RC24072 (W0610-045) October 10, 2006
Mathematics

IBM Research Report

On the Second-Order Feasibility Cone: Primal-Dual Representation and Efficient Projection

Alexandre Belloni
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Robert M. Freund
MIT Sloan School of Management
50 Memorial Drive
Cambridge, MA 02142

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

On the Second-Order Feasibility Cone: Primal-Dual Representation and Efficient Projection[*]

Alexandre Belloni[†] and Robert M. Freund[‡]

October, 2006

Abstract

We study the second-order feasibility cone $\mathcal{F} = \{y \in I\!R^n : \|My\| \leq g^T y\}$ 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 $x \in I\!R^n$ onto $\mathcal{F}$: $\min_y\{\|y - x\| : \|My\| \leq g^T y\}$, which aside from theoretical interest also arises as a necessary subroutine in the re-scaled 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 after the spectral decomposition of $M^T M - gg^T$ is computed. Here the width$(\mathcal{F})$, width$(\mathcal{F}^*)$ denotes the widths of $\mathcal{F}$ and $\mathcal{F}^*$, respectively. This is a substantial improvement over the complexity of a generic interior-point method.

[*] This research has been partially supported through the Singapore-MIT Alliance and an IBM PhD Fellowship.

[†] IBM T. J. Watson Research Center and MIT, 32-221, 1101 Kitchawan Road, Yorktown Heights, New York 10598, email: belloni@mit.edu

[‡] MIT Sloan School of Management, 50 Memorial Drive, Cambridge, MA 02142, USA, email: rfreund@mit.edu

1 Introduction and Main Results

Our notation is as follows: let K^* denote the dual of a convex cone $K \subset I\!R^k$, i.e., $K^* := \{z \in I\!R^k : z^T y \geq 0 \text{ for all } y \in K\}$. 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 [3]. Define the standard second-order cone in $I\!R^k$ to be $Q^k := \{y \in I\!R^k : \|(y_1, \dots, y_{k-1})\| \leq y_k\}$, where $\|\cdot\|$ denotes the Euclidean norm. Let $B(y, r)$ denote the Euclidean ball of radius r centered at y.

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

$$\mathcal{F} := \{y \in I\!R^n : \|My\| \leq g^T y\} = \{y \in I\!R^n : (My, g^T y) \in Q^{m+1}\}$$

and its dual cone $\mathcal{F}^*$. We make the following assumption about the data:

Assumption 1 $\text{rank}(M) \geq 2$ *and* $g \neq 0$*.*

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 i^{th} diagonal component of D and the i^{th} column of Q, respectively. By reordering the columns of Q we can presume that $D_1 \geq \cdots \geq D_n$ and $D_1, \dots, D_{n-1} \geq 0$. 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 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. We indicate how to amend our results and proofs to relax this hypothesis at the ends of Sections 2 and 3. Our main representation result is as follows:

Theorem 1 *Suppose that* $\mathcal{F}$ *is a regular cone. Then* $D_1, \dots, D_{n-1} > 0 > D_n$*, and:*
(i) $\mathcal{F} = \{y : y^T QDQ^T y \leq 0,\ y^T Q_n \geq 0\}$*;*
(ii) $\mathcal{F}^* = \{z : z^T QD^{-1}Q^T z \leq 0,\ z^T Q_n \geq 0\}$*;*
(iii) If $y \in \mathcal{F}$ *and* $\alpha \geq 0$*, then* $z := -\alpha QDQ^T y \in \mathcal{F}^*$*. 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* $y := -\alpha QD^{-1}Q^T z \in \mathcal{F}$*. 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 $\mathcal{S} := \{s \in I\!R^n : \sqrt{\sum_{j=1}^{n-1} D_i s_i^2} \le \sqrt{|D_n|} s_n\}$, so that $\mathcal{F}$ is the image of $\mathcal{S}$ under Q, and $\mathcal{F} = \{y : \sqrt{\sum_{i=1}^{n-1} D_i (Q_i^T y)^2} \le \sqrt{|D_n|} Q_n^T y\}$ and $\mathcal{F}^* = \{z : \sqrt{\sum_{i=1}^{n-1} (1/D_i)(Q_i^T z)^2} \le \sqrt{1/|D_n|} Q_n^T z\}$. 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 $M^T M - gg^T$. 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 $\{\alpha Q_n : \alpha \ge 0\}$. Last of all, note that $-\mathcal{F} = \{y : y^T Q D Q^T y \le 0,\ y^T Q_n \le 0\}$ and $-\mathcal{F}^* = \{z : z^T Q D^{-1} Q^T z \le 0,\ z^T Q_n \le 0\}$.*

It turns out that the eigendecomposition of $M^T M - gg^T = QDQ^T$, 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 QDQ^T and $QD^{-1}Q^T$ by $M^T M - gg^T$ and $(M^T M - gg^T)^{-1}$. Under the further hypothesis that $M^T M$ is invertible, the theorem can be restated as follows:

Corollary 1 *Suppose that $\mathcal{F}$ is a regular cone and* $\operatorname{rank}(M^T M) = n$. *Then:*
(i) $\mathcal{F} = \{y : \sqrt{y^T (M^T M) y} \le g^T y\}$;
(ii) $\mathcal{F}^* = \{z : \sqrt{z^T (M^T M)^{-1} z} \le \frac{g^T (M^T M)^{-1} z}{\sqrt{g^T (M^T M)^{-1} g - 1}}\}$;
(iii) If $y \in \mathcal{F}$ and $\alpha \ge 0$, then $z := -\alpha(M^T M - gg^T) y \in \mathcal{F}^*$. 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 $\alpha \ge 0$, then $y := -\alpha \left[(M^T M)^{-1} - \frac{(M^T M)^{-1} g g^T (M^T M)^{-1}}{g^T (M^T M)^{-1} g - 1}\right] z \in \mathcal{F}$. 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 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.

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

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

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 re-scaled perceptron algorithm in [1]: 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 [5] showed how to combine binary search and Newton's method to obtain double-logarithmic complexity. Using the representation results above, and extending ideas from [5], 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:

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

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 $M^TM - gg^T$, and it is straightforward to derive:

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

The main complexity result, which is proved in Section 3, is:

Theorem 2 *Suppose that $\mathcal{F}$ is a regular cone, and $x \in I\!R^n$ 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.*

Note that this is a substantial improvement over the complexity of a generic interior-point method which is $O(mn^2(\ln(1/\sigma) + \ln(1/\min\{\tau_{\mathcal{F}}, \tau_{\mathcal{F}^*}\})))$. We note also that the assumption that $\mathcal{F}$ is regular can be relaxed with no loss of strength of the results herein, but with substantial expositional overhead. These matters are discussed at the end of Section 3.

2 Proofs of Representation Results

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

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

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 $\bar{y} \in$ **int** $\mathcal{F}$. Using the singular-value decomposition, we can write $M = PR^T$ where $P \in I\!R^{m\times r}$, $R \in I\!R^{n\times r}$ $P^TP = I$ and $R^TR = E$ for some positive diagonal matrix E of rank $r = \mathrm{rank}(M)$. 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. Substituting $M = PR^T$ and squaring the previous equation and rearranging terms yields $2\beta(d^T RR^T \bar{y} - \bar{y}^T gg^T d) + \beta^2(d^T RR^T d - d^T gg^T d) = 0$, which is only true if $g^T d = 0 \Rightarrow R^T d = 0$. This in turn means that $\mathrm{rank}(R) = 1$, and so $\mathrm{rank}(M) = 1$, violating Assumption 1. Therefore there exists y satisfying $\|My\| < g^T y$. ■

The following straightforward characterization of $\mathcal{F}^*$, which was presented in more general form in [1], is included here for completeness.

Proposition 2 *Let* $\mathcal{T} = \left\{ M^T\lambda + g\alpha : \|\lambda\| \leq \alpha \right\}$. *Then* $\mathcal{F}^* = \mathbf{cl}\ (\mathcal{T})$.

Proof: ($\subseteq$) Let $\|\lambda\| \leq \alpha$. Then for every $x \in \mathcal{F}$ it follows that $\lambda^T Mx + \alpha g^T x \geq \alpha g^T x - \|Mx\|\|\lambda\| \geq 0$. Thus $\mathcal{T} \subset \mathcal{F}^*$, whereby $\mathbf{cl}\ (\mathcal{T}) \subseteq \mathcal{F}^*$ since $\mathcal{F}^*$ is closed. ($\supseteq$) Assume that there exists $y \in \mathcal{F}^* \backslash \mathbf{cl}\ (\mathcal{T})$. Thus there exists $h \neq 0$ satisfying $h^T y < 0$ and $h^T w \geq 0$ for all $w \in \mathbf{cl}\ (T)$. Notice that $\lambda^T Mh + \alpha g^T h \geq 0$ for all λ, α satisfying $\|\lambda\| \leq \alpha$, which implies that $\|Mh\| \leq g^T h$, and so $h \in \mathcal{F}$. On the other hand, since $y \in \mathcal{F}^*$, it follows that $h^T y \geq 0$, contradicting $h^T y < 0$. ■

The lack of closure of $\mathcal{T}$ can arise easily. Let $M = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$ and $g = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. In this case, $\mathcal{T} = \{(-\lambda_1 + \alpha, \lambda_2) \mid \|(\lambda_1, \lambda_2)\| \leq \alpha\}$. 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.

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 re-scale (M, g) by $1/\sqrt{|D_n|}$ in order to conveniently satisfy $D_n = -1$. (i) Define $\mathcal{H} := \{y : y^T QDQ^T y \leq 0,\ y^T Q_n \geq 0\}$. We first prove that $\mathcal{F} \subset \mathcal{H}$. If $y \in \mathcal{F}$, then $0 \geq y^T(M^T M - gg^T)y = y^T QDQ^T y$. We now prove that $y^T Q_n \geq 0$. Define $\lambda := -MQ_n$ and $\alpha = g^T Q_n$, whereby $\alpha \geq 0$ by the presumption above. Then

$$\|\lambda\| = \sqrt{(Q_n)^T M^T M Q_n} = \sqrt{(Q_n)^T(QDQ^T + gg^T)Q_n} = \sqrt{(g^T Q_n)^2 - 1} < |g^T Q_n| = g^T Q_n = \alpha\ ,$$

which also shows that $g^T Q_n \geq 1$. Direct arithmetic substitution shows that $M^T\lambda + g\alpha = Q_n$, whereby we have $y^T Q_n = y^T(M^T\lambda + g\alpha) \geq -\|My\|\|\lambda\| + g^T y\alpha \geq 0$. This shows that $y \in \mathcal{H}$. Next suppose that $y \in \mathcal{H}$. Then $y^T M^T My \leq (g^T y)^2$, whereby $y \in \mathcal{F}$ unless $g^T y < 0$. Supposing this is the case, it follows that $-g^T y \geq \|My\|$, and using the values of λ, α above, we have

$$0 \leq y^T Q_n = y^T M^T \lambda + y^T g\alpha \leq \|\lambda\|\|My\| - \alpha\|My\| = -\|My\|(\alpha - \|\lambda\|) \leq 0.$$

This then implies that $\|My\| = 0$ (since we showed above that $\alpha - \|\lambda\| > 0$), and hence $g^T y = 0$ as well since all inequalities above are then equalities. This contradiction establishes that $g^T y \geq 0$ and hence $y \in \mathcal{F}$, completing the proof of (i).

(ii) Having established (i), suppose that $D_i = 0$ for some $i \in \{1, \ldots, n-1\}$. Then $(\theta Q_i)^T QDQ^T(\theta Q_i) = 0$ 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 $\mathcal{J} := \{z : z^T Q D^{-1} Q^T z \le 0,\ z^T Q_n \ge 0\}$. Suppose that $z \in \mathcal{J}$ and $y \in \mathcal{F}$, in which case

$$\begin{aligned} y^T z = y^T Q Q^T z &= \textstyle\sum_{i=1}^{n-1} D_i^{\frac{1}{2}}(Q_i^T y) D_i^{-\frac{1}{2}}(Q_i^T z) + y^T Q_n z^T Q_n \\ &\ge -\sqrt{\textstyle\sum_{i=1}^{n-1} D_i (Q_i^T y)^2}\sqrt{\textstyle\sum_{i=1}^{n-1} D_i^{-1}(Q_i^T z)^2} + y^T Q_n z^T Q_n \ge 0 \end{aligned}$$

where the first inequality is an application of the Cauchy-Schwartz 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 $\mathcal{F} = \{y : \sqrt{y^T \bar{Q}\bar{D}\bar{Q}^T y} \le Q_n^T y\} = \{y : \|\bar{D}^{\frac{1}{2}}\bar{Q}^T y\| \le Q_n^T y\}$, and using Proposition 2 we know that $\mathcal{F}^* = \textbf{cl}\ \mathcal{T}$ where $\mathcal{T} = \{\bar{Q}\bar{D}^{\frac{1}{2}}\lambda + Q_n\alpha : \|\lambda\| \le \alpha\}$. Let $z \in \mathcal{T}$ where $z = \bar{Q}\bar{D}^{\frac{1}{2}}\lambda + Q_n\alpha$ and $\|\lambda\| \le \alpha$. Then

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

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

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

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

since $y \in \mathcal{F}$ ($y \in \partial\mathcal{F}$) implies $y^T Q D Q^T y \le (=)\ 0$, 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:

Proposition 3 *Suppose that* **int** $\mathcal{F} \neq \emptyset$ *and* $\mathrm{rank}(M^T M) = n$. *Then* $g^T(M^T M)^{-1} g > 1$ *and* $\bar{y} := (M^T M)^{-1} g \in$ **int** $\mathcal{F}$.

Proof: Let $\alpha := g^T(M^T M)^{-1} g > 0$ 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 re-scale $\hat{y}$ if necessary so that $g^T\hat{y} = \alpha$. Notice that $\bar{y}$ optimizes the function $f(y) = y^T M^T M y - 2g^T y$ whose optimal objective function value is $-\alpha$. Therefore

$$-\alpha \le \hat{y}^T M^T M \hat{y} - 2g^T\hat{y} < \alpha^2 - 2\alpha \ ,$$

which implies that $\alpha^2 > \alpha > 0$ and hence $\alpha > 1$. Next observe that $\|M\bar{y}\| = \sqrt{\bar{y}^T M^T M \bar{y}} = \sqrt{\alpha} < \alpha = g^T\bar{y}$, whereby $\bar{y} \in$ **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:

$$Q D^{-1} Q^T = (M^T M - g g^T)^{-1} = (M^T M)^{-1} - \frac{(M^T M)^{-1} g g^T (M^T M)^{-1}}{g^T (M^T M)^{-1} g - 1}$$

together with the fact from Proposition 3 that $g^T(M^TM)^{-1}g > 1$.

It remains to prove (ii). Let $\mathcal{K} := \{z \in I\!R^n : z^TQD^{-1}Q^Tz \le 0\}$. Then from Theorem 1 we have $\mathcal{K} = \mathcal{F}^* \cup -\mathcal{F}^*$. Let $\bar{y} = (M^TM)^{-1}g$ and note that $\bar{y} \in$ **int** $\mathcal{F}$ from Proposition 3. Define $\mathcal{H} := \{z \in I\!R^n : \bar{y}^Tz \ge 0\}$, and note that $\mathcal{H} \cap \mathcal{F}^* = \mathcal{F}^*$ and $\mathcal{H} \cap -\mathcal{F}^* = \{0\}$. Therefore $\mathcal{F}^* = \mathcal{K} \cap \mathcal{H} = \{z \in I\!R^n : z^TQD^{-1}Q^Tz \le 0,\ g^T(M^TM)^{-1}z \ge 0\}$. Using the Sherman-Morrison formula we obtain:

$$\mathcal{F}^* = \left\{ z^T \left((M^TM)^{-1} - \frac{(M^TM)^{-1}gg^T(M^TM)^{-1}}{g^T(M^TM)^{-1}g - 1} \right) z \le 0,\ g^T(M^TM)^{-1}z \ge 0 \right\}$$

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 "$z^TQD^{-1}Q^Tz \le 0$" in (ii) as "$\sum_{i \in N} D_i(Q^Tz)_i^2 \le 0$, $(Q^Tz)_i^2 = 0$ for $i \in Z$," and "$y := -\alpha QD^{-1}Q^Tz$" in (iv) as "$Q_i^Ty := -\alpha D_i^{-1}Q_i^Tz$ for $i \in N$ and Q_i^Ty is set* arbitrarily *for $i \in N$."*

Remark 3 The case when $\mathrm{rank}(M) = 1$. *In this case $M = fc^T$ for some f, c and $\|My\| = \|f\||c^Ty|$ for any y. This implies that $\mathcal{F} = \{y \in I\!R^n : (g - \|f\|c)^Ty \ge 0,\ (g + \|f\|c)^Ty \ge 0\}$. Therefore $\mathcal{F}$ is the intersection of either one or two halfspaces.*

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):

$$\begin{array}{l} y - \theta z = x \\ y \in \mathcal{F} \\ z \in \mathcal{F}^* \\ \|z\| \le 1 \\ \theta \ge 0,\ \theta\|z\| = \theta\ . \end{array} \tag{2}$$

Examining (2), 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 $G = \|y - x\| + x^Tz$. We also consider the following pair of conic problems that are "polar" to (1):

$$\begin{array}{llllll} \mathcal{P}^\circ : f^* & := \min_x & \|s - x\| & \qquad \mathcal{D}^\circ : f^* & := \max_w & -x^Tw \\ & \text{s.t.} & s \in -\mathcal{F}^* & & \text{s.t.} & \|w\| \le 1 \\ & & & & & w \in -\mathcal{F}\ , \end{array} \tag{3}$$

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

$$\begin{aligned} & s - \rho w = x \\ & s \in -\mathcal{F}^* \\ & w \in -\mathcal{F} \\ & \|w\| \leq 1 \\ & \rho \geq 0,\ \rho\|w\| = \rho\ ; \end{aligned} \tag{4}$$

here x is decomposed into $x = s - \rho w$ where now (s, w) is feasible for the problems (3) and $-\rho w \in \mathcal{F}$ and $s \in -\mathcal{F}^*$. Let G° denote the duality gap for (3), namely $G^\circ = \|s - x\| + x^T w$.

It is a straightforward exercise to show that conditions (2) together with the complementarity condition $y^T z = 0$ constitute necessary and sufficient optimality conditions for (1), and similarly (4) together with $s^T w = 0$ are necessary and sufficient for optimality for (3). Furthermore, solutions of (2) and (4) tranform to one-another:

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

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

Proposition 4 *Suppose (y, z, θ) satisfy (2) and (s, w, ρ) satisfy (4). Then (y, z) and (s, w) are feasible for their respective problems with respective duality gaps:*
(i) $G = y^T z$,
(ii) $G^\circ = s^T w$.
Furthermore,
(iii) if (y, z) is optimal for (1), then $t^* = \theta$
(iv) if (s, w) is optimal for (3), then $f^* = \rho$
(v) $(t^*)^2 + (f^*)^2 = \|x\|^2$.

Proof: To prove (i), observe $y^T z = z^T x + \theta\|z\|^2 = z^T x + \theta\|z\| = z^T x + \|y - x\| = G$, 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 (2) and $y^T z = 0$ if and only if (y, z) is optimal for (1), in which case it is easy to verify that $(s, w, \rho) \leftarrow (-\theta z, -y/\|y\|, \|y\|)$ satisfy (4) and (s, w) is optimal for (3). Therefore $\|x\|^2 = (y - \theta z)^T (y - \theta z) = y^T y + \theta^2 = \rho^2 + \theta^2 = (f^*)^2 + (t^*)^2$. ■

Proposition 5 *If $Q_n^T x \leq 0$, then $t^* \geq \tau_{\mathcal{F}^*}\|x\|$.*

Proof: We assume for the proof that $\|x\| = 1$, since t^*, f^* scale positively with $\|x\|$. 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 $\|x - c\| = \sqrt{x^T x + 2\frac{t^*}{f^*} Q_n^T x + \frac{t^{*2}}{f^{*2}} Q_n^T Q_n} \leq \sqrt{1 + \frac{t^{*2}}{f^{*2}}} = \frac{1}{f^*}$. Therefore $\frac{1}{f^*\|x - c\|} \geq 1$.

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

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

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

Proposition 6 *Given x satisfying $\|x\| = 1$ and $Q_n^T x \leq 0$, suppose that (s, w, ρ) satisfies (4) with duality gap $G^\circ \leq \sigma\tau_{\mathcal{F}^*}/2$ for (3), where $\sigma \leq 1$. Consider the assignment: $(y, z, \theta) \leftarrow (-\rho w, -s/\|s\|, \|s\|)$ (with the necessary modification that $y = 0$ if $s = 0$). Then (y, z, θ) satisfies (2), with duality gap $G \leq \sigma$ for (1).*

Proof: Note that $y^T z = \frac{(w^T s)\rho}{\|s\|} \leq \frac{\sigma\tau_{\mathcal{F}^*}\rho}{2\|s\|}$ and we have the following relations: (i) $w^T s \leq \sigma\tau_{\mathcal{F}^*}/2 \leq 1/2$, (ii) $\|s\| = \theta = \|y - x\| \geq t^* \geq \tau_{\mathcal{F}^*}$ from Proposition 5, and (iii) $\rho = \|s - x\| = s^T w - w^T x \leq 1/2 + f^* \leq 3/2$ from Proposition 4. Therefore $y^T z \leq \frac{\tau_{\mathcal{F}^*}\sigma}{2}\frac{3}{2}\frac{1}{\tau_{\mathcal{F}^*}} \leq \sigma$. ■

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 duality gap at most σ where $\sigma \leq 1$. Armed with Propositions 4, 5, and 6, we now show how to compute a feasible solution (y, z) of (1) with duality gap $G \leq \sigma$. Our method is best understood with the help of Figure 1. We know from Section 3.1 and the conditions (2) and/or (4) 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 $L_{\mathcal{F}}^+ := \{y \in I\!R^n : Q_n^T y \geq 0\}$ 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 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 $G \leq \sigma$. For each of the three regions in $L_{\mathcal{F}}^-$ we will instead work with the problem pair (3) and show how to compute a feasible solution (w, s) of (3) with duality gap $G^\circ \leq \sigma\tau_{\mathcal{F}^*}/2$, whereby from Proposition 6 we obtain a feasible solution (y, z) of (1) with duality gap $G \leq \sigma$. 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. Let $s = Qx$, therefore $x = Q^T s$ and $s_i = Q_i^T x$, $i = 1, \ldots, n$, and $\|s\| = 1$. For $x \in L_{\mathcal{F}}^+ \setminus \mathcal{F}$, define:

$$\varepsilon_{\mathcal{P}} = \varepsilon_{\mathcal{P}}(x) := \frac{Q_n^T x\sqrt{|D_n|}}{\sqrt{\sum_{i=1}^{n-1} D_i (Q_i^T x)^2}} \ , \qquad (5)$$

and notice that $x \in L_{\mathcal{F}}^+$ implies $\varepsilon_{\mathcal{P}} \geq 0$ and $x \notin \mathcal{F}$ 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

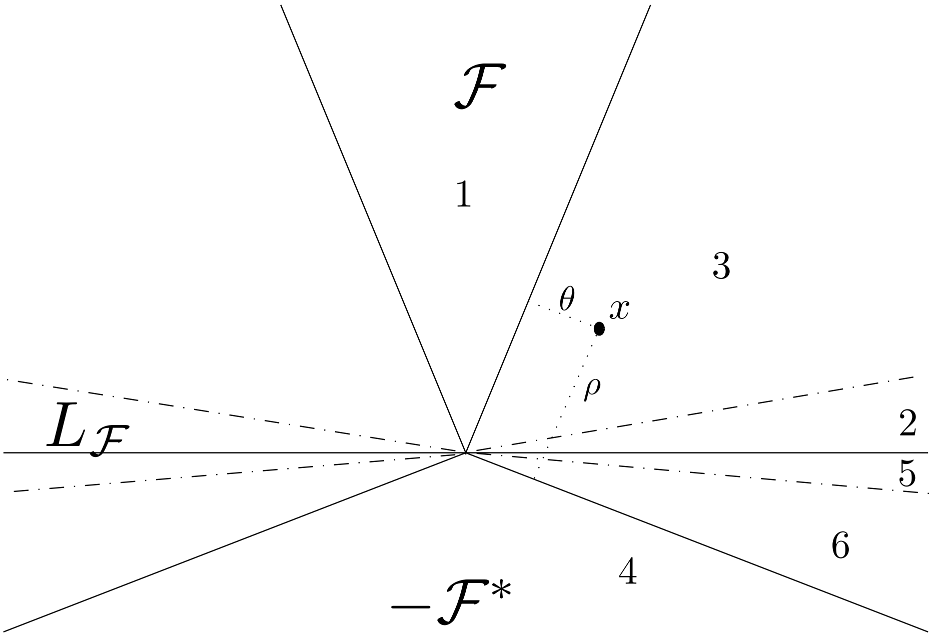

Figure 1: The geometry of the sets $\mathcal{F}$, $-\mathcal{F}^*$, and $L_{\mathcal{F}}$, and the six cases. The central axes of $\mathcal{F}$ and $-\mathcal{F}^*$ are the rays generated by $\pm Q_n$, respectively, which are orthogonal to the hyperplane $L_{\mathcal{F}}$. The regions corresponding to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

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 (2) with $y^T z = 0$ whereby from Proposition 4 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:

$$\begin{array}{rcl} [I + 1/|D_n| D] Q^T \hat{y} & = & Q^T x - e_n Q_n^T x \\ Q_n^T \hat{y} & = & 0 \end{array} \qquad (6)$$

where $e_n = (0, \ldots, 0, 1) \in I\!R^n$. Notice that the last row of the first equation system has all zero entries. Therefore this system is not over-determined, 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 $\alpha := \sqrt{\hat{y}^T Q D Q^T \hat{y}} / \sqrt{|D_n|}$, and then make the following assignments to variables:

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

Proposition 7 *Suppose that $\|x\| = 1$, $\sigma \leq 1$, and $\varepsilon_{\mathcal{P}} \leq \bar{\varepsilon} < 1$, and (y, z, θ) are computed according to Case 2 above. Then (y, z, θ) is feasible for (2) with duality gap $G \leq \bar{\varepsilon}/\tau_{\mathcal{F}}$ for (1).*

Applying Proposition 7 using $\bar{\varepsilon} = \bar{\varepsilon}_{\mathcal{P}} := \sigma\tau_{\mathcal{F}}$ ensures that the resulting duality gap satisfies $G \leq \bar{\varepsilon}/\tau_{\mathcal{F}} = \sigma$. 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 7: It is easy to establish that $(Q_1^T x, \ldots, Q_{n-1}^T x) \neq 0$ 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:

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

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 (6) that

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

(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 (2). Therefore from Proposition 4 (y, z) is feasible for (1) with duality gap

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

■

Case 3: $Q_n^T x \geq 0$ 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:

$$f(\gamma) := x^T Q[I + \gamma D]^{-1} D [I + \gamma D]^{-1} Q^T x = \sum_{i=1}^{n} \frac{D_i (x^T Q_i)^2}{(1 + D_i \gamma)^2} \ , \tag{7}$$

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 $\gamma \in [0, 1/|D_n|)$. 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 4 how to combine 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:

$$\begin{array}{rcl} y & \leftarrow & Q\,[I + \gamma D]^{-1}\,Q^T x \\ \theta & \leftarrow & \gamma\sqrt{y^T Q D^2 Q^T y} \\ z & \leftarrow & -\gamma Q D Q^T y/\theta \end{array} \tag{8}$$

We now show that (y, θ, z) satisfy (2). 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 (2) is satisfied. From Proposition 4, (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 binary search and Newton's method, which is postponed to Section 4 where we will prove:

Proposition 8 *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 (2) 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 (2) with $y^T z = 0$ whereby from Proposition 4 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 (5):

$$\varepsilon_{\mathcal{P}^*} = \varepsilon_{\mathcal{P}^*}(x) := \frac{-Q_n^T x\sqrt{1/|D_n|}}{\sqrt{\sum_{i=1}^{n-1}(1/D_i)(Q_i^T x)^2}} \ , \tag{9}$$

and notice that $x \in L_{\mathcal{F}}^-$ implies $\varepsilon_{\mathcal{P}^*} \geq 0$ and $x \notin -\mathcal{F}^*$ 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 analog of Case 2, with $\mathcal{F}$ replaced by $-\mathcal{F}^*$ and the pair (1) replaced by (3). Therefore the methodology of Case 2 can be used to compute (s, w, ρ) satisfying (4) and hence (s, w) is feasible for (3). Applying Proposition 7 to the context of the polar pair (3) with $\bar{\varepsilon} = \bar{\varepsilon}_{\mathcal{P}^*}$, it follows that the duality gap for (3) will be $G^\circ = s^T w$ 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 (s, w, ρ) to (y, z, θ) using Proposition 6, 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 analog of Case 3, with $\mathcal{F}$ replaced by $-\mathcal{F}^*$ and the pair (1) replaced by (3). Therefore the methodology of Case 3 can be used to compute (s, w, ρ) satisfying (4) and hence (s, w) is feasible for (3). Applying Proposition 8 to the context of the polar pair (3) with $\bar{\varepsilon} = \bar{\varepsilon}_{\mathcal{P}^*}$ and $g = \sigma\tau_{\mathcal{F}^*}/2$, it follows that a solution (s, w, ρ) of (4) with duality gap $G^\circ \leq g = \sigma\tau_{\mathcal{F}^*}/2$ for (3) 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 (s, w, ρ) to (y, z, θ) using Proposition 6, we obtain (y, z) feasible for (1) with duality gap $G \leq \sigma$.

Proof of Theorem 2: The spectral decomposition of $M^T M = Q D Q^T$ 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 4 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 set $Q_i^T z = 0$, and work in the lower-dimensional problem in the subspace spanned by Q_i, $i \in N$.*

4 Proof of Proposition 8

This section is devoted to the proof of Proposition 8. Our algorithmic approach is motivated by Ye [5], and consists of a combination of binary search and Newton's method to approximately solve $f(\gamma) = 0$ for the function f given in (7). An alternate approach would be to use interpolation methods as presented and analyzed in Meldman [2], for which global quadratic convergence is proved but there is no complexity analysis of associated constants. While Proposition 8 indicates that a solution (y, z, θ) of (2) 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 8 hold. We start with a review of Smale's main result for Newton's method in [4].

4.1 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 [4] can be re-stated as follows.

Theorem 3 *(Smale [4]) Let g be an analytic function. If $\hat{\gamma}$ satisfies*

$$\sup_{k>1} \left| \frac{g^{(k)}(\hat{\gamma})}{k! g'(\hat{\gamma})} \right|^{1/(k-1)} \leq \frac{1}{8} \left| \frac{g'(\hat{\gamma})}{g(\hat{\gamma})} \right| \ , \tag{10}$$

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 $k \geq 1$.

4.2 Properties of $f(\gamma)$

We employ the change of variables $s = Q^T x$, whereby from the hypotheses of Proposition 8 we have $s_n > 0$, $s^T D s > 0$, 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:

$$f(\gamma) = s^T (I + \gamma D)^{-2} D s = \sum_{i=1}^{n} \frac{D_i s_i^2}{(1 + \gamma D_i)^2} \tag{11}$$

Lemma 1 *Under the hypotheses of Proposition 8 f has the following properties:*
(i) $f(0) > 0$, $\lim\limits_{\gamma \to 1/|D_1|} f(\gamma) = -\infty$, and f has a unique root $\gamma^* \in (0, 1/|D_1|)$

(ii) f is analytic on $(-1/D_1, 1/|D_n|)$ and for $k \geq 1$ the k^{th} derivative of f is

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

(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) $\frac{1-\varepsilon_{\mathcal{P}}}{|D_n|+\varepsilon_{\mathcal{P}} D_1} \leq \gamma^* \leq \frac{1-\varepsilon_{\mathcal{P}}}{|D_n|}$ *where* $\varepsilon_{\mathcal{P}}$ *is given by (5)*

(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

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

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

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

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

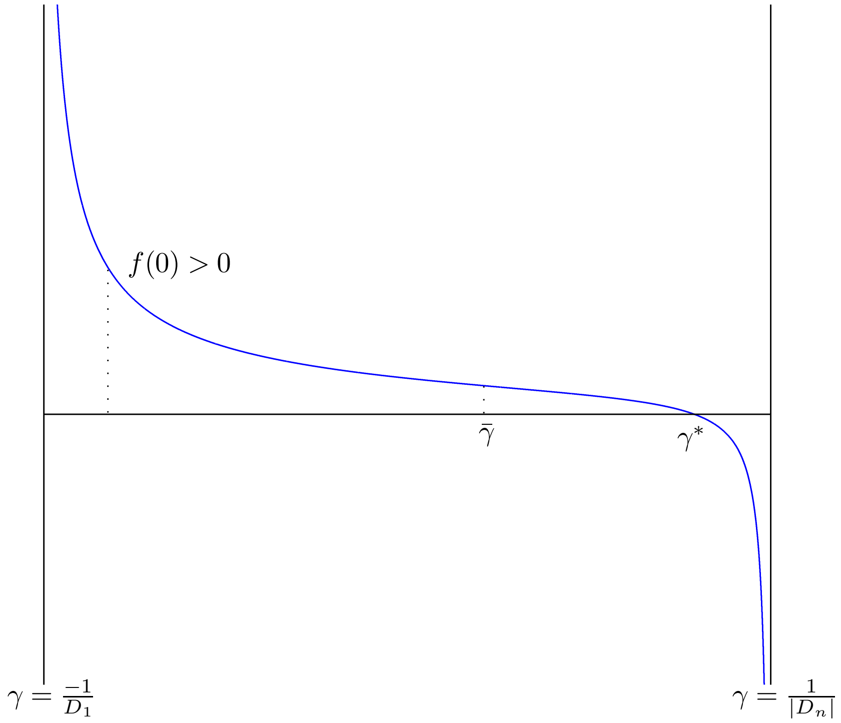

Figure 2: The function f on the interval $(-1/D_1, 1/|D_n|)$. Among many desirable properties, f is strictly decreasing, analytic, and has a unique root $\gamma^* \in (0, 1/|D_n|)$. Moreover, f is convex over $(-1/D_1, \bar{\gamma})$ and concave over $(\bar{\gamma}, 1/|D_n|)$, where $\bar{\gamma}$ is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

4.3 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 binary search to locate an approximate zero of f, motivated by the method of Ye [5]. Noticing from Remark 5 that the maximum in (ii) 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 $\left[\max\{0, M\}, \bar{U}\right]$. (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 $L^i := \frac{1}{D_1}\left(\left(\frac{13}{12}\right)^i - 1\right)$ 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 $R^i := \frac{1}{|D_n|} - \left(\frac{1}{|D_n|} - \bar{U}\right)\left(\frac{13}{12}\right)^i$ 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 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

$$\left| \frac{f(\hat{\gamma})}{f'(\hat{\gamma})} \right| = |\gamma^+ - \hat{\gamma}| \leq |\gamma^* - \hat{\gamma}| \leq b - a \ , \tag{12}$$

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

$$|f(\hat{\gamma})| \leq |f'(\hat{\gamma})||\gamma^* - \hat{\gamma}| \tag{13}$$

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

Lemma 2 *Under the hypotheses of Proposition 8 the intervals described herein have the following propeties:*
(i) the total number of left intervals and right intervals needed to cover $[0, \bar{U}]$ is $K_L := \left\lceil \frac{\ln(1/2) + 2\ln(1/\tau_{\mathcal{F}})}{\ln(13/12)} \right\rceil^+$ and $K_R := \left\lceil \frac{\ln(1/\bar{\varepsilon})}{\ln(13/12)} \right\rceil$, 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

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

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

$$\left(\frac{13}{12}\right)^{K_R} \leq \frac{13}{12\bar{\varepsilon}} \ . \tag{14}$$

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

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

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

$$\left(\frac{13}{12}\right)^{K_L} \leq \frac{13}{24\tau_{\mathcal{F}}^2} \ . \tag{15}$$

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 $k > 1$:

$$\frac{1}{8}\left|\frac{f'(\hat{\gamma})}{f(\hat{\gamma})}\right| \geq \frac{1/8}{b-a} = \frac{1/8}{(1/12)(a + 1/D_1)} \geq \frac{3}{2}\frac{D_1}{1+\hat{\gamma}D_1} \geq \left|\frac{f^{(k)}(\hat{\gamma})}{k!f'(\gamma_0)}\right|^{1/(k-1)} ,$$

where the first inequality uses (12), the second inequality uses $a \leq \hat{\gamma}$, and the third inequality uses Remark 5 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 $\hat{\gamma} = b$ and we have for all $k > 1$:

$$\begin{aligned}\frac{1}{8}\left|\frac{f'(\hat{\gamma})}{f(\hat{\gamma})}\right| &\geq \frac{1/8}{b-a} = \frac{1/8}{b - b(13/12) + \frac{1}{12|D_n|}} = \frac{1/8}{\frac{1}{12}(\frac{1}{|D_n|} - b)} = \frac{3}{2}\frac{|D_n|}{1 - b|D_n|} \\ &\geq \frac{3}{2}\frac{|D_n|}{1-\hat{\gamma}|D_n|} \geq \left|\frac{f^{(k)}(\gamma_0)}{k!f'(\gamma_0)}\right|^{1/(k-1)} ,\end{aligned}$$

where the first inequality uses (12), the second inequality uses $M \leq a \leq \hat{\gamma} \leq b$, and the third inequality uses Remark 5 and the fact that $\hat{\gamma} \geq M$ in conjunction with Lemma 1. Therefore $\hat{\gamma}$ is an approximate zero of f.

To prove (iii), for the right intervals

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

by the definition of K_R and the second inequality derives from (14). For the left intervals we can assume $M \geq 0$ (otherwise they are not constructed), in which case $D_1 \geq |D_n|$. In this case, we have

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

by the definition of K_L and the second inequality derives from (15).

■

Based on these properties, consider the following method for locating an approximate zero of f. Perform binary search on the endpoints of the intervals, testing the endpoints to locate an interval $[a, b]$ for which $f(a) \geq 0$ and $f(b) \leq 0$. 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 binary search on the left and right intervals there is no need to compute and evaluate f for all of the endpoints. 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.

4.4 Computing a Solution of (1) with Duality Gap at most σ

Under the hypotheses of Proposition 8, suppose that $[a, b]$ is one of the constructed intervals, and $f(a) \geq 0$ and $f(b) \leq 0$. 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 $\gamma \geq \gamma^*$ and consequently $f(\gamma) \leq 0$. Then the analysis in Case 3 shows that the assignment (8) 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 8 and the set-up 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 (8) will be feasible for (1) with duality gap at most g.

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

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

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

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

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

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

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 k^{th} iterate we have

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

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 (8) is not necessarily feasible for (1). However, invoking Theorem 3 we know that $\gamma_k + 2(1/2)^{2^{k-1}} |\gamma_1 - \gamma_0| \ge \gamma^*$, and 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 (8) 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*

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

Under the hypotheses of Proposition 8 and the set-up 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 (8) 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,

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

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

$$1 + \gamma_k D_n \leq 2(1 + (\gamma_k + \delta)D_n) \leq 2(1 + tD_n) \ \text{ for all } \ t \in [\gamma_k, \gamma_k + \delta] \ . \tag{17}$$

We also have from (16) that $\delta \leq 1/D_1 \leq 1/D_i \leq 1/D_i + \gamma_k$ for $i = 1, \ldots, n-1$, hence

$$1 + \gamma_k D_i + \delta D_i \leq 2(1 + \gamma_k D_i) \ , \ i = 1, \ldots, n-1 \ . \tag{18}$$

The duality gap of the assignment (8) 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 right-most expression. For the numerator we have:

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

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

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

and it follows that $-f(\gamma_k + \delta) \leq 8\delta|f'(\gamma_k)|$. To bound the denominator, simply notice from (18) 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

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

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

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

where the last inequality uses the second inequality of (16). ■

Proof of Proposition 8: Note from the discussion at the end of Section 4.3 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. ■

References

[1] 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, 2006.

[2] A. Meldman. A unifying convergence analysis of second-order methods for secular equations. *Mathematics of Computation*, 66(217):333–344, 1997.

[3] R. T. Rockafellar. *Convex Analysis.* Princeton University Press, Princeton, New Jersey, 1970.

[4] S. Smale. Newton's method estimates from data at one point. *The Merging of Disciplines: New Directions in Pure, Applied and Computational Mathematics*, pages 185–196, 1986.

[5] Y. Ye. A new complexity result for minimizing a general quadratic function with a sphere constraint. *Recent advances in global optimization*, pages 19–31, 1992.