thus $\inf_{(x,x^*)\in X\times X^*}\{f(x, x^*) + g(x, x^*)\} \geq 0$, and taking it into (6) one gets $f^*(\bar{x}^*, \bar{x}) + g^*(-\bar{x}^*, -\bar{x}) \leq 0$, i.e.,

$$(\varphi_T \circ B)^*(\bar{x}^*, A(\bar{x}+z)) - \langle \bar{x}^*, z\rangle - \langle z^*, \bar{x}\rangle + \frac{1}{2}\|-\bar{x}^* + z^*\|_*^2 + \frac{1}{2}\|-\bar{x}\|^2 \leq 0. \tag{7}$$

From Proposition 1 we have

$$(\varphi_T \circ B)^*(\bar{x}^*, A(\bar{x}+z)) = \min_{(y^*,y)\in Y^*\times Y}\{\varphi_T^*(y^*, y) : B^*(y^*, y) = (\bar{x}^*, A(\bar{x}+z))\},$$

with the minimum attained at some $(\bar{y}^*, \bar{y}) \in Y^* \times Y$. As the adjoint operator of B is $B^* : Y^* \times Y \to X^* \times Y$, $B^*(y^*, y) = (A^*y^*, y)$, it follows that $B^*(\bar{y}^*, \bar{y}) = (A^*\bar{y}^*, \bar{y}) = (\bar{x}^*, A(\bar{x}+z))$. Taking the last two relations into (7) we have

$$\begin{aligned}
0 &\geq \varphi_T^*(\bar{y}^*, \bar{y}) - \langle \bar{x}^*, z\rangle - \langle z^*, \bar{x}\rangle + \frac{1}{2}\|\bar{x}^* - z^*\|_*^2 + \frac{1}{2}\|\bar{x}\|^2 \\
&= \varphi_T^*(\bar{y}^*, A(\bar{x}+z)) - \langle \bar{y}^*, Az\rangle - \langle \bar{y}^*, A\bar{x}\rangle + \langle \bar{y}^*, A\bar{x}\rangle - \langle z^*, \bar{x}\rangle + \frac{1}{2}\|\bar{x}\|^2 \\
&\quad + \frac{1}{2}\|A^*\bar{y}^* - z^*\|_*^2 = \big(\varphi_T^*(\bar{y}^*, A(\bar{x}+z)) - \langle \bar{y}^*, A(\bar{x}+z)\rangle\big) \\
&\quad + \left(\langle A^*\bar{y}^* - z^*, \bar{x}\rangle + \frac{1}{2}\|A^*\bar{y}^* - z^*\|_*^2 + \frac{1}{2}\|\bar{x}\|^2\right) \geq 0,
\end{aligned}$$

where the last inequality comes from Proposition 3. Thus the inequalities above must be fulfilled as equalities, and so

$$\varphi_T^*(\bar{y}^*, A(\bar{x}+z)) - \langle \bar{y}^*, A(\bar{x}+z)\rangle = 0;$$

i.e., by Proposition 3, $\bar{y}^* \in T \circ A(\bar{x}+z)$ and

$$\langle A^*\bar{y}^* - z^*, \bar{x}\rangle + \frac{1}{2}\|A^*\bar{y}^* - z^*\|_*^2 + \frac{1}{2}\|\bar{x}\|^2 = 0,$$

i.e., $z^* - A^*\bar{y}^* \in \partial\frac{1}{2}\|\cdot\|^2(\bar{x})$. Further one has $A^*\bar{y}^* \in A^* \circ T \circ A(z+\bar{x}) = T_A(z+\bar{x})$ and $z^* - A^*\bar{y}^* \in J(\bar{x})$, and so $z^* \in T_A(z+\bar{x}) + J(\bar{x})$. As z and z^* have been arbitrarily chosen, Proposition 2 yields the conclusion. □

Remark 3. We compare in the following the constraint qualification (CQ) to some generalized interior-point regularity conditions given in the literature in order to ensure the maximality of the monotone operator T_A. Under the condition in [6] one gets the fulfillment of the ones considered in [1] and [15], which imply the ones in [13] and [24], that are actually equivalent (according to Theorem 7 in [24]) to

$$(CQ_Z) \qquad \bigcup_{\lambda>0} \lambda(D(T) - \mathrm{Im}(A)) \text{ is a closed linear subspace.}$$

By Corollary 3.6 in [16] this is nothing but $0 \in^{ic} (\mathrm{pr}_1(\mathrm{dom}(\varphi_T)) - \mathrm{Im}(A))$ or, equivalently, $\cup_{\lambda>0}\lambda(\mathrm{pr}_1(\mathrm{dom}(\varphi_T)) - \mathrm{Im}(A))$ is a closed linear subspace. This is actually the same with $\cup_{\lambda>0}\lambda\big(\mathrm{dom}(\varphi_T) - \mathrm{Im}(A) \times Y^*\big)$ being a closed linear subspace,

and so, taking into account that $B = A \times \mathrm{id}_{Y^*}$, $0 \in^{ic} (\mathrm{dom}(\varphi_T) - \mathrm{Im}(B))$. This yields, by Theorem 2.3.8(vii) in [25],

$$(\varphi_T \circ B)^*(w^*, Aw) = \min_{(y^*,y)\in Y^*\times Y} \{\varphi_T^*(y^*, y) : B^*(y^*, y) = (w^*, Aw)\},$$

which is equivalent to (CQ). Therefore $(CQ_Z) \Rightarrow (CQ)$. A counterexample to show that it is possible to have (CQ) satisfied and (CQ_Z) violated is given later.

Remark 4. The maximal monotonicity of T_A is valid also when imposing the constraint qualification

$$(\widetilde{CQ}) \qquad A^* \times \mathrm{id}_Y \times \mathrm{id}_{\mathbb{R}}(\mathrm{epi}(\varphi_T^*)) \text{ is closed.}$$

The only difference in the proof is that we use Corollary 1(i) instead of Proposition 1. One may notice that we have $(CQ_Z) \Rightarrow (\widetilde{CQ}) \Rightarrow (CQ)$; i.e., $(\widetilde{CQ})$ is still weaker than (CQ_Z).

The remaining part of the section is dedicated to the proof of the fact that (CQ) is indeed weaker than (CQ_Z).

Example 1. Let $X = \mathbb{R}$ and $Y = \mathbb{R} \times \mathbb{R}$. Then $X^* = \mathbb{R}$ and $Y^* = \mathbb{R} \times \mathbb{R}$. Consider the operator $T : \mathbb{R} \times \mathbb{R} \to 2^{\mathbb{R}\times\mathbb{R}}$ defined by

$$T(x, y) = \begin{cases} (-\infty, 0] \times \{0\} & \text{if } x = 0,\ y < 0, \\ (-\infty, 0] \times [0, +\infty) & \text{if } x = y = 0, \\ \{x\} \times \{0\} & \text{if } x > 0,\ y < 0, \\ \{x\} \times [0, +\infty) & \text{if } x > 0,\ y = 0, \\ \emptyset & \text{otherwise.} \end{cases}$$

It is not difficult to notice that, considering the following proper, convex, and lower semicontinuous functions $f, g : \mathbb{R} \to \overline{\mathbb{R}}$, $f(x) = (1/2)x^2 + \delta_{[0,+\infty)}(x)$, and $g = \delta_{(-\infty,0]}$, for any $(x, y) \in \mathbb{R} \times \mathbb{R}$ we have $T(x, y) = \big(\partial f(x), \partial g(y)\big)$, and thus T is a maximal monotone operator. Taking $A : \mathbb{R} \to \mathbb{R} \times \mathbb{R}$, $Ax = (x, x)$, one gets, for any $x \in \mathbb{R}$,

$$T_A(x) = A^* \circ T \circ A(x) = \partial f(x) + \partial g(x) = \begin{cases} \mathbb{R} & \text{if } x = 0, \\ \emptyset & \text{otherwise,} \end{cases}$$

and thus T_A is a maximal monotone operator, too.

Let us calculate the conjugate of φ_T to see if (CQ) is fulfilled. We have $\forall (x, y, x^*, y^*) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}$

$$\varphi_T(x, y, x^*, y^*) = \begin{cases} \left(\frac{x+x^*}{2}\right)^2 & \text{if } x \geq 0,\ x + x^* > 0,\ y \leq 0,\ y^* \geq 0, \\ 0 & \text{if } x \geq 0,\ x + x^* \leq 0,\ y \leq 0,\ y^* \geq 0, \text{ and} \\ +\infty & \text{otherwise,} \end{cases}$$

$$\varphi_T^*(x^*, y^*, x, y) = \begin{cases} x^2 & \text{if } x \geq 0,\ x \geq x^*,\ y^* \geq 0,\ y \leq 0, \\ +\infty & \text{otherwise.} \end{cases}$$

Thus the epigraph of the conjugate is

$$\mathrm{epi}(\varphi_T^*) = \bigcup_{x\geq 0} \Big((-\infty, x] \times [0, +\infty) \times \{x\} \times (-\infty, 0] \times [x^2, +\infty)\Big),$$

and so

$$A^* \times \mathrm{id}_{\mathbb{R}\times\mathbb{R}} \times \mathrm{id}_{\mathbb{R}}(\mathrm{epi}(\varphi_T^*)) = \mathbb{R} \times \bigcup_{x\geq 0} \Big(\{x\} \times (-\infty, 0] \times [x^2, +\infty)\Big),$$

which is closed; i.e., $(\widetilde{CQ})$ is valid. Thus it is closed regarding the subspace $\mathbb{R} \times \mathrm{Im}(A) \times \mathbb{R} = \mathbb{R} \times \Delta_{\mathbb{R}} \times \mathbb{R}$, too; i.e., (CQ) is satisfied for the chosen T and A. Here we used the notation $\Delta_X = \{(x, x) : x \in X\}$, in case $X = \mathbb{R}$.

Let us calculate now $\cup_{\lambda>0} \lambda(D(T) - \mathrm{Im}(A))$ in order to check the validity of (CQ_Z). It is clear that $D(T) = [0, +\infty) \times (-\infty, 0]$ and $\mathrm{Im}(A) = \Delta_{\mathbb{R}}$. We have $D(T) - \mathrm{Im}(A) = \{[x, +\infty) \times (-\infty, x] : x \in \mathbb{R}\}$, and so

$$\bigcup_{\lambda>0} \lambda(D(T) - \mathrm{Im}(A)) = \{[x, +\infty) \times (-\infty, x] : x \in \mathbb{R}\} = \{(x, y) \in \mathbb{R} : x \geq y\},$$

which is not a subspace, and thus (CQ_Z) is violated. Therefore, even if (CQ_Z) implies (CQ), the reverse implication does not always hold; i.e., (CQ) is indeed weaker than (CQ_Z).

3.2. Maximal monotonicity for the sum of two maximal monotone operators. An important special case of the problem treated in Theorem 1 is the situation when the sum of two maximal monotone operators is maximal monotone. This case is obtained from the general one by taking $Y = X \times X$, $A(x) = (x, x)$ for any $x \in X$ and $T : X \times X \to X^* \times X^*$, $T(x, y) = (T_1(x), T_2(y))$ when $(x, y) \in X \times X$, where T_1 and T_2 are maximal monotone operators on X. It is a simple verification to show that T is maximal monotone. Having these choices, for any $x \in X$ we have $T_A(x) = T_1(x) + T_2(x)$. Moreover, the condition on the domain of φ_T becomes $\mathrm{pr}_1(\mathrm{dom}(\varphi_{T_1})) \cap \mathrm{pr}_1(\mathrm{dom}(\varphi_{T_2})) \neq \emptyset$.

The literature concerning the maximal monotonicity of $T_1 + T_2$ is richer than the one in the more general case. Let us mention here, alongside the papers already cited above, also [3], [20], and [23]. A comprehensive study on this problem is available in [21] (see also [24]), where many sufficient conditions for the maximal monotonicity of the sum of two maximal monotone operators are compared and classified.

Our constraint qualification (CQ) becomes in this special case

(CQ^s) $\{(x^* + y^*, x, y, r) : \varphi^*_{T_1}(x^*, x) + \varphi^*_{T_2}(y^*, y) \leq r\}$ is closed regarding the subspace $X^* \times \Delta_X \times \mathbb{R}$.

Introducing the function (see [23])

$$\rho : X \times X^* \to \overline{\mathbb{R}},\ \rho(v, v^*) = \inf_{x^*, y^* \in X^*} \{\varphi_{T_1}(v, x^*) + \varphi_{T_2}(v, y^*) : x^* + y^* = v^*\},$$

by using Proposition 1, one can show that (CQ^s) is equivalent to

$$\begin{aligned} \rho^*(w^*, w) &= \min_{\substack{x, y \in X, \\ x^*, y^* \in X^*}} \{\varphi^*_{T_1}(x^*, x) + \varphi^*_{T_2}(y^*, y) : x^* + y^* = w^*,\ x = y = w\} \\ &= \min_{\substack{x^*, y^* \in X^*, \\ x^* + y^* = w^*}} \{\varphi^*_{T_1}(x^*, w) + \varphi^*_{T_2}(y^*, w)\}. \end{aligned}$$

In [23] Simons and Zălinescu prove that a sufficient condition for having this equality fulfilled is

$$\bigcup_{\lambda>0} \big(\mathrm{pr}_1(\mathrm{dom}(\varphi_{T_1})) - \mathrm{pr}_1(\mathrm{dom}(\varphi_{T_2}))\big)$$

being a closed linear subspace, which was the weakest constraint qualification guar-

anteeing the maximal monotonicity of $T_1 + T_2$ known so far. Consequently, (CQ) delivers the weakest constraint qualification for the maximal monotonicity of sum of two maximal monotone operators.

THEOREM 2. *Let T_1 and T_2 be maximal monotone operators on X such that $\mathrm{pr}_1(\mathrm{dom}(\varphi_{T_1})) \cap \mathrm{pr}_1(\mathrm{dom}(\varphi_{T_2})) \neq \emptyset$. If (CQ^s) is fulfilled, then $T_1 + T_2$ is a maximal monotone operator on X.*

Remark 5. The other constraint qualification we gave, $(\widetilde{CQ})$, becomes in this case

$$(\widetilde{CQ}^s) \qquad \{(x^* + y^*, x, y, r) : \varphi^*_{T_1}(x^*, x) + \varphi^*_{T_2}(y^*, y) \leq r\} \text{ is closed.}$$

One can prove that $(\widetilde{CQ}^s)$ is weaker than the other constraint qualifications mentioned within this subsection, except (CQ^s), which is implied by it.

Remark 6. One of the reviewers suggested that we try to obtain a constraint qualification for the maximal monotonicity of the precomposition with a linear operator assuming (CQ^s) to be known, by using the approach in [16]. For an appropriate choice of the monotone operators in the sum we have obtained Theorem 1 from Theorem 2. Let us notice that the same applies when using the approach in [9], quite different from the one in [16], for expressing the precomposition via a sum of monotone operators.

4. Brézis–Haraux-type approximation of the range of the subdifferential of the precomposition with a linear operator. Within this part X and Y are considered Banach spaces, unless otherwise specified. Let us mention that, unlike in the previous section, here we do not ask these spaces to be, moreover, reflexive. We rectify, weaken, and generalize a statement due to Riahi [18] concerning the so-called Brézis–Haraux-type approximation of the range of the sum of the subdifferentials of two proper convex lower semicontinuous functions, giving it for the operator T_A introduced in the previous section. Riahi's statement is recovered as a special case, under a weaker sufficient condition than in the original paper.

4.1. Some preliminaries. We need to introduce some notions and to recall some results which are dealt with only within this part. First, we define the so-called monotone operators of dense type, originally introduced by Gossez in [10], of type 3^*, also known as star monotone and of type (BH), and operators of type (NI). Let us stress once again that we work in nonreflexive Banach spaces.

Before this we need to introduce τ_1 as being the weakest topology on X^{**} which renders continuous the following real functions:

$$X^{**} \to \mathbb{R}: \quad x^{**} \mapsto \langle x^{**}, x^* \rangle \quad \forall x^* \in X^*,$$
$$X^{**} \to \mathbb{R}: \quad x^{**} \mapsto \|x^{**}\|.$$

The topology τ considered in $X^{**} \times X^*$ will be the product topology of τ_1 and the strong (norm) topology of X^* (see [10]) .

DEFINITION 5 (cf. [10]). *A monotone operator $T : Y \rightrightarrows Y^*$ is called of* dense type*, provided that the* closure operator $\overline{T} : Y^{**} \rightrightarrows Y^*$ *that is defined as follows:*

$$G(\overline{T}) = \{(x^{**}, x^*) \in Y^{**} \times Y^* : \exists (x_i, x_i^*)_i \in G(T) \text{ with } (\hat{x}_i, x_i^*) \xrightarrow{\tau} (x^{**}, x^*)\}$$

is maximal monotone.

Different from Riahi in [18], where these operators are called *densely maximal monotone*, we decided to call them of *dense type* as originally done by Gossez in [10]. By Lemme 2.1 in the same paper, for a monotone operator $T : Y \rightrightarrows Y^*$ of dense type, one has $(x^{**}, x^*) \in G(\overline{T})$ if and only if $\langle x^{**} - \hat{y},\, x^* - y^* \rangle \geq 0\ \forall (y, y^*) \in G(T)$.

DEFINITION 6 (cf. [6], [14], [18]). *A monotone operator $T : Y \rightrightarrows Y^*$ is called* 3^*-monotone *if $\forall x^* \in R(T)$ and $x \in D(T)$ there is some $\beta(x^*, x) \in \mathbb{R}$ such that* $\inf_{y^* \in T(y)} \langle x^* - y^*,\, x - y \rangle \geq \beta(x^*, x)$.

DEFINITION 7 (cf. [21], [22]). *An operator $T : Y \rightrightarrows Y^*$ is called of* type (NI) *if $\forall (x^{**}, x^*) \in Y^{**} \times Y^*$ one has* $\inf_{y^* \in T(y)} \langle \hat{y} - x^{**},\, y^* - x^* \rangle \leq 0$.

Some necessary results follow.

LEMMA 3 (cf. [18]). *Given the operator $T : Y \rightrightarrows Y^*$ of dense type and the nonempty subset $E \subseteq Y^*$ such that for any $x^* \in E$ there is some $x \in Y$ fulfilling* $\inf_{y^* \in T(y)} \langle y^* - x^*,\, y - x \rangle > -\infty$, *one has $E \subseteq \mathrm{cl}(R(T))$ and $\mathrm{int}(E) \subseteq R(\overline{T})$.*

PROPOSITION 4. *If $T : Y \rightrightarrows Y^*$ is 3^*-monotone and $A : X \to Y$ is a linear continuous mapping such that T_A is of dense type, then*

(i) $A^*(R(T)) \subseteq \mathrm{cl}(R(T_A))$,

(ii) $\mathrm{int}(A^*(R(T))) \subseteq R(\overline{T_A})$.

Proof. As T is 3^*-monotone, we have for any $s \in D(T)$ and any $s^* \in R(T)$ that there is some $\beta(s^*, s) \in \mathbb{R}$ such that $\beta(s^*, s) \leq \inf_{x^* \in T(x)} \langle s^* - x^*,\, s - x \rangle$.

To apply Lemma 3 for $E = A^*(R(T))$ and T_A, we need to verify if they satisfy its hypothesis. Take some $u^* \in A^*(R(T))$, and thus there is an $v^* \in R(T)$ such that $u^* = A^* v^*$. We have for any $u \in X$

$$\inf_{x^* \in T_A(x)} \langle x^* - u^*, x - u \rangle = \inf_{t^* \in T \circ A(x)} \langle A^* t^* - A^* v^*, x - u \rangle$$

$$= \inf_{t^* \in T \circ A(x)} \langle t^* - v^*, A(x - u) \rangle \geq \inf_{t^* \in T(t)} \langle t^* - v^*, t - Au) \rangle \geq \beta(v^*, Au) > -\infty.$$

Having this fulfilled for any u, we apply Lemma 3 which yields (i) and (ii). □

PROPOSITION 5 (cf. [10, Remark 1]). *In reflexive Banach spaces every maximal monotone operator is of dense type and coincides with its closure operator.*

The last result we give here carries the 3^*-monotonicity from T to T_A.

PROPOSITION 6. *If $T : Y \rightrightarrows Y^*$ is 3^*-monotone and $A : X \to Y$ is a linear continuous mapping, then T_A is 3^*-monotone, too.*

Proof. Take $x^* \in R(T_A)$; i.e., there is some $z \in X$ such that $x^* \in A^* \circ T \circ A(z)$. Thus there exists a $z^* \in T \circ A(z)$ satisfying $x^* = A^* z^*$. Clearly, $z^* \in R(T)$. Consider also an $x \in D(T_A)$ and denote $u = Ax \in D(T)$. When $y^* \in T_A(y)$ there is some $t^* \in T \circ A(y)$ such that $y^* = A^* t^*$. We have

$$\begin{aligned} \inf_{y^* \in T_A(y)} \langle x^* - y^*, x - y \rangle &= \inf_{t^* \in T \circ A(y)} \langle A^* z^* - A^* t^*, x - y \rangle \\ &= \inf_{t^* \in T \circ A(y)} \langle z^* - t^*, A(x - y) \rangle \\ &\geq \inf_{t^* \in T(v)} \langle z^* - t^*, u - v \rangle \geq \beta(z^*, u) \in \mathbb{R}, \end{aligned}$$

as T is 3^*-monotone. Therefore T_A is 3^*-monotone, too. □

4.2. Rectifying and extending Riahi's results. We give here the main results in this section concerning the so-called Brézis–Haraux-type approximation (see [21]) of the range of the operator T_A (respectively, of the subdifferential of the precomposition of a linear continuous mapping with a proper convex lower semicontinuous

function). Some results related to them were obtained by Pennanen in [14] but in reflexive spaces.

THEOREM 3. *If $T : Y \rightrightarrows Y^*$ is 3^*-monotone and $A : X \to Y$ is a linear continuous mapping such that T_A is of dense type, then*

(i) $\mathrm{cl}(A^*(R(T))) = \mathrm{cl}(R(T_A))$,

(ii) $\mathrm{int}(R(T_A)) \subseteq \mathrm{int}(A^*(R(T))) \subseteq \mathrm{int}(R(\overline{T_A}))$.

Proof. By Proposition 4(i) we have also $\mathrm{cl}(A^*(R(T))) \subseteq \mathrm{cl}(R(T_A))$ and $\mathrm{int}(A^*(R(T))) \subseteq \mathrm{int}(R(\overline{T_A}))$. Take some $x^* \in R(T_A)$. Then there are some $x \in X$ and $y^* \in T \circ A(x) \subseteq R(T)$ such that $x^* = A^* y^*$. Thus $x^* \in A^*(R(T))$, and so $R(T_A) \subseteq A^*(R(T))$, and so the same inclusion stands also between the closures (respectively, the interiors) of these sets. Relations (i) and (ii) follow immediately by Proposition 4. □

Remark 7. The previous statement generalizes Theorem 1 in [18], which can be obtained for $Y = X \times X$, $Ax = (x, x)$, and $T = (T_1, T_2)$. The next consequence extends Corollary 1 in [18] which arises for the same choice of Y, A, and T.

COROLLARY 2. *Assume X to be, moreover, reflexive and let $T : Y \rightrightarrows Y^*$ be 3^*-monotone and $A : X \to Y$ a linear continuous mapping such that T_A is maximal monotone. Then one has $\mathrm{cl}(A^*(R(T))) = \mathrm{cl}(R(T_A))$ and $\mathrm{int}(R(T_A)) = \mathrm{int}(A^*(R(T)))$.*

Proof. As X is reflexive, Proposition 5 yields that T_A is of dense type and that $\overline{T_A}$ and T_A coincide. We apply Theorem 3, which yields the conclusion. □

The next statement generalizes Corollary 2 in [18], providing, moreover, a weaker constraint qualification under which one can assert the Brézis–Haraux-type approximation of the range of the sum of the subdifferentials of two proper convex lower semicontinuous functions. First, we give the constraint qualification that guarantees our more general result,

$(\overline{CQ})$ $A^* \times \mathrm{id}_{\mathbb{R}}(\mathrm{epi}(f^*))$ is closed in the product topology of $(X^*, w(X^*, X)) \times \mathbb{R}$.

THEOREM 4. *Let the proper convex lower semicontinuous function $f : Y \to \overline{\mathbb{R}}$ and the linear continuous operator $A : X \to Y$ such that $f \circ A$ be proper and assume $(\overline{CQ})$ to be valid. Then one has*

(i) $\mathrm{cl}(A^*(R(\partial f))) = \mathrm{cl}(R(\partial(f \circ A)))$,

(ii) $\mathrm{int}(R(\partial(f \circ A))) \subseteq \mathrm{int}(A^*(R(\partial f))) \subseteq \mathrm{int}(D(\partial(A^* f^*)))$.

Proof. By Corollary 1(ii) we know that $(\overline{CQ})$ implies $A^* \circ \partial f \circ A = \partial(f \circ A)$. Again, $f \circ A$ is proper, convex, and lower semicontinuous, and so by Théoréme 3.1 in [10] we know that $\partial(f \circ A)$ is an operator of dense type, while according to Theorem B in [19] (see also [14], [18]) ∂f is 3^*-monotone. Applying Theorem 3 for $T = \partial f$ we get $\mathrm{cl}(A^*(R(\partial f))) = \mathrm{cl}(R(A^* \circ \partial f \circ A))$, which with $(\overline{CQ})$ yields (i), and $\mathrm{int}(R(A^* \circ \partial f \circ A)) \subseteq \mathrm{int}(A^*(R(\partial f))) \subseteq \mathrm{int}(R(\overline{A^* \circ \partial f \circ A}))$. Using $(\overline{CQ})$ the latter becomes

$$\mathrm{int}(R(\partial(f \circ A))) \subseteq \mathrm{int}(A^*(R(\partial f))) \subseteq \mathrm{int}(R(\overline{\partial(f \circ A)})). \tag{8}$$

As from Corollary 1(i) one may deduce that under $(\overline{CQ})$ $A^* f^* = (f \circ A)^*$, by Théoréme 3.1 in [10] we get $R(\overline{\partial(f \circ A)}) = D(\partial(f \circ A)^*) = D(\partial(A^* f^*))$. Taking this into (8) we get (ii). □

When one takes $Y = X \times X$, $Ax = (x, x)$, and $f(x, y) = g(x) + h(y)$, where $x, y \in X$, the constraint qualification $(\overline{CQ})$ becomes (see [2])

$(\overline{CQ}^s)$ $\mathrm{epi}(g^*) + \mathrm{epi}(h^*)$ is closed in the product topology of $(X^*, w(X^*, X)) \times \mathbb{R}$.

COROLLARY 3. *Let g and h be two proper convex lower semicontinuous functions on the Banach space X with extended real values such that* $\operatorname{dom}(g) \cap \operatorname{dom}(h) \neq \emptyset$. *Assume* $(\overline{CQ}^s)$ *satisfied. Then one has*

(i) $\operatorname{cl}(R(\partial g) + R(\partial h)) = \operatorname{cl}(R(\partial(g+h)))$,

(ii) $\operatorname{int}(R(\partial(g+h))) \subseteq \operatorname{int}(R(\partial g) + R(\partial h)) \subseteq \operatorname{int}(D(\partial(g^* \square h^*)))$.

A similar result has been obtained by Riahi in Corollary 2 in [18]. There he said that under the constraint qualification

$$(\overline{CQ}_R) \qquad \bigcup_{t>0} t(\operatorname{dom}(g) - \operatorname{dom}(h)) \text{ is a closed linear subspace of } X,$$

one gets $\operatorname{cl}(R(\partial g) + R(\partial h)) = \operatorname{cl}(R(\partial(g+h)))$ and $\operatorname{int}(R(\partial g) + R(\partial h)) = \operatorname{int}(D(\partial(g^* \square h^*)))$.

We prove that the latter is not always true when $(\overline{CQ}_R)$ stands. For a proper, convex, and lower semicontinuous function $g : X \to \overline{\mathbb{R}}$ Riahi's relation would become $\operatorname{int}(R(\partial g)) = \operatorname{int}(D(\partial g^*))$, which is equivalent, by Théoréme 3.1 in [10] to

$$\operatorname{int}(R(\partial g)) = \operatorname{int}(R(\overline{\partial g})). \tag{9}$$

From the above mentioned theorem we also have that ∂g is a monotone operator of dense type, and it is also known that it is maximal monotone, too. According to Simons [22] ∂g is also of type (NI). Finally, by Lemme 2.1 in [10] and by Theorem 20 in [22], we get that $\operatorname{int}(R(\overline{\partial g}))$ is convex, and so (9) yields $\operatorname{int}(R(\partial g))$ convex. Unfortunately this is not always true, as Example 2.21 in [17], originally given by Fitzpatrick, shows. Take $X = c_0$, which is a Banach space with the usual supremum norm, and $g(x) = \|x\| + \|x - (1, 0, 0, \dots)\|$, a proper, convex, and continuous function on c_0. Skipping the calculatory details, it follows that $\operatorname{int}(R(\partial g))$ is not convex, unlike $\operatorname{int} R(\overline{\partial g})$. Thus (9) is false and the same happens to Riahi's allegation.

Remark 8. As proven in Proposition 3.1 in [4] (see also [2]), $(\overline{CQ}_R)$ implies $(\overline{CQ}^s)$, but the converse is not true, as shown by Example 3.1 in the same paper. Therefore our Corollary 3 extends, by weakening the constraint qualification, and corrects Corollary 2 in [18].

5. Conclusions. Given a maximal monotone operator T on the reflexive Banach space Y and the linear continuous operator $A : X \to Y$, where X is a reflexive Banach space, too, we give a sufficient condition for the maximal monotonicity of $A^* \circ T \circ A$ weaker than the generalized interior-point regularity conditions known to us from the literature. Moreover, when Y, A, and T are chosen such that the assertion turns into the maximal monotonicity of the sum of two maximal monotone operators on a reflexive Banach space X, we prove that our constraint qualification is actually the weakest condition guaranteeing the mentioned result known to us. In the second part of the paper, where we work in nonreflexive Banach spaces, we rectify and extend a result due to Riahi, giving a weak constraint qualification in order to ensure the so-called Brézis–Haraux-type approximation of the range of $\partial(f \circ A)$, where $f : Y \to \overline{\mathbb{R}}$ is a proper convex lower semicontinuous function. For a special choice of functions we prove that the corrected version of Riahi's results holds under a weaker constraint qualification than required by him.

Acknowledgments. The authors are grateful to the anonymous reviewers for some valuable suggestions which improved the quality of the paper. Moreover, we are thankful to one of them for making us aware of the differences between the classes of monotone operators of *dense type* and of *type* (D).

REFERENCES

[1] J. M. Borwein, *Maximal monotonicity via convex analysis*, J. Convex Anal., 13 (2006), pp. 561–586.

[2] R. I. Boţ and G. Wanka, *A weaker regularity condition for subdifferential calculus and Fenchel duality in infinite dimensional spaces*, Nonlinear Anal., 64 (2006), pp. 2787–2804.

[3] F. E. Browder, *Nonlinear maximal monotone operators in Banach space*, Math. Ann., 175 (1968), pp. 89–113.

[4] R. S. Burachik and V. Jeyakumar, *A dual condition for the convex subdifferential sum formula with applications*, J. Convex Anal., 12 (2005), pp. 279–290.

[5] R. S. Burachik and B. F. Svaiter, *Maximal monotonicity, conjugation and the duality product*, Proc. Amer. Math. Soc., 131 (2003), pp. 2379–2383.

[6] L.-J. Chu, *On the sum of monotone operators*, Michigan Math. J., 43 (1996), pp. 273–289.

[7] S. Fitzpatrick, *Representing monotone operators by convex functions*, in Workshop/Miniconference on Functional Analysis and Optimization (Canberra, 1988), Proc. Centre Math. Anal. Austral. Nat. Univ. 20, Australian National University, Canberra, Australia, 1988, pp. 59–65.

[8] S. P. Fitzpatrick and S. Simons, *The conjugates, compositions and marginals of convex functions*, J. Convex Anal., 8 (2001), pp. 423–446.

[9] Y. García, M. Lassonde, and J. P. Revalski, *Extended sums and extended compositions of monotone operators*, J. Convex Anal., 13 (2006), pp. 721–738.

[10] J.-P. Gossez, *Opérateurs monotones non linéaires dans les espaces de Banach non réflexifs*, J. Math. Anal. Appl., 34 (1971), pp. 371–395.

[11] J.-B. Hiriart-Urruty, *ε-subdifferential calculus*, in Convex Analysis and Optimization (London, 1980), Res. Notes in Math. 57, P. Aubin and R. B. Vinter, eds., Pitman, Boston, 1982, pp. 43–92.

[12] J. E. Martínez-Legaz and M. Théra, *A convex representation of maximal monotone operators*, J. Nonlinear Convex Anal., 2 (2004), pp. 243–247.

[13] T. Pennanen, *Dualization of generalized equations of maximal monotone type*, SIAM J. Optim., 10 (2000), pp. 809–835.

[14] T. Pennanen, *On the range of monotone composite mappings*, J. Nonlinear Convex Anal., 2 (2001), pp. 193–202.

[15] J.-P. Penot, *The relevance of convex analysis for the study of monotonicity*, Nonlinear Anal., 58 (2004), pp. 855–871.

[16] J.-P. Penot and C. Zălinescu, *Some problems about the representation of monotone operators by convex functions*, ANZIAM J., 47 (2005), pp. 1–20.

[17] R. R. Phelps, *Lectures on maximal monotone operators*, Extracta Math., 12 (1997), pp. 193–230.

[18] H. Riahi, *On the range of the sum of monotone operators in general Banach spaces*, Proc. Amer. Math. Soc., 124 (1996), pp. 3333–3338.

[19] R. T. Rockafellar, *On the maximal monotonicity of subdifferential mappings*, Pacific J. Math., 33 (1970), pp. 209–216.

[20] R. T. Rockafellar, *On the maximality of sums of nonlinear monotone operators*, Trans. Amer. Math. Soc., 149 (1970), pp. 75–88.

[21] S. Simons, *Minimax and Monotonicity*, Springer-Verlag, Berlin, 1998.

[22] S. Simons, *The range of a monotone operator*, J. Math. Anal. Appl., 199 (1996), pp. 176–201.

[23] S. Simons and C. Zălinescu, *Fenchel duality, Fitzpatrick functions and maximal monotonicity*, J. Nonlinear Convex Anal., 6 (2004), pp. 1–22.

[24] C. Zălinescu, *A new proof of the maximal monotonicity of the sum using the Fitzpatrick function*, in Variational Analysis and Applications, Nonconvex Optim. Appl. 79, F. Giannessi and A. Maugeri, eds., Springer-Verlag, New York, 2005, pp. 1159–1172.

[25] C. Zălinescu, *Convex Analysis in General Vector Spaces*, World Scientific, Singapore, 2002.

SIAM J. OPTIM.
Vol. 17, No. 4, pp. 1253–1257
© 2007 Society for Industrial and Applied Mathematics

AN IMPLEMENTABLE ACTIVE-SET ALGORITHM FOR COMPUTING A B-STATIONARY POINT OF A MATHEMATICAL PROGRAM WITH LINEAR COMPLEMENTARITY CONSTRAINTS: ERRATUM*

MASAO FUKUSHIMA† AND PAUL TSENG‡

Abstract. In [M. Fukushima and P. Tseng, *SIAM J. Optim.*, 12 (2002), pp. 724–739], an ϵ-active set algorithm was proposed for solving a mathematical program with a smooth objective function and linear inequality/complementarity constraints. It is asserted therein that, under a uniform LICQ on the ϵ-feasible set, this algorithm generates iterates whose cluster points are B-stationary points of the problem. However, the proof has a gap and shows only that each cluster point is an M-stationary point. We discuss this gap and show that B-stationarity can be achieved if the algorithm is modified and an additional error bound condition holds.

Key words. MPEC, B-stationary point, ϵ-active set, error bound

AMS subject classifications. 65K05, 90C30, 90C33

DOI. 10.1137/050642460

1. Introduction. In a recent paper by the authors [3], an ϵ-active set algorithm was proposed for solving the following mathematical program with equilibrium constraints (MPEC):

$$\begin{array}{lll} \text{minimize} & f(z) & \\ \text{subject to} & G_i(z) \geq 0, & i = 1, \ldots, m, \\ & H_i(z) \geq 0, & i = 1, \ldots, m, \\ & G_i(z)H_i(z) = 0, & i = 1, \ldots, m, \\ & g_j(z) \leq 0, & j = 1, \ldots, p, \\ & h_l(z) = 0, & l = 1, \ldots, q, \end{array} \tag{1}$$

where f is a real-valued continuously differentiable function on $\Re^n$ and G_i, H_i, g_j, h_l are real-valued *affine* functions on $\Re^n$. In Theorem 4.1(a) of [3], it is asserted that every cluster point of iterates generated by the algorithm is a B-stationary point of (1). However, the proof has a gap and shows only that every cluster point is an M-stationary point. We will discuss this gap and a modified algorithm that achieves B-stationarity under an additional error bound condition.

The gap occurs on [3, page 734] in the line "If $\nu_k \to 0$, then $|\mathcal{K}'| = \infty$, $\delta_k \to 0$, and the updating formula for ϵ_k would imply $\epsilon_k \to 0$, so any cluster point $\bar{z}$ of $\{\hat{z}^k\}_{k \in \mathcal{K}'}$ would be a KKT point of the *relaxed problem* $\mathrm{R}(\bar{z})$, which is a B-stationary point of

*Received by the editors October 11, 2005; accepted for publication (in revised form) June 7, 2006; published electronically January 22, 2007. This research is supported by Scientific Research Grant-in-Aid from the Ministry of Education, Science, Sports and Culture of Japan, and by National Science Foundation grant DMS-0511283.

http://www.siam.org/journals/siopt/17-4/64246.html

†Department of Applied Mathematics and Physics, Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan (fuku@amp.i.kyoto-u.ac.jp).

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

MPEC (1) under the uniform LICQ." In particular, we have for all $k \in \mathcal{K}'$ that

$$v_i^k \geq -\nu_k \quad \text{and} \quad w_i^k \geq -\nu_k \quad \forall\, i \in \hat{A}^k \cap \hat{B}^k, \tag{2}$$

where $\hat{A}^k$, $\hat{B}^k$ are given by [3, eq. (7)] and v_i^k, w_i^k are multipliers associated with $\hat{z}^k$ (see [3, eqs. (5), (6)]).[1] Thus, if a subsequence $\{\hat{z}^k\}_{k\in\mathcal{K}''}$ ($\mathcal{K}'' \subseteq \mathcal{K}'$) converges to some $\bar{z}$, then by further passing to a subsequence if necessary, we can assume that the index sets $\hat{A}^k$ and $\hat{B}^k$ are constant (i.e., $\hat{A}^k = \bar{A}$, $\hat{B}^k = \bar{B}$ for some $\bar{A}$, $\bar{B}$) for all $k \in \mathcal{K}''$. Since $\bar{z}$ satisfies the uniform LICQ, $\{(v_i^k)_{i\in\bar{A}}, (w_i^k)_{i\in\bar{B}}\}_{k\in\mathcal{K}''}$ also converges to some $(\bar{v}_i)_{i\in\bar{A}}, (\bar{w}_i)_{i\in\bar{B}}$.[2] By (2),

$$\bar{v}_i \geq 0 \quad \text{and} \quad \bar{w}_i \geq 0 \quad \forall\, i \in \bar{A} \cap \bar{B}.$$

This together with [3, eqs. (5), (6)] implies that $\bar{z}$ is an *M-stationary point* (see [4, 5] and (5) below). If in addition

$$\bar{A} \cap \bar{B} = A_0(\bar{z}) \cap B_0(\bar{z}), \tag{3}$$

then $\bar{z}$ is a B-stationary point of (1). In general, however, we can only assert that $\bar{A} \cap \bar{B} \subseteq A_0(\bar{z}) \cap B_0(\bar{z})$. This is the gap.

2. A modified ϵ-active set algorithm. We now describe a way, based on the active set identification approach of Facchinei, Fischer, and Kanzow [1], to modify the ϵ-active set algorithm so that (3) holds under an additional error bound condition. To simplify the notation, we will consider only the complementarity constraints, i.e., we assume $p = q = 0$ in (1). The general case can be treated analogously. The Lagrangian associated with (1) is

$$L(z, v, w) := f(z) + \sum_{i=1}^{m} (G_i(z)v_i + H_i(z)w_i).$$

We assume that there exists a computable continuous function $R : \Re^n \times \Re^m \times \Re^m \to [0, \infty)$ providing a local Hölder error bound at each M-stationary point $\bar{z}$ that is not B-stationary, i.e., there exist scalars $\tau > 0$, $\gamma > 0$, and $\delta > 0$ (depending on $\bar{z}$) such that

$$\|(z, v, w) - (\bar{z}, \bar{v}, \bar{w})\| \leq \tau R(z, v, w)^\gamma \quad \text{whenever } \|(z, v, w) - (\bar{z}, \bar{v}, \bar{w})\| \leq \delta, \tag{4}$$

where the multiplier vectors $\bar{v}, \bar{w}$ satisfy

$$\nabla_z L(\bar{z}, \bar{v}, \bar{w}) = 0, \ \left\{ \begin{array}{l} \bar{v}_i \perp G_i(\bar{z}) \geq 0 \\ \bar{w}_i \perp H_i(\bar{z}) \geq 0 \end{array} \right\}, \ G_i(\bar{z})H_i(\bar{z}) = 0, \ \left\{ \begin{array}{c} \bar{v}_i\bar{w}_i \geq 0 \\ \bar{v}_i \geq 0 \text{ or } \bar{w}_i \geq 0 \end{array} \right\} \ \forall i. \tag{5}$$

Here, $a \perp b$ means $ab = 0$. Due to uniform LICQ, $\bar{v}, \bar{w}$ are uniquely determined by $\bar{z}$. In fact, (5) characterizes M-stationarity for any $\bar{z} \in \Re^n$. We also assume that

$$R(\bar{z}, \bar{v}, \bar{w}) = 0 \quad \Longleftrightarrow \quad (\bar{z}, \bar{v}, \bar{w}) \text{ satisfies (5)}. \tag{6}$$

[1] Throughout, we use the same notation as [3].

[2] This follows from [3, eq. (6)], $\|r^k\|_1 \leq \delta_k \to 0$ (see [3, eq. (5)]), and the fact that if $b^k = C^k u^k$ for all k and $b^k \to b \in \Re^q$, $C^k \to C \in \Re^{q\times p}$ with C having linearly independent columns, then $u^k \to u \in \Re^p$ with u being the unique solution of $b = Cu$.

The "residual" function $R(z,v,w)$ can be constructed analogous to the NLP and NCP cases [1, 2]. In particular, consider

$$
\begin{aligned}
R(z,v,w) \quad &:= \quad \|\nabla_z L(z,v,w)\| + \sum_{i=1}^{m}\Big(|\min\{G_i(z),|v_i|\}| + |\min\{H_i(z),|w_i|\}| \\
&+ \quad |G_i(z)H_i(z)| + |\min\{0,v_i w_i\}| + |\min\{0,v_i\}\min\{0,w_i\}|\Big).
\end{aligned} \tag{7}
$$

Then, R is continuous and satisfies (6). Arguing as in the proof of Corollary 6.6.4 in [2], we have that the local error bound (4) holds if the M-stationary point $\bar{z}$ is isolated and f and ∇f are continuous and subanalytic (G and H, by being affine, are automatically continuous and subanalytic). A referee suggests that the assumption of $\bar{z}$ being isolated is benign when G and H are affine. In particular, it is readily shown that the M-stationary points of (1) are isolated if f is strictly convex on the null space of the active constraint gradients. Alternatively, it can be shown that the local error bound (4) holds with $\gamma = 1$ if a certain second-order sufficient condition holds at $\bar{z}$. This is a topic for further research.

Let $\theta : (0,\infty) \to (0,\infty)$ be any continuous nondecreasing function satisfying $\lim_{t\downarrow 0} t/\theta(t^\gamma) = 0$ for any $\gamma > 0$. An example is $\theta(t) = -C/\log(\min\{t, 0.9\})$ with $C > 0$. Using (4), (6) and following [1, 2], the function

$$\Theta(z,v,w) := \theta(R(z,v,w))$$

has the active set identification property that, for any M-stationary point $\bar{z}$ that is not B-stationary and corresponding multiplier vectors $\bar{v}, \bar{w}$, we have

$$\lim_{(z,v,w)\to(\bar{z},\bar{v},\bar{w})} \frac{G_i(z)}{\Theta(z,v,w)} = \begin{cases} 0 & \text{if } G_i(\bar{z}) = 0, \\ \infty & \text{if } G_i(\bar{z}) > 0, \end{cases}$$

and similarly with "G_i" replaced by "H_i."

Let us define

$$\bar{A}^k := \left\{ i \in \{1,\dots,m\} : \frac{G_i(\hat{z}^k)}{\Theta(\hat{z}^k, v^k, w^k)} \le 1 \right\},$$

$$\bar{B}^k := \left\{ i \in \{1,\dots,m\} : \frac{H_i(\hat{z}^k)}{\Theta(\hat{z}^k, v^k, w^k)} \le 1 \right\},$$

where the ith component of v^k is v_i^k if $i \in \hat{A}^k$ and is zero otherwise (and w^k is defined analogously). Since $(\hat{z}^k, v^k, w^k)$ satisfies [3, eqs. (4)–(6)], if (2) holds, then $R(\hat{z}^k, v^k, w^k)$ would tend to zero as $\hat{z}^k \to \bar{z}$ and $\epsilon_k, \delta_k, \nu_k$ tend to zero and, for $\hat{z}^k$ sufficiently near $\bar{z}$, we would have (v^k, w^k) sufficiently near $(\bar{v}, \bar{w})$ (due to [3, A2]) and

$$\bar{A}^k = A_0(\bar{z}), \quad \bar{B}^k = B_0(\bar{z}), \tag{8}$$

as well as

$$A_\epsilon(\hat{z}^k) \supseteq \bar{A}^k \supseteq \hat{A}^k, \qquad B_\epsilon(\hat{z}^k) \supseteq \bar{B}^k \supseteq \hat{B}^k, \tag{9}$$

where $\epsilon \ge 0$ is defined as in [3] (see page 727 therein).[3] Let

$$\bar{\epsilon}_k := \max\left\{ \epsilon_k, \max_{i\in\bar{A}^k} G_i(\hat{z}^k), \max_{i\in\bar{B}^k} H_i(\hat{z}^k) \right\}. \tag{10}$$

[3]The first containment in (9) holds whenever $\Theta(\hat{z}^k, v^k, w^k) \le \epsilon$, which in turn holds whenever $R(\hat{z}^k, v^k, w^k)$ is sufficiently small. By (8) and [3, eq. (7)], the second containment in (9) holds whenever $A_0(\bar{z}) \supseteq A_{\epsilon_k}(\hat{z}^k)$, which in turn holds whenever $\hat{z}^k$ is near $\bar{z}$ and ϵ_k is sufficiently small. The other two containments can be argued similarly.

Since $\bar{\epsilon}_k \geq \epsilon_k$, [3, eq. (4)] implies that $\hat{z}^k \in \mathcal{F}_{\bar{\epsilon}_k}[A^k, B^k]$ for all k. In fact, it can be seen that $\hat{z}^k$ remains an approximate KKT point of the subproblem [3, eq. (3)] (in the sense of [3, eqs. (4)–(6)]) when ϵ_k is replaced by $\bar{\epsilon}_k$ and $\hat{A}^k, \hat{B}^k$ are correspondingly replaced by $A_{\bar{\epsilon}_k}(\hat{z}^k), B_{\bar{\epsilon}_k}(\hat{z}^k)$. Thus, we can modify Step 2 of the ϵ-active set algorithm by possibly making this replacement when we are in case (c) and (9) holds.

THE MODIFIED $\boldsymbol{\epsilon}$-ACTIVE SET ALGORITHM FOR MPEC (1).

This is the same as the ϵ-active set algorithm in [3, pp. 730–731], except that when we are in case (c) in Step 2, we do the following: If

$$\text{(9) holds,} \quad \bar{A}^k \cap \bar{B}^k \neq \hat{A}^k \cap \hat{B}^k, \quad \bar{\epsilon}_k < \bar{\epsilon} \tag{11}$$

($\bar{\epsilon}$ is a threshold which initially can be any positive scalar below ϵ), then repeat Step 2 with ϵ_k replaced by $\bar{\epsilon}_k$ (and with $\hat{A}^k$, $\hat{B}^k$ redefined accordingly, i.e., they are replaced by $A_{\bar{\epsilon}_k}(\hat{z}^k), B_{\bar{\epsilon}_k}(\hat{z}^k)$ in Step 2, (9), (11)), and update $\bar{\epsilon} \leftarrow \bar{\epsilon}/2$. Otherwise, if $\epsilon_k \leq \epsilon_{\rm tol}$ and $\nu_k \leq \nu_{\rm tol}$, then terminate; otherwise, determine ν_{k+1} and $\tilde{z}^k$ by [3, eq. (14)], and proceed to Step 3.

If (11) holds, then $\epsilon_k < \bar{\epsilon}_k$,[4] which in turn implies $\bar{A}^k = A_{\bar{\epsilon}_k}(\hat{z}^k)$ and $\bar{B}^k = B_{\bar{\epsilon}_k}(\hat{z}^k)$.[5] Thus, when Step 2 is repeated, the second relation in (11) is violated.

THEOREM 2.1. *Under assumptions* [3, A1–A3], *the following results hold for the sequence* $\{(z^k, \hat{z}^k, \tilde{z}^k, \epsilon_k, \nu_k)\}$ *generated by the modified* ϵ*-active set algorithm, with* $\bar{\mathcal{K}} := \{k : \textit{at iteration } k, \textit{ Step 2 is repeated}\}$.

(a) *Suppose that each M-stationary point* $\bar{z}$ *of MPEC* (1) *that is not B-stationary satisfies* (4), *where* $(\bar{v}, \bar{w})$ *satisfies* (5) *and* R *satisfies* (6). *If* $\epsilon_0 > 0$, $\nu_0 > 0$, $\epsilon_{\rm tol} = \nu_{\rm tol} = 0$, f *is Lipschitz continuous with constant* L *on a set* Z *containing* $\{z^k\}$ *and* $\{\tilde{z}^k\}$, *and* $|\bar{\mathcal{K}}| < \infty$ *(respectively,* $|\bar{\mathcal{K}}| = \infty$*), then* $\epsilon_k \downarrow 0$, $\nu_k \downarrow 0$, *and every cluster point of* $\{\hat{z}^k\}$ *(respectively,* $\{\hat{z}^k\}_{k\in\bar{\mathcal{K}}}$*) is a B-stationary point of MPEC* (1).

(b) *If* $\epsilon_0 = \nu_0 = 0$ *and* f *is quadratic, then there exists a* $\bar{k} \in \{0, 1, \ldots\}$ *such that* $\hat{z}^{\bar{k}}$ *is a B-stationary point of MPEC* (1).

Proof. The first paragraph of the proof is identical to the proof of [3, Thm. 4.1], except we define $\mathcal{K} := \{k :$ We enter Step 3 from case (a) or (b) in Step 2 at iteration $k\}$ and $\mathcal{K}' := \{k :$ We enter Step 3 from case (c) in Step 2 at iteration $k\}$. The proof of (b) is identical to the proof of [3, Thm. 4.1(b)]. We prove (a) below.

(a) Suppose $\nu_k \to 0$. Then $|\mathcal{K}'| = \infty$, $\delta_k \to 0$, and the updating formulas for ϵ_k and $\bar{\epsilon}$ imply $\epsilon_k \to 0$, so any cluster point $\bar{z}$ of $\{\hat{z}^k\}_{k\in\mathcal{K}'}$ is an M-stationary point of MPEC (1). First, suppose $|\bar{\mathcal{K}}| < \infty$, so that $\bar{\epsilon} > 0$ is constant after a while. Let $\{\hat{z}^k\}_{k\in\mathcal{K}''}$ ($\mathcal{K}'' \subseteq \mathcal{K}'$) be any subsequence converging to $\bar{z}$. Since [3, eqs. (4)–(6)] and (2) hold for all $k \in \mathcal{K}''$, we have from [3, A2] and the same argument as in section 1 that $\{(v^k, w^k)\}_{k\in\mathcal{K}''} \to (\bar{v}, \bar{w})$ satisfying (5). By (6), $R(\bar{z}, \bar{v}, \bar{w}) = 0$. Since R is continuous, $\{R(\hat{z}^k, v^k, w^k)\}_{k\in\mathcal{K}''} \to 0$. If $\bar{z}$ is not B-stationary for (1), then the error bound (4) would hold and this would imply that (8) and (9) hold for all $k \in \mathcal{K}''$ sufficiently large. Moreover, $\{\bar{\epsilon}_k\}_{k\in\mathcal{K}''} \to 0$, so that $\bar{\epsilon}_k < \bar{\epsilon}$ for all $k \in \mathcal{K}''$ sufficiently

[4] If $\epsilon_k = \bar{\epsilon}_k$, then (10) and [3, eq. (7)] would imply $\bar{A}^k \subseteq \hat{A}^k$ and $\bar{B}^k \subseteq \hat{B}^k$, so (9) would yield $\bar{A}^k = \hat{A}^k$ and $\bar{B}^k = \hat{B}^k$, contradicting (11).

[5] Why? Since $\epsilon_k < \bar{\epsilon}_k$, we have from (10) and the definition of $\bar{A}^k$ and $\bar{B}^k$ that

$$\bar{\epsilon}_k = \max\left\{\max_{i\in\bar{A}^k} G_i(\hat{z}^k), \max_{i\in\bar{B}^k} H_i(\hat{z}^k)\right\} \leq \Theta(\hat{z}^k, v^k, w^k).$$

Thus, if $i \notin \bar{A}^k$, then $G_i(\hat{z}^k) > \Theta(\hat{z}^k, v^k, w^k) \geq \bar{\epsilon}_k$. By (10), if $i \in \bar{A}^k$, then $G_i(\hat{z}^k) \leq \bar{\epsilon}_k$. This shows that $\bar{A}^k = A_{\bar{\epsilon}_k}(\hat{z}^k)$. An analogous argument shows that $\bar{B}^k = B_{\bar{\epsilon}_k}(\hat{z}^k)$.

large. Thus, at each such iteration $k \in \mathcal{K}''$, we would have upon entering Step 3 that $\bar{A}^k \cap \bar{B}^k = \hat{A}^k \cap \hat{B}^k$ (since (11) must be violated). Then it would follow from (2) and (8) that $\bar{z}$ is a B-stationary point of (1), a contradiction. Second, suppose $|\bar{\mathcal{K}}| = \infty$. Then, as we discussed earlier, for each iteration $k \in \bar{\mathcal{K}}$, the second relation in (11) is violated upon entering Step 3, i.e., $\bar{A}^k \cap \bar{B}^k = \hat{A}^k \cap \hat{B}^k$. Then, an argument similar to the one above shows that every cluster point $\bar{z}$ of $\{\hat{z}^k\}_{k \in \bar{\mathcal{K}}}$ is a B-stationary point of (1).

Suppose instead $\nu_k \not\to 0$, so that $|\mathcal{K}'| < \infty$, $|\mathcal{K}| = \infty$, and $\nu = \lim_{k \to \infty} \nu_k > 0$. The remainder of the proof is identical to the proof of [3, Thm. 4.1(a)], except that, due to ϵ_k being replaced by $\bar{\epsilon}_k$ in Step 2 for all iterations $k \in \bar{\mathcal{K}}$, instead of [3, eq. (22)] we have

$$f(z^{k+1}) \le f(\tilde{z}^k) + 2L\tau m(\epsilon_k - \epsilon_{k+1} + \Delta_k) \quad \forall\, k,$$

where $\Delta_k := \bar{\epsilon}_k$ if $k \in \bar{\mathcal{K}}$ and $\Delta_k := 0$ otherwise. Since (11) holds at each iteration $k \in \bar{\mathcal{K}}$ and $\bar{\epsilon}$ is halved at each such iteration, it follows that $\sum_{k=0}^{\infty} \Delta_k = \sum_{k \in \bar{\mathcal{K}}} \bar{\epsilon}_k < \infty$. Then it can be argued similarly as in the proof of [3, Thm. 4.1(a)] that $\{f(z^k)\}$ converges and so on. □

We illustrate the assumptions of Theorem 2.1 with the following example of (1):

$$\text{minimize} \quad f(z) \quad \text{subject to} \quad z_1 \ge 0,\ z_2 \ge 0,\ z_1 z_2 = 0.$$

This example satisfies assumption [3, A2] for any $\epsilon \ge 0$. If $f(z) = (z_2)^p$ $(p \ge 1)$, then assumption [3, A1] also holds and each M-stationary point, which is of the form $(\bar{z}_1, 0)$ with $\bar{z}_1 \ge 0$, is B-stationary. If $f(z) = z_1^4 + z_2^2 - z_2$, then assumptions [3, A1, A3] also hold and the M-stationary points, $(0,0)$ and $(0, \frac{1}{2})$, are isolated with $(0, \frac{1}{2})$ B-stationary. For R given by (7), the error bound (4) holds at $(0,0)$. However, if $f(z) = z_2^2 - z_2$, then the M-stationary point $\bar{z} = (0,0)$, with multipliers $\bar{v} = 0, \bar{w} = -1$, is not B-stationary and is not isolated. Moreover, for any continuous R satisfying (6), the error bound (4) does not hold at $(0,0)$. This is because, for any fixed $\delta > 0$, $(\delta, 0)$ is M-stationary with multipliers $v = 0, w = -1$, so $R((\delta, x_2), 0, -1) \to R(\delta, 0), 0, -1) = 0$ as $x_2 \to 0$. But $\|((\delta, x_2), 0, -1) - ((0,0), 0, -1)\| \to \delta$ as $x_2 \to 0$.

Acknowledgments. The authors thank Lifeng Chen for notifying them of the gap in the proof of [3, Thm. 4.1]. They also thank two referees for their helpful comments.

REFERENCES

[1] F. Facchinei, A. Fischer, and C. Kanzow, *On the accurate identification of active constraints*, SIAM J. Optim., 9 (1998), pp. 14–32.

[2] F. Facchinei and J.-S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Vol. II, Springer-Verlag, New York, 2003.

[3] M. Fukushima and P. Tseng, *An implementable active-set algorithm for computing a B-stationary point of a mathematical program with linear complementarity constraints*, SIAM J. Optim., 12 (2002), pp. 724–739.

[4] J. Outrata, *Optimality conditions for a class of mathematical programs with equilibrium constraints*, Math. Oper. Res., 24 (1999), pp. 627–644.

[5] H. Scheel and S. Scholtes, *Mathematical programs with complementarity constraints: Stationarity, optimality, and sensitivity*, Math. Oper. Res., 25 (2000), pp. 1–22.