SIAM J. OPTIM.
Vol. 19, No. 3, pp. 1073–1092
© 2008 Society for Industrial and Applied Mathematics

ON THE SECOND-ORDER FEASIBILITY CONE: PRIMAL-DUAL REPRESENTATION AND EFFICIENT PROJECTION*

ALEXANDRE BELLONI† AND ROBERT M. FREUND‡

Abstract. We study the second-order feasibility cone $\mathcal{F} = \{y \in \mathbb{R}^n : \|My\| \le g^T y\}$ for given data (M, g). We construct a new representation for this cone and its dual based on the spectral decomposition of the matrix $M^T M - gg^T$. This representation is used to efficiently solve the problem of projecting an arbitrary point $x \in \mathbb{R}^n$ onto $\mathcal{F}$: $\min_y\{\|y - x\| : \|My\| \le g^T y\}$, which aside from theoretical interest also arises as a necessary subroutine in the rescaled perceptron algorithm. We develop a method for solving the projection problem to an accuracy ε, whose computational complexity is bounded by $O(mn^2 + n \ln\ln(1/\varepsilon) + n \ln\ln(1/\min\{\text{width}(\mathcal{F}), \text{width}(\mathcal{F}^*)\}))$ operations. Here $\text{width}(\mathcal{F})$ and $\text{width}(\mathcal{F}^*)$ denote the width of $\mathcal{F}$ and $\mathcal{F}^*$, respectively. We also perform computational tests that indicate that the method is extremely efficient in practice.

Key words. second-order cone, convex cone, projection, computational complexity, Newton method

AMS subject classifications. 90C60, 90C51, 90C25, 49M15, 49M29

DOI. 10.1137/06067198X

1. Introduction and main results. Our notation is as follows: let K^* denote the dual of a convex cone $K \subset \mathbb{R}^k$, i.e., $K^* := \{z \in \mathbb{R}^k : z^T y \ge 0 \text{ for all } y \in K\}$. A convex cone K is *regular* if it is closed, has nonempty interior, and contains no lines, in which case K^* is also regular; see Rockafellar [7]. Define the standard second-order cone in $\mathbb{R}^k$ to be $\mathcal{Q}^k := \{y \in \mathbb{R}^k : \|(y_1, \ldots, y_{k-1})\| \le y_k\}$, where $\|\cdot\|$ denotes the Euclidean norm. Let $B(y, r)$ denote the Euclidean ball of radius r centered at y.

Given data $(M, g) \in (\mathbb{R}^{m\times n}, \mathbb{R}^n)$, our interest lies in the second-order feasibility cone

$$\mathcal{F} := \left\{y \in \mathbb{R}^n : \|My\| \le g^T y\right\} = \left\{y \in \mathbb{R}^n : \left(My, g^T y\right) \in \mathcal{Q}^{m+1}\right\}$$

and its dual cone $\mathcal{F}^*$.

We first take care of some trivial cases. When $\text{rank}(M) = 0$, it follows trivially that $\mathcal{F} = \{y \in \mathbb{R}^n : g^T y \ge 0\}$, whereby $\mathcal{F}$ is either a half-space or all of $\mathbb{R}^n$, depending on whether $g \ne 0$ or $g = 0$, respectively. When $\text{rank}(M) = 1$, $M = fc^T$ for some f, c, and $\|My\| = \|f\|\|c^T y|$ for any y. This implies that $\mathcal{F} = \{y \in \mathbb{R}^n : (g - \|f\|c)^T y \ge 0,\ (g + \|f\|c)^T y \ge 0\}$, and hence $\mathcal{F}$ is the intersection of either one or two half-spaces. We dispose of these trivial cases by making the following assumption about the data.

Assumption 1. $\text{rank}(M) \ge 2$ and $g \ne 0$.

We now describe our main representation result for $\mathcal{F}$ and $\mathcal{F}^*$. It is elementary to establish that $M^T M - gg^T$ has at most one negative eigenvalue, and we can write its eigendecomposition as $M^T M - gg^T = QDQ^T$, where Q is orthonormal ($Q^{-1} = Q^T$) and D is the diagonal matrix of eigenvalues. For notational convenience we denote D_i and Q_i as the ith diagonal component of D and the ith column of Q, respectively. By

*Received by the editors October 10, 2006; accepted for publication (in revised form) May 7, 2008; published electronically October 31, 2008. This research has been partially supported through the Singapore-MIT Alliance and an IBM Ph.D. Fellowship.

http://www.siam.org/journals/siopt/19-3/67198.html

†Fuqua School of Business, Duke University, Durham, NC 27708-0120 (abn5@duke.edu).

‡MIT Sloan School of Management, Cambridge, MA 02142 (rfreund@mit.edu).

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

reordering the columns of Q, we can presume that $D_1 \geq \cdots \geq D_n$ and $D_1, \ldots, D_{n-1} \geq 0$. By choosing either Q_n or $-Q_n$, we can further presume that $g^T Q_n \geq 0$. We implicitly assume Q and D can be computed to within machine precision (in the relative sense) in $O(mn^2)$ operations, consistent with computational practice.

Our interest lies mainly in the case when $\mathcal{F}$ is a regular cone, so we will hypothesize that $\mathcal{F}$ is a regular cone for the remainder of this section. This hypothesis implies that $n-1 \leq \text{rank}(M) \leq \min\{n, m\}$. (We indicate how to amend our results and proofs to relax this hypothesis at the end of sections 2 and 3.) Our main representation result is as follows.

THEOREM 1. *Suppose that $\mathcal{F}$ is a regular cone. Then $D_1, \ldots, D_{n-1} > 0 > D_n$, and*

(i) $\mathcal{F} = \{y : y^T Q D Q^T y \leq 0,\ y^T Q_n \geq 0\}$;
(ii) $\mathcal{F}^* = \{z : z^T Q D^{-1} Q^T z \leq 0,\ z^T Q_n \geq 0\}$;
(iii) *if $y \in \mathcal{F}$ and $\alpha \geq 0$, then $z := -\alpha Q D Q^T y \in \mathcal{F}^*$. Furthermore, if $y \in \partial\mathcal{F}$, then $z \in \partial\mathcal{F}^*$ and $z^T y = 0$;*
(iv) *if $z \in \mathcal{F}^*$ and $\alpha \geq 0$, then $y := -\alpha Q D^{-1} Q^T z \in \mathcal{F}$. Furthermore, if $z \in \partial\mathcal{F}^*$, then $y \in \partial\mathcal{F}$ and $z^T y = 0$.*

Note that (i) and (ii) of Theorem 1 describe easily computable representations of $\mathcal{F}$ and $\mathcal{F}^*$ that have the same computational structure, in that checking membership in each cone uses similar data, operations, etc., in a manner that is symmetric between the dual cones. Parts (iii) and (iv) indicate that the same matrices in (i) and (ii) can be used constructively to map points on the boundary of one cone to their orthogonal counterpart in the dual cone.

Remark 1 (geometry of $\mathcal{F}$ and $\mathcal{F}^*$). Examining (i) and the property that $D_n < 0$, the orthonormal transformation $y \to s := Q^T y$ maps $\mathcal{F}$ onto the axes-aligned ellipsoidal cone $\mathcal{S} := \{s \in \mathbb{R}^n : \sqrt{\sum_{j=1}^{n-1} D_i s_i^2} \leq \sqrt{|D_n|} s_n\}$ so that $\mathcal{F}$ is the image of $\mathcal{S}$ under Q, $\mathcal{F} = \{y : \sqrt{\sum_{i=1}^{n-1} D_i (Q_i^T y)^2} \leq \sqrt{|D_n|} Q_n^T y\}$, and $\mathcal{F}^* = \{z : \sqrt{\sum_{i=1}^{n-1} (1/D_i)(Q_i^T z)^2} \leq \sqrt{1/|D_n|} Q_n^T z\}$. This establishes that $\mathcal{F}$ is indeed simply an ellipsoidal cone whose axes are the eigenvectors of Q with dilations corresponding to the eigenvalues of $M^T M - g g^T$. From this perspective, the representation of $\mathcal{F}^*$ via (ii) makes natural geometric sense. Also, the central axis of both $\mathcal{F}$ and $\mathcal{F}^*$ is the ray $\{\alpha Q_n : \alpha \geq 0\}$. Last of all, note that $-\mathcal{F} = \{y : y^T Q D Q^T y \leq 0,\ y^T Q_n \leq 0\}$ and $-\mathcal{F}^* = \{z : z^T Q D^{-1} Q^T z \leq 0,\ z^T Q_n \leq 0\}$.

It turns out that the eigendecomposition of $M^T M - g g^T = Q D Q^T$, while useful both conceptually and algorithmically (as we shall see), is not even necessary for the above representation of $\mathcal{F}$ and $\mathcal{F}^*$. Indeed, Theorem 1 can alternatively be stated replacing $Q D Q^T$ and $Q D^{-1} Q^T$ by $M^T M - g g^T$ and $(M^T M - g g^T)^{-1}$. Under the further hypothesis that $\text{rank}(M) = n$, the theorem can be restated as follows.

COROLLARY 1. *Suppose that $\mathcal{F}$ is a regular cone and* $\text{rank}(M) = n$. *Then*

(i) $\mathcal{F} = \{y : \sqrt{y^T (M^T M) y} \leq g^T y\}$;
(ii) $\mathcal{F}^* = \{z : \sqrt{z^T (M^T M)^{-1} z} \leq \frac{g^T (M^T M)^{-1} z}{\sqrt{g^T (M^T M)^{-1} g - 1}}\}$;
(iii) *if $y \in \mathcal{F}$ and $\alpha \geq 0$, then $z := -\alpha (M^T M - g g^T) y \in \mathcal{F}^*$. Furthermore, if $y \in \partial\mathcal{F}$, then $z \in \partial\mathcal{F}^*$ and $z^T y = 0$;*
(iv) *if $z \in \mathcal{F}^*$ and $\alpha \geq 0$, then $y := -\alpha [(M^T M)^{-1} - \frac{(M^T M)^{-1} g g^T (M^T M)^{-1}}{g^T (M^T M)^{-1} g - 1}] z \in \mathcal{F}$. Furthermore, if $z \in \partial\mathcal{F}^*$, then $y \in \partial\mathcal{F}$ and $z^T y = 0$.*

The proofs of Theorem 1 and Corollary 1 are presented in section 2, along with proofs that all of the stated quantities are well defined: in particular, D^{-1} exists and $g^T (M^T M)^{-1} g - 1 > 0$ under the given hypotheses.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

These representation results are used to solve the following dual pair of optimization problems, where $x \in \mathbb{R}^n$ is a *given* point:

$$\begin{array}{lll} \mathcal{P}: \ t^* := \min_y & \|y - x\| & \mathcal{D}: \ t^* := \max_z \quad -x^T z \\ \text{s.t.} & y \in \mathcal{F}, & \text{s.t.} \quad \|z\| \le 1 \\ & & \quad\ z \in \mathcal{F}^*. \end{array} \tag{1}$$

The problem $\mathcal{P}$ is the classical projection problem onto the cone $\mathcal{F}$, whose solution is the point in $\mathcal{F}$ closest to x, and strong duality is easily established for this pair of problems. The problem $\mathcal{D}$ arises as a necessary subroutine in the rescaled perceptron algorithm in [2]: the subroutine needs to efficiently solve $\mathcal{D}$ using $x = x^k$ that arises at each outer iteration k of the algorithm. It is this latter problem that motivated our interest in efficiently representing $\mathcal{F}^*$ and solving both $\mathcal{P}$ and $\mathcal{D}$. Notice that $\mathcal{P}/\mathcal{D}$ involve intersections of a Euclidean ball and a second-order feasibility cone. This dual pair of problems is therefore a modest generalization of the trust region problem of optimizing a quadratic function over a Euclidean ball, for which Ye [10] showed how to combine a binary search and Newton's method to obtain double-logarithmic complexity. Using the representation results above and extending ideas from [10], we develop an algorithm for solving (1) in section 3. The complexity of the algorithm depends on the *widths* of the cones $\mathcal{F}$ and $\mathcal{F}^*$, where the width τ_K of a cone K is defined to be the radius of the largest ball contained in K that is centered at unit distance from the origin:

$$\tau_K := \max_{y,r} \{r : B(y, r) \subset K, \ \|y\| \le 1\}.$$

It readily follows from Theorem 1 that the widths of $\mathcal{F}$ and $\mathcal{F}^*$ are simple functions of the largest and smallest positive eigenvalues and the negative eigenvalue of $M^T M - gg^T$, and it is straightforward to derive the following:

$$\tau_{\mathcal{F}} = \sqrt{\frac{|D_n|}{|D_n| + D_1}} \quad \text{and} \quad \tau_{\mathcal{F}^*} = \sqrt{\frac{1/|D_n|}{1/|D_n| + 1/D_{n-1}}}.$$

The main complexity result, which is proved in section 3, is as follows.

THEOREM 2. *Suppose that $\mathcal{F}$ is a regular cone, and $x \in \mathbb{R}^n$ satisfying $\|x\| = 1$ is given. Then feasible solutions (y, z) of $(\mathcal{P}, \mathcal{D})$ satisfying a duality gap of at most σ are computable in $O(mn^2 + n \ln\ln(1/\sigma) + n \ln\ln(1/\min\{\tau_{\mathcal{F}}, \tau_{\mathcal{F}^*}\}))$ operations.*

In section 4, we complement this theoretical computational complexity bound with experimental computational results that indicate that the method is also extremely efficient in practice.

Last of all, we note that the hypothesis that $\mathcal{F}$ is regular can be removed with no loss of strength of the results herein but with substantial expositional overhead. The case when $\mathcal{F}$ is nonregular is discussed at the end of sections 2 and 3.

2. Proofs of representation results. Recall the eigendecomposition of $M^T M - gg^T = QDQ^T$, with $D_1 \ge \cdots \ge D_n$. A simple dimension argument establishes that $M^T M - gg^T$ has at most one negative eigenvalue, whereby $D_1, \ldots, D_{n-1} \ge 0$. By choosing either Q_n or $-Q_n$, we can ensure that $g^T Q_n \ge 0$. In preparation for the proof of Theorem 1, we first prove some preliminary results.

PROPOSITION 1. *Suppose that* **int** $\mathcal{F} \ne \emptyset$. *Then $D_n < 0$, and there exists y satisfying $\|My\| < g^T y$.*

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Proof. We first suppose that there exists $\bar{y}$ that satisfies $\|M\bar{y}\| < g^T\bar{y}$. In this case it easily follows that $0 > \bar{y}^T(M^TM - gg^T)\bar{y} = \bar{y}^TQDQ^T\bar{y}$, whereby $D_n < 0$. Next suppose that every $y \in \mathcal{F}$ satisfies $\|My\| = g^Ty$, and let $\bar{y} \in \mathbf{int}\ \mathcal{F}$. Since $\bar{y} \in \mathbf{int}\ \mathcal{F}$, we have $\|M(\bar{y} + \beta d)\| = g^T(\bar{y} + \beta d)$ for all $d \in B(0,1)$ and all sufficiently small positive β. Squaring the previous equation, then rearranging and cancelling terms yields $2(d^TM^TM\bar{y} - \bar{y}^Tgg^Td) + \beta(d^TM^TMd - d^Tgg^Td) = 0$, which is true only if $g^Td = 0 \Rightarrow Md = 0$. This in turn implies that $\text{rank}(M) = 1$, violating Assumption 1. Therefore there exists y satisfying $\|My\| < g^Ty$. ∎

One characterization of $\mathcal{F}^*$ is as follows:

$$\mathcal{F}^* = \mathbf{cl}\ \left\{M^T\lambda + g\alpha : \|\lambda\| \le \alpha\right\}. \tag{2}$$

This result admits an elementary proof by a separating hyperplane argument and has been part of the folklore of convex analysis for several decades. For a standard proof, see, for example, Theorem 3.1 of Berman [3] applied to the second-order cone. The lack of closure of $\mathcal{T} := \{M^T\lambda + g\alpha : \|\lambda\| \le \alpha\}$ can arise easily. Let $M = \left[\begin{smallmatrix} -1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$ and $g = \left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right]$. In this case, $\mathcal{T} = \{(-\lambda_1 + \alpha, \lambda_2) \mid \|(\lambda_1, \lambda_2)\| \le \alpha\}$. It is easy to verify that $(0,1) \notin \mathcal{T}$ but $(\varepsilon, 1) \in \mathcal{T}$ for every $\varepsilon > 0$ (set $\lambda_1 = \frac{1}{2\varepsilon} - \frac{\varepsilon}{2}$, $\lambda_2 = 1$, and $\alpha = \frac{1}{2\varepsilon} + \frac{\varepsilon}{2}$), which shows that $\mathcal{T}$ is not closed. For an analysis of cases when $\mathcal{T}$ is guaranteed to be closed, see Pataki [5].

Proof of Theorem 1. Since $\mathbf{int}\ \mathcal{F} \neq \emptyset$, Proposition 1 implies that $D_n < 0$, and so, for the sake of this proof, we rescale (M, g) by $1/\sqrt{|D_n|}$ in order to conveniently satisfy $D_n = -1$. (i) Define $\mathcal{H} := \{y : y^TQDQ^Ty \le 0,\ y^TQ_n \ge 0\}$ and $\mathcal{L} := \{y : y^TQDQ^Ty \le 0,\ y^Tg \ge 0\}$. We need to prove that $\mathcal{H} = \mathcal{F}$. It is straightforward to check that $\mathcal{F} = \mathcal{L}$ and, indeed, $\mathbf{int}\ \mathcal{F} = \mathbf{int}\ \mathcal{L} = \{y : y^TQDQ^Ty < 0,\ y^Tg > 0\}$ from Proposition 1, and this also readily establishes that $Q_n \in \mathbf{int}\ \mathcal{F}$. For any y we can write

$$y^TQ_n = y^T\left(-QDQ^T\right)Q_n = y^T\left(gg^T - M^TM\right)Q_n = \left(My, g^Ty\right)^T\left(-MQ_n, g^TQ_n\right), \tag{3}$$

where the final term's two parenthetic vectors lie in $\mathbb{R}^{m+1}$. Notice that $(-MQ_n, g^TQ_n) \in \mathbf{int}\ \mathcal{Q}^{m+1}$, since $Q_n \in \mathbf{int}\ \mathcal{F}$. If $y \in \mathcal{F}$, then both vectors in the last term of (3) are in $\mathcal{Q}^{m+1}$; hence $y^TQ_n \ge 0$ follows from the self-duality of $\mathcal{Q}^{m+1}$. Therefore $y \in \mathcal{H}$, showing that $\mathcal{F} \subset \mathcal{H}$. Next suppose that $y \in \mathcal{H}$. Then $y \in \mathcal{F}$ unless $g^Ty < 0$, in which case $-y \in \mathcal{F}$. Using (3) we have

$$0 \le y^TQ_n = \left(My, g^Ty\right)^T\left(-MQ_n, g^TQ_n\right) = -\left(-My, -g^Ty\right)^T\left(-MQ_n, g^TQ_n\right) \le 0,$$

again because the two vectors in the last term lie in the self-dual cone $\mathcal{Q}^{m+1}$. This implies that equality holds throughout, and hence $(-My, -g^Ty) = (0,0)$ since $(-MQ_n, g^TQ_n) \in \mathbf{int}\ \mathcal{Q}^{m+1}$, yielding the contradiction that $g^Ty = 0$. This establishes that $\mathcal{F} \subset \mathcal{H}$, completing the proof of (i).

(ii) Having established (i), suppose that $D_i = 0$ for some $i \in \{1, \ldots, n-1\}$. Then $(\theta Q_i)^TQDQ^T(\theta Q_i) = 0$ and $Q_n^T(\theta Q_i) = 0$, whereby $\theta Q_i \in \mathcal{F}$ for all θ, violating the hypothesis that $\mathcal{F}$ is regular. Therefore $D_i > 0$ for all $i \in \{1, \ldots, n-1\}$, and hence D^{-1} exists. Define $\mathcal{J} := \{z : z^TQD^{-1}Q^Tz \le 0,\ z^TQ_n \ge 0\}$. Suppose that $z \in \mathcal{J}$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

and $y \in \mathcal{F}$, in which case

$$y^T z = y^T Q Q^T z = \sum_{i=1}^{n-1} D_i^{\frac{1}{2}} \left(Q_i^T y\right) D_i^{-\frac{1}{2}} \left(Q_i^T z\right) + y^T Q_n z^T Q_n$$
$$\geq -\sqrt{\sum_{i=1}^{n-1} D_i \left(Q_i^T y\right)^2} \sqrt{\sum_{i=1}^{n-1} D_i^{-1} \left(Q_i^T z\right)^2} + y^T Q_n z^T Q_n \geq 0,$$

where the first inequality is an application of the Cauchy–Schwarz inequality and the second inequality follows, since $z \in \mathcal{J}$ and $y \in \mathcal{F}$ using part (i). Thus $z \in \mathcal{F}^*$, which shows that $\mathcal{J} \subset \mathcal{F}^*$. Next let $\bar{Q}$ denote the matrix of the first $n-1$ columns of Q, and let $\bar{D}$ denote the diagonal matrix composed of the $n-1$ diagonal components $D_1, \ldots, D_{n-1}$. Then from part (i) we have $\mathcal{F} = \{y : \sqrt{y^T \bar{Q} \bar{D} \bar{Q}^T y} \leq Q_n^T y\} = \{y : \|\bar{D}^{\frac{1}{2}} \bar{Q}^T y\| \leq Q_n^T y\}$, and using (2) we know that $\mathcal{F}^* = \mathbf{cl}\, \mathcal{T}$, where $\mathcal{T} = \{\bar{Q}\bar{D}^{\frac{1}{2}}\lambda + Q_n \alpha : \|\lambda\| \leq \alpha\}$. Let $z \in \mathcal{T}$, where $z = \bar{Q}\bar{D}^{\frac{1}{2}}\lambda + Q_n\alpha$ and $\|\lambda\| \leq \alpha$. Then

$$z^T Q D^{-1} Q^T z = \left(\bar{Q}\bar{D}^{\frac{1}{2}}\lambda + Q_n\alpha\right)^T Q D^{-1} Q^T \left(\bar{Q}\bar{D}^{\frac{1}{2}}\lambda + Q_n\alpha\right) = \lambda^T\lambda - \alpha^2 \leq 0,$$

and furthermore $Q_n^T z = \alpha \geq 0$, whereby $z \in \mathcal{J}$. Thus $\mathcal{T} \subset \mathcal{J}$. It then follows that $\mathcal{F}^* = \mathbf{cl}\, \mathcal{T} \subset \mathbf{cl}\, \mathcal{J} = \mathcal{J}$, which completes the proof of (ii).

To prove (iii), notice that $Q_n^T z = -\alpha D_n Q_n^T y \geq 0$ and

$$z^T Q D^{-1} Q^T z = \alpha^2 y^T Q D Q^T Q D^{-1} Q^T Q D Q^T y = \alpha^2 y^T Q D Q^T y \leq (=)\ 0,$$

since $y \in \mathcal{F}$ ($y \in \partial\mathcal{F}$) implies that $y^T Q D Q^T y \leq (=)\ 0$, and hence $z \in \mathcal{F}^*$ ($z \in \partial\mathcal{F}^*$) from part (ii). Furthermore $y^T z = -\alpha y^T Q D Q^T y = 0$ when $y \in \partial\mathcal{F}$, completing the proof of (iii). The proof of (iv) follows similar logic. ∎

Before proving Corollary 1 we first prove the following.

PROPOSITION 2. *Suppose that* $\mathbf{int}\ \mathcal{F} \neq \emptyset$ *and* $\operatorname{rank}(M) = n$. *Then* $g^T(M^TM)^{-1}g > 1$ *and* $\bar{y} := (M^TM)^{-1}g \in \mathbf{int}\ \mathcal{F}$.

Proof. Let $\alpha := g^T(M^TM)^{-1}g > 0$, since $g \neq 0$ from Assumption 1. From Proposition 1 we know there exists $\hat{y}$ satisfying $\|M\hat{y}\| < g^T\hat{y}$ and rescale $\hat{y}$ if necessary so that $g^T\hat{y} = \alpha$. Notice that $\bar{y}$ optimizes the function $f(y) = y^TM^TMy - 2g^Ty$, whose optimal objective function value is $-\alpha$. Therefore

$$-\alpha \leq \hat{y}^TM^TM\hat{y} - 2g^T\hat{y} < \alpha^2 - 2\alpha,$$

which implies that $\alpha^2 > \alpha > 0$, and hence $\alpha > 1$. Next observe that $\|M\bar{y}\| = \sqrt{\bar{y}^TM^TM\bar{y}} = \sqrt{\alpha} < \alpha = g^T\bar{y}$, whereby $\bar{y} \in \mathbf{int}\ \mathcal{F}$. ∎

Proof of Corollary 1. (i) is a restatement of the definition of $\mathcal{F}$, (iii) is a restatement of part (iii) of Theorem 1, and (iv) is a restatement of part (iv) of Theorem 1 using the Sherman–Morrison formula

$$QD^{-1}Q^T = \left(M^TM - gg^T\right)^{-1} = \left(M^TM\right)^{-1} - \frac{(M^TM)^{-1}gg^T\left(M^TM\right)^{-1}}{g^T\left(M^TM\right)^{-1}g - 1},$$

together with the fact from Proposition 2 that $g^T(M^TM)^{-1}g > 1$.

It remains to prove (ii). Let $\mathcal{K} := \{z \in \mathbb{R}^n : z^TQD^{-1}Q^Tz \leq 0\}$. Then from Theorem 1 we have $\mathcal{K} = \mathcal{F}^* \cup -\mathcal{F}^*$. Let $\bar{y} = (M^TM)^{-1}g$, and note that $\bar{y} \in \mathbf{int}\ \mathcal{F}$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

from Proposition 2. Define $\mathcal{H} := \{z \in \mathbb{R}^n : \bar{y}^T z \geq 0\}$, and note that $\mathcal{H} \cap \mathcal{F}^* = \mathcal{F}^*$ and $\mathcal{H} \cap -\mathcal{F}^* = \{0\}$. Therefore $\mathcal{F}^* = \mathcal{K} \cap \mathcal{H} = \{z \in \mathbb{R}^n : z^T Q D^{-1} Q^T z \leq 0,\ g^T (M^T M)^{-1} z \geq 0\}$. Using the Sherman–Morrison formula we obtain

$$\mathcal{F}^* = \left\{ z^T \left((M^T M)^{-1} - \frac{\left(M^T M\right)^{-1} g g^T \left(M^T M\right)^{-1}}{g^T \left(M^T M\right)^{-1} g - 1} \right) z \leq 0,\ g^T \left(M^T M\right)^{-1} z \geq 0 \right\},$$

which after rearranging yields the expression in (ii). ∎

Remark 2 (the case when $\mathcal{F}$ is not regular). Let Z and N partition the set of indices according to zero and nonzero values of D_i. If $D_n = 0$, then one can show that $\mathcal{F}$ is a half-subspace in the subspace spanned by the Q_i for $i \in Z$. If $D_n > 0$, then $\mathcal{F} = \{0\}$. If $D_n < 0$, then $\mathcal{F}$ has an interior, and we can interpret $D_i^{-1} = \infty$ for $i \in Z$. Then Theorem 1 remains valid if we interpret "$z^T Q D^{-1} Q^T z \leq 0$" in (ii) as "$\sum_{i \in N} D_i (Q^T z)_i^2 \leq 0,\ (Q^T z)_i^2 = 0$ for $i \in Z$," and "$y := -\alpha Q D^{-1} Q^T z$" in (iv) as "$Q_i^T y := -\alpha D_i^{-1} Q_i^T z$ for $i \in N$ and $Q_i^T y$ is set *arbitrarily* for $i \in N$."

3. An algorithm for approximately solving (1).

3.1. Basic properties of (1) and the polar problem pair. Returning to (1) where x is the given vector, consider the following conditions in (y, z, θ):

$$\begin{array}{l} y - \theta z = x, \\ y \in \mathcal{F}, \\ z \in \mathcal{F}^*, \\ \|z\| \leq 1, \\ \theta \geq 0,\ \theta \|z\| = \theta. \end{array} \tag{4}$$

Examining (4), we see that x is decomposed into $x = y - \theta z$, where $y \in \mathcal{F}$ and $-\theta z \in -\mathcal{F}^*$ and (y, z) is feasible for the problems (1). Let G denote the duality gap for (1), namely, $G = \|y - x\| + x^T z$. We also consider the following pair of conic problems that are "polar" to (1):

$$\begin{array}{lll} \mathcal{P}^\circ:\ f^* := \min_v & \|v - x\| & \mathcal{D}^\circ:\ f^* := \max_w \quad -x^T w \\ \text{s.t.} & v \in -\mathcal{F}^*, & \text{s.t.} \quad \|w\| \leq 1 \\ & & \qquad\ w \in -\mathcal{F}, \end{array} \tag{5}$$

together with the following conditions in (v, w, ρ):

$$\begin{array}{l} v - \rho w = x, \\ v \in -\mathcal{F}^*, \\ w \in -\mathcal{F}, \\ \|w\| \leq 1, \\ \rho \geq 0,\ \rho \|w\| = \rho; \end{array} \tag{6}$$

here x is decomposed into $x = v - \rho w$, where now (v, w) is feasible for the problems (5), $-\rho w \in \mathcal{F}$, and $v \in -\mathcal{F}^*$. Let G° denote the duality gap for (5), namely, $G^\circ = \|v - x\| + x^T w$.

It is a straightforward exercise to show that conditions (4) together with the complementarity condition $y^T z = 0$ constitute necessary and sufficient optimality conditions for (1), and similarly, (6) together with $v^T w = 0$ are necessary and sufficient for optimality for (5). Furthermore, the solutions of (4) and (6) transform to one another:

$$\begin{array}{rcl} (y, z, \theta) & \rightarrow & (v, w, \rho) = (-\theta z, -y/\|y\|, \|y\|), \\ (v, w, \rho) & \rightarrow & (y, z, \theta) = (-\rho w, -v/\|v\|, \|v\|), \end{array}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

with necessary modifications for the cases when $y = 0$ (set $w = 0$) and/or $v = 0$ (set $z = 0$).

PROPOSITION 3. *Suppose* (y, z, θ) *satisfy* (4) *and* (v, w, ρ) *satisfy* (6). *Then* (y, z) *and* (v, w) *are feasible for their respective problems with respective duality gaps:*

(i) $G = y^T z$;
(ii) $G^\circ = v^T w$.

Furthermore,

(iii) *if* (y, z) *is optimal for* (1), *then* $t^* = \theta$;
(iv) *if* (v, w) *is optimal for* (5), *then* $f^* = \rho$;
(v) $(t^*)^2 + (f^*)^2 = \|x\|^2$.

Proof. To prove (i), observe that $y^T z = z^T x + \theta\|z\|^2 = z^T x + \theta\|z\| = z^T x + \|y - x\| = G$, and a similar argument establishes (ii). To prove (iii), observe that $t^* = \|x - y\| = \|\theta z\| = \theta$ with similar arguments for (iv). To prove (v), notice that (y, z, θ) satisfy (4), and $y^T z = 0$ if and only if (y, z) is optimal for (1), in which case it is easy to verify that $(v, w, \rho) \leftarrow (-\theta z, -y/\|y\|, \|y\|)$ satisfy (6) and (v, w) is optimal for (5). Therefore $\|x\|^2 = (y - \theta z)^T(y - \theta z) = y^T y + \theta^2 = \rho^2 + \theta^2 = (f^*)^2 + (t^*)^2$. ∎

PROPOSITION 4. *If* $Q_n^T x \le 0$, *then* $t^* \ge \tau_{\mathcal{F}^*}\|x\|$.

Proof. We assume for the proof that $\|x\| = 1$, since t^*, f^* scale positively with $\|x\|$. If $f^* = 0$, the result follows trivially since $\tau_{\mathcal{F}^*} \le 1$, and $t^* = 1$ from Proposition 3. If $f^* > 0$, define $c = -\frac{t^*}{f^*}Q_n$, and note that $\|c\| = \frac{t^*}{f^*}$. By definition of the width, $B(c, \frac{t^*}{f^*}\tau_{\mathcal{F}^*}) \subset -\mathcal{F}^*$. Note that $\|x - c\| = \sqrt{x^T x + 2\frac{t^*}{f^*}Q_n^T x + \frac{t^{*2}}{f^{*2}}Q_n^T Q_n} \le \sqrt{1 + \frac{t^{*2}}{f^{*2}}} = \frac{1}{f^*}$. Therefore $\frac{1}{f^*\|x-c\|} \ge 1$.

Next observe that $c + \frac{\tau_{\mathcal{F}^*}\|c\|(x-c)}{\|x-c\|} \in -\mathcal{F}^*$, which is equivalent to $c + \frac{\tau_{\mathcal{F}^*}t^*(x-c)}{f^*\|x-c\|} \in -\mathcal{F}^*$. By the previous inequality, we have $c + \tau_{\mathcal{F}^*}t^*(x - c) \in -\mathcal{F}^*$. Thus we have

$$f^* \le \|c + \tau_{\mathcal{F}^*}t^*(x - c) - x\| = (1 - \tau_{\mathcal{F}^*}t^*)\|x - c\| \le (1 - \tau_{\mathcal{F}^*}t^*)\frac{1}{f^*}.$$

Therefore, $1 - t^{*2} = f^{*2} \le 1 - \tau_{\mathcal{F}^*}t^*$, which implies that $\tau_{\mathcal{F}^*} \le t^*$. ∎

PROPOSITION 5. *Given* x *satisfying* $\|x\| = 1$ *and* $Q_n^T x \le 0$, *suppose that* (v, w, ρ) *satisfies* (6), *with duality gap* $G^\circ \le \sigma\tau_{\mathcal{F}^*}/2$ *for* (5), *where* $\sigma \le 1$. *Consider the assignment* $(y, z, \theta) \leftarrow (-\rho w, -v/\|v\|, \|v\|)$ *(with the necessary modification that* $y = 0$ *if* $v = 0$*). Then* (y, z, θ) *satisfies* (4), *with duality gap* $G \le \sigma$ *for* (1).

Proof. Note that $y^T z = \frac{(w^T v)\rho}{\|v\|} \le \frac{\sigma\tau_{\mathcal{F}^*}\rho}{2\|v\|}$, and we have the following relations: (i) $w^T v \le \sigma\tau_{\mathcal{F}^*}/2 \le 1/2$, (ii) $\|v\| = \theta = \|y - x\| \ge t^* \ge \tau_{\mathcal{F}^*}$ from Proposition 4, and (iii) $\rho = \|v - x\| = v^T w - w^T x \le 1/2 + f^* \le 3/2$ from Proposition 3. Therefore $y^T z \le \frac{\tau_{\mathcal{F}^*}\sigma}{2}\frac{3}{2}\frac{1}{\tau_{\mathcal{F}^*}} \le \sigma$. ∎

3.2. The six cases. We assume here that the given x has unit norm, i.e., $\|x\| = 1$, and that we seek feasible solutions to (1) with a duality gap at most σ, where $\sigma \le 1$. Armed with Propositions 3, 4, and 5, we now show how to compute a feasible solution (y, z) of (1) with duality gap $G \le \sigma$. Our method is best understood with the help of Figure 1. We know from section 3.1 and the conditions (4) and/or (6) that we need to decompose x into the sum of a vector in $\mathcal{F}$ plus a vector in $-\mathcal{F}^*$ and that the central axes of $\mathcal{F}$ and $-\mathcal{F}$ are the rays corresponding to Q_n and $-Q_n$, respectively. Define the "dividing hyperplane" $L_{\mathcal{F}} := \{y : Q_n^T y = 0\}$ perpendicular to the central axes of $\mathcal{F}$ and $-\mathcal{F}$, and define $L_{\mathcal{F}}^+ := \{y \in \mathbb{R}^n : Q_n^T y \ge 0\}$ and $L_{\mathcal{F}}^- := -L_{\mathcal{F}}^+$. We divide $L_{\mathcal{F}}^+$ into three regions: region 1 corresponds to points in $\mathcal{F}$, region 2 corresponds to points in $L_{\mathcal{F}}^+$ "near" the dividing hyperplane (where our nearness criterion will be defined

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

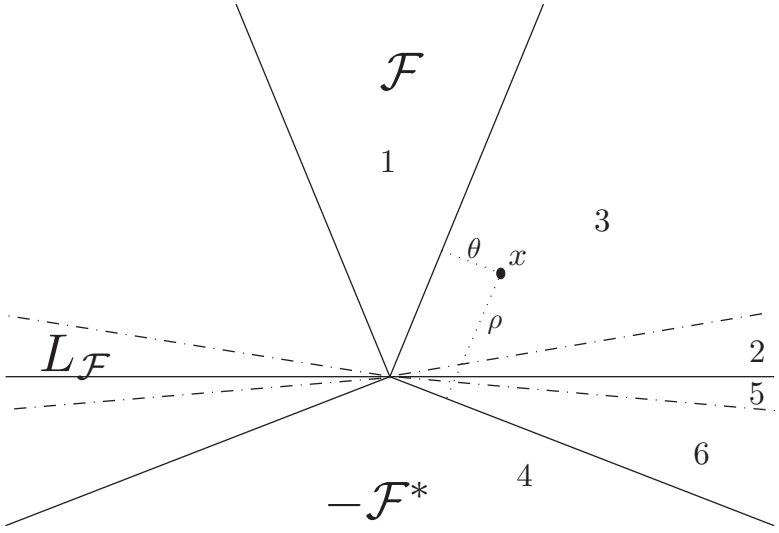


FIG. 1. *The geometry of the sets $\mathcal{F}$, $-\mathcal{F}^*$, and $L_{\mathcal{F}}$ and the six cases. The central axes of $\mathcal{F}$ and $-\mathcal{F}^*$ are the rays generated by $\pm Q_n$, respectively, which are orthogonal to the hyperplane $L_{\mathcal{F}}$. The regions corresponding to the six cases are shown as well.*

shortly), and region 3 corresponds to points in $L^+_{\mathcal{F}} \setminus \mathcal{F}$ that are "far" from $L_{\mathcal{F}}$. We divide $L^-_{\mathcal{F}}$ similarly, into regions 4, 5, and 6. For each of the three regions in $L^+_{\mathcal{F}}$, we will work with the problem pair (1) and show how to compute a feasible solution (y, z) of (1) with duality gap $G \le \sigma$. For each of the three regions in $L^-_{\mathcal{F}}$, we will instead work with the problem pair (5) and show how to compute a feasible solution (w, s) of (5) with duality gap $G^\circ \le \sigma\tau_{\mathcal{F}^*}/2$, whereby from Proposition 5 we obtain a feasible solution (y, z) of (1) with duality gap $G \le \sigma$. We will consider six cases, one for each of the regions described above and in Figure 1.

We first describe how we choose whether x is in region 2 or 3. For $x \in L^+_{\mathcal{F}} \setminus \mathcal{F}$, define

$$\varepsilon_{\mathcal{P}} = \varepsilon_{\mathcal{P}}(x) := \frac{Q_n^T x \sqrt{|D_n|}}{\sqrt{\sum_{i=1}^{n-1} D_i \left(Q_i^T x\right)^2}}, \tag{7}$$

and notice that $x \in L^+_{\mathcal{F}}$ implies that $\varepsilon_{\mathcal{P}} \ge 0$, $x \notin \mathcal{F}$ implies that $\varepsilon_{\mathcal{P}} < 1$, and smaller values of $\varepsilon_{\mathcal{P}}$ correspond to $Q_n^T x$ closer to zero and hence x closer to $L_{\mathcal{F}}$. We specify a tolerance $\bar\varepsilon_{\mathcal{P}}$ and determine whether x is in region 2 or 3 depending on whether $\varepsilon_{\mathcal{P}} \le \bar\varepsilon$ or $\varepsilon_{\mathcal{P}} > \bar\varepsilon$, respectively, where we set $\bar\varepsilon = \bar\varepsilon_{\mathcal{P}} := \sigma\tau_{\mathcal{F}}$.

Case 1: $Q_n^T x \ge 0$ *and* $x^T Q D Q^T x \le 0$. From Theorem 1 we know that $x \in \mathcal{F}$. Then it is elementary to show that $(y, z, \theta) \leftarrow (x, 0, 0)$ satisfy (4), with $y^T z = 0$, whereby from Proposition 3 the duality gap is $G = 0$.

Case 2: $Q_n^T x \ge 0$ *and* $x^T Q D Q^T x > 0$, $\varepsilon_{\mathcal{P}} \le \bar\varepsilon_{\mathcal{P}} := \sigma\tau_{\mathcal{F}}$. Let $\hat y$ solve the following system of equations:

$$\begin{aligned} [I + 1/|D_n| D] Q^T \hat y &= Q^T x - e_n Q_n^T x, \\ Q_n^T \hat y &= 0, \end{aligned} \tag{8}$$

where $e_n = (0, \ldots, 0, 1) \in \mathbb{R}^n$. Notice that the last row of the first equation system

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

has all zero entries. Therefore this system is not overdetermined, and one can write the closed-form solution $(Q^T\hat{y})_i = (Q^T x)_i/(1 + 1/|D_n|D_i)$ for $i = 1, \ldots, n-1$ and $(Q^T\hat{y})_n = 0$, in the transformed variables $\hat{s} := Q^T\hat{y}$. Having computed $\hat{y}$, next compute $\alpha := \sqrt{\hat{y}^T QDQ^T\hat{y}}/\sqrt{|D_n|}$, and then make the following assignments to variables:

$$\begin{aligned}
&\bar{y} \leftarrow \hat{y} + \alpha Q_n,\\
&\theta \leftarrow \sqrt{\bar{y}^T QD^2Q^T\bar{y}}/|D_n|,\\
&z \leftarrow -QDQ^T\bar{y}/(|D_n|\theta),\\
&y \leftarrow \bar{y} + Q_n^T x Q_n.
\end{aligned}$$

PROPOSITION 6. *Suppose that* $\|x\| = 1$, $\sigma \le 1$, *and* $\varepsilon_{\mathcal{P}} \le \bar{\varepsilon} < 1$ *and that* (y, z, θ) *are computed according to Case* 2 *above. Then* (y, z, θ) *is feasible for* (4) *with duality gap* $G \le \bar{\varepsilon}/\tau_{\mathcal{F}}$ *for* (1).

Applying Proposition 6 using $\bar{\varepsilon} = \bar{\varepsilon}_{\mathcal{P}} := \sigma\tau_{\mathcal{F}}$ ensures that the resulting duality gap satisfies $G \le \bar{\varepsilon}/\tau_{\mathcal{F}} = \sigma$. Note that the complexity of the computations in Case 2 is $O(mn^2)$ (assuming that square roots are sufficiently accurately computed in $O(1)$ operations).

Proof of Proposition 6. It is easy to establish that $(Q_1^T x, \ldots, Q_{n-1}^T x) \ne 0$, and hence $\alpha > 0$. This in turn implies that $Q_n^T\bar{y} = \alpha > 0$, and hence $\theta > 0$, so z is well defined. It is straightforward to verify that

$$\bar{y}^T QDQ^T\bar{y} = (\hat{y} + \alpha Q_n)^T QDQ^T(\hat{y} + \alpha Q_n) = \hat{y}^T QDQ^T\hat{y} - \alpha^2|D_n| = 0,$$

which shows via Theorem 1 that $\bar{y} \in \mathcal{F}$, and therefore $z \in \mathcal{F}^*$ and $z^T\bar{y} = 0$. It is also straightforward to verify that $\|z\| = 1$. Finally, we have from (8) that

$$\begin{aligned}
[I + 1/|D_n|D]Q^T\bar{y} &= [I + 1/|D_n|D]\left(Q^T\hat{y} + \alpha e_n\right)\\
&= [I + 1/|D_n|D]\left(Q^T\hat{y}\right) = Q^T\left(x - Q_n Q_n^T x\right)
\end{aligned}$$

(where the second equality above follows since the last row and column of the matrix are zero); hence $\bar{y} + 1/|D_n|QDQ^T\bar{y} = x - Q_n Q_n^T x$. Substituting the values of y, z, θ into this expression yields $y - \theta z = x$, which then shows that (y, z, θ) satisfy (4). Therefore from Proposition 3 (y, z) is feasible for (1) with duality gap

$$\begin{aligned}
G = z^T y = z^T\bar{y} + z^T Q_n Q_n^T x \le Q_n^T x &= \frac{\varepsilon_{\mathcal{P}}\sqrt{\sum_{i=1}^{n-1} D_i(Q_i^T x)^2}}{\sqrt{|D_n|}}\\
&\le \frac{\bar{\varepsilon}\sqrt{D_1}}{\sqrt{|D_n|}} \le \frac{\bar{\varepsilon}\sqrt{D_1 + |D_n|}}{\sqrt{|D_n|}} = \bar{\varepsilon}/\tau_{\mathcal{F}}. \qquad \square
\end{aligned}$$

Case 3*:* $Q_n^T x \ge 0$ *and* $x^T QDQ^T x > 0$, $\varepsilon_{\mathcal{P}} > \bar{\varepsilon}_{\mathcal{P}} := \sigma\tau_{\mathcal{F}}$. Here x is on the same side of the dividing hyperplane $L_{\mathcal{F}}$ as $\mathcal{F}$ but is neither in $\mathcal{F}$ nor close enough to $L_{\mathcal{F}}$ in the nearness measure. Consider the following univariate function in γ:

$$f(\gamma) := x^T Q[I + \gamma D]^{-1} D[I + \gamma D]^{-1} Q^T x = \sum_{i=1}^{n} \frac{D_i\left(x^T Q_i\right)^2}{(1 + D_i\gamma)^2}, \tag{9}$$

shown canonically in Figure 2.

Notice that $f(0) = x^T QDQ^T x > 0$, and since $D_n < 0$, we have $f(\gamma) \to -\infty$ as $\gamma \to 1/|D_n|$. Furthermore, $f'(\gamma) = -2\sum_{i=1}^{n} D_i^2(x^T Q_i)^2(1 + \gamma D_i)^{-3} < 0$ for $\gamma \in$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

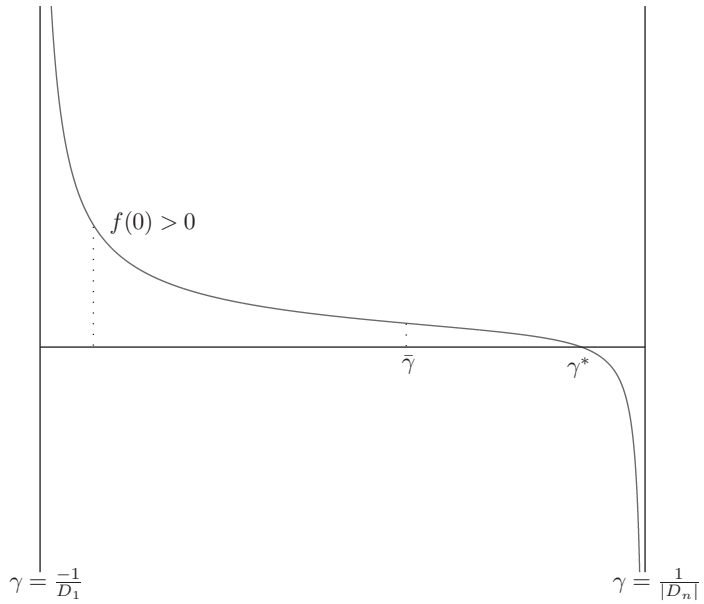


FIG. 2. *The function f on the interval $(-1/D_1, 1/|D_n|)$. Among many desirable properties, f is strictly decreasing and analytic and has a unique root $\gamma^* \in (0, 1/|D_n|)$. Moreover, f is convex over $(-1/D_1, \bar{\gamma})$ and concave over $(\bar{\gamma}, 1/|D_n|)$, where $\bar{\gamma}$ is the unique point satisfying $f''(\bar{\gamma}) = 0$. Note that one can have $\gamma^* \leq \bar{\gamma}$ or $\gamma^* \geq \bar{\gamma}$.*

$[0, 1/|D_n|)$. Therefore $f(\gamma)$ is strictly decreasing in the domain $[0, 1/|D_n|)$, whereby from the mean value theorem there is a unique value $\gamma^* \in (0, 1/|D_n|)$ for which $f(\gamma^*) = 0$. We show in section 5 how to combine a binary search and Newton's method to very efficiently compute $\gamma \in (0, 1/|D_n|)$ satisfying $f(\gamma) \leq 0$ and $f(\gamma) \approx 0$ (and $\gamma \approx \gamma^*$). Presuming that this can be done very efficiently, consider the following variable assignment:

$$\begin{array}{l} y \leftarrow Q\left[I + \gamma D\right]^{-1} Q^T x, \\ \theta \leftarrow \gamma\sqrt{y^T Q D^2 Q^T y}, \\ z \leftarrow -\gamma Q D Q^T y/\theta. \end{array} \tag{10}$$

We now show that (y, θ, z) satisfy (4). First note that $Q_n^T y = Q_n^T x/(1 - \gamma|D_n|) > 0$, and furthermore this shows that $\theta > 0$, and so z is well defined. By the hypothesis that $f(\gamma) \leq 0$ we have

$$y^T Q D Q^T y = x^T Q[I + \gamma D]^{-1} D[I + \gamma D]^{-1} Q^T x = f(\gamma) \leq 0,$$

which implies that $y \in \mathcal{F}$, and hence $z \in \mathcal{F}^*$ from Theorem 1. It is also straightforward to verify that $\|z\| = 1$. Finally, rearranging the formula for y yields $x = y + \gamma Q D Q^T y = y - \theta z$, which shows that (4) is satisfied. From Proposition 3, (y, z) is feasible for (1), and using the above assignments the duality gap works out to be

$$G = y^T z = -f(\gamma)/\sqrt{x^T Q D^2 [I + \gamma D]^{-2} Q^T x},$$

whereby G will be small if $f(\gamma) \approx 0$. To make this more precise requires a detailed analysis of a binary search and Newton's method, which is postponed to section 5 where we will prove the following.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

PROPOSITION 7. *Suppose that* $\|x\| = 1$, $1 > \varepsilon_{\mathcal{P}} > \bar{\varepsilon}$, *and* $g > 0$ *is a given gap tolerance. If* $Q_n^T x > 0$ *and* $x^T Q D Q^T x > 0$, *then a solution* (y, z, θ) *of* (4) *with duality gap* $G \leq g$ *for* (1) *is computable in* $O(n \ln\ln(1/\tau_{\mathcal{F}} + 1/\bar{\varepsilon} + 1/g))$ *operations.*

Substituting $\bar{\varepsilon} = \bar{\varepsilon}_{\mathcal{P}} := \sigma\tau_{\mathcal{F}}$ and $g = \sigma$, it follows that the complexity of computing a feasible of solution of (y, z) of (1) with duality gap at most σ is $O(n \ln\ln(1/\tau_{\mathcal{F}} + 1/\sigma))$=$O(n \ln\ln(1/\min\{\tau_{\mathcal{F}}, \tau_{\mathcal{F}^*}\} + 1/\sigma))$ operations.

Case 4: $Q_n^T x \leq 0$ *and* $x^T Q D^{-1} Q^T x \leq 0$. From Theorem 1 we know that $x \in -\mathcal{F}^*$. Then it is elementary to show that $(y, z, \theta) \leftarrow (0, -x/\|x\|, \|x\|)$ satisfy (4), with $y^T z = 0$, whereby from Proposition 3 the duality gap is $G = 0$.

Before describing how we treat Cases 5 and 6 (corresponding to regions 5 and 6), we need to describe how we choose whether x is in region 5 or 6. We use a parallel concept to that used to distinguish regions 2 and 3, except that $\mathcal{F}$ is replaced by $-\mathcal{F}^*$; see Figure 1. For $x \in L_{\mathcal{F}}^- \setminus -\mathcal{F}^*$, define the following quantity analogous to (7):

$$\varepsilon_{\mathcal{P}^*} = \varepsilon_{\mathcal{P}^*}(x) := \frac{-Q_n^T x \sqrt{1/|D_n|}}{\sqrt{\sum_{i=1}^{n-1} (1/D_i) \left(Q_i^T x\right)^2}}, \tag{11}$$

and notice that $x \in L_{\mathcal{F}}^-$ implies $\varepsilon_{\mathcal{P}^*} \geq 0$, $x \notin -\mathcal{F}^*$ implies $\varepsilon_{\mathcal{P}^*} < 1$, and smaller values of $\varepsilon_{\mathcal{P}^*}$ correspond to $Q_n^T x$ closer to zero and hence x closer to $L_{\mathcal{F}}$. We specify a tolerance $\bar{\varepsilon}_{\mathcal{P}^*}$ and determine whether x is in region 5 or 6 depending on whether $\varepsilon_{\mathcal{P}^*} \leq \bar{\varepsilon}$ or $\varepsilon_{\mathcal{P}^*} > \bar{\varepsilon}$, respectively, where we set $\bar{\varepsilon} = \bar{\varepsilon}_{\mathcal{P}^*} := \sigma\tau_{\mathcal{F}^*}^2/2$.

Case 5: $Q_n^T x \leq 0$ *and* $x^T Q D^{-1} Q^T x > 0$, *and* $\varepsilon_{\mathcal{P}^*} \leq \bar{\varepsilon}_{\mathcal{P}^*} := \sigma\tau_{\mathcal{F}^*}^2/2$. This case is an exact analogue of Case 2, with $\mathcal{F}$ replaced by $-\mathcal{F}^*$ and the pair (1) replaced by (5). Therefore the methodology of Case 2 can be used to compute (v, w, ρ) satisfying (6), and hence (v, w) is feasible for (5). Applying Proposition 6 to the context of the polar pair (5) with $\bar{\varepsilon} = \bar{\varepsilon}_{\mathcal{P}^*}$, it follows that the duality gap for (5) will be $G^\circ = v^T w$ and will satisfy $G^\circ \leq \bar{\varepsilon}/\tau_{\mathcal{F}^*} = \sigma\tau_{\mathcal{F}^*}^2/(2\tau_{\mathcal{F}^*}) \leq \sigma\tau_{\mathcal{F}^*}/2$. Converting (v, w, ρ) to (y, z, θ) using Proposition 5, we obtain (y, z) feasible for (1) with duality gap $G \leq \sigma$. Here the complexity of the computations is of the same order as Case 2.

Case 6: $Q_n^T x \leq 0$ *and* $x^T Q D^{-1} Q^T x > 0$, *and* $\varepsilon_{\mathcal{P}^*} > \bar{\varepsilon}_{\mathcal{P}^*} := \sigma\tau_{\mathcal{F}^*}^2/2$. In concert with the previous case, this case is an exact analogue of Case 3, with $\mathcal{F}$ replaced by $-\mathcal{F}^*$ and the pair (1) replaced by (5). Therefore the methodology of Case 3 can be used to compute (v, w, ρ) satisfying (6), and hence (v, w) is feasible for (5). Applying Proposition 7 to the context of the polar pair (5) with $\bar{\varepsilon} = \bar{\varepsilon}_{\mathcal{P}^*}$ and $g = \sigma\tau_{\mathcal{F}^*}/2$, it follows that a solution (v, w, ρ) of (6) with duality gap $G^\circ \leq g = \sigma\tau_{\mathcal{F}^*}/2$ for (5) is computable in $O(n \ln\ln(1/\tau_{\mathcal{F}^*} + 1/\bar{\varepsilon} + 1/g)) = O(n \ln\ln(1/\min\{\tau_{\mathcal{F}}, \tau_{\mathcal{F}^*}\} + 1/\sigma))$ operations. Converting (v, w, ρ) to (y, z, θ) using Proposition 5, we obtain (y, z) feasible for (1) with duality gap $G \leq \sigma$.

Proof of Theorem 2. The spectral decomposition of $M^T M - gg^T = QDQ^T$ is assumed to take $O(mn^2)$ operations. The computations in Cases 1 and 4 are trivial after checking the conditions of the cases, which is $O(mn^2)$ operations, and similarly for Cases 2 and 5. Regarding Cases 3 and 6, the discussion in the description of these cases establishes the desired operation bound. ∎

Remark 3 (the case when $\mathcal{F}$ is not regular, again). As in Remark 2, let Z and N partition the set of indices according to zero and nonzero values of D_i. Consider the case when $D_n < 0$ (the cases when $D_n > 0$ and $D_n = 0$ were discussed in Remark 2). We interpret $D_i^{-1} = \infty$ for $i \in Z$. Consider the orthonormal transformation $Q^T x$ and $Q^T y$, $Q^T z$ of the given vector x and the variables y, z. Then for $i \in Z$ simply set $Q_i^T y = Q_i^T x$ and $Q_i^T z = 0$ and work in the lower-dimensional problem in the subspace spanned by Q_i, $i \in N$.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

TABLE 1
Average computational results from 100 *randomly generated sparse problems.*

Dimension n	Iterations: Proposed method, Theoretical bound	Iterations: Proposed method, Actual	Iterations: SDPT3	Running time (seconds): Proposed method, EIG	Running time (seconds): Proposed method, Total	Running time (seconds): SDPT3	Range of widths: Minimum	Range of widths: Maximum
10	6.6	4.7	11.7	0.0005	0.0013	0.1089	1e-7	0.534015
20	7.2	4.8	13.9	0.0003	0.0011	0.1489	0.046400	0.510048
50	7.7	4.5	14.2	0.0011	0.0014	0.2828	0.087697	0.441248
100	8.0	4.3	18.8	0.0065	0.0075	1.1620	0.095081	0.414343
200	8.0	4.0	23.8	0.0556	0.0575	5.7393	0.176357	0.430163
500	8.0	3.8	18.8	1.0492	1.0571	33.9511	0.179974	0.317520

4. Comparison with interior-point methods. The primal-dual pair of problems (1) can be reformulated as second-order cone programs; one formulation of the primal problem is $\max_{y,t}\{-t : (My, g^T y) \in \mathcal{Q}^{m+1},\ (y - x, t) \in \mathcal{Q}^{n+1}\}$, for example. It then follows from interior-point complexity theory that approximate solutions to (1) with duality gap at most δ can be computed in $O(\ln(1/\delta))$ interior-point iterations. This iteration bound follows from Theorem 2.4.1 of Renegar [6], noting that an interior starting feasible solution with good symmetry is easy to precompute. Unlike the complexity bound for the proposed method in Theorem 2, the interior-point method bound has a stronger dependence on the duality gap δ (global linear convergence), but, unlike the proposed method, there is no dependence on the widths $\tau_{\mathcal{F}}, \tau_{\mathcal{F}^*}$. Therefore from a complexity viewpoint one cannot assert that one algorithm dominates the other. (There is a theory of local quadratic convergence for interior-point methods (see, for example, [1]) that could possibly be used to prove a weaker interior-point method dependence on δ for this class of problems.) In terms of computational practice, it is relevant to compare the two methods on randomly generated problems. We generated 100 relatively sparse random problem instances ((M, g) has density, respectively, 10% and 30% on average) for each of dimensions $(m, n) = (2n, n)$ for $n = 10$, 20, 50, 100, 200, and 500 and solved them using both our method and the conic convex interior-point method software SDPT3s [9]. All computation was performed in MatLab on a recent model laptop computer. We used MatLab's EIG command to compute the eigendecomposition for $M^T M - gg^T$ for our proposed method. Table 1 shows our computational results. Columns 2 and 3 of the table show the average theoretical iteration bound and the average actual iterations of our method. Columns 5, 6, and 7 report running time information. Note as expected that the eigendecomposition is the dominant computation in our method. Our method substantially outperforms SDPT3, as one would expect, since SDPT3 is not optimized for the problem class (1).

A fairer comparison can be done by presuming that the problem instances are preprocessed and transformed by the eigendecomposition, replacing $Q^T y$ with y, whereby the second-order cone formulation takes the diagonal form:

$$\max_{y,t} \left\{ -t : \left(\sqrt{D_1} y_1, \ldots, \sqrt{D_{n-1}} y_{n-1}, \sqrt{|D_n|} y_n \right) \in \mathcal{Q}^n,\ (y - x, t) \in \mathcal{Q}^{n+1} \right\},$$

and SDPT3 naturally exploits the sparse structure of this problem. We generated 100 random problem instances each, for a range of n from $n = 20$ to $n = 5000$, where each instance was generated by randomly choosing D but ensuring that $\tau_{\mathcal{F}} = \tau_{\mathcal{F}^*} = 10^{-7}$. These instances need no eigendecomposition for our method; also SDPT3 can take good advantage of the problem's natural sparsity as well. Table 2 shows our

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

TABLE 2

Average computational results from 100 *randomly generated diagonal problems with* $\tau_{\mathcal{F}} = \tau_{\mathcal{F}^*} = 10^{-7}$.

Dimension n	Iterations: Proposed method: Theoretical bound	Iterations: Proposed method: Actual	Iterations: SDPT3	Running time (seconds): Proposed method	Running time (seconds): SDPT3
10	8.0	5.0	23.9	0.0004	0.2320
20	8.0	5.0	26.9	0.0005	0.2778
50	8.0	5.0	28.1	0.0005	0.4869
100	8.0	5.0	18.3	0.0018	1.1063
200	8.0	5.0	17.6	0.0033	1.5806
500	8.0	4.9	20.2	0.0154	0.5290
1000	8.0	4.9	16.9	0.0567	1.0957
2000	8.0	5.0	16.4	0.2153	1.8111
5000	8.0	5.2	19.9	1.2656	9.3998

computational results. Our method still substantially outperforms SDPT3 but not as dramatically when n is very large. However, the running time numbers in Table 2 are the running time until the stopping criteria are met for each method. The stopping criteria for SDPT3 includes stopping when the duality gap is sufficiently small or when insufficient progress is made in satisfying primal/dual feasibility/optimality. For the diagonal problems generated, this latter stopping criteria is unfortunately encountered quite often: the relative error of the final solution from SDPT3 was at least 0.01 for 65% of the diagonal problem instances and was at least 0.001 for 81% of the instances. In fact, SDPT3 stopped with a relative error of at most 10^{-6} in only 2% of the instances. In contrast, our proposed method terminated with a relative error of 10^{-12} in all instances.

5. Proof of Proposition 7. This section is devoted to the proof of Proposition 7. Our algorithmic approach is motivated by Ye [10], and it consists of a combination of a binary search and Newton's method to approximately solve $f(\gamma) = 0$ for the function f given in (9). An alternate approach would be to use interpolation methods as presented and analyzed in Melman [4], for which global quadratic convergence is proved but there is no complexity analysis of associated constants. While Proposition 7 indicates that a solution (y, z, θ) of (4) with duality gap $G \leq g$ for (1) can be computed extremely efficiently, unfortunately our proof is not nearly as efficient as we or the reader might wish. We assume throughout this section that the hypotheses of Proposition 7 hold. We start with a review of Smale's main result for Newton's method in [8].

5.1. Newton's method and Smale's results. Let g be an analytic function, and consider the Newton iterate from a given point $\hat{\gamma}$:

$$\gamma^+ = \hat{\gamma} - \frac{g(\hat{\gamma})}{g'(\hat{\gamma})},$$

and let $\{\gamma_k\}_{k\geq 0}$ denote the sequence of points generated starting from $\hat{\gamma} = \gamma_0$.

DEFINITION 1. *A point* γ_0 *is said to be an approximate zero of* g *if*

$$|\gamma_k - \gamma_{k-1}| \leq (1/2)^{2^{k-1}-1}|\gamma_1 - \gamma_0| \quad \textit{for } k \geq 1.$$

For an approximate zero γ_0, let $\gamma^* = \lim_{k\to\infty} \gamma_k$. Then γ^* is a zero of g, and Newton's method starting from γ_0 converges quadratically to γ^* from the very first iteration. The main result in [8] can be restated as follows.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

THEOREM 3 (Smale [8]). *Let g be an analytic function. If $\hat{\gamma}$ satisfies*

$$\sup_{k>1}\left|\frac{g^{(k)}(\hat{\gamma})}{k!g'(\hat{\gamma})}\right|^{1/(k-1)} \leq \frac{1}{8}\left|\frac{g'(\hat{\gamma})}{g(\hat{\gamma})}\right|, \tag{12}$$

then $\hat{\gamma}$ is an approximate zero of g. Furthermore, if $\hat{\gamma}$ is an approximate zero of g, then $|\gamma_k - \gamma^*| \leq 2(1/2)^{2^{k-1}}|\gamma_1 - \gamma_0|$ for all $k \geq 1$.

5.2. Properties of $f(\gamma)$. We employ the change of variables $s = Q^T x$, whereby from the hypotheses of Proposition 7 we have $s_n > 0$, $s^T Ds > 0$, and $\varepsilon_{\mathcal{P}} = s_n\sqrt{|D_n|}/\sqrt{\sum_{j=1}^{n-1} D_i s_i^2} > \bar{\varepsilon}$. We consider computing a zero of our function of interest:

$$f(\gamma) = s^T (I+\gamma D)^{-2} Ds = \sum_{i=1}^{n} \frac{D_i s_i^2}{(1+\gamma D_i)^2}. \tag{13}$$

LEMMA 1. *Under the hypotheses of Proposition 7, f has the following properties:*

(i) *$f(0) > 0$, $\lim_{\gamma\to 1/|D_n|} f(\gamma) = -\infty$, and f has a unique root $\gamma^* \in (0, 1/|D_n|)$.*

(ii) *f is analytic on $(-1/D_1, 1/|D_n|)$, and for $k \geq 1$ the kth derivative of f is*

$$\begin{aligned} f^{(k)}(\gamma) &= (-1)^k(k+1)!s^T (I+\gamma D)^{-(k+2)} D^{k+1}s \\ &= (-1)^k(k+1)!\sum_{i=1}^{n} \frac{D_i^{k+1}s_i^2}{(1+\gamma D_i)^{k+2}}. \end{aligned}$$

(iii) $\displaystyle\sup_{k>1}\left|\frac{f^{(k)}(\gamma)}{k!f'(\gamma)}\right|^{1/(k-1)} \leq \frac{3}{2}\max\left\{\frac{D_1}{1+\gamma D_1},\ \frac{|D_n|}{1-\gamma|D_n|}\right\}.$

(iv) $\displaystyle\frac{1-\varepsilon_{\mathcal{P}}}{|D_n|+\varepsilon_{\mathcal{P}}D_1} \leq \gamma^* \leq \frac{1-\varepsilon_{\mathcal{P}}}{|D_n|}$, *where $\varepsilon_{\mathcal{P}}$ is given by (7).*

(v) *There exists a unique value $\bar{\gamma} \in (-1/D_1, 1/|D_n|)$ such that f is convex on $(-1/D_1, \bar{\gamma}]$ and concave on $[\bar{\gamma}, 1/|D_n|)$.*

Proof. (i) follows from the mean value theorem and the observation that f is decreasing on $(0, 1/|D_1|)$, and (ii) follows using a standard derivation. To prove (iii) observe that

$$\begin{aligned} \left|\frac{f^{(k)}(\gamma)}{k!f'(\gamma)}\right|^{1/(k-1)} &= \left|\frac{(k+1)!}{2k!}\right|^{1/(k-1)} \left|\frac{s^T (I+\gamma D)^{-(k+2)} D^{k+1}s}{s^T (I+\gamma D)^{-3} D^2 s}\right|^{1/(k-1)} \\ &\leq \frac{3}{2}\left|\frac{s^T (I+\gamma D)^{-3/2} D\left[(I+\gamma D)^{-1} D\right]^{k-1} D (I+\gamma D)^{-3/2} s}{s^T (I+\gamma D)^{-3/2} D^2 (I+\gamma D)^{-3/2} s}\right|^{1/(k-1)} \\ &\leq \frac{3}{2}\max_{v\neq 0}\left|\frac{v^T P^{k-1} v}{v^T v}\right|^{1/(k-1)} \\ &= \frac{3}{2}\max_{i=1,\ldots,n}\left\{\frac{|D_i|}{1+\gamma D_i}\right\}, \end{aligned}$$

where $P = (I+\gamma D)^{-1} D$. Therefore

$$\left|\frac{f^{(k)}(\gamma)}{k!f'(\gamma)}\right|^{1/(k-1)} \leq \frac{3}{2}\max_{i=1,\ldots,n}\left\{\frac{|D_i|}{1+\gamma D_i}\right\} \leq \frac{3}{2}\max\left\{\frac{D_1}{1+\gamma D_1},\ \frac{|D_n|}{1-\gamma|D_n|}\right\},$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

which proves (iii). To prove the first inequality of (iv), note that

$$f(\gamma) = \sum_{i=1}^{n} \frac{D_i s_i^2}{(1+\gamma D_i)^2} \geq \frac{1}{(1+\gamma D_1)^2} \sum_{i=1}^{n-1} D_i s_i^2 - \frac{|D_n| s_n^2}{(1+\gamma D_n)^2}.$$

The right-hand side of the expression above equals zero only at $\tilde{\gamma} := \frac{1-\varepsilon_{\mathcal{P}}}{|D_n|+\varepsilon_{\mathcal{P}} D_1}$. This implies that $f(\tilde{\gamma}) \geq 0$, whereby $\tilde{\gamma} \leq \gamma^*$, since f is strictly decreasing. For the second inequality note that $\varepsilon_{\mathcal{P}} \in (0,1)$ since $s_n > 0$ and $s^T D s > 0$. We have $f(\gamma) < \sum_{i=1}^{n-1} s_i^2 D_i \ - |D_n| s_n^2/(1+\gamma D_n)^2$, and substituting $\gamma = \frac{1-\varepsilon_{\mathcal{P}}}{|D_n|}$ into this strict inequality yields $f(\frac{1-\varepsilon_{\mathcal{P}}}{|D_n|}) < 0$, which then implies that $\gamma^* < \frac{1-\varepsilon_{\mathcal{P}}}{|D_n|}$. To prove (v), examine the derivatives of f in (ii), and notice that $f^{(k)}(\gamma) < 0$ for any odd value of k, whereby f'' is strictly decreasing. Let $\bar{\gamma}$ be the unique point in $(-1/D_1, 1/|D_n|)$ such that $f''(\bar{\gamma}) = 0$. Since f'' is strictly decreasing, f is convex on $(-1/D_1, \bar{\gamma})$ and concave on $(\bar{\gamma}, 1/|D_n|)$. ∎

Figure 2 illustrates the geometry underlying some of the analytical properties of f described by Lemma 1.

Remark 4. In the interval $(\frac{-1}{D_1}, \frac{1}{2|D_n|} - \frac{1}{2D_1}]$ the maximum in (iii) of Lemma 1 is $\frac{D_1}{1+\gamma D_1}$, and in the interval $[\frac{1}{2|D_n|} - \frac{1}{2D_1}, \frac{1}{|D_n|})$ the maximum is $\frac{|D_n|}{1+\gamma D_n}$.

5.3. Locating an approximate zero of f by binary search. From Lemma 1 we know that $\gamma^* \in (0, \bar{U}]$, where $\bar{U} := (1-\bar{\varepsilon})/|D_n|$. We will cover this interval with subintervals and use a binary search to locate an approximate zero of f, motivated by the method of Ye [10]. Noticing from Remark 4 that the maximum in (iii) of Lemma 1 depends on the "midpoint" $M := \frac{1}{2|D_n|} - \frac{1}{2D_1}$, we will consider two types of subintervals: the *left intervals* will cover $[0, \max\{0, M\}]$, and the *right intervals* will cover $[\max\{0, M\}, \bar{U}]$. (Of course, in the case when $M \leq 0$, there is no need to create the left intervals.)

The left intervals will be of the form $[L^{i-1}, L^i]$, where $L^i := \frac{1}{D_1}((\frac{13}{12})^i - 1)$ for $i = 0, 1, \ldots$. If $M \leq 0$, we do not consider creating these intervals. The right intervals will have the form $[R^i, R^{i-1}]$, where $R^i := \frac{1}{|D_n|} - (\frac{1}{|D_n|} - \bar{U})(\frac{13}{12})^i$ for $i = 0, 1, \ldots$.

Let $[a, b]$ denote one of these intervals (either $[L^{i-1}, L^i]$ or $[R^i, R^{i-1}]$ for some i). Note that if $f(a) \geq 0$ and $f(b) \leq 0$, then $\gamma^* \in [a, b]$. Supposing that this is the case, it follows from Lemma 1 that f is either convex on $[a, \gamma^*]$ or concave on $[\gamma^*, b]$ (or both), and consider starting Newton's method from $\hat{\gamma} = a$ in the first case or $\hat{\gamma} = b$ in the second case. Then the Newton step

$$\gamma^+ = \hat{\gamma} - \frac{f(\hat{\gamma})}{f'(\hat{\gamma})}$$

satisfies

$$\left| \frac{f(\hat{\gamma})}{f'(\hat{\gamma})} \right| = |\gamma^+ - \hat{\gamma}| \leq |\gamma^* - \hat{\gamma}| \leq b - a, \tag{14}$$

where the first inequality follows from either the convexity of f on $[a, \gamma^*]$ or the concavity of f on $[\gamma^*, b]$. In particular, we have

$$|f(\hat{\gamma})| \leq |f'(\hat{\gamma})||\gamma^* - \hat{\gamma}|, \tag{15}$$

which relates the value of the function at an approximate solution and the error in our approximation.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

LEMMA 2. *Under the hypotheses of Proposition 7 the intervals described herein have the following properties:*

(i) *The total number of left intervals and right intervals needed to cover* $[0, \bar{U}]$ *is* $K_L := \lceil \frac{\ln(1/2)+2\ln(1/\tau_{\mathcal{F}})}{\ln(13/12)} \rceil^+$ *and* $K_R := \lceil \frac{\ln(1/\bar{\varepsilon})}{\ln(13/12)} \rceil$, *respectively.*

(ii) *Let* $[a, b]$ *denote one of these intervals, and suppose that* $f(a) \geq 0$ *and* $f(b) \leq 0$. *Then either* a *or* b *is an approximate zero of* f, *and* $\gamma^* \in [a, b]$.

(iii) $R^{i-1} - R^i \leq \frac{1}{12|D_n|}$ *for* $i = 1, \ldots, K_R$ *and* $L^i - L^{i-1} \leq \frac{1}{12|D_n|}$ *for* $i = 1, \ldots, K_L$.

Proof. We first prove (i) for the right intervals. We have $R^0 = \bar{U}$ and

$$R^{K_R} = \frac{1}{|D_n|} - \frac{\bar{\varepsilon}}{|D_n|}\left(\frac{13}{12}\right)^{K_R} \leq \frac{1}{|D_n|} - \frac{\bar{\varepsilon}}{|D_n|}\frac{1}{\bar{\varepsilon}}\min\left\{1, \frac{|D_n|}{2D_1} + \frac{1}{2}\right\}$$
$$= \max\left\{0, \frac{1}{2|D_n|} - \frac{1}{2D_1}\right\} = \max\{0, M\},$$

and thus the right intervals cover $[\max\{0, M\}, \bar{U}]$. Note that, using the above reasoning, one easily shows that, because $K_R \leq 1 + \ln(1/\bar{\varepsilon})/\ln(13/12)$, one also has

$$\left(\frac{13}{12}\right)^{K_R} \leq \frac{13}{12\bar{\varepsilon}}. \tag{16}$$

For the left intervals, first consider the case when $M \geq 0$. Then $|D_n| \leq D_1$ and $\tau_{\mathcal{F}} \leq \frac{1}{\sqrt{2}}$, whereby there is no need to take the nonnegative part in the definition of K_L. We have $L^0 = 0$ and

$$L^{K_L} = \frac{1}{D_1}\left(\left(\frac{13}{12}\right)^{K_L} - 1\right) \geq \frac{1}{D_1}\left(\frac{1}{2\tau_{\mathcal{F}}^2} - 1\right) = \frac{1}{D_1}\left(\frac{D_1 + |D_n|}{2|D_n|} - 1\right) = M,$$

and thus the left intervals cover $[0, M] = [0, \max\{0, M\}]$. Note that, using the above reasoning, one easily shows that, because $K_L \leq 1 + \frac{\ln(1/2)+2\ln(1/\tau_{\mathcal{F}})}{\ln(13/12)}$, one also has

$$\left(\frac{13}{12}\right)^{K_L} \leq \frac{13}{24\tau_{\mathcal{F}}^2}. \tag{17}$$

When $M \leq 0$ there is nothing to prove.

To prove (ii), we consider the two cases of $[a, b]$ being either a left or right interval. If $[a, b]$ is a left interval, then $M \geq 0$ and $b = a(13/12) + \frac{1}{12D_1}$. In this case, for one of $\hat{\gamma} = a$ or $\hat{\gamma} = b$, we have for all $k > 1$:

$$\frac{1}{8}\left|\frac{f'(\hat{\gamma})}{f(\hat{\gamma})}\right| \geq \frac{1/8}{b-a} = \frac{1/8}{(1/12)(a + 1/D_1)} \geq \frac{3}{2}\frac{D_1}{1 + \hat{\gamma}D_1} \geq \left|\frac{f^{(k)}(\hat{\gamma})}{k!f'(\hat{\gamma})}\right|^{1/(k-1)},$$

where the first inequality uses (14), the second inequality uses $a \leq \hat{\gamma}$, and the third inequality uses Remark 4 and the fact that $\hat{\gamma} \leq M$ in conjunction with Lemma 1. Therefore $\hat{\gamma}$ is an approximate zero of f. If $[a, b]$ is a right interval, then $a = b(13/12) - \frac{1}{12|D_n|}$ and $M \leq a \leq b$. In this case, for one of $\hat{\gamma} = a$ or $\hat{\gamma} = b$, we have for all $k > 1$:

$$\frac{1}{8}\left|\frac{f'(\hat{\gamma})}{f(\hat{\gamma})}\right| \geq \frac{1/8}{b-a} = \frac{1/8}{b - b(13/12) + \frac{1}{12|D_n|}} = \frac{1/8}{\frac{1}{12}\left(\frac{1}{|D_n|} - b\right)} = \frac{3}{2}\,\frac{|D_n|}{1 - b|D_n|}$$
$$\geq \frac{3}{2}\,\frac{|D_n|}{1 - \hat{\gamma}|D_n|} \geq \left|\frac{f^{(k)}(\hat{\gamma})}{k!f'(\hat{\gamma})}\right|^{1/(k-1)},$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

where the first inequality uses (14), the second inequality uses $M \le a \le \hat{\gamma} \le b$, and the third inequality uses Remark 4 and the fact that $\hat{\gamma} \ge M$ in conjunction with Lemma 1. Therefore $\hat{\gamma}$ is an approximate zero of f.

To prove (iii), for the right intervals

$$R^{i-1} - R^i = \frac{\bar{\varepsilon}}{13|D_n|}\left(\frac{13}{12}\right)^i \le \frac{\bar{\varepsilon}}{13|D_n|}\left(\frac{13}{12}\right)^{K_R} \le \frac{13}{12}\frac{1}{13|D_n|} = \frac{1}{12|D_n|},$$

by the definition of K_R, and the second inequality derives from (16).

For the left intervals, we can assume $M \ge 0$ (otherwise they are not constructed), in which case $D_1 \ge |D_n|$. In this case, we have

$$L^i - L^{i-1} = \frac{1}{13D_1}\left(\frac{13}{12}\right)^i \le \frac{1}{13D_1}\left(\frac{13}{12}\right)^{K_L} \le \frac{1}{13D_1}\frac{13}{24\tau_{\mathcal{F}}^2} = \frac{1}{24}\left(\frac{1}{D_1} + \frac{1}{|D_n|}\right) \le \frac{1}{12|D_n|},$$

by the definition of K_L, and the second inequality derives from (17). ∎

Based on these properties, consider the following method for locating an approximate zero of f. Perform a binary search on the end points of the intervals, testing the end points to locate an interval $[a, b]$ for which $f(a) \ge 0$ and $f(b) \le 0$. Then either a or b is an approximate zero of f. Then initiate Newton's method from *both* a and b either in parallel or iterate-sequentially. Notice that, in order to perform a binary search on the left and right intervals, there is no need to compute and evaluate f for all of the end points. In fact, the operation complexity of a binary search will be $O(n \ln K_L)$ and $O(n \ln K_R)$, respectively, since each function evaluation of f requires $O(n)$ operations.

5.4. Computing a solution of (1) with duality gap at most σ. Under the hypotheses of Proposition 7, suppose that $[a, b]$ is one of the constructed intervals, $f(a) \ge 0$, and $f(b) \le 0$. Then, from Lemmas 1 and 2, $\gamma^* \in [a, b]$ and either f is convex on $[a, \gamma^*]$ or concave on $[\gamma^*, b]$ (or both). We first analyze the latter case, i.e., when f is concave on $[\gamma^*, b]$, whereby b is an approximate zero of f, and we analyze the iterates of Newton's method for k iterations starting at $\gamma_0 = b$. Let $\gamma := \gamma_k$ be the final iterate. It follows from the concavity of f on $[\gamma^*, b]$ that $\gamma \ge \gamma^*$ and consequently $f(\gamma) \le 0$. Then the analysis in Case 3 shows that the assignment (10) yields a feasible solution of (1) with duality gap $G = -f(\gamma)/\sqrt{s^T D^2 [I + \gamma D]^{-2} s}$. The following result bounds the value of this duality gap.

LEMMA 3. *Let $g \in (0, 1]$ be the desired duality gap for* (1)*, and let*

$$k = 1 + \left\lceil \frac{\ln\ln\left(\left(\frac{1}{3g}\right)\left(\frac{1}{\tau_{\mathcal{F}}^2} + \frac{1}{\bar{\varepsilon}^2}\right)\right) - \ln\ln 2}{\ln 2} \right\rceil.$$

Under the hypotheses of Proposition 7 *and the setup above where b is an approximate zero of f, let $\gamma_0 := b$ and $\gamma_1, \ldots, \gamma_k$ be the Newton iterates, and define $\gamma := \gamma_k$. Then the assignment* (10) *will be feasible for* (1) *with duality gap at most g.*

Proof. We have $|f(\gamma)| \le |f'(\gamma)|\,|\gamma^* - \gamma|$ from the concavity of f on $[\gamma^*, b]$. Also, we have

$$|f'(\gamma)| = 2\sum_{i=1}^{n} \frac{D_i^2 s_i^2}{(1 + \gamma D_i)^3} \le 2\sum_{i=1}^{n-1} \frac{D_i^2 s_i^2}{(1 + \gamma D_i)^2} + 2\frac{D_n^2 s_n^2}{(1 + \gamma D_n)^2}\frac{1}{(1 + \gamma D_n)}.$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Substitute $\frac{1}{1+\gamma D_n} = 1 + \frac{-\gamma D_n}{1+\gamma D_n}$ to obtain

$$|f'(\gamma)| \le 2\sum_{i=1}^{n} \frac{D_i^2 s_i^2}{(1+\gamma D_i)^2} - 2\frac{\gamma D_n^3 s_n^2}{(1+\gamma D_n)^3}.$$

Let $G = y^T z$ denote the duality gap. Then

$$\begin{aligned}
G &= \frac{-f(\gamma)}{\sqrt{s^T D^2 [I+\gamma D]^{-2} s}} \le \frac{|f'(\gamma)|\ |\gamma^* - \gamma|}{\sqrt{s^T D^2 [I+\gamma D]^{-2} s}} \\
&\le \frac{2\sum_{i=1}^n \frac{D_i^2 s_i^2}{(1+\gamma D_i)^2} + 2\frac{\gamma |D_n|^3 s_n^2}{(1+\gamma D_n)^3}}{\sqrt{s^T D^2 [I+\gamma D]^{-2} s}} |\gamma^* - \gamma| \\
&= \left(2\sqrt{s^T D^2 [I+\gamma D]^{-2} s} + 2\frac{\gamma |D_n|^3 s_n^2}{(1+\gamma D_n)^3 \sqrt{s^T D^2 [I+\gamma D]^{-2} s}} \right) |\gamma^* - \gamma| \\
&\le \left(2D_1 + 2\frac{|D_n|}{1+\gamma D_n} + 2\frac{\gamma D_n^2 s_n}{(1+\gamma D_n)^2} \right) |\gamma^* - \gamma|,
\end{aligned}$$

where we used $\sqrt{s^T D^2 [I+\gamma D]^{-2} s} \ge |D_n| s_n / (1+\gamma D_n)$ in the last inequality. Next note that $\gamma \le \bar{U} = \frac{1-\bar{\varepsilon}}{|D_n|}$, which implies that $\frac{1}{\bar{\varepsilon}} \ge \frac{1}{1+\gamma D_n}$. Therefore, recalling that γ is the kth iterate, we have

$$\begin{aligned}
G &\le 2|\gamma^* - \gamma| \left(D_1 + \frac{|D_n|}{\bar{\varepsilon}} + \frac{(1-\bar{\varepsilon}) D_n^2}{|D_n| \bar{\varepsilon}^2} \right) \\
&\le 2|\gamma^* - \gamma| |D_n| \left(\frac{1}{\tau_{\mathcal{F}}^2} + \frac{1}{\bar{\varepsilon}^2} \right) \\
&\le 4|\gamma_1 - \gamma_0| |D_n| \left(\frac{1}{\tau_{\mathcal{F}}^2} + \frac{1}{\bar{\varepsilon}^2} \right) \left(\frac{1}{2} \right)^{2^{k-1}} \\
&\le 4\frac{1}{12|D_n|} |D_n| \left(\frac{1}{\tau_{\mathcal{F}}^2} + \frac{1}{\bar{\varepsilon}^2} \right) \left(\frac{1}{2} \right)^{2^{k-1}} \\
&= \frac{1}{3} \left(\frac{1}{\tau_{\mathcal{F}}^2} + \frac{1}{\bar{\varepsilon}^2} \right) \left(\frac{1}{2} \right)^{2^{k-1}},
\end{aligned}$$

where we used Theorem 3 for the third inequality and Lemma 2 for the fourth inequality. Substituting the value of k above yields $G \le g$. ∎

Last of all, we analyze the case when f is convex on $[a, \gamma^*]$, whereby a is an approximate zero of f, and we analyze the iterates of Newton's method for k iterations starting at $\gamma_0 = a$. Let γ_k be the final iterate. It follows from the convexity of f on $[a, \gamma^*]$ that $\gamma_k \le \gamma^*$ and consequently $f(\gamma_k) \ge 0$, in which case the assignment (10) is not necessarily feasible for (1). However, invoking Theorem 3, we know that $\gamma_k + 2(1/2)^{2^{k-1}} |\gamma_1 - \gamma_0| \ge \gamma^*$, we also know that $\bar{U} \ge \gamma^*$, and we can set $\gamma := \min\{\gamma_k + 2(1/2)^{2^{k-1}} |\gamma_1 - \gamma_0|, \bar{U}\}$. Then the analysis in Case 3 shows that the assignment (10) yields a feasible solution of (1), with duality gap $G = -f(\gamma)/\sqrt{s^T D^2 [I+\gamma D]^{-2} s}$. The following result bounds the value of this duality gap.

LEMMA 4. *Let $g \in (0,1]$ be the desired duality gap for* (1)*, and let*

$$k = 1 + \left\lceil \frac{\ln \ln \left(\left(\frac{16}{3g} \right) \left(\frac{1}{\tau_{\mathcal{F}}^2} + \frac{1}{\bar{\varepsilon}^2} \right) \right) - \ln \ln 2}{\ln 2} \right\rceil.$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Under the hypotheses of Proposition 7 and the setup above where a is an approximate zero of f, let $\gamma_0 := a$ and $\gamma_1, \ldots, \gamma_k$ be the Newton iterates, and define $\gamma := \min\{\gamma_k + 2(1/2)^{2^{k-1}}|\gamma_1 - \gamma_0|, \bar{U}\}$. Then the assignment (10) will be feasible for (1) with duality gap at most g.

Proof. Define $\delta := \gamma - \gamma_k$, and it follows that $\delta \geq 0$ and $\gamma_k + \delta \leq \bar{U}$. Furthermore,

$$\begin{aligned}\delta &\leq 2(1/2)^{2^{k-1}}|\gamma_1 - \gamma_0| \\ &\leq \frac{2}{\left(\frac{16}{3g}\right)[1/\tau_{\mathcal{F}}^2 + 1/\bar{\varepsilon}^2]\,12|D_n|} \\ &\leq \frac{\min\{\bar{\varepsilon}^2, \tau_{\mathcal{F}}^2\}}{|D_n|} \leq \frac{\min\{\bar{\varepsilon}, \tau_{\mathcal{F}}^2/(1-\tau_{\mathcal{F}}^2)\}}{|D_n|} = \min\{\bar{\varepsilon}/|D_n|, 1/D_1\}.\end{aligned} \tag{18}$$

Therefore $\delta \leq \bar{\varepsilon}/|D_n|$, whereby $1 + \gamma_k D_n + 2\delta D_n = 1 - (\gamma_k + \delta)|D_n| - \delta|D_n| \geq 1 + \bar{\varepsilon} - 1 - \bar{\varepsilon} = 0$, where we also used $\gamma_k + \delta \leq \bar{U} = (1 - \bar{\varepsilon})/|D_n|$. Therefore

$$1 + \gamma_k D_n \leq 2(1 + (\gamma_k + \delta)D_n) \leq 2(1 + tD_n) \quad \text{for all} \quad t \in [\gamma_k, \gamma_k + \delta]. \tag{19}$$

We also have from (18) that $\delta \leq 1/D_1 \leq 1/D_i \leq 1/D_i + \gamma_k$ for $i = 1, \ldots, n-1$; hence

$$1 + \gamma_k D_i + \delta D_i \leq 2(1 + \gamma_k D_i), \ i = 1, \ldots, n-1. \tag{20}$$

The duality gap of the assignment (10) is

$$G = y^T z = \frac{-f(\gamma)}{\sqrt{s^T D^2 [I + \gamma D]^{-2} s}} = \frac{-f(\gamma_k + \delta)}{\sqrt{s^T D^2 [I + (\gamma_k + \delta) D]^{-2} s}}.$$

We now proceed to bound the numerator and denominator of the rightmost expression. For the numerator we have

$$-f(\gamma_k + \delta) = |f(\gamma_k + \delta)| = \left| f(\gamma_k) + \int_{\gamma_k}^{\gamma_k + \delta} f'(t)dt \right|.$$

However, observe that $f(\gamma_k) \geq 0$, $f(\gamma_k + \delta) \leq 0$, and $f'(t) \leq 0$ for all $t \in [0, 1/|D_n|)$, whereby

$$|f(\gamma_k + \delta)| \leq \int_{\gamma_k}^{\gamma_k + \delta} |f'(t)|dt.$$

Using (19) for $t \in [\gamma_k, \gamma_k + \delta]$, we have

$$\begin{aligned}|f'(t)| &= 2\sum_{i=1}^{n-1} \frac{D_i^2 s_i^2}{(1 + tD_i)^3} + 2\frac{D_n^2 s_n^2}{(1 + tD_n)^3} \\ &\leq 2\sum_{i=1}^{n-1} \frac{D_i^2 s_i^2}{(1 + \gamma_k D_i)^3} + 16\frac{D_n^2 s_n^2}{(1 + \gamma_k D_n)^3} \leq 8|f'(\gamma_k)|,\end{aligned}$$

and it follows that $-f(\gamma_k + \delta) \leq 8\delta|f'(\gamma_k)|$. To bound the denominator, simply notice from (20) and $1 + \gamma_k D_n + \delta D_n \leq 1 + \gamma_k D_n$ that $\sqrt{s^T D^2 [I + (\gamma_k + \delta) D]^{-2} s} \geq (1/2)\sqrt{s^T D^2 [I + \gamma_k D]^{-2} s}$. Therefore

$$G = \frac{-f(\gamma_k + \delta)}{\sqrt{s^T D^2 [I + (\gamma_k + \delta) D]^{-2} s}} \leq 16\frac{\delta|f'(\gamma_k)|}{\sqrt{s^T D^2 [I + \gamma_k D]^{-2} s}}.$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Next notice from the logic from the proof of Lemma 3 that

$$\frac{|f'(\gamma_k)|}{\sqrt{s^T D^2 [I+\gamma_k D]^{-2} s}} \leq 2|D_n| \left(\frac{1}{\tau_{\mathcal{F}}^2} + \frac{1}{\bar{\varepsilon}^2}\right);$$

therefore

$$G \leq 32\delta |D_n| \left(\frac{1}{\tau_{\mathcal{F}}^2} + \frac{1}{\bar{\varepsilon}^2}\right) \leq 32|D_n| \left(\frac{1}{\tau_{\mathcal{F}}^2} + \frac{1}{\bar{\varepsilon}^2}\right) \frac{2}{\left(\frac{16}{3g}\right)[1/\tau_{\mathcal{F}}^2 + 1/\bar{\varepsilon}^2]\, 12|D_n|} = g,$$

where the last inequality uses the second inequality of (18). ∎

Proof of Proposition 7. Note from the discussion at the end of section 5.3 that the operation complexity of the binary search is $O(n \ln K_L + n \ln K_R) = O(n \ln \ln(1/\tau_{\mathcal{F}} + 1/\bar{\varepsilon}))$ from Lemma 2. The number of Newton steps is $O(\ln \ln(1/\tau_{\mathcal{F}} + 1/\bar{\varepsilon} + 1/g))$ from Lemmas 3 and 4, with each Newton step requiring $O(n)$ operations, yielding the desired complexity bound. ∎

Acknowledgments. We are grateful to two anonymous referees for their suggestions on ways to improve the paper.

REFERENCES

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

[2] A. Belloni, R. M. Freund, and S. Vempala, *Efficiency of a Re-scaled Perceptron Algorithm for Conic Systems*, Working paper OR 379-06, MIT Operations Research Center, Cambridge, MA, 2006.

[3] A. Berman, *Cones, Matrices, and Mathematical Programming*, Springer-Verlag, New York, 1973.

[4] A. Melman, *A unifying convergence analysis of second-order methods for secular equations*, Math. Comp., 66 (1997), pp. 333–344.

[5] G. Pataki, *On the closedness of the linear image of a closed convex cone*, Math. Oper. Res., 32 (2007), pp. 395–412.

[6] J. Renegar, *A Mathematical View of Interior-Point Methods in Convex Optimization*, SIAM, Philadelphia, 2001.

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

[8] S. Smale, *Newton's method estimates from data at one point*, in The Merging of Disciplines: New Directions in Pure, Applied and Computational Mathematics, R. Ewing, K. Gross, and C. Martin, eds., Springer-Verlag, New York, 1986, pp. 185–196.

[9] J. Sturm, *Using SeDuMi* 1.02*, a Matlab toolbox for optimization over symmetric cones*, Optim. Methods Softw., 11 & 12 (1999), pp. 625–653.

[10] Y. Ye, *A new complexity result for minimizing a general quadratic function with a sphere constraint*, in Recent Advances in Global Optimization, C. Floudas and P. Pardalos, eds., Princeton University Press, Princeton, NJ, 1992, pp. 19–31.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.